Big Data Analytics for Large Scale Wireless Networks: Challenges and Opportunities

HONG-NING DAI*, Macau University of Science and Technology
RAYMOND CHI-WING WONG, the Hong Kong University of Science and Technology (HKUST)
HAO WANG, Norwegian University of Science and Technology
ZIBIN ZHENG, Sun Yat-sen University
ATHANASIOS V. VASILAKOS, Lulea University of Technology

The wide proliferation of various wireless communication systems and wireless devices has led to the arrival of big data era in large scale wireless networks. Big data of large scale wireless networks has the key features of wide variety, high volume, real-time velocity and huge value leading to the unique research challenges that are different from existing computing systems. In this paper, we present a survey of the state-of-art big data analytics (BDA) approaches for large scale wireless networks. In particular, we categorize the life cycle of BDA into four consecutive stages: Data Acquisition, Data Preprocessing, Data Storage and Data Analytics. We then present a detailed survey of the technical solutions to the challenges in BDA for large scale wireless networks according to each stage in the life cycle of BDA. Moreover, we discuss the open research issues and outline the future directions in this promising area.

Additional Key Words and Phrases: Big Data; Data mining; Wireless Networks

ACM Reference format:
Hong-Ning Dai, Raymond Chi-Wing Wong, Hao Wang, Zibin Zheng, and Athanasios V. Vasilakos. 2019. Big Data Analytics for Large Scale Wireless Networks: Challenges and Opportunities. 1, 1, Article 1 (May 2019), 46 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, we have seen the proliferation of wireless communication technologies, which are widely used today across the globe to fulfill the communication needs from the extremely large number of end users. The interconnection of various wireless communication systems together forms a *large scale* wireless networks, where “large scale” means the high density of network stations (or nodes) and the large coverage area. Meanwhile, there is a surge of big volume of mobile data traffic generated from wireless networks that consist of a wide diversity of wireless devices, such as smart-phones, mobile tablets, laptops, RFID tags, sensors, smart meters and smart appliances. It is predicted that in a report of [cis 2017] there is a growth of the mobile data traffic from 10 EB/month (1EB = 1×10^{18} bytes) in 2017 to 49 EB/month in 2021, representing that we are entering a “*big data era*” [Cui et al. 2016].

*The corresponding author

Essentially, big data has the following salient features called “4Vs” differentiating it from other concepts, such as “very large data”, “large volume data” and “massive data” [Zikopoulos and Eaton 2011]:

(1) *Volume.* The quantity of generated and stored data (usually refer to the data volume from Terabytes to Petabytes);
(2) *Variety.* The type and nature of the data (structured, semi-structured, unstructured, text and multimedia);
(3) *Velocity.* The speed at which the data is generated and processed to meet the demands (e.g., real time).
(4) *Value.* The analytical results based on big data can bring huge both business value and social value.

Although there are other two ’V’s, i.e., “Variability” and “Veracity” [Hilbert 2016], we mainly use the above “4Vs” to describe big data generated from wireless networks. Since there are various types of large scale wireless networks, we only enumerate several exemplary networks including mobile communication networks, vehicular networks, mobile social networks, and Internet of Things (IoT). The wireless devices include not only various wired interfaces and wireless interfaces but also sensors consisting of temperature sensor, light sensor, acoustic sensor, vibration sensor, chemical sensor, accelerator and RFID tags [Wang and Liu 2011], which can generate high volume data in real-time fashion. In summary, big data generated from large scale wireless networks is often featured with wide variety, high volume, real-time velocity and huge value.

The growth of big data in large scale wireless networks brings not only the *challenges* in designing scalable wireless networks but also the *value*, which is beneficial to many areas, such as network operation, network management, network security, network optimization, intelligent traffic system, logistic management and social behavior study. It requires *big data analytics* (BDA) dedicated for large scale wireless networks to harness the benefits. Data generated from large scale wireless networks should be collected, filtered, stored and analyzed until the “value” is extracted.

1.1 Comparisons between this paper and existing surveys

There are a number of studies related to BDA including data mining, machine learning and distributed computing. For example, a survey on big data processing models from the *data mining perspective* is presented in [Wu et al. 2014]. The study of [Hu et al. 2014] presents a tutorial on BDA from the *scalability of BDA platforms*. Ref. [Chen et al. 2014] gives a survey on BDA from the aspect of *enabling technologies*. However, these previous surveys concentrated on BDA for general computing systems (e.g., data warehouse) and are not dedicated for large scale wireless networks, which have different features. For example, data generated in wireless networks is usually in real-time fashion and in heterogeneous types. Therefore, the conventional BDA approaches in general computing systems cannot be applicable to large scale wireless networks.

Recently, several surveys on BDA in wireless networks have been published. The paper [Alsheikh et al. 2014] presents a survey on using machine learning methods in wireless sensor networks (WSNs). The study of [Bi et al. 2015b] gives a short overview on big data analytics in wireless communication systems. In [Jiang et al. 2017], an overview on machine learning in next-generation wireless networks is presented. Ref. [Kibria et al. 2018] presents an overview on BDA and artificial intelligence in next-generation wireless networks. Qian et al. [Qian et al. 2017] survey a limited number of studies from data, transmission, network and application layers in wireless networks including communication networks and Internet of Things (IoT).

However, these studies are *too specific* to a certain type of wireless networks (either WSNs or mobile communication networks). Essentially, different wireless networks being featured of heterogeneous data types require different BDA approaches. For instance, vehicular networks have less computational and energy constraints than wireless sensor networks. In contrast to the above-mentioned surveys, we attempt to provide an in-depth survey on BDA for large scale

Table 1. Comparison of this article with existing surveys

Research issues	**References**	**This survey**
General BDA	[Wu et al. 2014] [Hu et al. 2014] [Chen et al. 2014]	✓
Wireless Sensor Networks (WSN)	[Alsheikh et al. 2014]	✓
Mobile Communication Networks	[Bi et al. 2015b] [Jiang et al. 2017] [Kibria et al. 2018] [Qian et al. 2017]	✓
Vehicular networks	✗	✓
Mobile Social Networks	✗	✓
Internet of Things (IoT)	[Qian et al. 2017]	✓

wireless networks with the inclusion of up-to-date studies. Our survey has a good horizontal and vertical coverage of research issues in BDA for large scale wireless networks. In the horizontal dimension, we mainly focus on four phases in the life cycle of BDA. In the vertical dimension, we consider four representative wireless networks (namely WSNs, Mobile Communication Networks, Vehicular networks and IoT). Table 1 highlights the differences between this survey and other existing surveys.

1.2 Contributions

We first conduct a comprehensive literature collection and analytics with consideration of timeliness, relevance and quality. The research methodology is presented in Section 2. We then introduce typical data sources of large scale wireless networks and discuss the necessities BDA for large scale wireless networks in Section 3.

The core contribution of this paper is to present the state-of-the-art of BDA in the context of large scale wireless networks in two dimensions: 1) life cycle of BDA and 2) different types of wireless networks. In order to give readers a clear roadmap about the BDA procedures, we introduce the life cycle of BDA. As shown in Fig. 1, we categorize the life cycle of BDA into four consecutive stages: Data Acquisition, Data Preprocessing, Data Storage and Data Analytics. Note that the data flow along the above four stages may not strictly go forward. In other words, there might be some backward links from one stage to the preceding stage. For example, the data flow in the data analytics stage may go back to the data storage stage since some statistic modeling algorithms require the comparison of the current data with the historical data. It is also worth mentioning that there are other taxonomies of the phases of BDA proposed for other computing systems [Casado and Younas 2015; Hu et al. 2014]. In this paper, we categorize the life cycle of BDA into the above four stages since this categorization can accurately capture the key features of BDA in large scale wireless networks, which are significantly different from other computing systems. We next briefly describe them as follows.

1. *Data acquisition.* Data acquisition consists of data collection and data transmission. In particular, data collection involves acquiring raw data from various data sources with dedicated data collection technologies, for example, reading RFID tags by RFID readers in IoT. Then, the data is then transmitted to the data storage system via wired or wireless networks. Details about data acquisition are given in Section 4.
2. *Data preprocessing.* After collecting raw data, the raw data needs to be preprocessed before keeping them in data storage systems because of the big volume, duplication, uncertainty features of the raw data [Wang et al.

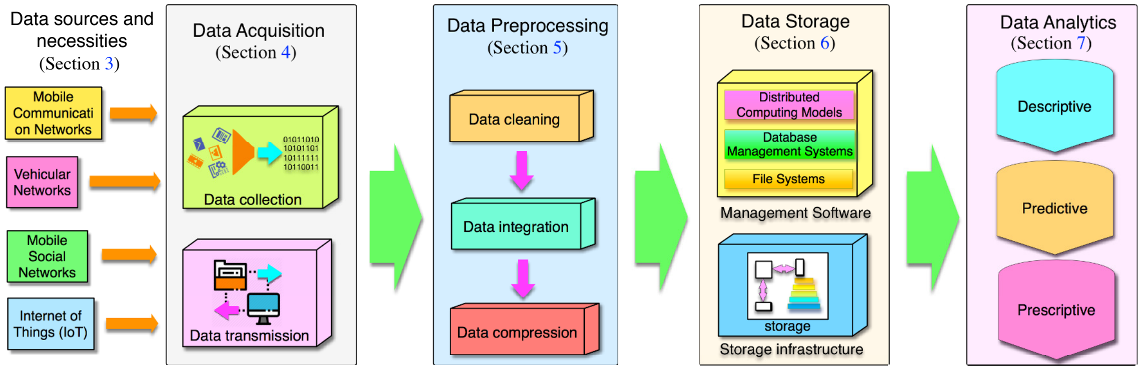


Fig. 1. Life cycle of Big Data Analytics in large scale wireless networks

2012]. The typical data preprocessing techniques include data cleaning, data integration and data compression. We present more details on data preprocessing in Section 5.

3. *Data storage.* Data storage refers to the process of storing and managing massive data sets. We divide the data storage system into two layers: storage infrastructure and data management software. The infrastructure not only includes the storage devices but also the network devices connecting the storage devices together. In addition to the networked storage devices, data management software is also necessary to the data storage system. Details about data storage are given in Section 6.
4. *Data analytics.* In this phase, various data analytics schemes are used to extract valuable information from the massive data sets. We roughly categorize the data analytics schemes into three types: (i) descriptive analytics, (ii) predictive analytics and (iii) prescriptive analytics. Details on data analytics are presented in Section 7.

It is worth mentioning that we also consider different types of wireless networks in each of the above stages. In addition, we present some open research issues and discuss the future directions in this promising area in Section 8. Finally, we conclude this paper in Section 9.

2 RESEARCH METHODOLOGY

2.1 Reference databases and search criteria

Fig. 2 gives the schematic illustration of the methodology adopted in this paper. In particular, we query seven reference databases to obtain the relevant articles/books: 1) ACM Digital Library, 2) Claviate Analytics Web of Science, 3) IEEE Xplore, 4) ScienceDirect, 5) Scopus, 6) SpringerLink and 7) Wiley Online Library. Moreover, we also exploit mainstream search engines to obtain the relevant literature.

Furthermore, in order to include relevant papers as many as possible, we establish a keyword-dataset consisting of keywords and their synonyms. For example, Internet of things may be relevant to wireless sensor networks, Machine-to-Machine communications, cyber-physical systems, smart city, RFID, etc. We search relevant literature according to the search string with the “OR” connection of keywords and their synonyms. Table 2 gives the representative search keywords and their synonyms.

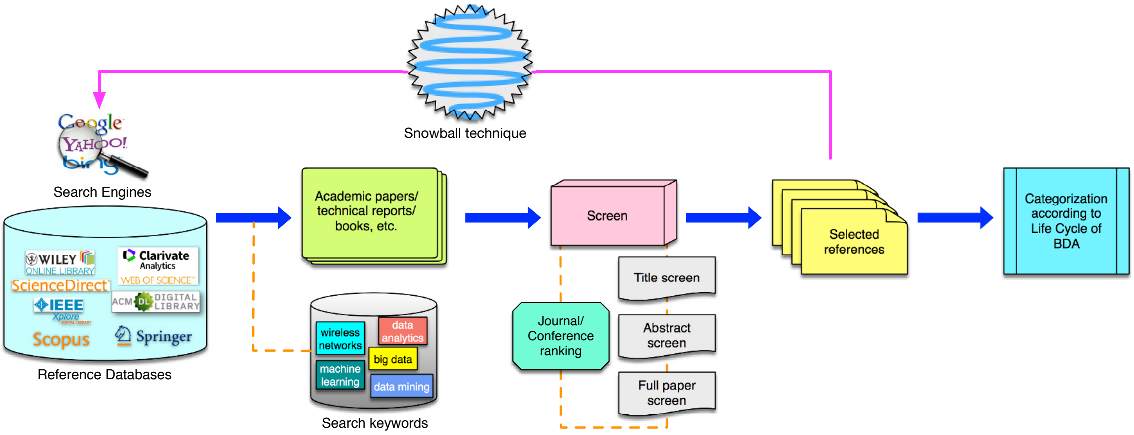


Fig. 2. Methodology adopted in this article

2.2 Data extraction

In the second step, we then read titles and abstracts for initial screening. If necessary, we will read other parts of some papers. It is worth mentioning that we are selective when including relevant high-quality papers while we exclude irrelevant and non-peer-reviewed papers (e.g., papers published in predatory journals). Therefore, we apply journal/conference rankings in the screening process (e.g., Scientific Journal Rankings, Journal Citation Reports, Excellence in Research for Australia, etc.). Moreover, to reflect timeliness of this research area, we choose time window from 2002 to 2018 with an exception of papers related to background knowledge of specific technologies (e.g., storage technologies as shown in Section 6).

Table 2. Representative search keywords and synonyms

Keywords	**Synonyms**
Big data analytics	Big data, Massive data, Machine learning, Deep learning, Data mining, Data analysis, Data science, data cleaning, etc.
Mobile communication networks	Wireless networks, Wireless communications, Cellular networks, WiFi, 802.11, Mobile networks, Mobile communications, 5G, 4G, etc.
Mobile social networks	Social networks, Community, Online Social Networks (OSN), Graph mining, Social media, etc.
Vehicular networks	Vehicular technology, Vehicle, Vehicle-to-vehicle (V2V), transportation systems, intelligent transportation systems (ITS), traffic flow, etc.
Internet of Things	Internet of Things, IoT, Machine-to-Machine (M2M), Wireless Sensor Networks, WSNs, RFID, Cyber Physical Systems, etc.

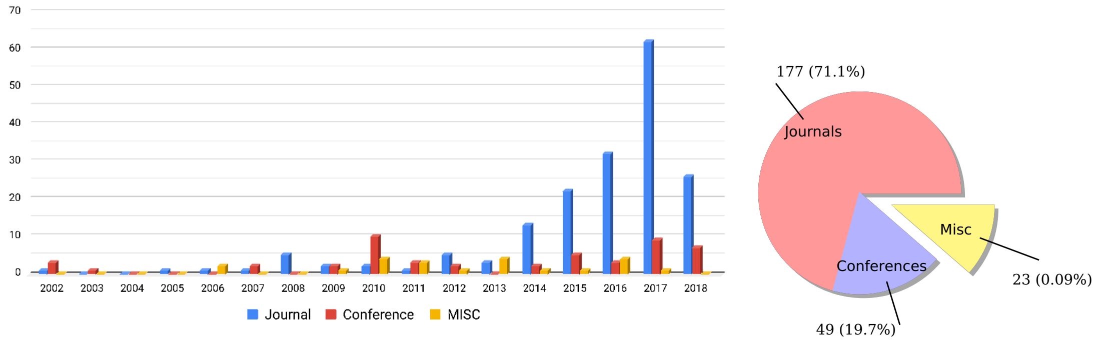
In the third step, we then thoroughly review the articles obtained from initial screening and identify the relevance to big data analytics in wireless networks. In addition, we also extend the systematic literature review by using snowballing technique (as shown in Fig. 2). The main idea of snowballing technique is to use the references or citations of a paper to further include other relevant studies. The advantages of snowballing technique include: i) complimenting traditional systematic review methods, ii) locating hidden while important literature and iii) focusing specific relevant

Table 3. Data elements extracted from the references

No.	Element	Description
1	Bibliographic information	Authors, title, publication year, source of the paper
2	Type of paper	journal, conference, book, technical report, white paper
3	Categorization	four stages in BDA life cycle and four types of wireless networks
4	Novelty	Are new approaches proposed?
5	Validation	Have experimental results validated the observations?
6	Research challenges	Have research challenges been addressed?
7	Open directions	Are any implications given? Are any open research issues raised?

topics [Wohlin 2014]. The usage of references is named as the backward snowballing method (BSM) while the usage of citations is named as the forward snowballing method (FSM). In this article, we use both BSM and FSM. Finally, we obtain 249 references after this step. Table 3 gives the reference extraction guideline, which include major data elements when we select the articles.

2.3 Distribution of references

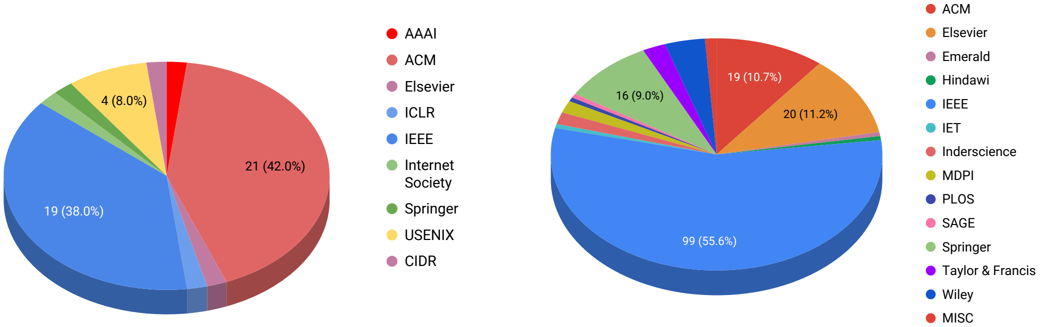


(a) Publication years (horizontal axis) vs No. of selected papers (vertical axis).

(b) Publication types

Fig. 3. Distribution of selected references

Fig. 3 presents an overview of the selected papers in terms of their publication years and types. In particular, Fig. 3(a) shows the distribution of the 249 papers from 2002-2018. We observe that there are few papers between 2002 and 2009 while the number of papers has grown steadily from 2013 to 2018. In 2017, there are 72 papers published, which nearly count for 31% of all the selected papers. It implies that *big data analytics in wireless networks is becoming a hot topic*. According to the published venues, we further categorize papers into the following types: 1) journal, 2)



(a) Conference articles vs publishers

(b) Journal articles vs publishers

Fig. 4. Proportions of conference and journal articles vs publishers

conference/symposium/workshop, 3) miscellaneous (including books, technical reports, etc.). Fig. 3(b) shows the distribution of papers according to the types of papers. We notice that there are more journal papers than other types of papers. It may owe to the fact that we prefer journal articles to other types of papers during the screening process since journal papers typically contain more technical details than other types of papers though we admit that many top-tier conference (such as USENIX NSDI) papers also contain adequate technical details.

Since journal and conference articles occupy the largest proportion (i.e., nearly 90%) among all the references, we further analyze the proportions of conference/journal articles versus publishers. It is shown in Fig. 4(a) that ACM has published the majority of conference articles (42%) followed by IEEE (38%); both of them occupy nearly 80% of total number of conference articles. With regard to journal articles, IEEE has the largest proportion of journal articles (55.5%) among all the publishers followed by Elsevier (11.2%), ACM (10.7%) and Springer (9%) as shown in Fig. 4(b).

Table 4. Summary of the topics in BDA of wireless networks

	Data Acquisition	**Data Preprocessing**	**Data Storage**	**Data Analytics**	**Misc**	**Sub-total**
General topics of BDA	6	12	48	23	24	113
Mobile communication networks	12	4	3	5	11	35
Mobile social networks	6	5	6	8	4	29
Vehicular networks	6	4	6	4	7	27
Internet of Things	12	11	10	7	5	45
Total						249

Table 4 summarizes all the topics in BDA for wireless networks. We categorize the topics mainly according to four types of wireless networks: 1) mobile communication networks, 2) mobile social networks, 3) vehicular networks and 4) Internet of Things. All of them counts for the largest proportion (i.e., 136 papers) among all the references. We observe from Table 4 that "Internet of Things" is one of the hottest topics and it has received extensive attention recently. In addition to four types of wireless networks, we put the remaining 113 papers in the category of general topics of BDA; this class includes references that cover the general topics of BDA ranging from data sources, data acquisition, data preprocessing, data storage and data analytics.

In addition to the horizontal categorization, Table 4 also shows the vertical categorization of the topics according to the four stages in the life cycle of BDA, i.e., data acquisition, data preprocessing, data storage and data analytics. It is worth mentioning that we put other relevant topics such as data sources, necessities and challenges of BDA into *Misc* (i.e., miscellaneous) type. We try to include as many synonyms of key terms as possible when we analyze the papers. Moreover, we also avoid double counting when the content of a paper covers two different types of topics (in this case, we choose the closest topic for this paper after thoroughly reviewing it).

