Miro Enev*, Alex Takakuwa, Karl Koscher, and Tadayoshi Kohno

Automobile Driver Fingerprinting

Abstract: Today's automobiles leverage powerful sensors and embedded computers to optimize efficiency, safety, and driver engagement. However the complexity of possible inferences using in-car sensor data is not well understood. While we do not know of attempts by automotive manufacturers or makers of after-market components (like insurance dongles) to violate privacy, a key question we ask is: *could* they (or their collection and later accidental leaks of data) violate a driver's privacy? In the present study, we experimentally investigate the potential to identify individuals using sensor data snippets of their natural driving behavior. More specifically we record the in-vehicle sensor data on the controller-area-network (CAN) of a typical modern vehicle (popular 2009 sedan) as each of 15 participants (a) performed a series of maneuvers in an isolated parking lot, and (b) drove the vehicle in traffic along a defined $\sim$ 50 mile loop through the Seattle metropolitan area. We then split the data into training and testing sets, train an ensemble of classifiers, and evaluate identification accuracy of test data queries by looking at the highest voted candidate when considering all possible one-vs-one comparisons. Our results indicate that, at least among small sets, drivers are indeed distinguishable using only in-car sensors. In particular, we find that it is possible to differentiate our 15 drivers with 100% accuracy when training with all of the available sensors using 90% of driving data from each person. Furthermore, it is possible to reach high identification rates using less than 8 minutes of training data. When more training data is available it is possible to reach very high identification using only a single sensor (e.g., the brake pedal). As an extension, we also demonstrate the feasibility of performing driver identification across multiple days of data collection.

Keywords: Security and Privacy, Vehicle Sensors, Driver Identification, Machine Learning

DOI 10.1515/popets-2015-0029
Received 2015-04-15; revised 2015-07-15; accepted 2015-07-15.

***Corresponding Author: Miro Enev:** University of Washington, E-mail: miro@cs.washington.edu
Alex Takakuwa: University of Washington, E-mail: alextaka@cs.washington.edu
Karl Koscher: University of California, San Diego, E-mail: supersat@cs.ucsd.edu

1 Introduction

Cars have evolved past their purely mechanical roots into "smart" cyberphysical platforms built on sophisticated sensing and computing systems which coordinate to improve safety, efficiency, and engagement. More recently, vehicles have also gained the ability to communicate with car manufacturers, 3rd parties, and the road infrastructure via mobile telecommunications which enables two-way streaming of [sometimes pre-processed or subsampled] data.

On one hand, computation, sensing, and connectivity technologies are unlocking new levels of innovation in the automotive marketplace which aim to improve the driving experience. On the other hand, these developments have been met with growing concern from consumers and policy makers given the potential conflicts of interest over data ownership and privacy [4, 7].

While there has been significant effort spent on maximizing the utility of the data streams generated from connected cars, relatively little is known about the privacy implications and potential for misuse. In the current work, we aim to provide an experimental grounding for the policy discussion necessary to advance the balance between utility and privacy in vehicle data sharing scenarios.

To this end, we investigate the potential to perform unintended privacy breaking inferences using data collected from sensors in a typical car (2009 sedan). The data we use as the basis for our experiments *already* exists on the car's internal network, and as we describe in Section 2, drivers are increasingly opting to share/stream this data to 3rd parties (including insurance companies and start-ups). While we expect that most 3rd party collectors of vehicle data are trustworthy, we seek to evaluate the potential for abuse of this trust by measuring the potential to extract private information from the sensor data (note that even well intentioned data collectors can expose individuals to risks due to the possibility for subsequent breaches or subpoenas).

Tadayoshi Kohno: University of Washington, E-mail: yoshi@cs.washington.edu

Experiments

We logged the data streams from 16 sensors that already broadcast over the car's internal computer network as 15 drivers (7 female, 8 male) navigated through (1) three laps around a closed-course section composed of parking and weaving maneuvers, and (2) ~ 50 miles of open road driving. We collect the open road data along the same pre-defined course and start each recording session at the same time of day to normalize against traffic conditions.

Using the open and closed road sensor recordings as a database, we investigated the potential of identifying the driver from snippets of query sensor data unseen in training (using a classifier ensemble with cross-validation 90%-10% training-testing splits).

Goal

The goal of our experiments is to evaluate the accuracy when trained machine classifiers are asked to identify/fingerprint the driver from queries which vary in both recording duration and the number of available sensors streams. In Section 3 (Threat Model) we discuss scenarios for which driver fingerprinting could lead to important privacy compromises. Note that we use the terms "identification" and "fingerprinting" not to imply uniqueness, but rather to imply a measure that would allow reidentification of a driver among a set.

Results

Initially, we expected some differentiation to be possible between drivers since various classes of human behaviors have been shown to have between subject variability; however, we were surprised to see that very high identification accuracy was possible among 15 people with very short amounts of collected data and/or sensors. Several highlights of our findings include:

- 100% driver ID among 15 drivers is possible using 15 sensors and the entire database of driving data.
- 100% driver ID among 15 drivers is possible using just the brake pedal and the entire database for training.
- 100% ID among 15 drivers is possible given short training datasets (8 mins, 15 mins, 1 hour) and multiple sensors; 87% accuracy is achievable using a single sensor (brake pedal) and only the first 15 minutes of open-road driving as a training database.

Implications

Our results suggest that vehicle sensor data has significant potential for enabling powerful inferences–some of which may be undesirable from the perspective of the driver whose actions were captured in the data streams. While we do not expect to diminish consumer appetite for interactive, personalized, and connected experiences in their vehicles we believe that our work sheds new light on the potential privacy risks with computerized automobiles. Our work is specifically relevant to manufacturers, drivers, and participating stakeholders in the existing marketplace which capture messages (sensor data) sent over the car's internal computer network while the car is being driven. Lastly, we hope that our efforts can help inform the design of policy and mechanisms to balance the utility and privacy tensions emerging in modern automotive contexts.

2 Background

During the first century following their invention, cars were exclusively non-digital technologies. It was not until the early 1980s, with the rise of microprocessors and the introduction of the California Clean Air act that electronic control units became widely integrated in the fabric of the vehicle. Since then, the number of digital components has grown dramatically with modern cars typically containing on the order of hundreds of embedded CPUs and sensors linked through sophisticated internal networks.

Most recently, the data generated by the internal networks of the vehicle is also being connected to remote 3rd parties. From a hardware perspective this connectivity is accomplished either via built in telematics units (2G/GSM, 3G, 4G, and LTE), via cellular connectivity built into data collection devices (e.g., insurance dongles), via 'carried in' connection solutions which may rely on a driver's smart-phone for internet connectivity, and in some instances OBD-II dongles can be removed from the vehicle and connected to a computer or returned to the provider.

The potential to monetize the vehicle's sensor data streams and the ubiquity of connectivity options have been driving factors toward a novel data market for vehicular data. Below we describe some of the participants in the rapidly growing car data-sharing economy.

Usage Based Insurance

Many insurance companies offer "pay as you drive" discounts which enable rate reductions for consistently "safe" driving behavior (e.g., Progressive's Snapshot, State Farm's In-Drive, Allstate's Drive Wise, etc.). Interested drivers opt-in to installing a dongle attachment to the OBD-II port which either locally analyzes sensor data or transmits the data for upstream processing. The features which insurances companies use to make rate reduction judgments are not always public, but some companies make aspects of their policy known – Progressive, for instance, claims that it considers bad behaviors to be instances of hard braking, driving between midnight and 4am, and driving at high speed [10, 13].

Fleet Monitoring

Commercial fleet operators often install trackers in their vehicles (OBD-II dongle connector with built in GPS and telematics) which offer remote web-based reporting tools as well as customized triggers that send alerts and notifications whenever drivers idle excessively, drive over the speed limit, or stray beyond pre-defined geographical areas known as "geofences" [11].

