Big Sensor Data Systems for Smart Cities

Li-Minn Ang *Senior Member, IEEE*, Kah Phooi Seng *Member, IEEE*, Adamu Murtala Zungeru *Member, IEEE,* and Gerald Ijemaru, *Member, IEEE*

***Abstract*— Recent advances in large-scale networked sensor technologies and the explosive growth in Big Data computing have made it possible for new application deployments in smart cities ecosystems. In this paper, we define big sensor data systems, and survey progress made in the development and applications of big sensor data research. We classify the existing research based on their characteristics and smart city layer challenges. Next, we discuss several applications for big sensor data systems, and explore the potential of large-scale networked sensor technologies for smart cities in the Big Data era. We conclude the paper by discussing future work directions highlighting some futuristic applications. The aim of this survey paper is to be useful for researchers to get insights into this important area, and motivate the development of practical solutions towards deployment in smart cities.**

***Index Terms*— Big data, networked sensors, survey, smart cities, sensor technologies**

I. Introduction

BIG data techniques are targeted towards solving system-level problems that cannot be solved by conventional methods and technologies. With the emergence of new networked sensor technologies including large scale wireless sensor networks, body sensor networks, Internet/Network/ Vehicle/Web-of-Things, the next generation of big data systems will need to deal with machine-generated data from these forms of networked sensor systems. This is becoming an important research area. An estimation from IBM is that the volume of machine-generated data sources will increase to 42% of all data by 2020, up from 11% in 2005 [1]. Much less research has been conducted for big sensor data systems compared with conventional big data systems. Also, the term "big sensor data" is less used compared to the term "big data". For example, on the IEEE Xplore database, a search for the phrase "big sensor data" returned ten papers when compared with the phrase "big data" which returned several thousands of papers. In general, researchers have referred to big sensor data as the high volumes of data being generated from wireless sensor networks [2],[3]. In this paper, we will use a wider definition for a big sensor data framework as referring to the data sources from all kinds of networked sensing technologies.

Fig. 1 shows the data sources for a big sensor data framework encompassing three domains: (1) Sensor/Things Domain, (2) Internet/Communications Domain, and (3) People Domain. The following gives a brief summary of the various domains. The conventional big data model using stored data faces challenges within the Internet/Communications domain. Useful technologies to resolve this challenge is to employ distributed processing and storage techniques (e.g. Hadoop, MapReduce) or cloud computing technologies. The big sensor data framework would need to include two other components: Sensors/Things and People/Creature. The authors in [4] give a term of "technical sensors" to any sensor type which is not human generated. This is divided into two categories: (a) in-situ sensors which measure data in the immediate surroundings (e.g. ground-based environmental sensors, RFID), and (b) remote sensors which measure data from a distance (e.g. satellite-based sensing, LiDAR, hyperspectral, UAV). Mobile sensing can be achieved by embedding in-situ sensors onto mobile platforms (e.g. bicycles, buses, mobile robots).

The Internet/Communications Domain contains the network and communication infrastructure to interconnect the various sensors and devices. This includes the Internet, mobile and fixed networks. The intersection of the Internet/ Communications Domain with the Sensor/Things Domain results in technologies like wireless sensor networks (WSNs) and the Internet-of-Things (IoT). The boundary for the IoT does not include the People/Creature Domain. Historical data sources and other forms of stored data sets (text-based, audio-based, visual-based, etc.) would be included in this domain. The People/Creature Domain is an important part of the big sensor data framework. People can generate data voluntarily or as a by-product of another activity. Examples of the former case include using social media platforms (e.g. Twitter, Instagram, Flickr) and wearable body sensors to consciously share data. This is given the term of "People as sensors". Examples of the latter case include involuntarily releasing data (e.g. geolocation) as mobile phone calls/text messages are made. This is termed as collective sensing or crowd sensing. Sensors attached to animals/creatures for sensing and tracking [5],[6] would also be classified into this domain.

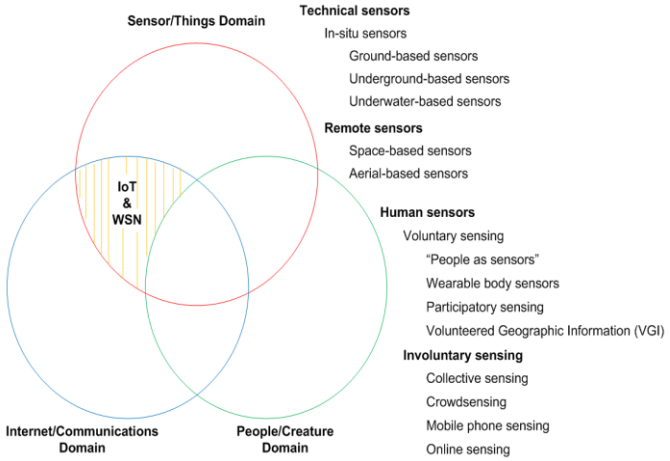

Fig. 1. Big sensor data framework showing data sources encompassing three domains: (1) Internet/Communications, (2) Sensor/Things, and (3) People/Creature.

Recently, there has been much interests for developing smart cities in many countries. Many smart cities have been established in the world such as in Santander [7], Barcelona [8], and Singapore [9]. In Australia, the state of Tasmania is experimenting with the world's first economy-wide intelligent sensor network technology (Sense-T) [10]. Smart city applications would generate huge amounts of data from a wide variety of data sources ranging from environmental sensors, mobile phones, localization sensors, to data generated by people from social networks. While there is general agreement that the use of these big sensor data would lead to improved services in smart cities, many research challenges remain on how to integrate and utilize these big sensor data. The authors in [11] discusses how digital devices and infrastructure have been used in smart cities to produce Big Data and enable real-time analysis of city life, and new modes of urban governance. Other challenges as outlined in [12]-[14] that face the design, development and deployment of big data applications for smart cities include: data sources and integration, data and information sharing, quality of data, smart city population, cost, smart network infrastructure, advanced algorithms, citizen awareness, and Big Data management.

Fig. 2 shows a smart cities framework proposed by the authors in [15] showing the logical information flow and challenges by layers. The framework consists of five inter-connected core layers: Connectivity, Datacenter, Analytics, Applications, and End-User Layers. The Connectivity Layer provides the base networking technologies including sensors, collectors, and wireless communications (e.g. cellular networks, LPWAN, WPAN). The Datacenter Layer provides the repository and storage, which is often based on cloud technologies. The Analytics Layer gives the value generation and predictive analytics from different types of Big Data. The final End User Layer absorbs the outcomes of the smart city. The five layers of the smart city framework provide a logical flow that enables the stakeholders in the smart cities to view the flow of information. The circular flow of information within the framework results in a feedback loop that aids in better understanding of the best practices in other Smart City initiatives. The challenges of each layer of the framework are further discussed and elaborated in Sections III-V.

This paper aims to survey the progress made in the development and applications of big sensor data research towards its deployment in smart cities ecosystems. The connecting theme would be the challenges faced by the five layers in the smart cities framework. We classify the existing research on big sensor data systems based on their characteristics and layer challenges for smart cities. This survey is targeted towards getting insights for using big sensor data systems for deployment in smart cities. To the best of our knowledge, the classification presented in this paper would have a different emphasis from other previous surveys on Big Data [16]-[19], smart cities [20]-[22] and/or IoTs [23]-[25]. The remainder of the paper is organized as follows. Section II presents an overview of the advancements in big sensor systems and presents a classification of the works based on the layer challenges. In Sections III to V, we review some research works focusing on the Connectivity, Datacenter, and Analytics layers in these areas. Section VI continues the discussion for several applications for big sensor data research in the design of smart cities ecosystems. Future work directions highlighting some futuristic applications are given in Section VII. Finally, Section VIII concludes the paper.