3 DATA SOURCES AND NECESSITIES OF BDA

In this section, we first introduce several typical examples of big data sources of large scale wireless networks in Section 3.1. These data sources include: (1) Mobile Communication Networks as introduced in Section 3.1.1, (2) Vehicular networks as introduced in Section 3.1.2, (3) Mobile Social Networks as introduced in Section 3.1.3 and (4) Internet of Things (IoT) as introduced in Section 3.1.4. We next discuss necessities of BDA in wireless networks in Section 3.2.

3.1 Data sources

3.1.1 Mobile Communication Networks. Mobile communication networks are experiencing a shift from single, simple, low data-rate transmission to multiple, complex, high data-rate transmission with the evolution of mobile communication systems. For example, the downlink data rate of a wireless device in the 5G mobile networks is greater than 20 Gbps, which is about 100 times of that in 4G mobile systems [Shafi et al. 2017]. This shift also exhibits in the wide diversity of data types (e.g., 4k video streams, high fidelity audio, RAW pictures, heart rate, spatial and temporal data in 5G mobile networks). Note that the data generated by cellular networks are not only user data but also system-level data (including cell-level, core-network-level data, etc.) [Xu et al. 2018].

With the growing demands of the bandwidth-ravenous applications, several new wireless access architectures, such as coordinated multipoint (CoMP) [Bassoy et al. 2017], massive MIMO [Molisch et al. 2017], Non-Orthogonal Multiple Access (NOMA) [Shin et al. 2017] and cloud-based radio access network (C-RAN) [Checko et al. 2015] have been proposed. Besides, there is a trend of the fusion of cellular networks with other wireless networks, such as wireless LAN (WLANs), wireless personal area networks (WPANs), and small-cell networks together to form a heterogeneous network (HetNet) [Mehmeti and Spyropoulos 2017].

The proliferation of various coexisting wireless networks in HetNets also results in the wide diversity of data sources. In particular, data sources in HetNets can be categorized into the following types:

- *Subscriber-related data* contains control data and contextual data. Examples include call setup time, call success rate, call drop rate, signaling, packet jitter, delay, etc. In addition, it also includes subscriber specific data [Xu et al. 2018].
- *Network-related data source* contains both radio (Physical) measurement data and Base Stations (BSs) layer-2 (Link) measurement data, where BSs are referred to macro BSs, micro BSs, pico BSs, femto BSs, WiFi APs, etc. Moreover, it also includes network specific data such as faults, configurations, accounting information and performance information.

- *Application Data* contains social media, smartphone sensor data, mobility status, locations, weather, etc.

3.1.2 Vehicular Networks. Vehicular safety and transportation optimization have received extensive attention recently since cars and other private vehicles are playing an important role in our daily life. Vehicular Networks (VNets) were proposed [Cooper et al. 2017; MacHardy et al. 2018] to fulfill safety and efficiency requirements of Intelligent Transportation Systems (ITS). There are two typical communications in a VNet: vehicle-to-vehicle (V2V) communications and vehicle-to-infrastructure (V2I) communications. Various wireless communication technologies were proposed to support V2V and V2I communications. These technologies include IEEE 802.11 (WLAN or WiFi), IEEE 802.11p Dedicated Short Range Communications (DSRC) and Wireless Access in a Vehicular Environment (WAVE) [Bila et al. 2017] and the aforementioned 2G-4G communication technologies.

VNets provide vehicle drivers and other road users (e.g., road operators and pedestrians) with a wide range of information, which can be used to enhance the road safety, the public security, the traveling comfort of passengers and the efficiency of optimizating traffic flows [Koesdwiady et al. 2016]. In particular, we categorize the data sources generated from VNets into the following types.

- *Traffic flow data* contains vehicular speed, density of vehicles, vehicular flow and traffic bottlenecks, which can be used to design an optimal road network and minimize the traffic congestion [Liu et al. 2016].
- *Public safety/security data* includes the route information of suspicious vehicles (e.g., conducting a terrorism behavior) and the event messages of emergency vehicles (e.g., an ambulance uses a lane preemptively).
- *Vehicular safety warning messages* include intersection collision avoidance, turn signals (left turn or right turn), lane change warning, blind spot warning, etc.
- *Ride quality monitoring information* includes the roughness of a road surface (affecting the ride quality) and the slipperiness of a road surface (affecting the ride safety) [Liu et al. 2017].
- *Location-aware social network information* mainly includes not only the messages or micro-blogs of some emergencies, such as traffic jams, malfunctioning traffic signals and accidents but also the traveling information, such as the location and the prices of the nearest restaurant and petrol stations [Giridhar et al. 2016].

The above wide range of data was usually generated from various sensors (such as accelerometer, laser sensors and GPS), cameras, wireless devices of vehicles, smart phones and RFIDs (that are used for re-identification at electronic toll collection transponders). Besides, most of the data is generated in real time and in a time-critical way. For example, a road dangerous warning message could be sent to drivers within several hundred milliseconds [Bila et al. 2017].

3.1.3 Mobile Social Networks (MSNs). Mobile social networks (MSNs) can provide mobile users with various social applications and services due to the proliferation of wireless networks and various mobile devices [Su et al. 2016; Xu et al. 2015]. There are various data sources generated from MSNs including login information, personal profiles, rating information or interests (e.g., “likes”), contextual data (e.g., tags, status, location), photos, video, etc. We categorize them into the following types.

(1) *Service Provider-related Data* mainly refers to the data originates from the service usage of social networks. They include the following sub-types: (i) *Login data.* Social network service providers require the prior user authentication to prevent from the identity theft. (ii) *Connection data.* The connections to MSNs result in a large volume of digital traces caused by protocols of different layers of Open Systems Interconnection (OSI) model. For example, the location information can be acquired through the GPS or IP address. (iii) *Application*

Table 5. Summary of data sources of large scale wireless networks

Data source	**Volume**	**Variety**	**Velocity**	**Value**
Mobile Communication Networks	TB	*User data*: unstructured, semi-structured *Operator data*: structured	very fast (real time)	User satisfaction, operation efficiency, system reliability
Vehicular Networks	TB	structured, semi-structured	very fast (real time)	Transportation safety, transportation efficiency, ride quality
Mobile Social Networks	PB	structured, semi-structured, unstructured	fast	Personal interests, user behavior, social welfare, demography
Internet of Things	TB to PB	structured, semi-structured, unstructured	fast	Environment protection, industrial productivity, public safety

data. In addition, data originates from the use of third party services, e.g., playing online-games, which can be offered by either the same social service provides or other service providers [Richthammer et al. 2014].

(2) *User-related Data* mainly refer to the data related to the personality and social interactions of users including the following sub-types: (i) *User profile data* are profile-centric data which can describe personality aspects of users, e.g., address, education, favorites, hobbies, etc. (ii) *Ratings/interests.* This type of data is mainly expressed interests of users, e.g., “Likes” and ratings of photos/posts shared by others. (iii) *Social Network data.* One of the important features of social networks is *small-world phenomenon*, which refers to the network-like relationships among people. The connections of social networks can be either unidirectional or bidirectional. (iv) *Contextual data.* There are some typical examples of contextual data including tagging peoples’ names in their social networks, the status or the location of a shared item related to an event.

3.1.4 Internet of Things (IoT). Internet of Things (IoT) can connect various *things* to Internet so that data can be collected from the ambiance. The typical killer applications of IoT include the logistic management with Radio-Frequency Identification (RFID) technology [ISO 2013], environmental monitoring with WSNs [Fei et al. 2017], smart homes [Stojkoska and Trivodaliev 2017], e-health [Stankovic 2017], smart grids [Yu and Xue 2016], Maritime Industry [Wang et al. 2015], smart city [Mehmood et al. 2017], etc.

Wireless sensor networks (WSNs) can be regarded a sub-category of IoT technologies [Ayaz et al. 2018]. WSNs were first proposed to support military applications (e.g., surveillance in war zones) while WSNs have a wider range of applications rather than military surveilance. WSNs can be used in environment monitoring [Bai et al. 2018; Boubrima et al. 2017], ITS and smart cities [Memos et al. 2018]. A sensor node usually consists of (i) a power module offering the reliable power, (ii) a sensor module gathering sensory data (via converting raw light, vibration and chemical signals into digital readings), (iii) a micro-controller processing the data received from the sensor and (iv) a wireless transceiver unit transferring the data to another sensor node or a sink. There are a number of wireless communication technologies proposed to support the data communications of WSNs including Bluetooth, IEEE 802.15.4 [Raza et al. 2017], etc. Data sources of WSNs have similar features to the aforementioned networks including a wide range of data types (including temperature, light, pressure and speed), various physical dimensions and data heterogeneity.

As another one of core technologies in IoT, RFID systems have been widely used in supply chain management, inventory control systems, retails, access control, libraries and e-health systems [Ertek et al. 2017]. In particular, RFID allows a sensor (also called a reader) to read a unique identification from a short distance without contacting with the tag [Want 2006]. Data sources generated by RFID usually have the following characteristics: (a) RFID data contains

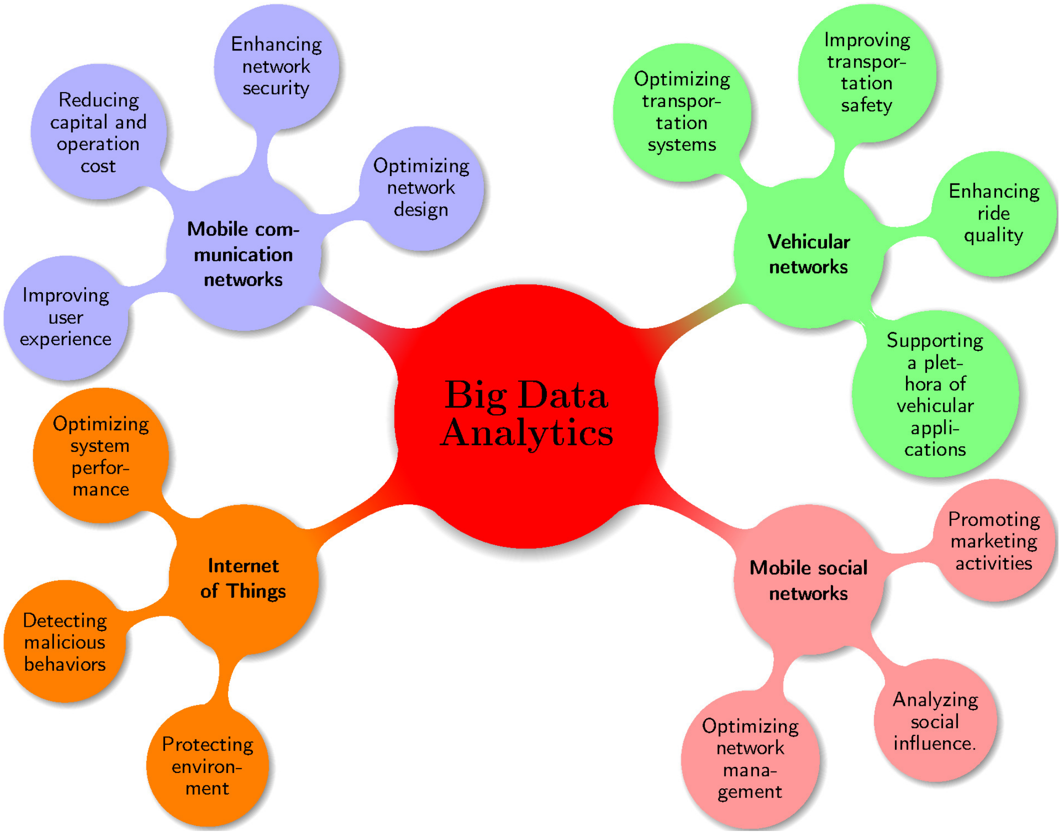

Big Data Analytics
Mobile communication networks
Enhancing network security
Reducing capital and operation cost
Optimizing network design
Improving user experience
Vehicular networks
Optimizing transportation systems
Improving transportation safety
Enhancing ride quality
Supporting a plethora of vehicular applications
Internet of Things
Optimizing system performance
Detecting malicious behaviors
Protecting environment
Mobile social networks
Promoting marketing activities
Analyzing social influence.
Optimizing network management

Fig. 5. Necessities of big data analytics for large scale wireless networks

noise and redundant information; (b) RFID data is temporal and streaming; (c) RFID data is processed on the fly; (d) RFID data volume is enormous.

Table 5 summarizes the key features of the aforementioned data sources. It is obvious that most data sources generate an enormous and heterogeneous data, which requires sophisticated data analytics. Therefore, we next discuss the necessities of BDA for large scale wireless networks.

3.2 Necessities of big data analytics for large scale wireless networks

There are an enormous amount of data generated everyday, which necessitates the demand that such "big data" need to be extensively analyzed so that some valuable and informative information can be obtained. In particular, we summarize the reasons of BDA for different types of wireless networks as shown in Fig. 5.

1. Mobile communication networks

One of the challenges that mobile network operators are facing is the growing cost (including capital expenditure and operating expenditure) and the energy consumption with the growth of user demands. It is necessary to optimize the cost and the energy consumption in mobile communication networks. The recent advances in data analytics have promoted the network optimization based on BDA [Jiang et al. 2017]. The benefits of BDA in mobile communication networks are summarized as follows.

- *Improving user experience.* A mobile user always expects seamless network connectivity, omnipotent service, zero latency and low cost of service. BDA-based network optimization schemes can potentially improve the quality of user experience (QoE) and the quality of service (QoS) by analyzing both mobile service data and user data [Bi et al. 2015b].
- *Reducing capital and operational cost.* Mobile network operators (MNOs) can also benefit from BDA. First, MNOs can easily obtain both the data generated by users and the data generated by various network components. BDA can provide MNOs with deep insights before MNOs make the formal decisions [Zheng et al. 2016]. In particular, BDA technologies can extract intelligence and important information from various data, which can either be instantaneous or historical. The insightful information can help MNOs give both long-term strategies and make immediate decisions to reduce the capital and operational costs.
- *Enhancing network security.* BDA can also help to improve the network security. For example, the study of [Sanctis et al. 2016] shows that using BDA can help to identify anomalies and malicious behaviors. Moreover, BDA can also be used in intrusion detection for next-generation networks [Gai et al. 2016].
- *Optimizing network design.* BDA can help to optimize network design in both software-defined networks (SDN) [Cui et al. 2016] and self-organizing network (SON) [Mohajer et al. 2017]. Moreover, it can be used to manage wireless traffic effectively [Li et al. 2016].

2. Vehicular networks

The application of BDA in vehicular networks can bring a number of benefits as follows.

- *Optimizing transportation systems.* BDA is the core of ITS. Specifically, BDA can provide ITS with important insights, such as planning public transit lanes, adjusting the length of traffic lights and predicting traffic flow [Polson and Sokolov 2017].
- *Improving transportation safety.* BDA also plays an important role in transportation safety. First, BDA can help to provide public with useful transportation information, such as traffic jams, road blockage due to an event and malfunctioning traffic infrastructures. Second, BDA also offers drivers real-time information about driving safety, such as collision avoidance, turn assistant and pedestrian crossing information [Liu et al. 2017].
- *Enhancing ride quality.* On one hand, BDA can offer drivers with the traffic situation and road information, through which drivers can plan a route with the minimum delay or can avoid traffic congestion. Moreover, BDA can help to analyze the user riding experience, consequently improving ride quality [Furtado et al. 2017].
- *Supporting a plethora of vehicular applications.* BDA can support a number of vehicular applications, such as weather and road information, interactive games and roadside services of nearby restaurants or gas-stations since most of them require the data from vehicular networks [MacHardy et al. 2018].

3. Mobile social networks

With the revolution of web technologies and the proliferation of various mobile devices, enormous user data is generated from various mobile social services, including forums, blogs, microblogs, multimedia sharing services and wikis, though which people are virtually connected to form mobile social networks. BDA on mobile social networks can help us acquire valuable insights on personal interests, user behavior, social relations and concerns on media [Lv et al. 2017]. In particular, BDA on MSNs can bring us a number of benefits in the following aspects.

- *Promoting marketing activities.* Obtaining useful information of MSNs, we make better marketing decisions , promote product advertisements [Bao et al. 2016] and enhance recommendation systems[Amato et al. 2019].
- *Analyzing social influence.* BDA on MSNs can also help to predict political election [Stieglitz et al. 2018], offer early warning of epidemics [Atefeh and Khreich 2015] and detect real-time events [Nguyen and Jung 2017]. Moreover, it can be used to detect anomalous behaviors in social networks [Ruan et al. 2016].
- *Optimizing network management.* BDA on MSNs is beneficial to the network optimization [Su et al. 2016] and the location-based service design [Hristova et al. 2016]. Furthermore, it can be used to increase the routing efficiency and reliability in mobile ad hoc networks [Zhang et al. 2017b].

4. Internet of Things

In particular, BDA on IoT can bring us a number of benefits in the following aspects.

- *Protecting environment.* Sensors mounted in IoT can collect various ambient data, which can then be used to identify possible environmental hazards and offer decision support on environmental protecting policies[Montori et al. 2018]. Moreover, the real-time analysis on industrial environmental data can also help to make an immediate response to emergencies [Zhu et al. 2017].
- *Detecting malicious behaviors.* The analysis on massive data in IoT can be used to detect malicious behaviors. For example, it is shown in [Zheng et al. 2018b] that the analysis on smart grid data can help to detect electricity theft consequently securing smart grids. Moreover, the IoT data in the whole food supply-chain is also beneficial to prevent mischievous actions and guarantee food safety [Leng et al. 2018a].
- *Optimizing system performance.* The data analysis on the IoT-enabled supply chain can help to improve the system efficiency and reduce the turnaround time [Dweekat et al. 2017]. In addition, BDA on IoT-enabled intelligent manufacturing shops [Zhong et al. 2017] can also help to make accurate logistic plan and schedules. As a result, the system efficiency can be greatly improved.

4 DATA ACQUISITION

We first discuss the challenges in Section 4.1. We then introduce current solutions to these challenges in four types of wireless networks in Section 4.2. We finally discuss research opportunities in Section 4.3.

4.1 Challenges in data acquisition

There are the following challenges in data acquisition (including data collection and data transmission).

- *Difficulty in data representation.* Due to the high diversity of data sources, the data sets in large scale wireless networks have different types, heterogeneous structures and various dimensions. Take mobile communication networks as an example. There are both user data (such as voice, text and video) and system data (including cell-level and core-network-level data). How to represent these structured, semi-structured and un-structured data becomes one of major challenges in BDA for large scale wireless networks.
- *Effective data collection.* Data collection refers to the procedure of obtaining raw data from various types of wireless networks. This process must be effective and valid since the inaccurate data collection will affect the subsequent data analysis procedure.
- *Efficient data transmission.* How to transmit the tremendous volumes of data to data storage infrastructure in an efficient way becomes a challenge due to the following reasons: (i) *high bandwidth consumption* since the

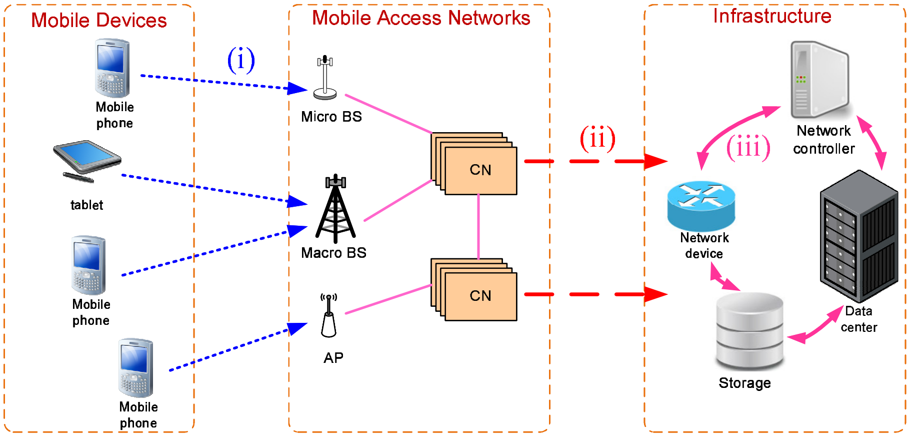


Fig. 6. Data transmission in mobile communication networks

transmission of big data becomes a major bottleneck of wireless systems; (ii) *energy efficiency* is one of major constraints in many wireless systems, such as wireless sensor networks and IoT.

We then present current solutions in data acquisition in large scale wireless networks.

4.2 Current solutions in data acquisition

4.2.1 Mobile communication networks (MCNs). It is a critical issue to represent various types of data in MCNs. In order to support further data preprocessing and analytics, radio waveform data and signaling data are usually converted and represented in matrices or vectors [He et al. 2016] while call detail records (CDRs) and user profiles are represented in records in DBMS [Imran et al. 2014]. User data is stored at mobile devices, base stations and processing units while system data is usually stored at base stations and processing units.

