Transportation Research Part F: Psychology and Behaviour 85 (2022) 83–102

Contents lists available at ScienceDirect

Transportation Research Part F: Psychology and Behaviour

journal homepage: www.elsevier.com/locate/trf

A data–information–knowledge cycle for modeling driving behavior

Christelle Al Haddad *, Constantinos Antoniou

Chair of Transportation Systems Engineering, Technical University of Munich, Arcisstrasse 21, 80333 Munich, Germany

ARTICLE INFO

Keywords:
Data collection
Information extraction
Impacts of AVs
Behavioral modeling
Data analytics
Data fusion

ABSTRACT

When talking about automation, "autonomous vehicles", often abbreviated as AVs, come to mind. In transitioning from the "driver" mode to the different automation levels, there is an inevitable need for modeling driving behavior. This often happens through data collection from experiments and studies, but also information extraction, a key step in behavioral modeling. Particularly, naturalistic driving studies and field operational trials are used to collect meaningful data on drivers' interactions in real–world conditions. On the other hand, information extraction methods allow to predict or mimic driving behavior, by using a set of statistical learning methods. In simple words, the way to understand drivers' needs and wants in the era of automation can be represented in a data–information cycle, starting from data collection, and ending with information extraction. To develop this cycle, this research reviews studies with keywords "data collection", "information extraction", "AVs", while keeping the focus on driving behavior. The resulting review led to a screening of about 161 papers, out of which about 30 were selected for a detailed analysis. The analysis included an investigation of the methods and equipment used for data collection, the features collected, the size and frequency of the data along with the main problems associated with the different sensory equipment; the studies also looked at the models used to extract information, including various statistical techniques used in AV studies. This paved the way to the development of a framework for data analytics and fusion, allowing the use of highly heterogeneous data to reach the defined objectives; for this paper, the example of impacts of AVs on a network level and AV acceptance is given. The authors suggest that such a framework could be extended and transferred across the various transportation sectors.

1. Introduction

In the era of automation, an increasing interest in human–machine interactions has been witnessed across various disciplines. In transportation, this is the case of so-called "autonomous vehicles" (AVs), where a main research objective is to be able to assess their impacts on a network level[1]; among the advantages of AVs are improved mobility through enhancing first and last–mile connectivity to transit services (Moorthy, De Kleine, Keoleian, Good, & Lewis, 2017; Ohnemus & Perl, 2016), but also their presumable positive impact on traffic safety (Nair & Bhat, 2021). With improved technology and advances in big data analytics, it is now possible to obtain data

* Corresponding author.
E-mail addresses: christelle.haddad@tum.de (C. Al Haddad), c.antoniou@tum.de (C. Antoniou).

[1] Please note that in this paper, the terms "autonomous vehicles" and "automated vehicles" are used interchangeably, mostly to refer to highly automated driving systems. These terms would be used in the same way they were found in the papers to be analyzed.

https://doi.org/10.1016/j.trf.2021.12.017
Received 11 March 2021; Received in revised form 13 November 2021; Accepted 28 December 2021
Available online 12 January 2022

from different sensors and sources, and merge it in such a way that it is useful for analysis. This is usually the case of naturalistic driving studies, where driving data is collected by means of a set of sensors and often video cameras, resulting in thousands of driving hours and millions of kilometers of continuous driving (Antin et al., 2019; Knoefel, Wallace, Goubran, & Marshall, 2018). This leads to many challenges, such as data heterogeneity, quality (Wijnands, Thompson, Nice, Aschwanden, & Stevenson, 2019; Yadawadkar et al., 2018), abundance (Antin et al., 2019; Fridman et al., 2019; Blanco et al., 2016; Simons-Morton et al., 2015), etc. In an attempt to understand the process of driving behavior modeling and impact assessment, one should consider the different steps starting from the proper data collection, and ending with the analytics and fusion of heterogeneous data, which would then allow the extraction of the required knowledge. An analysis of the literature shows that there is a gap in representing these different steps as part of a data–information–knowledge cycle, which would encompass the various aspects starting from data and ending with the knowledge.

The main objective of this paper is therefore to better represent this data–knowledge cycle, through a thorough literature review, which aims to give insights into its different components, including the analytics and fusion frameworks, which could be transferable to different modes and research objectives. To the best of the authors' knowledge, this has not been done before, as previous studies focused on specific aspects of data collection or information extraction, separately.

When planning for a new research project, in which data has not already been collected, or in which data is not derived from a previous project, there is a need to start from the first step of acquiring data through an inevitable data collection scheme, followed by many key components such as processing data or storing it, after which knowledge can be generated. Previous research in this area has focused on either data collection, or knowledge extraction, separately, but rarely, if ever, both aspects were mentioned and discussed together. Having this overview would be crucial as it could help better planning for this cycle in which data is first collected, and then useful knowledge for modeling driving behavior could be generated. This is important from a policy point of view since it would allow to have this entire overview and help to better plan new projects, considering the different challenges that pertain to different components of this cycle. New type of generated knowledge could for instance be the different driving styles, or driving maneuvers, resulting from in-vehicle data collection, or even user acceptance on ADAS, based on questionnaires or interviews, etc. When it comes to autonomous vehicles, conducting experiments and collecting data on human–machine interactions, can help gathering useful information which would feed into models that can be replicated at larger scales. Different challenges identified from previous research could pave the way to a better planning for future research. Data collection for instance is often associated with challenges pertaining to data processing, data quality, data privacy, or other external considerations. Putting all of these challenges in one framework would be an easy tool or checklist that can be used before planning for future research on driving behavior modeling.

The contribution of this work would then consist of this holistic framework of analytics and fusion, which can be extended depending on the research question. In essence, the objectives and findings of this work could be structured along following research questions:

1. How is driving behavior data collected?
2. How is knowledge extracted to model driving behavior?
3. How can a data–knowledge cycle be represented to include various aspects of analytics and fusion for driving behavior modeling?

In the next sections, the methodology for this extensive literature is presented (Section 2), followed by literature findings (Section 3) on data collection, and information extraction. After that, the proposed data–information–knowledge framework is presented (Section 4), focusing on aspects of data analytics and fusion, but also on transferability across different modes. Finally, a conclusion is given (Section 5), shedding light on the main findings and future work.

2. Methodology

In this section, the methodology followed in this research is presented in detail. To answer the research questions defined in the introduction, an extensive review has been conducted, which will be reported following some common key items from the PRISMA guidelines (Moher, Liberati, Tetzlaff, & Altman, 2010), such as the eligibility criteria, information sources, search strategy, etc., study selection. A collection of relevant literature was done by searching in Scopus, Google Scholar and IEEE Xplore, with an aim to collect studies focusing on in–vehicle data collection and information extraction. Therefore, and to answer the research questions defined, following keywords were used in the different search engines: "data collection", "information extraction" (to get insights on data collection), "autonomous vehicles", but also "autonomous driving" (to get insights on driving behavior for highly automated vehicles). Particularly, different combinations of these keywords were entered in the search engines, namely "autonomous vehicles" AND "data collection", "autonomous vehicles" AND "information extraction"; the search was also done using "autonomous driving" in place of "autonomous vehicles"[2]. The search was completed by September 2020, and included literature in English, focusing on transportation topics. Additionally, about five references in the literature were reviewed and included ("backwards snowballing" principle). A total of 161 studies were eventually collected, covering road transportation, which were first classified by mode (passenger cars, buses, trucks, bikes, or not specified, usually referring to studies collecting and describing highway environments without being specific to a mode.), and level of automation (conventional vehicles, and automated vehicles such as semi–autonomous, fully–autonomous). Upon initial screening, various topics were identified, based on which a classification was made, with following categories: "data collection",

[2] While "autonomous vehicles" as a term could refer to highly automated vehicles, it might be the case that some studies were missed for not using the term "automated vehicles", which can be a limitation of the keywords search.

Table 1
Initial set of screened papers (conventional vehicles).

Study	Cars	Buses	Bikes	Not specified	Data collection	Driving behavior	NDS	Statistical analysis	Big data
Jacob and Rabha (2018)	•				•				
Yan et al. (2019)	•					•			
Morgenstern et al. (2020, 2009)	•						•		
Bosi et al. (2019)	•								•
Ehsani et al. (2020), Itkonen et al. (2020), Koppel et al. (2020), Kovaceva et al. (2020), Li et al. (2020), Petzoldt (2020), Yasmin et al. (2020), Ding et al. (2019), Muronga and Ruxwana (2017), Dingus et al. (2016), Dingus et al. (2006), Simmons et al. (2016), Wallace et al. (2016), Tivesten and Dozza (2015), Fitch et al. (2014), Montgomery et al. (2014), Tian et al. (2014), Tivesten and Dozz (2014), Wege et al. (2013), Myers et al. (2011), Adornato et al. (2009), Klauer et al. (2006), Sayer et al. (2005)	•					•	•		
Wang and Ho (2018)	•					•		•	
Wu and Jovanis (2013), Wu and Jovanis (2012), Liang et al. (2012), Jovanis et al. (2011)	•						•	•	
Samiee et al. (2014)	•							•	•
Antin et al. (2019), Barnard et al. (2016), Blatt et al. (2015), Simons-Morton et al. (2015), Klauer et al. (2014), Ott et al. (2012), Neale et al. (2005)	•				•	•	•		
Ma et al. (2021), Xia et al. (2018), Warren et al. (2019)	•				•	•		•	
Das et al. (2020), Liang (2020, 2020), Rasch et al. (2020), Alekseenko et al. (2019), Arvin et al. (2019), Hochin et al. (2019), Kuo et al. (2019), Li et al. (2019), Thapa et al. (2019), Wang et al. (2019), Yadawadkar et al. (2018), Precht et al. (2017), Carney et al. (2015), Guo et al. (2015), Hallmark et al. (2015), Victor et al. (2015), Foss and Robert D.Goodwin (2014), Jonasson and Rootzén (2014), Bagdadi (2013), Guo and Fang (2013), Valero-Mora et al. (2013), Ahlstrom et al. (2012), Davis et al. (2012), Guo et al. (2010), Shankar et al. (2008), Lin et al. (2008)	•					•	•	•	
Dawson (2019), Wallace et al. (2017)	•					•	•		•
Chun et al. (2019)	•					•		•	•
Barbier et al. (2019), Chhabra et al. (2019), Yadawadkar et al. (2018)	•				•	•	•	•	
Rosales et al. (2017)	•					•	•	•	•
Fridman et al. (2019)	•				•	•	•	•	•
Blanco et al. (2016)		•			•	•	•		
Barnard et al. (2016), Soccolich et al. (2013), Hickman and Hanowski (2012)		•				•	•		
Aihara et al. (2019), Barr et al. (2011)		•				•	•	•	
Dozza et al. (2016), Dozza and Werneke (2014), Espié et al. (2013)			•		•	•	•		
Kovaceva et al. (2019)			•			•	•	•	
Bachechi and Po (2019), Fan et al. (2019), Kaur et al. (2019), Piedad et al. (2019), Pop and Prostean (2019), Zhao et al. (2019), Abodo et al. (2018), Bellini et al. (2018), Kaushik et al. (2018), Mo et al. (2017), Al-Najada and Mahgoub (2017), McLaughlin et al. (2008)				•				•	
Fernandez-Rojas et al. (2019), Liu and Li (2019), Moharm et al. (2019), Pucci and Vecchio (2019), Zhu et al. (2019), Figueiras et al. (2018), Gohar et al. (2018), Park et al. (2018), Torre-Bastida et al. (2018), Schatzinger and Lim (2017)				•					•
Kaushik et al. (2018)				•	•			•	

(*continued on next page*)

Table 1 (*continued*)

Study	Cars	Buses	Bikes	Not specified	Data collection	Driving behavior	NDS	Statistical analysis	Big data
Mishra et al. (2020), Sangster et al. (2013), Lee et al. (2004)				•		•	•		
Guo et al. (2018), Zhao et al. (2017)				•		•		•	
Guo (2019), McLaughlin et al. (2008)				•			•	•	
Guan et al. (2019), Guleng et al. (2019), Kang et al. (2019), Nallaperuma et al. (2019), Serok et al. (2019), Sivasankaran and Balasubramanian (2019), Zhang et al. (2019)				•				•	•
Knoefel et al. (2018)				•	•	•	•		
Sun et al. (2018, 2013)				•	•			•	•
Zhou et al. (2019)				•		•		•	•

"driving behavior", "Naturalistic Driving Studies (NDS)", "statistical analysis", and "big data. Initial screening was made by reading the abstract first, then scanning the contents, and finally going more in depth into the paper when otherwise unclear. The mode classification was important to see the most dominant modes across these studies. The other categories were useful to highlight the fields of contribution made by each paper. Data collection referred to such studies where procedures of the experiments were described, along with the devices and sensors used, size of data, and aspects of data handling. Driving behavior referred to all studies whose aim were to classify different driving styles or traits that help better understand driving characteristics. Naturalistic driving studies were ones where the main data was part of an NDS, as described by the authors themselves. Further, statistical analysis were studies where statistical models were elaborated to extract information and features, useful generally to model driving behavior. Finally, big data referred to studies focusing on big data tools and methods for modeling, processing, analyzing and visualizing transport and mobility.