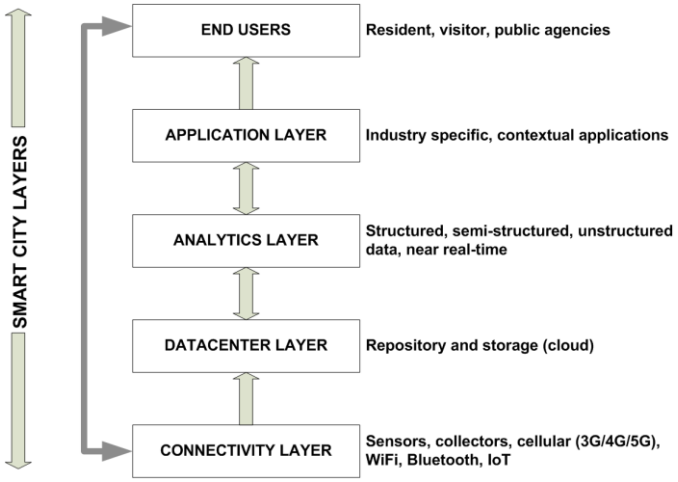

Fig. 2. Smart city framework showing logical flow and challenges by layers.

II. Overview and Classification

This section discusses big sensor data systems and classify these works into different categories, based on their layer challenges in smart cities. This classification is shown in Table I, which also gives a summary of the overall scope of this paper.

TABLE I.
CLASSIFICATION OF BIG SENSOR SYSTEMS FOR SMART CITIES.

Classification	References
Connectivity Layer	
Network infrastructures	[26],[27],[29],[30]
Interoperability and standards	[31],[32],[33]
Power consumption	[34],[35],[36],[37],[38]
Scalability	[26]
Datacenter Layer	
Cloud storage services	[39],[40],[41],[42]
Storage capacity challenges	[43],[44],[45],[46],[47], [48],[49]
Data analysis and processing challenges	[51],[52],[53],[54]
Data management challenges	[55]
Data extraction and cleaning challenges	[56],[57],[58]
Data integration and aggregation challenges	[59],[60]
Analytics Layer	
Cross-domain machine learning	[61],[62],[63],[64]
Data inference challenges	[65],[66],[67],[68],[69], [70],[71],[72]
Real-time and real-world applications	[68],[69]
New uses of available sensing infrastructures	[73]
Applications	
Environmental monitoring	[57],[67],[68],[69]
Intelligent transportation	[74],[75],[76],[86]
Disaster management	[77],[78],[79],[91]
Smart building management	[80],[81],[85]
Real-time urban monitoring	[82],[83],[84],[87],[88], [89],[90]

III. Connectivity Layer

This section discusses four challenges for designing big sensor data systems at the Connectivity Layer: (1) Network infrastructures; (2) Interoperability and standards; (3) Power consumption, and (4) Scalability.

A. Network Infrastructures

Currently, an important feature for any smart device is how to reach the Internet. The connectivity and communication of systems and/or platforms in most cases require the IoT. There are several options, all of them with their own advantages and drawbacks. Some of these techniques include:

1) Low Power Wide Area Networks (LPWAN): LoRaWAN, LTE-Cat M, SigFox, Ingenu RPMA, Dash7, nWave;
2) Cellular technologies: GPRS, 2G, 3G, 4G (LTE) and the future 5G;
3) Wireless Personal Area Networks (WPAN): IEEE802.15.4 (ZigBee), IEEE802.15.1 (Bluetooth), Wi-Fi, BLE.

The communication of some of the devices or systems requires 3/4/5G wireless networks. Smart meters and home devices may require ZigBee, 6LOWPAN, BLE, WiFi, which is dependent on the type of sensors and IoT devices used in the system. The challenge here is on how the real-time data collected by the multi-sensors devices and systems find access to the Internet. As reported in [26], practical realization of most services in the smart cities are not affected by technical issues, but are affected by the lack of widely accepted service architectures and communication protocols [27]. It is therefore recommended that for any type of network infrastructures used in the smart cities, robustness and addressing should be considered. Because of the highly populated sensor devices in an urban area like smart cities, unique addresses for different sensor devices becomes challenging. Hence new addressing mechanisms or novel routing techniques that do not require identifications (IDs) for each device are required. Table II shows a summary of network infrastructure technologies for smart cities ecosystems and their characteristics.

Also, since the wireless network of the sensor devices will have to rely on devices in the smart city to deliver information in a multi-hop manner, communication protocols should be robust to device failures and prevent single point-to-point situations, for information not to be lost in case a sensor dies [26],[29]. Another challenge would be to do with costs. Big sensor data systems will require the collection of data from hundreds of thousands of sensors which could be embedded in "Things" (e.g. garbage cans, street lights, etc.) which could be located and scattered anywhere in the environment. It is prohibitively expensive to implement a sensor relay infrastructure everywhere in the city for data gathering. The authors in [30] proposed an opportunistic approach using taxi cabs as "data mules" to perform the data collection. Sensor data throughout the city is gathered as the cabs go about its daily travels. Their results showed that an average of 120 vehicles could achieve an 80% coverage in less than 24 hours for the town area of Rome.

TABLE II.
NETWORK INFRASTRUCTURE TECHNOLOGIES FOR SMART CITIES.

Characteristics	**Cellular Technologies**	**Low Power Wide Area Networks (LPWAN)**	**Wireless Personal Area Networks (WPAN)**
Transmission range	Up to 100km	3-50km	10-100meters
Data rate	200Kbps-1Mbps	0.1Kbps-8Kbps	40Kbps-250Kbps
Security and Reliability	Highly secured (SIM based) and Ultra reliable	Less secure and reliable	Not secure and less reliable in connectivity
Scalability	Can handle massive volumes of devices (in millions)	Low volume of devices (in thousands)	Less scalable (connecting few devices mostly in the range of hundreds)
Tolerable delay	30min for data	20min for data	1-5min for data
Feasibility	Possible to realize, but require special sensors	Possible to realize, but require special sensors	Easy to realize, but require low cost sensors
Cost	Expensive in installation of cellular devices	Less expensive as compared with cellular devices	Low cost of installation
Energy source	Mostly battery powered (low battery life of less than a year)	Battery powered (long battery life of up to 20 years)	Mostly energy harvester or mains powered
Services	Mostly waste management and backbone for transferring collected data to the cloud	Smart parking, smart bicycles, smart lighting	Air quality monitoring, traffic congestion, smart parking
Data transfer	Backbone for transferring collected data to the cloud	Backbone for transferring collected data to the cloud	Enhance sensor and node implementation

B. Interoperability and Standards

Some researchers point towards interoperability and standards as the true enabler of the smart-cities ecosystems [31]. The continuous ability to send and receive data among interconnected devices in the city without quality of services degradation is challenging in the urban area. Standards remain the essential tools that make possible the design of interoperable systems/platforms. Some of the interoperability issues in big sensor systems involve connection admission control, specification of protocol suites, end-to-end quality of service, provision of basic and enhanced services, user selection of transit networks and content providers, user data element format, processing, and retrieval and storage of information.