Both user data and system data in cellular networks are mainly collected in the *pull-based* manner [Biral et al. 2015], in which data is collected *proactively* by agents distributed in the whole network. In conventional cellular networks, system data is usually stored at some centralized servers offered by mobile operators. However, the proliferation of massive data in mobile networks leads to the big challenge in managing both user data and system data in heterogeneous networks. Content-centric networks are one of possible solutions to this challenge [Su and Xu 2015]. The main idea of content-centric networks is to cache the popular contests at the intermediate servers (base stations, gateways or routers) so that the user demands for the same content can be fulfilled locally [Wang et al. 2014]. As a result, a lot of traffic can be significantly reduced.

The collected data will be further transmitted to the storage infrastructure. As shown in Fig. 6, data transmission procedure of wireless networks typically consists of the following sub-phases [Bi et al. 2015b]: (i) transmission from mobile devices to base stations (or APs), which are connected with core networks (CNs); (ii) transmission from CNs to storage infrastructure and computing infrastructure (i.e., data centers); (iii) transmission between data centers. In sub-phase (i), the communications are mainly conducted through wireless connections, which have the *limited capacity*, are vulnerable to interference and are susceptible to eavesdroppers compared with conventional wired networks. Sub-phase (ii) mainly consists of wired links connecting base stations to CNs and the links connecting CNs to data centers.

Similar to Sub-phase (ii), Sub-phase (iii) consists of wired links, which usually have the higher bandwidth and are more robust than wireless links [Bhaumik 2012].

The connections between BSs and CUs is named as *front-haul* links [Checko et al. 2015]. CUs are connected with mobile core networks (CNs) through *back-haul* links. It is challenging to manage network resources in both front-haul and back-haul networks effectively in order to support big data applications in next generation mobile networks [Bi et al. 2015b]. C-RAN [Checko et al. 2015] is one of the most promising architectures with cost-efficient and energy-efficient solutions. SON [Mohajer et al. 2017], HetNets [Mehmeti and Spyropoulos 2017] and software defined networking (SDN) [Fan et al. 2016b; Kuang et al. 2016] are other proposals.

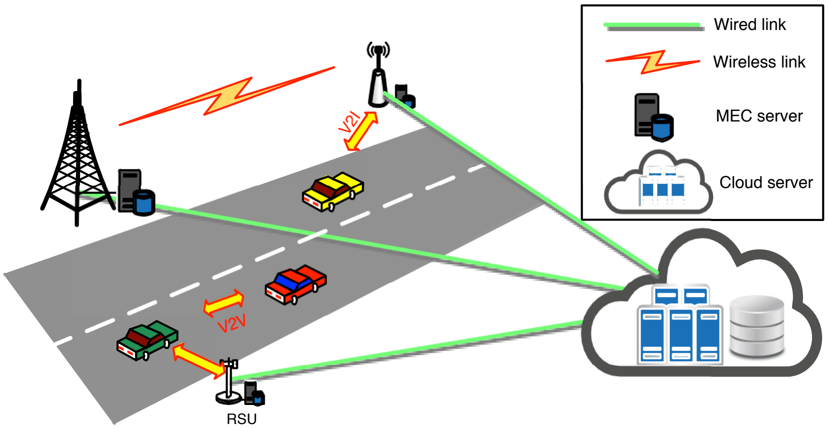


Fig. 7. Data acquisition in vehicular networks

4.2.2 Vehicular networks (VNets). Fig. 7 shows different types of vehicular communications to support data acquisition. For example, data can be transmitted through vehicle to vehicle (V2V), vehicle to road side unit (V2R), vehicle to infrastructure (V2I) manners. The vehicular data can be collected and preprocessed at Mobile Edge Computing (MEC) servers located at RSUs, BSs or at remote clouds. However, data collection of VNets is suffering from the rapid changed network topology due to the movement of vehicles. How to design delay-tolerant data gathering schemes in distributed VNets has received extensive attention recently. In [Mansour and Moussaoui 2015], a new data gathering and distribution scheme named Collaborative Data Collection Protocol (CDCP) was proposed. The main feature of CDCP lies in the optimized delay and it can consequently be used in non-delay tolerant applications. In [Brik et al. 2016], a novel Distributed Data Gathering Protocol (DDGP) was proposed for data collection conducted by vehicles in highways. The main idea of DDGP is to allow vehicles to access the channel in a distributed manner according to their geo-locations. Besides, DDGP removed those redundant, dated and undesired data. As a result, DDGP improves both the reliability and the efficiency of the data collection process compared with other existing schemes. In addition, data collected in VNets are usually represented in data records or text logs[Ilarri et al. 2015], which will be further preprocessed and stored at vehicles, RSUs, BSs and at ITS centers.

How to reduce the amount of data is another challenge in data acquisition of VNets. In [Płaczek 2017], a data transmission scheme based on the estimation of data flow at control nodes. With the support of traffic flow estimation and uncertainty estimation, the amount of transmitted data can be greatly reduced. The study of [Sahoo et al. 2017] proposed a hierarchical aggregation scheme to reduce the amount of data to be transferred in VNets. Besides, data compression schemes [Gandhi et al. 2009] used in WSNs can also be applied to VNets.

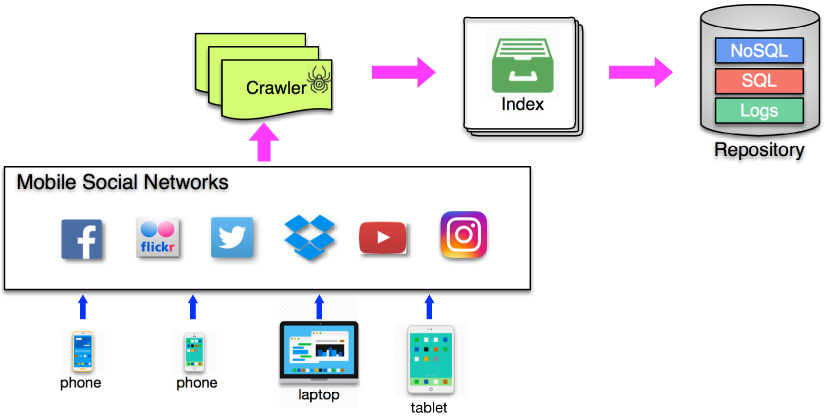


Fig. 8. Data acquisition in mobile social networks

4.2.3 Mobile Social Networks (MSNs). Conventional data gathering methods used in web technologies such as web crawlers to social networks can be used to collect the data from mobile social networks. Fig. 8 shows an example of using web crawlers to obtain mobile social data. Besides web crawlers, log files stored at web servers also gather useful information, such as the clicks, hits, access time and other important attributes, which are either represented in log files or non-SQL data format (like json) [Leskovec and Sosič 2016] [Hajarian et al. 2017]. In addition to web crawlers and log files, we can use various sensors to acquire user-related data from mobile devices. This emerging technology namely mobile crowdsensing (MCS) has received extensive attention recently. There are many challenges in MCS such as coverage constraint [Han et al. 2017], incentive mechanisms [Yang et al. 2017b], privacy preservation [Alsheikh et al. 2017] and energy consumption [Wang et al. 2018a].

The collected data from mobile social networks will then be sent through mobile communication networks to mobile social service providers for the further analysis. Similar to mobile communication networks, mobile social services providers also have data centers to store the massive data from mobile social networks while the above research challenges should be addressed. Since the data transmission follows the procedure similar to that of mobile communication networks, we omit the discussions here.

4.2.4 Internet of Things. RFID tags allow the uniquely identifiers attaced at objects to be read in a short distance by a RFID reader in wireless manner. RFIDs tags can be categorized as data-on-network (DON) and data-on-tag (DOT) types [Diekmann et al. 2007]. Table 6 compares the two different types of RFIDs. In particular, Electronic Product Code (EPC) tag is one of the typical DON RFIDs. In an EPC tag, there is no extra storage except for the product ID information. Usually, an EPC tag can be passively read by an RFID reader periodically and there is no onboard battery on an EPC tag. This design can greatly save the cost. Due to the unstable channel conditions such as the shiedings and reflections [Kennedy et al. 2017], the EPC reads may contain errors and noise. Therefore, filters are needed to preprocess the raw EPC reads. Then, readers aggregate the preprocessed data into events, which are stored in EPC systems and can be further accessed by other EPC applications. With the decreased cost of RFID transponders, DOT types of RFIDs will

Table 6. Comparison between data-on-network tags and data-on-tag tags

	DON tags	**DOT tags**
Information	ID	object related data, such as time, location, temperature, etc.
Storage	low storage capacity	high storage capacity
Energy Supply	No	Yes (some)
Data Access	network connection	presence of objects
Security	Access control at infrastructure	encryption at tags
Cost	low	high

become affordable in the future. Compared with DON tags, DOT tags have the higher storage and can be supplied by onboard batteries. Besides, some sensors (e.g., temperature sensors) can be mounted with DOT tags that can store the sensor information (such as temperature and humidity) [Want 2006]. Some of these DOT tags can actively transmit the information to other nodes.

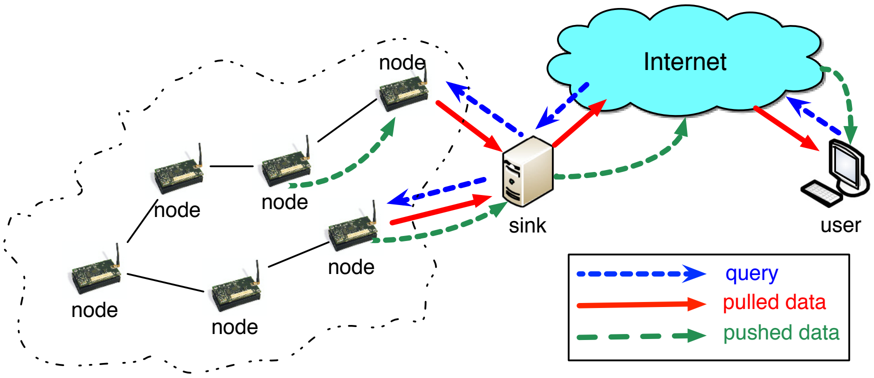


Fig. 9. Pull-based methods vs Push-based methods

Another key IoT technology is WSNs. There are two major approaches for data collection in WSNs: *pull-based* methods and *push-based* methods [Fahmy 2016]. Fig. 9 compares the two methods. In the pull-based schemes, users first send queries to the sink, which then broadcasts the queries to sensor nodes. Then, the data is sent to the sink according the specified queries and the sink next forwards the data to the users. In the push-based schemes, the sensors autonomously decide when to send the sensor data to the sink that consequently forwards the data to the users.

One of major challenges in data transmission in WSNs is the energy consumption since most of sensor nodes have the limited energy (i.e., supplied by batteries). How to design energy-efficient routing/transmission schemes in WSNs has received extensive attention recently. The study of [Chen et al. 2017] proposed an energy-efficient one-to-many broadcasting scheme for data collection in WSNs. Ref. [Abdul-Salaam et al. 2017] developed an energy-efficient data collection scheme for position-free WSNs. In addition to energy consumption, the transmission delay has also been considered in recent works [Takaishi et al. 2014; Yang et al. 2017a].

Conventional RFID and WSNs that are suffering from short communication range (typically less than hundreds of meters) cannot support the wide-coverage scenarios like smart metering, smart cities and smart grids [Xu et al. 2017]. Low Power Wide Area (LPWA) networks essentially provide a solution to the wide coverage demand. Typically LPWA technologies include Sigfox, LoRa, Narrowband IoT (NB-IoT) [Mekki et al. 2018]. One of LPWA advantages is low power consumption. For example, NB-IoT has a ten-year battery life [Xu et al. 2017]. Moreover, LPWA has a longer communication range than RFID and WSNs. Specifically, LPWA technologies have the communication range from 1km to 10 km. Moreover, they can also support a large number of concurrent connections (e.g., NB-IoT can support 52,547 connections [Xu et al. 2017]). However, one of limitations of LPWA technologies is the low data rate (e.g., NB-IoT can only support a data rate upto 200 kps). Therefore, LPWA technologies shall complement with conventional RFID and WSNs so that they can support the various data acquisition requirements.

In addition to LPWA technologies, Wireless LAN is another alternative solution to data acquisition in IoT. The recent work in [Iqbal et al. 2018] presents a data acquisition scheme based on IEEE 802.11AH standard to ensure reliable seismic data acquisition.

4.3 Opportunities in data acquisition

Although the challenging issues as mentioned in Section 4.1 have been partially or fully addressed, it is worth mentioning that there are still many issues not well addressed. We identify three research opportunities as follows:

- *Heterogeneity of data types.* There are different types of data in each type of wireless networks. For example, user data in mobile communication networks is usually unstructured or semi-structured in contrast to structured operator data. There is no general representation method or tool to depict the heterogeneous types of data. This may result in difficulties in data preprocessing, data storage and data analytics. Therefore, research on handling heterogeneous data types will be a new trend in this area.
- *Security in data transmission.* Data transmission in IoT, RFID and WSNs is often vulnerable to malicious attacks due to the limitation of IoT objects, RFID tags and sensor nodes. For example, the security module of narrowband IoT is removed to save the cost of NB-IoT objects [Xu et al. 2017], which however makes the susceptibility of IoT objects to security threats. Recent research efforts like secure key generations based on reciprocity and randomness of wireless channels [Xu et al. cess] and protective jamming schemes [Hu et al. 2018] have shown the effectiveness in IoT.
- *Energy harvesting in data transmission.* In addition to the security issues during the data transmission in IoT, how to provide the low-power RFIDs or other objects in IoT with the energy is another challenge. Recently, RF-enable wireless energy transfer technology [Bi et al. 2015a] provides an attractive solution by powering the RFID tags with energy over the air. However, in order to overcome the higher attenuation of RF energy, directional wireless power transfer is expected [Wang et al. 2016].

5 DATA PREPROCESSING

We first discuss the challenges in Section 5.1. We then introduce exiting studies in data acquisition in four types of wireless networks in Section 5.2. We discuss the research opportunities in Section 5.3.

5.1 Challenges in data preprocessing

Data acquired from large scale wireless networks has the following characteristics:

- *Various data types.* There are various data types generated from large scale wireless networks, including text, sensed values, audio, video, etc. The data is structured, semi-structured and non-structured.
- *Erroneous and noisy data.* The data obtained from wireless networks are often erroneous and noisy mainly due to the following reasons: (a) intermittent loss of communications, (b) the failure of wireless nodes or sensors and (c) interference during the process of data collection [Li et al. 2009]. For example, wireless communications are often interfered by various channel conditions, such as blockage, fading and shadowing effects. Besides, in wireless sensor networks, the data collection may fail when sensors deplete their batteries.
- *Data duplication.* Data generated in large scale wireless networks often contain excessive duplicated information, exhibiting in both temporal and spatial dimensions. For example, the enormous duplicated RFID data will be obtained when several readers read multiple RFID tags at different time moments [Ertek et al. 2017]; this data duplication often results in data inconsistency. Besides, as shown in wireless health-care systems [Zhang et al. 2017a], a vast volume of duplicated medical data has been generated in real-time fashion. It is worth mentioning that data redundancy is often beneficial in DBMS in order to improve the data reliability while data duplication results in data inconsistency especially in data preprocessing.

The above features lead to the following research challenges in data preprocessing.

- *Integration of various types of data.* As mentioned above, data generated in large scale wireless networks has the various types and heterogeneous features. It is necessary to integrate the various types of data so that efficient BDA schemes can be implemented. However, it is quite challenging to integrate various categories of data.
- *Duplication reduction.* An aforementioned challenge lies in the temporal and spatial duplication of the raw data generated in wireless networks. The data duplication often leads to the data inconsistency, which has the impacts on the subsequent data analysis.
- *Data cleaning and data compression.* In addition to data duplication, data of large scale wireless networks is often erroneous and noisy; it inevitably makes data cleaning more difficult. Thus, we need to design effective schemes to compress data and clean the errors of data.

We next present existing studies (i.e., solutions) in data preprocessing in BDA for large scale wireless networks.

5.2 Existing studies in data preprocessing

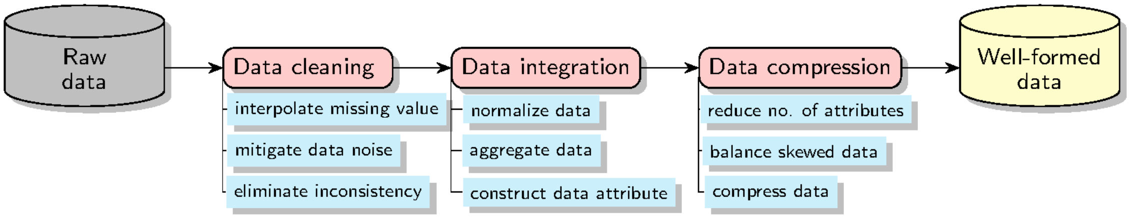


Fig. 10. Data preprocessing procedure

Fig. 10 shows the three major steps in converting raw collected data into well-formed data: 1) data cleaning, 2) data integration and 3) data compression, each of which consists of different preprocessing techniques.

5.2.1 Mobile communication networks.

The typical data preprocessing techniques include *data integration*, *data cleaning* and *data duplication elimination.* Data integration approaches typically include data warehouse method [Lenzerini 2002] and data federation method [Haas et al. 2002]. However, both the two methods may not be applicable to massive data generated from heterogeneous networks. In this sense, we may exploit data integration schemes dedicated for massive data [Gubanov 2017] of mobile communication networks. In order to remove the erroneous, inaccurate, incomplete data, data cleaning is necessary. The typical data cleaning schemes (models) include regression models, probabilistic models and outlier detection models [Han et al. 2012]. Data duplication is a common issue in big data of wireless networks. The typical solutions include duplication detection [Zhang et al. 2002] and data compression [Gandhi et al. 2009]. In addition to the above data preprocessing schemes, other approaches such as dimensionality reduction and feature selection are also useful in data preprocessing [Khatib et al. 2016; Sanctis et al. 2016]. Moreover, a data cleaning model has been proposed in [Fan et al. 2016a] to understand user preference.

5.2.2 Vehicular networks.

In vehicular networks, vehicles and other units such as Road Side Unit (RSU) produce a tremendous amount of data. Data compression is one of typical strategies to reduce the volume of the data. In particular, Feldman et al. in [Feldman et al. 2012] proposed a method named Core-Set to compress the streaming data generated from distributed vehicular networks. The main idea of Core-Set is to construct a small set that approximately represents the original data. Core-Set can also be used to compress the diversity of massive data sets generated from in-car GPS and cameras. Moreover, a cooperative data sensing and compression approach was proposed in [Yu et al. 2010] to compress the data and reduce the communication traffic. This method is mainly based on exploiting the spatial correlation of the data.

In addition to data compression, data of vehicular networks can also contain many errors or incompleted data instances. Therefore, we can apply the aforementioned data cleaning schemes such as regression models, probabilistic models, outlier detection models [Han et al. 2012] to remove the errors and the duplicated data. For instance, it is shown in [Fogue et al. 2014] that there is no noise or inaccuracies detected in data of automotive accidents after applying the above data cleaning and other data mining approaches.

5.2.3 Mobile Social Networks.

Data preprocessing schemes of mobile social networks include data integration, cleaning, and duplication elimination. Similarly, we can use the typical data warehouse method and data federation method to integrate the service-provider-related data of mobile social networks. However, it is quite challenging to integrate the *user-related data* since the above conventional methods cannot be used to integrate mobile sensing data and online social network (OSN) data [Mehrotra et al. 2014]. Besides, the privacy preservation during the data integration is also necessary [Aggarwal and Abdelzaher 2011]. Moreover, the missing values can be restored by data augmentation [Jorgensen et al. 2018].

Data duplication is a critical issue in mobile social networks. There are several proposals in addressing the above challenges. In [Zheng et al. 2010], the integration of local information (obtained from GPS) and activity recommendations was investigated. In particular, only valuable information is extracted. Moreover, Mehrotra et al. [Mehrotra et al. 2014] proposed and implemented a middleware named SenSocial, which can obtain mobile sensory data and integrate with OSN automatically. The two implemented prototypes demonstrated that SenSocial can significantly reduce the efforts in developing OSN applications.

5.2.4 Internet of Things.

RFID data is generally noisy, erroneous and redundant because of complicated wireless channel conditions and cross-reads from multiple readers [Ertek et al. 2017]. Hence, it is necessary to preprocess the raw RFID data. Data preprocessing approaches on RFID data include *data cleaning* and *data compression*. In [Baba et al. 2017], an Indoor RFID Multi-variate Hidden Markov Model (IR-MHMM) was proposed to identify data uncertainty and clean the cross-reads of the RFID data. RFID data also contains valid reads that nevertheless is non-of-interest for analysis. The study of [Ma et al. 2018] proposed a machine-learning based method to filter out this useless data.

