Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications

Alex Neilson[a], Indratmo[a,*], Ben Daniel[b], Stevanus Tjandra[c]

[a] *Department of Computer Science, MacEwan University, Canada*
[b] *Higher Education Development Centre, University of Otago, New Zealand*
[c] *Traffic Safety Section, City of Edmonton, Canada*

Abstract

Research in Big Data and analytics offers tremendous opportunities to utilize evidence in making decisions in many application domains. To what extent can the paradigms of Big Data and analytics be used in the domain of transport? This article reports on an outcome of a systematic review of published articles in the last five years that discuss Big Data concepts and applications in the transportation domain. The goal is to explore and understand the current research, opportunities, and challenges relating to the utilization of Big Data and analytics in transportation. The review shows the potential of Big Data and analytics to garner insights and improve transportation systems through the analysis of various forms of data obtained from traffic monitoring systems, connected vehicles, crowdsourcing, and social media. We discuss some platforms and software architecture for the transport domain, along with a wide array of storage, processing, and analytical techniques, and describe challenges associated with the implementation of Big Data and analytics. This review contributes broadly to the various ways in which cities can utilize Big Data in transportation to guide the creation of sustainable and safer traffic systems. Since research in Big Data and transportation is, by and large, at infancy, this article does not prescribe recommendations to the various challenges identified, which also constitutes the limitation of the article.

*Corresponding author
Email address: indratmo@macewan.ca (Indratmo)

Keywords: Big Data, smart city, intelligent transportation system, connected vehicle, road traffic safety, Vision Zero

1. Introduction

Research in Big Data and analytics offers opportunities to apply the evidence-based approach to decision-making in many domains. In the transportation domain, Big Data has the potential to improve the safety and sustainability of transportation systems. Many cities have installed monitoring equipment, such as cameras, roadside sensors, and wireless sensor networks, to observe traffic conditions and promote traffic safety. This equipment collects a massive amount of traffic data and enables transportation departments to gain a better understanding of traffic flow in the respective areas. The availability of traffic data allows both historical and streaming data analysis, which can reveal meaningful traffic patterns, identify congestion, and assist in understanding the causes of collisions or near misses. The various forms of analytics and approaches employed in Big Data, such as machine learning, can be used to sift through the vast amount of traffic data to extract useful knowledge and enable the transportation authority to take preventive actions and make appropriate decisions.

Traffic data contains hidden values that can improve and support safe and sustainable transportation systems. For instance, by using roadside sensors, vehicle speed data can be collected and analyzed to identify traffic congestion. When traffic congestion is detected, travel alerts can be provided to drivers to help them find alternate routes and hence reduce the congestion [1]. Analysis of vehicle wait times at traffic lights can produce insightful information and lead to better ways to optimize traffic light policies and improve traffic flow [2]. Analysis of video data can detect and classify objects (e.g., vehicles or pedestrians), identify their trajectories, and recognize significant traffic events, such as veering, abrupt braking, and near misses [3]. Such analysis can help decision makers take the necessary actions to improve road safety, prevent collisions, and

save lives.

Traffic data, arguably, fits the characteristics of Big Data, often characterized along the following dimensions: volume, variety, velocity, veracity, and value [4, 5]. First, various equipment installed on roads to monitor traffic generates a vast amount of data. The volume of traffic data will grow more significantly when connected vehicles communicate and exchange information with other devices within themselves, with the road infrastructure, or with other nearby vehicles. Connected vehicles could generate approximately 30 gigabytes of data per day [6]. At this rate, the traffic data would exceed a terabyte of data over approximately one month. Second, traffic data comes in a variety of structured and unstructured data, such as JPG, JSON, XML, GPS, PDF, image, video, and social media posts [7]. Third, the velocity of traffic data is substantial as various sources produce new data continually. Fourth, data veracity refers to uncertainties that are inherent in traffic data, such as inaccurate or incomplete data [8]. Finally, traffic data contains invaluable information but at a low density. For example, video data may reveal the cause of a collision at an intersection. However, since collisions do not occur all the time, most of the data capture only normal vehicle movement in the area.

Some applications in the transportation domain, such as autonomous vehicles, require real-time data processing and reliable communication networks. Considering the volume of traffic data, Big Data and analytics can benefit from edge computing, which allows data processing and computation to happen near the data sources (e.g., cameras, sensors, mobile devices, vehicles), thereby reducing the bandwidth consumption and network latency between the end-users and the cloud computing platforms that store, manage, and analyze data [9, 10, 11, 12]. Edge computing has the potential to offer an efficient way to tackle issues relating to the exponential growth of data, limited communication bandwidth, and high computational resources in the cloud.

The primary goal of this review is to gain knowledge of the current research and applications of Big Data in transportation. The review is intended to give researchers and transportation departments insights into Big Data and analytics

to guide and support the development of better transportation systems. Additionally, this review can assist cities that have adopted Vision Zero [13] in working towards eliminating traffic fatalities and major injuries (i.e., collision injuries that result in admission to hospital).

We organize the rest of this article as follows. Section 2 describes the systematic review protocol including research questions, search terms used, inclusion/exclusion criteria, databases examined, and articles included in this review. Section 3 reviews architectures of Big Data and intelligent transportation systems. Section 4 presents opportunities relating to the utilization of Big Data and analytics to support sustainable transportation systems, whereas Section 5 describes the associated challenges. Section 6 discusses our point of view related to the research questions. Finally, Section 7 summarizes the main points of our literature review.

2. Systematic review protocol

A systematic review protocol helps researchers to develop a high-level overview of knowledge on a particular research area [14]. It provides a methodical process of identifying, screening, and synthesizing a body of published work in pursuit of holistic evidence relevant to particular research questions. We employed a systematic review to provide a useful overview of the current body of research on Big Data and analytics in the transport domain. This section describes the protocol we developed and used to *guide*, rather than restrict, the review of the research work. The protocol utilized in this article includes the description of the research questions, databases searched, search terms, and inclusion/exclusion criteria.

2.1. Research questions

The research questions that guided our review fall into two broad classes: Big Data concepts and applications. In these research questions, the Department of Transportation (DOT) refers to an organization that manages transportation in

its jurisdiction (e.g., states, provinces, cities) and Big Data applications in the domain of transport:

1. What is the current state of scientific research on the application of Big Data and analytics in DOTs?
2. What kinds of Big Data systems are being used by DOTs?
3. How are DOTs using insights derived from Big Data and analytics?
4. How can DOTs use Big Data and analytics to develop more efficient and sustainable transportation systems?
5. What are the current opportunities and challenges in using Big Data for DOTs?
6. What types of data are being collected, stored, and managed by DOTs, and how is this data being used?

2.2. Databases and search terms

The following databases were used to search for articles and papers: Association for Computing Machinery (ACM) Digital Library, Computers and Applied Sciences Complete, Web of Science (ISI), Scopus, Academic Search Complete, Directory of Open Access Journals (DOAJ), and the Institute of Electrical and Electronic Engineers (IEEE) Xplore Digital Library. We selected these databases due to their depth in content, coverage of computer science, and accessibility in the domain of Big Data technical literature.