Make Every Car Smart

Startups like Automatic and Zubie are promising to turn any car into a "smart car" by beaming the data from the vehicle's internal network to the driver's phone (via OBD-II dongle and Bluetooth – ~100USD) where it can be processed within an installed application. Some example functions are remembering where you parked, alerting your carpool/friends about your arrival time, novice driver coaching/tracking, gamification (break your fuel efficiency record for home commutes), switching your phone to do-not-disturb mode, deciphering diagnostic codes, and many more. A very compelling aspect of these technologies is that they can make any OBD-II equipped vehicle (manufactured after 1995) a participant in the driver's technology ecosystem and enable rich interactions with other smart devices and apps (e.g., warm up home when leaving work [via Nest], log my trips to a Google Spreadsheet, text my spouse for item requests when I'm at the grocery store etc.) [2, 14].

The Car as a Mobile Operating System

Car manufacturers are also offering powerful integrated infotainment and assistance systems that synthesize sensor data and unlock powerful functions with telematics and remote connectivity (GM MyLink, Ford Sync, Mercedes DriveStyle). In addition car manufactures are partnering with phone manufactures to enable seamless interactions with existing mobile platforms. For instance Apple's CarPlay (expected to launch on 40 vehicle brands in 2015) enables Apple phones to take over the display panel interface and run a light version of iOS which allows voice control (Siri) interactions with the audio and audio-streaming services, messages, maps and phone functions of their iPhone while locking the driver out of more distracting functions. [1]

3 Threat Model

We consider a threat model in which the adversary has access to information *already* being communicated on the car's internal computer network. In our model, an adversary is thus a passive eavesdropper on the car's internal computer network. Although this seems to imply that some malicious entity has access to the car, we note that the number of 3rd party companies which offer solutions based on analyzing vehicle OBD-II port data is growing (e.g, Automatic.com, Mojo.io), and that present/future systems could upload raw data to servers where it could be compromised or abused. So although we use the term "adversary," we note that the entity that we call the "adversary" may not be intentionally malicious. For example, the "adversary" may be collecting and storing information sent on the car's internal computer network for debugging or other purposes but, because of a data breach or subpoena, later exposes that data to a different party who *does* wish to use the data to compromise the driver's privacy. We take this perspective because we do not want to imply that existing automotive manufacturers or after-market vendors are being malicious, but–as we discuss below–we do note that there may be incentives for them to try to fingerprint drivers.

3.1 Fingerprinting

We chose "driver fingerprinting" as the metric for privacy. We say that an adversary compromises a driver's privacy if he or she can–based on some information collected on the car's internal computer network–fingerprint (or identify) the driver. Our use of the terms "fingerprint" and "identify" are, however, distinct from common English language uses of the terms. Rather,

like device or operating system fingerprinting, a driver fingerprint is one that would allow an adversary to (for example) re-identify a driver among a (possibly small) set of other candidate drivers.

3.2 Data Sources

Our understanding is that existing components (e.g., after-market auto insurance dongles, built-in radios like the telematics unit, or phone interconnected dashboards like Apple's CarPlay; see also Section 2) can access a car's internal computer network and read data for various purposes. Because the communications on a car's internal computer network are readily accessible to any other component on that network, and because existing after-market and built-in components are already reading those communications, we believe that it is reasonable to assume that in the future more and more parties will gain access to the car's internal network communications. Our goal is to understand whether, even with the limited data *already* available on the car's internal computer network, an adversary might be able to compromise the driver's privacy. For simplicity, in the remainder of the threat model we assume a dongle is reading the data, though there may be many other ways to obtain data from the internal network, for example through built-in or after-market components.

3.3 Scenarios

Our goal is to understand the potential privacy implications of making the data from a car's internal computer network available to a potential adversary. Thus, while most of our work focuses on a technical study of the topic, we do present several example scenarios in which the capabilities that we explore could be used against a car owner or another driver of a car.

- Suppose a red-light camera captures a photo of a car driving through a red light. Also suppose that the driver's face is obscured. The owner, Alice, says that she was not driving the car but rather loaned it to Bob. The car has an after-market insurance dongle connected to the car's internal computer network. The police, perhaps in collaboration with the car owner's insurance company or via a subpoena, obtain access to the data stored on that dongle (or perhaps the data is uploaded to the cloud automatically). Using that data, the police could obtain evidence strongly indicative of the fact that Alice–not Bob–was driving the car at the time that the car ran the red light.
- Suppose that Alice and Bob rent a car together, but the rental agreement states that Alice is the only authorized driver. She signs the agreement, and then pulls out of the rental car parking lot and drives for two hours. They stop for coffee and Bob decides to drive. Using the techniques that we explore in this paper, a dongle attached to the car could detect (with high probability) that Alice is no longer the driver.
- Suppose that Alice owns a car. Her car has an after-market insurance dongle plugged into it. Her son, Bob, just got his license. Alice chooses (or chose) not to purchase the extra night time insurance coverage for Bob, and hence Bob is only allowed to drive during the day. But one night he does drive. The insurance company dongle detects this and, as a result, cancels Alice's insurance.
- Alice wants to know whether Bob's significant other is driving Bob's car–something that Alice explicitly disallowed. To detect this, Alice installs a monitoring dongle in Bob's car, and Alice gets a real-time text message if that dongle detects a driver other than Bob.
- Alice and Bob own a car with a dongle attached to the car's internal network. When the dongle detects that Alice is driving, the dongle's back-end service sends Alice a text message with a targeted ad for her favorite restaurant. The dongle does the same thing to Bob when Bob is driving.

These scenarios, although hypothetical, suggest that driver fingerprintability–even among small sets of drivers–can raise privacy issues. We note that some may argue that at least the first three scenarios may represent valid uses of the data already available on a car's internal computer network; others may argue that all five scenarios demonstrate violations of privacy. The potential for a debate over the first three scenarios suggest that it is an important issue to discuss.[1] Understanding the feasibility of driver fingerprinting, as we do in this paper, will help inform that debate.

1 Similar debates have occurred around other technologies as well, e.g., RFID toll both transponders [3].

3.4 Anecdotes

Anecdotally, we find that companies have, in the past, already used or know that they could use the data available on vehicles in ways that some might consider privacy-violating. For example, Elon Musk (Tesla Motors CEO) recently used vehicle sensor data to dispute the claims of a New York Times journalist about the limited range of his car's electric batteries (Musk demonstrated that during a road test the NYT journalist took a detour and did not fully charge the vehicle) [12]. Similarly, Ford sales executive Jim Farley was quoted as saying: "We know everyone who breaks the law. We know when you are doing it. We have GPS in your car, so we know what you are doing." [8]. These anecdotes further suggest a need to better understand the privacy implications of the increasing computerization of modern automobiles; we explore that question in this paper.

3.5 Limiting the Adversary

In our investigations we intentionally limit the adversary to only passive eavesdropping on communications already happening on the car's internal computer network. We make this limitation in order to ensure that we are putting the adversary in a challenging position; if the adversary can compromise privacy in this situation, then it surely can also compromise privacy when given even more data or capabilities. In some of our analyses we even further limit the capability of an adversary to only access data pertaining to specific components in the car. We find that even such a limited adversary can fingerprint drivers with high probability. Our results suggest an inherent difficulty in preventing private information flow to any party connected to the car's internal computer network, at least under the current design of automobiles built on top of the standard automotive network protocols.

4 Experimental Data Collection

Recall that the goal of our work is to experimentally measure the degree of driver differentiation possible using the data generated from the existing sensors in the vehicle. To this end, we collected data from the internal communication network (CAN bus) of a single car which was driven by 15 volunteer participants in an isolated parking lot as well as along a 50 mile open-road course. To minimize bias due to traffic conditions, we performed the data collection during the same time of day for each driver. To further normalize the driving context, we requested that all drivers listen to the same radio station (uptempo pop music).

4.1 Vehicle and Selected Sensors

The vehicle we used in our data collection was a a 2009-edition modern sedan. In particular we connect to the diagnostic port (OBD-II) and log the messages broadcast by various manufacturer installed electric control units (ECUs) and sensors during driving behavior collection.

