Big Data Analytics in Intelligent Transportation Systems: A Survey

Li Zhu, Fei Richard Yu, *Fellow, IEEE*, Yige Wang, Bin Ning, *Fellow, IEEE*, and Tao Tang

***Abstract*— Big data is becoming a research focus in intelligent transportation systems (ITS), which can be seen in many projects around the world. Intelligent transportation systems will produce a large amount of data. The produced big data will have profound impacts on the design and application of intelligent transportation systems, which makes ITS safer, more efficient, and profitable. Studying big data analytics in ITS is a flourishing field. This paper first reviews the history and characteristics of big data and intelligent transportation systems. The framework of conducting big data analytics in ITS is discussed next, where the data source and collection methods, data analytics methods and platforms, and big data analytics application categories are summarized. Several case studies of big data analytics applications in intelligent transportation systems, including road traffic accidents analysis, road traffic flow prediction, public transportation service plan, personal travel route plan, rail transportation management and control, and assets maintenance are introduced. Finally, this paper discusses some open challenges of using big data analytics in ITS.**

***Index Terms*— Big data analytics, intelligent transportation systems (ITS), machine learning, transportation.**

I. Introduction

RECENTLY, Big Data has become a hot topic in both academia and industry. It represents large and complex data sets obtained from all kinds of sources. Many of the most popular data process techniques contain Big Data techniques, including data mining, machine learning, artificial intelligence, data fusion, social networks and so on [1]. Many people use Big Data analytics in various fields, and have achieved great success [2]. For example, in business field, some enterprises use Big Data to understand the consumer behavior more accurately so as to optimize the product price, improve operational efficiency and reduce personnel costs [3]. In social network field [3], through Big Data analytics of instant messaging, online social networking, microblog and sharing space, some companies such as Facebook, Twitter and Linkedin can understand the user's current behavior, social connections and rules of social behavior, and then promote some products. In health care field, by processing, and querying of health care data, doctors can analyze the pathogenic characteristics, assessment of the patient's physique so as to develop more humane treatment plans and suggestions and reduce incidence of patients [4]. In smart grid field, via the analysis of smart grid data, grid operators can know which parts of the electricity load and power frequency are too high, and even can diagnose which lines are in failure state. The results of these data analysis can be contributed to the upgrading of the electrical grid, renovation and maintenance work [5]. With successful application of Big Data analytics in so many fields, intelligent transportation systems also start looking at Big Data with great interests.

Intelligent transportation systems (ITS) have been developed since the beginning of 1970s. It is the future direction of the transportation system. ITS incorporate advanced technologies which include electronic sensor technologies, data transmission technologies, and intelligent control technologies into the transportation systems [6]. The purpose of ITS is to provide better services for drivers and riders in transportation systems [7]–[9].

In ITS, data can be obtained from diverse sources, such as smart card, GPS, sensors, video detector, social medias, and so on. Using accurate and effective data analytics of seemingly disorganized data can provide better service for ITS [10], [11]. With the development of ITS, the amount of data generated in ITS is developing from Trillionbyte level to Petabyte. Given such amount of data, traditional data processing systems are inefficient, and cannot meet the data analytics requirement. This is because they do not foresee the rapid growth of data amount and complexity.

Big Data analytics provides ITS a new technical method. ITS can benefit from Big Data analytics in the following aspects.

1. Vast amounts of diverse and complex data generated in ITS can be handled by Big Data analytics. Big Data analytics has resolved three problems: data storage, data analysis and data management. Big Data platforms such as Apache Hadoop and Spark are capable to processing massive amounts of data, and they have been widely used in academia and industry [12], [13].

Manuscript received August 1, 2017; revised January 14, 2018; accepted February 24, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61603026, in part by the Beijing Natural Science Foundation under Grant L171004, in part by the Technological Research and Development Program of China Railway Corporation under Grant 2016X008-B, in part by the State Key Laboratory of Rail Traffic Control and Safety under Grant RCS2017ZT006, and Project KIE017001531, and in part by the Beijing Key Laboratory of Urban Rail Transit Automation and Control. The authors declare that there is no conflict of interest regarding the publication of this paper. The Associate Editor for this paper was J. E. Naranjo. *(Corresponding author: Fei Richard Yu.)*

L. Zhu, Y. Wang, B. Ning, and T. Tang are with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: lizhu@bjtu.edu.cn; 15120287@bjtu.edu.cn; bning@bjtu.edu.cn; ttang@bjtu.edu.cn).

F. R. Yu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: richard.yu@carleton.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2815678

2. Big Data analytics can improve the ITS operation efficiency. Many subsystems in ITSs need to handle large amount of data to give information or provide decision to manage traffic. Through fast data collection and analysis of current and historical massive traffic data, traffic management department can predict traffic flow in real time. Public transportation Big Data analytics can help management department to learn the riders journey patterns in the transportation network, which can be used for better public transportation service planning. Big Data analytics of transportation APP developers can help the users to reach their destination in a most suitable route and with the shortest possible time.

3. Big Data analytics can improve the ITS safety level. Using advanced sensor and detection techniques, massive amount of real time transportation information can be obtained. Through Big Data analytics, we can effectively predict the occurrence of traffic accident. When accidents happens, or emergency rescue is needed, the real time response capability in the Big Data analytics based system can greatly improve the emergency rescue ability. Big Data analytics can also offer new opportunities to identify assets problems, such as pavement degradation, ballast aging, etc. It can help make maintenance decision in an appropriate time, and prevent the vehicle or infrastructure from being in a failure state.

Although applications of Big Data analytics in ITS have the great vision, many critical research issues and significant challenges remain need to be addressed. To the best of our knowledge, a systematic summary of Big Data analytics from data sources and collection methods, data analytics methods and platforms, to Big Data analytics applications in ITS has not been done before. In this survey, we first discuss the sources of Big Data in ITS and how we can collect the generated Big Data. The framework of conducting Big Data analytics in ITS is discussed. We also summarize the data analytics methods and platforms in ITS. Some case studies of Big Data analytics applications in ITS are introduced as well.

The rest of paper is organized as follows. The architecture of conducting Big Data analytics in ITS is discussed in Section II. Section III summarizes the data source and collection methods. Big Data analytics methods are discussed in Section IV. Section V introduces the cases studies of ITS Big Data analytics applications in details. We present the Big Data analytics platforms in Section VI. Some open challenges of using Big Data analytics in ITS are discussed in Section VII. Finally, We conclude the paper in Section VIII.

II. The Architecture of Conducting Big Data Analytics in ITS

A. Big Data Characteristics in ITS

Intelligent transportation system incorporates advanced technologies which include electronic sensor technologies, data transmission technologies, and intelligent control technologies into the transportation systems [6]. The purpose of ITS is to provide better services for drivers and riders in transportation systems [7]. According to [7], ITS includes six fundamental components: advanced transportation management systems, advanced traveler information systems, advanced vehicle control systems, business vehicle management, advanced public transportation systems, and advanced urban transportation systems. Literature review [7]–[9] indicates that most of these components are specific to vehicles and road transportation. Therefore, we focus on ITS in-road transportation in this survey paper.

The data collected by the intelligent transportation systems (ITS) are increasingly complex and are with Big Data features. Big companies including Gartner IBM and Microsoft put forward that that Big Data could be described by three Vs, i.e., volume, variety, and velocity [14], [15].

Volume refers to the quantities of data produced by various sources and are still expanding. With the growth of the amount of traffic, and detectors, the volume of data in transportation has increased significantly. In addition, travelers, goods and vehicles generate more data when tracking transponders are used. The data generated from infrastructures, environmental and meteorological monitoring is also increasing as a critical part of transportation data.

Variety is mainly focused on all kinds of data produced by detectors, sensors, and even social media. The variety of transport-related data has increased remarkably. For example, modern vehicles can report internal system telemetry in real time and the information of all crew members and passengers.

The velocity of data in transportation has increased due to improved communications technologies, increased processing power and speed of monitoring and processing. For example, ticketing and tolling transactions that use smart cards or tags are now immediately reported, whereas paper-based ticketing needs human processing to acquire helpful data from the transactions.

B. The Architecture of Conducting Big Data Analytics in ITS

The architecture of conducting Big Data analytics in ITS is shown in Fig. 1. It can be divided into three layers, which are data collection layer, data analytics layer, and application layer.

- Data collection layer: Data collection layer is the basis of the architecture, since it provides the necessary data for the upper layer. The data come from diverse sources such as induction loop detectors, microwave radars, video surveillance, remote sensing, radio frequency identification data, and GPS, etc. Details about collection of Big Data will be introduced in next sections.
- Data analytics layer: Data analytics layer is the core layer of architecture. This layer is primarily to receive data from the data collection layer, and then apply various Big Data analytics approaches and the corresponding platform to complete data storage, management, mining, analysis, and sharing. Details about the Big Data analytics approaches and platform will be introduced in next sections.
- Application layer: Application layer is the topmost layer in this architecture. It applies the data process results from the data analytics layer in different transportation circumstances, for example, traffic flow prediction, traffic guidance, signal control, and emergency rescue, etc.

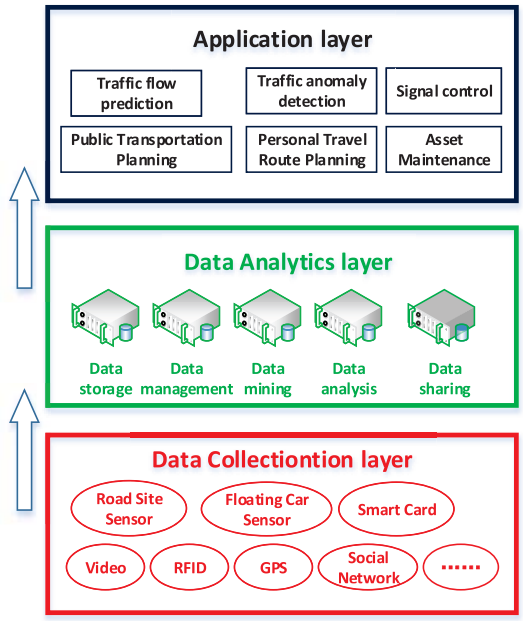


Fig. 1. Architecture of conducting Big Data analytics in ITS.

Using advanced data collection techniques, the data collection layer monitors people, vehicles, roads and the environment. The original traffic data which includes structured data, semi-structured and mixed data is transmitted to the data analytics layer via wired or wireless communication. After the data analytics layer receives the original traffic data, it first classifies the data, removes duplicate data, cleans the data and distributes the useful and accurate data in a distributed manner. Then it uses mathematics and engineering theory to extract the hidden information, mainly including descriptive analysis and predictive analysis. Using the analysis results, the application layer can predict the trend of future traffic flow and passengers flow, analyze the traffic accident prone locations, adjust the signal distribution, and implement traffic control to provide decision support for the city management department.

III. Big Data Collection in ITS

People unconsciously participate in the collection, transmission and application of Big Data in ITS. The technology development in ITS has led to an increase in the complexity, diversity and amount of data created and collected from vehicle, and people movements. According to different sources in ITS, Big Data in ITS can be primarily categorized into the following types, and the collected data is illustrated in Table I.