To retrieve potentially relevant articles, we used the following search terms in each of the databases:

- "big data" AND "applications" AND "transportation" AND "smart city"
- "big data" AND "smart transportation" AND "road safety"
- "big data" AND "transportation" AND "connected vehicle"

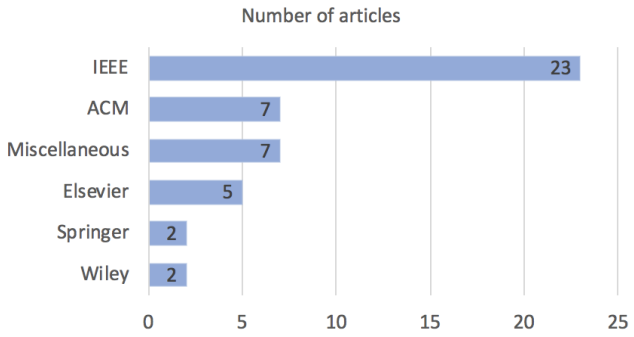

Figure 1: Publishers of the cited papers in this article.

2.3. Article inclusion and exclusion criteria

Articles in the search results were screened using these main criteria:

- The article has to be relevant to at least one of the research questions.
- The article needs to be current, published in the last five years.
- The full text of the article must be available and accessible online.
- The article must have a reasonable number of references from credible sources.
- The article must be published from a reputable publisher.

We followed a time frame of five years used in a recent Big Data literature review [15]. Focusing on articles published in the last five years is reasonable at this point, as Big Data was considered an emerging technology in 2012 [16] and started to provide more productivity for organizations by 2015. We used the last two criteria to indicate our intent to focus on published, peer-reviewed articles in conference proceedings and academic journals. The majority of the articles included in our review are published by well-known publishers, such as IEEE, ACM, and Elsevier (see Fig. 1). These organizations distribute a wide range of quality publications on technical knowledge and are accessible to professionals in computer science and information systems.

2.4. Search results

The following results represented a snapshot of our initial search. Applying the search terms to the databases returned 173 articles. During the initial screening, we relied on the abstract to determine the relevancy of content of the articles to the research questions. This process reduced the number of potential articles to 56 items. The most common reasons for excluding articles were the lack of relevance to the research questions and that the full text of the articles was not available or electronically accessible. Out of the 56 articles that passed the initial screening, we reviewed 28 articles due to their high relevance to the research questions. We also included some articles that discuss Big Data and edge computing, that were found in the references of the reviewed articles, and that were relevant to our discussion. One of these articles is older than five years but was included because it pertains to the research questions. The total number of papers cited in this article is 46. Table 1 presents the research questions, keywords, and relevant papers.

3. Big data systems and their utilization in transportation

This section details how Big Data architectures and systems attempt to address some of the challenges and realize the potential of Big Data in transportation. The research questions addressed in this section are: (1) What is the current state of scientific research on the application of Big Data and analytics in DOTs? (2) What kinds of Big Data systems are being used by DOTs? (3) How are DOTs using insights derived from Big Data and analytics? (4) How can DOTs use Big Data and analytics to develop more efficient and sustainable transportation systems?

3.1. Big data systems and architectures

In the context of transportation, Big Data architectures or platforms described in the literature can be used to implement an Intelligent Transportation System (ITS). An ITS is a transportation management system that integrates

Table 1: List of research questions, keywords, and relevant papers.

Research Questions	Keywords and References
What is the current state of scientific research on the application of Big Data and analytics in DOTs?	Privacy, security, optimal path, historical, real-time, predictive, visual, and video/image analytics [3, 8, 17, 26, 28, 29, 34, 44]
What kinds of Big Data systems are being used by DOTs?	Intelligent transportation system, Hadoop, MapReduce, batch and stream data processing, NoSQL [15, 18, 19, 20, 21, 22, 33, 41]
How are DOTs using insights derived from Big Data and analytics?	Urban planning, collision and near miss analysis, traffic congestion, safety, and optimization [3, 6, 19, 20, 21, 22, 23, 30, 35]
How can DOTs use Big Data and analytics to develop more efficient and sustainable transportation systems?	Urban planning, collision and near miss analysis, real-time update and information sharing [3, 6, 22, 23, 24, 27, 31, 33, 35, 42]
What are the current opportunities and challenges in using Big Data for DOTs?	Data collection, quality, storage, processing, privacy, security, connected and autonomous vehicle [2, 17, 19, 22, 26, 30, 36, 37, 43, 44]
What types of data are being collected, stored, and managed by DOTs, and how is this data being used?	Speed, location, video, image, traffic intensity, social media, crowdsourcing, machine learning, historical, real-time, predictive, visual, and video/image analytics [1, 2, 3, 7, 20, 23, 27, 37, 38]

technology, such as information systems, sensors, and electronics, with transportation infrastructure to make a transportation system more efficient, safe, environmentally friendly, and accessible to users [17].

For example, a smart city Big Data architecture, which can be used as an ITS, consists of seven layers [18]: (1) Data sources (e.g., sensors, devices) (2) Data normalization (e.g., extract, transform, and load) (3) Data brokering (e.g., using a context broker) (4) Data storage (e.g., distributed storage) (5) Data analytics (e.g., statistical and numerical analysis and real-time analysis) (6) Data visualization (e.g., event mapping) (7) Decisions (e.g., real-time and long-term actions). We can categorize these layers broadly into data collection and preparation, and data analytics and utilization to support decision-making processes. The data source and normalization layers collect data from various sources, prepare the data, and load it into databases. The data brokering stage addresses the heterogeneity of this data, dealing with varied formats, sources, and frequencies of data updates, and combines and integrates the data appropriately. The data storage layer manages the storage and retrieval of integrated data to support data analytics. The data analytics layer performs various analysis to extract useful knowledge and meaningful patterns from data. The data visualization layer presents the analysis results in graphical forms to users to help them make informed decisions. The decision stage of the architecture supports both real-time and long-term actions. This architecture was used to develop an ITS system where GPS data from taxis in Rome was collected and analyzed [18]. After the system cleans the data (e.g., removing GPS locations from outside the city), it visualizes the results on a map so that users can compare the density of routes in real time. This visualization allows users to choose optimal routes to avoid and reduce traffic congestion.

The Traffic Telco Big Data architecture consists of a data layer, a processing layer, and an application layer [19]. In addition to using telecommunication data, the architecture uses map data and social media data. The data layer can be implemented using a combination of data storage platforms, such as a relational database management system, the Hadoop Distributed File System,

Apache Spark for fast querying, and Apache Hive for SQL-like queries. The processing layer is where the data is anonymized to maintain privacy. The application layer provides a user interface that allows interaction with data and supports data analytics and visualizations. At the application layer, users can generate a heat map, find an optimal route based on travel time, identify traffic congestion, and perform a simulation of how traffic flow would change when a particular road is blocked.

Another architecture, called Hut, incorporates machine learning to provide context for real-time data [20]. Hut is an Internet of Things (IoT) architecture for smart cities that can be used to process transportation data. The architecture was used to manage and process data collected from 3,000 traffic sensors in the city of Madrid, Spain. Node-RED is used to acquire data from many different sources. Streams of this data are given to a message broker (e.g., Apache Kafka) and then stored in a cloud storage system (e.g., Amazon S3). Data is retrieved from the storage system by a batch analytics platform where machine learning (e.g., Apache Spark's MLib) is used to gain insight into the data and create models for both real-time and predictive analytics.