As previously mentioned, there are many more available sensor streams in our experimental vehicle than the ones we choose to log. The motivation for our sensor stream selection was to focus our analysis on the control actions of the driver and the dynamic state of the vehicle (without added knowledge of external surroundings). The list of 16 sensors we record from is available in Table 1. These sensors are likely to be present in many vehicles and provide a baseline from which to measure information leakage in modern automotive contexts. Furthermore, we expect that the values produced by the members of this sensor subset will be dependent on driver behavior (as opposed to being exclusively coupled to the behaviors of internal vehicle systems). Note that the equipment we use to collect the data is passive, and we are only intercepting broadcasts (i.e., we did not modify any of the sensors).

4.2 Driver Recruitment

Prior to recruitment we first obtained approval from the University of Washington's Human Subjects Division (IRB#: 44435 "Methodologies for Driver Behavior Fingerprinting from Sensor Data Collected During Vehicle Operation"). Subsequently, we recruited subjects via public fliers and email lists which described the experimental setup and offered a $75 compensation fee for an expected maximum study duration of 3.5 hours (average duration was 3 hours).

From the pool of interested responders we selected candidates which: (1) held a valid driver's license, (2) held a valid university ID (for insurance purposes), and (3) had driven a vehicle in the past month. In addition to these inclusion criteria, we did our best to select participants so as to achieve equal male and female rep-

resentation. Of our final set of 15 participants 8 were males (average age 27.7), and 7 were females (average age 32.5). The youngest participant in the study was 24 years old, and the eldest was 47.

4.3 Driving Data Setup

Next we helped subjects become familiar with the vehicle and subsequently began the two part data collection process. During data collection an experimenter was always present in the vehicle to record vehicle sensors (using a laptop computer), provide instructions, aid with questions/concerns, and offer assistance in case of an accident. The data for each driver was collected at the same time of the day (familiarization and closed course start at 12:30 PM PST, open road drive begins at 1pm PST) to exclude the impact of special traffic situations (e.g., rush-hour) on the driving style.

Furthermore, the results hold when the training dataset is from the closed-course section and testing data is from the open-road section (and vice-versa) suggesting that the effects of context (traffic) do not dominate the identification signature.

Vehicle Familiarization

Since we did not expect our volunteer drivers to be familiar with our car, we guided each of them through a brief inspection and orientation process prior to the beginning of the driving portion of the study. Participants were instructed to familiarize themselves with all dashboard indicators, controls (e.g., wipers, turn signals, hazard lights, car horn), and subjects also had an opportunity to perform adjustments (seat, steering wheel, rear-view mirrors). We note that none of these adjustments (nor any interactions with actuators/sensors outside of our allowed list 1) were used in our data collection or for driver identification. We hypothesize that the use of these sensors would only have made fingerprinting easier, however we did not use them because we wanted to focus exclusively on actions connected to driving behavior independent of the particular features of the experimental vehicle's interior.

Driving Part 1 - Parking Lot Maneuvers

The closed course portion of the experiment was intended to help us collect technical driving behavior without the interference of other drivers and traffic conditions. Subjects were asked to complete a series of 3 laps (the first lap was practice, laps two and three logged in driving database) each of which consisted of the following sequence of maneuvers: (1) parallel park, (2) forward weave through 5 cones, (3) 3-point turn, (4) reverse weave through 5 cones. All of the closed course experiments were completed in a subsection of a parking lot reserved for long term storage of work vehicles after seeking permission from our University's Fleet Services.

Driving Part 2 - Open-Road Loop

For the final part of the study participants were asked to drive along a predefined interurban loop spanning roughly 50 miles (∼2 hours). The course was designed to incorporate a diversity of road types including highway, city, residential, and industrial driving segments.

5 Analysis Methods

Below we describe the sequence of steps for extracting sensor values from the car's internal network packets, signal pre-processing, feature extraction, and running queries of test data snippets against the trained set of pairwise classifiers (multi-class classification).

5.1 Sensor Values from CAN data

As described in Section 2 sensor values are broadcast on the vehicle's control area network with periodic timings. For the sensors in Table 1 we capture the raw hexadecimal payload, add a timestamp and extract the decimal interpretation. In some instances, the raw values have to be linearly transformed in order to adhere to the expected range for each sensor; the transformation coefficients for addition/multiplication are available from the manufacturer's documentation.

1 While traffic conditions are an uncontrolled variable, and a potential source of bias (e.g., heavy traffic vs light traffic) our results hold when the training dataset is from the closed-course section and testing data is from the open-road section (and vice-versa) suggesting that the effects of context (traffic) do not dominate the identification signatures.

Table 1. Sensors Included in Data Collection

Sensor	Control Module	Range	Update Rate	Summary
Brake Pedal Position	Brake	$0 - 100\%$	15ms	Degree to which driver is depressing the brake pedal.
Steering Wheel Angle	Brake	$-2048 - 2048^{\circ}$	20ms	Positive when steering wheel is rotated counterclockwise.
Lateral Acceleration	Brake	$-32 - 32 \frac{deg}{sec}^2$	25ms	Measurement from an accelerometer, positive in left direction.
Yaw Rate	Brake	$-128 - 128 \frac{deg}{sec}$	30ms	Vehicle rotation around vertical axis, positive in left turn.
Gear Shift Lever	Transmission	$1 - 6$	50ms	An indication of the state of the transmission shift lever position as selected by the driver.
Vehicle Speed	Transmission	$0 - 317.46 \frac{miles}{hr}$	100ms	Vehicle speed computed using the angular velocity of the primary (high torque) axle.
Estimated Gear	Transmission	$1 - 6$	60ms	An estimate of the gear that the transmission has achieved (will not change its value until a shift is complete).
Shaft Angular Velocity	Transmission	$0 - 16383.8rpm$	25ms	Speed of the transmission output shaft; on front wheel drive configurations this signal represents the average speed of the front axles.
Accelerator Pedal Position	Engine	0-100%	20ms	Degree to which driver is depressing the accelerator pedal.
Engine Speed (RPMs)	Engine	$0 - 16383.8rpm$	15ms	High-resolution engine speed in revolutions per minute.
Driver Requested Torque	Engine	$-848 - 1199Nm$	60ms	Value is based on the acceleration and brake pedal characteristics.
Maximum Engine Torque	Engine	$-848 - 1199Nm$	125ms	This signal is the calculated maximum torque that the engine can provide under the current circumstances (altitude, temperature, etc.), based on wide-open throttle conditions.
Fuel Consumption Rate	Engine	$0 - 102 \frac{liters}{hr}$	125ms	Instantaneous fuel consumption rate computed based on the average over the last sample period (e.g., 100 ms).
Throttle Position	Engine	$0 - 100\%$	30ms	Zero represents the near closed bore position (idle, coast) and 100% represents full available power.
Turn Signal	N/A	$2/1/0$	N/A	Off, left, or right turn signal.

List of sensors used in analysis. Note that ranges are based on sensor hardware and may not necessarily reflect the empirical levels reachable during normal operation.

Table 2. Data Segment Details

Data Segment	Avg. Duration	Distance	Details
Parking Lot	7.65 min	0.42 mi	Parallel Park x2, Forward Weave x2, 3-point Turn x2, Reverse Weave x2
Open Drive Part 1	17.81 min	5.3 mi	College Campus (1.4 mi), Interstate (3 mi), Downtown (0.9 mi)
Open Drive Part 2	135.27 min	44.8 mi	Downtown (1.4 mi), Interstate (4.5 mi), Residential (7.5 mi), Interstate(13.8 mi), Highway(7.1 mi), Shopping Mall (7 mi), Residential(3 mi), College Campus(.5 mi)
All	160.73 min	50.52 mi	-

Each subject drove the vehicle through 3 segments (Parking Lot, Open Drive Part 1, and Open Road Part 2). Details about the road types/manoeuvres, travel duration, and travel distance for each segment are provided above.