TABLE I
Big Data in ITS

Source	Tools	Data
Smart Card	Smart Card	OD Flows, Travel Time
GPS	GPS	Vehicle Position, Vehicle Density, Vehicle Speed
Video	Video Camera	Vehicle Position, Vehicle Speed, Vehicle Density, Vehicle Classification
Road Site Sensor	Induction Loops, Road Tubes, Microwave Radar, LIDAR/Infared Acoustic, Toll Plazas	Vehicle Position, Vehicle Speed, Vehicle Density, Vehicle Classification
Floating Car Sensor	License Plate Recognition, Transponders	Travel Time, OD Flows
Wide Area Sensor	GPS, Cell phone Tracking, Airborne Sensors	Travel Time, OD Flows
Connected and Autonomous Vehicles (CAVs)	Diverse Sensors	Coordinate, speed, acceleration, safety data,
Passive Collection	Social Media, Mobile Phone Data	Travel Time, OD Flows
Other Sources	Smart Grid, Smart Meters, Cellular Service, Dedicated Tests	Electric and Energy Consumption, Location, Channel Data

A. Big Data From Smart Cards

Automatic Fare Collection (AFC) systems has been widely deployed in urban rail systems, which makes the smart card data become the main data source for investigating the passengers movement patterns [16]–[18]. In AFC systems passengers are required to use smart cards when they take buses or trains. The electronic readers will capture passenger details such as boarding time, OD information, etc., when they touch their smart cards. Smart cards in AFC systems generate huge amount of data records every day in big cities. For instance, Transportation for London (TfL) collects smart card data from 8 million trips every day at London metro stations.

Substantial work has been done to use smart card data to study the spatial and temporal patterns of public transportation passenger travel behaviour [19]–[22]. Due to its potential capacity of offering comprehensive spatial-temporal information on travel behaviour [17], [21], smart card data is becoming a significant component of public transportation services planning and management.

B. Big Data From GPS

GPS is the most popular tool for location tracking. Traffic data can be collected more efficiently and safely with location tracking via GPS. Combining geographic information system (GIS) or other map displaying technologies, GPS provides a promising tool for data collection, and the collected data can be used for addressing many traffic issues, such as travel mode

detection [23], [24], travel delay measurement [25] and traffic monitoring [26].

C. Big Data From Videos

Video cameras are widely deployed in ITS. As demonstrated in advanced traffic management systems (ATMS), video image detection systems (VIDS) are good alternatives compared with conventional sensors for tasks like vehicle identification and traffic flow detection. One advantage of VIDS is the low cost [27]. Freeway imaging sensors that use massive video data have been successfully deployed to carry out incident detection and have shown high accuracy in certain circumstance [28]. Apart from general traffic management [29], transportation engineers and planners that collect more accurate vehicle video data can improve the image process system so as to be better at making general transportation demand regarding vehicle emission models.

D. Big Data From Sensors

Sensor equipment installed in ITS is used to collect data such as vehicle speeds, vehicle density, traffic flows, and trip times. Traditional on-road sensors, (e.g., infrared and microwave detectors), have been evolving to obtain, compute and transfer traffic data [30]. As presented in [30], data collection from sensors can be divided into three sources: roadside data, floating car data, and wide area data [31].

Roadside data mainly refers to the data collected by sensors located along roadside. Traditional roadside sensors such as inductive magnetic loops, pneumatic road tubes, piezoelectric loops arrays and microwave radars have been used for many years. New generation roadside sensors such as ultrasonic and acoustic sensor systems, magnetometer vehicle detectors, infrared systems, light detection and ranging (LIDAR), and video image processing and detection systems gradually appear with recent advanced technology developments.

Floating car data (FCD) mainly refers to the vehicle mobility data at different locations in ITS, where customized detectors are embedded in vehicles [32]. Some onboard sensors provide confident and efficient information for travel route selection and estimations. With developments of vehicle sensor technique, popular FCD sensors techniques include: automatic vehicle identification (AVI), license plate recognition(LPR), and transponders such as probe vehicles and electronic toll tags.

Wide area data refers to the wide area traffic flow data that is collected by diverse sensor tracking techniques such as photogrammetric processing, sound recording, video processing, and space-based radar.

E. Big Data From CAV and VANET

Connected and autonomous vehicles (CAV) are new technologies in ITS area that combines radical changes of vehicles design and their interactions with the road infrastructure. Connected and autonomous vehicles incorporate a range of different technologies, facilitating the safe, efficient movement of people and goods. CAV enabled traffic system has demonstrated great potential to mitigate congestion, reduce travel delay, and enhance safety performance [33], [34]. CAVs can generate big amount of environmentally relevant real-time transportation data, such as coordinate, speed, acceleration, safety data [33]. Using latest network technologies such as Software Defined Networking, data can be obtained more efficiently [35] These data can be used to create actionable information to support and facilitate green transportation choices, and apply to the real-time adaptive signal control [36], [37].

Vehicle Ad Hoc Network (VANET) is a kind of mobile ad hoc network that uses vehicles and infrastructure elements as nodes to increase the coverage area and the communication capabilities. As an important part of ITS, VANET generates large amounts of data [38]. Data preparation and real-time results are challenging tasks for large-scale analysis. Using Big Data analytics, we can address most of data related VANET challenges [39], such as data filting [40], congestion and accidents alerting [41], and Traffic Flow prediction [42].

F. Big Data From Passive Collection

Compared with the actively collected data in transportation research, the rapid development of mobile technologies have enabled the collection of a massive amount of passive data. Passive data refers to those data not collected through active collection. It is generated for purposes that are not intended but can be potentially used for research [43], [44]. Chen *et al.* [45] and Zeyu *et al.* [46] propose to combine passive Big Data such as mobile phone data, internet access data and active data to study human mobility, travel behavior, and transportation planning. In [47], contextual information such as current time, cell phone ID, user identity are used for predicting the stay time of mobile users.

Social media data is the most popular passive data, and it refers to applications or websites where people interact with each other to create, share, and exchange information and ideas. Social media networks such as Linkedin, Facebook and Twitter have been developed rapidly recently. They have become relevant interests of transportation professionals as they provide information flows between providers and consumers in real time [48]. Though data collected via social media networks is generally unstructured and requires complicated processing, it provides significant transportation information when attitudes are expressed in different kind of transportation, and responses to travel disruptions are found in social media [49]–[52].

G. Big Data From Other Sources

There are some sources of data that cannot be classified into the above categories. For example, real-time infrastructure state is considered as an important source of data [53]. The best known example is the smart grid [54], which will allow us to collect daily electricity consumption information for electric vehicles and train traction in urban rail transportation system.

Another important data source is the data from dedicated test in ITS. For example, in our previous work, we carry out field tests in a real train ground communication system in urban rail transportation Communication Based Train Control (CBTC) system [55], [56]. A large amount of channel gain data is obtained from the field test. The data is processed

to model the stochastic characteristic of channel state, and the model is used to optimize the CBTC system performance.

IV. Big Data Analytics Methods ITS

Machine learning is most popular modelling and analytics theory in Big Data ecosystems, which makes it easy to derive patterns and models from large amount of data. In ITS areas, machine learning theory has also be widely used to conduct data analytic. Depending on the completeness of data set that is available for learning, Machine learning models can be categorized into supervised, unsupervised and reinforcement learning algorithms. With the recent rapid development of Artificial Intelligence, the powerful deep learning models have also been adopted to ITS recently.

A. Supervised Learning

Labeled training data is used in supervised learning algorithms [57]. The models use input data and the target outputs (labels) to learn the function or map between them. Combined with the learned model and the input data, the unseen outputs can be predicted. Among all the supervised learning models, linear regression, decision trees, neural networks, and support vector machines, are the most frequently used in ITSs.

The function of regression is to explain the relationship between one dependent variable and one or more independent variables. Linear regression is the most commonly used supervised learning [58]. Linear regression is incredibly simple, robust, easy to interpret, and easy to code. Despite its simplicity, linear regression is particularly successful in various ITS scenarios, such as traffic flow prediction [59], traffic speed estimation [60], and transportation travel route evaluation [61].

A decision tree is a decision support tool that uses a tree-like graph to model decisions and their possible consequences [62]. Due to their portability, robustness and transparency, decision trees are widely used in various ITS scenarios, such as traffic accident detection [63], accident severity analysis [64] and travel mode choice [65].

Artificial Neural network (ANN) is a popular example of flexible and robust supervised learning for both classification and regression [57]. With enough hidden layers of processing nodes and training data, ANN can learn any non-linear relations between input and target data. As a data modeling tool, it has also been adopted in ITS such as traffic flow prediction [66], travel time prediction [67], traffic accident detection [68] and remaining parking spaces forecasting [69].

Support vector machine (SVM) is another popular supervised learning algorithms that use labelled data for regression and classification. Among all the Big Data analytics model tools in ITS, SVMs have attracted great interests in research area. It has been successfully used in travel time prediction [70], bus arrival time prediction [71], and traffic accident detection [72].

A typical example of using supervised learning in ITS is introduced in [72], where SVM is used to predict traffic incidents. Given the training subset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i) \ldots\}$, where x_i is the input of the training sample which consists of the values of the traffic flow parameters such as volume, speed, occupancy and so on, and y_i is the class label of x_i. With a kernel function $K(x, x')$, according to the SVM classifier theory, the support vector α_i can be obtained as,

$$\max_{\alpha_i} \; -\frac{1}{2}\sum_{i=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) + \sum_{i=1}^{l} \alpha_i$$
$$s.t. \; \sum_{i=1}^{l} y_i \alpha_i = 0, \tag{1}$$

Then, we get the decision function $g(x)$ to compute the label for the sample x as,

$$g(x) = sgn(\sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b). \tag{2}$$

If x is an incident sample, $g(x) = 1$. Otherwise, we have $g(x) = -1$.

B. Unsupervised Learning

Unsupervised learning normally also referred as clustering focus on learning natural group from unlabeled multidimensional data [57]. K-means is the most popular unsupervised learning tool, and it has been widely adopted in highway transportation planning [73], and travel time prediction [74].

With a set of historical data, authors of [74] gives a classic example of using unsupervised learning to predict travel time. The procedures are as follows,

1. Compute the travel time frequency ε. It means the number of time that the travel time appears.

2. Define a tuple $\Gamma(\tau_i, \varepsilon_i, \nu_i)$ that contains distinct features, where τ_i is the travel time, ε_i is the travel time frequency, and ν_i is the travel velocity.

3. Find the greatest value in the data based on the travel time frequency. A tuple $\Gamma(\tau_p, \varepsilon_p, \nu_p)$ is chosen as a centroid of Cluster 1, where ε_p is the maximum travel time frequency, τ_p is the corresponding maximum travel time associated with ε_p, and ν_p is the travel velocity associated with ε_p.

4. Compare each tuple $\Gamma(\tau_i, \varepsilon_i, \nu_i)$ with the centroid $\Gamma(\tau_p, \varepsilon_p, \nu_p)$ of Cluster 1 by compute their distance. Choose the tuple $\Gamma(\tau_q, \varepsilon_q, \nu_q)$ with the maximum distance.