In WSNs, sensor data is usually uncertain and erroneous due to the depletion of battery power, imprecise measurement of sensors and network failures. To address these issues, it is necessary to employ data cleaning schemes. However, data cleaning in sensor data is challenging since there are strong temporal and spatial correlations between sensor data. There are several approaches proposed to address this challenge. In particular, an autocorrelation-based scheme was propsed in [Bhandari et al. 2017] to preprocess time-series temperature data and remove duplicated data. In [Tasnim et al. 2017], a novel data cleaning mechanism was proposed to clean erroneous data in environmental sensing applications in WSNs. In WSNs, energy-saving is a critical issue in data-cleaning algorithms. In [Deng et al. 2018], an energy-efficient data-cleaning scheme was proposed. In addition, an interpolation method was proposed in [Zheng et al. 2018b] to recover the missing values of smart grids. Moreover, the work of [AlemÃąn et al. 2018] presents a context-aware data cleaning scheme to estimate and restore the missing values of WSNs while minimizing the error.

5.3 Opportunities in data preprocessing

Although most of the challenges as mentioned in Section 5.1 have been solved, it is worth mentioning that there are still many issues not well addressed. We just enumerate some of research opportunities as follows:

- *Privacy preservation in data preprocessing.* Most of existing studies preprocess data without consideration of the protection of sensitive information of users. For example, in order to improve user experience in text input in mobile devices, mobile operating systems often collect the frequently-used terms from users and preprocess them [Wang et al. 2018b]. It however would violate user's privacy. How to design privacy-preserved data preprocessing schemes becomes a critical issue.
- *Security assurance in data preprocessing*. Data preprocessing has been conducted in different sub-systems throughout large scale wireless networks. The fragmentation and heterogeneity of wireless networks nevertheless often result in the difficulty in guaranteeing the security during data preprocessing. For example, it is shown in [Roman et al. 2013] that IoT systems are also vulnerable to be malicious attacks due to the failure of security firmware updates in time. Although typical solutions such as authentication, authorization and communication encryption can partially amend security vulnerability, the general solutions across heterogeneous wireless systems are still expected.
- *Energy-efficiency in data preprocessing*. RFID tags and wireless sensors are suffering from the limited energy. Therefore, heavy-weighted data preprocessing schemes may not be feasible at these IoT nodes. Many existing studies only consider uploading raw data to remote clouds, which preprocess the data. However, it may cause extra delay to upload and process data at the clouds. Mobile Edge Computing (MEC) can help to offload processing tasks at local MEC serves so that the delay can be greatly reduced.

6 DATA STORAGE

Data storage plays an important role in big data analytics for large scale wireless networks. We first summarize the challenges of data storage in Section 6.1. Fig. 11 illustrates a general data storage system used for big data analytics for large scale wireless networks. In particular, the data storage architecture can be categorized into two layers: storage

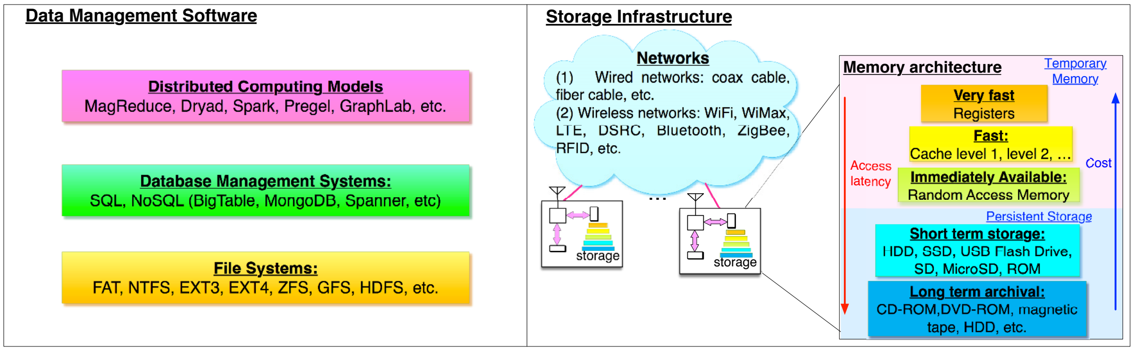

Data Management Software
Distributed Computing Models
MagReduce, Dryad, Spark, Pregel, GraphLab, etc.
Database Management Systems:
SQL, NoSQL (BigTable, MongoDB, Spanner, etc)
File Systems:
FAT, NTFS, EXT3, EXT4, ZFS, GFS, HDFS, etc.
Storage Infrastructure
Networks
(1) Wired networks: coax cable, fiber cable, etc.
(2) Wireless networks: WiFi, WiMax, LTE, DSRC, Bluetooth, ZigBee, RFID, etc.
storage
storage
Memory architecture
Temporary Memory
Very fast
Registers
Fast:
Cache level 1, level 2, …
Immediately Available:
Random Access Memory
Access latency
Cost
Persistent Storage
Short term storage:
HDD, SSD, USD Flash Drive, SD, MicroSD, ROM
Long term archival:
CD-ROM,DVD-ROM, magnetic tape, HDD, etc.

Fig. 11. Data storage system

infrastructure to be introduced in Section 6.2 and data management software to be presented in Section 6.3. We finally discuss the research opportunities in Section 6.4.

6.1 Challenges in data storage

Data storage plays an important role in data analysis. However, designing an efficient and scalable data storage system is challenging in large scale wireless networks. We summarize the challenges in data storage as follows.

- *Reliability and persistency of data storage.* Data storage systems must ensure the reliability and the persistency of data. However, it is challenging to fulfill the above requirements of big data systems while balancing the cost due to the tremendous amount of data [Guerra et al. 2011].
- *Scalability.* Besides the storage reliability, another challenging issue lies the scalability of storage systems for BDA. The various data types, the heterogeneous structures and the large volume of massive data sets of wireless networks lead to conventional databases being infeasible in BDA.
- *Efficiency.* Another concern with data storage systems is the efficiency. In order to support the vast number of concurrent accesses or queries from the data analytics phase, data storage needs to fulfill the efficiency, the reliability and the scalability together, which is extremely challenging.

We next summarize the solutions to the above challenges in this section; we categorize these solutions according to storage infrastructure and data management software (as shown in Fig. 11) to be presented in Section 6.2 and Section 6.3, respectively.

6.2 Storage Infrastructure

Storage devices can be categorized into the following types according to the storage methods [Goda and Kitsuregawa 2012; Placek and Buyya 2006]:

(1) *Persistent Storage* devices mainly include: i. long term storage (for archival usage): magnetic Harddisk Drives (HDDs), magnetic taps, CD-ROMs, DVD-ROMs, etc.; ii. short term storage: HDDs, Solid-State Drives (SSD), USB flash drives, Secure Digital (SD) cards, micro SD cards, Read-Only-Memory (ROM), etc.

(2) *Temporary Memory* devices mainly include: i. immediately available memory: Random Access Memory (RAM); ii. Fast memory: Caches (level 1, level 2, etc.) at CPUs or other processing units; iii. Very fast memory: registers at CPUs or other processing units.

The memory architecture shown in Fig. 11 also indicates the different performance metrics and the cost of different storage devices. In particular, persistent storage devices usually have the lower cost than temporary memory devices while the access latency of persistent storage devices is also significantly higher than that of temporary memory. For example, HDDs are usually 100,000 slower than registers inside CPUs [Patterson and Hennessy 2013]. To balance the cost and the access latency of heterogeneous storage devices, multi-tier hybrid storage architectures have been proposed recently in [Cheng et al. 2015; Guerra et al. 2011; Li et al. 2015; Strunk 2012].

There are various wireless devices including smart-phones, tablets, PCs, laptops, vehicles, GPS navigators, sensors, RFID tags, IoT objects and wearable devices, etc. Due to the heterogeneity of wireless devices, each wireless device may only include subsets of the aforementioned storage devices but not all of them. For example, a smart-phone may consist of the persistent storage devices, such as ROM and a micro SD card, as well as the temporary memory devices, such as RAM and registers while there is no HDD in a smart-phone due to the bulky size of HDDs. However, an EPC RFID tag may only contain a Complementary metal-oxide-semiconductor (CMOS) integrated circuits with Electrically Erasable Programmable Read-Only Memory (EEPROM) with limited storage capacity [Landt 2005].

It is worth mentioning that the wireless devices are interconnected together to form the storage infrastructure for large scale wireless networks as shown in Fig. 11. There are several ways of managing the network infrastructure of storage systems: (i) Directed Attached Storage (DAS), (ii) Network Attached Storage (NAS) and (iii) Storage Area Network (SAN). Essentially, the distributed storage systems often introduce redundancy to increase reliability with erasure coding repair schemes [Weatherspoon and Kubiatowicz 2002], which however inevitably cause the extra network traffic. There are a number of solutions proposed to address this challenge [Al-Awami and Hassanein 2017; Chen et al. 2015, 2016; Jun et al. 2016; Sathiamoorthy et al. 2014; Zhu et al. 2015].

6.3 Data management software

Data management software plays an important role in constructing the scalable, effective, reliable storage system to support BDA. As shown in Fig. 11, we categorize the data management software into three layers: (i) file systems, (ii) database management systems and (iii) distributed computing models, which are explained as follows.

(i) File systems

We start the introduction from simple file systems at a single personal computer (PC) to distributed file systems. In a PC, a file system is mainly used for the purpose of data storage and retrieval. The typical file systems for PCs include: a) The family of File Allocation Table (FAT) file systems: FAT 12, FAT 16, FAT 32; b) NTFS (New Technology File System) offering better security than FAT; c) index-node file system: Unix File Systems (GPFS, ZFS, APFS, etc.) and Linux File Systems (ext2, ext3, ext4); d) Contiguous allocation file systems: ISO 9660:1988 and Joliet ("CDFS"), which are mainly used for CD-ROM and DVD-ROM disks.

In order to support the file sharing and the collaboration, there are a number of distributed file systems including Andrew File System (AFS) and Network File System (NFS) [Arpaci-Dusseau and Arpaci-Dusseau 2015], etc. However, most of these distributed file systems can only be accessed within a local area network and are not scalable to support the large scale data storage in wide area networks.

Google File System (GFS) [Ghemawat et al. 2003] was proposed by Google to support the large data intensive applications (e.g., search engine) in distributed environments. In particular, GFS divides the data into equal-size blocks (typically with 64MB per block), which have been stored on different machines with several copies to ensure the fault

Table 7. Comparison between SQL Database and NoSQL Database

	SQL Database	**NoSQL Database**
Schema	Relations (structured or fixed types)	Non-structured and varied types
Storage Models	tables (in rows or records)	Combination of varied types
Scalability	lowly scalable	highly scalable
ACID	Guaranteed	Supported by some of them
Programming Interfaces	Common (e.g., using SQL)	Varied APIs depending on databases
Examples	IBM DB2, Oracle, Microsoft SQL server, MySQL, Postgres, SQLite	Dynamo [DeCandia 2007], BigTable [Chang et al. 2008], Cassandra [Lakshman and Malik 2009], Hbase [Apache 2016], MongoDB [Chodorow and Dirolf 2010], Megastore [et al. 2011] and Spanner [Corbett 2013]

tolerance. However, GFS is suffering from a number of drawbacks, such as inefficiency with smaller files and the degraded performance with the increased number of random writes. It is shown in [McKusick and Quinlan 2009] that Google has partially solved the above defects of GFS.

Hadoop Distributed File System (HDFS) [Shvachko et al. 2010] proposed by Apache uses GFS for reference. HDFS is designed to store massive data sets reliably and support big data applications. The main idea of Hadoop is to partition data and computation across many servers and to execute computations in the parallel manner. Essentially, HDFS offers the scalable storage infrastructure to support Hadoop computational tasks.

In addition to GFS and HDFS, there are other distributed file systems, such as XtreemFS [Hupfeld et al. 2008], C# Open Source Managed Operating System (Cosmos) proposed by Microsoft [Chaiken et al. 2008] and Haystack proposed by Facebook [Beaver et al. 2010]. Most of them can partially or fully support the storage of large scale data sets.

(ii) Database management systems

Database management systems (DBMS) concern how to organize the data in an efficient and effective manner. We roughly categorize DBMS into two types: (i) traditional relational DBMS and (ii) non-relational DBMS. In short, we name traditional relational DBMS as Structured Query Language (SQL) Database since most of them can support SQL queries. Similarly, we name non-relational DBMS as No-SQL database. Table 7 summarizes the main differences between SQL DBMS and NoSQL DBMS.

SQL databases have been a primary data management approach since 1970s. There are several typical examples of SQL databases including commercial databases, such as Oracle, Microsoft SQL server and IBM DB2, and other open-source alternatives (e.g., MySQL and PostgreSQL). Before operating on data, a *schema* must be defined first. The schema usually defined tables, field types, primary keys, indexes, relationships, triggers and stored procedures.

SQL databases usually store data in tables of records, resulting in the poor scalability. There is a need to partition the data and distribute the operation load among different nodes (i.e. the database servers) with the growth of data. However, it is challenging to allocate data in a distributed manner [Rahimi and Haug 2010]. One of benefits of SQL databases is that most of SQL databases can guarantee ACID (Atomicity, Consistency, Isolation, Durability) properties, which are crucial to many commercial applications. Besides, the standardized SQL also facilitates the application development process since most of programming languages can operate on databases through general programming interfaces (such as ODBC and JDBC).

Different from SQL databases, there is no schema to specify the data design in NoSQL databases. Besides, NoSQL databases also support various types of data, such as records, text, and binary objects. Compared with traditional relational databases, most of NoSQL databases are usually highly scalable and can support the tremendous amount of data. Thus, NoSQL databases have received extensive attention recently especially in BDA. We next briefly review several typical NoSQL databases.

- *Key-Value databases.* In Key-Value databases, data is organized as key-value pairs. Similar to the primary key in SQL databases, each key in a Key-Value database is also unique. Through partitioning and replication mechanisms, such databases have higher scalability than traditional databases. Dynamo [DeCandia 2007] is one of typical Key-Value databases though there are many other variants similar to Dynamo.
- *Column-Oriented databases.* Instead of storing data in rows, Column-Oriented databases store and manage data in columns. Bigtable [Chang et al. 2008] invented by Google is one of the typical column-oriented databases. In Bigtable, data is organized in a sparse and distributed *map*, in which column keys, row keys and time keys are the indexes. Bigtable is essentially implemented on GFS with the integration of other technologies. There are other alternatives to Bigtable, including Cassandra [Lakshman and Malik 2009] and Hbase [Apache 2016].
- *Document databases.* Document databases can support more complicated data types than SQL databases and key-value databases. Besides, there is no schema in document databases, consequently increasing the variety of data types. Typical document databases include MongoDB [Chodorow and Dirolf 2010], CouchDB [Anderson et al. 2010] and SimpleDB [Chaganti and Helms 2010].
- *Hybrid databases.* Hybrid databases integrate the benefits of SQL databases and No-SQL databases and obtain the better performance than both SQL databases and No-SQL databases. Megastore [et al. 2011] and Spanner [Corbett 2013] are two of typical examples of hybrid databases. They achieve the high scalability of NoSQL databases while guaranteeing data consistency (e.g., ACID properties) of SQL databases.

(iii) Distributed computing models

Though distributed databases [Rahimi and Haug 2010] were proposed to improve the performance of DBMS through distributing operations over distributed data storages, they cannot fulfill the growing demands of BDA on the tremendous amount of data in large scale distributed systems. As a result, there are a number of distributed computing models proposed for BDA. In this paper, we roughly categorize those models into the following three types as shown in Fig. 12.

- *General Purpose Models* have been proposed to support BDA in large scale distributed database systems [Koch et al. 2016]. Typical general purpose models include MapReduce [Dean and Ghemawat 2008], a distributed programming model, is mainly used to big data analytics. MapReduce consists of *Map* functions and *Reduce* functions. Specifically, the *Map* function first processes and sorts key-value pairs, and then save the intermediate data to temporary storage. The *Map* function next consolidates the intermediate data (i.e., key-value pairs). Hadoop MapReduce [Apache 2014] is the open source implementation of Google MapReduce. One of limitations of MapReduce lies in the lack of iterations or recursions, which are however required by many data analysis applications, such as data mining, graph analysis and social network analysis. There are some extensions to MapReduce to address this concern, including HaLoop [Bu et al. 2010], Berkeley Orders of Magnitude (BOOM) Analysis [Alvaro et al. 2010], Twister [Ekanayake et al. 2010], iHadoop [Elnikety et al. 2011] and iMapReduce

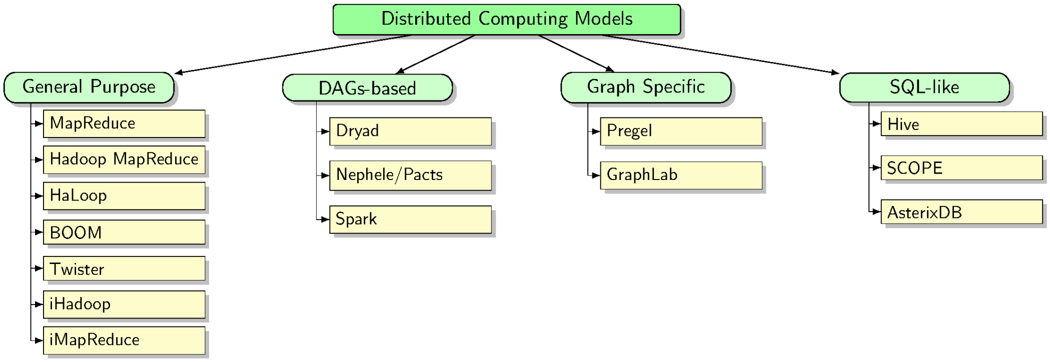


Fig. 12. Classification of distributed computing models

[Zhang et al. 2012]. MapReduce has the merits including the scalability (e.g., it can be easily extended to a scenario of massive data sets and deployed in commodity PCs) and the simplicity (e.g., it can easily implemented by programmers without any parallel/distribute computing experience).

- *Directed Acyclic Graph (DAG)-based Models*. Dryad [Isard et al. 2007] models an application as a Directed Acyclic Graph (DAG), in which a vertex represents a computation and a directed edge represents a communication between vertices. In contrast to MapReduce, Dryad improves the flexibility while sacrificing model simplicity. Nephele/PACTs system [Battré et al. 2010] is a parallel data processing system including two components: Nephele and Parallelization Contracts (PACTs) where Nephele is a parallel computing system and PACT is a programming model. Nephele/PACTs has some similarity to Dryad owe to the flexible provision of dataflows on top of DAGs. Spark [Zaharia et al. 2010] is a cluster computing framework and can be easily extended to support different applications. Spark has the similar scheduling scheme to Dryad (based on DAGs), but it assigns tasks to machines based on the data locality information.
- *Graph Models*. Pregel [Malewicz 2010] adopts a vertex-centric model, in which each vertex undertakes the computing tasks. Moreover, each edge consists of a source vertex and a destination vertex. In Pregel, every edge or every vertex is linked with a value. Similar to Pregel, GraphLab [Low et al. 2012] is also a vertex-centric approach. The main differences between Pregel and GraphLab are (i) different access privileges at vertices and edges and (ii) asynchronous/synchronous iterations.
- *SQL-like Models*. SQL is a declarative language widely used in relational DBMS. Recently, there are a number of studies on designing distributed big data processing systems to support SQL or SQL-like language. In particular, Hive [Thusoo et al. 2010] developed by Facebook can undertake extensive data processing tasks. Hive supports SQL-like queries by using HiveQL. Essentially, Hive compiles HiveQL queries, packs them into MapReduce jobs and executes them on Hadoop. Hive has an excellent scalability and can be used to construct data storage systems while it has poor performance in processing interactive queries. Other alternatives include Structured Computations Optimized for Parallel Execution (SCOPE) [Chaiken et al. 2008] and AsterixDB [Alsubaiee 2014].

6.4 Opportunities in data storage

Most of the challenges in data storage have been solved based on solutions in Section 6.2 and Section 6.3. However, it is worth mentioning that there are still many issues not well addressed. We just identify three research opportunities as follows:

- *Data storage and distributing computing models for heterogeneous networks.* There are few studies contributed to data storage and distributing computing models dedicated to different types of wireless networks. Different types of wireless networks may require different types of data storage and distributing computing models.
- *Lightweight computing models.* Most of previous studies just assume that data is either stored at a cloud (or a fog) server or at a mobile device and sophisticated cloud-based computing models are used. However, many wireless devices (like IoT or sensors) may have the constraints such as limited storage and computational power. Heavy-weighted storage or computing models may not be feasible in this scenario.
- *Privacy preservation and security assurance.* Despite recent advances in distributed storage and computing systems, privacy-leakage concerns have received extensive attentions. For example, it is shown in [Bazai et al. 2017] that the output of MapReduce operations may cause the potential privacy leakage. To address the privacy leakage, Rao et al. [Ram Mohan Rao et al. 2018] proposed a number of privacy-preservation techniques including data anonymity, data diversity, randomization and cryptographic schemes. Meanwhile, Wang et al. [Wang et al. 2018] proposed an efficient and verifiable erasure coding based storage to secure HDFS-like systems. In addition to these schemes, recent advances in blockchain technologies [Zheng et al. 2018a] also show the strength in data privacy preservation and security assurance.