From an initial screening of abstracts, it was obvious that most papers could either answer the first research question (on the collection of driving behavior data), or the second (on knowledge extraction for modeling driving behavior). Furthermore, we did not find any holistic contribution which elaborated the different steps going from data collection (and the challenges faced there) to information extraction (based on that same collected dataset). We therefore split the initially collected papers in two subsections, one for data collection (mostly found in papers addressing conventional vehicles), and the other for information extraction (in which we focused on findings in the papers tackling AVs). The aim was to eventually combine findings from each of these sub–sections in order to answer the third research question, which would then be a bridge between both, and a transition to future research on AV behavioral modeling.

A full list of the primarily selected papers is presented in Tables 1 and 2, for conventional and autonomous vehicles respectively. Finally, these papers were screened, and a subset of 27 studies were selected, to be analyzed in further detail. These were studies that fit best the scope of the research objective: modeling driving behavior by looking at data collection aspects, and information extraction. This means the primary focus was given on driving behavior as a common interest factor. For example, some studies were removed as they were not concerned with driving behavior; this includes studies on image classification and vehicle detection (Ghandour, Krayem, & Gizzini, 2019), work zone sign detection (Seo, Wettergreen, & Zhang, 2012), traffic sign estimation (Vu, Yang, Farrell, & Barth, 2013), text recognition (Balaji, Kumar, & Sujatha, 2017), road investigation under weather conditions (Cheng, Wang, & Zheng, 2017), driver and vehicle recognition (Mo, Gao, & Zhao, 2017). Moreover, studies which presented the same or similar outcomes from the same authors, describing the same projects, were removed from the final selection. The selected papers were finally presented in Tables 3 (for conventional vehicles) and 4 for (autonomous vehicles and driving simulator studies), elaborated in Sections 3.1 and 3.2, respectively. The presented methodology is summarized in Fig. 1.

3. Literature findings

3.1. Data collection

In this section, the main review findings on data collection are presented, with an aim to answer the first research question on how driving behavior data is collected. These are based on the selected studies from the initial set of screened papers, where in–vehicle data was collected. Particularly, highlights are provided for used methods and equipment, features collected, and size and frequency of data. Studies selected for analysis are presented in details in Table 3 and are the ones mostly focusing on data collection processes aiming at driving behavior investigation.[3]

[3] In this table, highlights of the papers are presented, including useful findings (+), but also challenges or limitations (-). These highlights are of course based on a subjective classification by the authors of this paper, and some challenges (example the huge datasets collected) could be as well considered as great assets and strengths of the same studies.Distances reported to miles have been converted to kilometers (Kms) to only keep one unit in the table, for comparison purposes..

Table 2
Initial set of screened papers (automated vehicles)

Study	Cars	Buses	Bikes	Not specified	Data collection	Driving behavior	NDS	Statistical analysis	Big data
Shahrdar et al. (2019)	•				•				
Bloom et al. (2017), Rowley et al. (2018), Abberley et al. (2019)	•							•	
Hecht et al. (2020)	•				•	•			
Nica et al. (2019)	•				•			•	
Aldibaja et al. (2018), Chen et al. (2019)	•							•	•
Endsley (2017), Gao and Shi (2019), Gaspar and Carney (2019), Orlovska et al. (2020)	•				•	•	•		
Zhao et al. (2015)	•				•	•		•	
Park et al. (2019)	•				•	•			•
Sun et al. (2020)	•				•			•	•
Kouchak and Gaffar (2017)		•			•				
Huang et al. (2020)				•		•			
Seo et al. (2012), Balaji et al. (2017), Bai et al. (2018), Ghandour et al. (2019)				•				•	
Gurudatt and Umesh (2017), Tian et al. (2018), Kumar et al. (2019), Ma et al. (2019), Chen (2019)				•					•
Balado et al. (2019), Chao et al. (2020)				•	•			•	
Koo and Kim (2019)				•	•				•
Zec et al. (2018)				•				•	•

3.1.1. Methods and equipment

As previously mentioned, studies focusing on in–vehicle data collection, for the purpose of driving behavior analysis, are mostly field test trials, or naturalistic driving studies. The latter are studies where data is collected unobtrusively, by instrumenting vehicles and monitoring drivers' behavior, as they "normally" drive, including the collection of "baseline data", reflecting normal driving (Carsten, Kircher, & Jamson, 2013). The aim is to investigate associations between different variables, but also to extract risk factors in safety–critical events, and classify drivers according to different profiles. Such studies cover usually road transportation modes, particularly passenger car vehicles. In a simplified manner, the collected data covers different components, which will be presented here under: **vehicle data**, **environment and context data**, and **driver data**.

Vehicle data is collected through vehicle instrumentation, including video camera[4], and sensor technology, often integrated in a Data Acquisition System (DAS) in cars (Antin et al., 2019; Knoefel et al., 2018; Guo, Fang, & Antin, 2015; Simons-Morton et al., 2015; Carney, McGehee, Harland, Weiss, & Raby, 2015; Valero-Mora et al., 2013; Myers, Trang, & Crizzle, 2011; Fridman et al., 2019), trucks (Blanco et al., 2016; Hickman & Hanowski, 2012), and bikes (Dozza, Piccinini, & Werneke, 2016; Espié, Boubezoul, Aupetit, & Bouaziz, 2013). DAS often includes several units, cameras, and sensors like accelerometers, gyroscope and rate sensors, GPS, radar and radar interface box or RIB (Antin et al., 2019), and an OBD connector to measure on-board-diagnostics of the vehicle; sometimes audio data is recorded as well (Blanco et al., 2016).

External, context, or environment-related data is supplemental, out–of–vehicle data, which could include roadway (Victor et al., 2015) and weather information (Carney et al., 2015; Knoefel et al., 2018). While weather data can be measured in–vehicle by meteorological sensors, it can also be referred to as context or external data if obtained from other sources, and later merged to the existing data. This is also the case for instance for accidents datasets, which can be added a posteriori if obtained from police reports.

Finally, **driver data** pertains to drivers' demographics and health conditions, and includes questionnaires, assessments, or diaries, as often done in bike and truck experiments (Dozza et al., 2016), or even post-experiment interviews (Espié et al., 2013). Additionally, driver data can be collected from mobile phone records, where participants' mobile phones could be paired with the vehicles (Fridman et al., 2019).

3.1.2. Features collected

Distinct data types are collected from the methods and equipment used, allowing the collection of different features. Data collected can be classified under vehicle data, environmental or context data, and driver-related data. Vehicle data is mostly **dynamic data** (in–vehicle sensor data and video and images data); these are time–series data including kinematics variables or driving parameters such as: acceleration, speed, position on the road, distance to other cars, type of road, radar and GPS and computer data (Knoefel et al., 2018), yaw rate, network data (Guo et al., 2015), steering wheel rotation angle, brake pressure [as in Prologue (Valero-Mora et al., 2013)]. Video and image data can be collected from multiple cameras (forward, and rear windshields) providing images of the drivers'

[4] Although video data can record data from the road ahead or the drivers' faces, etc., this would still be classified as vehicle data, since the data source is the vehicle itself, as the camera is installed in the vehicle.

Table 3
Selected papers focusing on data collection aspects

	Mode	Data collection equipment					Features collected				Size/Frequency	Remarks
		Sensors	Cameras	GPS	Radar	OBD	Video/ Image	Vehicle Kinematics	Subjective data	Other		
Ma et al. (2021)	Cars, buses, trucks	•							•	•		(+) Objective and subjective factors were considered to analyze factors contributing to perceptual bias of aggressive driving (+) Objective factors include penalty points, subjective factors include self–assessment of aggressive driving.
Antin et al. (2019)	Cars SUVs, pickups, trucks	•	•	•	•	•	•	•	•	•	* Videos: 15 Hz * Cabin images: 1/10 min * Time-series: asynchronously * 51 M Kms of driving data: 5 PB of data.	(-) Large and complex database
Fridman et al. (2019)	Cars	•	•	•		•	•	•	•	•	*511 K Kms of driving data: 100 000 GB of data *7.1 billion video frames *CAN sensors: 1 GHz processor *Cameras 30 Hz *Data has to be time–stamped to allow perfect synchronization of multiple data streams in post-processing	(+) Computer vision–based analysis of human behavior (+) ADAS functions including ACC, pilot assist, blind spot monitor (+) Semi–automated annotation (-) Huge data
Warren et al. (2019)	Cars	•	•	•			•	•		•		(+) Phone sensors can complement traditional data collection techniques (+) Less costly and time consuming (+) In-phone sensors (+) Clustered drivers based on driving behavior: flag what deviates from the norm
Wijnands et al. (2019)	Cars	•	•				•	•		•	* 30 frames per second	(+) Detection approach on a mobile phone (+) Early fusion of spatial and temporal information (+) Balance between high prediction accuracy and real time inference requirements (+) Avoids computationally expensive pre-processing steps
Knoefel et al. (2018)	Cars	•	•	•		•	•	•		•	*15 M Kms: 1 TB data storage * GPS and computer data:>1 Hz	
Yadawadkar et al. (2018)	Cars	•	•		•	•	•	•	•			(+) Identifies driver distraction and drowsiness (+) Insights into data from collection from DAS to feature extraction (+) No video data (-) Data reductionists reviewed coded and evaluated events

(*continued on next page*)

Table 3 (*continued*)

	Mode	Data collection equipment					Features collected				Size/Frequency	Remarks
		Sensors	Cameras	GPS	Radar	OBD	Video/ Image	Vehicle Kinematics	Subjective data	Other		
												(-) Timing of data across variables asynchronous, leading to missing variables at each collection time point (-) Missing value replaced by last corresponding known value
Blanco et al. (2016)	Trucks	•	•	•	•		•	•	•	•	*1.2 M kms: 8 TB data storage	(+) Additional data from driver incident button, activity registers, extended medical assessments, and actigraphy or sleep devices (-) Data volume
Dozza et al. (2016)	E-bikes	•	•	•	•		•	•	•	•	*Sensor data: 100 Hz *Video data: 30 Hz, GPS data: 10 Hz	(+) Push-buttons for critical events, trip diaries, and post-experiment interviews help complementing objective data
Carney et al. (2015)	Cars						•	•	•	•	*Videos: 4 Hz	
Guo et al. (2015)	Cars	•	•	•			•	•				
Simons-Morton et al. (2015)	Cars	•	•	•			•	•	•	•		(-) Data volume
Espié et al. (2013)	Powered two-wheelers	•	•	•		•	•	•	•		*Vehicle dynamics: 1 kHz frame rate, with 4 μs time data stamping Video data: at 12.5 Hz GPS at 1 Hz.	(+) Combine subjective data with objective data (-) Cost
Guo and Fang (2013)	Cars	•	•	•	•		•	•	•			
Valero-Mora et al. (2013)	Cars	•	•	•	•		•				* Vehicle data: 100 Hz; synchronized automatically with the video data (25 Hz) * Eye-tracking data: 60 Hz; needs manual synchronization with vehicle and video data	(+) Highly instrumented vehicles can complement studies using a large number of standardized vehicles (-) Large amounts of data can be challenging to manage
Hickman and Hanowski (2012)	Trucks and buses	•	•	•			•	•				(+) On-board monitoring systems to identify safety–critical events
Ott et al. (2012)	Cars		•				•			•		(-) Uncontrollable environmental factors may affect the validity of the road test
Myers et al. (2011)	Cars			•		•	•	•	•	•		
Lin et al. (2008)	Taxis	•	•	•	•		•	•	•			(+) Investigation of causes of rear-end conflicts (-) Data volume
Neale et al. (2005)	Cars	•	•	•	•	•	•	•		•	* 3.2 M Kms * 43 K hours of data	(+) Hard drive large enough to store data for several weeks (+) Independent sensing systems (+) Detection systems for headway, side obstacle (+) Incident box for drivers to flag incidents

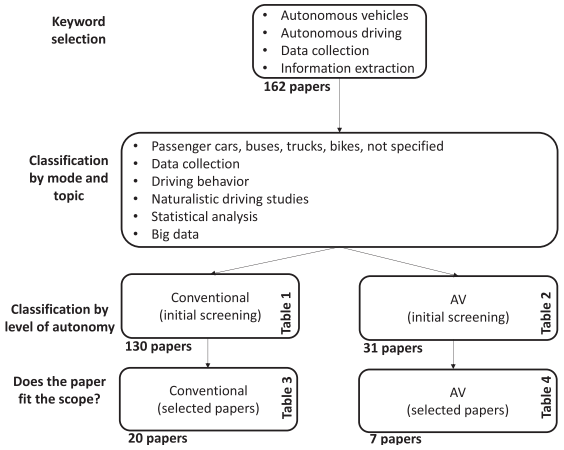

Fig. 1. Methodology for paper selection (*own illustration*)

face, or the cabin conditions as in Antin et al. (2019). In addition to video data, audio data is sometimes recorded (Carney et al., 2015). This data category can be considered dynamic, since it is recorded continuously and collected real-time. Supplemental data includes environmental and context data like maps, weather, or other data like roadway (workzone), data, or crash investigation or reports. Mobile phone records can also be used as an additional data source (Antin et al., 2019). Such data types (weather, roadway databases, etc.), cannot be considered real–time or continuous in the same manner as in-vehicle data, and therefore will be referred to as **static data** in this research. In particular, while map and weather data can be derived using GPS coordinates and can be registered and updated real–time, they are considered static in this representation, as usually, and based on previous research, their corresponding time–series are not usually used real–time for classifying driving behavior. As mentioned previously, both can be categorized as context data as they are used to enrich the existing datasets. For instance, weather data can be used as an indication of the task complexity, and it might be more interesting to know the weather condition, e.g. rainy or sunny, simply for a longer period of time, for instance a trip duration.