5. Build two clusters where the centroid of Cluster 1 is tuple $\Gamma(\tau_p, \varepsilon_p, \nu_p)$ and that of Cluster 2 is tuple $\Gamma(\tau_q, \varepsilon_q, \nu_q)$.

6. Define the cluster memberships of all the tuples by assigning them to the nearest cluster centroid.

7. Re-estimate the cluster centre using the arithmetic mean.

8. Repeat step 6 and 7.

9. After complete preparation of clusters, desired predicted time is calculated separately for each cluster as, $\varsigma_k = \sum_{j=1}^{N} \varepsilon_j * \tau_j / \sum_{j=1}^{N} \varepsilon_j$. Where ς_k is the travel time obtained from kth cluster, N is the total number of tuple in the associated cluster, ε_j is the travel time frequency of the jth tuple, and τ_j is the travel time of the jth tuple.

10. The final predicted approximate travel time is obtained by computing the arithmetic mean of ς_1 and ς_2.

C. Reinforcement Learning

Different from supervised and unsupervised learning, as shown in Fig. 2, the aim of the reinforcement learning is to minimize the long term cost through exploration and

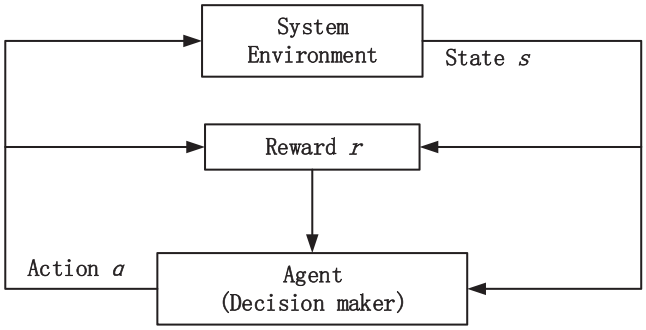


Fig. 2. Reinforcement learning.

learn the optimal policy by interacting with the experimental data [57]. Reinforcement learning is highly relevant to control and optimization theory, and it has been proved to be quite feasible in traffic signal control in ITS [75]–[78].

Using reinforcement learning in ITS requires a formulation of the ITS control and optimization problem in the language of reinforcement learning, specifically, defining a state space S, an action space A and a reward R. One classic example of using reinforcement learning in ITS traffic signal control is formulated in [76]. The state of traffic at an intersection with n lanes is formally defined as the discrete traffic state encoding (DTSE). For each lane approaching the intersection, the DTSE discretizes a length l of the lane segment, beginning at the stop line, into cells of length c. The selection of c will change the behavior of system. The DTSE is composed of three vectors. The first vector B represents the presence of a vehicle or not in the cell. The second vector R represents the speed of the vehicle, and the third vector P is the current traffic signal phase (i.e., the most recent action selected). Thus, the system states can be defined as, $S \in (BR)^{\frac{L}{c}n} P$.

After the agent has observed the state of the environment, it must choose one action from the set of all available actions. The possible actions are North-South Green (a_1), East-West Green (a_2)), North-South Advance Left Green (a_3)), East-West Advance Left Green (a_4)). The set of all possible actions A is defined as $A = \{a_1, a_2, a_3, a_4\}$. At time t, the agent chooses an action $a(t)$, where $a(t) \in A$.

After the agent has observed the state of the environment s_t, it performs an action $a(t)$, and receives the reward. The reward r_{t+1} is a consequence of performing a selected action from a specific state. In this formulation, the reward is defined as change in cumulative vehicle delay between actions.

The reinforcement learning algorithm used in this formulation is Q-Learning [57], which is used to develop an optimal action-selection policy. The optimal policy is achieved by using the convolutional neural network to approximate the action-value function. The action-value function $Q(s_t, a_t)$ maps states to action utilities (i.e., what is the value of each action from a given state). The basis of Q-learning is the value iteration update defined as,

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_A Q(s_{t+1}, a_t) - Q(s_t, a_t)). \quad (3)$$

Where the learning rate α controls the degree to which new action-value estimates are weighted against old estimates and the discount factor γ determines how immediate rewards are weighted against future rewards. After the action-value function has been sufficiently learned, the optimal policy can be determined by selecting the action with the highest value.

D. Deep Learning

Deep learning models exploit much more system features and complex architecture than traditional Artificial Neural Network, and can achieve better performance than traditional machine learning models. They have been widely applied in ITSs. For example, a deep Restricted Boltzmann Machine and Recurrent Neural Network architecture is utilized to model and predict traffic congestion evolution based on GPS data from taxi [79]. Using deep neural networks, fault diagnoses on bogies with Big Data is carried out in [80]. Chen [81] carry out the vehicle detection task using the rich feature of convolutional neural network(CNN) learned from ImageNet dataset. Duan *et al.* [82] use stacked auto-encoders for traffic data imputation. In traffic flow area, deep learning model has become a popular tool to predict traffic flow density [83]–[86].

Literature [85] gives a typical deep learning based approach to do the traffic flow prediction. Stacked autoencoders (SAEs) are used to learn generic traffic flow features. Considering SAEs with K layers, the first layer is trained as an autoencoder, with the training set as inputs. After obtaining the first hidden layer, the output of the jth hidden layer is used as the input of the $(k+1)$th hidden layer. In this way, multiple autoencoders can be stacked hierarchically. To use the SAE network for traffic flow prediction, a logistic regression layer is added on top of the network for supervised traffic flow prediction. The whole deep architecture model is shown in Fig.3.

The data collected from all freeways are used as the input. Considering the temporal relationship of traffic flow, the traffic flow data at previous time intervals, i.e., $\chi_{t-1}, \chi_{t-2}, \ldots, \chi_{t-l}$ are used to predict the traffic flow at time interval t. The proposed model accounts for the spatia land temporal correlations of traffic flow inherently.

E. Ontology Based Methods

An ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really exist in a particular domain of discourse. Ontology based methods can accurately describe data semantics and infer implicit data semantic relations. Compared with the traditional data extraction from the bottom up, ontology data integration has a top-down feature and uses ontology modeling to share semantic views of data and map heterogeneous data from different data sources to minimize or even eliminate the ambiguous understanding of shared data. Ontology based method has been widely applied in ITSs. For example, Zhai *et al.* [87] propose an information retrieval system for ITS based on a fuzzy ontology framework. This framework includes three parts: concepts, properties of concepts and values of properties, and it focuses on information about traffic accidents. Fernandez and Ito [88] propose a driver behavior model based in ontology for intelligent transportation system. The driver behavior ontology has the knowledge related to driver characteristics, perception and cognitive state to perform different driving

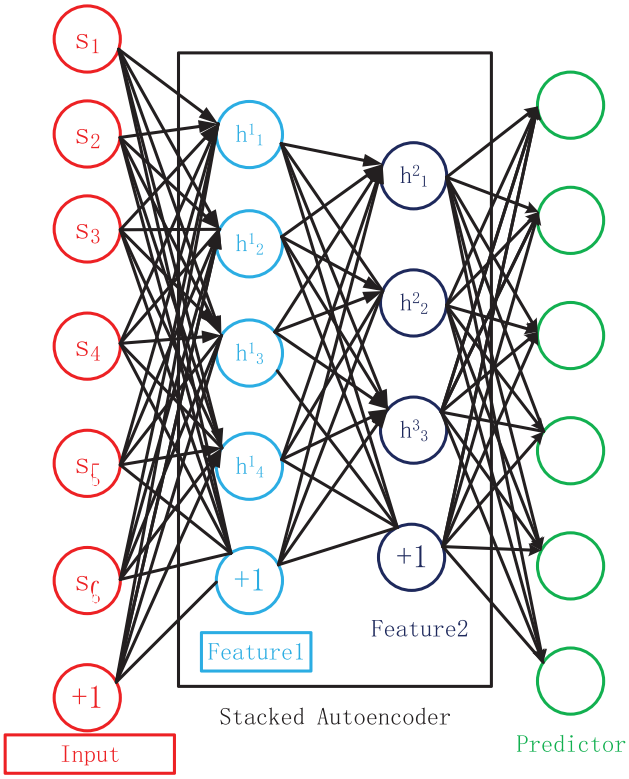


Fig. 3. Deep architecture model for traffic flow prediction.

tasks. It can be used to predict traffic accidents and optimize the road congestion. Fernandez and Ito [89] propose to use the ontology to manage the sensor information in intelligent transportation systems and convert the sensor data into semantic data. The system performs the automatic traffic light settings can use the data to predict and avoid traffic accidents. Gregor *et al.* [90] propose a systematic methodology to create ontology in ITS domain. This ontology will serve as the basis of semantic information to a semantic service that allows the connection of new equipment to an urban network. Zhao *et al.* [91] introduce an ontology-based Knowledge Base, which contains maps and traffic regulations. By accessing to the Knowledge Base, the intelligent vehicles can be aware of over speed situations and make decisions at intersections in comply with traffic regulations. Chen *et al.* [92] depict an ontology-based approach for safety management in Cooperative ITS (C-ITS), primarily in an automotive context. It provides the support for ontology driven ITS development and its formal information model. Yang and Wang [93] take advantages of the semantic completeness of the ontology to build urban traffic ontology model, which resolve the problems as ontology mergence and equivalence verification in semantic fusion of traffic information integration. The model can increase the function of semantic fusion, and reduce the amount of data integration of urban traffic information as well enhance the efficiency and integrity of traffic information query.

V. Big Data Applications in ITS

Big Data provides technical supports for the development and applications of ITS. By efficient, accurate and timely data collection, analyzing and processing in road and rail transportation system, the Big Data applications can provide the public with convenient and high efficient transportation. In order to identify problems, improving ITS efficiency, reducing costs and deriving valuable insights, Big Data applications in ITS can be divided into the following six categories.

A. Road Traffic Accidents Analysis

Evidence shows that in the world around 1.2 million people are killed and 50 million injured from traffic accidents every year [94]. Accurate traffic accident data analysis results can provide traffic department with important information to make policies so as to prevent accidents.

Many studies focused on using Big Data analytics in traffic accidents analysis. Using measured traffic flow data, Golob and Recker [95] study the relationships among weather, lighting conditions, traffic flow, and urban freeway accidents, with a multivariate statistical model. In [10], Bayesian inference and Random forest are adopted in a real-time crash prediction model to reduce crash risks. Xiong *et al.* [96] introduce classification and regression trees (CART), logistic regression and multivariate adaptive regression splines (MARS) to perform analytical operation on motor vehicle accident injury data. Lee and Mannering [97] present a method which uses zero-inflated count models and nested logit models to analyze run-off-roadway accident frequency and severity on a 96.6km section of highway in Washington State. The results show that some measures can be taken to reduce run-off-roadway accident frequencies. Karlaftis and Golias [98] apply a rigorous non-parametric statistical methodology which is hierarchical tree-based regression (HTBR) to analyze the influences of terrain and traffic characteristics on accident rates of rural roads. The methodology can also be used to predict the accident rates of highway. Chang, et. analyze the relationship between highway geometric variables and traffic accidents by using a negative binomial regression model and a classification and regression tree model. The parameters come from the 2001-2002 accident data of National Freeway 1 in Taiwan [99]. Bédard *et al.* [100] determine the respective effect of driver, crash, and vehicle characteristics to the fatality risk of drivers by using a multivariate logistic regression algorithm, the results indicate that increasing seat belt use, reducing vehicle speed, and decreasing the number and severity of driver-side impacts could prevent traffic accidents.