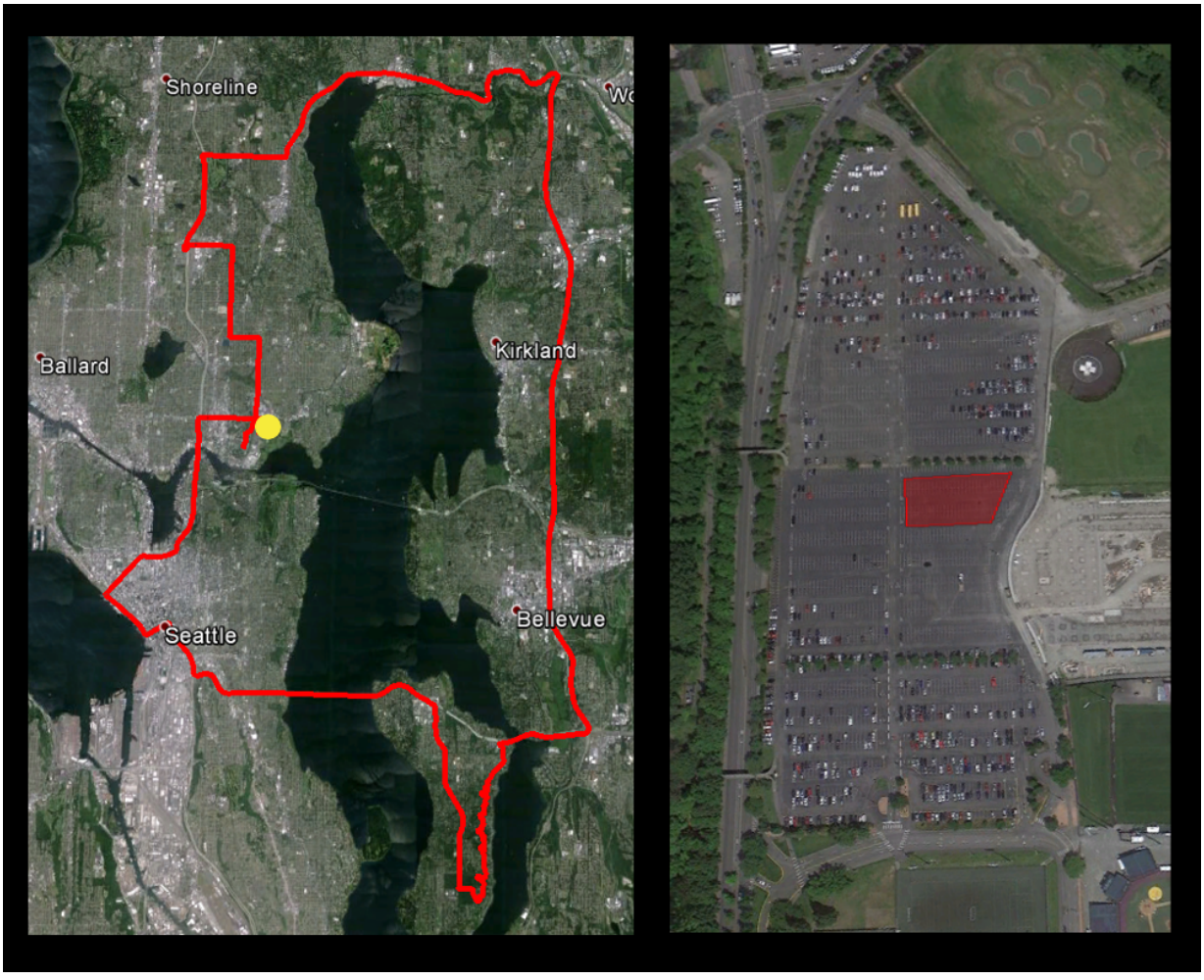

Fig. 1. Data Collection: Drive Loop and Parking Lot Locations. The open-road section started and ended within the University of Washington campus (yellow dot; reverse-clockwise traversal). The closed-course (parking lot) driving involved several laps in an enclosed section of parking lot; each lap consisted of a sequence of maneuvers including parallel parking, forward and reverse weaving through cones, and a 3-point turn.

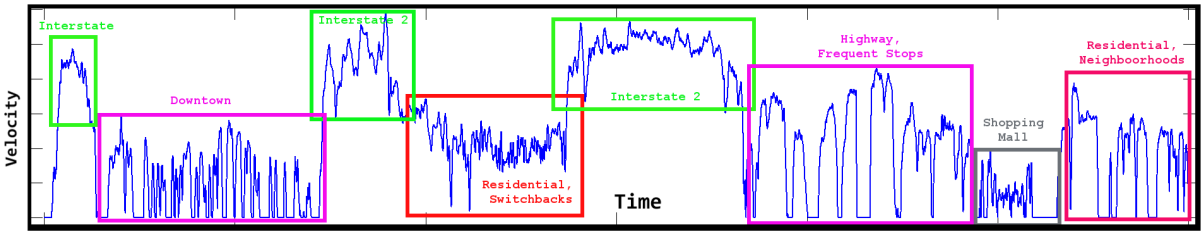

Fig. 2. Velocity data shown throughout the entire open road drive (excludes parking lot). Note the difference in velocity across the different road segments; segments are shown highlighted with boxes of different colors (i.e., interstate = green, urban = pink).

5.2 Signal Pre-processing

Once the decimal values have been processed and linearly transformed within the expected ranges, we resample each sensor to 60Hz by applying quadratic interpolation and decimation as necessary depending on the inherent sampling rate of each sensor.

After the data is uniformly sampled, we smooth each sensor stream by applying wavelet denoising to remove high frequency artifacts. This operation involves multi-level stationary wavelet decomposition and subsequent reconstruction using the Haar wavelet (a.k.a. Daubechies 1) with the default denoising threshold of the MATLAB *iswt* command [6, 17].

5.2.1 Derived Sensors

We were interested in testing the potential for using derived features in addition to the raw sensor readings we could collect from the diagnostic port. To this end we computed the derivative of acceleration (jerk) in the forward and lateral directions. Jerk is a feature that has been commonly used in the optimal control literature [21] and we also anticipated that it may capture the behavior of drivers that try to maximize smoothness. To compute forward jerk we used the second derivative of forward velocity, and lateral jerk was computed via the derivative of lateral acceleration. In both instances we applied an additional layer of smoothing to remove high frequency artifacts from non-continuous sampling.

5.3 Sliding Windows

After pre-processing data from each sensor, we divided it into overlapping sliding windows from which we extracted our features. The sliding window length (number of samples) and percentage of overlap with previous and successive windows were free variables which we set to default values and subsequently optimized in the Results section.

5.4 Features

The features we derive from the pre-processed signals are intended to capture the statistical and morphological characteristics of each sensor data stream. For each sensor and time segment (sliding window) we end up with 48 features which include :

- **Statistical features** – minimum, maximum, average, quartiles, standard deviation, autocorrelation, kurtosis, skewness
- **Descriptive features** – piece-wise average approximation PAA (10 subdivisions)
- **Frequency features** – Fast Fourier Transform (first ten Hz power components, average power in 10-20Hz, average power in 20-70Hz, and average power of > 70Hz components). The compression of higher frequencies and the emphasis of capturing the raw values of lower frequencies was based on the expected rate of actuation (i.e., highly unlikely that drivers perform many actions at a rate higher than 10 times per second). Also, we mention that we resample at 60Hz, so any data over 30Hz is a harmonic (Nyquist criterion).

5.5 Machine Learning and Multi-Class Classification

The features computed from each sliding window comprise a single sample vector used in training or testing of a machine classifier ensemble. Below we describe the members of the classifier ensemble, the division of training and testing samples, and the method used for multi-class classification.

5.5.1 Training vs Testing Segmentation

Given a database of driving data (e.g., parking lot sensor recordings) we train each classifier using the majority of available data (90%) and test (perform queries with unseen data) using the remaining subset (10%). We used ten way cross-validation to ensure that each subset of the database was used for both testing and training. We also ensure that no overlapping sliding windows span between the training and test set by removing samples that are on the border of the 90%/10% split. In each cross-validation slice we used the test data to compute a scaling value that we applied to all data (both train and test). The scaling is intended to reduce the influence of outlier samples and is based on the following formula

$$X_{scaled} = \frac{X_{raw} - mean(X_{raw})}{std(X_{raw}) + \epsilon}$$

where epsilon is a very small positive value ($1e^{-6}$) to avoid division by zero.

5.5.2 Classifier ensemble

In our analysis we used the following four machine learning algorithms for binary classification.