Finally, driver data includes characteristics from surveys, but also assessment or medical examinations. This data type will also be considered static, since it also does not change in a continuous real-time manner. For instance, Simons-Morton et al. (2015) administered a stress inducing test to test drivers' stress responsivity; while these test results can theoretically change, these tests and therefore their corresponding data are often collected only once (or more times) during the experiments and are therefore cross–sectional. Additionally, biometric data of the driver, such as heart rate data or other physiological measurements, can be continuously collected (using for instance wearables); this would then be considered as dynamic and objective data. The presented data (vehicle, environment, and driver) can be further classified into **objective data** (which does not depend on the drivers' own judgments and perceptions, but is rather collected through sensors, or other objective assessments), or **subjective data** [including self–reported information including participants' diaries, own points of views on safety–critical events through interviews or questionnaires, or even expert assessment of skills, and video coding of events (Hickman & Hanowski, 2012). Based on the collected data, features can be extracted covering mostly crash and near–crash data (Antin et al., 2019), and crash risk assessment (Knoefel et al., 2018). Safety–critical events are often calculated upon exceeding specific thresholds. For instance abnormal driving is triggered by high acceleration or other kinematic factors: Guo and Fang (2013) recorded 8 s before and 4 s after the trigger. In other words, going from the raw collected data, derived data is often calculated, by using statistical methods to evaluate risk or measurements of interest. For instance, statistical modeling of collected data can help reducing the data (e.g. PCA), or assess risk and driver profiles (Guo & Fang, 2013).

Other road transport modes collect similar features through comparable data collection equipment; for example in a truck study, both audio and video data were used, in addition to actigraphy devices to monitor sleep quantity, since fatigue is often a parameter of interest for professional drivers and long driving hours (Blanco et al., 2016). For Powered-two wheelers, participants' points of view are often of interest. Subjective data is therefore collected by interviewing participants after the experiments to better understand critical events (Espié et al., 2013; Dozza et al., 2016). Also, other modes often collect additional data that drivers themselves flag, when they see themselves in safety–critical situations, by pushing an incident button (Blanco et al., 2016; Dozza et al., 2016). Overall, to summarize the type and source of collected data, we can present it on two axes: on the x–axis, describing the frequency with which the data can be collected (grouped under static and dynamic), and on the y–axis, presenting whether the data is rather "unbiased" or more subject to personal judgments and perceptions (grouped under subjective and objective). This classification, stemming from analyzing previous research, can be useful for representing the different dimensions of the data and can be visualized in Fig. 2 below.

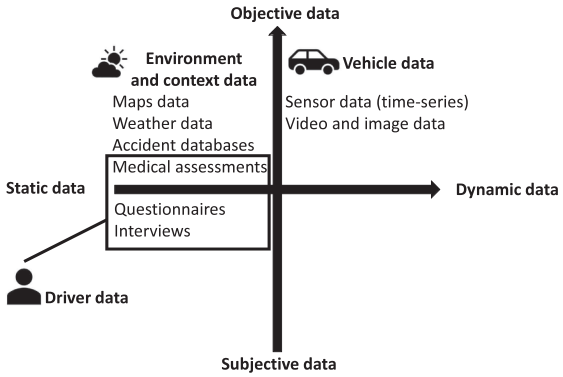

Fig. 2. Collected data by source and type (*own illustration*)

3.1.3. Size and frequency of data

Field operational tests and NDS often result in up to millions of kilometers of driving data, covering millions of trips, for an equivalent of hundreds of thousands of hours, which often translates into several thousands of crash or near-crash events[5]. As part of the Second Strategic Highway Research Program (SHRP 2), over 50 million kilometers of continuous data was collected from over 3500 drivers across the Unites States, an equivalent of over 900,000 h of in-vehicle time, and 5.5 million trips. The study captured more than 1900 light-vehicle crashes and 6900 near-crashes, an equivalent of five petabytes of data (Antin et al., 2019). In the Candrive study (Knoefel et al., 2018), data was collected data from 256 drivers in Ottawa, Canada, monitored for up to five years each, and amounting to a total of more than 15 million kilometers driven, the equivalent of one terabyte of storage data. The naturalistic teenage driving study (NTDS) itself collected 18 months of driving data (Simons-Morton et al., 2015). In the 100–car Naturalistic Driving study (Guo & Fang, 2013), data was collected throughout one year, resulting in three million vehicle kilometers, the equivalent of 43,000 h of data. Another study (Lin et al., 2008) collected data from 50 taxis in urban areas for 10 months using Video Drive Recorders (VDRs) in Beijing, China, collecting a total number of 2440 of valid events, including 40 accidents. Studies featuring other road vehicles also collected huge amounts of data. Dozza et al. (2016) collected 1500 km of biking data, including 88 critical events in Sweden in all environments. Hickman and Hanowski (2012) collected data from 183 commercial truck and bus fleets comprising 13,306 vehicles and included 1085 crashes, 8375 near–crashes, 30,661 crash-relevant conflicts, and 211,171 baseline events. Blanco et al. (2016) collected more than 14,500 driving hours of valid truck data from approximately 2,200 driving shifts and 26,000 on-duty hours of daily activity register data from more than one million kilometer of driving, an equivalent of eight terabytes of data storage. Overall, what these numbers can tell us is that collected in–vehicle data often results in several thousands of hours of driving data, millions of kilometers of data, and non-negligible storage needs.

However, for the above studies, the ratio of storage (in terabytes) to driving data collected (in hours) is not constant. This variation is due to the varying sensor frequencies, but also whether or not video data has been collected. This variation in frequency is a challenge for data collection and processing; sensors and cameras often collect data and images at different frequency. For instance, Dozza et al. (2016) collected data continuously at 100 Hz for all signals, video data at 30 Hz, and GPS data at 10 Hz. Valero-Mora et al. (2013) (PROLOGUE) also collected vehicle data at 100 Hz, video data at 25 Hz, and eyetracking data at 60 Hz. In this study, while vehicle data was automatically synchronized, eyetracking needed to be manually synchronized with vehicle and video data. In the 2BeSafe project (Espié et al., 2013), vehicle dynamics were collected at 1000 Hz, while video data was at 12.5 Hz and GPS data at 1 Hz. In SHRP 2 (Antin et al., 2019), video data was collected at a frequency of 15 Hz and sensor data at 10 Hz. In Candrive (Knoefel et al., 2018), GPS and computer data was collected at a frequency above 1 Hz. In Blanco et al. (2016), accelerometer frequency at 10 Hz. This only highlights the need for data synchronization for subsequent data analysis; for a perfect synchronization of multiple data streams in post/processing, data has to be timestamped (Fridman et al., 2019).

The main highlights of these studies, as noted in the "Remarks" column of Table 3 are: (i) data collection often results in a huge volume of data, which is challenging to manage, in terms of both time and costs, (ii) data quality is of utmost importance, e.g., missing data can be a challenge in asynchronous data, (iii) statistical techniques (data reduction, clustering, annotation and fusion of spatial and temporal info) can avoid computationally expensive pre–processing steps, (iv) phone sensors can complement traditional data collection techniques, (v) additional driver data (diaries, interviews, and flagged events) can help complement collected vehicle data and boosts interpretability.

3.2. Knowledge extraction

Having presented the highlights of data collection methods for behavioral modeling in Section 3.1, this section will present the

[5] In this section, the size and frequency of data collected often depend on the NDS based on which the studies were made, rather than being individual data collection studies made by the authors themselves

Table 4
Selected papers focusing on knowledge extraction

	Vehicle type	Driving behavior	Data collection			Prediction Models				Remarks
			Real-time	In-vehicle sensing	Sensor Fusion	Supervised Learning	Unsupervised Learning	Deep Learning	Other	
Mohammadnazar et al. (2021)	Connected	•	•	•	•		•			(+) Classifying driving styles by extracting volatility measures (+) K–means and K–medoids are used for grouping drivers under aggressive, normal, and calm clusters
Gite et al. (2019)		•	•	•	•		•	•		(+) Discusses signal quality of data for eye movement (+) Provides noise suppression method and smoothing filters to deal with noisy data: accuracy of extraction (+) Improves the image pattern recognition and prediction time (+) Gives a few extra seconds to anticipate the driver's correct action
Zhu et al. (2019)						•	•	•	•	(+) Summarizes techniques and approaches for big data in ITS (+) Importance of data collection quality: accuracy completeness, reliability (+) Need to invest in data collection technology (-) Data storage, processing, privacy
Guo et al. (2018)		•	•				•			(+) Can apply the networks to large scale GPS dataset to assess driver behavior and impacts (+) Improper vehicle lateral position maintenance, speeding and inconsistent or excessive acceleration and deceleration have been identified (-) Size of the data (3 TB) (-) Data may include user and system errors
Kamal et al. (2018)	Partially connected	•	•	•	•					(+) Method for highly anticipative driving. (+) Road–speed profile by extracting info form traffic big data broadcast from surrounding vehicles. (+) Accuracy evaluated for different penetration levels.
Zhao et al. (2017)		•	•	•	•			•		(+) Predicting steering angle of front wheel and speed of vehicle. (+) Using DBN, which proved to be more stable than BP NN.
Zhao et al. (2015)		•	•	•	•				•	(+) Data representation: converted sensor data into machine-understandable data and query to retrieve knowledge from the Knowledge Base. (+) Can be further extended by adding more knowledge such as traffic light data and traffic regulations to improve driving safety. (-) Shifts of GPS positions and missing number of lanes on some roads: test several times to find maximum shift distance to set up allowed shift threshold for finding out the position for updating target nodes (-) Delays of data transmission in the collected GPS sensor data

hline

findings of the review focusing on information extraction for autonomous vehicles, and aiming to answer the second research question. Accordingly, statistical methods used for driving maneuver prediction along with the validation measures have been analyzed in details; a summary of the results is given in Table 4[6]. Studies were differentiated by automation levels, as fully-autonomous cars (Zhao, Gong, Lu, Xiong, & Weijie, 2017; Rowley et al., 2018), semi-autonomous (Gite, Agrawal, & Kotecha, 2019), or simulated autonomous cars (Zhao, Ichise, Mita, & Sasaki, 2015). Real-time data collection (Gite et al., 2019; Zhao et al., 2017; Guo, Liu, Zhang, & Wang, 2018; Zhao et al., 2015) in general entailed in-vehicle sensing (Gite et al., 2019; Zhao et al., 2017; Zhao et al., 2015) where sensors were in general fused (Gite et al., 2019; Zhao et al., 2017; Rowley et al., 2018; Zhao et al., 2015). The selected studies for this section focus on data-driven information extraction methodology for driving behavior. Applications for driving behavior include early anticipation of the driver's maneuver (Gite et al., 2019), and driver's behavior prediction and risk patterns (Guo et al., 2018). Other studies on autonomous vehicle applications that did not particularly focus on driving behavior discussed different ITS applications (Zhu, Yu, Wang, Ning, & Tang, 2019), including a framework for speed limit detection and warnings (Zhao et al., 2015), or provided a testbed for autonomous cars, using statistics of historical data of crashes (Rowley et al., 2018).

The selected studies covered a range of data-driven methods including unsupervised learning, supervised learning, and deep learning methods. For instance, Guo et al. (2018) used a hybrid unsupervised learning framework by combining feature learning-Autoencoder- and feature clustering by Self Organizing Mapping (AESOM) to extract latent features and classify driving behavior. Deep learning was also quite popular for driving prediction for autonomous driving. This includes RNN-LSTM for driver's maneuver prediction (Gite et al., 2019), and Deep Belief Network (DBN) (Zhao et al., 2017). In the latter, sensors including camera, lidar, wave radar, GPS, accelerometer were used to collect environmental information. The authors used location, speed of surrounding vehicles and speed and steering angle of vehicle of previous time to predict speed and steering angle of vehicle at the current moment. Models were usually evaluated using performance metrics like precision, recall and accuracy (Gite et al., 2019) The key findings identified in the analysis are the importance of data quality (in terms of reliability and completeness (Zhu et al., 2019), resolution (Rowley et al., 2018), or due to system and user errors (Guo et al., 2018), but also of data storage, privacy, and processing (Zhu et al., 2019). To deal with the quality of the data and signals, methods have been proposed such as noise suppression method and smoothing filters to deal with noisy data (Gite et al., 2019), which could result into improving image pattern recognition, and prediction time, and therefore translate into having a few seconds more to anticipate the driver's correct action. The challenges identified in this section also pertain to the size of the data (Guo et al., 2018), as mentioned in Section 3.1; the advantage though of such statistical methods would be the possibility to apply them to larger scales datasets to assess driver behavior and impacts, as discussed in Guo et al. (2018).