As shown in Fig. 3, the challenge is on how information is exchanged through networks and used through user-based and network-based applications and services, (*A, B, C, D, … M*) and (*a, b, c, d, … m*) respectively. Some of the standards available to devices used in the smart cities are standards for cellular network - 3rd Generation Partnership Project (3GPP), Committee T1, International Telecommunication Union Telecommunication Standard (ITU-T), European Telecommunications Standards Institute (ETSI), Institute of Electrical and Electronics Engineer (IEEE), Internet Engineering Task Force (IETF) and the NSG Metadata Foundation (NMF). Considering the 3GPP which affects most of the device-to-device communication in the smart cities, [31]-[33] proposes that, to meet with the new connectivity

requirements of the emerging massive IoT segment, there is a need for new standards that will address the following: (1) low device cost; (2) improved battery life; (3) improved coverage, and (4) support for massive numbers of IoT connections.

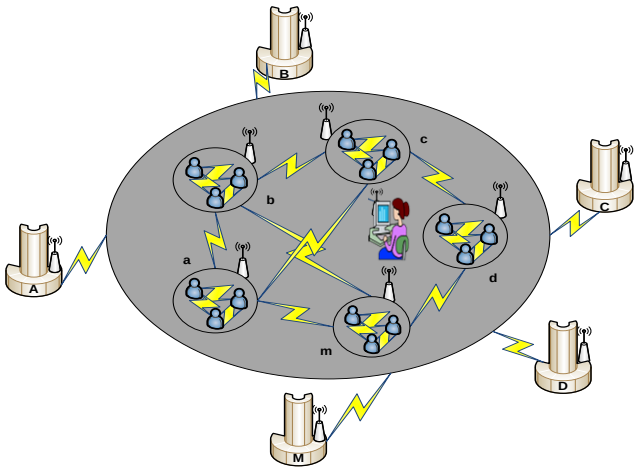

Fig. 3. Interoperability of systems.

C. Power Consumption

The end points of big sensor data systems are small node devices which suffer from the same power constraints as wireless sensor nodes. Thus, they can only be equipped with limited power sources due to severe hardware constraints. Some possible solutions towards the power consumption challenge have been proposed. In [34], researchers proposed a smart energy city. The smart energy city is highly energy and resource efficient and is increasingly powered by renewable energy resources. The system relies on integrated and resilient resource systems, which is insight-driven and innovative approaches to strategic planning. Also as discussed in [35], the energy in a smart city will have to rely on a smart, sustainable and resilient energy system built in an integrated planning approach for energy planning, active buildings, smart grids, smart supply technologies, and sustainable mobility.

Another important requirement is to minimize the energy consumption for the sensing process. The energy consumption for sensing is determined by its sampling rate. New techniques like compressive sensing (CS) have been shown to be effective to reduce the node energy consumption [36],[37]. Although traditional compression techniques can reduce the amount of data to be transmitted by removing redundancies, they still need to sample the data at high rates, and thus incurs high energy consumption for sampling. Furthermore, the compression process itself incurs additional energy and requires more computational power from sensor nodes. CS techniques work by trading off a simpler data acquisition requirement at the node level with a heavier computational requirement to reconstruct the CS sampled data at the base station. This asymmetric arrangement is well-suited for big sensor data systems where substantial processing power (without energy constraints) is available at the central processing station. The authors in [38] proposed another approach for reducing node energy consumption for data collection by scheduling nodes to sleep (i.e. turn off their radios). The challenge is that sleeping nodes cannot participate in network functions (e.g. routing), with the possibility that parts of the network become partitioned and are not reachable by any node. The authors showed that their management protocol could maintain both full connectivity and higher than 90% coverage in large-scale sensor networks.

D. Scalability

To observe physical phenomena and make all devices in the urban area to become smart devices, big sensor data systems require a high-density deployment of sensors and a large number of sensors for proper connectivity [26]. However, the large number of sensor devices will in turn reduce the transmission range of the devices, and poses a challenge for deployment in smart cities. Hence, fully distributed protocols which operate with limited knowledge of the topology, need to be developed to provide scalability as pointed out in Table II. Furthermore, since high level data is more important than individual pieces of data from each sensor device, routing protocols should support in-network combination of the data from a large number of devices without hampering energy consumption. This also brings about the need for standards.

IV. Datacenter Layer

The deployment of big sensor data systems in smart cities face the same barriers as other standard Big Data systems. This section discusses six challenges for designing big sensor data systems at the Datacenter Layer: (1) Cloud storage services; (2) Storage capacity; (3) Data analysis and processing; (4) Data management; (5) Data extraction and cleaning, and (6) Data integration and aggregation.

A. Cloud Storage Services

Probably the most common solution for smart cities ecosystems is to use cloud storage services. There exists several options for cloud storage services such as public cloud, private cloud and hybrid cloud storage, ranging from specific service providers to custom ad-hoc solutions. We shall briefly discuss each of the cloud storage services, and draw a comparison in terms of their scalability, security, performance, reliability and cost.

(i) Public Cloud Storage: Here, the services are provided by a fast growing list of service providers which include AT&T, Amazon, Amazon Simple Storage Service (S3), and Amazon Glacier for cold storage, Microsoft Azure, Iron Mountain Inc., Microsoft Corp., Nirvanix Inc., Google, Rackspace, Hosting Inc. and many others [39],[40]. Their storage infrastructure usually consists of low-cost storage nodes with directly attached commodity drives with an object-based storage stack that manages the distribution of content across nodes. Data in the cloud is typically accessed via Internet protocols, mostly Representational State Transfer (REST) [41] and to a lesser degree Simple Object Access Protocol (SOAP). Resilience and redundancy is achieved by storing each object on at least two nodes. Usage is charged on a dollar-per-gigabyte-per-month basis and depending on the service provider, there may

be additional fees for the amount of data transferred and access charges [40].

(ii) Private Cloud Storage: Runs on dedicated infrastructure in the data center and addresses the two main concerns of security and performance, but otherwise offers the same benefits of public cloud storage. Private storage clouds are usually for a single tenant, even though larger enterprises may use multi-tenancy features to segregate access by departments or office locations. Unlike their public cloud storage counterparts, scalability requirements are more modest [40],[42] so internal cloud storage offerings are more likely to have traditional storage hardware under the hood. An example of a private cloud storage offering is the Hitachi Data Systems Cloud Service for Private File Tiering. Based on the Hitachi Content Platform (HCP), it resides in the customer's data center but is owned and managed by Hitachi. Besides an initial setup fee, the customer pays for it by usage. Similarly, Nirvanix hNode provides a fully managed, pay-as-you-go, internal cloud offering within the data center, based on the same technology that powers the Nirvanix Storage Delivery Network (SDN).

(iii) *Hybrid Cloud Storage*: Users of this service manage resources both externally and in-house. Hybrid cloud solutions must meet certain key requirements to make hybrid cloud storage work. One of which is a mechanism that keeps active and frequently used data on-site while simultaneously moving inactive data to the cloud. These types of clouds also depend on policy engines to define when specific data gets moved into or pulled out of the cloud. Table III shows a summary of cloud storage services and their characteristics adapted from [40].

TABLE III.
CLOUD STORAGE SERVICES AND THEIR CHARACTERISTICS.