B. Road Traffic Flow Prediction

Timely and accurate traffic flow information is critical for transportation management. Big Data analytics in ITS has an advantage in traffic flow prediction [101]–[103]. According to [9], a classic road traffic flow prediction model using Big Data analytics is shown in Fig 4. The original ITS data is first preprocessed to get the effective data set. Using selected data mining or analysis method, traffic flow model is established with the preprocess data. The traffic flow model gives decision supports to traffic management department and get feedback from real traffic flows to calibrate the model.

Many scholars have studied traffic flow prediction using Big Data analytic. Lv *et al.* [85] propose a deep learning based

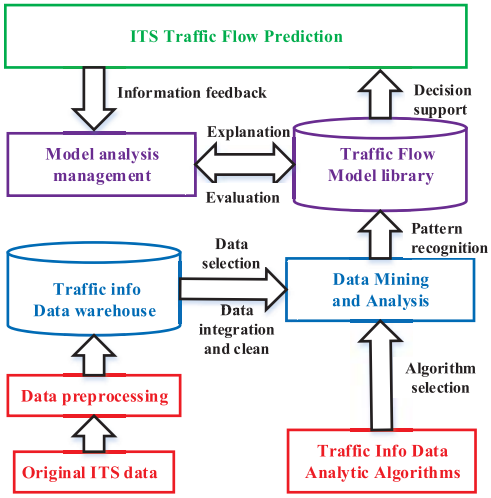


Fig. 4. A typical traffic flow prediction model.

traffic flow prediction method which use the greedy layerwise unsupervised learning algorithm. Stacked auto encoder (SAE) model is used to learn generic traffic flow features. The results show that the deep learning based model has superior performance for traffic flow prediction. Liu *et al.* [104] analyze multidimensional parameters and the traffic flow prediction models is developed from different dimensions based on SVMs. Dong *et al.* [105] propose a pre-selection space time model to estimate the traffic flow at locations with little data detectors. Canaud *et al.* [106] present a probability hypothesis density filtering based model for real-time traffic flow prediction. Pan *et al.* [107] put forward a modified stochastic cell transmission model to support short-term traffic flow prediction. Antoniou *et al.* [108] propose an approach for local traffic flow state estimation and prediction based on data-driven computational approaches. Using the seemingly unrelated time-series equation (SUTSE), Ghosh *et al.* [109] present a new multivariate structural time-series (MST) model to predict traffic flow. The SUTSE model can respectively track the change of each traffic flows and their components as time goes by, and the results show it has a superior prediction accuracy. Xu *et al.* [110] propose a novel online algorithm which is a context-aware adaptive traffic prediction algorithm. The algorithm can learn from the current traffic condition and use the historical traffic data to predict the future traffic flow. The experiments indicate that this algorithm do better than the current solutions. Lu *et al.* [111] build a traffic flow state clustering model which adopts the simulated annealing genetic algorithm using fuzzy c-means (SAGA-FCM). This model is based on traffic speed data and occupancy data which comprehensively considers the temporal, spatial, and historical correlations of traffic flow Big Data.

With the recent rapid development of AI technology, deep learning methods have been widely applied to predict traffic flow. Huang *et al.* [84] introduce deep belief network into transportation system. Ma *et al.* [79] combined deep restricted Boltzmann machines (RBM) with RNN and formed a RBM-RNN model that inherits the advantages of both RBM and RNN. They also [86] use LSTM to predict traffic and demonstrate that LSTM achieve better performance compared with traditional neural networks in both stability and accuracy regarding traffic speed prediction by using loop detector data collected in the Beijing road network. Lv *et al.* [85] propose a novel deep-learning- based traffic prediction model that considered spatiotemporal relations, and employed stack autoencoder (SAE) to extract traffic features.

C. Public Transportation Services Planning

Public transportation Big Data analytics can help to understand transportation riders journey patterns across the transportation network. The riders journey patterns can be used to inform decisions to transportation operators about the services planning.

With heterogeneous sources of traffic measurements data, Lu *et al.* [112] present a path flow based nonlinear optimization model to estimate dynamic OD demand that does not need explicit dynamic link information. Using triangulated mobile phone records of millions of anonymous users, authors of [113] present a method to predict average daily OD trips. The applicability of the proposed model is verified by the spatial and temporal distributions of trips get from local and national surveys. Using complete daily set of smart card data from London Metro and iBus vehicle location system, Gordon [114] derives the boarding and alighting times of every passenger, and transfer information is derived from passenger trips belong to different public transportation modes. The full journey matrices are established from the data, and are validated by traditional O-D matrices. The approach is efficient enough to be performed daily and provide the transportation operators travel behavior of their services. The Big Data analytics results in these works can help the emerging intelligent traffic management applications generate proactive, coordinated traffic information provision. Tao [115] investigate the temporal and spatial dynamics of Bus Rapid Transit (BRT) trips against non-BRT trips during five typical calendar events. The smart card data is first pre-processed to build OD flow matrices and bus trip route for BRT and non-BRT trips respectively. Service management department can identify important implications for evidence-based BRT policies. In [116], operational Big Data from Automated Fare Collection (AFC) systems is used for transportation planning management in Istanbul, Turkey. Works in MIT [117] shows the potential value of London AFC data in rail transportation planning and operations. The applications developed in their work provide rail transportation operators and planner an easy-to-update management tool that evaluates rail service in several aspects at near real-time. Toole *et al.* [118] use mobile phone data from open source data repositories to implement a travel demand model. Routable road networks, validated OD matrices and trip tables can be extract from the Call Data Record (CDR) data with the model. Their work serve as universal guide to help the transportation operators perform public transportation planning.

D. Personal Travel Route Planning

The transportation Apps start with great vision. Report suggests that only telling passengers the arriving time of

the next bus could make them more satisfied with the bus service [119]. Based on the data from smart phones and vehicle GPS data, some transportation APPs provide riders with real time traffic information [120], others provide most suitable driving routes with minimum travel time [121]. Combined with public transportation data with information from users through their smart phones, transportation APPs can even provide riders with real time public transportation journey planning [122]. Fully integrated Apps even let people plan trips that move from trains to buses and private cars or bicycles at the ends [123].

Big Data analytics in these transportation APPs generates huge economic benefits by reducing travel time, traffic congestion, pollution, and greenhouse-gas emissions. For example, opening up Transportation for London (TfL) data has been valued at 15-58 million pounds per year and has resulted in over 200 travel Apps being developed by private companies [124].

E. Rail Transportation Management and Control

Rail transportation systems have been transformed with advanced IT technology. They are the main beneficiary of Big Data analytics. This is because that Rail transportation systems are generally closed systems that carry out sophisticated processing of large volumes of data, such as real time train speed and position, train departure and arrival time of a certain station, and passenger OD information. Big Data analytics can make the rail transportation operators be better at train control and improve the rail transportation system operation efficiency.

In industry, Big Data analytics is starting to play an important role in rail transportation system. As a typical public rail transportation system, the Bay Area Rapid Transit (BART) maintains supervision over all phases of its system, including train operations, passenger services, power delivery, and wayside facilities. Big Data analytics is a key element within all of these functions. Schultz in [125] point out that the critical role of BART's operational analytics is ensuring schedule reliability. Using Markov chain model, a multimodal transportation network in London is developed with better information clusters for transportation efficiency [126]. In [127], Big Data analytics is applied in Utrecht, Netherlands to predict the traffic and improve operations with data from mobile phones, smart cards and computers.

In academia, substantial work has been done about using Big Data analytics in rail transportation management and control. Using the passenger OD information of Shanghai rail transportation line 1, Jiang *et al.* [128] evaluate the train timetable efficiency. This method is verified in a real rail transportation system that involves more than 1 million passenger trips and 600 trains. Yin *et al.* [129] present a smart train operation (STO) method which combines the advantages of automatic train operation (ATO) and manual driving. The fusion of expert knowledge and data mining algorithm is applied in STO method. The results suggest that in energy consumption and riding comfort, this proposed method is better than ATO, and in punctuality and parking accuracy, it is also better than manual driving. Chen *et al.* [130] propose two simplified models about the relationship between train stop error and the train control parameters by using train speed data, time data and distance data before stopping, and introduce one online learning algorithm - polynomial adaline algorithm to increase the parking accuracy. The results show that the proposed simplified models and the online learning algorithms are effective in reducing the parking error and correct the bias of train stop error distribution. Zhou [131] apply two typical machine learning algorithms Gaussian processes and Boosting to improve train stop accuracy by utilizing a number of the initial velocity data and distance data before stopping, the results show that Gaussian process regression algorithm gets the best performance. Hou *et al.* [132] propose three train stop control algorithms which chooses initial braking position data, braking force data and their combined data as control input. Based on terminal iterative learning control (TILC), these algorithms use the stop position error in previous braking process to improve train stop accuracy. Chen *et al.* [133] use a new machine learning technique and propose novel online learning control algorithms to realize train automatic stop control. The algorithm includes heuristic online learning algorithm (HOA), gradient-descent based online learning algorithm (GOA), and RL-based online learning algorithm (RLA). The required parameters come from the track-side balises. The results suggest that this method can limit the stopping errors in the range of ±0.30m under regular interferences.

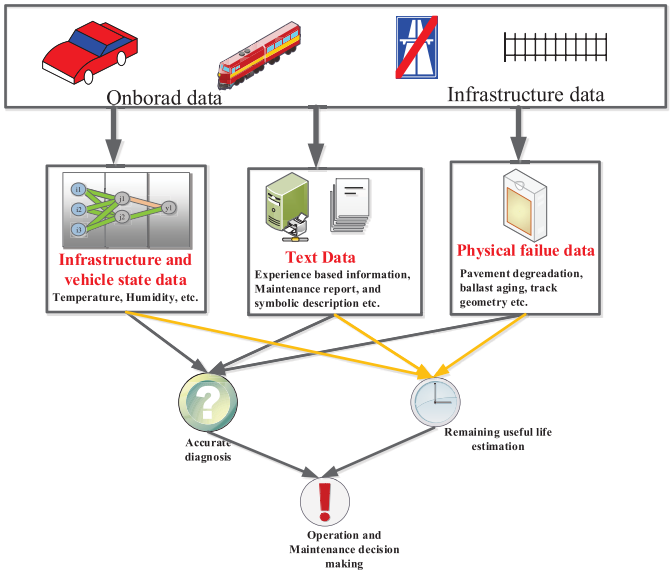


Fig. 5. A typical framework of using Big Data analytics for asset maintenance.