3.3. Lessons learned and proposed contribution

The analyzed studies in Sections 3.1 and 3.2 pointed out to the opportunities provided by data collection studies, and the available methods for extracting information and features, needed to predict driving behavior, which is essential for assessing the impacts of AVs. However, the review pointed out to challenges of the data, such as its size, which imposes the development of certain protocols to control data quality (for both data collection, and data processing), but also to reduce data and extract relevant features. In the case of autonomous vehicles, the goal would be that the vehicle component can accurately predict maneuvers and based on certain behaviors (in the last couple of seconds), engage in the safe one at every moment. A set of data–driven models can improve model predictions for different types of behavior, steering angle, eye, and face tracking, etc.

While the insights provided from the literature helped in giving an overview on methods for data collection and information extraction for driving behavior modeling, there is still a gap in representing these different components in a holistic framework including various aspects that need to be considered. This paper will therefore propose a data–knowledge cycle, based on the above findings, aiming to better represent features of analytics and fusion, for driving behavior modeling.

4. Proposed data–information–knowledge framework

4.1. Data analytics framework

By looking at different components from the moment data is collected, until it becomes useful for behavioral modeling, and further assessments, a data analytics framework can be drawn, and is visualized in Fig. 3. In this framework, the data collection component includes data captured in (this case) in-vehicle experiments where in-vehicle sensors collect data continuously/realtime for monitoring driving behavior. This can include GPS, cameras, sensors which often do not collect data at the same frequency. This would then inevitably include pre–processing of the data to ensure first that there are no proper communication issues and that signals correctly measure data and make it available, but then also processing it and fusing different sensors where possible, to ensure time-stamped synchronization. In experiments, data often includes subjective data as well like questionnaires from participants or drivers, which would then have to be properly linked to the field data; these however are not dynamic in general, or at least less dynamic, and so they would need to be managed differently.

Data processing includes aspects of data quality, which can be checked through different methods, ranging from filtering, noise cleaning, to manually controlling for consistency in the collection; for example for eyetracking measurements, synchronization is done

[6] In this table, highlights of the papers are presented, including useful findings (+), but also challenges or limitations (-)

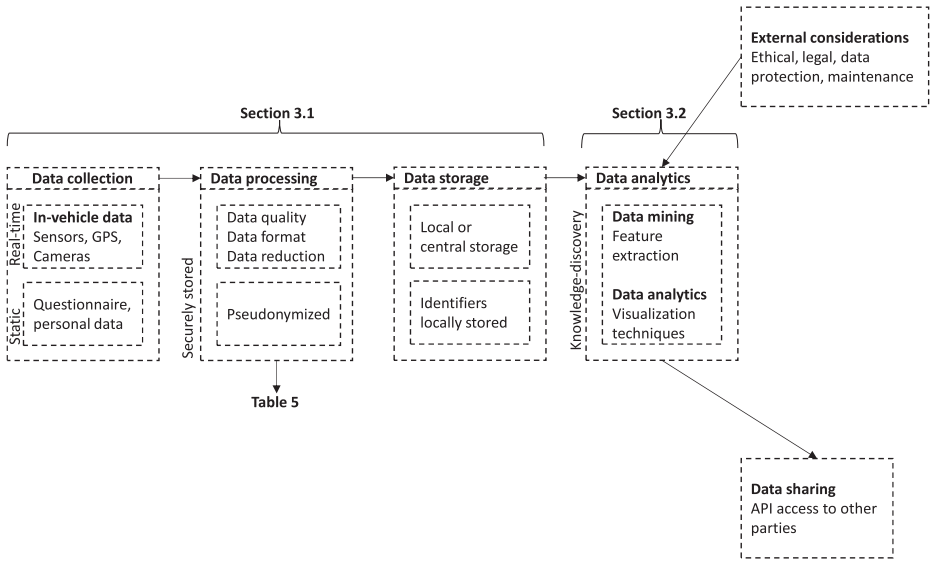

Fig. 3. Proposed data analytics framework *(own illustration)*

Table 5
Suggested data processing tasks

Data processing task	Description
Data quality	Labeling or encoding data from test vehicles Handling missing data (sensor and communication failure) Temporal order for time-series: needed to deal with possible network requests from the collection end to the cloud server that do not arrive in the correct order, or when data is received by the server, but it's acknowledgement does not reach the data collection end Handling the timezone information carefully Data verification for errors (removing outliers and irrelevant data, cleaning datasets, rectifying GPS data) In case of inconsistency, the vehicle data logger should be checked to recognize and fix issues as soon as possible According to the desired format A description of the data variables should be provided by the technical partners generating the data and should be sufficient for future reference
Data reduction	Reducing data volume mostly for video data. Video data may be pre-processed in a way to reduce data volume without compromising the quality of the video Metadata of the videos (event, timestamps, trip info etc.) should also be attached with each video for ease of future analysis
Data pseudonymization	Assigning a unique identifier for each participant to comply with GDPR, and linking the data from participants to vehicle data

with vehicle and video data.

Data mining can include methods like classification and clustering, feature extraction using Machine Learning methods, pattern recognition, predictive analysis, and visualization techniques with dashboard-based elements. The idea would then be that once data is made available, data could be processed in such a way to predict the needs of the drivers accurately and safely.

The different components presented also need to follow ethical, legal, and privacy standards of the country where the collection is taking place. Looking at previous studies, we can see a pattern in data management where ethical and legal considerations are at the backbone of data collection. Data handling as well, including data storage, and sharing, would need to follow specific standards; in Europe, this means a compliance with the EU Regulation 2016/679, or the General Data Protection Regulation (GDPR), which came in effect from 25 May 2018 (European Commission, 2018), aiming at protecting personal data. Protocols of anonymization or pseudonymization of data at the source should therefore be part of the framework. For instance, for data storage, different techniques exist either involving private or public storage, which depend on the usability and purpose. For instance, personal and identifiable data should be locally stored (not publicly), for complying with GDPR. Only pseudonymized data can be associated with the vehicle data and stored in the public storage (pseudonymized or anonymized, depending on regulations). Similarly, for data sharing (and eventually maintenance), different access levels may be defined, according to defined agreements, in order to make different parts of the data accessible to different parties. Specific processing tasks and their descriptions are suggested and elaborated in Table 5.

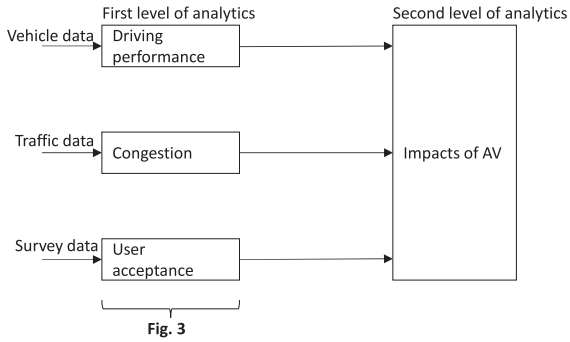

Fig. 4. Proposed data fusion framework adapted from Akbar et al. (2018) *(own illustration)*

4.2. Data fusion framework

While the data analytics framework presented in Section 4.1 described data fusion processes, these were only at sensor levels, as part of pre-processing or processing steps. A major challenge that has not been addressed is the process of combining heterogeneous data, in a way to obtain meaningful information, and extract an additional layer of information. When thinking of data collected within experiments such as NDS or pilot studies for AVs, the heterogeneity of data can be translated into elements of driving data, questionnaire data, but also other contextual data (traffic data, accident data). A combination or fusion of information is therefore necessary to develop models that can answer the defined research questions, for instance, automation acceptance, users' interactions with AVs, and AV impacts on a network level; for the latter, for instance, heterogeneous findings of AV user acceptance and congestion, can be input by means of corresponding metrics to the network simulation, when testing the impact of different AV penetration levels. Data fusion can therefore be achieved at several levels: at a sensor level, or after the first layer of analytics.

Akbar et al. (2018) developed a methodology with two levels of analytics, where events were defined from individual data streams in the first level, to probabilistic complex events after the second level; this can also be referred to as the fusion of these various events, using Bayesian Networks (BNs). In the first level, events of interest are defined and extracted in real time, while in the second level, BNs can take uncertainty while detecting complex events. Specifically for this study Akbar et al. (2018), used data streams included traffic, weather, and social media data streams from Madrid, Spain. The approach used followed a hybrid framework based on complex event processing (CEP) and Bayesian Networks (BNs) to extract high-level knowledge in the form of probabilistic complex event (in this case the probability of congestion in real-time). The approach was qualitatively (using web-interface) and quantitatively evaluated using F-measure showing an accuracy of over 80%. A generalized framework for fusion of different data streams, adapted from Akbar et al. (2018), is depicted in Fig. 4. In this proposed framework, for instance, different information sources (such as vehicle data, traffic data, or survey data) can be used in order to modify input parameters of well–known driving behavior models, in order to see their impacts on a broader network level. For instance, traffic safety or traffic efficiency can be investigated by finding the total network travel time, or the number of conflicts in the network, based on the different penetration rates of the AVs in the network. This can be done following an experimental design and by changing different input parameters which are obtained at the first level of analytics. A similar assessment or experimental design (using a full factorial design) has been used by Kostovasili and Antoniou (2017) using SUMO traffic simulation including parameters for driving aggressiveness (speed acceptance, maximum acceleration, normal deceleration, reaction time). A similar approach would allow to generate different scenarios and assess the impacts of different driving behaviors on a network level. Examples of human factors that can be incorporated in different car–following models has been extensively reviewed by Saifuzzaman and Zheng (2014).

Akiwowo and Eftekhari (2013) also used Bayesian data fusion to improve false positive rates for cocaine detection. After pre-processing the raw data and identifying and extracting relevant features, feature outputs were used for decision and as input into a Bayesian data fusion module, which then output the probability that a sample belonged to a class based on the observed features; decision was made based on the class with the higher probability. The results showed that the Bayesian fusion module greatly improves the detection rates of individual feature. In the process, the authors defined multiple data fusion levels including single sensor fusion at a sensor level (different sampling rates and raw data combined to be able to extract features), feature level, and decision level. In the context of modeling driving behavior data, events can be derived depending on data streams and objectives. Data fusion can be of interest after the analytics phase (sensor level fusion would already take place in the processing component of the data analytics). Driving behavior data presented in this paper mostly included vehicle data, survey data, but could eventually include other data types which could enrich the existing knowledge layer, for instance social media data. An inference using a similar approach as Akbar et al. (2018) can be used to estimate the probability of AV acceptance using pilot data for driving behavior and acceptance, enriched by additional data streams (e.g. social media, to infer the general perception towards AVs for example).

4.3. Transferability across modes

While this study focused on road transportation, the presented frameworks can possibly be extended to other transportation modes. Though limited, studies researching driving behavior in other transportation modes include similar equipment and collect data that is

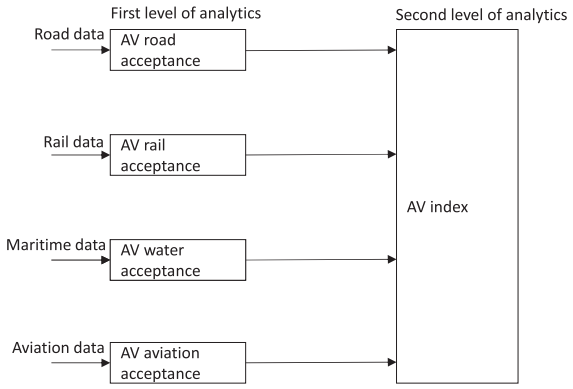

Fig. 5. Data fusion framework across different transportation modes, adapted from Akbar et al. (2018) *(own illustration)*

similar to the one for road transport (depicted in Fig. 2). Although automation is well–established in other modes, there are several opportunities to transfer knowledge between transport modes, as highlighted in Papadimitriou et al. (2020).

Rail studies for instance also include objective and subjective data, such as GPS data, surveys (Larue & Wullems, 2019; Guo, Wei, Liao, & Chu, 2016), which can help evaluating rail driving behavior at crossings, or even video analytics (Zaman, Liu, & Zhang, 2018). Such studies also aim to assess risky behavior, or crash or near-crash data, using also advanced analytics algorithms.

In the maritime sector, most studies remain not behavioral, and target mostly monitoring systems and detection, or predicting destinations (Park, Koo, & Kim, 2018; Kim & Lee, 2019). Still multi-sensor data Gao and Shi (2019) was used for positioning, navigation status, speed information, etc., and key features were extracted, with benefits for ship traffic flow and navigational behavior learning (providing a foundation for subsequent research on ship handling behavior and intelligent ship collision avoidance).

Also in the flying sector, while limited, studies describe the data collection process for driving behavior monitoring, a research study proposed a framework to be used for flying naturalistic data, including multi-channel video sensors to measure pilot behavior, and external sensors to measure flight operational data (Oh, 2017). Similarly to road transport, the collected information would include driver data (in this case the pilot data), vehicle data (here, flight operational data like current location, altitude, attitude, air speed, real-time fuel burn data), and environment or context data (here external data like weather data, air traffic data, and other data connected to a central information centre).