Characteristic	Public Cloud Storage	Private Cloud Storage	Hybrid cloud Storage
Scalability	Very high	Limited	Very high
Security	Good	Most secure. All storage is on premises	Very secure
Performance	Low to medium	Very good	Good. Active content is cached on premises
Reliability	Medium. Dependent on internet connectivity and service provider availability	High. All equipment is on premises	Medium to high. Cached content is kept on premises
Cost	Very good. Pay-as-you-go model and no need for on premises storage infrastructure	Good. Require on premises resources (data center space, electricity and cooling)	Improved. Allows moving some of storage resources to a pay-as-you-go model

B. Storage Capacity Challenges

This section gives a brief analysis of Big Data oriented databases such as MongoDB, Cassandra, HyperTable, PostgreSQL, and CouchDB. Due to the rapid advances in sensor technologies, the number of sensors and the amount of sensor data have been increasing with incredible rates. Processing and analyzing such Big Data require enormous computational and storage costs with a traditional SQL database. The scalability and availability requirements for sensor data storage platform solutions resulted in the use of NoSQL databases, which have the ability to efficiently distribute data over many servers and dynamically add new attributes to data records [43]. Open source NoSQL databases provide efficient alternatives for large amounts of sensor data storage and is a suitable solution for smart cities. Table IV shows a summary of three different categories of NoSQL databases with details adapted from [44]. The different categories of NoSQL databases presented in Table IV provide high availability, performance, and scalability for Big Data.

The authors in [43],[45] proposed a two-tier architecture with a data model and alternative mobile web mapping solution using the NoSQL database CouchDB, which is available on most operating systems. The authors in [45] discussed the possibilities of using NoSQL databases such as MongoDB and Cassandra in large-scale sensor network systems. Their analysis of the two showed that while Cassandra is the better choice for large critical sensor applications, MongoDB is the better choice for a small or medium sized noncritical sensor application. On the other hand, MongoDB has a moderate performance when using virtualization; by contrast, the read performance of Cassandra is heavily affected by virtualization. MongoDB is a document-oriented database with support for storing JSON-style documents [46], and provides high performance, high availability, and easy scalability [45]. Furthermore, documents stored in MongoDB can be mapped to programming language data types. MongoDB servers can be replicated with automatic master failover. MongoDB has been investigated in several studies and used in various types of commercial and academic projects [47]–[49]. The major reasons for using MongoDB, as discussed by the authors in [44] are that it provides high performance write support for QuickServer, and permits scalability of databases for big sensor infrastructures.

TABLE IV.
CATEGORIES OF NoSQL DATABASES.

Category	Characteristics	Examples
Key-value Stores	Stores values indexed for retrieval by keys. Hold structured or unstructured data	Redis, Riak, Amazon SimpleDB, Scalaris.
Column-oriented databases	Contain one extendable column of closely related data	Cassandra, HBase, DynamoBD, HyperTable
Document-based stores	Store and organize data as collections of documents. Allows additions of fields to a document	MongoDB, CouchDB

C. Data Analysis and Processing Challenges

Analysing big sensor data requires use of statistical methods as well as data-mining or machine-learning algorithms. There are many user-friendly machine-learning frameworks such as RapidMiner and Weka [50]. However, these traditional frameworks do not scale to big data due to their memory constraints. Several open source big data projects have implemented many of these algorithms. One of these

frameworks is Mahout, which is a distributed machine-learning framework and licensed under the Apache Software Foundation License. The goal of Mahout is basically to build a scalable machine-learning library to be used on Hadoop. The study by [51] compared the performance of various classification and clustering algorithms using Mahout-library on two different processing systems: Hadoop and Granules. Their results showed that the processing time of Granules implementation is shorter than Hadoop. In another study described by [52], it was observed that Mahout demonstrates bad performance and no gain for files smaller than 128 MB.

The comparison results of Spark and Hadoop performances presented by authors in [44] showed that Spark outperformed Hadoop when executing simple programs such as WordCount and Grep. In another similar study, it has been shown that the *k*-means algorithm on Spark ran about five times faster than that on MapReduce. The performance comparisons of Hadoop, Spark, and DataMPI using *k*-means and Naïve Bayes benchmarks as the workloads are described in [53]. OpenStack is a popular cloud computing technology that offers many opportunities for Big Data processing with scalable computational clusters and advanced data storage systems for applications and science researchers [44]. The cloud computing stack can be categorized into three service models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS), where IaaS is the most flexible and basic cloud computing model. IaaS providers include Amazon EC2, Rackspace Cloud, and Google Compute Engine (GCE). OpenStack is an IaaS-cloud computing software project based on the code developed by Rackspace and NASA. OpenStack offers a scalable, flexible, and open source cloud computing management platform.

MapReduce, Google's Big Data processing paradigm, has been implemented in open source projects like Hadoop. Hadoop has been the most popular MapReduce implementation and is used in many projects in many areas of the Big Data industry. The Hadoop Ecosystem also provides many other Big Data tools such as Hadoop Distributed File System for storing data on clusters, Pig [43],[44] an engine for parallel data flow execution on Hadoop, HBase, Google's BigTable, Hive, a data warehouse software on Hadoop, and data analysis software like Mahout. The major advantages of the Hadoop MapReduce framework are scalability, cost effectiveness, flexibility, speed, and resilience to failures [54]. On the other hand, Hadoop does not fully support complex iterative algorithms for machine learning and online processing.

D. Data Management Challenges

The sporadic load characteristics of Big Data applications, coupled with increasing demand for data storage while guaranteeing round the clock availability, and varying degrees of consistency requirements pose new challenges for data management in the cloud. These modern application demands systems capable of providing scalable and consistent data management as a service in the cloud. Big sensor data management for smart cities is dependent upon the following key factor and challenges: data privacy, data security, governance and ethical size of big data structures, legal and technological implications, and policy issues [40],[55].

The traditional database and software technologies can no longer be used in the management of big sensor data systems, hence the distributed database system is designed to take care of the inefficiencies. However, this could not be sustained owing to the crippling effect on performance caused by partial failures and synchronization overhead [44]. The emergence of a different class of scalable data management systems such as Google's BigTable, PINUTS from Yahoo!, Amazon's Dynamo, Apache Hbase (open source BigTable clone), Apache Cassandra, Riak (open source Dynamo clone), etc. [54] are concerned with managing and accessing petabytes of data.

E. Data Extraction and Cleaning Challenges

Data captured from the physical world through sensor devices tends to be noisy, incomplete, and unreliable. Traditional data cleaning techniques for conventional Big Data (e.g. data warehousing) do not take into account the strong spatial and temporal correlations typically present in sensor-based data types. The authors in [56] considered the problem where the training data is severely skewed for the non-polluted range and proposed an entropy maximization step to normalize the data distribution, and to avoid over-fitting prior to training the neural network. The authors in [57] performed two preprocessing steps as part of the data extraction and cleaning. The first preprocessing removed duplicate data points and data points that contained missing information. The second preprocessing removed erroneous and outlier data points.

Other useful approaches for data cleaning for sensor data have used techniques like spatiotemporal regression and Kalman filters. The authors in [58] proposed three models to detect and identify erroneous data among inconsistent observations based on the inherent structure of various sensor measurements from a group of sensors. The first model used multivariate Gaussian model to explore the correlated data changes of a group of sensors. The second model used principal component analysis (PCA) to capture the sparse geometric relationship among sensors, and the third model used kernel functions to map the original data into a high dimensional feature space prior to using the PCA model (i.e. kernel PCA technique). Their results demonstrated good detection rates with limited false alarms.