F. Asset Maintenance

In ITS, there are substantial asset that is dependent on large amounts of data to operate and maintain. Proper asset maintenance approach is very important for protecting ITS capital and reduce maintenance costs. Big Data analytics can help identify problems more quickly and accurately, and minimize maintenance costs. A typical framework of using Big Data analytics for asset maintenance decision making [134] is shown in Fig. 5. Onboard and Infrastructure data is collected from different sensors. Physical failure data such as pavement degradation, ballast aging, track geometry etc. can be used directly. Text data such as experience based information

and maintenance report, symbolic description etc., can be processed to extract important information. Infrastructure and vehicle state data such as temperature, humidity, etc. can be processed with data driven method, and obtain the condition indicators. The results from the three process methods is integrated and get the accurate diagnosis of asset condition and determine the remaining useful life of asset, which can be used for end users to make maintenance or operation decisions.

One example of Big Data analytics based maintenance is conducted by Dutch railways on Axle Box Acceleration (ABA). With one Terabyte of track degradation data, a self-learning and adapted mechanisms is performed [135]. Thaduri *et al.* [134] introduce a hybrid modelling approach to provide accurate diagnosis asset condition so as to determine the remaining useful life of asset. The proposed method gives an insight in providing maintenance decision making for end users. Based on semantic data models, a railway asset monitoring system is implemented in [136] and prove to be more capable for data integration, extensibility, and compatibility compared with traditional approaches. Using data collected by multiple inspection vehicles in 330,000 km of railroad track, Zarembski [137] introduce the procedure of data collection, storage and plan the rail track maintenance with Big Data analytics so as to optimize its capital infrastructure and keep costs under control. Using the data from smart phones and GPS co-ordinates, Network Rail in UK successfully improve the track defect position from 1 mile to 5 meters, which significantly reduce the time to fix the rail track [138]. Using huge amount of historical system state data, in combination with train type data, maintenance action data, inspection schedule data, and system failure data, Li *et al.* [139] explore several machine learning based analytical methods to automatically learn regulations and construct failure estimation models. The models can use the real-time data to estimate if the current conditions will lead to system failure. A bilevel feature extraction-based text mining method is proposed in [140], where features extracted at semantic and syntax levels are used. The proposed method significantly improves the fault diagnosis precision for all fault classes.

VI. Big Data Platforms in ITS

Big Data analytics in ITS have been evolving with the help from advanced Big Data platforms. The Big Data platform leverages distributed file system and parallel computing capability to enable fast data process. It is capable of making sense of Big Data as well as supporting large-scale system optimization.

Apache Hadoop is the most popular open source software framework for distributed process and storage of large amount of data sets. Hadoop is a universal Big Data process platform, where various kinds data process or data analytical operations can be carried out. The distributed process capability makes Hadoop well-suited for analyzing the data in ITS, such as smart card data, diverse sensors, social media, GPS data etc. Apache Spark is the latest open-source platform for large amount of data sets processing that peculiarly adapts to machine learning tasks [141]. Spark adopts the same distributed storage technology as Hadoop, and it allows user programs to load data into a clusters memory and query it repeatedly. Spark is well-suited to machine learning approaches. The Big Data analytics approaches we introduced in the last subsection are machine learning based, and they can definitely be performed in both Hadoop and Spark platforms. The Big Data platform with the data analytics approaches running on it, will play a huge role in Big Data analytics in ITS.

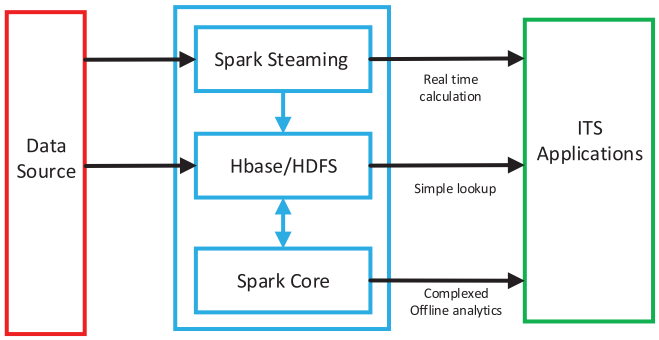


Fig. 6. A typical framework of using Apache Spark platform in ITS.

A typical framework of using Spark platform in ITS is shown in Fig. 6. Data from different sources are collected by HBase (Hadoop Database) APIs, and they are sent to the data center. Spark Streaming processes the data in real time. Some real time tasks, such as vehicle speed detection, vehicle identification, real time warning etc. can be implemented. HBase is a distributed open source database. It will perform high level feature extraction, and create index for massive data sets, so as to improve the effectiveness and efficiency of data retrieval. Spark Core is the foundation of spark system, and it can carry out off-line tasks with distributed computation capability. Critical tasks such as traffic management and control, accident analysis etc. can be conducted under Spark Core engine.

Different from the general-purpose Big Data platform, in transportation area, several platforms have been proposed to process the transportation data.

Mian *et al.* [142] propose a platform with multiple engines to support various types of analytic for traffic data. Zareian *et al.* [143] propose a monitor system named K-Feed for performance analysis of applications deployed on cloud. Shtern *et al.* [144] propose a conceptual architecture for a data engine, Godzilla, to perform real-time traffic data process and support analytical operation over transportation data. They design a multi-cluster approach to handle large amount of growing data under various kind of workloads and different number of users. Khazaei *et al.* [145], propose a platform to perform analytical operation on urban transportation data. The platform can be used by traffic-related software developers or directly by traffic engineers and researchers to gain insights of traffic patterns. Chaolong *et al.* [146] study the development trend of the virtual data center and its technical advantages, proposes a scheme of virtual system of smart transportation data center based on VMware vSphere. A Big Data simulation platform is proposed in [147] for Greater Toronto Area. The platform enables Big Data transportation applications to run in real time.

Real time data streaming process function is a necessary part of Big Data process platform in ITS. Because there are many real-time applications such as traffic monitor and control, and public transportation schedule. Based on the tradition Big Data process system, substantial real-time data streaming systems have been proposed in ITS. Guerreiro *et al.* [148] propose an ETL (extract, transform and load) architecture for intelligent transportation systems, addressing an application scenario on dynamic toll charging for highways. The proposed architecture is capable of handling real-time and historical data using Big Data technologies such as Spark on Hadoop and MongoDB. A data stream processing platform is proposed in [149], which supports a mechanism for sharing multiparty data sources, software components, and even intermediate results. They give an example of using this platform to conduct traffic management. A comprehensive and flexible architecture based on distributed computing platform for real-time traffic control is proposed in [150]. They have partly realized the architecture in a prototype platform that employs Kafka, a state-of-the-art Big Data tool for building data pipelines and stream processing.

Data injection is another critical part of Big Data process system. It is used to transfer data between Big Data process system and relational databases or mainframes. As a popular data injection system, Apache Sqoop has been widely adopted in ITS. For example, Sqoop is used with Hadoop in traffic management system in [151]. It has also been deployed to process vehicle diagnostics data and deliver useful outcomes that can be used by actors in automotive ecosystems [152]. In [153], Apache Sqoop is used to ingest ITS relational data. Apache Flume is another popular data injection system that processes unstructured data, and it has been adopted to process log data in ITS [154].

VII. Open Challenges

Although Big Data analytics has made great achievements in ITS, there are still substantial open challenges have not been fully studied. They need to be tackled in future works. This section introduce the main open challenges of using Big Data analytics in ITS as follows.

- Data collection: Due to the frequent movement of vehicles and pedestrians, data collected in transportation may be inaccurate, incomplete or unreliable in particular locations or at certain times. For instance, not all vehicles are embedded with the techniques needed to provide real-time location data, and road traffic data from road sensors can be missing. One possible way to tackle the challenger is to invest new data collection technologies and improve the data collection capability. With the development IoT, new sensor techniques are invented annually, which can help improve data collection and data quality. In addition, the adoption of data capturing automation to minimize manual data entry is also essential to data quality improvement.
- Data privacy: In the era of Big Data, the most challenging and concerned problem is privacy [155]. Personal privacy may be leaked during data transmission, storage and usage [156]. Data collected from transportation systems used to be non-personal data, such as vehicle location, traffic flow data. However, privacy problems have been concerned since personal data collection by the public and private sectors grows over time. For example, the location of individuals and vehicles can be easily collected. If these data are not strictly protected, people who steal these data would harm the owner of the data. Therefore, privacy protection is an important thing for Big Data applications in ITS. To prevent unauthorized disclosure of the personal private information, governments should develop complete data privacy laws which include what data can be published, the scope of the data publishing and using, the basic principles of data distribution, data availability and other areas [157]. The transportation departments should strictly regulate the personal data definition, strengthen the management of data security certification, and use more advanced algorithms to improve the data security level.
- Data storage: Currently, the data volume has jumped from TB level to PB level, and the growth in data storage capacity is far behind the data growth. Especially in ITS, it will produce a variety of data from the various sensors every day. Traditional data storage infrastructure and database tools have been unable to cope with the increasingly large and complex mass data [158]. Therefore, designing the most reasonable data storage architecture has become a key challenge. The main public cloud storage providers, such as Google and Microsoft, continue to improve their services with integrated Big Data capabilities, and multi-cloud storage and hybrid storage are emerging as key areas for Big Data storage. Their compute bursting capabilities have advantages in many forms of compute-intensive analytics workloads. In addition, combining intelligence with storage is also a good solution. Enterprises are looking for smart management tools which can provide integrated analytics within storage. This enables them to conduct resource monitoring and make full use of storage infrastructure.
- Data processing: Timeliness is crucial to Big Data applications in ITS, these applications include traffic data preprocessing, traffic state recognition, real-time traffic control, dynamic route guidance and real-time bus scheduling. Traffic data which contain different formats from diverse sources, must be compared with the historical data, then processed within a short time [159]. The data processing system must be able to process more complicated and increasingly expanding data. How to guarantee the process timeliness with so large and fast data is a big challenge. Many general Big Data frameworks that handle real time data sources, such as Apache Storm, Apache Flink, Apache Samza, Apache Spark Streaming and Kafka Streams, have appeared recently. In addition, dedicated Big Data processing frameworks for ITS have also been developed, such as platform for real-time traffic control, and estimating the average speed and the congested sections of a highway. These processing framework provide good solutions to real time data processing.

- Data opening: To enable transportation service users and App developers to find and re-use data effectively, data need to be archived and made publicly accessible in good quality. Data quality refers to its accuracy, completeness, reliability, and consistency [160], [161]. Without good data quality, Big Data will be misleading to decision-making and even produce harmful results. However, opening up data with good quality might require time and money. There is a trade-off between opening up data quickly at low cost and making high quality data available at high costs, which makes opening up good quality data one more big challenge. Effective solutions include the adoption of automatic data capturing and/or utilization of artificial intelligence to verify the data. Additionally, the transportation departments should have a data management process enacted to ensure pristine and accurate data.