- **Support Vector Machine** – radial basis function kernel (sigma 1), interior-point method (quadratic programming solver) (libsvm 3.1 package)
- **Random Forest** – 1000 classifier trees (randomforest-matlab 4.5-29 package)
- **Naive Bayes** – Kernel smoothing density estimate, uniform prior (MATLAB NaiveBayes.fit)
- **KNN, k-nearest neighbor** – parameters: q = 9, using euclidean distance metric with majority rule tie break (MATLAB knnclassify)

5.6 Pairwise Comparisons - Qweighted

Since all of the classifiers we utilize are binary, and we need to distinguish between many possible individual drivers ($N > 2$) we need to be able to support multi-class classification. The method we use to enable multi-class classification is to train a set of pairwise classifiers (one for each pair of subjects). This approach has been shown to produce more accurate results than the one-against-all approach for a wide variety of learning algorithms because it (1) requires less training data, and (2) enables training using less total memory [26].

6 Results

We began our analysis with an expectation that drivers may intermittently exhibit unique behaviors but no intuition about how this might translate into quantifiable identification accuracy between the participants in our database.

An initial proof of concept experiment found statistically significant differences in the raw sensor data of several subjects who self-reported differences in driving style. Motivated by this result we applied our multi-class machine learning query framework to a subset of our database (parking lot) which yielded a promising starting baseline for achievable accuracy. We subsequently optimized the free parameters of our analysis workflow and honed in to the best performing classifiers.

Once our framework was tuned, we found compelling evidence that drivers are indeed distinguishable from sensor data; furthermore we observe that not much data and a few sensors are sufficient for identification.

Table 3. Identification Accuracy

Sensor(s)	Parking Lot	Drive Part1	Drive Part2	All Data
Brake Pedal	50.00	87.33	100	100
Steer Angle	31.33	64.67	83.33	86.67
Accel. Pedal	15.33	18.00	30.00	31.33
Max Torque	75.33	60.67	100	91.33
Lat. Accel.	25.33	62.00	91.3	72.67
Top 3 Sensors	80.06	92.67	100	100
Top 5 Sensors	84.67	99.33	100	100
All Sensors	91.33	100	100	100

Driver identification accuracy matrix using various combinations of sensor(s) and driving section(s). Top sensors are based on ranking described in Section 6.3.

6.1 Parameter Optimization

To find the best settings for the two key parameters (sub-window size, and overlap percentage) we did a search in parameter space with all classifiers, sensors and features using cross validation (90%, 10 splits%). The sub-window sizes we checked ranged from 200 milliseconds to 15 seconds[2] and the overlap percentages between successive windows were allowed to vary between 10-50%. No overlap was allowed between test and training windows.

While this was a time intensive process, it was also worthwhile given its impact on performance. For our database the best driver identification was achieved using 3 second windows with 25% overlap–91.33% accuracy. The runner up combination was 2 second windows with 33% overlap–86.67% accuracy. The average accuracy across all tested combinations was 74.27%.

The significant boost in performance with tuned settings highlights the importance of finding a window size that spans the duration of driving events. Indeed one important conclusion of our work is that the 3 second envelope may be the optimal length for capturing separate driving [micro] events (especially when using a sliding window approach to feature extraction).

2 Sub-window sizes checked [in seconds] include: .25 .5 .75 1 1.5 2 3 5 10 15.

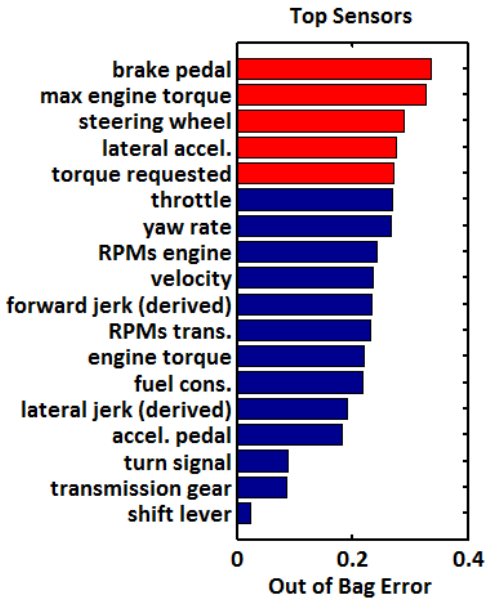

Fig. 3. Top sensors shown in sorted order of their ability to differentiate between drivers (top 5 sensors are shown in red). The brake pedal position is the most telling indicator of a driver's style. The next most relevant sensor is the max engine torque.

6.2 Classifier Ensemble Pruning

Another interesting result of our efforts in optimization was that the Random Forest classifier almost always outperformed the other members of the ensemble (better in 97.33% of test cases, tied for first in 99.13% of evaluated instances).

We attribute the significant gap in performance between these classifiers to their unique mathematical machinery and specifically to each model's ability to handle large, redundant, and/or irrelevant sets of features. While some classifiers were very sensitive to the training features (support vector machines) the Random Forest classifier did very well because it performs an internal feature selection step.

Due to the dominant performance of the Random Forest model in our subsequent analysis we do not report results from the other members of the classifier ensemble in favor of computational complexity as well as for reporting simplicity.

6.3 Top Sensors and Features for Driver ID

Given the optimized parameters and classifier model, we wanted to find which sensors and features were most important for accurate identification. To this end we combined all available data (parking lot and both open-road driving sections) and tested the identification accuracy of each sensor individually using all available features (Random Forest classifier). The results of this experiment are shown in Figure 3 and one interesting conclusion is that braking actions produce the most identifiable aspect of driving behavior in our database (via the brake pedal position sensor).

Next we explored the importance of the individual features in our feature set. This analysis follows the variable importance method and again used the combined dataset which included all sensors from all drives (parking lot and both open-road driving sections). For each individual feature (m) to be tested, we randomly permuted its value along branches of the Random Forest and averaged the correct classifications using out-of-bag error to determine the importance score for feature m.

Our hypothesis is that the sensor and feature ranking results are likely to hold for other vehicles. While the maximum achievable torque at every instant may not be accessible from every vehicle, we believe that the brake pedal and steering wheel will be among the top sensors for driver identification because (1) they represent the most direct information about the actions of the driver, and (2) seem to capture the most unique aspects of a driver's strategy/execution. As for the feature ranking, the top features seem to capture the range of sensor values in the time windows of analysis (though we expect the exact order of feature importance to be very sensitive to differences in analysis methods). The top 5 features were: min, std, max, range, 4th quartile.

6.4 Query Results vs. Course, Sensors

Next we computed the driver identification accuracy on the various segments of our course (parking lot, vs drive part 1, vs drive part 2) using different sets of sensors. Table 3 shows the accuracy achievable in the various combinations. Below we highlight some key results:

- **Single Sensor** – 87.33% accuracy can be reached within the set of participants using a single sensor (brake pedal) using the first part of the open-road drive (∼15 minute average duration), and 100% accuracy is achieved using the brake sensor when the second portion of the open-road drive (∼1.5 hour average duration).
- **Parking Lot** – 91.33% accuracy can be reached within the set of participants using all available sensors on the closed-road technical maneuvers in the parking lot (∼8 minutes average duration)

- **Driving Part 1** – 100.00% accuracy can be reached within the set of participants using all available sensors on the first open-road section (∼15 minutes average duration) which includes urban and highway segments
- **Driving Part 2** – 100.00% accuracy can be reached within the set of participants using all available sensors on the second open-road section (∼1.5 hours average duration) which includes residential, city, and highway segments

To summarize, our investigation shows that not much time and not many sensors are needed to accurately identify a driver in our database.

6.5 Extension: Fingerprint Stability

As an extension we also explored whether a single driver could be consistently identified across multiple days of data recording and differences in the course.

To this end we selected one driver and collected 5 round trips from the University to a nearby town (22 mile trip). Using this dataset as a query (and our original dataset as training) we applied our analytic methods to find that our test driver's unique fingerprint was consistent across multiple days and roads (91% accuracy, same driver different roads and days of data collection). The data in these fingerprint stability experiments used all the sensors and the optimized parameters developed in the work described in Section 6.1 above.