F. Data Integration and Aggregation Challenges

The authors in [59] defined two types of sensor-based applications depending on the information that is needed at the sink: (1) "functional" – where only statistical summary values (e.g. maximum, average, median) are required, and (2) "recoverable" – where the full dataset is required. For functional applications, data integration and aggregation can be easily performed during the data collection and transmission process as part of a hierarchical data gathering tree. On the other hand, the challenges of recoverable applications pose more difficulty for data aggregation due to

the lack of prior knowledge on the spatial data correlation structure. Their work for a "functional" sensor application, proposed a data aggregation approach based on compressed sensing (CS). The authors showed that their CS scheme based on diffusion wavelets gave high fidelity recovery for aggregated sensor data while achieving significant energy savings.

A further example of the challenges faced in this module is by the work in [60] for a wireless body area sensor network (WBASN) that enables continuous monitoring of ambulatory patients at home while they recover from noncritical conditions. A focus in this work is to work within the power limitations of wearable sensors which often employ button-cell batteries to achieve compactness and small form factor. This work used a two tier hierarchical network for data gathering consisting of a first tier of wearable sensors used for vital signs collection and a second tier point-to-point link between the WBASN coordinator device and a number of fixed access points (APs).

V. Analytics Layer

This section discusses four challenges for designing big sensor data systems at the Analytics Layer: (1) Cross-domain machine learning; (2) Data inference; (3) Real-time applications; and (4) New uses of sensing infrastructures.

A. Cross-Domain or Transfer Machine Learning

An assumption made in traditional machine learning techniques is that the training and testing data are obtained from the same data domain and has the same distribution. This may not be the case in big sensor systems where data is collected and drawn from a variety of sensing sources and domains. Furthermore, both real-time and historical sources may be required to be utilized. Compared with using conventional machine learning techniques, much less research has been conducted for using transfer learning techniques in networked sensor-based systems. The work in [61] proposed an application of cross-domain learning to reduce the calibration effort of learning a model for calculating where a client device is located in a wireless network. The authors proposed an approach for transferring the learning model trained on data obtained from one spatial sensing space (area of a building) to be applied in another area. Their approach learnt a mapping function between the signal space and the location space by solving an optimization problem based on manifold learning techniques.

The works in [62] and [63] proposed to apply transfer learning techniques towards solving sensor-based activity recognition in indoor environments. The authors in [64] remarked that many transfer learning works are only focused on learning from a single source domain to a target domain. However, the challenge on how to apply the knowledge learnt from multiple source domains (a characteristic of big sensor data) to a target domain remains to be clearly addressed. The authors proposed a technique based on a consensus regularization framework where a local classifier is trained by considering both local data available in one source domain and the prediction consensus with classifiers learnt from other source domains.

B. Data Inference Challenges

An important requirement for big sensor data systems is towards using a variety of heterogeneous data sources or historical data to infer missing data or predict future trends in the spatial-temporal sensing field. Fig. 4 shows a big sensor model which has been adapted from the 5*V*'s model [65]. There are six modules in the model divided into three data transformation processes: (1) data collection; (2) data inference; and (3) value generation. The inputs into the data collection process are the raw sensor data values $s(x,y,t)$ harvested from (multiple) sensor farms. The output of the process are the cleaned and aggregated data values $s'(x,y,t)$. The inputs into the data inference process are the cleaned and aggregated data values $s'(x,y,t)$ from the data collection process. The output of the process is a snapshot of the sensing field $S(t)$ at a particular time. Thus, the objective of the big data research for the data inference process is to perform an inference process to give values to the sensing field where no direct data are available. The challenge is to find suitable models and techniques to integrate the various sources of data to solve the big data problem. A way to visualize this is to picture a set of virtual sensors deployed at desirable locations in the sensing field. The task then is to infer the sensing values for the virtual sensors, using information from available physical sensors in combination with other data sources. Different approaches can be employed to perform the data inference, ranging from conventional statistical-based approaches to newer machine-learning approaches.

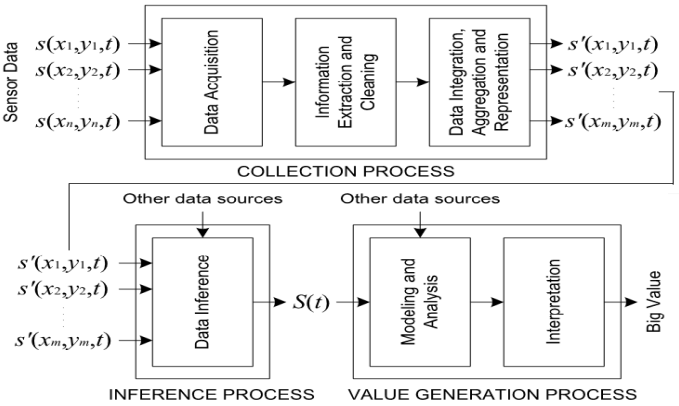

Fig. 4. Big sensor model for networked sensors divided into three main data transformation processes: (1) Data collection, (2) Data inference, and (3) Value generation.

The authors in [66] used a statistical-based approach (Gaussian Process Regression or Kriging) and the authors in [67] used Generalized Additive Models (GAMs) to model pollution concentrations at locations not covered by the sensor nodes. On the other hand, the authors in [68],[69] used a machine-learning approach in combination with a variety of data sources (meteorological, traffic, human mobility, Point-of-Interest, road network data) as well as historical AQI data to resolve their inference problem. Other examples of the challenges faced in the data inference process can be found in

the works by [70], [71] in the field of geosensor networks.

The authors in [70] identified two challenges related to sensor network deployments. The first challenge is the data sparsity problem where the granularity and density of the deployed sensors do not match the required sensing requirements, and the second challenge is that sensor network deployments were mostly collected from a regular topology. In this work, the authors proposed using Kriging and non-stochastical interpolation techniques to estimate data at unmonitored locations, and then applied for irregular topologies. A more recent work by the authors in [71] used a Markovian approach to model spatial correlation in sensor network data. The usefulness of their model is that it makes no assumption that only nodes that are close to each other will be correlated. If some strange application results in nodes that are close to each other being uncorrelated but nodes far away being correlated, then their model will not fail.

Since the output of the data inference process is used as input into the value generation process, the final outcome from the value generation process is also dependent on the performance and effectiveness of the data inference process. The authors in [72] give a useful framework for data integration and fusion using kernel-based methods to combine together different modalities and data types. They used an application from systems biology and showed the diverse data modalities which can be extracted from a single instance of DNA (e.g. the high dimensional expression data, the sparse protein-protein-interaction data, the sequence data, the annotation data, and the text mining data). Their approach used the kernel trick where data which has diverse data structures is all transformed into kernel matrices with the same size for combination.

C. Real-Time and Real-World Applications

Many applications for smart cities (e.g. for earthquake/disaster early warning system, air pollution monitoring) require (near) real-time performance to serve its function. This will drive the "Velocity" characteristic for big sensor data systems. Currently, most (if not all) research on big sensor data systems do not consider this aspect (are performed offline), and research is conducted using historical or past data. Our investigation and survey of this area showed this. A major challenge for applying machine learning techniques in urban deployments is that it is difficult to perform the training (tagging) to test applications for real-world systems. Many studies may deduce some benefits from a period of study-learning. However, very few of them continue to perform the empirical and hypothesis formulated in real-world scenarios.