VIII. Conclusions

In this paper, we presented the development of Big Data and the relevant knowledge of ITS. The framework of conducting Big Data analytics in ITS was discussed. We summarized the data source and collection methods, data analytics methods and platforms, and Big Data analytics application categories in ITS. We presented several applications of Big Data analytics in ITS, including asset maintenance, road traffic flow prediction, road traffic accidents analysis, public transportation service planning, personal travel route planning and rail transportation management and control. Several open challenges of using Big Data analytics in ITS were discussed in this paper, including data collection, data privacy, data storage, data processing, and data opening. Big Data analytics will have profound impacts on the design of intelligent transportation system, and make it safer, more efficient and profitable.

References

[1] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016.

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[3] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[4] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.

[5] M. Mayilvaganan and M. Sabitha, "A cloud-based architecture for big-data analytics in smart grid: A proposal," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res. (ICCIC)*, Dec. 2013, pp. 1–4.

[6] L. Qi, "Research on intelligent transportation system technologies and applications," in *Proc. Workshop Power Electron. Intell. Transp. Syst.*, 2008, pp. 529–531.

[7] S.-H. An, B.-H. Lee, and D.-R. Shin, "A survey of intelligent transportation systems," in *Proc. Int. Conf. Comput. Intell.*, Jul. 2011, pp. 332–337.

[8] N.-E. El Faouzi, H. Leung, and A. Kurian, "Data fusion in intelligent transportation systems: Progress and challenges—A survey," *Inf. Fusion*, vol. 12, no. 1, pp. 4–10, 2011.

[9] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[10] Q. Shi and M. Abdel-Aty, "Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 380–394, Sep. 2015.

[11] N. Mohamed and J. Al-Jaroodi, "Real-time big data analytics: Applications and challenges," in *Proc. Int. Conf. High Perform. Comput. Simulation*, Jul. 2014, pp. 305–310.

[12] X. Lin, P. Wang, and B. Wu, "Log analysis in cloud computing environment with hadoop and spark," in *Proc. 5th IEEE Int. Conf. Broadband Netw. Multimedia Technol. (IC-BNMT)*, Nov. 2013, pp. 273–276.

[13] M. Zaharia *et al.*, "Fast and interactive analytics over Hadoop data with spark," *USENIX Login*, vol. 37, no. 4, pp. 45–51, 2012.

[14] D. Corrigan, P. Zikopoulos, K. Parasuraman, T. Deutsch, D. Deroos, and J. Giles, *Harness the Power of Big Data the IBM Big Data Platform*. 1st ed. New York, NY, USA: McGraw-Hill, Nov. 2012.

[15] L. Basche, "Says solving 'big data' challenge involves more than just managing volumes of data," Bus. Wire, San Francisco, CA, USA, Tech. Rep., Jun. 2011.

[16] M. Bagchi and P. R. White, "The potential of public transport smart card data," *Transp. Policy*, vol. 12, no. 5, pp. 464–474, 2005.

[17] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, 2011.

[18] Y. Liu, X. Weng, J. Wan, X. Yue, and H. Song, "Exploring data validity in transportation systems for smart cities," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 26–33, 2017.

[19] H. Nishiuchi, J. King, and T. Todoroki, "Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data," *Int. J. Intell. Transp. Syst. Res.*, vol. 11, no. 1, pp. 1–10, 2013.

[20] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile," *Transp. Res. C, Emerg. Technol.*, vol. 24, pp. 9–18, Oct. 2012.

[21] K. A. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transp. Res. Rec., J. Transp. Res. Board*, pp. 63–72, Dec. 2008.

[22] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile," *Transp. Res. C, Emerg. Technol.*, vol. 24, pp. 9–18, 2012.

[23] H. Gong, C. Chen, E. Bialostozky, and C. T. Lawson, "A GPS/GIS method for travel mode detection in New York City," *Comput., Environ. Urban Syst.*, vol. 36, no. 2, pp. 131–139, 2012.

[24] X. Wang, S. Zhao, and L. Dong, "Research and application of traffic visualization based on vehicle GPS big data," in *Proc. Int. Conf. Intell. Transp.*, 2016, pp. 293–302.

[25] C. Asensio, J. López, R. Pagán, I. Pavón, and M. Ausejo, "GPS-based speed collection method for road traffic noise mapping," *Transp. Res. D, Transp. Environ.*, vol. 14, no. 5, pp. 360–366, 2009.

[26] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobsond, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 4, pp. 568–583, Aug. 2010.

[27] K. G. Courage, M. Doctor, S. Maddula, and R. Surapaneni, "Video image detection for traffic surveillance and control," Transp. Res. Center, Univ. Florida, Gainesville, FL, USA, Tech. Rep. TD100:FL96-119, Mar. 1996.

[28] C. Grant, B. Gillis, and R. Guensler, "Collection of vehicle activity data by video detection for use in transportation planning," *J. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 343–361, 2000.

[29] M. Kadaieaswaran, V. Arunprasath, and M. Karthika, "Big data solution for improving traffic management system with video processing," *Int. J. Eng. Sci.*, vol. 7, no. 2, 2017.

[30] J. Lopes, J. Bento, E. Huang, C. Antoniou, and M. Ben-Akiva, "Traffic and mobility data collection for real-time applications," in *Proc. IEEE Intell. Transp. Syst. (ITSC)*, Sep. 2010, pp. 216–223.

[31] C. Antoniou, R. Balakrishna, and H. Koutsopoulos, "Emerging data collection technologies and their impact on traffic management applications," in *Proc. 10th Int. Conf. Appl. Adv. Technol. Transp.*, Athens, Greece, 2008.

[32] E. Huang, "Algorithmic and implementation aspects of on-line calibration of dynamic traffic assignment," Ph.D. dissertation, Dept. Civil, Environ. Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 2010.

[33] E. Uhlemann, "Autonomous vehicles are connecting… [connected vehicles]," *IEEE Veh. Technol. Mag.*, vol. 10, no. 2, pp. 22–25, Jun. 2015.

[34] C. Chen, T. H. Luan, X. Guan, N. Lu, and Y. Liu. (2017). "Connected vehicular transportation: Data analytics and traffic-dependent networking." [Online]. Available: https://arxiv.org/abs/1704.08125

[35] X. Wang, C. Wang, J. Zhang, M. Zhou, and C. Jiang, "Improved rule installation for real-time query service in software-defined Internet of vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 2, pp. 225–235, Feb. 2017.

[36] J. Hu, L. Kong, W. Shu, and M.-Y. Wu, "Scheduling of connected autonomous vehicles on highway lanes," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 5556–5561.

[37] R. King, "Traffic management in a connected or autonomous vehicle environment," in *Proc. Auto. Passenger Veh.*, May 2015, pp. 1–20.

[38] P. Bedi and V. Jindal, "Use of big data technology in vehicular ad-hoc networks," in *Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI)*, Sep. 2014, pp. 1677–1683.

[39] J. Contreras-Castillo, S. Zeadally, and J. A. G. Ibañez, "Solving vehicular ad hoc network challenges with big data solutions," *IET Netw.*, vol. 5, no. 4, pp. 81–84, Jul. 2016.

[40] M. Tan, I. W. Tsang, and L. Wang, "Towards ultrahigh dimensional feature selection for big data," *J. Mach. Learn. Res.*, vol. 15, pp. 1371–1429, Apr. 2014.

[41] A. Mahajan and A. Kaur, "Predictive urban traffic flow model using vehicular big data," *Indian J. Sci. Technol.*, vol. 9, no. 42, pp. 1–8, 2016.

[42] H. A. Najada and I. Mahgoub, "Anticipation and alert system of congestion and accidents in vanet using big data analysis for intelligent transportation systems," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2016, pp. 1–8.

[43] L. Gong, X. Liu, L. Wu, and Y. Liu, "Inferring trip purposes and uncovering travel patterns from taxi trajectory data," *Cartography Geograph. Inf. Sci.*, vol. 43, no. 2, pp. 103–114, 2016.

[44] C. Kang, Y. Liu, X. Ma, and L. Wu, "Towards estimating urban population distributions from mobile call data," *J. Urban Technol.*, vol. 19, no. 4, pp. 3–21, 2012.

[45] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transp. Res. C, Emerg. Technol.*, vol. 68, pp. 285–299, Jul. 2016.

[46] J. Zeyu, Y. Shuiping, Z. Mingduan, C. Yongqiang, and L. Yi, "Model study for intelligent transportation system with big data," *Proc. Comput. Sci.*, vol. 107, pp. 418–426, 2017.

[47] S. Liu, H. Cao, L. Li, and M. C. Zhou, "Predicting stay time of mobile users with contextual information," *IEEE Trans. Automat. Sci. Eng.*, vol. 10, no. 4, pp. 1026–1036, Oct. 2013.

[48] A. Gal-Tzur, S. M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, and I. Shoor, "The potential of social media in delivering transport policy goals," *Transp. Policy*, vol. 32, pp. 115–123, Mar. 2014.

[49] F. Alesiani, K. Gkiotsalitis, and R. Baldessari, "A probabilistic activity model for predicting the mobility patterns of homogeneous social groups based on social network data," in *Proc. 93rd Annu. Meeting Transp. Res. Board*, 2014.

[50] Y. Chen, A. Frei, and H. Mahmassani, "From personal attitudes to public opinion: Information diffusion in social networks toward sustainable transportation," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2, pp. 28–37, Nov. 2014.

[51] B. Pender, G. Currie, A. Delbosc, and N. Shiwakoti, "Social media use during unplanned transit network disruptions: A review of literature," *Transp. Rev.*, vol. 34, no. 4, pp. 501–521, 2014.

[52] X. Zheng *et al.*, "Big data for social transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 620–630, Mar. 2016.

[53] C. D. Cottrill and S. Derrible, "Leveraging big data for the development of transport sustainability indicators," *J. Urban Technol.*, vol. 22, no. 1, pp. 45–64, 2015.

[54] G. R. Grob, "Future transportation with smart grids & sustainable energy," in *Proc. 6th Int. Multi-Conf. Syst., Signals Devices (SSD)*, 2009, pp. 1–5.

[55] L. Zhu, F. R. Yu, B. Ning, and T. Tang, "Cross-layer handoff design in MIMO-enabled WLANs for communication-based train control (CBTC) systems," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 719–728, May 2012.

[56] L. Zhu, F. R. Yu, B. Ning, and T. Tang, "Handoff performance improvements in MIMO-enabled communication-based train control systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 582–593, Jun. 2012.

[57] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine Learning: An Artificial Intelligence Approach*. New York, NY, USA: Springer, 2013.

[58] G. A. Seber and A. J. Lee, *Linear Regression Analysis*, vol. 936. Hoboken, NJ, USA: Wiley, 2012.

[59] H. Sun, H. Liu, H. Xiao, R. He, and B. Ran, "Use of local linear regression model for short-term traffic forecasting," *Transp. Res. Rec., J. Transp. Res. Board*, pp. 143–150, Jan. 2003.

[60] Z. Shan, D. Zhao, and Y. Xia, "Urban road traffic speed estimation for missing probe vehicle data based on multiple linear regression model," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 118–123.