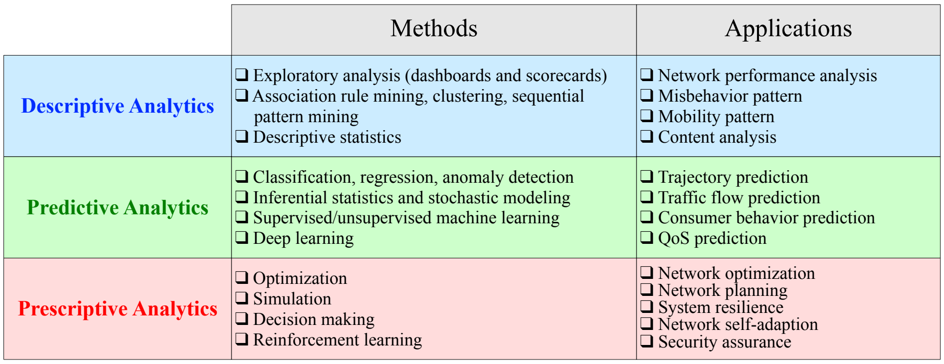
7 DATA ANALYTICS

The goal of data analysis is to extract useful information from large scale wireless networks. In this section, we first identify the challenges in data analytics in Section 7.1, then review the common solutions to these challenges in Section 7.2. In Section 7.3, we then enumerate the applications of the data analysis tools in the aforementioned typical wireless networks. We finally discuss the research opportunities in Section 7.4.

7.1 Challenges in data analytics

It is quite challenging in BDA for large scale wireless networks due to the tremendous volume, the heterogeneous structures, the high dimension and the wide data diversity. The major challenges are summarized as follows.

- *Data temporal and spatial correlation.* Different from conventional data warehouses, data of large scale wireless networks is usually spatially and temporally correlated. How to manage the data and extract valuable information becomes a new challenge.
- *Efficient data mining schemes.* The tremendous volume of data of large wireless networks leads to the challenge in designing efficient data mining schemes due to the following reasons: (i) it is not feasible to apply conventional multi-pass data mining schemes due to the huge volume of data, (ii) it is critical to mitigate the data errors and uncertainty due to the erroneous features of wireless systems.
- *Privacy concern.* It is quite challenging to pertain the privacy of data during the analysis process. Though there are a number of conventional privacy-preserving data analytical schemes, they may not be applicable to the wireless data with the huge volume, heterogeneous structures, various types and spatio-temporal correlations.



	Methods	Applications
Descriptive Analytics	❑ Exploratory analysis (dashboards and scorecards) ❑ Association rule mining, clustering, sequential pattern mining ❑ Descriptive statistics	❑ Network performance analysis ❑ Misbehavior pattern ❑ Mobility pattern ❑ Content analysis
Predictive Analytics	❑ Classification, regression, anomaly detection ❑ Inferential statistics and stochastic modeling ❑ Supervised/unsupervised machine learning ❑ Deep learning	❑ Trajectory prediction ❑ Traffic flow prediction ❑ Consumer behavior prediction ❑ QoS prediction
Prescriptive Analytics	❑ Optimization ❑ Simulation ❑ Decision making ❑ Reinforcement learning	❑ Network optimization ❑ Network planning ❑ System resilience ❑ Network self-adaption ❑ Security assurance

Fig. 13. Classification of Data Analytics Methods

- *Real-time requirement.* Wireless data is often generated in real-time fashion. it is challenging to design and implement data analytics schemes in supporting real-time wireless applications while maintaining high performance (like prediction accuracy).

We next survey the common data analytics methods to address these challenges.

7.2 Common data analytics methods

We categorize the common data analytics methods into the following types (as shown in Fig. 13).

7.2.1 Descriptive methods. Descriptive analytics mainly utilizes existing data sets to reveal the properties of data (i.e., what happened). We further categorize the descriptive methods into the following subcategories.

- *Association Rule Mining* is to determine the dependence between two (or more than two) features. Typical association rule mining algorithms include Apriori algorithm and Frequent Pattern Growth (FP-Growth) algorithm [Han et al. 2012]. The main idea of Apriori algorithm is to identify the frequent individual items in the database and then to extend them to larger item sets as long as the frequency of the appearance of those item sets is sufficiently high in the database. One of disadvantages of Apriori algorithm lies the cost of generating candidate items. FP-Growth algorithm is based on an extended prefix-tree structure, which can store frequent patterns (hence this structure is also named as frequent-pattern tree).
- *Clustering* is a method of grouping objects such that objects in the same group have the higher similarity to each other than to those in other groups. The group consisting of similar objects is also named a *cluster*. Typical clustering approaches include k-means and Density-based spatial clustering of applications with noise (DBSCAN) [Han et al. 2012]. The main idea of k-means is to divide n objects into k clusters such that objects with the nearest mean belong to the same cluster. DBSCAN partitions and associate the points into three groups: (1) the points, each of which contains a large number of neighbors; (2) the points that are reachable from the points from group (1); (3) the outliers that are not reachable from both the points in (1) and (2).
- *Sequential Pattern Mining* is the process that discovers relevant patterns between data examples where the values are delivered in a sequence [Mooney and Roddick 2013]. There are several typical sequential pattern mining algorithms: Generalized Sequential Pattern (GSP) and Sequential Pattern Discovery Using Equivalent Class (SPADE) and Prefix-Projected Sequential Pattern Mining (PrefixSpan) [Han et al. 2012]. GSP conducts

multiple database passes. The first pass counts all single items (with sequence equal to 1). The second pass counts a set of candidate 2-sequences and one extra pass is conducted to identify their frequency. The set of candidate 2-sequences is used to generate the set of candidate 3-sequences. This process is repeated until no more frequent sequences are found. The main idea of SPADE is to divide the original problem into a number of sub-problems, which are independently solved in main-memory. During this process, efficient lattice search techniques are often used. PrefixSpan has further improved GSP and SPADE. In PrefixSpan, a sequence database is recursively partitioned into a number of sub-sets. Then, sequential patterns grow in each sub-database by selecting frequent segments only.

- *Descriptive statistics* can summarize features of data by exploiting measures of data. The typical measures of descriptive statistics include: (i) *central tendency* such as mean, median and mode (e.g., bimodal and multimodal); (ii) *dispersion* (or variability) including variance, covariance, tailedness (or kurtosis) and skewness [Trochim et al. 2016]. Besides, there are two typical descriptive statistical methods: *univariate* analysis and *bivariate* analysis. Univariate analysis mainly describes the distribution of a single variable while bivariate analysis is mainly used to the relationship between pairs of variables with sample data sets consisting of more than one variable.

7.2.2 *Predictive analytics.*

Predictive analytics mainly utilizes historical data to anticipate the trends of data (i.e., what will occur in the future). The predictive methods can be categorized into the following subcategories.

- *Classification* is the process of assigning items in a collection to target categories. Classification is considered as an instance of supervised learning while clustering is considered as an example of unsupervised learning algorithms The typical classification algorithms include naïve Bayes, Bayes networks, k-Nearest Neighbors (k-NN), support vector machines (SVMs) and C4.5 [Wu et al. 2008]. Naïve Bayes is a simple scheme to classify features based on Bayes' theorem. Bayes classifier can minimize the probability of mis-classification. The main idea of k-NN is to classify an object by considering the closeness to its k nearest neighbors. SVM is to find a hyperplane that maximizes the gap between the classes. C4.5 is an algorithm used to generate a decision tree, which can then be used for classification.
- *Regression* is a procedure of establishing a function to determine the relationship between target features. The regression is similar to classification though regression is mainly used to predict continuous values but classification is used to predict discrete values. Typical regression algorithms include linear regression and logistic regression.
- *Anomaly Detection* (or outlier detection) is to identify objects that do not comply with an expected pattern as given. Anomaly detection approaches can be categorized into the following types [Buczak and Guven 2016]: (a) *supervised* anomaly detection schemes, (b) *unsupervised* anomaly detection schemes, and (c) *semi-supervised* anomaly detection schemes. The main difference between supervised schemes and unsupervised schemes mainly lies in the fact that the supervised schemes require a labeled (normal or abnormal) data set and a trained classifier. Semi-supervised schemes also use the trained classifier while it only consists of normal data.
- *Inferential statistics* can infer hidden distribution properties from the sample data. Different from descriptive statistics, inferential statistics deduces the data from a larger data set than the observed data set. A typical statistical inference procedure includes *testing hypotheses* and *deriving estimates* [Bandyopadhyay and Forster 2011].

- *Stochastic modeling methods* have recently received extensive attention since they can capture the dynamic features of data traffic, predict user mobility and track objects. There are several typical stochastic models: dynamic Bayesian networks (DBNs), Markov models, Kalman filters and Extended Kalman filters [Klaine et al. 2017]. Most of these stochastic methods require collecting a certain amount of user data to provide stochastic models with parameter estimation (e.g., parameter estimation in transition matrices of Markov chains).
- *Supervised learning algorithms* require training data with labels first. Predefined inputs and desired outputs are also given in supervised learning. The goal of supervised learning is to find the relationship between the inputs, the outputs and other arguments. Typical supervised learning algorithms include support vector machines (SVMs), naïve Bayes, Decision tree learning, k-Nearest Neighbors (k-NN), hidden Markov model, Bayesian networks, neural networks and Ensemble methods [Klaine et al. 2017; Qiu et al. 2016; Russell and Norvig 2009].
- *Unsupervised learning algorithms* do not require labeled training data and the desired outputs. The main goal of unsupervised learning is to classify items to target categories (i.e., clustering). Typical unsupervised learning algorithms include k-means, singular value decomposition (SVD) and Principal Component Analysis (PCA) [Abdi and Williams 2010]. The k-means algorithm is mainly used to assign data into different categories (or types). The main idea of PCA is to transform multiple correlated variables into new linearly orthogonal variables. These linearly uncorrelated variables is also called principal components.
- *Deep learning algorithms.* Conventional learning methods are mainly conducted in *shallow-structured* learning architectures, which are not suitable for identifying complicated patterns. Recently, deep learning architectures have received extensive attention [Sze et al. 2017]. There are two typical deep learning approaches: Convolutional Neural Networks (CNNs) and Deep Belief Networks (DBNs) [Fadlullah et al. 2017].

7.2.3 Prescriptive analytics. Prescriptive analytics extends the results of both descriptive and predictive analytics to make right decisions in order to achieve predicted outcomes (i.e., what should we do to achieve the goal?). The prescriptive methods can be categorized into the following subcategories.

- *Simulation.* The proliferation of massive data of both descriptive and predictive analytics can be used to predict possible outcomes by mimicking possible scenarios with consideration of various constraints. Typical simulation tools include Monte Carlo simulation [Kroese et al. 2013], vehicular traffic flow simulations [Treiber and Kesting 2013], Operator Training Simulator (OTS) for industrial systems [Gerlach et al. 2015], etc.
- *Optimization.* Optimization techniques include linear programming, integer programming, and nonlinear programming. The optimization problem usually concerns maximizing or minimizing a certain outcomes while satisfying given constraints. The typical optimal outcomes in wireless networks include throughput, delay, outage, energy consumption, operation cost and QoS [Kibria et al. 2018].
- *Decision making.* A typical decision making process includes 1) identifying objectives (predictive outcomes), 2) giving a number of alternatives, 3) selecting the best alternative fulfilling the optimal outcomes via problem modeling (or simulation). Multiple criteria decision making (MCDM) has been typically used in decision making process. MCDM can help to find optimal outcomes in complicated scenarios with consideration of various criteria and conflicting objectives. Recently, MCDM models have been used in IoT systems [Nunes et al. 2016], smart grids [Kumar et al. 2017] and traffic flow control [Jiang et al. 2018].
- *Reinforcement learning algorithms* enable software agents to learn via the interactions with the context (or ambience) so that the cumulative reward can be maximized. One of the most popular reinforcement learning

algorithms is Q-learning [Klaine et al. 2017; Russell and Norvig 2009], which is a model-free reinforcement learning technique. The main goal of Q-learning is to find an optimal policy on selecting actions for any finite Markov decision process. The procedure can be modeled by a quality value (i.e., Q-value). Q-learning has the strength that it works without the environment model.

7.3 Applications of data analytics

We next offer an overview on the applications of the data analytics in large scale wireless networks.

7.3.1 Mobile communication networks. BDA can be used to extract important features from mobile communication networks. We enumerate some of typical BDA applications in mobile communication networks as follows.

- *Network performance analysis.* Classification and regression analysis can be exploited to analyze and predict the mobile traffic. For example, a machine learning based approach was proposed in [El Khayat et al. 2010] to enhance the throughput of wireless networks via distinguishing packet loss causes. Moreover, the study of [Zhang et al. 2015] proposed a new scheme named Robust statistical Traffic Classification (RTC) by combining both supervised and unsupervised ML techniques to address the *zero-day* problem. Furthermore, a multiclass-classification learning approach was proposed in [Joung 2016] to solve the antenna-selection problem in mobile communications.
- *Network security.* Both unsupervised clustering and supervised classification algorithms have been used to detect network intrusions or other malicious attacks. In particular, AdaBoost, Naive Bayes, Random Forest were proposed to detection the intrusion in 802.11 networks [Kolias et al. 2016]. Moreover, machine learning-based techniques can also be used in constructing intrusion detection systems for mobile clouds in heterogeneous networks [Gai et al. 2016].
- *Mobility (Trajectory) prediction.* Mobility prediction has received extensive attention recently since it can offer the supports for various wireless services and mobile applications. There are a number of efforts been done in this area. Conventional approaches include using Markov models, Kalman filters and Extended Kalman filters to predict the user mobility. However, these approaches suffer from the poor accuracy. In [Xia et al. 2017], a novel nonlinear SVM-based framework using massive spatio-temporal data was proposed. This approach was demonstrated to have the better accuracy than other existing methods.

7.3.2 Vehicular networks. We enumerate several typical BDA applications in vehicular networks as follows.

- *Misbehavior detection.* In particular, an intrusion detection systems based on deep neural networks was proposed to identify the malicious behaviors in vehicular networks [Kang and Kang 2016]. Experimental results demonstrated the accuracy of the proposed approach. Moreover, in [Scalabrin et al. 2017], a novel method based on Bayesian networks was proposed to analyze the anomaly in traffic data.
- *Network performance.* The network topology and communication links in vehicular networks have experienced the frequent transition due to the high mobility of vehicles. The optimization of data transmissions in vehicular networks requires the accurate prediction of the moves of vehicles [Ye et al. 2018]. The study of [Lai et al. 2015] proposed a novel routing information system called the machine learning-assisted route selection (MARS) system to estimate necessary routing information. Experimental results show that MARS can significantly improve the network performance.

- *Traffic prediction.* Traffic prediction in vehicular networks has received extensive attention recently. In [Yu et al. 2016], a short-term traffic condition prediction model based on k-nearest neighbor algorithm was proposed. The study in [Ko et al. 2016] proposed a Markov process based method to predict traffic conditions between roads. Moreover, a deep learning-based approach [Polson and Sokolov 2017] was proposed to predict short-term traffic flow. Furthermore, Zhao et al. [Zhao et al. 2017] proposed a novel method based on long short-term memory (LSTM) to predict the traffic flow. The benefit of this approach is the capability of capturing time-series features of traffic flow.

7.3.3 Mobile Social Networks. We enumerate several typical BDA applications of mobile social networks as follows.

- *Community prediction.* Community activity prediction is useful to many applications, e.g., recommendation systems and network performance enhancement [Rossetti et al. 2017]. In [Puranik and Narayanan 2017], two methods were proposed to analyze the dynamics of community structures. The study of [Hao et al. 2017] proposed an efficient algorithm of k-clique community detection using formal concept analysis (FCA). Experimental results show that the proposed algorithm has higher accuracy and lower computational cost than traditional methods.
- *Content analysis.* Social media contains a wide variety of data types from text, pictures, audio to video. As a result, the integration of multiple data analysis approaches together can improve the existing methods. In [Peng et al. 2017], the graph theory was used to investigate the influence of social content via a social relationship graph.
- *Human behavior study.* Understanding human behavior via mobile social networks is extremely valuable for service providers and governors. In [Xu et al. 2016], a study linking cyberspace and the physical world with social ecology via massive mobile cellular data is presented. Analytical results show that there are strong correlations between the mobile traffic and the human mobility; these results turn out to be related to social ecology.

7.3.4 Internet of Things. We can apply data analysis approaches to extract valuable information from Internet of Things. We enumerate several typical applications as follows.

- *Event detection.* Event detection is one of the most important functions of WSNs. Applying data mining or machine learning approaches to WSNs can help to design effective event detection mechanisms. In particular, anomaly detection plays an important role in ensuring the reliability of industrial systems. In [Chen et al. 2015], simulation results show that the distributed general anomaly detection scheme is scalable to large scale WSNs and outperforms other existing methods in terms of detection accuracy and efficiency. Moreover, the study of [Saeedi Emadi and Mazinani 2018] proposed a DBSCAN-based scheme and an SVM-based scheme to detect anomalies in WSNs. Furthermore, it is crucial to extract events in the context of RFID data management applications. The work of [Ding et al. 2017] proposed an RFID-enabled graphical deduction model to track objects with attritutes like time-sensitive state and position. Moreover, a machine learning based approach was proposed in [Dagli et al. 2015] to detect complex events in RFID systems.
- *Localization.* Localization is one of the most important functions of WSNs. Through localization, sensor nodes can determine the location of each other. There are two types of localization algorithms in WSNs: range-based algorithms and range-free algorithms. Range-based algorithms have relatively accurate localization while they require additional hardware (e.g., GPS) to obtain the distance information. Range-free algorithms require no

extral hardware support but suffer from the poor localization accuracy. In [Phoemphon et al. 2018], a novel range-free localization algorithm based on the integration of Fuzzy Logic (FL) and Extreme Learning Machines (ELMs) was proposed. Experimental results demonstrate the effectiveness of the proposed scheme. One of the most important issues in IoT is the localization of tags and smart objects. However, the localization for IoT is more challenging than that in WSNs due to the following reasons: (i) GPS devices do not work indoor where RFID tags are typically used; (ii) RFID tags and IoT objects have the very limited power supply; (iii) RFID tags and IoT objects have more limited computational capability than nodes in WSNs. To solve the challenge, there are many machine learning based approaches proposed in the literature. For example, [Kung et al. 2015] proposed a neural-network based RFID localization method, which uses the training data derived from both reference-tag coordinates and the coordinates generated by localization algorithms. Experimental results show that the proposed protocol can accurately locate critical objects.

- *Network optimization/security.* Due to the energy constraint of WSNs, how to design the network protocols to minimize the energy consumption is one of the challenges in WSNs. There are a number of efforts using machine learning methods to optimize the network performance of WSNs. For example, reinforcement learning based geographic routing algorithms were proposed in [Alsheikh et al. 2014]. Machine learning algorithms can also be used to improve the network security of IoT and WSNs. For example, [Huang et al. 2017] developed an efficient intrusion detection approach via learning traffic patterns. Experimental results show that this scheme has high detection accuracy. Moreover, machine learning approaches can also be used to design secure network protocol in WSNs[Ahad et al. 2016].
- *Consumer behaviour prediction.* Consumer behaviour prediction plays an important role in many business application. In [Zuo 2016], a Bayesian network based approach was proposed to predict the customer purchase behaviour; the analysis is based on massive RFID data collected through RFID tags attached at customers. In [Zheng et al. 2018b], a novel method based on deep convolutional neural networks (CNN) was proposed to identify electricity theft (i.e., a malicious consumer behaviour) in smart grids.