A few examples where authors have continued to perform testing in real-world scenarios are the works for deriving air pollution maps in Beijing [68],[69] and Zurich [67]. In the future, we anticipate the research and development of big sensor data systems where real-time analytics will be performed on large volumes of recently acquired data from multiple sensor farms, and using a number of diverse and historical sources. These challenges for real-time and real-world deployments remain an important issue to be resolved.

D. New Utilization of Available Sensing Infrastructures

Another challenge is to find ways to utilize available sensing infrastructures to provide new information for new tasks or applications. An example is the work by [73] for real-time rainfall monitoring in the Netherlands. The authors used the received signal level data from microwave links in cellular communication networks, and derived the path-averaged rainfall intensity from the signal's attenuation between transmitter and receiver.

VI. Applications for Big Sensor Data Research in Smart Cities Ecosystems

Applications of big sensor data research in various domains for smart cities ecosystems can be found in environmental monitoring [57],[67]-[69], intelligent transportation [74]-[76],[86] disaster management [77]-[79],[91], smart building management [80],[81],[85], and real-time urban monitoring [82]-[84],[87]-[90]. Table V gives a summary of a few examples using the sensor classifications as shown in Fig. 1.

Currently, a major success for big sensor data research in smart cities is for inferring or predicting the air quality index (AQI) levels. In smart cities and urban areas, the monitoring of air pollution is very challenging because of multiple factors which affect the air quality such as meteorology, traffic volume, land use, and urban structures. The authors in [68],[69] discussed an air quality monitoring system for Beijing. Their approach used a cross-domain machine learning data inference process to infer AQI levels for regions without monitoring stations but containing other available data sources. A co-training based semi-supervised learning approach was employed where unlabeled data were used to improve the inference accuracy. Two classifiers (a spatial and a temporal classifier) were built. The spatial classifier was based on a backpropagation neural network and used the static features like the road network and POI features to model the spatial correlation of air quality amongst different cells. The temporal classifier was based on a linear-chain conditional random field (CRF) and used the dynamic features like meteorological, traffic, and human mobility features to model the temporal dependency of air quality in an individual cell. The researchers reported an accuracy of 82% for the detection of PM_{10} levels and inferred the AQIs for the entire Beijing in five minutes.

A different approach was taken by the researchers in [57] which proposed a big data analytics model to identify clusters or communities of buildings with large $PM_{2.5}$ and NO_x amounts of emissions or "hot spots" to understand the trends of air pollution in New York City. They considered a data set where each data element is represented by a building. For each of the N data elements, there is a corresponding geographic location which was obtained from a separate set of land use and geographic data at the tax lot level. Each building is abstracted as a feature vector sensor source possessing spatial correlations with other neighboring buildings in the sensing space. The authors represented the built environment as a graph signal model $G = (V,W)$ where $V = \{v_0,\ldots v_{N-1}\}$ is the set

TABLE V.
SUMMARY FOR BIG SENSOR DATA RESEARCH USE-CASES IN SMART CITIES.

Ref.	City	Application remarks	Technical Sensors	Human Sensors	Data Collected	Analytical Method
[68], [69]	Beijing	Infer the PM_{10} levels for the entire city with near real-time performance	Real-time air quality (AQI) sensors		Real-time and historical AQI levels, meteorological data, traffic data, human mobility data, Point-of-Interest data, road network data	Semi-supervised co-training using two classifiers: (1) spatial classifier based on backpropagation network, (2) temporal classifier based on linear-chain conditional random field (CRF)
[57]	New York	Understand the trends of air pollution in New York City	Energy consumption of heating oil data for large buildings		Heating oil data, land-use data, geographic data	Community network detection algorithm based on Louvain method for modularity maximization
[67]	Zurich	Derive accurate ultrafine particles (UFPs) pollution maps with high spatiotemporal resolution	Air quality sensors on mobile sensing nodes		50 million UFP measurements collected over a period of two years	Generalized Additive Models (GAMs) to construct land-use regression (LUR) model
[84]	Barcelona	Expose the underlying daily routines and patterns of citizens using bicycle	RFID sensors on bicycle station		Station geolocation, available bicycles, vacant parking lots	Clustering using Expectation-Maximization
[85]	Christchurch	Incorporate sensors into building infrastructure	Environmental sensors		Air pollution levels, noise levels, water use, traffic flow	
[86]	Amsterdam	Traffic flow management to decrease vehicle loss hours by as much as 10%	2400 vehicle detector stations, 60 number plate recognition cameras		Traffic flow	
[87]	New Jersey	Human mobility characterization from cellular network data		Call Detail Records (CDRs) from cellular network	Anonymous phone identifier, data/time of voice call or text message, elapsed call time, cellular antennas, billing ZIP code	Classification algorithm for likelihood of sequence of antennas on a particular route
[88]	Copenhagen	Electric bicycling system to collect fine-grained environmental information (e.g. heat map)		Crowdsourcing environmental sensors (CO, NOx, temperature, noise, humidity) on bicycle	Environmental data	
[89]	Rome	Uncover the movements of tourists from georeferenced photos		Georeferenced photos on photo-sharing website Flickr	Number of photos, number of photographers, number of phone calls made by foreigners over time period	Spatial data clustering
[90]	Rome	Real-time urban monitoring using anonymous monitoring of mobile cellular networks and GPS positioning of buses and taxis	GPS	Cellphone network (radio channel measurements)	Instantaneous position of each mobile element	Geographic interpolation of traffic intensity
[91]	Sao Carlos	Decision support system for flood risk management in Brazil	*In-situ* water sensors	Human participatory sensing, Volunteered Geographic Information (VGI)	Water levels (“Gauge Height”)	

of nodes and W is the weighted adjacency matrix of the graph. Each data element (building) corresponds to a node v_n in the graph model and the entry $W_{i,j}$ is the weight of a directed edge that reflects the degree of relation (spatial) of the jth building to the ith building. The analysis was performed using a complex network systems technique based on the Louvain method for community detection. They reported that their graph signal model could better quantify and rank the combined impact of a building's own heating oil consumption and the consumption of its neighbors on surrounding air quality compared with the conventional method.

A recent work by [67] proposed a mobile measurement system for the city of Zurich to derive accurate ultrafine particles (UFPs) pollution maps with high spatiotemporal resolution. UFPs are particles with a diameter of less than 100nm. Their system collected a very large scale dataset of over 50 million UFP measurements using mobile sensing nodes over more than two years (from April 2012 to April 2014). The mobile measurement system consists of ten sensor nodes installed on top of public transport vehicles, which cover a large urban area (100m x 100m) on a regular schedule. The authors developed land-use regression (LUR) models to produce accurate pollution maps with high spatiotemporal resolution. Their LUR model used a set of explanatory variables (land-use and traffic characteristics data) based on Generalized Additive Models (GAMs) to model pollution concentrations at locations not covered by the mobile sensor nodes. The authors evaluated the dependencies between the explanatory variables and the measurements, and exploited these spatiotemporal relationships to predict the pollution levels for all locations without measurements but with available land-use and traffic information.