A distributed smart traffic system uses a Lambda architecture to integrate both stream and batch data processing [21]. This framework incorporates three layers: the speed layer, batch layer, and serving layer. The architecture prescribes that data be acquired through a web service that provides a common interface for heterogeneous data sources. Through this web service, data is sent to the speed, batch, and serving layers and ultimately to the end-user application. The speed layer uses Spark Streaming to enable the creation of real-time views, whereas the batch layer uses HBase and Hadoop for distributed processing of historical data. The serving layer combines both real-time and historical data views and manages the views in an SQL database. The serving layer addresses the different needs of data access by enabling both real-time and historical data access either separately or at the same time if required by users. Visualization of data is provided by an end-user application, called Locality-Enhanced Geographic Information System (LEGIS), which provides a map visualization of

traffic data. The system also provides a video analysis tool, which can analyze traffic video to count the number of vehicles and then passes the data to the web service layer. The framework uses OpenStreetMap to gain access to information about intersections, road types, and speed limits.

Taking advantage of the widespread use of smartphones, the Smartphone Road Monitoring System (SRoM) uses a crowdsourcing approach to collecting real-time transportation data, such as traffic conditions and driving behavior [22]. This approach addresses the high costs and limited scalability of using fixed or mobile traffic sensors by utilizing smartphones to collect data. The app collects data using both active and passive mode. The active mode, intended for passengers and pedestrians, allows users to report traffic events explicitly. The passive mode automatically collects traffic data such as current location, direction, speed, and potholes (by detecting vibration).

Table 2 summarizes the features of Big Data architectures that support transportation data processing and analysis. This table highlights whether a particular architecture offers historical, real-time, predictive, visual, and video and image analytics tools.

3.2. Big data utilization

This section presents a summary of how cities or DOTs can use Big Data in transportation to create safer and more sustainable transportation systems. We highlight three main ideas in this section: sharing real-time traffic information, using traffic data for urban planning, and improving traffic safety by analyzing near misses and collisions (see Fig. 2).

3.2.1. Sharing real-time traffic information

Sharing traffic information with users in real time is an obvious way to improve the sustainability of transportation systems. Having relevant information at hand allows users to make an informed decision. Traffic congestion due to poor weather or accidents may be avoided if people receive the information promptly [22, 23]. They can take alternate routes or change their usual way

Table 2: Big Data platforms and the kinds of analytics supported.

Big Data Platform	Supported Analytics				
	Historical	Real Time	Predictive	Visual	Video/Img
Traffic Telco Big Data Architecture [19]	Yes	Yes	Yes	Yes	No
Hut Architecture [20]	Yes	Yes	Yes	Yes	Not stated
Distributed Smart Traffic System [21]	Yes	Yes	Yes	Yes	Yes
Smartphone Road Monitoring System [22]	Yes	Yes	Yes	Yes	No
Intelligent Transportation System [18]	Yes	Yes	Yes	Yes	Not stated

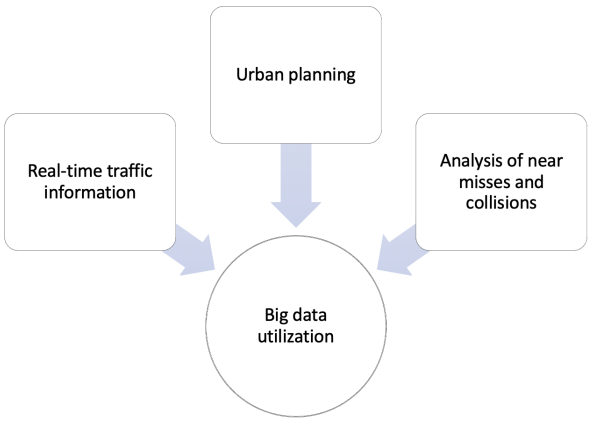

Figure 2: Big data utilization to improve the safety and sustainability of transportation systems.

to commute (e.g., taking a train instead of a car). Helping drivers find available parking spots in busy downtown areas means that fewer cars will stay unnecessarily longer on the road, reducing the use of fuel and traffic congestion. Providing real-time assistance for emergency vehicles to get to their destinations may save more lives as people can receive help sooner. These examples show how the real-time distribution of relevant information to users of transportation systems can meet the ever-increasing load of transportation systems.

Traffic information can be distributed through various communication channels, including radios, mobile apps, and electronic billboards [22]. Traffic alerts on mobile apps, especially, must be designed so that they do not distract users from driving and paying attention to the road. An ITS may supplement real-time traffic information with predictive analytics of traffic conditions. For example, when the weather is a factor affecting traffic flow in a city, transportation officials can share traffic analytics that predicts congestion due to weather and other relevant factors [23].

3.2.2. Urban planning

Historical analysis of transportation data can support evidence-based approaches to urban planning. Decision makers no longer have to rely solely on their intuition or unreliable assumptions while developing plans for new transportation infrastructure. Traffic flow and patterns on the current street infrastructure can inform future infrastructure construction [18]. Changes in traffic flow may be simulated before new construction begins or new policies take place. Vehicle wait times at traffic lights can be optimized to improve traffic flow and reduce emissions [2]. The sustainability of transportation networks can be improved by analyzing patterns of public transportation usage and then optimizing routes and scheduling of public transportation [24].

3.2.3. Analysis of near misses and collisions

Analysis of near misses and collisions may improve the safety of transportation systems. Visualization of locations and types of traffic collisions can help

identify high-collision areas and causes of collisions [25]. The aggregation of data attributes, such as the frequencies and causes of collisions and types of vehicles involved in the collisions, can be visualized on a map to help the transportation authority assess if changes in traffic safety policies or infrastructure are required. Proactive safety analysis of traffic video of streets or intersections can classify objects, such as vehicles or pedestrians, by metrics like turning movement, speed, and direction [3]. Video analysis may reveal near misses or other potential risks with the existing road infrastructure. Having such information would allow the transportation authority to take the necessary actions to prevent collisions and improve traffic safety, such as putting new stop signs at previously uncontrolled intersections or installing cameras at specific locations to enforce the speed limit in the areas.

4. Opportunities

Connected vehicles represent a significant opportunity for Big Data in transportation in the areas of safety and sustainability. A connected vehicle uses GPS, communication technologies, and sensors to communicate with devices within the vehicle and with other vehicles, as well as with transportation networks [26]. Data collected from these sources may be utilized for the internal operation of systems or shared with others for collective benefits.

A connected car can offer personalized services and an integrated information system to the driver. It may include having to-do lists, appointments, and music playlists on mobile devices to be accessible by connected vehicles. The audio system may play favorite songs automatically and thereby provide fun entertainment. The car's dashboard may show today's meeting schedules and help coordinate timeline scheduling and other time pressing issues. There is a possibility of a predictive capability where the car may alert the driver when they are near frequently visited stores. Also, a connected vehicle may consult the documentation on a cloud platform to perform engine diagnostics, to schedule maintenance, and to calculate optimal routes based on different criteria, such as

shortest distance, quickest time, and fuel-efficient paths [26]. These are examples of how a connected vehicle can use relevant data for internal usage and employ various analytical techniques available in Big Data literature.

Further, the potential of connected vehicles extends beyond providing personalized services to drivers. Connected vehicles also offer an excellent opportunity to improve traffic safety and sustainability of transportation systems by sharing relevant information automatically with nearby cars and transportation networks [26, 28]. For instance, when a car detects an accident on the road or a potential hazard (e.g., reckless driving or bad weather conditions), a connected vehicle may alert the surrounding cars so that they can slow down or take the necessary action to avoid damage or collisions. Traffic conditions may be shared to help drivers find the most efficient routes to their destinations. Emergency vehicles may receive real-time assistance in getting to the needed locations quickly and safely. Such benefits are far reaching and can improve the sustainability of transportation systems, as they reduce the amount of time cars are staying on the road, which means reduced fuel consumption, traffic congestion, and commute time.