Considering the knowledge that could possibly be gained by instrumenting vehicles for different transport modes (or by conducting pilots for AVs in different modes), extracted knowledge from each mode could then be combined to create an overall transferable finding. For instance, in case the objective is to develop an index for AV acceptance, while that could be a first level of analytics for a transportation mode, a fusion of multiple indices across different modes could result in an overall AV index. For example, different field experiments or surveys can give insights into the acceptance for AVs in a given region. While most research is done for road–based AVs, this could also be relevant for the acceptance of AVs for other modes of transport, such as rail, water, or air transportation modes. While for the latter, less interaction between the operator and vehicles is expected in any case, still there might be some relevant insights that could be found on the acceptance of automation for these modes. Such insights on the trust of automation for professional drivers can help transport planners better understand or assess the acceptance for these modes in different cities or regions. Theoretically then, a first level of analytics for AV acceptance would help assess AV acceptance of different transport modes. Taking into account the rich information provided by this first level of analytics, an overall AV index could then be drawn from these different indices found, highlighting the factors influencing this acceptance for instance, such as trust, or relevant demographic variables. This is depicted in Fig. 5, which was also drawn based on the principles described in Akbar et al. (2018). While that might have challenges and obvious limitations (such as the assumptions drawn for such results to hold true; for instance the need of consistent pilot data, collected in areas with comparable populations, like the same city or country), the aim of this example was to rather provide an insight on how findings of heterogeneous types could and should be exploited; transport modes can considerably learn from each other mostly in terms of automation and trust (Papadimitriou et al., 2020).

5. Conclusion

Having extensively reviewed studies on data collection and information extraction for driving behavior modeling, this study has highlighted several findings and helped answer the research questions drawn in Section 1. Data collection is often done by means of various methods and equipment, often combining vehicle data, environment and context data, and driver data. These can be further classified as dynamic or static, subjective, or objective, depending on the features collected. In understanding the nature of this data, it is important to note that driver data is at the intersection between objective and subjective data, but also static and dynamic data. Findings of the review point out to challenges of the data, such as its size, which imposes the development of certain protocols to control data quality (for both data collection, and data processing), but also to reduce data and extract the relevant features.

In the case of autonomous vehicles, the goal would be that the vehicle component can accurately predict maneuvers and based on certain behaviors (in the last couple of seconds), engage in the safe one in each moment. Different levels of automation, semi, full, or

even driving simulators, have been used in studies with real time data collection, in–vehicle sensing and sensor fusion, aiming to classify driving behavior. In these studies, a set of data-driven models (supervised learning, unsupervised learning, and deep learning methods) were used to improve model prediction for different types of behavior, steering angle, eye, and face tracking, etc.

After identifying the different methods for data collection and information extraction, this paper proposed a data-knowledge cycle, in order to better represent features of analytics and fusion, for driving behavior modeling. The first component is a data analytics framework, starting from a data collection component (with different sources of data: static, dynamic, etc.), followed by a data processing component (with detailed suggested tasks for quality, format, reduction, and data pseudonymization for data protection purposes), then a data storage component (with different storage strategies), and ending with a data mining and analytics component. Additionally, overarching principles or external considerations including ethical, legal, and data protection, overrule and provide guidelines for the different components, eg., pseudonymization before storing and uploading the data, but also regarding data sharing and access to other parties etc.

Besides the data analytics framework, data fusion methods adapted from Akbar et al. (2018) were highlighted for use according to the desired objectives. In this manuscript, an example of impact of AVs has been used as second level of analytics, with vehicle, survey, and traffic data, as a first level of analytics. Finally, the paper provided insights for transferring these findings to other modes. Despite limited research in other sectors focusing on driving behavior, different transport modes can arguably learn from each other, as suggested in Papadimitriou et al. (2020). The study however does not come without limitations. The keywords choice in the review inevitably influences the obtained papers and therefore findings. Moreover, the applicability of the results can be challenging as it would require a large scale study, where such frameworks become useful. Future research could extend these findings to a multimodal context, one parameter of interest might be for instance AV acceptance, where a first layer analytics could be the AV acceptance for each sector (e.g. for each of road, rail, maritime, and air transport), and the overall or second layer of analytics could be a certain overall AV acceptance index. Future work could use these frameworks of analytics and fusion, enriching possibly the former with additional considerations, and using the latter according to the defined objectives and drawn layers of analytics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 815001 (project DriveToTheFuture).

References

Abberley, L., Crockett, K., & Cheng, J. (2019). Modelling road congestion using a fuzzy system and real-world data for connected and autonomous vehicles. In *2019 Wireless Days (WD)* (pp. 1–8). https://doi.org/10.1109/WD.2019.8734238

Abodo, F., Rittmuller, R., Sumner, B., & Berthaume, A. (2018). Detecting Work Zones in SHRP 2 NDS Videos Using Deep Learning Based Computer Vision. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 679–686). https://doi.org/10.1109/ICMLA.2018.00108

Adornato, B., Patil, R., Filipi, Z., Baraket, Z., & Gordon, T. (2009). Characterizing naturalistic driving patterns for plug-in hybrid electric vehicle analysis. In *2009 IEEE Vehicle Power and Propulsion Conference* (pp. 655–660). https://doi.org/10.1109/VPPC.2009.5289786

Ahlstrom, C., Victor, T., Wegean, C., & Erik, S. (2012). Processing of eye/head-tracking data in large-scale naturalistic driving data sets. *IEEE Transactions on Intelligent Transportation Systems, 13*, 553–564. https://doi.org/10.1109/TITS.2011.2174786

Aihara, K., Bin, P., & Imura, H. (2019). On the relationship between accuracy of bus position estimated by crowdsourcing and participation density. In N. Streitz, & S. Konomi (Eds.), *Distributed, Ambient and Pervasive Interactions* (pp. 101–112). Cham: Springer International Publishing.

Akbar, A., Kousiouris, G., Pervaiz, H., Sancho, J., Ta-Shma, P., Carrez, F., & Moessner, K. (2018). Real-time probabilistic data fusion for large-scale iot applications. *IEEE Access, 6*, 10015–10027.

Akiwowo, A., & Eftekhari, M. (2013). Feature-based detection using bayesian data fusion. *International Journal of Image and Data Fusion, 4*, 308–323.

Al-Najada, H., & Mahgoub, I. (2017). Real-time incident clearance time prediction using traffic data from internet of mobility sensors. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)* (pp. 728–735).

Aldibaja, M., Suganuma, N., Yoneda, K., Yanase, R., & Kuramoto, A. (2018). Supervised calibration method for improving contrast and intensity of lidar laser beams. In S. Lee, H. Ko, & S. Oh (Eds.), *Multisensor Fusion and Integration in the Wake of Big Data, Deep Learning and Cyber Physical System* (pp. 210–218). Cham: Springer International Publishing.

Alekseenko, A., Dang, H. Q., Bansal, G., Sanchez-Medina, J., & Miyajim, C. (2019). ITS+DM Hackathon (ITSC 2017): Lane Departure Prediction With Naturalistic Driving Data. *IEEE Intelligent Transportation Systems Magazine, 11*, 78–93. https://doi.org/10.1109/MITS.2018.2880264

Antin, J. F., Lee, S., Perez, M. A., Dingus, T. A., Hankey, J. M., & Brach, A. (2019). Second strategic highway research program naturalistic driving study methods. *Safety Science, 119*, 2–10. https://doi.org/10.1016/j.ssci.2019.01.016. http://www.sciencedirect.com/science/article/pii/S0925753518301012.

Arvin, R., Kamrani, M., & Khattak, A. J. (2019). The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. *Accident Analysis & Prevention, 132*, 105226. https://doi.org/10.1016/j.aap.2019.07.002. http://www.sciencedirect.com/science/article/pii/S0001457519306517.

Bachechi, C., & Po, L. (2019). *Implementing an urban dynamic traffic model. In IEEE/WIC/ACM International Conference on Web Intelligence WI'19* (pp. 312–316). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3350546.3352537

Bagdadi, O. (2013). Assessing safety critical braking events in naturalistic driving studies. *Transportation Research Part F: Traffic Psychology and Behaviour, 16*, 117–126. https://doi.org/10.1016/j.trf.2012.08.006. http://www.sciencedirect.com/science/article/pii/S1369847812000770.

Bai, Z., Cai, B., ShangGuan, W., & Chai, L. (2018). Deep learning based motion planning for autonomous vehicle using spatiotemporal lstm network. In *2018 Chinese Automation Congress (CAC)* (pp. 1610–1614). IEEE.

Balado, J., Martínez-Sánchez, J., Arias, P., & Novo, A. (2019). Road environment semantic segmentation with deep learning from mls point cloud data. *Sensors, 19*, 3466.

Balaji, Y., Kumar, M. B., & Sujatha, Y. (2017). Text information extraction and analysis for autonomous vehicle. In *2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)* (pp. 1–6). https://doi.org/10.1109/SPICES.2017.8091281

Barbier, C., Guyonvar'ch, L., Guillaume, A., & Tattegrain, H. (2019). Is the self-confrontation method applicable to naturalistic driving studies? *Safety Science, 119*, 29–39. https://doi.org/10.1016/j.ssci.2018.11.005. http://www.sciencedirect.com/science/article/pii/S0925753518300055.

Barnard, Y., Utesch, F., Nes, N. v., Eenink, R., & Baumann, M. (2016). The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters. volume 8. doi:10.1007/s12544-016-0202-z.

Barr, L.C., Yang, C.Y.D., Hanowski, R.J., & Olson, R.L. (2011). An Assessment of Driver Drowsiness, Distraction, and Performance in a Naturalistic Setting.

Bellini, P., Bilotta, p., Stefano andNesi, Paolucci, M., & Soderi, M. (2018). Real-time traffic estimation of unmonitored roads. In 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech) (pp. 935–942).

Blanco, M., Hickman, J.S., Olson, R.L., Bocanegra, J.L., Greening, M., Madison, P.H., Holbrook, G.T., Hanowski, R.J., Nakata, A., & Bowman, D. (2016). Investigating Critical Incidents, Driver Restart Period, Sleep Quantity, and Crash Countermeasures in Commercial Vehicle Operations Using Naturalistic Data Collection.

Blatt, A., Pierowicz, J., Flanigan, M., Lin, P.-S., Kourtellis, A., Lee, C., Jovanis, P., Jenness, J., Wilaby, M., Campbell, J. et al. (2015). Naturalistic driving study: Field data collection. Technical Report.

Bloom, C., Tan, J., Ramjohn, J., & Bauer, L. (2017). Self-driving cars and data collection: Privacy perceptions of networked autonomous vehicles. In Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017) (pp. 357–375).

Bosi, I., Ferrera, E., Brevi, D., & Pastrone, C. (2019). In-vehicle iot platform enabling the virtual sensor concept: A pothole detection use-case for cooperative safety. In Proceedings of the 4th International Conference on Internet of Things, Big Data and Security - Volume 1: IoTBDS, (pp. 232–240). INSTICC SciTePress. doi: 10.5220/0007690602320240.

Carney, C., McGehee, D., Harland, K., Weiss, M., & Raby, M. (2015). Using naturalistic driving data to assess the prevalence of environmental factors and driver behaviors in teen driver crashes.

Carsten, O., Kircher, K., & Jamson, S. (2013). Vehicle-based studies of driving in the real world: The hard truth? *Accident Analysis & Prevention, 58*, 162–174.

Chao, Q., Bi, H., Li, W., Mao, T., Wang, Z., Lin, M.C., & Deng, Z. (2020). A survey on visual traffic simulation: Models, evaluations, and applications in autonomous driving. Computer Graphics Forum,. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13803. doi:10.1111/cgf.13803. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13803.

Chen, S., Leng, Y., & Labi, S. (2019). A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information. Computer-Aided Civil and Infrastructure Engineering, https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12495. doi:10.1111/mice.12495. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/mice.12495.

Chen, S.-C. (2019). Multimedia for autonomous driving. *IEEE MultiMedia, 26*, 5–8.

Cheng, G., Wang, Z., & Zheng, J. Y. (2017). Big-video mining of road appearances in full spectrums of weather and illuminations. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1–6). https://doi.org/10.1109/ITSC.2017.8317601

Chhabra, R., Verma, S., & Rama Krishna, C. (2019). Detecting aggressive driving behavior using mobile smartphone. In C. R. Krishna, M. Dutta, & R. Kumar (Eds.), *Proceedings of 2nd International Conference on Communication, Computing and Networking* (pp. 513–521). Singapore: Springer Singapore.

Chun, S., Hamidi Ghalehjegh, N., Choi, J., Schwarz, C., Gaspar, J., McGehee, D., & Baek, S. (2019). Nads-net: A nimble architecture for driver and seat belt detection via convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Das, A., Khan, M. N., & Ahmed, M. M. (2020). Detecting lane change maneuvers using shrp2 naturalistic driving data: a comparative study machine learning techniques. *Accident Analysis & Prevention, 142*, 105578.