Brief descriptions are given next for other examples of big sensor data research in smart cities [84]-[91]. The works in [84], [85] and [86] used technical sensors embedded in the environment. The work in [84] for Barcelona aimed to expose the daily routines and patterns of people using the city bicycling program. The system collected data on when a bike is picked up or parked. The work in [85] embeds sensor into the city infrastructure. The data collected would be useful for studying the impact of air pollution on respiratory disease, and generating data to inform cycle way development. The work in [86] for Amsterdam used 2400 vehicle detector stations and 60 number plate recognition cameras to decrease the vehicle loss hours. The works in [87],[88] and [89] used humans as sensors and collect data as people go about their daily routines. The work in [87] used Call Detail Records (CDRs) from a cellular network to characterize the human mobility. The work in [88] collects fine-grained environmental information in the city using data mined from crowdsourced bicycles. The work in [89] used the geolocation from photos on the Flickr social networking website to uncover the movements of tourists in Rome. The works in [90] and [91] are examples where both technical and human sensors are used. The work in [90] used a combination of GPS data and radio channel measurements from a cellphone network to give the instantaneous position of each mobile element. The work in [91] is for flood risk management in Brazil and uses a combination of in-situ water sensors and human participatory sensing to give the water level height.

VII. Future Work Direction

New futuristic applications for smart cities are envisioned with the emergence of new technological concepts such as big data analytics, semantic sensor networks, sensor-cloud computing, context-aware sensing, etc. The authors in [92] considered some open challenges and identified some challenges for the future IoT such as energy efficient sensing, security and reprogrammable networks, participatory sensing, and new networking protocols. This section broadens the discussion towards the big sensor data framework, and in particular will focus on the new potential and transformative applications which will see more interaction and information sharing among the Sensor/Things, Internet/Communications, and People/Creature domains. We highlight three potential future directions which would form and emerge from the intersection of the three domains: (1) Internet-of-People; (2) Context-aware sensing networks; and (3) Sensor-cloud infrastructure.

A. Internet-of-People

The authors in [93] have introduced the concept of the Internet-of-People (IoP). Their proposal advocates the use of smartphones to learn about its owner and context to improve the connection between people and the IoT (e.g. to enable context-aware sensing). Our vision for the IoP goes further in that we can see that people could not only contribute to data collection and sensing (e.g. human as sensors) but could also actively participate in performing collaborative and intelligent information processing for complex and unexpected tasks/events. Currently, a popular method to engage people for information processing tasks is mobile crowdsourcing using smartphones. In our case, an example IoP scenario would be for a smart city to utilize a fleet of small inexpensive drones to capture aerial images after a natural disaster (e.g. earthquake) to look for survivors. The cost of the drones can be made much cheaper by having the complex visual processing performed by volunteer humans (i.e. human as processing nodes) to tag useful images in real-time, and transmit to search and rescue teams to better focus their operations. Another application of the IoP would be to serve as quality assurance for automated sensing systems. In one scenario, a random selection from the results of automated classification systems could be sent to humans as processors to set a performance baseline and ensure that the system performance would be contained within the specified limits.

B. Context-aware Semantic Sensor Networks

The second potential direction from the intersection of the domains would be improving and developing new context-aware sensing networks and intelligent services for the IoT [94], [95] and IoP. The work in [94] surveyed many research efforts and placed emphasis on the development of middleware solutions to manage the large deployment of sensors to allow sensing systems to adapt their behaviors

towards changing environments and provide more useful and meaningful interpretation. Currently, many context-aware applications are targeted towards humans as the user of the services (e.g. location-awareness, personalization features). In the future, more context-aware applications would be targeted towards deriving advantages for the smart things/objects [94] in the sensing system. The derived advantages could be towards achieving two aims: (1) Towards advantages for the object sensing infrastructure; and/or (2) Towards advantages for the overall object applications. An example of (1) would be using the context information for increasing the energy efficiency of sensor node processing and communications. This can be performed in two ways (using Intra-PHY or Inter-PHY adaptation) [96]. In Intra-PHY, the PHY (physical layer) is not changed, but the parameters that affect energy efficiency are adapted based on context to realize the minimum energy consumption. In Inter-PHY, the node design implementation incorporates multiple PHYs, and the node can decide to use the PHY communication medium (e.g. Wi-Fi vs Bluetooth) suitable for the context. An example of (2) could be a smart grid sensing structure [97] taking in inputs from intelligent sensor networks for context-awareness to maximize the power supply and consumption efficiencies.

C. Sensor-cloud Infrastructure

A third future direction would be towards addressing the storage, processing, and retrieval issues for the huge amounts of sensor data (and context information) that would be generated. A potential solution to these issues would be new developments and improvements of the sensor-cloud infrastructure [98]-[100]. The sensor-cloud infrastructure is formed from the convergence of Internet/cloud and sensor-based technologies to combine their respective capabilities. It is modeled as having two gateways (sensor and cloud gateways) to link the sensing (physical) and storage (cyber) sides [98]. The sensor gateway performs data collection from the different types of sensing nodes, compresses the data, and transmits it to the cloud gateway. The cloud gateway decompresses the data for storage on the cloud servers. The harvesting of sensor information to maximize the collective intelligence from the vast sensor data storage in the cloud remains a significant challenge. As discussed in Section V, the development of new transfer machine learning techniques would serve as useful tools for the cross-domain analytics. Other than technological considerations, future works also need to consider the practical scenarios for cost effectiveness of big sensor data systems in smart cities. The authors in [99] considered a practical emphasis of the sensor cloud as a cloud of virtual sensors built on top of physical sensors to serve as a "sensing as a service" paradigm for next-generation mobile computing. In this model, the sensor-cloud decouples the network owner and the user, where the user need not be the owner. The sensor-cloud may consist of many sensing systems with different owners. Users purchase sensing services on demand from the sensor-cloud. This makes it more practical and cost-effective. The security/privacy issues of the stored data in the sensor-cloud is another critical challenge [100].

VIII. Conclusion

This paper has surveyed the area of big sensor data research towards the development of smart cities ecosystems. The major requirements and challenges for big sensor data research have been systematically explored using a layered smart city framework. Although, many techniques for handling a single form of sensor data type have been well established, what is still missing are approaches and models for handling and integrating the diverse and historical sources of cross-domain and multi-modal spatial-temporal data which characterizes big sensor data systems to produce valuable outcomes for human society. Our study has shown that most current research on big sensor data systems are performed offline and do not consider (near) real-time implementations. The convergence of the three domains in the big sensor framework, and the emergence of the IoP as well as the current IoT enables new sensing applications to be designed and developed. Together with the other challenges discussed in this study, this remains important issues to be resolved for practical deployments in smart cities ecosystems.

References

[1] http://www.ibm.com/big-data/au/en/big-data-and-analytics/operations-management.html. [Online].

[2] C. Yang, C. Liu, X. Zhang, S. Nepal, and J. Chen, "A time efficient approach for detecting errors in big sensor data on cloud", *IEEE Trans. Parallel & Distributed Systems*, vol. 26, no. 2, pp. 329-339, 2015.

[3] C. D'Este, C. Sharman, and A. Rahman, "Distributed feature selection with big sensor data," in *Proc. MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, 2014.

[4] G. Sagl, B. Resch, and T. Blaschke, "Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities," *Sensors*, pp. 17013-17035, 2015.

[5] Y.G. Sahin, "Animals as mobile biological sensors for forest fire detection," *Sensors*, vol.7, pp. 3084-3099, 2007.

[6] M. Baratchi, N. Meratnia, P.J.M. Havinga, A.K. Skidmore, and B.A.G. Toxopeus, "Sensing solutions for collecting spatio-temporal data for wildlife monitoring applications: A review," *Sensors*, vol. 13, pp. 6054-6088, 2013.