Finally, autonomous vehicles aim to bring potential benefits of connected vehicles further by eliminating the need to have human drivers completely. While this idea may seem futuristic, with advances in technologies, autonomous vehicles could improve traffic safety; for example, by reducing the number of collisions due to human errors and fatigue. However, user acceptance and development of autonomous vehicle technologies remain challenging at this time [29].

4.1. Attributes of transportation data

There are a number of metrics that can be derived or calculated from transportation data. Data collected by roadside sensors may contain information about speed of vehicles, vehicle types, GPS locations, wait times at traffic lights, traffic density, and traffic intensity [2, 20, 22]. Traffic density is the average number of vehicles per mile or kilometer, whereas traffic intensity is the average

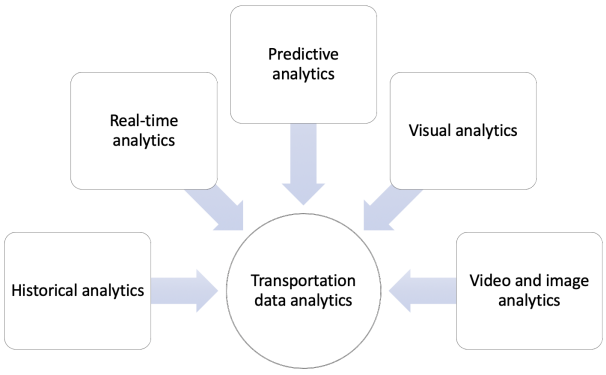

Figure 3: Various kinds of transportation data analytics.

number of vehicles that pass through a specific location per unit of time.

Video and image data is also part of transportation data. This data is recorded using cameras installed in fixed locations or mounted on cars or aircraft. Traffic images can be used to identify vehicle license plate and vehicle type [6]. They are also used for traffic enforcement (e.g., speed limit, traffic light, stop sign) and contain information about vehicle speed, location, time, and trajectory.

4.2. Transportation data analytics

Transportation data gathered from various sources contains rich information. Various analytics can be performed on transportation data, including historical, real-time, predictive, visual, and video and image analytics (see Fig. 3).

4.2.1. Historical analytics

Historical transportation data can be analyzed to discover trends and patterns to help make longer-term decisions and conduct urban planning. For example, traffic flow, driver behavior, and usage of current street infrastructure can be analyzed before building new transportation infrastructure [18]. Additionally, transportation data can be analyzed and used by policy makers to adjust policies, such as public transportation routes or traffic signals [30].

Urban planning can also benefit from data collected from the public [24]. This project examined public transportation routes using crowdsourced geolocation data where users added their public transportation routes to an online map. Such data allows identification of high-demand routes, rush hours, and other relevant information useful for public transport planning and routing. This project used clustering algorithms to analyze user preferences for routes and applied a routing algorithm to the clusters to develop optimal transportation routes for public.

4.2.2. Real-time analytics

Real-time analytics refers to the capability of processing, analyzing, and disseminating information (or knowledge or intelligence) as soon as the relevant data is available in the system. It can provide up-to-the-second information to decision-makers to allow them to make better and quicker business decisions. The need to update users drives the requirement of real-time analytics, and its latency may range from a few seconds to a few minutes. For example, a real-time adaptive traffic control system requires to immediately run analytics and get the results to decide if the signal timing needs to be changed to adapt to the new traffic condition data. The dissemination of information to users could also be in a significantly-longer interval than the data collection interval to provide more accurate information and to avoid infobesity that may confuse users or lead to an unexpected reaction of the users. For example, although traffic detectors can collect and update traffic data continuously every minute or shorter, like 20 seconds, traffic authorities may need collision likelihood update only every five minutes for their decision making. The gap between the availability of the data and the need to update the users determines the allowable analytics latency that drives the selection of analytical techniques employed.

Research into real-time data analytics is a significant area in Big Data. In the transportation domain, real-time analytics can be performed to assess current traffic flow; to detect traffic congestion, accidents, or hazardous road conditions; and to assist emergency vehicles in finding optimal routes [20, 22, 31, 32, 33].

Algorithms that analyze the optimal route choice by calculating the shortest path are also found in the literature [34].

Detecting traffic conditions and patterns in real time can be done using data gathered from roadside sensors, GPS, smartphones, and connected vehicles. For example, acceleration data from smartphones can be analyzed to detect traffic accidents in real time [22]. Analytics on a real-time data stream allows the detection of significant traffic events, such as reduction in average traffic speed [20]. Real-time traffic data can also be combined with weather data to help identify hazardous road conditions (e.g., snow), which may assist drivers and emergency responders in avoiding such conditions and allow faster emergency vehicle routing and improved response time [22].

4.2.3. Predictive analytics

Having collected a huge amount of data, Big Data systems offer great potential for development of predictive models. A predictive model uses data mining and machine learning algorithms to learn about data and make predictions about possible traffic events, for example, a traffic jams prediction [23]. This predictive system uses a data mining platform called KNIME to analyze weather data and traffic data collected from roadside sensors in Santander, Spain. Using its predictive model, the system can predict a traffic jam for the next 15 minutes.

Predictive analytics allows the relevant authority to prepare for likely traffic events. For example, when traffic congestion is expected due to a major sporting event, the transportation authority can provide special buses or trains to get to the event and encourage people to take public transport. When poor weather is predicted to cause many collisions, people can be warned to be extra careful while driving. Police may deploy more personnel in areas where crimes are likely to occur during specific time. Essentially, predictive analytics enables the relevant authority to take proactive and preventive actions to maintain and improve public safety.

4.2.4. Visual analytics

Visual analytics relies on human vision to recognize patterns and other meaningful information by transforming and mapping data onto visual forms. Well-designed visualization can help users gain insight into data with little cognitive effort. Outliers or clusters of data can be identified easily, as they become apparent when plotted in a chart. For example, an application called Traffic Accidents Analyzer analyzes traffic collision data in New York City and visualizes the data on a map [25]. The visualization shows high-collision areas and object types involved in collisions (e.g., bicycle, bus, taxi, and passenger vehicle), along with other attributes, such as time, location, injuries/fatalities, and cause of an accident. Users can see the frequency of cause of collisions or object types involved in collisions on a bar graph and collision locations on a map.

In visual analytics, it is common to provide multiple, coordinated views to allow users to manage the complexity of data. For example, a project visualizing social media and smart card transportation data in Tokyo provides three coordinated views: HeatMap, AnimatedRibbon, and TweetBubble [27]. Each view focuses on a specific aspect of the data. The HeatMap view focuses on the temporal dimension of data; that is, how busy or not a particular passenger line is over a specified period (i.e., daily or monthly view). The AnimatedRibbon view uses animation and color-encoded ribbons, overlaid on routes on a map, to visualize the number of passengers and change from the average on a particular route. The TweetBubble visualization aggregates trending keywords from tweets and provides an overview of the usage of the words in bubbles. These views show their usefulness in the analysis of traffic flow and passenger behavior in response to natural disasters (e.g., earthquake) or events (e.g., parade).