Davis, J. D., Papandonatos, G. D., Miller, L. A., Hewitt, S. D., Festa, E. K., Heindel, W. C., & Ott, B. R. (2012). Road test and naturalistic driving performance in healthy and cognitively impaired older adults: Does environment matter? *Journal of the American Geriatrics Society, 60*, 2056–2062. https://doi.org/10.1111/j.1532-5415.2012.04206.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2012.04206.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1532-5415.2012.04206.x.

Dawson, J. D. (2019). Practical and statistical challenges in driving research. *Statistics in Medicine, 38*, 152–159. https://doi.org/10.1002/sim.7903. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7903, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7903.

Ding, N., Zh, S.u., Wan, H.g., & Jiao, N. (2019). Effects of reverse linear perspective of transverse line markings on car-following headway: A naturalistic driving study. *Safety Science, 119*, 50–57. https://doi.org/10.1016/j.ssci.2018.08.021. http://www.sciencedirect.com/science/article/pii/S0925753517320118.

Dingus, T., Klauer, S., Lewis, V., Petersen, A., Lee, S., Sudweeks, J., Perez, M., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z., Jermeland, J., & Knipling, R. (2006). The 100-car naturalistic driving study: Phase ii - results of the 100-car field experiment.

Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences, 113, 2636–2641. https://www.pnas.org/content/113/10/2636. doi:10.1073/pnas.1513271113. arXiv:https://www.pnas.org/content/113/10/2636.full.pdf.

Dozza, M., Piccinini, G.F.B., & Werneke, J. (2016). Using naturalistic data to assess e-cyclist behavior. Transportation Research Part F: Traffic Psychology and Behaviour, 41, 217–226. http://www.sciencedirect.com/science/article/pii/S1369847815000662. doi: 10.1016/j.trf.2015.04.003. Bicycling and bicycle safety.

Dozza, M., & Werneke, J. (2014). Introducing naturalistic cycling data: What factors influence bicyclists' safety in the real world? *Transportation Research Part F: Traffic Psychology and Behaviour, 24*, 83–91. https://doi.org/10.1016/j.trf.2014.04.001. http://www.sciencedirect.com/science/article/pii/S1369847814000394.

Ehsani, J. P., Seymour, K. E., Chirles, T., & Kinnear, N. (2020). Developing and testing a hazard prediction task for novice drivers: A novel application of naturalistic driving videos. *Journal of Safety Research*. https://doi.org/10.1016/j.jsr.2020.03.010. http://www.sciencedirect.com/science/article/pii/S0022437520300402.

Endsley, M. R. (2017). Autonomous driving systems: A preliminary naturalistic study of the tesla model s. *Journal of Cognitive Engineering and Decision Making, 11*, 225–238. https://doi.org/10.1177/1555343417695197. arXiv:https://doi.org/10.1177/1555343417695197.

Espié, S., Boubezoul, A., Aupetit, S., & Bouaziz, S. (2013). Data collection and processing tools for naturalistic study of powered two-wheelers users' behaviours. *Accident Analysis & Prevention, 58*, 330–3398. https://doi.org/10.1016/j.aap.2013.03.012. http://www.sciencedirect.com/science/article/pii/S0001457512002485.

European Commission (2018). Data protection in the EU. https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en.

Fan, Y., Zhou, L., Fan, L., & Yang, J. (2019). Multiple obstacle detection for assistance driver system using deep neural networks. In *International Conference on Artificial Intelligence and Security* (pp. 501–513). Springer.

Fernandez-Rojas, R., Perry, A., Singh, H., Campbell, B., Elsayed, S., Robert, H., & Abbass, A. H. (2019). Contextual awareness in human-advanced-vehicle systems: A survey. *IEEE Access, 7*, 33304–33328. https://doi.org/10.1109/ACCESS.2019.2902812

Figueiras, P., Herga, Z., Guerreiro, G., Rosa, A., Costa, R., & Jardim-Gonçalves, R. (2018). Real-time monitoring of road traffic using data stream mining. In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)* (pp. 1–8).

Fitch, G.M., Grove, K., Hanowski, R.J., & Perez, M.A. (2014). Compensatory behavior of drivers when conversing on a cell phone: Investigation with naturalistic driving data. Transportation Research Record, 2434, 1–8. doi: 10.3141/2434-01. doi:10.3141/2434-01. arXiv:https://doi.org/10.3141/2434-01.

Foss, A. H., & Goodwin, Robert D. (2014). *Distracted driver behaviors and distracting conditions among adolescent drivers: Findings from a naturalistic driving study. jadohealth* (p. 54). https://doi.org/10.1016/j.jadohealth.2014.01.005

Fridman, L., Brown, D. E., Glazer, M., Angell, W., Dodd, S., Jenik, B., Terwilliger, J., Patsekin, A., Kindelsberger, J., Li Ding, S. S., Mehler, A., Sipperley, A., Pettinato, A., Seppelt, B., Angell, L., Mehler, B., & Reimer, B. (2019). MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction With Automation. *IEEE Access, 7*, 102021–102038. https://doi.org/10.1109/ACCESS.2019.2926040

Gao, M., & Shi, G.-Y. (2019). Ship spatiotemporal key feature point online extraction based on ais multi-sensor data using an improved sliding window algorithm. Sensors, 19. https://www.mdpi.com/1424-8220/19/12/2706. doi:10.3390/s19122706.

Gaspar, J., & Carney, C. (2019). The effect of partial automation on driver attention: A naturalistic driving study. *Human Factors, 61*, 1261–1276. https://doi.org/10.1177/0018720819836310. PMID: 30920852, arXiv:https://doi.org/10.1177/0018720819836310.

Ghandour, A., Krayem, H., & Gizzini, A. (2019). Autonomous vehicle detection and classification in high resolution satellite imagery. doi:10.1109/ACIT.2018.8672712.

Gite, S., Agrawal, H., & Kotecha, K. (2019). Early anticipation of driver's maneuver in semiautonomous vehicles using deep learning. *Progress in Artificial Intelligence, 8*, 293–305. https://doi.org/10.1007/s13748-019-00177-z

Gohar, M., Muzammal, M. M., & Rahman, A. U. (2018). Smart tss: Defining transportation system behavior using big data analytics in smart cities. *Sustainable Cities and Society, 41*, 114–119. https://doi.org/10.1016/j.scs.2018.05.008. http://www.sciencedirect.com/science/article/pii/S2210670717309757.

Guan, Z.-W., Liu, X.-Y., Yang, L., Zhao, H.-L., Liu, X.-F., & Du, F. (2019). Using loop detector big data and artificial intelligence to predict road network congestion. In W. Wang, K. Bengler, & X. Jiang (Eds.), *Green Intelligent Transportation Systems* (pp. 179–187). Singapore: Springer Singapore.

Guleng, S., Wu, C., Yoshinaga, T., & Ji, Y. (2019). Traffic big data assisted broadcast in vehicular networks. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems RACS '19* (pp. 236–240). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3338840.3355683.

Guo, F. (2019). Statistical methods for naturalistic driving studies. *Annual Review of Statistics and Its Application, 6*, 309–328. https://doi.org/10.1146/annurev-statistics-030718-105153. arXiv:https://doi.org/10.1146/annurev-statistics-030718-105153.

Guo, F., & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention, 61*, 3–9. https://doi.org/10.1016/j.aap.2012.06.014. http://www.sciencedirect.com/science/article/pii/S0001457512002382.

Guo, F., Fang, Y., & Antin, J.F. (2015). Older driver fitness-to-drive evaluation using naturalistic driving data. Journal of Safety Research, 54, 49.e29–54. http://www.sciencedirect.com/science/article/pii/S0022437515000456. doi: 10.1016/j.jsr.2015.06.013. Strategic Highway Research Program (SHRP 2) and Special Issue: Fourth International Symposium on Naturalistic Driving Research.

Guo, F., Klauer, S.G., Hankey, J.M., & Dingus, T.A. (2010). Near crashes as crash surrogate for naturalistic driving studies. Transportation Research Record, 2147, 66–74. doi: 10.3141/2147-09. doi:10.3141/2147-09. arXiv:https://doi.org/10.3141/2147-09.

Guo, J., Liu, Y., Zhang, L., & Wang, Y. (2018). Driving behaviour style study with a hybrid deep learning framework based on gps data. *Sustainability, 10*. https://doi.org/10.3390/su10072351. https://www.mdpi.com/2071-1050/10/7/2351.

Guo, M., Wei, W., Liao, G., & Chu, F. (2016). The impact of personality on driving safety among chinese high-speed railway drivers. *Accident Analysis & Prevention, 92*, 9–14. https://doi.org/10.1016/j.aap.2016.03.014. http://www.sciencedirect.com/science/article/pii/S0001457516300859.

Gurudatt, P. K., & Umesh, V. (2017). Novel architecture for cloud based next gen vehicle platform–transition from today to 2025. In *2017 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)* (pp. 151–155).

Hallmark, S.L., Tyner, S., Oneyear, N., Carney, C., & McGehee, D. (2015). Evaluation of driving behavior on rural 2-lane curves using the shrp 2 naturalistic driving study data. Journal of Safety Research, 54, 17.e1–27. http://www.sciencedirect.com/science/article/pii/S0022437515000493. doi: 10.1016/j.jsr.2015.06.017. Strategic Highway Research Program (SHRP 2) and Special Issue: Fourth International Symposium on Naturalistic Driving Research.

Hecht, T., Feldhtter, A., Draeger, K., & Bengler, K. (2020). What do you do? an analysis of non-driving related activities during a 60minutes conditionally automated highway drive. In R. C. S. C. A. Ahram, & Tareqand Taiar (Eds.), *Human Interaction and Emerging Technologies* (pp. 28–34). Cham: Springer International Publishing.

Hickman, J. S., & Hanowski, R. J. (2012). An assessment of commercial motor vehicle driver distraction using naturalistic driving data. *Traffic Injury Prevention, 13*, 612–619. https://doi.org/10.1080/15389588.2012.683841. PMID: 23137092, arXiv:https://doi.org/10.1080/15389588.2012.683841.

Hochin, T., Shinohara, Y., & Nishizaki, Y. (2019). Detection of driver's eye fixation on a moving target by using line fitting. In *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science Engineering (BCD)* (pp. 94–99).

Huang, X., Zhang, S., & Peng, H. (2020). Developing robot driver etiquette based on naturalistic human driving behavior. *IEEE Transactions on Intelligent Transportation Systems, 21*, 1393–1403. https://doi.org/10.1109/TITS.2019.2913102

Itkonen, T. H., Lehtone, E., & Selpi. (2020). Characterisation of motorway driving style using naturalistic driving data. *Transportation Research Part F: Traffic Psychology and Behaviour, 69*, 72–79. https://doi.org/10.1016/j.trf.2020.01.003. http://www.sciencedirect.com/science/article/pii/S136984781930419X.

Jacob, J., & Rabha, P. (2018). Driving data collection framework using low cost hardware. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jonasson, J. K., & Rootzén, H. (2014). Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach. *Accident Analysis & Prevention, 62*, 102–109. https://doi.org/10.1016/j.aap.2013.09.013. http://www.sciencedirect.com/science/article/pii/S0001457513003667.

Jovanis, P. P., Aguero-Valverde, J., Wu, K.-F., & Shankar, V. N. (2011). *Analysis of naturalistic driving event data*.

Kamal, M. A. S., Hayakawa, T., & Imura, J.-I. (2018). Road-speed profile for enhanced perception of traffic conditions in a partially connected vehicle environment. *IEEE Transactions on Vehicular Technology, 67*, 6824–6837.

Kang, M. J., Kwon, O. H., & Park, S. H. (2019). Development of a crash risk prediction model using the k-nearest neighbor algorithm. In J. J. Park, V. Loia, K.-K. R. Choo, & G. Yi (Eds.), *Advanced Multimedia and Ubiquitous Engineering* (pp. 835–840). Singapore: Springer Singapore.

Kaur, S., Singh, S., & Kaur, D. (2019). Frequency regulation in smart grids using electric vehicles considering real-time pricing. In C. R. Krishna, M. Dutta, & R. Kumar (Eds.), *Proceedings of 2nd International Conference on Communication, Computing and Networking* (pp. 323–334). Singapore: Springer Singapore.

Kaushik, K., Wood, E., & Gonder, J. (2018). Coupled Approximation of U.S. Driving Speed and Volume Statistics using Spatial Conflation and Temporal Disaggregation. *Transportation Research Record, 2672*, 1–11. https://doi.org/10.1177/0361198118758391. arXiv:https://doi.org/10.1177/0361198118758391.

Kim, K. I., & Lee, K. M. (2019). Data-driven prediction of ship destinations in the harbor area using deep learning. In W. Lee, & C. K. Leung (Eds.), *Big Data Applications and Services 2017* (pp. 81–90). Singapore: Springer Singapore.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J., & Ramsey, D.J. (2006). The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data.

Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England Journal of Medicine, 370*, 54–59. https://doi.org/10.1056/NEJMsa1204142. PMID: 24382065, arXiv:https://doi.org/10.1056/NEJMsa1204142.