As validation, we also excluded this driver from the training database (reduced to N=14) and attempted to query with the new test data from the 5 round trips (not included in the original database). This led to very low confidence results (average of 6.53% gap in pairwise comparisons between top candidate and runner up) randomly distributed among the set of candidate drivers ($8.2\% +/- 4\%$ probability of attribution to any of the 14 drivers in the database). These results suggest that the fingerprinting method can be used to reliably interpret whenever query data belongs to a driver not present in the training database.

7 Related Work

To our knowledge, the two most similar prior works that have also targeted driver identity inference from sensor data were conducted by Miyajima et al. and Nishiwaki et al. in 2007 [23, 24]. Both of these efforts were completed by a similar set of authors and they differ in that the Miyajima et al. citation refers to an article while Nishiwaki et al. appears as a book chapter.

Miyajima et al. and Nishiwaki et al. [23, 24] developed an identification method based on frequency analysis (ceptstrum based) of sensor data collected from two independent experiments; the first experiment used data collected from a driving virtual simulator (86% identification accuracy among 11 subjects), and the second experiment leveraged data previously collected from the CAIR dataset (76.8% identification accuracy among 274 subjects). The CAIR dataset recorded multimedia data such as audio, video and vehicle sensor information as drivers responded to prompted dialogue questions; the main objective of this dataset was to study the human-machine speech interface during driving behavior [22].

While the driver identification results of Miyajima et al. and Nishiwaki et al. are important, we note that their driving datasets were based on either simulated data, or collected with expensive/uncommon sensors (i.e., laser range finders, video) in a highly instrumented van with a large computer rack. Our work, focuses on a stock sensors in a modern sedan, and focuses on natural driving behavior without introducing potentially distracting tasks such as prompted dialogue.

Choi et al. also looked at performing inferences about the driver using in-vehicle CAN bus information, however the primary emphasis of the work was on measuring driver attentiveness during normal driving and driving with distractor tasks. The authors evaluated the the potential to identify among 9 drivers and reached 31.45% accuracy using a hidden Markov model (HMM) [16].

Two other related efforts which investigated the potential to identify drivers in simulated virtual environments were performed by Zhang et al. and by Wakita et al. [29, 31]. Wakita et al. reached 73% driver identification among a set of 30 drivers which were instructed to follow a guide vehicle. Zhang et al. collected simulator data from 20 male subjects across multiple day session (sessions were performed on the same route and lasted approximately 30 minutes) and reached an identification accuracy level of 85% using HMMs. On one hand, these simulated experiments enabled a controlled setting which removed potentially confounding factors present in real world driving (e.g., traffic variation). On the other hand, the participants self reported that the simulator did not capture the authentic experience– specifically emphasizing that braking imitated real driving poorly. Unlike these efforts which attempt to mimic

biofidelity, our data comes from real driving scenarios; furthermore, we try to balance for the lack of a controlled environment by collecting data during a closed-road (parking lot) session in which there are no external factors to influence driver behavior.

Lastly, Van Ly et al. attempted to perform driver identification distinguishing between two drivers using sensor data collected from inertial sensors in a mobile device [28]. This work initially shows that a mounted phone sensor's accelerometer is highly correlated with acceleration and braking activity, and subsequently the authors use the phone data to distinguish between the two drivers along a diverse multi-hour course involving residential and highway segments (using a modern sedan). Their results indicate that the highest achievable performance using acceleration, braking, and turn data using their dataset is roughly 60% using unsupervised k-means and supervised SVM classifiers.

While past results indicate that driver identification above chance levels may be possible it is not clear to what degree this inference can be made from the information flowing through an unmodified vehicle. We aim to address this gap and investigate the level of driver identification possible using the sensor data in a stock vehicle driven by 15 drivers along open and closed courses. Unlike past work we do not use simulation data nor mobile phones and the sensor streams we tap are those pre-installed by the manufacturer without including additional instrumentation (i.e., laser range finders).

8 Discussion

While we anticipated some level of de-anonymization success, our results are surprising given the apparent potential of vehicle sensor data present in stock vehicles to distinguish between individuals given limited time and restricted access to sensors. We view this as a significant result since it implies that even simple devices – such as insurance dongles attached to a car's internal computer network – have the potential to violate privacy. Moreover, we expect that future vehicles are likely to have even richer sensor streams including video data and location awareness, which only increase the potential for privacy breaking attacks. However, we note that as more functionality eventually becomes autonomous, the ability to fingerprint decreases; at the ultimate end of the spectrum, with a fully autonomous vehicle, we imagine that we could at most fingerprint the algorithm and not the passenger.

8.1 Scaling to Large N and Different Vehicles

One natural question about our work is whether the techniques we have presented will enable driver identification when applied to large sets of individuals. While applying our techniques to only a few drivers can still be a significant privacy concern, as noted in our threat model section, we believe that it should be possible to apply driver identification on very large scales. Specifically, we speculate that several ideas can be applied [together] to restrict the candidate pool of matching drivers given a query sample of sensor data :

- *Clustering techniques* (and other unsupervised structural methods) can be used to limit the set of candidate matches to a given query.
- If the rough *geographic location* of a car and driver are known, it would be possible to further restrict the search space.
- Access to longitudinal data should facilitate identification (i.e., given *enough data* everyone can be distinguished).

One issue that we did not experimentally explore in our work, is how a driver's fingerprint transfers between different vehicles and vehicle types. While we consider this analysis out of scope for our study and core threat model, we conjecture that drivers are likely to retain the majority of their driving signature (strategy and execution patterns) independent of the vehicle in question. An interesting direction for future research would be to develop driver identification models that can adapt to different vehicle dynamics/makes/models.

8.2 Towards Utility and Privacy Balance in Vehicle Data

Given the diversity and scale of the automotive ecosystem we believe that developing a balance between utility and privacy in sensor data exchanges will require a combination of legal and technical solutions. Policy debates are already ushered on by consistent calls for increased consumer privacy protections, however the diversity of existing legal opinions highlights the complexity of creating regulatory frameworks in intelligent automotive systems.

2 Further experimentation is necessary to validate or refute these claims.

Technical methods have also been suggested to mitigate tensions, however matching the available solutions to each deployment context is a difficult problem. Below we touch on some of the interesting developments in the legal and technical spheres which we consider relevant to the future of utility and privacy in vehicle sensor data exchanges.

8.2.1 Existing Legal Perspectives for Vehicle Data Privacy

From a legal perspective there are varying stances on vehicle sensor data ownership, processing, and management. One of the central policy challenges is mitigating the risks of data reuse for unforeseen, and potentially adversarial, purposes which raises significant privacy concerns. Within the United States, 13 states have adopted the stance that a vehicle's sensor data is private and the property of the car owner [9], however within these 13 states there are marked differences on what constitutes acceptable data retrieval without owner consent [3]. Of course, if a driver may still authorize another entity to access the vehicle's sensor data, e.g., as part of a contract.

8.2.2 Technical Defenses

From a technical perspective there are significant efforts to develop de-anonymization tools which defend individuals from privacy attacks. Typically these efforts have focused on providing theoretical guarantees in limited contexts where information releases are managed by a statistical database (or data vaults) capable of obfuscating data or injecting noise to prevent the linkage of data entries to specific persons [19, 27]. While these approaches offer strong protections, their use cases are somewhat constrained by the information request and release mechanisms required to enforce privacy polices. More aligned with the streaming nature of vehicle sensor data, is the work towards privacy preserving transformations of real-time streams (e.g., SensorSift [20]) intended to remove sensitive aspects of the data while allowing useful inferences to still extract utility from the sifted data.

Another defensive technique specific to the driver de-identification problem, would be to embed random sensor signals (e.g., break pedal actuation) to the output nodes of the CAN bus (e.g., OBD-II port). In this way the vehicle state would not actually be interrupted by signal injections (i.e., the break signals would not be executed) but would be observed by any upstream subscribers – hence introducing noise in the ability to acquire a driver's unique fingerprint.