4.2.5. Video and image analytics

Analysis of traffic video and images is an emerging practice in Big Data and transportation analytics. Video analytics supports basic recognition tasks, such as license plate and vehicle type, based on photos of passing vehicles [6]. This ability can be used for traffic enforcement, for example, by improving the

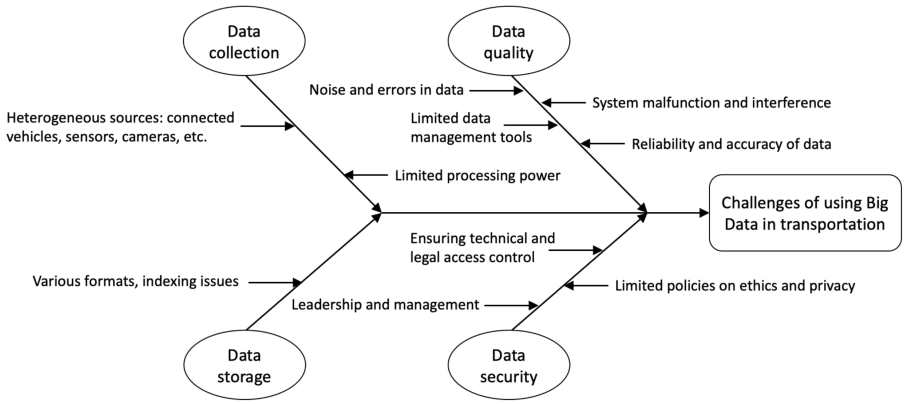

Figure 4: Key challenges of utilizing Big Data in transportation.

efficiency of processing speeding tickets.

A collaborative project between Microsoft, City of Bellevue, and University of Washington employs tracker technology that uses machine learning to classify objects in video by movement (e.g., turning), direction, type (e.g., bus, car, bicycle, pedestrian), and speed [3]. The goal of this project is to detect and analyze near misses and to understand their causes. This goal is relevant to Vision Zero, which aims to reduce the number of traffic fatalities and major injuries to zero.

Video data is also useful for proactive safety analysis of traffic data [35]. This project analyzed approximately 473 hours of video data from 20 roundabouts in the province of Quebec, Canada. The analysis uses various criteria, such as time to collision, motion prediction, and vehicle interactions. Such analysis can help reveal high-collision areas and allow the transportation authority to make the necessary adjustment to reduce the number of collisions in the areas.

5. Challenges in using big data in transportation

Given the nature of transportation Big Data, there are several challenges identified in the literature. These challenges can be summarized as data collection, quality, storage, and security issues (see Fig. 4).

5.1. Data collection

One of the main challenges in utilizing Big Data in transportation is that traffic data is collected from various sources. Some sources, such as roadside sensors, provide ready-to-use traffic data, which can be analyzed easily. Other sources, such as logs of user activities on smartphones, may require some analytical processing before we can derive meaningful information from this data. Furthermore, these sources may be controlled by third parties (e.g., telecommunication companies) so that the data is not readily accessible by transportation departments. Though increasing availability of this data opens up numerous opportunities, making sense of these data can pose many challenges, including access permission to use such data for research, data governance, ethics, and privacy. Since this data comes in various formats, issues around system interoperability, data cleansing, and processing can be difficult to achieve.

Roadside sensors, such as cameras and detectors (e.g., magnetic loop, laser, ultrasonic, and infrared), are capable of tracking the movement of vehicles along a road or an intersection. They can particularly record vehicle types (e.g., car, truck), speeds, and time stamps. While these sensors are designed specifically for monitoring and collecting traffic data, they come with high costs of deployment and maintenance, they are statically fixed, and cannot be easily scaled up for large cities [36]. Mobile sensors such as cameras or radar, on the other hand, provide more flexibility than roadside sensors as they can be deployed on vehicles or aircraft to record traffic data. However, mobile mounted devices are often bulky and costly to operate and maintain, and their effectiveness can be limited by poor weather [22]. Nonetheless, these different data collection apparatus have different technical capabilities when generating data, in terms of costs and efficiency. In addition, although using cloud computing for the analysis of Big Data to the cloud than in the devices at the network edge has been an efficient way to process data due to large scale computing power, the bandwidth of the networks that carry data to and from the cloud has not corresponding increased [10].

Transportation data can also be derived from GPS-enabled smartphones

[36]. For example, Uber and Lyft use smartphones to manage and monitor their operations, which involve a lot of traffic data [37]; they can also use this data to facilitate ridesharing services. Since these apps keep continuous track of movement of registered cars, it is possible to analyze and estimate traffic conditions at real time; for example, based on car speeds on certain routes. The INRIX mobile app is another example of applications that collect transportation data, such as speed, location, and time, from participating mobile phones [22]. Data from mobile apps, however, may not be readily available for transportation departments, especially if it is collected for a specific purpose (e.g., ridesharing) because privacy regulations may prevent such data from being used for another purpose (e.g., transportation management). Issues of privacy and consent to use such data might be difficult for research purposes, making this form of Big Data "dark data" (or unstructured, untagged, and unprocessed data that is not being used to derive useful insights).

Cellular network signals can also be used to serve indirectly as a source of transportation data [30]. Using triangulation, for example, it is possible to estimate users locations and their movement from their streams of telecommunication data. These movement patterns can reveal rush hours and busy routes in cities. Such information can help drivers find alternate routes or plan their commute time better. However, given the private information contained in streams of telecommunication data, protection of privacy must be considered and steps to anonymize data must be taken carefully [19].

Dedicated sensors embedded in vehicles that utilize technologies, such as wireless networks and GPS [26], are other ways to generate transportation data usage. A concern of using GPS-enabled devices, either through smartphones or vehicle-embedded sensors, is the possibility of sharing such data with unconsented third parties, such as Google Maps, Waze, or Apple Maps, which raises privacy concerns [19]. This raises issues of personal safety accruing to commercialization of personal data and possibility of data leakage.

Transportation smart cards are another source of transportation data [27]. Smart cards are payment cards that can be used to access public transport

systems, such as buses and trains. Smart card log data contains information about how users travel origin, destination, start time, and exit time. This data can be analyzed to identify busy routes, rush hours, and the number of users to provide better service to the public. Unlike other personal data, individuals protection and privacy regulations might override the desire to analyze such data to inform better ways of improving transport services.

It is likely that with the introduction of connected vehicles additional challenges can be expected in the collection and processing of data for highway applications, as well as rendering results to users and applications.

Finally, transportation data can also be collected from posts on social media [19, 27]. Social media applications allow users to share their thoughts, reviews, experience, and news with their social networks. For example, when people experience a traffic jam or see an accident, they may post the information to social media. They might complain or notify the relevant authority when they see unsafe driving behavior, a pothole, or damaged infrastructure. Thus, social media can serve as a communication channel for cities to collect transportation-related data from the public. However, data obtained from social media are primarily reactive, raising issues of authenticity, data integrity, and credibility. Government agencies cannot take risk with these shortcomings.

5.2. Data quality

The wide availability of transportation data also means that there are many quality issues that need to be dealt with before using such data. For instance, transportation data inherently contains a lot of noise and uncertainty. The quality of data can be compromise because of sensor malfunctioning or interferences, which may result in missing or conflicting data [8]. Absence of tools and systems in place to manage data veracity also increase the possibility of making this data unstable. Furthermore, conflicting data need to be resolved by cross validating data from multiple sources. For example, when there is conflicting information from roadside sensors, it may be possible to resolve it by checking the same information available on video or smartphone data. Thus,

the challenge of managing data collection from multiple sources also represents an opportunity to manage and deal with uncertainty in transportation data.