Knoefel, F., Wallace, B., Goubran, R., & Marshall, S. (2018). Naturalistic driving: A framework and advances in using big data. *Geriatrics, 3*. https://doi.org/10.3390/geriatrics3020016. URL https://www.mdpi.com/2308-3417/3/2/16.

Koo, Y., & Kim, S. (2019). Distributed logging system for ros-based systems. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 1–3).

Koppel, S., Liu, P., Griffiths, D., Hua, P., Stephan, K., Logan, D., Porter, M., Mazer, B., Gélinas, I., Vrkljan, B., Marshall, S., & Charlton, J. (2020). A comparison of older drivers' driving patterns during a naturalistic on-road driving task with patterns from their preceding four-months of real-world driving. *Safety Science, 125*, 104652. https://doi.org/10.1016/j.ssci.2020.104652. http://www.sciencedirect.com/science/article/pii/S0925753520300497.

Kostovasili, M., & Antoniou, C. (2017). Simulation-based evaluation of evacuation effectiveness using driving behavior sensitivity analysis. *Simulation Modelling Practice and Theory, 70*, 135–148.

Kouchak, S.M., & Gaffar, A. (2017). Determinism in future cars: Why autonomous trucks are easier to design. In 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1–6). doi:10.1109/UIC-ATC.2017.8397598.

Kovaceva, J., Isaksson-Hellman, I., & Murgovski, N. (2020). Identification of aggressive driving from naturalistic data in car-following situations. *Journal of Safety Research*. https://doi.org/10.1016/j.jsr.2020.03.003. http://www.sciencedirect.com/science/article/pii/S0022437520300335.

Kovaceva, J., Nero, G., Bärgman, J., & Dozza, M. (2019). Drivers overtaking cyclists in the real-world: Evidence from a naturalistic driving study. *Safety Science, 119*, 199–206. https://doi.org/10.1016/j.ssci.2018.08.022. http://www.sciencedirect.com/science/article/pii/S0925753517321008.

Kumar, S. S., Babu, R. M., Vineeth, R., Varun, S., Sahil, A. N., & Sharanraj, S. (2019). Autonomous traffic light control system for smart cities. In S.-L. Peng, N. Dey, & M. Bundele (Eds.), *Computing and Network Sustainability* (pp. 325–335). Singapore: Springer Singapore.

Kuo, J., Lenné, M. G., Mulhall, M., Sletten, T., Anderson, C., Howard, M., Rajaratnam, S., Magee, M., & Collins, A. (2019). Continuous monitoring of visual distraction and drowsiness in shift-workers during naturalistic driving. *Safety Science, 119*, 112–116. https://doi.org/10.1016/j.ssci.2018.11.007. http://www.sciencedirect.com/science/article/pii/S092575351830211X.

Larue, G. S., & Wullems, C. (2019). A new method for evaluating driver behavior and interventions for passive railway level crossings with pneumatic tubes. *Journal of Transportation Safety & Security, 11*, 150–166. https://doi.org/10.1080/19439962.2017.1365316. arXiv:https://doi.org/10.1080/19439962.2017.1365316.

Lee, S.E., Olsen, E.C.B., & Wierwille, W.W. (2004). A comprehensive examination of naturalistic lane-changes.

Li, G., Wang, Y., Zhu, F., Sui, X., Wang, N., Qu, X., & Green, P. (2019). Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in china. *Journal of Safety Research, 71*, 219–229. https://doi.org/10.1016/j.jsr.2019.09.012. http://www.sciencedirect.com/science/article/pii/S0022437519306267.

Li, R., Brand, H., Gopinath, A., Kamarajugadda, S., Yang, L., Wang, W., & Li, B. (2020). Driver drowsiness behavior detection and analysis using vision-based multimodal features for driving safety. In *WCX SAE World Congress Experience*. SAE International. https://doi.org/10.4271/2020-01-1211.

Li, S., Li, P., Yao, Y., Han, X., Xu, Y., & Chen, L. (2020). Analysis of drivers' deceleration behavior based on naturalistic driving data. *Traffic Injury Prevention, 21*, 42–47. https://doi.org/10.1080/15389588.2019.1707194. PMID: 31986072, arXiv:https://doi.org/10.1080/15389588.2019.1707194.

Liang, X. (2020). Research on the correlation of dangerous driving behaviors based on naturalistic driving experiment. *IOP Conference Series: Materials Science and Engineering, 780*, 072034. https://doi.org/10.1088/1757-899x/780/7/072034. https://doi.org/10.1088%2F1757-899x%2F780%2F7%2F072034.

Liang, Y., Lee, J. D., & Yekhshatyan, L. (2012). How dangerous is looking away from the road? algorithms predict crash risk from glance patterns in naturalistic driving. *Human factors, 54*(6), 1104–1116.

Lin, Q., Feng, R., Cheng, B., Lai, J., Zhang, H., & Mei, B. (2008). Analysis of causes of rear-end conflicts using naturalistic driving data collected by video drive recorders. In SAE Technical Paper. SAE International. URL https://doi.org/10.4271/2008-01-0522. doi:10.4271/2008-01-0522.

Liu, X., & Li, C. (2019). An intelligent urban traffic data fusion analysis method based on improved artificial neural network. *Journal of Intelligent & Fuzzy Systems, 37*, 4413–4423.

Ma, S., Zhang, X., Wang, S., Zhang, X., Jia, C., & Wang, S. (2019). Joint feature and texture coding: Toward smart video representation via front-end intelligence. *IEEE Transactions on Circuits and Systems for Video Technology, 29*, 3095–3105.

Ma, Y., Gu, X., Yu, Y., Khattakc, A. J., Chen, S., Tang, K., et al. (2021). Identification of contributing factors for driver's perceptual bias of aggressive driving in china. *Sustainability, 13*, 766.

McLaughlin, S. B., Hankey, J. M., & Dingus, T. A. (2008). A method for evaluating collision avoidance systems using naturalistic driving data. *Accident Analysis & Prevention, 40*, 8–16. https://doi.org/10.1016/j.aap.2007.03.016. http://www.sciencedirect.com/science/article/pii/S0001457507000632.

Mishra, R., Fu, X., He, S., Du, J., Wang, X., & Ge, T. (2020). Variations in naturalistic driving behavior and visual perception at the entrances of short, medium, and long tunnels. *Journal of Advanced Transportation*. https://doi.org/10.1155/2020/7630681

Mo, W., Gao, Y., & Zhao, Q. (2017). Confusable vehicle feature extraction and recognition based on cascaded svm. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)* (pp. 2154–2158). https://doi.org/10.1109/CompComm.2017.8322918

Mohammadnazar, A., Arvin, R., & Khattak, A. J. (2021). Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. *Transportation research part C: emerging technologies, 122*, 102917.

Moharm, K. I., Zidane, E. F., El-Mahdy, M. M., & El-Tantawy, S. (2019). Big data in its: Concept, case studies, opportunities, and challenges. *IEEE Transactions on Intelligent Transportation Systems, 20*, 3189–3194. https://doi.org/10.1109/TITS.2018.2868852

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., et al. (2010). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Int J Surg, 8*, 336–341.

Montgomery, J., Kusano, K. D., & Gabler, H. C. (2014). Age and gender differences in time to collision at braking from the 100-car naturalistic driving study. *Traffic Injury Prevention, 15*, S15–S20. https://doi.org/10.1080/15389588.2014.928703. PMID: 25307380, arXiv:https://doi.org/10.1080/15389588.2014.928703.

Moorthy, A., De Kleine, R., Keoleian, G., Good, J., & Lewis, G. (2017). Shared autonomous vehicles as a sustainable solution to the last mile problem: A case study of ann arbor-detroit area. *SAE International Journal of Passenger Cars-Electronic and Electrical Systems, 10*, 328–336.

Morgenstern, T., Schott, L., & Krems, J. F. (2020). Do drivers reduce their speed when texting on highways? a replication study using european naturalistic driving data. *Safety Science, 128*, 104740. https://doi.org/10.1016/j.ssci.2020.104740. http://www.sciencedirect.com/science/article/pii/S0925753520301375.

Muronga, K., & Ruxwana, N. (2017). Naturalistic driving studies: The effectiveness of the methodology in monitoring driver behaviour.

Myers, A. M., Trang, A., & Crizzle, A. M. (2011). Naturalistic study of winter driving practices by older men and women: Examination of weather, road conditions, trip purposes, and comfort. *Canadian Journal on Aging/ La Revue canadienne du vieillissement, 30*, 577–589. https://doi.org/10.1017/S0714980811000481

Nair, G. S., & Bhat, C. R. (2021). Sharing the road with autonomous vehicles: Perceived safety and regulatory preferences. *Transportation research part C: emerging technologies, 122*, 102885.

Nallaperuma, D., Nawaratne, R., Bandaragoda, T., Adikari, A., & Nguyen, S. (2019). Online incremental machine learning platform for big data-driven smart traffic management. *IEEE Transactions on Intelligent Transportation Systems, 20*, 4679–4690.

Neale, V.L., Dingus, T.A., Klauer, S., Sudweeks, J., & Goodman, M.J. (2005). An overview of the 100-car naturalistic study and findings.

Nica, A. N., Trascau, M., Rotaru, A. A., Andreescu, C., Sorici, A., Florea, A. M., & Bacue, V. (2019). Collecting and processing a self-driving dataset in the upb campus. In *2019 22nd International Conference on Control Systems and Computer Science (CSCS)* (pp. 202–209). IEEE.

Oh, C.-G. (2017). Naturalistic flying study as a method of collecting pilot communication behavior data. In *2017 Cognitive Communications for Aerospace Applications Workshop (CCAA)* (pp. 1–4). https://doi.org/10.1109/CCAAW.2017.8001876

Ohnemus, M., & Perl, A. (2016). Shared autonomous vehicles: Catalyst of new mobility for the last mile? *Built Environment, 42*, 589–602.

Orlovska, J., Novakazi, F., Lars-Ola, B., Karlsson, M., Wickman, C., & Söderberg, R. (2020). Effects of the driving context on the usage of automated driver assistance systems (adas) -naturalistic driving study for adas evaluation. Transportation Research Interdisciplinary. *Perspectives, 4*, 100093. https://doi.org/10.1016/j.trip.2020.100093. http://www.sciencedirect.com/science/article/pii/S259019822030004X.

Ott, B. R., Papandonatos, G. D., Davis, J. D., & Barco, P. P. (2012). Naturalistic validation of an on-road driving test of older drivers. *Human Factors, 54*, 663–674. https://doi.org/10.1177/0018720811435235. PMID: 22908688, arXiv:https://doi.org/10.1177/0018720811435235.

Papadimitriou, E., Schneider, C., Tello, J. A., Damen, W., Vrouenraets, M. L., & Ten Broeke, A. (2020). Transport safety and human factors in the era of automation: What can transport modes learn from each other? *Accident Analysis & Prevention, 144*, 105656.

Park, M., Koo, Y., & Kim, S. (2018). Motion control block implementation for driving computing system. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 653–656). https://doi.org/10.1109/BigComp.2018.00118

Park, S.J., Hong, S., Kim, D., Hussain, I., & Seo, Y. (2019). Intelligent in-car health monitoring system for elderly drivers in connected car. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) (pp. 40–44). Cham: Springer International Publishing.

Patil, R., Adornato, B., & Filipi, Z. (2009). Impact of naturalistic driving patterns on phev performance and system design. In SAE Technical Paper. SAE International. URL https://doi.org/10.4271/2009-01-2715. doi:10.4271/2009-01-2715.

Petzoldt, T. (2020). Drivers' behavioural (non)adaptation after a texting-related crash. *Safety Science, 127*, 104715. https://doi.org/10.1016/j.ssci.2020.104715. http://www.sciencedirect.com/science/article/pii/S0925753520301120.

Piedad, E. J., Le, T.-T., Aying, K., Pama, F. K., & Tabale, I. (2019). Vehicle Count System based on Time Interval Image Capture Method and Deep Learning Mask R-CNN. In *TENCON 2019–2019 IEEE Region 10 Conference (TENCON)* (pp. 2675–2679). https://doi.org/10.1109/TENCON.2019.8929426

Pop, M.-D., & Prostean, O. (2019). Bayesian Reasoning for OD Volumes Estimation in Absorbing Markov Traffic Process Modeling. In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1–6). https://doi.org/10.1109/ICBDSC.2019.8645611

Precht, L., Keinath, A., & Krems, J. F. (2017). Identifying effects of driving and secondary task demands, passenger presence, and driver characteristics on driving errors and traffic violations - using naturalistic driving data segments preceding both safety critical events and matched baselines. *Transportation Research Part F:*

Traffic Psychology and Behaviour, 51, 103–144. https://doi.org/10.1016/j.trf.2017.09.003. http://www.sciencedirect.com/science/article/pii/S1369847817304631.

Pucci, P., & Vecchio, G. (2019). Big data: Hidden challenges for a fair mobility planning. In *Enabling Mobilities: Planning Tools for People and Their Mobilities* (pp. 43–58). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-19581-6_4.