7.4 Opportunities in data analytics

Although many challenging issues as mentioned in Section 7.1 have been partially or fully addressed, there are still many issues not well addressed. We just enumerate some of research opportunities as follows:

- *Energy-efficiency and time-sensitiveness.* For example, most of previous studies focus on the accuracy of data analytics. Many efforts have been taken on improving the analysis accuracy or reducing the analysis error. Few studies consider the practical issues such as energy-efficiency and time-sensitiveness when data analytics schemes are deployed at wireless nodes.
- *Privacy preservation in data analytics.* Due to the limited computational capability of some wireless nodes, many data analytics tasks are conducted at remote clouds, which are however owned by third parties. Many studies often ignore the key step in removing privacy-sensitive attributes from the collected data and directly upload the data to remote clouds.
- *Security assurance in data analytics.* Although machine learning algorithms have shown their strength in data analytics in massive data, they are also suffering from vulnerabilities to malicious attacks. For example, it is shown in [Wang and Gong 2018] that hyper parameters in machine learning models can be stolen so as to breach proprietary rights and divulge confidential information. Moreover, machine learning models are

Table 8. Summary of solutions to challenges in BDA for large scale wireless networks

Challenges	**Mobile Communication Networks**	**Vehicular Networks**	**Mobile Social Networks**	**Internet of Things**
Data Acquisition: · Data representation · Data collection · Data transmission	[Imran et al. 2014] [He et al. 2016] [Biral et al. 2015] [Su and Xu 2015] [Wang et al. 2014] [Checko et al. 2015] [Fan et al. 2016b] [Mohajer et al. 2017] [Kuang et al. 2016] [Mehmeti and Spyropoulos 2017]	[Ilarri et al. 2015] [Mansour and Moussaoui 2015] [Brik et al. 2016] [Płaczek 2017] [Sahoo et al. 2017]	[Richthammer et al. 2014] [Leskovec and Sosič 2016] [Hajarian et al. 2017]	[Kennedy et al. 2017] [Fahmy 2016] [Chen et al. 2017] [Xu et al. 2017] [Mekki et al. 2018] [Iqbal et al. 2018]
Data Preprocessing: · Integration · Duplication reduction · Cleaning and compression	[Gubanov 2017] [Zhang et al. 2002] [Sanctis et al. 2016] [Fan et al. 2016a]	[Lenzerini 2002] [Haas et al. 2002] [Han et al. 2012] [Yu et al. 2010] [Feldman et al. 2012] [Fogue et al. 2014]	[Lenzerini 2002][Haas et al. 2002] [Aggarwal and Abdelzaher 2011] [Zheng et al. 2010] [Mehrotra et al. 2014] [Jorgensen et al. 2018]	[Ertek et al. 2017] [Baba et al. 2017] [Bhandari et al. 2017] [Ma et al. 2018] [Tasnim et al. 2017] [Deng et al. 2018] [Zheng et al. 2018b] [AlemÃąn et al. 2018]
Data storage: · Reliability & Persistency · Scalability · Efficiency	[Weatherspoon and Kubiatowicz 2002] [Chen et al. 2015] [Su and Xu 2015] [Al-Awami and Hassanein 2017]	[Sathiamoorthy et al. 2014] [Li et al. 2015]	[Cheng et al. 2015] [Su et al. 2016] [Hu et al. 2017]	[Guerra et al. 2011] [Chen et al. 2016] [Jun et al. 2016] [Sharma and Wang 2017]
Data Analytics: · Temporal-spatial correlation · Efficiency · Privacy · Real-time	[El Khayat et al. 2010] [Zhang et al. 2015] [Joung 2016] [Kolias et al. 2016] [Gai et al. 2016][Xia et al. 2017]	[Kang and Kang 2016] [Scalabrin et al. 2017] [Yu et al. 2016] [Ko et al. 2016] [Polson and Sokolov 2017] [Zhao et al. 2017] [Ye et al. 2018]	[Xu et al. 2016] [Rossetti et al. 2017] [Puranik and Narayanan 2017] [Hao et al. 2017] [Peng et al. 2017]	[Chen et al. 2015] [Zuo 2016] [Ding et al. 2017] [Huang et al. 2017] [Phoemphon et al. 2018] [Saeedi Emadi and Mazinani 2018] [Zheng et al. 2018b]

vulnerable to malicious attacks such as trojaning attack [Liu et al. 2018] and poisoning attack [Jagielski et al. 2018]. Hence, security countermeasures to remedy these vulnerabilities are expected.

8 FUTURE RESEARCH DIRECTIONS

Table 8 summarizes the solutions to challenges in big data analytics for large scale wireless networks in the aspects of data acquisition, data preprocessing, data storage and data analytics. Although many research challenges have been solved, there are still many research issues to be solved. We next discuss the future directions in big data analytics for large scale wireless networks.

Fig. 14 shows an overview on future directions in big data analytics for large scale wireless networks.

8.1 Distributed data processing models

Although a lot of efforts are done in developing distributed data processing models for large scale wireless networks, there are still many open research issues in this area.

- *Stream data processing*. Due to the tremendous volume of real-time data (e.g., sensor data from WSNs), it is impossible to store and process the entire data in memory. As a result, many conventional data analysis algorithms that require accessing the whole data sets do not work in this scenario. As mentioned in Section 5.2.1, some data stream preprocessing schemes [Aggarwal 2006] can be used in mobile communication networks and wireless sensor networks. However, there are few studies on data analysis of data streams as far as we know [Krempl 2014].
- *In-network processing*. Since there are a large number of wireless nodes distributed in large scale wireless networks, the integration of the data generated from distributed nodes is necessary for data processing. However, the data fusion among the distributed networks inevitably causes significant communication cost. One of the

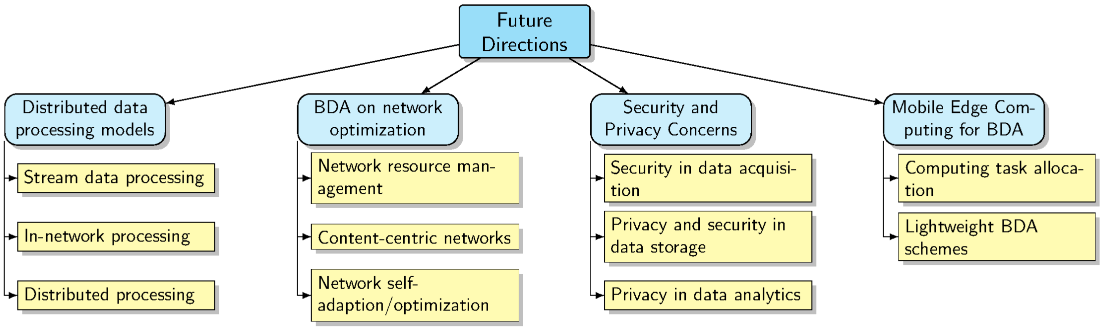

Future Directions
Distributed data processing models
Stream data processing
In-network processing
Distributed processing
BDA on network optimization
Network resource management
Content-centric networks
Network self-adaption/optimization
Security and Privacy Concerns
Security in data acquisition
Privacy and security in data storage
Privacy in data analytics
Mobile Edge Computing for BDA
Computing task allocation
Lightweight BDA schemes

Fig. 14. Future research directions for BDA of wireless networks

solutions to this challenge is to conduct in-network processing among the whole network, in which data is processing at each node instead of at centralized servers. To further reduce the computational cost, it is usually advantageous to organize the nodes into clusters [Diallo et al. 2015]. However, how to choose the cluster to fulfill the big data requirement becomes a new challenge.

- *Distributed processing*. Among the existing distributed data processing models, MapReduce [Dean and Ghemawat 2008] and its alternatives have many advantages, such as simple, fault tolerant and scalable. However, one of its limitations lies in the inefficiency compared with other parallel processing models. Conventional parallel computing models, such as Message Passing Interface (MPI), OpenMP (Open Multi-Processing), FPGA-oriented programming and GPU-based parallel processing (e.g., NVIDIA's CUDA), have the higher performance than MapReduce-like computing models. To integrate MapReduce-like models with parallel computing models can potentially improve the performance further. Besides, by exploiting the benefits of some dedicated processing platforms, we can further improve the existing data analysis algorithms. For example, there are some efforts in FPGA-oriented Convolutional neural network (CNN) [Zhang et al. 2015] and FPGA deep learning algorithm [Lacey et al. 2016], which have the better performance than those implemented in common PCs.

8.2 Big data analytics on network optimization

BDA can also help to optimize the network design of large scale wireless networks. We enumerate some of open issues on the impacts of BDA on network optimization as follows.

- *Network resource management.* Based on BDA of network data, network administrators can predict the network resource demands. For an example as illustrated in [Zheng et al. 2016], we can easily predict that there potentially exists a congested traffic in some locations of a city when a social event like a marathon takes place. Therefore, mobile operators can allocate more radio resources (i.e., more spectrum) to the hotspot so that the peak traffic can be absorbed smoothly [Jiang et al. 2017; Zheng et al. 2016].
- *Content-centric networks.* As suggested in many previous works [Bi et al. 2015b; Su and Xu 2015; Su et al. 2016], to store some popular contents (also named caches) at base stations can significantly reduce the real-time traffic and consequently improve the network performance. However, how to determine the cache becomes a challenge. Essentially, we can acquire cache information through analyzing the application data. However, it

is quite challenging to obtain the accurate user information due to the privacy preservation of user application data and the heterogeneous data types of various applications.

- *Network self-adaption/self-optimization.* BDA is extremely useful in network self-adaption or self-optimization in self-organizing networks (SONs) [Mohajer et al. 2017; Wang et al. 2015a]. For example, Fan et al. proposed a self-optimization method with integration of a fuzzy neural network (NN) and reinforcement learning; this method can fulfill the requirements of coverage and capacity of SONs. In summary, more efforts shall be conducted in this new area.

8.3 Security and Privacy Concerns

Security and privacy are important issues in BDA in wireless networks. Security and privacy are closely correlated while they are different from each other in the following aspects [Abouelmehdi et al. 2018]: 1) Security is to ensure the *confidentiality*, *integrity* and *availability* of data; 2) Privacy is to guarantee the proper usage of the data without the disclosure of user private information in the absence of user consent. We point out the future directions in security and privacy of BDA in wireless networks as follows.

- *Security in data acquisition.* During this phase, the wiretapping behavior can take place anywhere and consequently leads to the information leakage. Therefore, substantial efforts on protecting the confidential information of wireless networks are necessary. Usually, we can apply encryption schemes in wireless networks [Granjal et al. 2015]. However, it is infeasible to apply cryptography-based techniques in IoT due to the constraints of the energy and computational capability of smart objects [Sadeghi et al. 2015; Yang et al. 2017c]. Therefore, new lightweight protection schemes are expected to be developed for IoT [Liu et al. 2017b].
- *Privacy and security in data storage.* Once invasion on data storage system is successful, more personal confidential information can be disclosed. Thus, it is more critical to protect the stored data in this phase. Fortunately, it is easier to employ encryption algorithms to ensure security at data storage than that in data acquisition (transmission). However, it is still challenging to enforce the privacy-preserved operations in data storage [Wang et al. 2015b], especially when data storage service is offered by a third party [Lin et al. 2017]. Mobile Edge Computing (MEC) [Mach and Becvar 2017] can essentially offer a solution to the privacy-preservation in data storage by offloading the data from the untrusted third party to the trusted MEC server (deployed in proximity to the user).
- *Privacy in data analytics.* One of the major concerns of this phase lies in the balance between the privacy and the efficiency of data analysis [Lu et al. 2014]. For example, to protect the private user documents, usually the documents are encrypted and stored at a server (or a cloud). However, operations on the encrypted documents are time-consuming, which consequently lead to the in-efficiency in data analytics [Au et al. 2018]. There are still substantial efforts in the aspects of data publishing[Zhang et al. 2017], data mining output and distributed data privacy [Mendes and Vilela 2017] needed to be done.

8.4 Mobile Edge Computing for BDA

Due to inherent constraints such as limited power and inferior computational capability of wireless nodes, it is preferable to submit computing tasks from wireless nodes to remote cloud servers that have superior computational capability without resource constraints. However, cloud computing is also suffering from the limitations such as high latency, performance bottleneck, context unawareness and privacy exposure [Liu et al. 2017a]. MEC (or Fog computing) serves as a complement to cloud computing by overcoming the aforementioned limitations [Tran et al. 2017]. The main idea of MEC is to offload computing tasks from remote clouds to various MEC servers deployed at base stations, IoT gateways and WiFi APs in a proximity to end users. In this way, the computing-intensive and delay-tolerant tasks will be executed at remote cloud servers while the delay-critical, computing less-intensive and context-aware tasks can be offloaded to edge servers. However, there are many challenges in MEC for BDA of wireless networks.

- *Computing task allocation.* There are various computing resources in wireless networks such as supercomputers at remote clouds, edge (fog) servers and mobile devices. It is necessary to determine how to allocate the computation resources at different computing devices. However, it is challenging to allocate and coordinate various computing resources distributed in large scale networks.
- *Lightweight BDA schemes.* It is shown [Lin et al. 2018] that AlexNet (i.e., a typical convolutional neural network) has the model size of 240MB, consequently resulting in huge communication cost from the server to the edge node. The resource limitation of edge servers and mobile devices motivates the research in designing lightweight BDA schemes and compressing BDA models [Leng et al. 2018b]. It requires the efforts in hardware design, optimization, data compression, distributed computing and machine learning to achieve this goal [Cheng et al. 2018].

9 CONCLUSION

In this paper, we present a detailed survey on big data analytics (BDA) for large scale wireless networks. We first introduce the research methodology used in this paper. We then introduce data sources of several exemplary wireless networks including mobile communication networks, vehicular networks, mobile social networks, Internet of things. We next discuss the necessities and the challenges in BDA for large scale wireless networks. Based on our proposed four-stage life cycle of BDA for large scale wireless networks, we present a detailed survey on the existing solutions to the challenges in BDA. However, numerous research issues in this area are still open and need further efforts, such as improving distributed processing models, designing wireless networks with consideration of BDA and balancing the performance and privacy preservation trade-off in BDA.

ACKNOWLEDGEMENT

The work described in this paper was partially supported by Macao Science and Technology Development Fund under Grant No. 0026/2018/A1. The authors would like to thank Gordon K.-T. Hon for his constructive comments.

REFERENCES

2013. ISO/IEC 18000. (2013). http://en.wikipedia.org/wiki/ISO/IEC_18000

2017. *Cisco Visual Networking Index: Forecast and Methodology, 2016-2021.* Technical Report.

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.

G. Abdul-Salaam, A. Hanan Abdullah, and M. Hossein Anisi. 2017. Energy-Efficient Data Reporting for Navigation in Position-Free Hybrid Wireless Sensor Networks. *IEEE Sensors Journal* 17, 7 (2017), 2289–2297.

Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. 2018. Big healthcare data: preserving security and privacy. *Journal of Big Data* 5, 1 (Jan 2018), 1.

Charu C. Aggarwal. 2006. *Data Streams: Models and Algorithms (Advances in Database Systems).* Springer, Secaucus, NJ, USA.

Charu C. Aggarwal and Tarek Abdelzaher. 2011. *Integrating Sensors and Social Networks.* Springer.

Nauman Ahad, Junaid Qadir, and Nasir Ahsan. 2016. Neural networks in wireless networks: Techniques, applications and guidelines. *Journal of Network and Computer Applications* 68 (2016), 1 – 27.

Louai Al-Awami and Hossam S. Hassanein. 2017. Robust decentralized data storage and retrieval for wireless networks. *Computer Networks* 128 (2017), 41 – 50. DOI: https://doi.org/10.1016/j.comnet.2017.02.004 Survivability Strategies for Emerging Wireless Networks.

C. S. AlemÃąn, N. Pissinou, S. Alemany, K. Boroojeni, J. Miller, and Z. Ding. 2018. Context-Aware Data Cleaning for Mobile Wireless Sensor Networks: A Diversified Trust Approach. In *2018 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 226–230. DOI: https://doi.org/10.1109/ICCNC.2018.8390320

M. A. Alsheikh, Y. Jiao, D. Niyato, P. Wang, D. Leong, and Z. Han. 2017. The Accuracy-Privacy Trade-off of Mobile Crowdsensing. *IEEE Communications Magazine* 55, 6 (2017), 132–139.

M. A. Alsheikh, S. Lin, D. Niyato, and H. P. Tan. 2014. Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications. *IEEE Communications Surveys Tutorials* 16, 4 (2014), 1996–2018.

Sattam et al. Alsubaiee. 2014. AsterixDB: A Scalable, Open Source BDMS. *Proc. VLDB Endow.* 7, 14 (Oct. 2014), 1905–1916.

Peter Alvaro, Tyson Condie, Neil Conway, Khaled Elmeleegy, Joseph M. Hellerstein, and Russell Sears. 2010. Boom Analytics: Exploring Data-centric, Declarative Programming for the Cloud. In *Proceedings of the 5th European Conference on Computer Systems (EuroSys)*. ACM.

Flora Amato, Vincenzo Moscato, Antonio Picariello, and Francesco Piccialli. 2019. SOS: A multimedia recommender System for Online Social networks. *Future Generation Computer Systems* 93 (2019), 914 – 923. DOI: https://doi.org/10.1016/j.future.2017.04.028

J. Chris Anderson, Jan Lehnardt, and Noah Slater. 2010. *CouchDB: The Definitive Guide Time to Relax* (1st ed.). O'Reilly Media, Inc.

Apache. 2014. Hadoop MapReduce. (2014). https://hadoop.apache.org/

Apache. 2016. Apache HBase. (2016). https://hbase.apache.org/

Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau. 2015. *Operating Systems: Three Easy Pieces* (0.91 ed.). Arpaci-Dusseau Books.

Farzindar Atefeh and Wael Khreich. 2015. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence* 31, 1 (2015), 132–164.

Man Ho Au, Kaitai Liang, Joseph K. Liu, Rongxing Lu, and Jianting Ning. 2018. Privacy-preserving personal data operation on mobile cloud – Chances and challenges over advanced persistent threat. *Future Generation Computer Systems* 79 (2018), 337 – 349.

M. Ayaz, M. Ammad-uddin, I. Baig, and e. H. M. Aggoune. 2018. Wireless Sensor's Civil Applications, Prototypes, and Future Integration Possibilities: A Review. *IEEE Sensors Journal* 18, 1 (2018), 4–30.

Asif Iqbal Baba, Hua Lu, Torben Bach Pedersen, and Manfred Jaeger. 2017. Cleansing Indoor RFID Tracking Data. *SIGSPATIAL Special* 9, 1 (July 2017), 11–18.

X. Bai, Z. Wang, L. Sheng, and Z. Wang. 2018. Reliable Data Fusion of Hierarchical Wireless Sensor Networks With Asynchronous Measurement for Greenhouse Monitoring. *IEEE Transactions on Control Systems Technology* PP, 99 (2018), 1–11. DOI: https://doi.org/10.1109/TCST.2018.2797920

Prasanta S. Bandyopadhyay and Malcolm R. Forster. 2011. *Philosophy of Statistics*. Elsevier.

Y. Bao, X. Wang, Z. Wang, C. Wu, and F. C. M. Lau. 2016. Online influence maximization in non-stationary Social Networks. In *IEEE/ACM 24th International Symposium on Quality of Service (IWQoS)*. IEEE, 1–6.

S. Bassoy, H. Farooq, M. A. Imran, and A. Imran. 2017. Coordinated Multi-Point Clustering Schemes: A Survey. *IEEE Communications Surveys Tutorials* 19, 2 (2017), 743–764.

Dominic Battré, Stephan Ewen, Fabian Hueske, Odej Kao, Volker Markl, and Daniel Warneke. 2010. Nephele/PACTs: A Programming Model and Execution Framework for Web-scale Analytical Processing (SoCC). In *Proceedings of the 1st ACM Symposium on Cloud Computing*. ACM.

Sibghat Ullah Bazai, Julian Jang-Jaccard, and Xuyun Zhang. 2017. A Privacy Preserving Platform for MapReduce. In *Applications and Techniques in Information Security*. Springer, Singapore, 88–99.

Doug Beaver, Sanjeev Kumar, Harry C. Li, Jason Sobel, and Peter Vajgel. 2010. Finding a Needle in Haystack: Facebook's Photo Storage. In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation (OSDI)*. USENIX Associatation.

Siddhartha Bhandari, Neil Bergmann, Raja Jurdak, and Branislav Kusy. 2017. Time Series Data Analysis of Wireless Sensor Network Measurements of Temperature. *Sensors* 17, 6 (2017).

Sourjya et al. Bhaumik. 2012. CloudIQ: A Framework for Processing Base Stations in a Data Center. In *Proceedings of ACM MobiCom*. ACM.

S. Bi, C. K. Ho, and R. Zhang. 2015a. Wireless powered communication: opportunities and challenges. *IEEE Comm. Magazine* 53, 4 (2015), 117–125.

S. Bi, R. Zhang, Z. Ding, and S. Cui. 2015b. Wireless communications in the era of big data. *IEEE Communications Magazine* 53, 10 (October 2015), 190–199.

C. Bila, F. Sivrikaya, M. A. Khan, and S. Albayrak. 2017. Vehicles of the Future: A Survey of Research on Safety Issues. *IEEE Transactions on Intelligent Transportation Systems* 18, 5 (2017), 1046–1065.

Andrea Biral, Marco Centenaro, Andrea Zanella, Lorenzo Vangelista, and Michele Zorzi. 2015. The challenges of M2M massive access in wireless cellular networks. *Digital Communications and Networks* 1, 1 (2015), 1 – 19. DOI: https://doi.org/10.1016/j.dcan.2015.02.001

A. Boubrima, W. Bechkit, and H. Rivano. 2017. Optimal WSN Deployment Models for Air Pollution Monitoring. *IEEE Transactions on Wireless Communications* 16, 5 (2017), 2723–2735.

Bouziane Brik, Nasreddine Lagraa, Abderrahmane Lakas, and Abbas Cheddad. 2016. DDGP: Distributed Data Gathering Protocol for vehicular networks. *Vehicular Communications* 4 (2016), 15 – 29.

Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D. Ernst. 2010. HaLoop: Efficient Iterative Data Processing on Large Clusters. *Proc. VLDB Endow.* 3, 1-2 (Sept. 2010).

A. L. Buczak and E. Guven. 2016. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys Tutorials* 18, 2 (Secondquarter 2016), 1153–1176.

Rubén Casado and Muhammad Younas. 2015. Emerging Trends and Technologies in Big Data Processing. *Concurr. Comput. : Pract. Exper.* 27, 8 (2015), 2078–2091.