Transportation datasets often contain noise and errors that require cleaning and preprocessing. For example, datasets from the New York Taxi & Limousine Commission contain trip information, such as pickup and drop-off locations and their time stamps [38]. The quality of transportation data affects data processing and accurate interpretation. Zhang and colleagues [37] demonstrate how low data quality can cause issues in real-time data processing. In particular, inaccurate or incomplete traffic data (e.g., due to data loss) causes uncertainty and becomes disorganized when processed in parallel and concurrently. They suggest a potential solution to this problem where data processing is delayed until all the necessary data instances are completely ready. One of the key advantages of working with Big Data is the ability to more efficiently gather data from a variety of sources and process it in real time to solve immediate problems. However, there is a trade-off between maintaining data quality and processing data as quickly as possible. For instance, data reliability is important because it determines accuracy and precision.

GPS signal interference from high buildings, a malfunction of the GPS devices, or human errors (e.g., taxi drivers forgetting to log the end of a trip) can lead to inaccuracies in both location and time data. It is imperative that such inaccuracies in data quality need to be resolved or minimized before performing analytics on the data.

5.3. Data storage

Similar to challenges of diversity of data types and formats when working with Big Data reported in the literature [39, 40], it is not surprising that transportation data formats are varied. Typical formats of transportation data include JPG, JSON, XML, PDF, GPS, video, and social media data. This variety of formats presents a challenge in data storage as one single type of database may not be optimal to store and manage all kinds of data [7]. For example, GPS and location-related data may work well as a document-oriented

database (e.g., MongoDB), whereas social media data may work better if stored as a graph-oriented database (e.g., Neo4J).

In addition, certain data types may require a specific type of database. Movement data, such as GPS locations and time stamps, can be stored and managed efficiently using parallel and distributed Moving Objects Databases (MODs) [41]. MODs can store, update, and keep track of locations of a large number of moving objects, such as people and vehicles. Current challenges in MODs include query processing and centralized and distributed MODs.

Using a single type of data storage can limit the system ability to meet different user expectations. Some data storage platforms are designed for large batch data processing, whereas others are optimized for real-time data processing. If a Big Data system only supports one type of data storage, the system will not be able to satisfy different user requirements for data processing. Furthermore, with the emerging common use of non-relational databases, we now have more options to offer multiple data storage solutions to meet the user need.

An example of using multiple data storage platforms is demonstrated in a project that manages streaming data from connected vehicles [37]. In this project, three different storage platforms are used to deal with different kinds of data: (1) a relational database (PostgreSQL) to manage relational and processed data; (2) a non-relational database (HBase) to store raw data; and (3) in-memory caching systems (Memcached and Redis) to deliver real-time performance. Using various platforms, the system is able to meet the need for timely data processing such as vehicle tracking using in-memory data storage, whereas efficient batch processing can be performed efficiently using HBase and Hadoop.

Hadoop uses a distributed storage system called the HDFS or Hadoop Distributed File System [18]. Babar and Arif [42] proposed a prototype of big data analytics architecture that uses the Hadoop Ecosystem. The architecture consists of three layers: data acquisition and aggregation, data computation and processing, and decision-making and application. The system preprocesses data to deal with missing values, noise, and other problems, and then stores and maintains the preprocessed data in the HDFS. The system utilizes MapReduce

of Hadoop. First, it uses the map function to create key/value pairs on the data and then uses the reduce function to aggregate values associated with a key and store the results in the HDFS. The smaller dataset can then be processed more efficiently in a parallel manner over the HDFS distributed storage system. However, Hadoop might not be secure in terms of protecting individual privacy, because the system is designed for handling less sensitive data [40]. As a proof of concept, the prototype was used to analyze three open datasets: the vehicle traffic and parking datasets in the City of Aarhus, Denmark, and the water usage in the City of Surrey, Canada. The analysis uses some threshold values to identify or monitor traffic congestion, the availability of parking space, and the water consumption in a city.

5.4. Data security

The security of data is another significant challenge in using Big Data in transportation, as transportation data can be a target of hackers [17]. Safeguarding Big Data is a complex task, which requires appropriate authentication, access control, and encryption measures [43]. Some challenges in transportation Big Data security currently being researched include the secure use of cloud-based services for devices and networks, the identification of internal and external threats, and the standardization of data security [44].

Providing and maintaining appropriate access control to Big Data is challenging due to the need to distribute and share data. Securing data becomes critical especially when different organizations need to share data over computer networks that have different levels of security [2]. There is a tension between security and accessibility of data. Another security challenge in Big Data is due to the distributed nature of data storage and processing on Big Data platforms, such as Hadoop and Cassandra, which have their own security concerns.

One of the most notable security aspects required by Big Data applications is privacy. It is a critical aspect to be addressed because users share more and more personal data that raises legal and policy requirements for data handling and sharing. A recent report funded by the European Union reviewed and

identified existing public and private sector policies related to data privacy and security on Big Data in the transportation domain and its applications [45].

The development of policies on data privacy and security requires leadership support from management. Many organizations collect and store data in different and disparate databases. To effectively utilize this data, we need new tools to harvest, clean, and consolidate the data into a useful format. Since various departments may own different data within a single organization, leadership support is necessary for coordinating, managing, and governing the data. The challenges of harvesting and consolidating scattered data within an organization are likely to be influenced by many factors, including the history and ownership of the data, the availability of governance structures and data sharing protocols, and the size of the organization. Ultimately, leveraging the potentials of Big Data within the transport systems requires leadership support and reliable infrastructure that can make use of data already collected and capture new forms of data.

6. Discussion

The primary purpose of this literature review is to explore the research and applications of Big Data in the transportation domain. We aim to examine the current practices on the use of Big Data and analytics to improve the efficiency and sustainability of transportation systems, rather than to identify and prescribe specific solutions to the challenges of Big Data in transportation. During the analysis of the literature, it became apparent that Big Data projects in the transportation domain are still infancy. These projects are typically part of academic research conducted at universities or research institutions and have a limited scale of deployment. The proposed architectures in the literature discuss various components of Big Data systems at the conceptual level, often supported by testbed systems as a proof-of-concept, but most of the designs have not progressed to a large-scale implementation and deployment in the real world. Consequently, such projects have not delivered real, measurable benefits

to the transportation systems.

This situation is understandable, as managing transportation systems is a complex task and involves many stakeholders. Deploying a Big Data system requires a coordinated, multi-year effort and commitment from various departments in a city, and therefore, such projects are usually part of smart city initiatives. The city may need to expand the existing infrastructure with new sensors and communication networks. The residents may have concerns about privacy and security of personal data collection and usage. Moreover, the reliability and robustness of the system must be tested thoroughly, as traffic flow is fluid and can be unpredictable at times.

Hadoop is the most common platform used in Big Data research projects [15]. Hadoop is an open system consisting of a collection of open source libraries for processing large data sets across thousands of computers in clusters. Hadoop is used widely for Big Data because it does not rely on specialized hardware, and virtually, most of the software can be plugged into it and used for data processing and visualization. Since transport data comes in different formats, Hadoop is suitable for gathering data from disparate data sources in different formats. It is also natural for research projects to use open-source software to test software architecture or algorithms. There are a few exceptions, where big vendors, such as IBM or Microsoft, collaborate with research institutions to develop smart city projects. Currently, the evidence of how Big Data can deliver measurable benefits still lacks due to the limited scope of the existing deployment. We expect that as Big Data is maturing, best practices will emerge and improve the return on investment.