[61] N. Zenina and A. Borisov, "Regression analysis for transport trip generation evaluation," *Inf. Technol. Manage. Sci.*, vol. 16, no. 1, pp. 89–94, 2013.

[62] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[63] H. J. Payne and S. Tignor, "Freeway incident detection algorithms based on decision trees with states," *Transp. Res. Rec.*, vol. 682, pp. 30–37, Jan. 1978.

[64] J. AbellánN, G. LóPez, and J. De OñA, "Analysis of traffic accident severity using decision rules via decision trees," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 6047–6054, 2013.

[65] C. Xie, J. Lu, and E. Parkany, "Work travel mode choice modeling with data mining: Decision trees and neural networks," *Transp. Res. Rec., J. Transp. Res. Board*, pp. 50–61, Jan. 2003.

[66] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach," *Transp. Res. C, Emerg. Technol.*, vol. 13, no. 3, pp. 211–234, 2005.

[67] J. Van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data," *Transp. Res. C, Emerg. Technol.*, vol. 13, nos. 5–6, pp. 347–369, Oct./Dec. 2005.

[68] X. Jin, R. L. Cheu, and D. Srinivasan, "Development and adaptation of constructive probabilistic neural network in freeway incident detection," *Transp. Res. C, Emerg. Technol.*, vol. 10, no. 2, pp. 121–147, 2002.

[69] X. Zhu, J. Guo, and W. Huang, "Short-term forecasting of remaining parking spaces in parking guidance systems," in *Proc. 95th Annu. Meeting Transp. Res. Board*, 2016.

[70] L. Vanajakshi and L. R. Rilett, "Support vector machine technique for the short term prediction of travel time," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2007, pp. 600–605.

[71] Y. Bin, Y. Zhongzhen, and Y. Baozhen, "Bus arrival time prediction using support vector machines," *J. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 151–158, 2006.

[72] J. Xiao and Y. Liu, "Traffic incident detection using multiple-kernel support vector machine," *Transp. Res. Rec., J. Transp. Res. Board*, pp. 44–52, Dec. 2012.

[73] Y. Meng and X. Liu, "Application of K-means algorithm based on ant clustering algorithm in macroscopic planning of highway transportation hub," in *Proc. 1st IEEE Int. Symp. Inf. Technol. Appl. Edu. (ISITAE)*, Nov. 2007, pp. 483–488.

[74] R. P. D. Nath, H.-J. Lee, N. K. Chowdhury, and J.-W. Chang, "Modified k-means clustering for travel time prediction based on historical traffic data," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, 2010, pp. 511–521.

[75] B. Abdulhai, R. Pringle, and G. J. Karakoulas, "Reinforcement learning for true adaptive traffic signal control," *J. Transp. Eng.*, vol. 129, no. 3, pp. 278–285, 2003.

[76] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, "Reinforcement learning-based multi-agent system for network traffic signal control," *IET Intell. Transp. Syst.*, vol. 4, no. 2, pp. 128–135, 2010.

[77] A. L. C. Bazzan, "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control," *Auto. Agents Multi-Agent Syst.*, vol. 18, no. 3, pp. 342–375, Jun. 2009.

[78] L. Li, Y. Lv, and F.-Y. Wang, "Traffic signal timing via deep reinforcement learning," *IEEE/CAA J. Automat. Sinica*, vol. 3, no. 3, pp. 247–254, Apr. 2016.

[79] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PLoS ONE*, vol. 10, no. 3, p. e0119044, 2015.

[80] H. Hu, B. Tang, X. Gong, W. Wei, and H. Wang, "Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2106–2116, Aug. 2017.

[81] T. Chen, "Going deeper with convolutional neural network for intelligent transportation," Ph.D. dissertation, Dept. Elect. Comput. Eng., Worcester Polytech. Inst., Worcester, MA, USA, 2015.

[82] Y. Duan, Y. Lv, W. Kang, and Y. Zhao, "A deep learning based approach for traffic data imputation," in *Proc. IEEE 17th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 912–917.

[83] N. Polson and V. Sokolov. (2016). "Deep learning for short-term traffic flow prediction." [Online]. Available: https://arxiv.org/abs/1604.04527

[84] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2191–2201, Oct. 2014.

[85] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[86] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.

[87] J. Zhai, Y. Cao, and Y. Chen, "Semantic information retrieval based on fuzzy ontology for intelligent transportation systems," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2008, pp. 2321–2326.

[88] S. Fernandez and T. Ito, "Driver behavior model based on ontology for intelligent transportation systems," in *Proc. IEEE 8th Int. Conf. Service-Oriented Comput. Appl. (SOCA)*, Oct. 2015, pp. 227–231.

[89] S. Fernandez and T. Ito, "Using SSN ontology for automatic traffic light settings on inteligent transportation systems," in *Proc. IEEE Int. Conf. Agents (ICA)*, Sep. 2016, pp. 106–107.

[90] D. Gregor *et al.*, "A methodology for structured ontology construction applied to intelligent transportation systems," *Comput. Standards Interfaces*, vol. 47, pp. 108–119, Aug. 2016.

[91] L. Zhao, R. Ichise, S. Mita, and Y. Sasaki, "Ontologies for advanced driver assistance systems," *J. Jpn. Soc. Artif. Intell.*, 2015, accessed: Aug. 12, 2016. [Online]. Available: http://www.ei.sanken.osaka-u.ac.jp/sigswo/papers/SIG-SWO-035/SIG-SWO-035-03.pdf

[92] D. Chen, F. Asplund, K. Östberg, E. Brezhniev, and V. Kharchenko, "Towards an ontology-based approach to safety management in cooperative intelligent transportation systems," in *Proc. 10th Int. Conf. Depend. Complex Syst. Depcos-Relcomex*, 2015, pp. 107–115.

[93] W.-D. Yang and T. Wang, "The fusion model of intelligent transportation systems based on the urban traffic ontology," *Phys. Proc.*, vol. 25, no. 49, pp. 917–923, 2012.

[94] T. Toroyan, "Global status report on road safety," *Injury Prevention*, vol. 15, no. 4, p. 286, 2009.

[95] T. F. Golob and W. W. Recker, "Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions," *J. Transp. Eng.*, vol. 129, no. 4, pp. 342–353, 2003.

[96] G. Xiong, F. Zhu, H. Fan, X. Dong, W. Kang, and T. Teng, "Novel ITS based on space-air-ground collected big-data," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Oct. 2014, pp. 1509–1514.

[97] J. Lee and F. Mannering, "Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis," *Accident Anal. Prevention*, vol. 34, no. 2, pp. 149–161, 2002.

[98] M. G. Karlaftis and I. Golias, "Effects of road geometry and traffic volumes on rural roadway accident rates," *Accident Anal. Prevention*, vol. 34, no. 3, pp. 357–365, 2002.

[99] L.-Y. Chang and W.-C. Chen, "Data mining of tree-based models to analyze freeway accident frequency," *J. Safety Res.*, vol. 36, no. 4, pp. 365–375, 2005.

[100] M. Bédard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities," *Accident Anal. Prevention*, vol. 34, no. 6, pp. 717–727, 2002.

[101] R. Li, C. Jiang, F. Zhu, and X. Chen, "Traffic flow data forecasting based on interval type-2 fuzzy sets theory," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 2, pp. 141–148, Apr. 2016.

[102] D. Chen, "Research on traffic flow prediction in the big data environment based on the improved RBF neural network," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2000–2008, Aug. 2017.

[103] S. Jeon and B. Hong, "Monte Carlo simulation-based traffic speed forecasting using historical big data," *Future Generat. Comput. Syst.*, vol. 65, pp. 182–195, Dec. 2016.

[104] X.-L. Liu, P. Jia, S.-H. Wu, and B. Yu, "Short-term traffic flow forecasting based on multi-dimensional parameters," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 11, no. 4, pp. 140–146, 2011.

[105] H.-H. Dong, X.-L. Sun, L.-M. Jia, H.-J. Li, and Y. Qin, "Traffic condition estimation with pre-selection space time model," *J. Central South Univ.*, vol. 19, no. 1, pp. 206–212, 2012.

[106] M. Canaud, L. Mihaylova, J. Sau, and N.-E. El Faouzi, "Probabilty hypothesis density filtering for real-time traffic state estimation and prediction," *Netw. Heterogeneous Media*, vol. 8, no. 3, pp. 825–842, 2013.

[107] T. L. Pan, A. Sumalee, R. X. Zhong, and N. Indra-Payoong, "Short-term traffic state prediction based on temporal–spatial correlation," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1242–1254, Sep. 2013.

[108] C. Antoniou, H. N. Koutsopoulos, and G. Yannis, "Dynamic data-driven local traffic state estimation and prediction," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 89–107, Sep. 2013.

[109] B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate short-term traffic flow forecasting using time-series analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 2, pp. 246–254, Jun. 2009.

[110] J. Xu, D. Deng, U. Demiryurek, C. Shahabi, and M. van der Schaar, "Mining the situation: Spatiotemporal traffic prediction with big data," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 702–715, Jun. 2015.

[111] H.-P. Lu, Z.-Y. Sun, and W.-C. Qu, "Big data and its applications in urban intelligent transportation system," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 15, no. 5, pp. 45–52, 2015.

[112] C.-C. Lu, X. Zhou, and K. Zhang, "Dynamic origin–destination demand flow estimation under congested traffic conditions," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 16–37, Sep. 2013.

[113] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 240–250, Sep. 2015.

[114] J. B. Gordon, "Intermodal passenger flows London's public transport network: Automated inference full passenger journeys using fare-transaction vehicle-location data," Ph.D. dissertation, Dept. Urban Studies Planning, Dept. Civil Environ. Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 2012.

[115] S. Tao, "Investigating the travel behaviour dynamics of bus rapid transit passengers," Ph.D. dissertation, School Geograp., Planning Environ. Manage., Univ. Queensland, Brisbane, Qld, Australia, 2015.

[116] I. Gokasar and K. Simsek, "Using 'Big data' for analysis and improvement of public transportation systems in Istanbul," Tech. Rep., 2014.

[117] J. Chan *et al.*, "Rail transit OD matrix estimation and journey time reliability metrics using automated fare data," PhD thesis, Dept. Civil Environ. Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 2007.

[118] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, "The path most traveled: Travel demand estimation using big data resources," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 162–177, Sep. 2015.

[119] B. Ferris, K. Watkins, and A. Borning, "OneBusAway: Results from providing real-time arrival information for public transit," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2010, pp. 1807–1816.

[120] [Online]. Available: http://inrix.com/mobile-apps/

[121] [Online]. Available: https://www.waze.com/

[122] [Online]. Available: http://moovitapp.com/

[123] [Online]. Available: http://gaode.com/

[124] "Opening up to open data, " in *Proc. Int. Assoc. Public Transp.*, 2014.

[125] B. Schultz, Ed., "Operational analytics keeps bay area trains on track," *All Analytics*, May 2012. [Online]. Available: http://www.allanalytics.com/author.asp?section_id=1411&doc_id=244062

[126] M. Faizrahnemoon, A. Schlote, L. Maggi, E. Crisostomi, and R. Shorten, "A big-data model for multi-modal public transportation with application to macroscopic control and optimisation," *Int. J. Control*, vol. 88, no. 11, pp. 2354–2368, 2015.