Lastly, some automotive manufacturers are starting to mediate access to CAN packets through gateways which can limit the information observable at the OBD-II port [5, 30]. If this feature becomes more common it would thwart methods that rely on data exfiltration from the diagnostics port (though it may be possible to gain access to sensor data using other trusted nodes on the network).

9 Conclusion

Through our work we hope to inform stakeholders with concrete results of information leakage (via privacy braking inference) in a realistic vehicular context. Unlike past work, our analysis focused only on stock sensors in a typical vehicle (2009 sedan) that has not been instrumented beyond what has been installed by the manufacturer. As our results indicate, it is possible to accurately identify drivers using limited amounts of sensor data collected from a restricted set of sensors (e.g., 87% accuracy in distinguishing between 15 drivers, using just the brake pedal position from 15 minutes of open-road driving data [13.5 minutes training, 1.5 minutes test data], 99% accuracy is achievable when using the top 5 sensors). Furthermore, an extension of our work suggests that a driver's fingerprint (driving strategy and unique patterns of execution) are consistent across different days and road types (see Section 6.5).

These results suggest that drivers should be wary of sharing their vehicle data streams without substantial guarantees for superior service. Similarly the consumers and collectors of said data should have a responsibility to offer users with privacy controls and develop safeguards for data processing and retention which keep up with the evolving threat model landscape.

3 Connecticut requires warrants [18], Oregon allows unconsented disclosure to "facilitate medical research of the human body's reaction to motor vehicle crashes" or "to diagnose, service, or repair a motor vehicle" [25], and Arkansas prohibits insurance companies from access to the data in accidents to prevent the insurer from assuming vehicle ownership [15].

10 Acknowledgments

This research received funding from the Alfred P. Sloan Foundation, the Intel Pervasive Computing Science & Technology Center, and NSF Grant CNS-0963695. We kindly thank Melody Kadenko for her significant and continuous support in all stages of this project. We are also grateful to Patrick Johnson, and University of Washington Commuter Services for their assistance in arranging our use of the E1 parking lot for experiments.

References

[1] Apple carplay. http://www.apple.com/ios/carplay/?cid=wwa-us-kwg-features-com.

[2] Automatic. https://www.automatic.com/.

[3] Buyer beware: Ez pass toll tags used by private attorneys to catch cheating spouses. http://www.texasturf.org/2012-06-01-03-09-30/latest-news/755-buyer-beware-ez-pass-toll-tags-used-by-private-attorneys-to-catch-cheating-spouses.

[4] Epic comments on electronic data recorders. https://epic.org/privacy/drivers/edr_comm81304.html.

[5] Exclusive: Say Goodbye to Chip Tuning Open CAN Bus Going Away in Two Model Cycles. http://www.thetruthaboutcars.com/2014/05/exclusive-say-goodbye-to-chip-tuning-open-can-bus-going-away-in-two-model-cycles/.

[6] Inverse discrete stationary wavelet transform 1-d. http://www.mathworks.com/help/wavelet/ref/iswt.html.

[7] Markey report reveals automobile security and privacy vulnerabilities. http://www.markey.senate.gov/news/press-releases/markey-report-reveals-automobile-security-and-privacy-vulnerabilities.

[8] The next data privacy battle may be waged inside your car. http://www.nytimes.com/2014/01/11/business/the-next-privacy-battle-may-be-waged-inside-your-car.html.

[9] Privacy of data from event data recorders: State statutes, national conference of state legislatures. http://www.ncsl.org/issues-research/telecom/privacy-of-data-from-event-data-recorders.aspx.

[10] Progressive snapshot privacy statement. https://www.progressive.com/auto/snapshot-privacy-statement/.

[11] Teletrac. http://www.teletrac.com.

[12] Tesla CEO Elon Musk disputes N.Y. Times article on Model S range. http://articles.latimes.com/2013/feb/11/autos/la-fi-hy-autos-tesla-model-s-ny-times-musk-battle-20130211.

[13] Wikipedia: Usage based insurance. http://en.wikipedia.org/wiki/Usage-based_insurance.

[14] Zubie. http://zubie.com/.

[15] Arkansas Code Title 23, Chapter 112, Section 107. Motor vehicle event data recorder – data ownership.

[16] S. Choi, J. Kim, D. Kwak, P. Angkititrakul, and J. H. L. Hansen. Analysis and classification of driver behavior using in-vehicle can-bus information.

[17] R. R. Coifman and D. L. Donoho. *Translation-invariant de-noising*. Springer, 1995.

[18] Connecticut General Statutes Chapter 246b. Motor vehicle event data recorders.

[19] C. Dwork. Differential privacy. In *Encyclopedia of Cryptography and Security*, pages 338–340. Springer, 2011.

[20] M. Enev, J. Jung, L. Bo, X. Ren, and T. Kohno. Sensorsift: balancing sensor data privacy and utility in automated face understanding. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 149–158. ACM, 2012.

[21] T. Flash and N. Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *The journal of Neuroscience*, 5(7):1688–1703, 1985.

[22] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura. Multimedia data collection of in-car speech communication. In *7th European Conference on Speech Communication and Technology/2nd INTERSPEECH Event in Aalborg, Denmark on September 3-7, 2001 (EUROSPEECH 2001). 2001, p. 2027-2030*, 2001.

[23] C. Miyajima, Y. Nishiwaki, K. Ozawa, T. Wakita, K. Itou, K. Takeda, and F. Itakura. Driver modeling based on driving behavior and its evaluation in driver identification. *Proceedings of the IEEE*, 95(2):427–437, 2007.

[24] Y. Nishiwaki, K. Ozawa, T. Wakita, C. Miyajima, K. Itou, and K. Takeda. Driver identification based on spectral analysis of driving behavioral signals. In *Advances for In-Vehicle and Mobile Systems*, pages 25–34. Springer US, 2007.

[25] Oregon Revised Statutes Chapter 105 Motor Vehicle Event Data Recorders. Retrieval or use of data for responding to medical emergency, for medical research or for vehicle servicing or repair.

[26] S.-H. Park and J. Fürnkranz. Efficient pairwise classification. In *Machine Learning: ECML 2007*, pages 658–665. Springer, 2007.

[27] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[28] M. Van Ly, S. Martin, and M. M. Trivedi. Driver classification and driving style recognition using inertial sensors. In *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pages 1040–1045. IEEE, 2013.

[29] T. Wakita, K. Ozawa, C. Miyajima, K. Igarashi, I. Katunobu, K. Takeda, and F. Itakura. Driver identification using driving behavior signals. *IEICE TRANSACTIONS on Information and Systems*, 89(3):1188–1194, 2006.

[30] M. Wolf, A. Weimerskirch, and C. Paar. Security in automotive bus systems. In *Workshop on Embedded Security in Cars*, 2004.

[31] X. Zhang, X. Zhao, and J. Rong. A study of individual characteristics of driving behavior based on hidden markov model. *Sensors & Transducers (1726-5479)*, 167(3), 2014.