There are many potential benefits of Big Data in transportation. However, these benefits have not materialized or assessed thoroughly. Big Data is an emerging technology, so it will take some time to see successful applications at a large scale (e.g., citywide) that have run for an extended period (e.g., five year or longer) in order to reap the benefits of these projects. It is also necessary to explore ways to develop systematic methods for assessing the benefits of Big Data systems in transportation.

Managing traffic flow requires not only timely but also selective and preventive notifications. For example, notifying all drivers about traffic congestion on a road segment may result in moving the congestion to the proposed alternate route. The system must inform only enough drivers to take an alternative way to reduce the congestion. It may also be necessary to send preventive notifications based on historical patterns (rush hours, weekdays, and weekends) in certain areas. In this way, drivers can take the necessary actions to manage their schedules, for example, by leaving home five-minutes earlier on Monday mornings.

Getting access to personal calendar, appointments, and daily schedules may allow a Big Data system to provide personalized service to users, such as suggesting an alternate route that works better for them. However, users will expect that their privacy and security to be protected in case of a data breach. There is a tension between personalized service and privacy protection. A potential solution may include on-demand and temporary access to personal data where a system may gain access to personal information only when the users explicitly request customized service. Once the system delivers the required service, it should remove personal data immediately from the system.

Furthermore, when connected vehicles communicate with road infrastructure, the cloud, or other cars, it is hard to know what kinds of information contained in these traces of communication. While the little bits of data may seem trivial, the aggregate data over an extended period may contain rich data, such as driving patterns and daily schedules of these vehicles. Protecting this data from unauthorized access is of high priority to ensure user trust and acceptance of the system. For instance, it is possible to track the movement of traffic, co-locating vehicles, and linking this information to people and places. Policies to guide how such data can be utilized to optimize traffic safety are necessary to prevent the chances of violating individuals' privacy.

7. Conclusion

Public transport systems are essential determinants of quality of life. Many countries face challenges in attempts to improve the quality of their public transport systems. Developing efficient and sustainable policies is key to durable transport systems. A report by the International Transport Forum [46] indicated that the growth and availability of vast amount of data in the transportation domain could lead to new policy-relevant insights and operational improvements of traffic. In particular, the growth in various forms of technologies, especially the increasing digitizing of transport infrastructure networks, can improve forecasting, promote reliability, and increase efficiency. Moreover, the availability of multiple types of data within the transport domain enables decision-makers to gather and triangulate various forms of evidence in real-time or near real-time to make better decisions.

The volume of transport-related data is likely to increase significantly because of growing rural-urban migration and globalization (global movement of people and goods), resulting in an increase in the amount of traffic in big cities. Vehicles generate more data from mobile devices and tracking transponders. Conventional ways of collecting and analyzing this data to yield useful insights will be a challenge. New forms of data harvesting, linking, and processing are required to enable cities, policymakers, and urban planners to gain useful insight to support decision making.

Big Data and analytics offer many opportunities to inform the development of sustainable transportation systems. The vast amount of transportation data collected by a Big Data system enables different kinds of data analytics, including historical, real-time, predictive, visual, and video and image analytics. The outcomes of this data analytics can support urban planning, provide real-time assistance on the road, and improve traffic safety. Furthermore, connected vehicles offer great potential to advance the collection, analysis, and utilization of transportation data. The challenges associated with the utilization of Big Data include dealing with heterogeneous data sources and limited communication

bandwidth, managing veracity inherent in transportation data, using multiple storage systems, and ensuring that privacy and data security are managed properly across different networks.

Acknowledgements

The project is part of institutional collaboration between MacEwan University, the City of Edmonton, and the University of Otago. We thank Elizabeth Cayen for proofreading an earlier version of this article.

Declarations of interest: none

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] C. Dobre, F. Xhafa, Intelligent services for big data science, Future Generation Computer Systems 37 (2014) 267 – 281. doi:10.1016/j.future.2013.07.014.

[2] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, J. Al-Jaroodi, Applications of big data to smart cities, Journal of Internet Services and Applications 6 (1) (2015) 25. doi:10.1186/s13174-015-0041-5.

[3] F. Loewenherz, V. Bahl, Y. Wang, Video analytics towards vision zero, Institute of Transportation Engineers. ITE Journal 87 (3) (2017) 25–28.

[4] A. B. Ayed, M. B. Halima, A. M. Alimi, Big data analytics for logistics and transportation, in: Proceedings of the 4th International Conference on Advanced Logistics and Transport (ICALT), 2015, pp. 311–316. doi:10.1109/ICAdLT.2015.7136630.

[5] M. Chen, S. Mao, Y. Liu, Big data: A survey, Mobile Networks and Applications 19 (2) (2014) 171–209. doi:10.1007/s11036-013-0489-0.

[6] W. Yuan, P. Deng, T. Taleb, J. Wan, C. Bi, An unlicensed taxi identification model based on big data analysis, IEEE Transactions on Intelligent Transportation Systems 17 (6) (2016) 1703–1713. doi:10.1109/TITS.2015.2498180.

[7] G. Kemp, G. Vargas-Solar, C. F. D. Silva, P. Ghodous, C. Collet, Aggregating and managing big realtime data in the cloud - application to intelligent transport for smart cities, in: Proceedings of the 1st International Conference on Vehicle Technology and Intelligent Transport Systems - Volume 1: VEHITS, 2015, pp. 107–112. doi:10.5220/0005491001070112.

[8] A. Artikis, M. Weidlich, A. Gal, V. Kalogeraki, D. Gunopulos, Self-adaptive event recognition for intelligent transport management, in: Proceedings of the IEEE International Conference on Big Data, 2013, pp. 319–325. doi:10.1109/BigData.2013.6691590.

[9] W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge computing: Vision and challenges, IEEE Internet of Things Journal 3 (5) (2016) 637–646. doi:10.1109/JIOT.2016.2579198.

[10] W. Shi, S. Dustdar, The promise of edge computing, Computer 49 (5) (2016) 78–81. doi:10.1109/MC.2016.145.

[11] Y. Mao, C. You, J. Zhang, K. Huang, K. B. Letaief, A survey on mobile edge computing: The communication perspective, IEEE Communications Surveys Tutorials 19 (4) (2017) 2322–2358. doi:10.1109/COMST.2017.2745201.

[12] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, X. Yang, A survey on the edge computing for the internet of things, IEEE Access 6 (2018) 6900–6919. doi:10.1109/ACCESS.2017.2778504.

[13] The vision zero network [online]. Accessed 23 September 2018.

[14] J. Chandler, S. Hopewell, Cochrane methods - twenty years experience in developing systematic review methods, Systematic Reviews 2 (1) (2013) 76. doi:10.1186/2046-4053-2-76.

[15] M. Volk, S. Bosse, K. Turowski, Providing clarity on big data technologies: A structured literature review, in: Proceedings of the IEEE 19th Conference on Business Informatics (CBI), Vol. 01, 2017, pp. 388–397. doi:10.1109/CBI.2017.26.

[16] Gartner inc.'s 2012 hype cycle for emerging technologies [online]. Accessed 24 September 2018.

[17] Y. Lin, P. Wang, M. Ma, Intelligent transportation system (its): Concept, challenge and opportunity, in: Proceedings of the IEEE 3rd international conference on big data security on cloud, 2017, pp. 167–172. doi:10.1109/BigDataSecurity.2017.50.