[127] N. Van Oort, "Big data opportunities in public transport: Enhancing public transport by ITCS," in *Proc. IT-TRANS*, Karlsruhe, Germany, Feb. 2014.

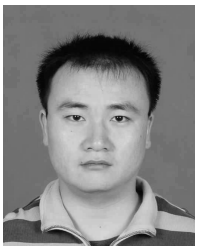
[128] Z. Jiang, C.-H. Hsu, D. Zhang, and X. Zou, "Evaluating rail transit timetable using big passengers' data," *J. Comput. Syst. Sci.*, vol. 82, no. 1, pp. 144–155, 2016.

[129] J. Yin, D. Chen, and Y. Li, "Smart train operation algorithms based on expert knowledge and ensemble CART for the electric locomotive," *Knowl.-Based Syst.*, vol. 92, pp. 78–91, Jan. 2016.

[130] D. Chen, T. Tang, C. Gao, and R. Mu, "Research on the error estimation models and online learning algorithms for train station parking in urban rail transit," *China Railway Sci.*, vol. 31, no. 6, pp. 122–127, 2010.

[131] J. Zhou, "Applications of machine learning methods in problem of precise train stopping," *Comput. Eng. Appl.*, vol. 46, no. 25, pp. 226–230, 2010.

[132] Z. Hou, Y. Wang, C. Yin, and T. Tong, "Terminal iterative learning control based station stop control of a train," *Int. J. Control*, vol. 84, no. 7, pp. 1263–1274, 2011.
[133] D. Chen, R. Chen, T. Tang, and Y. Li, "Online learning algorithms for train automatic stop control using precise location data of balises," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1526–1535, Sep. 2013.
[134] A. Thaduri, D. Galar, and U. Kumar, "Railway assets: A potential domain for big data analytics," *Proc. Comput. Sci.*, vol. 53, no. 1, pp. 457–467, 2015.
[135] A. Núñez, J. Hendriks, Z. Li, B. De Schutter, and R. Dollevoet, "Facilitating maintenance decisions on the dutch railways using big data: The aba case study," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2014, pp. 48–53.
[136] J. Tutcher, "Ontology-driven data integration for railway asset monitoring applications," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2014, pp. 85–95.
[137] A. M. Zarembski, "Some examples of big data in railroad engineering," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2014, pp. 96–102.
[138] *Asset Management Services*, Network Rail, London, U.K., 2013.
[139] H. Li *et al.*, "Improving rail network velocity: A machine learning approach to predictive maintenance," *Transp. Res. C, Emerg. Technol.*, vol. 45, pp. 17–26, Aug. 2014.
[140] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, "Bilevel feature extraction-based text mining for fault diagnosis of railway systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 49–58, Jan. 2017.
[141] X. Meng *et al.*, "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, no. 34, pp. 1–7, 2016.
[142] R. Mian, H. Ghanbari, S. Zareian, M. Shtern, and M. Litoiu, "A data platform for the highway traffic data," in *Proc. MESOCA*, 2014, pp. 47–52.
[143] S. Zareian, R. Veleda, M. Litoiu, M. Shtern, H. Ghanbari, and M. Garg, "K-feed—A data-oriented approach to application performance management in cloud," in *Proc. IEEE 8th Int. Conf. Cloud Comput.*, Jun. 2015, pp. 1045–1048.
[144] M. Shtern, R. Mian, M. Litoiu, S. Zareian, H. Abdelgawad, and A. Tizghadam, "Towards a multi-cluster analytical engine for transportation data," in *Proc. Int. Conf. Cloud Auton. Comput. (ICCAC)*, 2014, pp. 249–257.
[145] H. Khazaei, S. Zareian, R. Veleda, and M. Litoiu, "Sipresk: A big data analytic platform for smart transportation," in *Proc. EAI Int. Conf. Big Data Anal. Smart Cities*, 2015, pp. 419–430.
[146] J. Chaolong, H. Wang, and L. Wei, "Study of smart transportation data center virtualization based on vmware vsphere and parallel continuous query algorithm over massive data streams," *Proc. Eng.*, vol. 137, no. 6, pp. 719–728, 2016.
[147] I. R. Kamel, H. Abdelgawad, and B. Abdulhai, "Transportation big data simulation platform for the greater toronto area (GTA)," in *Smart City 360°*. New York, NY, USA: Springer, 2016, pp. 443–454.
[148] G. Guerreiro, P. Figueiras, R. Silva, R. Costa, and R. Jardim-Goncalves, "An architecture for big data processing on intelligent transportation systems: An application scenario on highway traffic flows," in *Proc. IEEE 8th Int. Conf. Intell. Syst. (IS)*, Sep. 2016, pp. 65–72.
[149] E. Bouillet *et al.*, "Data stream processing infrastructure for intelligent transport systems," in *Proc. IEEE 66th Veh. Technol. Conf. (VTC-Fall)*, Oct. 2007, pp. 1421–1425.
[150] S. Amini, I. Gerostathopoulos, and C. Prehofer, "Big data analytics architecture for real-time traffic control," in *Proc. 5th IEEE Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, Jun. 2017, pp. 710–715.
[151] G. Zeng, "Application of big data in intelligent traffic system," *IOSR J. Comput. Eng.*, vol. 17, no. 1, pp. 1–4, 2015.
[152] M. Tahmassebpour and A. M. Otaghvari, "Increase efficiency big data in intelligent transportation system with using IoT integration cloud," *J. Fundam. Appl. Sci.*, vol. 8, no. 3S, pp. 2443–2461, 2016.
[153] M. Chowdhury, A. Apon, and K. Dey, *Data Analytics for Intelligent Transportation Systems*. Amsterdam, The Netherlands: Elsevier, 2017.
[154] Z. Ji, I. Ganchev, M. O'Droma, L. Zhao, and X. Zhang, "A cloud-based car parking middleware for IoT-based smart cities: Design and implementation," *Sensors*, vol. 14, no. 12, pp. 22372–22393, 2014.
[155] M. Smith, C. Szongott, B. Henne, and G. von Voigt, "Big data privacy issues in public social media," in *Proc. IEEE Int. Conf. Digit. Ecosyst. Technol.*, Jun. 2012, pp. 1–6.
[156] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, "Enabling public auditability and data dynamics for storage security in cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 5, pp. 847–859, May 2011.
[157] O. Tene and J. Polonetsky, "Big data for all: Privacy and user control in the age of analytics," *Northwestern J. Technol. Intell. Property*, vol. 11, no. 5, 2012, Art. no. 1.
[158] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Sci.*, vol. 332, no. 6025, pp. 60–65, 2011.
[159] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," *J. Parallel Distrib. Comput.*, vols. 79–80, pp. 3–15, May 2013.
[160] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS J. Photogram. Remote Sens.*, vol. 115, pp. 134–142, May 2016.
[161] J. Li and X. Liu, "An important aspect of big data: Data usability," *J. Comput. Res. Develop.*, vol. 50, no. 6, pp. 1147–1162, 2013.



Li Zhu received the Ph.D. degree in traffic control and information engineering from Beijing Jiaotong University, Beijing, China, in 2012. He is currently a Faculty Member at Beijing Jiaotong University and a Visiting Scholar at Carleton University, Ottawa, ON, Canada, and The University of British Columbia, Vancouver, BC, Canada. His research interests include intelligent transportation systems, train-ground communication technology in communication base train ground communication systems, and cross layer design in train-ground communication systems.

Fei Richard Yu (S'00–M'04–SM'08–F'18) received the Ph.D. degree in electrical engineering from The University of British Columbia in 2003. From 2002 to 2006, he was with Ericsson, Lund, Sweden, and a start-up in California, USA. He joined Carleton University in 2007, where he is currently a Professor. His research interests include wireless cyber-physical systems, connected/autonomous vehicles, security, distributed ledger technology, and deep learning. He received the IEEE Outstanding Service Award in 2016, the IEEE Outstanding Leadership Award in 2013, the Carleton Research Achievement Award in 2012, the Ontario Early Researcher Award (formerly Premiers Research Excellence Award) in 2011, the Excellent Contribution Award at IEEE/IFIP TrustCom 2010, the Leadership Opportunity Fund Award from Canada Foundation of Innovation in 2009, and the Best Paper Awards at IEEE VTC 2017 Spring, ICC 2014, Globecom 2012, IEEE/IFIP TrustCom 2009, and International Conference on Networking 2005.

Dr. Yu is a registered Professional Engineer in ON, Canada, and a fellow of the Institution of Engineering and Technology. He has served as the Technical Program Committee Co-Chair of numerous conferences. He serves on the Editorial Boards of several journals, including the Co-Editor-in-Chief for *Ad Hoc and Sensor Wireless Networks*, a Lead Series Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He is a Distinguished Lecturer, the Vice President (Membership), and an Elected Member of the Board of Governors of the IEEE Vehicular Technology Society.

Yige Wang received the bachelor's degree in electrical engineering from Shandong University of Technology, Shandong, China, in 2015. He is currently pursuing the degree with Beijing Jiaotong University. His research interests include train-ground communication technology in communication base train ground communication systems and intelligent transportation systems.

Bin Ning (F'14) received the master's and Ph.D. degrees in engineering from Northern Jiaotong University in 1987 and 2005, respectively. He was a Visiting Scholar in electronics and electrical power engineering with Brunel University, London, U.K. From 2002 to 2003, he was with UC Berkeley as a Senior Visiting Scholar. He is currently a Professor with Beijing Jiaotong university. His research mainly focus on high speed train control system and railway transportation train control system, including main locomotive signal, communication based train control system, intelligent transportation, fault-tolerant design of signal system, fault diagnosis, system reliability, and security design. He is responsible for several key scientific and technical projects, and made great research achievement.

Dr. Ning is a member of the China Overseas Returned Scholars Association and a member of the Editorial Board of the Journal of Railways in China. He is a fellow of the Association of International Railway Signaling Engineers, The Institute of Engineering and Technology, and the China Railway Society. He is the Deputy Director of the China Traffic System Engineering Society and the Beijing Railway Society. He is the Chair of Technical Committee on Railroad Systems and Applications of the IEEE Intelligent Transportation Systems Society. He was an Associate Editor of IEEE TRANSACTION ON INTELLIGENT TRANSPORTATION SYSTEMS (2010–2012) and *Acta Automatica Sinica* (2011–2012).

Tao Tang received the Ph.D. degree in engineering from Chinese Academy of Science in 1991. He is currently a Professor with Beijing Jiaotong University and an Associate Director of the Rail Traffic Control and Safety State Key Laboratory. His research interests include communication based train control, high speed train control system, and intelligent transportation system.

Dr. Tang is a member of experts Group of High Technology Research and Development Program of China (863 Program) and the Leader in the Field of Modern Transportation Technology Experts Group. He is also a Specialist of National Development and Reform Commission and Beijing Urban Traffic Construction Committee.