11 Appendix – Additional Figures

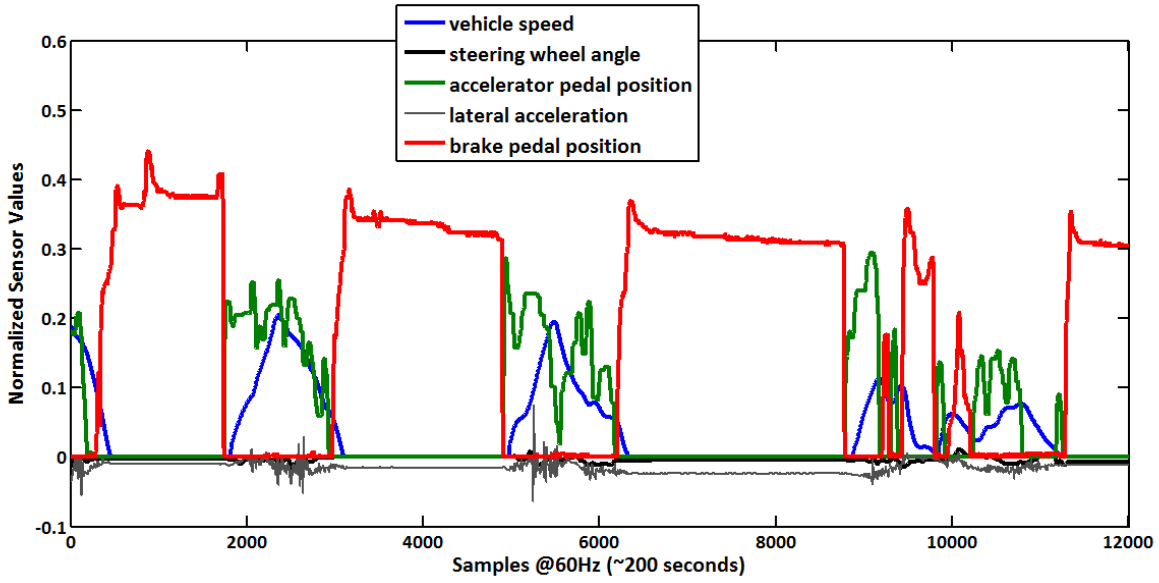

Fig. 4. Sensor data during a segment of the downtown portion of the drive. Inner city traffic lights produce a predictable acceleration and deceleration pattern evident in the velocity plot (blue curve), brake pedal (red curve) and accelerator pedal (green curve).

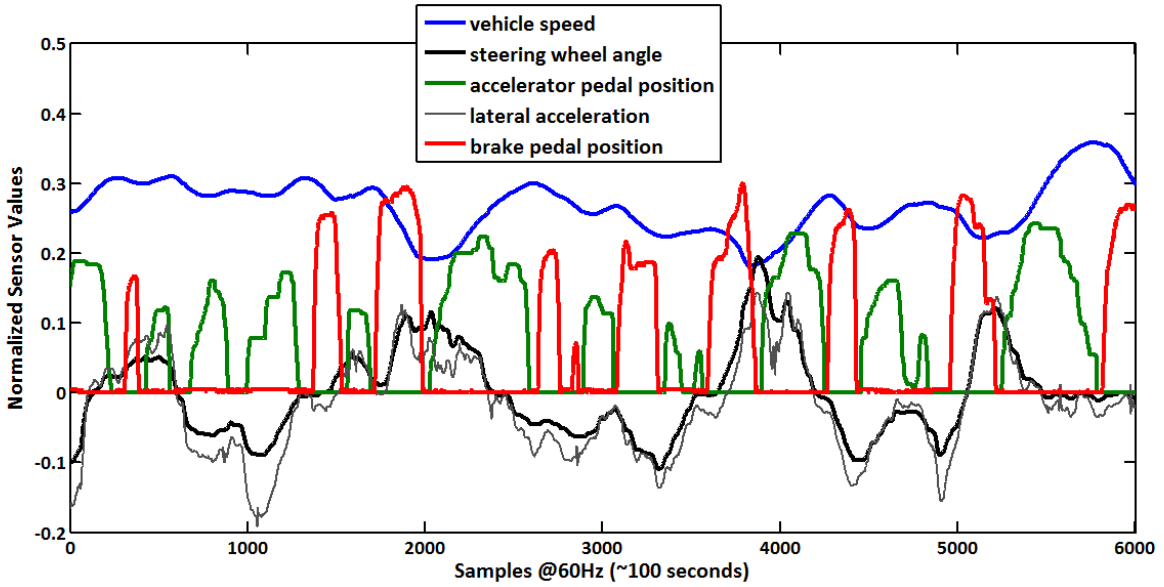

Fig. 5. Sensor data along a winding section of residential roadway requires some technical driving. Note the high amount of steering wheel activity required (black curve) and its close correlation with the lateral acceleration measurements (gray curve). One consistent aspect of this driver's behavior is the amount of braking (red curve) in the early part of turns and the subsequent accelerations (green curve) during turn exits.

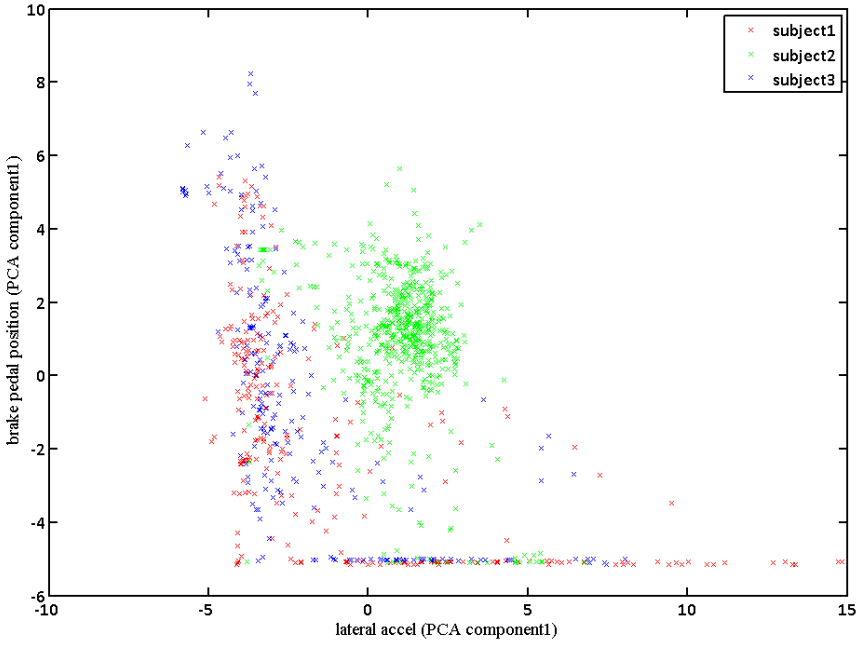

Fig. 6. Diversity in style between three subjects S1,S2, and S3; S1 and S2 are the most similar in the dataset, and S1 and S3 are the most disimilar. While each participant is driving each sensor produces a time series. This time series is divided up into 3 second windows, and 49 sub-features (e.g., min, max, std, mean, quartiles, fft, etc.) are computed. Each of the 3 seconds windows ends up being a point on the graph (different colors for different subjects) after dimensionality reduction (in the form of Principal Component Analysis) is applied to reduce the 49 sub-features onto the 1D principal component axis.

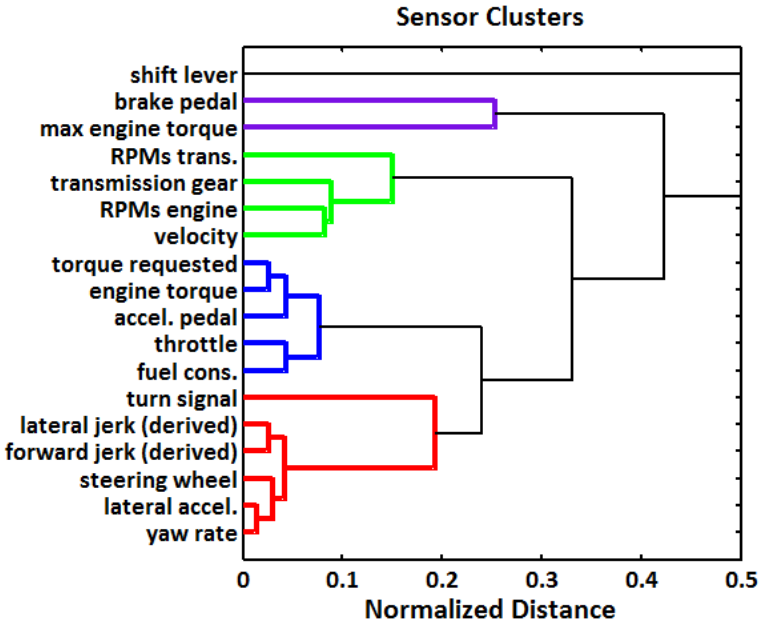

Fig. 7. Sensors can be clustered into four groups: (1) **acceleration** - shown in blue [accelerator pedal, torque requested, etc.], (2) **turning** - shown in red [steering angle, lateral accel, etc.], (3) **vehicle state** - shown in green [velocity, gear, RPMs], (4) **deceleration** - shown in purple [brake pedal, maximum achivable torque]