Prabhakar Chaganti and Rich Helms. 2010. *Amazon SimpleDB Developer Guide* (1st ed.). Packt Publishing.

Ronnie Chaiken, Bob Jenkins, Per-Å ke Larson, Bill Ramsey, Darren Shakib, Simon Weaver, and Jingren Zhou. 2008. SCOPE: Easy and Efficient Parallel Processing of Massive Data Sets. *Proc. VLDB Endow.* 1, 2 (Aug. 2008), 1265–1276.

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. 2008. Bigtable: A Distributed Storage System for Structured Data. *ACM Trans. Comput. Syst.* 26, 2 (June 2008).

A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann. 2015. Cloud RAN for Mobile Networks - A Technology Overview. *IEEE Communications Surveys Tutorials* 17, 1 (First 2015), 405–426.

C. A. Chen, M. Won, R. Stoleru, and G. G. Xie. 2015. Energy-Efficient Fault-Tolerant Data Storage and Processing in Mobile Cloud. *IEEE Transactions on Cloud Computing* 3, 1 (2015), 28–41.

F. Chen, T. Xiang, Y. Yang, and S. S. M. Chow. 2016. Secure Cloud Storage Meets with Secure Network Coding. *IEEE Trans. Comput.* 65, 6 (2016), 1936–1948.

Min Chen, Shiwen Mao, and Yunhao Liu. 2014. Big data: A survey. *Mobile networks and applications* 19, 2 (2014), 171–209.

P. Y. Chen, S. Yang, and J. A. McCann. 2015. Distributed Real-Time Anomaly Detection in Networked Industrial Sensing Systems. *IEEE Transactions on Industrial Electronics* 62, 6 (2015), 3832–3842.

Zhuangbin Chen, Anfeng Liu, Zhetao Li, Young-June Choi, Hiroo Sekiya, and Jie Li. 2017. Energy-Efficient Broadcasting Scheme for Smart Industrial Wireless Sensor Networks. *Mobile Information Systems* (2017), 1–17. DOI: https://doi.org/doi:10.1155/2017/7538190

Yue Cheng, M. Safdar Iqbal, Aayush Gupta, and Ali R. Butt. 2015. CAST: Tiering Storage for Data Analytics in the Cloud. In *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*. ACM.

Y. Cheng, D. Wang, P. Zhou, and T. Zhang. 2018. Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges. *IEEE Signal Processing Magazine* 35, 1 (Jan 2018), 126–136.

Kristina Chodorow and Michael Dirolf. 2010. *MongoDB: The Definitive Guide* (1st ed.). O'Reilly Media, Inc.

C. Cooper, D. Franklin, M. Ros, F. Safaei, and M. Abolhasan. 2017. A Comparative Survey of VANET Clustering Techniques. *IEEE Communications Surveys Tutorials* 19, 1 (2017), 657–681.

James C. et al. Corbett. 2013. Spanner: Google's Globally Distributed Database. *ACM Trans. Comput. Syst.* 31, 3 (Aug. 2013), 8:1–8:22.

L. Cui, F. R. Yu, and Q. Yan. 2016. When big data meets software-defined networking: SDN for big data and big data for SDN. *IEEE Network* 30, 1 (January 2016), 58–65.

Cihan H. Dagli, Nijat Mehdiyev, Julian Krumeich, David Enke, Dirk Werth, and Peter Loos. 2015. Determination of Rule Patterns in Complex Event Processing Using Machine Learning Techniques. *Procedia Computer Science* 61 (2015), 395 – 401.

Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51, 1 (Jan. 2008), 107–113.

Giuseppe et al. DeCandia. 2007. Dynamo: Amazon's Highly Available Key-value Store. In *Proceedings of ACM SOSP*. ACM.

Changyi Deng, Ruifeng Guo, Chao Liu, Ray Y. Zhong, and Xun Xu. 2018. Data cleansing for energy-saving: a case of Cyber-Physical Machine Tools health monitoring system. *International Journal of Production Research* 56, 1-2 (2018), 1000–1015.

O. Diallo, J. J. P. C. Rodrigues, M. Sene, and J. Lloret. 2015. Distributed Database Management Techniques for Wireless Sensor Networks. *IEEE Transactions on Parallel and Distributed Systems* 26, 2 (2015), 604–620.

T. Diekmann, A. Melski, and M. Schumann. 2007. Data-on-Network vs. Data-on-Tag: Managing Data in Complex RFID Environments. In *Proc. of 40th Annual Hawaii International Conference on System Sciences*. IEEE.

K. Ding, P. Jiang, P. Sun, and C. Wang. 2017. RFID-Enabled Physical Object Tracking in Process Flow Based on an Enhanced Graphical Deduction Modeling Method. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 11 (Nov 2017), 3006–3018.

Abdallah Jamal Dweekat, Gyusun Hwang, and Jinwoo Park. 2017. A supply chain performance measurement approach using the internet of things: Toward more practical SCPMS. *Industrial Management & Data Systems* 117, 2 (2017), 267–286.

Jaliya Ekanayake, Hui Li, Bingjing Zhang, Thilina Gunarathne, Seung-Hee Bae, Judy Qiu, and Geoffrey Fox. 2010. Twister: A Runtime for Iterative MapReduce. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC)*. ACM.

Ibtissam El Khayat, Pierre Geurts, and Guy Leduc. 2010. Enhancement of TCP over wired/wireless networks with packet loss classifiers inferred by supervised learning. *Wireless Networks* 16, 2 (2010), 273–290.

E. Elnikety, T. Elsayed, and H. E. Ramadan. 2011. iHadoop: Asynchronous Iterations for MapReduce. In *IEEE CloudCom*. IEEE.

G. Ertek, X. Chi, and A. N. Zhang. 2017. A Framework for Mining RFID Data From Schedule-Based Systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 11 (2017), 2967–2984.

Jason Baker et al. 2011. Megastore: Providing Scalable, Highly Available Storage for Interactive Services. In *Proceedings of the Conference on Innovative Data system Research (CIDR)*. www.cidrdb.org.

Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani. 2017. State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems. *IEEE Communications Surveys Tutorials* 19, 4 (Fourthquarter 2017), 2432–2455.

Hossam Mahmoud Ahmad Fahmy. 2016. *Wireless Sensor Networks: Concepts, Applications, Experimentation and Analysis*. Springer.

B. Fan, S. Leng, and K. Yang. 2016b. A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks. *IEEE Network* 30, 1 (January 2016), 6–10.

Y. C. Fan, Y. C. Chen, K. C. Tung, K. C. Wu, and A. L. P. Chen. 2016a. A Framework for Enabling User Preference Profiling through Wi-Fi Logs. *IEEE Transactions on Knowledge and Data Engineering* 28, 3 (2016), 592–603.

Z. Fei, B. Li, S. Yang, C. Xing, H. Chen, and L. Hanzo. 2017. A Survey of Multi-Objective Optimization in Wireless Sensor Networks: Metrics, Algorithms, and Open Problems. *IEEE Communications Surveys Tutorials* 19, 1 (2017), 550–586.

Dan Feldman, Andrew Sugaya, and Daniela Rus. 2012. An Effective Coreset Compression Algorithm for Large Scale Sensor Networks. In *Proceedings of the 11th International Conference on Information Processing in Sensor Networks (IPSN)*. ACM.

M. Fogue, P. Garrido, F. J. Martinez, J. C. Cano, C. T. Calafate, and P. Manzoni. 2014. A System for Automatic Notification and Severity Estimation of Automotive Accidents. *IEEE Transactions on Mobile Computing* 13, 5 (2014), 948–963.

V. Furtado, E. Furtado, C. Caminha, A. Lopes, V. Dantas, C. Ponte, and S. Cavalcante. 2017. A data-driven approach to help understanding the preferences of public transport users. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 1926–1935.

Keke Gai, Meikang Qiu, Lixin Tao, and Yongxin Zhu. 2016. Intrusion detection techniques for mobile cloud computing in heterogeneous 5G. *Security and Communication Networks* 9, 16 (2016), 3049–3058.

Sorabh Gandhi, Suman Nath, Subhash Suri, and Jie Liu. 2009. GAMPS: Compressing Multi Sensor Data by Grouping and Amplitude Scaling. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. ACM.

Inga Gerlach, Volker C. Hass, and Carl-Fredrik Mandenius. 2015. Conceptual Design of an Operator Training Simulator for a Bio-Ethanol Plant. *Processes* 3, 3 (2015), 664–683.

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google File System. In *Proceedings of ACM SOSP*. ACM.

Prasanna Giridhar, Md Tanvir Amin, Tarek Abdelzaher, Dong Wang, Lance Kaplan, Jemin George, and Raghu Ganti. 2016. ClariSense+: An enhanced traffic anomaly explanation service using social network feeds. *Pervasive and Mobile Computing* 33 (2016), 140 – 155.

K. Goda and M. Kitsuregawa. 2012. The History of Storage Systems. *Proc. IEEE* 100, Special Centennial Issue (May 2012), 1433–1440.

J. Granjal, E. Monteiro, and J. Silva. 2015. Security for the Internet of Things: A Survey of Existing Protocols and Open Research issues. *IEEE Communications Surveys Tutorials* 17, 3 (2015), 1294 – 1312.

M. Gubanov. 2017. PolyFuse: A Large-Scale Hybrid Data Fusion System. In *IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 1575–1578.

Jorge Guerra, Himabindu Pucha, Joseph Glider, Wendy Belluomini, and Raju Rangaswami. 2011. Cost Effective Storage Using Extent Based Dynamic Tiering. In *Proceedings of the 9th USENIX Conference on File and Stroage Technologies (FAST)*. USENIX Association.

L. M. Haas, E. T. Lin, and M. A. Roth. 2002. Data integration through database federation. *IBM Systems Journal* 41, 4 (2002), 578–596.

Mohammad Hajarian, Azam Bastanfard, Javad Mohammadzadeh, and Madjid Khalilian. 2017. Introducing fuzzy like in social networks and its effects on advertising profits and human behavior. *Computers in Human Behavior* 77 (2017), 282 – 293.

G. Han, L. Liu, S. Chan, R. Yu, and Y. Yang. 2017. HySense: A Hybrid Mobile CrowdSensing Framework for Sensing Opportunities Compensation under Dynamic Coverage Constraint. *IEEE Communications Magazine* 55, 3 (March 2017), 93–99.

Jiawei Han, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques* (third edition ed.). Morgan Kaufmann, Boston, USA.

F. Hao, G. Min, Z. Pei, D. S. Park, and L. T. Yang. 2017. K-Clique Community Detection in Social Networks Based on Formal Concept Analysis. *IEEE Systems Journal* 11, 1 (March 2017), 250–259.

Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu. 2016. Big Data Analytics in Mobile Cellular Networks. *IEEE Access* 4 (2016), 1985–1996.

Martin Hilbert. 2016. Big Data for Development: A Review of Promises and Challenges. *Development Policy Review* 34, 1 (2016), 135–174.

Desislava Hristova, Matthew J. Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo. 2016. Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*. ACM.

H. Hu, Y. Wen, T. S. Chua, and X. Li. 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access* 2 (2014), 652–687.

H. Hu, Y. Wen, and D. Niyato. 2017. Public Cloud Storage-Assisted Mobile Social Video Sharing: A Supermodular Game Approach. *IEEE Journal on Selected Areas in Communications* 35, 3 (March 2017), 545–556.

L. Hu, H. Wen, B. Wu, F. Pan, R. Liao, H. Song, J. Tang, and X. Wang. 2018. Cooperative Jamming for Physical Layer Security Enhancement in Internet of Things. *IEEE Internet of Things Journal* 5, 1 (Feb 2018), 219–228. DOI: https://doi.org/10.1109/JIOT.2017.2778185

K. Huang, Q. Zhang, C. Zhou, N. Xiong, and Y. Qin. 2017. An Efficient Intrusion Detection Approach for Visual Sensor Networks Based on Traffic Pattern Learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 10 (2017), 2704–2713.

Felix Hupfeld, Toni Cortes, Björn Kolbeck, Jan Stender, Erich Focht, Matthias Hess, Jesus Malo, Jonathan Marti, and Eugenio Cesario. 2008. The XtreemFS Architecture&Mdash;a Case for Object-based File Systems in Grids. *Concurrency and Computation: Practice and Experience* 20, 17 (Dec. 2008), 2049–2060.

S. Ilarri, T. Delot, and R. Trillo-Lado. 2015. A Data Management Perspective on Vehicular Networks. *IEEE Communications Surveys Tutorials* 17, 4 (2015), 2420–2460.

A. Imran, A. Zoha, and A. Abu-Dayya. 2014. Challenges in 5G: how to empower SON with big data for enabling 5G. *IEEE Network* 28, 6 (Nov 2014), 27–33.

N. Iqbal, S. Al-Dharrab, A. Muqaibel, W. Mesbah, and G. StÃijber. 2018. Analysis of Wireless Seismic Data Acquisition Networks using Markov Chain Models. In *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 1–5. DOI: https://doi.org/10.1109/PIMRC.2018.8580821

Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly. 2007. Dryad: Distributed Data-parallel Programs from Sequential Building Blocks. *SIGOPS Oper. Syst. Rev.* 41, 3 (March 2007), 59–72.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In *Proceedings of 2018 IEEE Symposium on Security and Privacy, SP 2018*. IEEE, 19–35. DOI: https://doi.org/10.1109/SP.2018.00057

C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, and L. Hanzo. 2017. Machine Learning Paradigms for Next-Generation Wireless Networks. *IEEE Wireless Communications* 24, 2 (April 2017), 98–105.

D. Jiang, L. Huo, Z. Lv, H. Song, and W. Qin. 2018. A Joint Multi-Criteria Utility-Based Network Selection Approach for Vehicle-to-Infrastructure Networking. *IEEE Transactions on Intelligent Transportation Systems* (2018), 1–15. DOI: https://doi.org/10.1109/TITS.2017.2778939

Terrence D. Jorgensen, K. Jean Forney, Jeffrey A. Hall, and Steven M. Giles. 2018. Using modern methods for missing data analysis with the social relations model: A bridge to social network analysis. *Social Networks* 54 (2018), 26 – 40.

J. Joung. 2016. Machine Learning-Based Antenna Selection in Wireless Communications. *IEEE Communications Letters* 20, 11 (2016), 2241–2244.

Sang-Woo Jun, Ming Liu, Sungjin Lee, Jamey Hicks, John Ankcorn, Myron King, Shuotao Xu, and Arvind. 2016. BlueDBM: Distributed Flash Storage for Big Data Analytics. *ACM Trans. Comput. Syst.* 34, 3 (2016), 7:1–7:31.

Min-Joo Kang and Je-Won Kang. 2016. Intrusion detection system using deep neural network for in-vehicle network security. *PloS one* 11, 6 (2016), 1–17.

T. F. Kennedy, R. S. Provence, J. L. Broyan, P. W. Fink, P. H. Ngo, and L. D. Rodriguez. 2017. Topic models for RFID data modeling and localization. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 1438–1446.

E. J. Khatib, R. Barco, P. Munoz, I. D. La Bandera, and I. Serrano. 2016. Self-healing in mobile networks with big data. *IEEE Communications Magazine* 54, 1 (January 2016), 114–120.

Mirza Golam Kibria, Kien Nguyen, Gabriel Porto Villardi, Kentaro Ishizu, and Fumihide Kojima. 2018. Big Data Analytics and Artificial Intelligence in Next-Generation Wireless Networks. *IEEE Access* 6 (2018), 32328–32338.

P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza. 2017. A Survey of Machine Learning Techniques Applied to Self-Organizing Cellular Networks. *IEEE Communications Surveys Tutorials* 19, 4 (Fourthquarter 2017), 2392–2431.

Eunjeong Ko, Jinyoung Ahn, and Eun Yi Kim. 2016. 3D Markov Process for Traffic Flow Prediction in Real-Time. *Sensors* 16, 2 (2016).

Jannis Koch, Christian L. Staudt, Maximilian Vogel, and Henning Meyerhenke. 2016. An empirical comparison of Big Graph frameworks in the context of network analysis. *Social Network Analysis and Mining* 6, 1 (Sep 2016), 84.

A. Koesdwiady, R. Soua, and F. Karray. 2016. Improving Traffic Flow Prediction With Weather Information in Connected Cars: A Deep Learning Approach. *IEEE Transactions on Vehicular Technology* 65, 12 (2016), 9508–9517.

C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis. 2016. Intrusion Detection in 802.11 Networks: Empirical Evaluation of Threats and a Public Dataset. *IEEE Communications Surveys Tutorials* 18, 1 (Firstquarter 2016), 184–208.

Georg et al. Krempl. 2014. Open Challenges for Data Stream Mining Research. *SIGKDD Explor. Newsl.* 16, 1 (Sept. 2014), 1–10.

Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. 2013. *Handbook of monte carlo methods*. Vol. 706. John Wiley & Sons.

L. Kuang, L. T. Yang, X. Wang, P. Wang, and Y. Zhao. 2016. A tensor-based big data model for QoS improvement in software defined networks. *IEEE Network* 30, 1 (January 2016), 30–35.

Abhishek Kumar, Bikash Sah, Arvind R. Singh, Yan Deng, Xiangning He, Praveen Kumar, and R.C. Bansal. 2017. A review of multi criteria decision making (MCDM) towards sustainable renewable energy development. *Renewable and Sustainable Energy Reviews* 69 (2017), 596 – 609.

Hsu-Yang Kung, Sumalee Chaisit, and Nguyen Thi Mai Phuong. 2015. Optimization of an RFID location identification scheme based on the neural network. *International Journal of Communication Systems* 28, 4 (2015), 625–644.

Griffin Lacey, Graham W. Taylor, and Shawki Areibi. 2016. Deep Learning on FPGAs: Past, Present, and Future. *CoRR* abs/1602.04283 (2016). http://arxiv.org/abs/1602.04283

Wei Kuang Lai, Mei-Tso Lin, and Yu-Hsuan Yang. 2015. A Machine Learning System for Routing Decision-Making in Urban Vehicular Ad Hoc Networks. *International Journal of Distributed Sensor Networks* 11, 3 (2015).

Avinash Lakshman and Prashant Malik. 2009. Cassandra: A Structured Storage System on a P2P Network. In *Proceedings of SPAA*. ACM.

J. Landt. 2005. The history of RFID. *IEEE Potentials* 24, 4 (Oct 2005), 8–11.

Cong Leng, Hao Li, Shenghuo Zhu, and Rong Jin. 2018b. Extremely Low Bit Neural Network: Squeeze the Last Bit Out with ADMM. In *AAAI*. AAAI Press.

Kaijun Leng, Linbo Jin, Wen Shi, and Inneke Van Nieuwenhuyse. 2018a. Research on agricultural products supply chain inspection system based on internet of things. *Cluster Computing* (Feb 2018).

Maurizio Lenzerini. 2002. Data Integration: A Theoretical Perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*. ACM, 233–246.

Jure Leskovec and Rok Sosič. 2016. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 1 (2016), 1.

Ming Li, Deepak Ganesan, and Prashant Shenoy. 2009. PRESTO: Feedback-driven Data Management in Sensor Networks. *IEEE/ACM Trans. Netw.* 17, 4 (Aug. 2009), 1256–1269.

Zhichao Li, Ming Chen, Amanpreet Mukker, and Erez Zadok. 2015. On the Trade-Offs Among Performance, Energy, and Endurance in a Versatile Hybrid Drive. *ACM Trans. Storage* 11, 3 (2015), 13:1–13:27.

Zhenhua Li, Weiwei Wang, Tianyin Xu, Xin Zhong, Xiang-Yang Li, Yunhao Liu, Christo Wilson, and Ben Y. Zhao. 2016. Exploring Cross-Application Cellular Traffic Optimization with Baidu TrafficGuard. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. USENIX Association.

J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao. 2017. A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. *IEEE Internet of Things Journal* 4, 5 (2017), 1125–1142.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. 2018. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. In *International Conference on Learning Representations (ICLR)*. ICLR.

H. Liu, F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, and Y. Zhang. 2017a. Mobile Edge Cloud System: Architectures, Challenges, and Approaches. *IEEE Systems Journal* PP, 99 (2017), 1–14. DOI: https://doi.org/10.1109/JSYST.2017.2654119

H. Liu, T. Taniguchi, Y. Tanaka, K. Takenaka, and T. Bando. 2017. Visualization of Driving Behavior Based on Hidden Feature Extraction by Using Deep Learning. *IEEE Transactions on Intelligent Transportation Systems* 18, 9 (2017), 2477–2489.

Ruilin Liu, Hongzhang Liu, Daehan Kwak, Yong Xiang, Cristian Borcea, Badri Nath, and Liviu Iftode. 2016. Balanced traffic routing: Design, implementation, and evaluation. *Ad Hoc Networks* 37, Part 1 (2016), 14 – 28.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2018. Trojaning Attack on Neural Networks. In *Proceeindgs of Network and Distributed System Security Symposium, NDSS 2018*. Internet Society, 1–15. DOI: https://doi.org/10.14722/ndss.2018.23300

Z. Liu, X. Huang, Z. Hu, M. K. Khan, H. Seo, and L. Zhou. 2017b. On Emerging Family of Elliptic Curves to Secure Internet of Things: ECC Comes of Age. *IEEE Transactions on Dependable and Secure Computing* 14, 3 (2017), 237–248.

Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M. Hellerstein. 2012. Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. *Proc. VLDB Endow.* 5, 8 (April 2012), 716–727.

R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao. 2014. Toward efficient and privacy-preserving computing in big data era. *IEEE Network* 28, 4 (July 2014), 46–50.

Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo. 2017. Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. *IEEE Transactions on Industrial Informatics* 13, 4 (Aug 2017), 1891–1899.

Haishu Ma, Yi Wang, and Kesheng Wang. 2018. Automatic detection of false positive RFID readings using machine learning algorithms. *Expert Systems with Applications* 91 (2018), 442 – 451.

P. Mach and Z. Becvar. 2017. Mobile Edge Computing: A Survey on Architecture and Computation Offloading. *IEEE Communications Surveys Tutorials* 19, 3 (thirdquarter 2017), 1628–1656.

Z. MacHardy, A. Khan, K. Obana, and S. Iwashina. 2018. V2X Access Technologies: Regulation, Research, and Remaining Challenges. *IEEE Communications Surveys Tutorials* PP, 99 (2018), 1–20.

Grzegorz et al. Malewicz. 2010. Pregel: A System for Large-scale Graph Processing. In *Proceedings of ACM SIGMOD*. ACM.

Louiza Mansour and Samira Moussaoui. 2015. CDCP: Collaborative Data Collection Protocol in Vehicular Sensor Networks. *Wireless Personal Communications* 80, 1 (2015), 151–165.

Marshall Kirk McKusick and Sean Quinlan. 2009. GFS: Evolution on Fast-forward. *ACM Queue* 7, 7 (Aug. 2009), 10:10–10:20.

F. Mehmeti and T. Spyropoulos. 2017. Performance Analysis of Mobile Data Offloading in Heterogeneous Networks. *IEEE Transactions on Mobile Computing* 16, 2 (2017), 482–497.

Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S. Guizani. 2017. Internet-of-Things-Based Smart Cities: Recent Advances and Challenges. *IEEE Communications Magazine* 55, 9 (2017), 16–24.

Abhinav Mehrotra, Veljko Pejovic, and Mirco Musolesi. 2014. SenSocial: A Middleware for Integrating Online Social Networks and Mobile Sensing Data Streams. In *Proceedings of the 15th International Middleware Conference (Middleware)*. ACM.

Kais Mekki, Eddy Bajic, Frederic Chaxel, and Fernand Meyer. 2018. A comparative study of LPWAN technologies for large-scale IoT deployment. *ICT Express* (2018).

Vasileios A. Memos, Kostas E. Psannis, Yutaka Ishibashi, Byung-Gyu Kim, and B.B. Gupta. 2018. An Efficient Algorithm for Media-based Surveillance System (EAMSuS) in IoT Smart City Framework. *Future Generation Computer Systems* 83 (2018), 619 – 628.

R. Mendes and J. P. Vilela. 2017. Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access* 5 (2017), 10562–10582.

Amin Mohajer, Morteza Barari, and Houman Zarrabi. 2017. Big Data based Self-Optimization Networking: A Novel Approach Beyond Cognition. *Intelligent Automation & Soft Computing* 0, 0 (2017), 1–7. DOI: https://doi.org/10.1080/10798587.2017.1312893

A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda. 2017. Hybrid Beamforming for Massive MIMO: A Survey. *IEEE Communications Magazine* 55, 9 (2017), 134–141.

F. Montori, L. Bedogni, and L. Bononi. 2018. A Collaborative Internet of Things Architecture for Smart Cities and Environmental Monitoring. *IEEE Internet of Things Journal* 5, 2 (April 2018), 592–605.

Carl H. Mooney and John F. Roddick. 2013. Sequential Pattern Mining – Approaches and Algorithms. *ACM Comput. Surv.* 45, 2 (March 2013), 19:1–19:39.

Duc T. Nguyen and Jai E. Jung. 2017. Real-time event detection for online behavioral analysis of big social data. *Future Generation Computer Systems* 66 (2017), 137 – 145.

Luiz H. Nunes, Júlio C. Estrella, Alexandre N. Delbem, Charith Perera, and Stephan Reiff-Marganiec. 2016. The Effects of Relative Importance of User Constraints in Cloud of Things Resource Discovery: A Case Study. In *Proceedings of the 9th International Conference on Utility and Cloud Computing (UCC '16)*. ACM, 245–250.

David A. Patterson and John L. Hennessy. 2013. *Computer Organization and Design: The Hardware/Software Interface* (5th ed.). Morgan Kaufmann.

Sancheng Peng, Aimin Yang, Lihong Cao, Shui Yu, and Dongqing Xie. 2017. Social influence modeling using information theory in mobile social networks. *Information Sciences* 379 (2017), 146 – 159. DOI: https://doi.org/10.1016/j.ins.2016.08.023

Songyut Phoemphon, Chakchai So-In, and Tri Gia Nguyen. 2018. An enhanced wireless sensor network localization scheme for radio irregularity models using hybrid fuzzy deep extreme learning machines. *Wireless Networks* 24, 3 (Apr 2018), 799–819.

Martin Placek and Rajkumar Buyya. 2006. *A Taxonomy of Distributed Storage Systems*. Technical Report GRIDS-TR- 2006-11. Grid Computing and Distributed Systems Laboratory, The University of Melbourne, Australia.

Bartłomiej Płaczek. 2017. Efficient data collection for self-organising traffic signal systems based on vehicular sensor networks. *International Journal of Ad Hoc and Ubiquitous Computing* 26, 1 (2017), 56–69.

Nicholas G. Polson and Vadim O. Sokolov. 2017. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies* 79 (2017), 1 – 17.

Tejas Puranik and Lata Narayanan. 2017. Community Detection in Evolving Networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ACM, 385–390.

Lijun Qian, Jinkang Zhu, and Sihai Zhang. 2017. Survey of wireless big data. *Journal of Communications and Information Networks* 2, 1 (Mar 2017), 1–18.

Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing* 2016, 1 (2016), 1–16. DOI: https://doi.org/10.1186/s13634-016-0355-x

Saeed K Rahimi and Frank S Haug. 2010. *Distributed database management systems: A Practical Approach*. John Wiley & Sons.

P. Ram Mohan Rao, S. Murali Krishna, and A. P. Siva Kumar. 2018. Privacy preservation techniques in big data analytics: a survey. *Journal of Big Data* 5, 1 (22 Sep 2018), 33.

U. Raza, P. Kulkarni, and M. Sooriyabandara. 2017. Low Power Wide Area Networks: An Overview. *IEEE Communications Surveys Tutorials* 19, 2 (2017), 855–873.

C. Richthammer, M. Netter, M. Riesner, J. Sänger, and G. Pernul. 2014. Taxonomy of social network data types. *EURASIP Journal on Information Security* 2014, 1 (2014), 11.

Rodrigo Roman, Jianying Zhou, and Javier Lopez. 2013. On the Features and Challenges of Security and Privacy in Distributed Internet of Things. *Comput. Netw.* 57, 10 (July 2013), 2266–2279.

Giulio Rossetti, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. 2017. Tiles: an online algorithm for community discovery in dynamic social networks. *Machine Learning* 106, 8 (Aug 2017), 1213–1241.

Xin Ruan, Zhenyu Wu, Haining Wang, and Sushil Jajodia. 2016. Profiling online social behaviors for compromised account detection. *IEEE transactions on information forensics and security* 11, 1 (2016), 176–187.

Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach (3rd Edition)* (3 ed.). Prentice Hall.

Ahmad-Reza Sadeghi, Christian Wachsmann, and Michael Waidner. 2015. Security and Privacy Challenges in Industrial Internet of Things. In *Proceedings of the 52nd Annual Design Automation Conference (DAC)*. ACM.

Hossein Saeedi Emadi and Sayyed Majid Mazinani. 2018. A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks. *Wireless Personal Communications* 98, 2 (Jan 2018), 2025–2035.

J. Sahoo, S. Cherkaoui, A. Hafid, and P. K. Sahu. 2017. Dynamic Hierarchical Aggregation for Vehicular Sensing. *IEEE Transactions on Intelligent Transportation Systems* 18, 9 (2017), 2539–2556.

M. De Sanctis, I. Bisio, and G. Araniti. 2016. Data mining algorithms for communication networks control: concepts, survey and guidelines. *IEEE Network* 30, 1 (January 2016), 24–29.

M. Sathiamoorthy, A. G. Dimakis, B. Krishnamachari, and F. Bai. 2014. Distributed Storage Codes Reduce Latency in Vehicular Networks. *IEEE Transactions on Mobile Computing* 13, 9 (2014), 2016–2027.

M. Scalabrin, M. Gadaleta, R. Bonetto, and M. Rossi. 2017. A Bayesian forecasting and anomaly detection framework for vehicular monitoring networks. In *IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder. 2017. 5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice. *IEEE Journal on Selected Areas in Communications* 35, 6 (2017), 1201–1221.

S. K. Sharma and X. Wang. 2017. Live Data Analytics With Collaborative Edge and Cloud Processing in Wireless IoT Networks. *IEEE Access* 5 (2017), 4621–4635.

W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor. 2017. Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges. *IEEE Communications Magazine* 55, 10 (2017), 176–183.

Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. The Hadoop Distributed File System. In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. IEEE.

John A Stankovic. 2017. Research directions for cyber physical systems in wireless and mobile healthcare. *ACM Transactions on Cyber-Physical Systems* 1, 1 (2017), 1.

Stefan Stieglitz, Milad Mirbabaie, BjÃűrn Ross, and Christoph Neuberger. 2018. Social media analytics âĂŞ Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management* 39 (2018), 156 – 168.

Biljana L Risteska Stojkoska and Kire V Trivodaliev. 2017. A review of Internet of Things for smart home: Challenges and solutions. *Journal of Cleaner Production* 140 (2017), 1454–1464.

John D. Strunk. 2012. Hybrid Aggregates: Combining SSDs and HDDs in a Single Storage Pool. *SIGOPS Oper. Syst. Rev.* 46, 3 (2012).

Z. Su and Q. Xu. 2015. Content distribution over content centric mobile social networks in 5G. *IEEE Communications Magazine* 53, 6 (June 2015), 66–72.

Z. Su, Q. Xu, and Q. Qi. 2016. Big data in mobile social networks: a QoE-oriented framework. *IEEE Network* 30, 1 (January 2016), 52–57.

V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer. 2017. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc. IEEE* 105, 12 (2017), 2295–2329.

D. Takaishi, H. Nishiyama, N. Kato, and R. Miura. 2014. Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks. *IEEE Transactions on Emerging Topics in Computing* 2, 3 (Sept 2014), 388–397.

S. Tasnim, N. Pissinou, and S. S. Iyengar. 2017. A novel cleaning approach of environmental sensing data streams. In *2017 14th IEEE Annual Consumer Communications Networking Conference (CCNC)*. IEEE, 632–633.

A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy. 2010. Hive - a petabyte scale data warehouse using Hadoop. In *IEEE 26th International Conference on Data Engineering (ICDE)*. IEEE.

T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili. 2017. Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges. *IEEE Communications Magazine* 55, 4 (April 2017), 54–61.

Martin Treiber and Arne Kesting. 2013. *Traffic Flow Dynamics*. Springer.

William M.K. Trochim, Jim Donnelly, and Kanika Arora. 2016. *Research Methods The Essential Knowledge Base* (2nd ed.). Cengage Learning.

Binghui Wang and Neil Zhenqiang Gong. 2018. Stealing Hyperparameters in Machine Learning. In *Proceedings of 2018 IEEE Symposium on Security and Privacy, SP 2018*. IEEE, 36–52. DOI:https://doi.org/10.1109/SP.2018.00038

B. Wang, B. Li, and H. Li. 2015b. Panda: Public Auditing for Shared Data with Efficient User Revocation in the Cloud. *IEEE Transactions on Services Computing* 8, 1 (Jan 2015), 92–106.

F. Wang and J. Liu. 2011. Networked Wireless Sensor Data Collection: Issues, Challenges, and Approaches. *IEEE Communications Surveys Tutorials* 13, 4 (Fourth 2011), 673–687.

G. Wang, J. Xin, L. Chen, and Y. Liu. 2012. Energy-Efficient Reverse Skyline Query Processing over Wireless Sensor Networks. *IEEE Transactions on Knowledge and Data Engineering* 24, 7 (July 2012), 1259–1275.

Hao Wang, Ottar Osen, Guoyuan Li, Wei Li, Hong-Ning Dai, and Wei Zeng. 2015. Big Data and Industrial Internet of Things for the Maritime Industry in Northwestern Norway. In *IEEE Region 10 Conference (TENCON)*. IEEE.

Jiangtao Wang, Yasha Wang, Daqing Zhang, and Sumi Helal. 2018a. Energy Saving Techniques in Mobile Crowd Sensing: Current State and Future Opportunities. *IEEE Communications Magazine* (2018), 1–7. DOI:https://doi.org/10.1109/MCOM.2018.1700644

Ning Wang, Xiaokui Xiao, Yin Yang, Ta Duy Hoang, Hyejin Shin, Junbum Shin, and Ge Yu. 2018b. PrivTrie: Effective Frequent Term Discovery under Local Differential Privacy. In *IEEE International Conference on Data Engineering (ICDE)*. IEEE.

T. Wang, N. S. Nguyen, J. Wang, T. Li, X. Zhang, N. Mi, B. Zhao, and B. Sheng. 2018. RoVEr: Robust and Verifiable Erasure Code for Hadoop Distributed File Systems. In *27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–9.

X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung. 2014. Cache in the air: exploiting content caching and delivery techniques for 5G systems. *IEEE Communications Magazine* 52, 2 (February 2014), 131–139.

X. Wang, X. Li, and V. C. M. Leung. 2015a. Artificial Intelligence-Based Techniques for Emerging Heterogeneous Network: State of the Arts, Opportunities, and Challenges. *IEEE Access* 3 (2015), 1379–1391.

Z. Wang, L. Duan, and R. Zhang. 2016. Adaptively Directional Wireless Power Transfer for Large-Scale Sensor Networks. *IEEE Journal on Selected Areas in Communications* 34, 5 (May 2016), 1785–1800.

R. Want. 2006. An introduction to RFID technology. *IEEE Pervasive Computing* 5, 1 (Jan 2006), 25–33.

Hakim Weatherspoon and John Kubiatowicz. 2002. Erasure Coding Vs. Replication: A Quantitative Comparison. In *Revised Papers from the First International Workshop on Peer-to-Peer Systems (IPTPS)*. IEEE, 328–338.

Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. ACM, 38.

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. 2008. Top 10 Algorithms in Data Mining. *Knowl. Inf. Syst.* 14, 1 (2008), 1–37.

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. 2014. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering* 26, 1 (2014), 97–107.

Zhuoqun Xia, Zhenzhen Hu, and Junpeng Luo. 2017. UPTP Vehicle Trajectory Prediction Based on User Preference Under Complexity Environment. *Wireless Personal Communications* 97, 3 (2017), 4651–4665.

F. Xu, Y. Li, M. Chen, and S. Chen. 2016. Mobile cellular big data: linking cyberspace and the physical world with social ecology. *IEEE Network* 30, 3 (May 2016), 6–12.

J. Xu, J. Yao, L. Wang, Z. Ming, K. Wu, and L. Chen. 2017. Narrowband Internet of Things: Evolutions, Technologies and Open Issues. *IEEE Internet of Things Journal* PP, 99 (2017), 1–13.

Q. Xu, Z. Su, K. Zhang, P. Ren, and X. S. Shen. 2015. Epidemic Information Dissemination in Mobile Social Networks With Opportunistic Links. *IEEE Transactions on Emerging Topics in Computing* 3, 3 (Sept 2015), 399–409.

W. Xu, S. Jha, and W. Hu. 2019 (early access). LoRa-Key: Secure Key Generation System for LoRa-based Network. *IEEE Internet of Things Journal* (2019 (early access)), 1–10. DOI: https://doi.org/10.1109/JIOT.2018.2888553

W. Xu, Y. Xu, C. H. Lee, Z. Feng, P. Zhang, and J. Lin. 2018. Data-Cognition-Empowered Intelligent Wireless Networks: Data, Utilities, Cognition Brain, and Architecture. *IEEE Wireless Communications* 25, 1 (February 2018), 56–63.

G. Yang, S. He, Z. Shi, and J. Chen. 2017b. Promoting Cooperation by the Social Incentive Mechanism in Mobile Crowdsensing. *IEEE Communications Magazine* 55, 3 (March 2017), 86–92.

S. Yang, U. Adeel, Y. Tahir, and J. A. McCann. 2017a. Practical Opportunistic Data Collection in Wireless Sensor Networks with Mobile Sinks. *IEEE Transactions on Mobile Computing* 16, 5 (May 2017), 1420–1433.

Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao. 2017c. A Survey on Security and Privacy Issues in Internet-of-Things. *IEEE Internet of Things Journal* 4, 5 (Oct 2017), 1250–1258.

H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu, and M. Wu. 2018. Machine Learning for Vehicular Networks: Recent Advances and Application Examples. *IEEE Vehicular Technology Magazine* 13, 2 (2018), 94–101.

Bin Yu, Xiaolin Song, Feng Guan, Zhiming Yang, and Baozhen Yao. 2016. k-Nearest Neighbor Model for Multiple-Time-Step Prediction of Short-Term Traffic Condition. *Journal of Transportation Engineering* 142, 6 (2016), 1–10.

Xinghuo Yu and Yusheng Xue. 2016. Smart grids: A cyber–physical systems perspective. *Proc. IEEE* 104, 5 (2016), 1058–1070.

X. Yu, H. Zhao, L. Zhang, S. Wu, B. Krishnamachari, and V. O. K. Li. 2010. Cooperative Sensing and Compression in Vehicular Sensor Networks for Urban Monitoring. In *IEEE International Conference on Communications (ICC)*. IEEE.

Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster Computing with Working Sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud)*. USENIX Association.

Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong. 2015. Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks. In *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA)*. ACM.

J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu. 2015. Robust Network Traffic Classification. *IEEE/ACM Transactions on Networking* 23, 4 (2015), 1257–1270.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.* 42, 4 (Oct. 2017), 25:1–25:41.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and Redundancy Detection in Adaptive Filtering. In *Proceedings of ACM SIGIR*. ACM.

Yanfeng Zhang, Qixin Gao, Lixin Gao, and Cuirong Wang. 2012. iMapReduce: A Distributed Computing Framework for Iterative Computation. *Journal of Grid Computing* 10, 1 (2012), 47–68.

Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri. 2017a. Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data. *IEEE Systems Journal* 11, 1 (2017), 88–95.

Y. Zhang, L. Song, C. Jiang, N. H. Tran, Z. Dawy, and Z. Han. 2017b. A Social-Aware Framework for Efficient Information Dissemination in Wireless Ad Hoc Networks. *IEEE Communications Magazine* 55, 1 (January 2017), 174–179.

Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. 2017. LSTM network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems* 11, 2 (2017), 68–75.

K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang. 2016. Big data-driven optimization for mobile networks toward 5G. *IEEE Network* 30, 1 (January 2016), 44–51.

Vincent W. Zheng, Yu Zheng, Xing Xie, and Qiang Yang. 2010. Collaborative Location and Activity Recommendations with GPS History Data. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*. ACM.

Zibin Zheng, Shaoan Xie, Hong-Ning Dai, Xiangping Chen, and Huaimin Wang. 2018a. Blockchain Challenges and Opportunities: A Survey. *International Journal of Web and Grid Services* 14, 4 (2018), 352 – 375.

Z. Zheng, Y. Yang, X. Niu, H. N. Dai, and Y. Zhou. 2018b. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. *IEEE Transactions on Industrial Informatics* 14, 4 (2018), 1606–1615.

Ray Y. Zhong, Chen Xu, Chao Chen, and George Q. Huang. 2017. Big Data Analytics for Physical Internet-based intelligent manufacturing shop floors. *International Journal of Production Research* 55, 9 (2017), 2610–2621.

Guoxun Zhu, Ye Tian, Yuyong Zhou, and Rencai Dong. 2017. Technical configurations of the Internet of Things for environmental monitoring at large-scale coal-fired power plants. *International Journal of Sustainable Development & World Ecology* 24, 5 (2017), 450–455. DOI: https://doi.org/10.1080/13504509.2016.1273261

Y. Zhu, J. Lin, P. P. C. Lee, and Y. Xu. 2015. Boosting Degraded Reads in Heterogeneous Erasure-Coded Storage Systems. *IEEE Trans. Comput.* 64, 8 (2015), 2145–2157.

Paul Zikopoulos and Chris Eaton. 2011. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (1st ed.). McGraw-Hill Osborne Media.

Yi Zuo. 2016. Prediction of Consumer Purchase Behaviour Using Bayesian Network: An Operational Improvement and New Results Based on RFID Data. *Int. J. Knowl. Eng. Soft Data Paradigm.* 5, 2 (April 2016), 85–105.