[18] C. Chilipirea, A.-C. Petre, L.-M. Groza, C. Dobre, F. Pop, An integrated architecture for future studies in data processing for smart cities, Microprocessors and Microsystems 52 (2017) 335 – 342. doi:10.1016/j.micpro.2017.03.004.

[19] C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, M. F. Mokbel, Towards real-time road traffic analytics using telco big data, in: Proceedings of the International Workshop on Real-Time Business Intelligence and Analytics, 2017, pp. 5:1–5:5. doi:10.1145/3129292.3129296.

[20] P. Ta-Shma, A. Akbar, G. Gerson-Golan, G. Hadash, F. Carrez, K. Moessner, An ingestion and analytics architecture for iot applied to smart city use cases, IEEE Internet of Things Journal 5 (2) (2018) 765–774. doi:10.1109/JIOT.2017.2722378.

[21] D. Serrano, T. Baldassarre, E. Stroulia, Real-time traffic-based routing, based on open data and open-source software, in: Proceedings of the IEEE

3rd World Forum on Internet of Things (WF-IoT), 2016, pp. 661–665. doi:10.1109/WF-IoT.2016.7845419.

[22] S. Aleyadeh, S. M. Oteafy, H. S. Hassanein, Scalable transportation monitoring using the smartphone road monitoring (srom) system, in: Proceedings of the 5th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications, 2015, pp. 43–50. doi:10.1145/2815347.2815349.

[23] J. Treboux, A. J. Jara, L. Dufour, D. Genoud, A predictive data-driven model for traffic-jams forecasting in smart santader city-scale testbed, in: Proceedings of the IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2015, pp. 64–68. doi:10.1109/WCNCW.2015.7122530.

[24] A. Golubev, I. Chechetkin, K. S. Solnushkin, N. Sadovnikova, D. Parygin, M. Shcherbakov, Strategway: Web solutions for building public transportation routes using big geodata analysis, in: Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, 2015, pp. 91:1–91:4. doi:10.1145/2837185.2843851.

[25] E. Abdullah, A. Emam, Traffic accidents analyzer using big data, in: Proceedings of the International Conference on Computational Science and Computational Intelligence (CSCI), 2015, pp. 392–397. doi:10.1109/CSCI.2015.187.

[26] R. Y. Ali, V. M. V. Gunturi, S. Shekhar, A. Eldawy, M. F. Mokbel, A. J. Kotz, W. F. Northrop, Future connected vehicles: Challenges and opportunities for spatio-temporal computing, in: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2015, pp. 14:1–14:4. doi:10.1145/2820783.2820885.

[27] M. Itoh, D. Yokoyama, M. Toyoda, Y. Tomita, S. Kawamura, M. Kitsuregawa, Visual fusion of mega-city big data: An application to traffic and tweets data analysis of metro passengers, in: Proceedings of the

IEEE International Conference on Big Data, 2014, pp. 431–440. doi:10.1109/BigData.2014.7004260.

[28] S. Tbatou, A. Ramrami, Y. Tabii, Security of communications in connected cars modeling and safety assessment, in: Proceedings of the 2nd International Conference on Big Data, Cloud and Applications, 2017, pp. 56:1–56:7. doi:10.1145/3090354.3090412.

[29] H. A. Najada, I. Mahgoub, Autonomous vehicles safe-optimal trajectory selection based on big data analysis and predefined user preferences, in: Proceedings of the IEEE 7th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), 2016, pp. 1–6. doi:10.1109/UEMCON.2016.7777922.

[30] D. Tosi, S. Marzorati, Big data from cellular networks: Real mobility scenarios for future smart cities, in: Proceedings of the IEEE Second International Conference on Big Data Computing Service and Applications, 2016, pp. 131–141. doi:10.1109/BigDataService.2016.20.

[31] J. Magtoto, A. Roque, Real-time traffic data collection and dissemination from an android smartphone using proportional computation and freesim as a practical transportation system in metro manila, in: Proceedings of the TENCON 2012 IEEE Region 10 Conference, 2012, pp. 1–5. doi:10.1109/TENCON.2012.6412332.

[32] Q. Shi, M. Abdel-Aty, Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways, Transportation Research Part C: Emerging Technologies 58 (2015) 380 – 394, big Data in Transportation and Traffic Engineering. doi:10.1016/j.trc.2015.02.022.

[33] W. Xu, N. R. Juri, A. Gupta, A. Deering, C. Bhat, J. Kuhr, J. Archer, Supporting large scale connected vehicle data analysis using hive, in: Proceedings of the IEEE International Conference on Big Data, 2017, pp. 2296–2304. doi:10.1109/BigData.2016.7840862.

[34] A. Attanasi, E. Silvestri, P. Meschini, G. Gentile, Real world applications using parallel computing techniques in dynamic traffic assignment and shortest path search, in: Proceedings of the IEEE 18th International Conference on Intelligent Transportation Systems, 2015, pp. 316–321. doi:10.1109/ITSC.2015.61.

[35] P. St-Aubin, N. Saunier, L. Miranda-Moreno, Large-scale automated proactive road safety analysis using video data, Transportation Research Part C: Emerging Technologies 58 (2015) 363 – 379. doi:10.1016/j.trc.2015.04.007.

[36] F. Wang, L. Hu, D. Zhou, R. Sun, J. Hu, K. Zhao, Estimating online vacancies in real-time road traffic monitoring with traffic sensor data stream, Ad Hoc Networks 35 (2015) 3 – 13, special Issue on Big Data Inspired Data Sensing, Processing and Networking Technologies. doi:10.1016/j.adhoc.2015.07.003.

[37] M. Zhang, T. Wo, T. Xie, X. Lin, Y. Liu, Carstream: An industrial system of big data processing for internet-of-vehicles, Proceedings of the VLDB Endowment 10 (12) (2017) 1766–1777. doi:10.14778/3137765.3137781.

[38] K. Zhao, S. Tarkoma, S. Liu, H. Vo, Urban human mobility data mining: An overview, in: Proceedings of the IEEE International Conference on Big Data, 2016, pp. 1911–1920. doi:10.1109/BigData.2016.7840811.

[39] B. Daniel, Big data and analytics in higher education: Opportunities and challenges, British Journal of Educational Technology 46 (5) (2015) 904–920. doi:10.1111/bjet.12230.

[40] B. K. Daniel, Big data and data science: A critical review of issues for educational research, British Journal of Educational Technologydoi:10.1111/bjet.12595.

[41] Z. Ding, B. Yang, Y. Chi, L. Guo, Enabling smart transportation systems:

A parallel spatio-temporal database approach, IEEE Transactions on Computers 65 (5) (2016) 1377–1391. doi:10.1109/TC.2015.2479596.

[42] M. Babar, F. Arif, Smart urban planning using big data analytics based internet of things, in: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers, 2017, pp. 397–402. doi:10.1145/3123024.3124411.

[43] S. Shukla, B. K, S. V. S, A framework for smart transportation using big data, in: Proceedings of the International Conference on ICT in Business Industry Government, 2016, pp. 1–3. doi:10.1109/ICTBIG.2016.7892720.

[44] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, S. Guizani, Internet-of-things-based smart cities: Recent advances and challenges, IEEE Communications Magazine 55 (9) (2017) 16–24. doi:10.1109/MCOM.2017.1600514.

[45] J. Tanis, T. Teoh, A. Aavik, A. Burgess, J. César, R. Russotto. Big data policies [online] (2018). Accessed October 2018.

[46] Big data and transport: Understanding and assessing options [online] (2015). Accessed December 2017.