Rasch, A., Panero, G., Boda, C.-N., & Dozza, M. (2020). How do drivers overtake pedestrians? Evidence from field test and naturalistic driving data. *Accident Analysis & Prevention, 139*, 105494. https://doi.org/10.1016/j.aap.2020.105494. http://www.sciencedirect.com/science/article/pii/S0001457519305391.

Rosales, A., and; Guojun Wang, M.Z.A.B., Xing, T.W. X., & Alelaiwi, A. (2017). Naturalistic driving data for a smart cloud-based abnormal driving detector. In 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1–8). doi:10.1109/UIC-ATC.2017.8397449.

Rowley, J., Liu, A., Sandry, S., Gross, J., Salvador, M., Anton, C., & Fleming, C. (2018). Examining the driverless future: An analysis of human-caused vehicle accidents and development of an autonomous vehicle communication testbed. (pp. 58–63). doi:10.1109/SIEDS.2018.8374759.

Saifuzzaman, M., & Zheng, Z. (2014). Incorporating human-factors in car-following models: a review of recent developments and research needs. *Transportation research part C: emerging technologies, 48*, 379–403.

Samiee, S., Azadi, S., Kazemi, R., Nahvi, A., & Eichberger, A. (2014). Data fusion to develop a driver drowsiness detection system with robustness to signal loss. *Sensors, 14*, 17832–17847. https://doi.org/10.3390/s140917832. https://www.mdpi.com/1424-8220/14/9/17832.

Sangster, J., Rakha, H., & Du, J. (2013). Application of naturalistic driving data to modeling of driver car-following behavior. Transportation Research Record, 2390, 20–33. doi: 10.3141/2390-03. doi:10.3141/2390-03. arXiv:https://doi.org/10.3141/2390-03.

Sayer, J.R., Devonshire, J.M., & Flannagan, C.A.C. (2005). The effects of secondary tasks on naturalistic driving performance.

Schatzinger, S., & Lim, C. Y. R. (2017). Taxi of the future: Big data analysis as a framework for future urban fleets in smart cities. In A. Bisello, D. Vettorato, R. Stephens, & P. Elisei (Eds.), *Smart and Sustainable Planning for Cities and Regions: Results of SSPCR 2015* (pp. 83–98). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-44899-2_6.

Seo, Y.-W., Wettergreen, D., & Zhang, W. (2012). Recognizing temporary changes on highways for reliable autonomous driving. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 3027–3032). https://doi.org/10.1109/ICSMC.2012.6378255

Serok, N., Levy, O., Havlin, S., & Blumenfeld-Lieberthal, E. (2019). Unveiling the inter-relations between the urban streets network and its dynamic traffic flows: Planning implication. *Environment and Planning B: Urban Analytics and City Science, 46*, 1362–1376. https://doi.org/10.1177/2399808319837982

Shahrdar, S., Park, C., & Nojoumian, M. (2019). Human trust measurement using an immersive virtual reality autonomous vehicle simulator. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 515–520).

Shankar, V., Jovanis, P.P., Aguero-Valverde, J., & Gross, F. (2008). Analysis of naturalistic driving data: Prospective view on methodological paradigms. Transportation Research Record, 2061, 1–8. doi: 10.3141/2061-01. doi:10.3141/2061-01. arXiv:https://doi.org/10.3141/2061-01.

Simmons, S. M., Hicks, A., & Cair, J. K.d. (2016). Safety-critical event risk associated with cell phone tasks as measured in naturalistic driving studies: A systematic review and meta-analysis. *Accident Analysis & Prevention, 87*, 161–169. https://doi.org/10.1016/j.aap.2015.11.015. http://www.sciencedirect.com/science/article/pii/S0001457515301305.

Simons-Morton, B.G., Klauer, S.G., Ouimet, M.C., Guo, F., Albert, P.S., Lee, S.E., Ehsani, J.P., Pradhan, A.K., & Dingus, T.A. (2015). Naturalistic teenage driving study: Findings and lessons learned. Journal of Safety Research, 54, 41.e29–44. http://www.sciencedirect.com/science/article/pii/S0022437515000420. doi: 10.1016/j.jsr.2015.06.010. Strategic Highway Research Program (SHRP 2) and Special Issue: Fourth International Symposium on Naturalistic Driving Research.

Sivasankaran, S.K., & Balasubramanian, V. (2019). Data mining based analysis of hit-and-run crashes in metropolitan city. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) (pp. 113–122). Cham: Springer International Publishing.

Soccolich, S. A., Blanco, M., Hanowski, R. J., Olson, R. L., Morgan, J. F., Guo, F., & Wu, S.-C. (2013). An analysis of driving and working hour on commercial motor vehicle driver safety using naturalistic data collection. *Accident Analysis & Prevention, 58*, 249–258. https://doi.org/10.1016/j.aap.2012.06.024. http://www.sciencedirect.com/science/article/pii/S0001457512002485.

Sun, C., Liu, W., Chu, D., Li, W., Lu, Z., & Wang, J. (2018). A novel method of symbolic representation in diving data mining: A case study of highways in china. Concurrency and Computation: Practice and Experience, 30, e4976. https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.4976. doi:10.1002/cpe.4976. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.4976. E4976 CPE-18-0859.R1.

Sun, C., Vianney, J. M. U., Li, Y., Chen, L., Li, L., Wang, F.-Y., Khajepour, A., & Cao, D. (2020). Proximity based automatic data annotation for autonomous driving. *IEEE/CAA Journal of Automatica Sinica, 7*, 395–404.

Thapa, R., Hallmark, S., Smadi, O., & Goswamy, A. (2019). Assessing driving behavior upstream of work zones by detecting response points in speed profile: A naturalistic driving study. *Traffic Injury Prevention, 20*, 854–859. https://doi.org/10.1080/15389588.2019.1663348. PMID: 31647333, arXiv:https://doi.org/10.1080/15389588.2019.1663348.

Tian, J., Chin, A., & Yanikomeroglu, H. (2018). Connected and autonomous driving. *IT Professional, 20*, 31–34. https://doi.org/10.1109/MITP.2018.2876928

Tian, R., Li, L., Yang, K., Chien, S., Chen, Y., & Sherony, R. (2014). Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on tasi 110-car naturalistic driving data collection. In *2014 IEEE Intelligent Vehicles Symposium Proceedings* (pp. 623–629). https://doi.org/10.1109/IVS.2014.6856599

Tivesten, E., & Dozz, M. (2014). Driving context and visual-manual phone tasks influence glance behavior in naturalistic driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 26*, 258–272. https://doi.org/10.1016/j.trf.2014.08.004. http://www.sciencedirect.com/science/article/pii/S1369847814001211.

Tivesten, E., & Dozza, M. (2015). Driving context influences drivers' decision to engage in visual-manual phone tasks: Evidence from a naturalistic driving study. *Journal of Safety Research, 53*, 87–96. https://doi.org/10.1016/j.jsr.2015.03.010. http://www.sciencedirect.com/science/article/pii/S0022437515000225.

Torre-Bastida, A.I., and; Ibai Laöa, J.D.S., Ilardia, M., Bilbao, M.N., & Campos-Cordobés, S. (2018). Big data for transportation and mobility: recent advances, trends and challenges. IET Intelligent Transport Systems, 12, 742–755. doi:10.1049/iet-its.2018.5188.

Valero-Mora, P. M., Tontsch, A., Welsh, R., Morris, A., Reed, S., Touliou, K., & Margaritis, D. (2013). Is naturalistic driving research possible with highly instrumented cars? lessons learnt in three research centres. *Accident Analysis & Prevention, 58*, 187–194. https://doi.org/10.1016/j.aap.2012.12.025. http://www.sciencedirect.com/science/article/pii/S0001457512004472.

Victor, T., Dozza, M., Bärgman, J., Boda, C.-N., Engström, J., Flannagan, C., Lee, J.D., & Markkula, G. (2015). Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk. Technical Report.

Vu, A., Yang, Q., Farrell, J.A., & Barth, M. (2013). Traffic sign detection, state estimation, and identification using onboard sensors. In 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013) (pp. 875–880). doi:10.1109/ITSC.2013.6728342.

Wallace, B., Knoefel, F., Goubran, R.A., Porter, M.M., Smith, A.W.B., & Marshall, S. (2017). Features that distinguish drivers: Big data analytics of naturalistic driving data.

Wallace, B., Puli, A., Goubran, R., Knoefel, F., Marshall, S., Porter, M., & Smith, A. (2016). Measurement of distinguishing features of stable cognitive and physical health older drivers. *IEEE Transactions on Instrumentation and Measurement, 65*, 1990–2001. https://doi.org/10.1109/TIM.2016.2526617

Wang, G., Sun, P., & Zhang, Y. (2019). Utilizing random forest and neural network to extract lane change events on shanghai highway. In CICTP 2019 (pp. 318–330).

Wang, Y., & Ho, I. W.-H. (2018). Joint deep neural network modelling and statistical analysis on characterizing driving behaviors. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1–6). https://doi.org/10.1109/IVS.2018.8500376

Warren, J., Lipkowitz, J., & Sokolov, V. (2019). Clusters of driving behavior from observational smartphone data. *IEEE Intelligent Transportation Systems Magazine, 11*, 171–180. https://doi.org/10.1109/MITS.2019.2919516

Wege, C., Wil, S.l., & Victor, T. (2013). Eye movement and brake reactions to real world brake-capacity forward collision warnings–a naturalistic driving study. *Accident Analysis & Prevention, 58*, 259–270. https://doi.org/10.1016/j.aap.2012.09.013. http://www.sciencedirect.com/science/article/pii/S000145751200320X.

Wijnands, J. S., Thompson, J., Nice, K. A., Aschwanden, G. D., & Stevenson, M. (2019). Real-time monitoring of driver drowsiness on mobile platforms using 3d neural networks. *Neural Computing and Applications*, 1–13.

Wu, K.-F., & Jovanis, P. P. (2012). Crashes and crash-surrogate events: Exploratory modeling with naturalistic driving data. *Accident Analysis & Prevention, 45*, 507–516. https://doi.org/10.1016/j.aap.2011.09.002. http://www.sciencedirect.com/science/article/pii/S0001457511002399.

Wu, K.-F., & Jovanis, P.P. (2013). Defining and screening crash surrogate events using naturalistic driving data. Accident Analysis & Prevention, 61, 10–22. http://www.sciencedirect.com/science/article/pii/S0001457512003600. doi: 10.1016/j.aap.2012.10.004. Emerging Research Methods and Their Application to Road Safety Emerging Issues in Safe and Sustainable Mobility for Older Persons The Candrive/Ozcandrive Prospective Older Driver Study: Methodology and Early Study Findings.

Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., & Whitney, D. (2018). Predicting driver attention in critical situations. In *Asian conference on computer vision* (pp. 658–674). Springer.

Yadawadkar, S., Mayer, B., Lokegaonkar, S., Islam, M. R., Ramakrishnan, N., Song, M., & Mollenhauer, M. (2018). Identifying distracted and drowsy drivers using naturalistic driving data. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 2019–2026). IEEE.

Yan, Y., Dai, Y., Li, X., Tang, J., & Guo, Z. (2019). Driving risk assessment using driving behavior data under continuous tunnel environment. *Traffic injury prevention, 20*, 807–812.

Yasmin, S., Hu, J., & Luo, S. (2020). A car-following driver model capable of retaining naturalistic driving styles. *Journal of Advanced Transportation*. https://doi.org/10.1155/2020/6520861

Zaman, A., Liu, X., & Zhang, Z. (2018). Video analytics for railroad safety research: An artificial intelligence approach. *Transportation Research Record, 2672*, 269–277. https://doi.org/10.1177/0361198118792751. arXiv:https://doi.org/10.1177/0361198118792751.

Zec, E. L., Mohammadiha, N., & Schliep, A. (2018). Statistical sensor modelling for autonomous driving using autoregressive input-output hmms. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1331–1336). IEEE.

Zhang, J., Ma, Z., Zhu, X., & Lin, Y. (2019). Analysis of driving control model of normal lane change based on naturalistic driving data. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 104–109).

Zhao, C., Gong, J., Lu, C., Xiong, G., & Weijie, M. (2017). Speed and steering angle prediction for intelligent vehicles based on deep belief network. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 301–306). https://doi.org/10.1109/ITSC.2017.8317929

Zhao, L., Ichise, R., Mita, S., & Sasaki, Y. (2015). An ontology-based intelligent speed adaptation system for autonomous cars. In T. Supnithi, T. Yamaguchi, J. Z. Pan, V. Wuwongse, & M. Buranarach (Eds.), *Semantic Technology* (pp. 397–413). Cham: Springer International Publishing.

Zhao, S., Zhao, Q., Bai, Y., & Li, S. (2019). A traffic flow prediction method based on road crossing vector coding and a bidirectional recursive neural network. *Electronics, 8*. https://www.mdpi.com/2079-9292/8/9/1006.

Zhou, T., Shi, W., Liu, X., Tao, F., Qian, Z., & Zhang, R. (2019). A novel approach for online car-hailing monitoring using spatiotemporal big data. *IEEE Access, 7*, 128936–128947. https://doi.org/10.1109/ACCESS.2019.2939787

Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2019). Big data analytics in intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems, 20*, 383–398. https://doi.org/10.1109/TITS.2018.2815678