[7] L. Sanchez, L. Munoz, J.A. Galache, P. Sotres, J.R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis, and D. Pfisterer, "SmartSantander: IoT experimentation over a smart city testbed", *Computer Networks*, vol. 61, pp.217-238, 2014.

[8] T. Bakici, E. Almirall, and J. Wareham, "A smart city initiative: the case of Barcelona," *Journal of the Knowledge Economy*, vol. 4, no. 2, pp. 135-148, 2013.

[9] K. Kloeckl, O. Senn, and C. Ratti, "Enabling the real-time city: LIVE Singapore!," *Journal of Urban Technology*, vol. 19, no. 2, pp. 89-112, 2012.

[10] http://www.sense-t.org.au/. [Online].

[11] R. Kitchin, "The real-time city? Big data and smart urbanism," *GeoJournal*, vol. 79, no. 1, pp. 1–14, 2014.

[12] E. A. Nuaimi, H. A. Neyadi, N. Mohamed, and J. A. Jaroodi. "Applications of big data to smart cities," *Journal of Internet Services and Applications*, vol. 6, no. 25, pp. 1-15, 2015.

[13] B. Bajracharya, D. Cattell, I. and Khanjanasthiti, "Challenges and opportunities to develop a smart city: A case study of Gold Coast Australia," in *Proc. REAL CORP*, pp. 119-129, 2014.

[14] K. Su, J. Li, and H. Fu, "Smart city and the applications", in *Proc. Electronics, Communications and Control (ICECC)*, pp. 1028–1031, 2011.

[15] M. Kumar, "Building agile data driven smart cities," *IDC White Paper*, 2015.

[16] C.L.P. Chen and C-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314-347, 2014.

[17] M. Chen, S. Mao, and Y. Liu, "Big Data: A survey", *Mobile Network*

Applications, vol. 19, pp. 171-209, 2014.

[18] H. Hu, Y. Wen, T-S. Chua, and X. Li, "Toward scalable systems for Big Data analytics: A technology tutorial", *IEEE Access*, vol. 2, pp. 652-687, 2014.

[19] C-W. Tsai, C-F. Lai, H-C. Chao, and A. V. Vasilakos, "Big Data analytics: A survey", *Journal of Big Data*, 2015.

[20] M. Angelidou, "Smart cities: A conjuncture of four forces", *Cities*, vol. 47, pp. 95-106, 2015.

[21] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter cities and their innovation challenges", *Computer*, vol. 44, pp. 32-39, 2011.

[22] J. Wang, C. Li, and X. Zhang, "Survey of data-centric smart city", *Journal of Computer Research and Development*, vol. 51, no. 2, pp. 239-259, 2014.

[23] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey", *Computer Networks*, vol. 54, no. 15, pp. 2787-2805, 2010.

[24] C. Perera, A.B. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of Things: A survey", *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 415-454, 2014.

[25] C. Perera, A.B. Zaslavsky, P. Christen, and D. Georgakopoulos, "Sensing as a service model for smart cities supported by Internet of Things", *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 1, pp. 81–93, 2014.

[26] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Network*, vol. 38, no. 4, pp. 393-422, 2002.

[27] A. Zanella, N. Bui, A. Castellani, L. Vangelista and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 21, pp. 22-32, 2014.

[28] H. Assasa, S.V. Yadhav and L. Westberg, "Service mobility in mobile networks," in *Proc. IEEE 8th International Conference on Cloud Computing*, pp. 397-404, 2015.

[29] A.M. Zungeru, L.-M. Ang and K.P. Seng, "Termite-hill: Performance optimized swarm intelligence based routing algorithm for wireless sensor networks," *Journal of Network and Computer Applications*, vol. 35, no. 6, pp. 1901-1917, 2012.

[30] M. Bonola, L. Bracciale, P. Loreti, R. Amici, A. Rabuffi, and G. Bianchi, "Opportunistic communication in smart city: Experimental insight with small-scale taxi fleets as data carriers," *Ad Hoc Networks*, Article in Press (available online), 2016.

[31] P.T. De-Sousa and P. Stuckmann, "Telecommunication network interoperability," *Telecommunication Systems and Technologies*, vol. 2, 2007.

[32] S. Fang, L.D. Xu, Y. Zhu, J. Ahati, H. Pei, J. Yan, and Z. Liu, "An integrated system for regional environmental monitoring and management based on internet of things," *IEEE Trans. Industrial Informatics*, vol. 10, no. 2, pp. 1596-1605, 2014.

[33] Ericsson, "Cellular networks for massive IoT," *Ericsson White Paper*, January 2016.

[34] P.S. Nielsen, S. Ben Amer, and K. Halsnaes, "Definition of smart energy city and state of the art of 6 transform cities using key performance indicators," *Deliverable 1.2*, 2013.

[35] H.-G. Schwartz, "Member states initiative for smart cities," *GEODE Autumn Seminar*, November 2012.

[36] S. Li, L.D. Xu, and X. Wang, "Compressed sensing signal and data acquisition in wireless sensor networks and Internet of Things," *IEEE Trans. Industrial Informatics*, vol. 9, no. 4, pp. 2177-2186, 2013.

[37] W. Chen and I.J. Wassell, "Energy efficient signal acquisition in wireless sensor networks: A compressive sensing framework," in *Proc. 6th International Symposium on Wireless and Pervasive Computing*, pp. 1-6, 2011.

[38] H. Wang, H.E. Roman, L. Yuan, Y. Huang, and R. Wang, "Connectivity, coverage and power consumption in large-scale wireless sensor networks," *Computer Networks*, vol. 75, pp. 212-225, 2014.

[39] R. Buyya, S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol.25, no. 6, pp. 599-616, 2009.

[40] Hitachi Data Systems, "Determining the type of cloud to use and your cloud ROI". [Online].

[41] A. Huth, and J. Cebula, "The basics of cloud computing," *United States Computer Emergency Readiness Team, US-CERT*, pp. 1-4, 2011.

[42] R. Buyya, J. Broberg, and A. Goscinski, *Cloud Computing Principles and Paradigms*, John Wiley & Sons, 2011.

[43] R. Cattell, "Scalable SQL and NoSQL data stores," *ACM SIGMOD Record*, vol. 39, no. 4, pp. 12–27, 2010.

[44] N. Leavitt, "Will NoSQL databases live up to their promise?", *Computer*, pp. 12-14, 2010.

[45] G. Aydin, I.R. Hallac, and B. Karakus, "Architecture and implementation of a scalable sensor data storage and analysis system using cloud computing and Big Data technologies," *Journal of Sensors*, 2015.

[46] P. Membrey, E. Plugge, and D. Hawkins, *The Definitive Guide to MongoDB: the noSQL Database for Cloud and Desktop Computing*, *Apress*, 2010.

[47] A. Boicea, F. Radulescu, and L. I. Agapin, "MongoDB vs oracle-database comparison," in *Proc. 3rd Int. Conf. on Emerging Intelligent Data and Web Technologies (EIDWT '12)*, pp. 330–335, 2012.

[48] E. Dede, M. Govindaraju, D. Gunter, R. S. Canon, and L. Ramakrishnan, "Performance evaluation of a MongoDB and Hadoop platform for scientific data analysis," in *Proc. 4th ACM Workshop on Scientific Cloud Computing (ScienceCloud '13)*, pp. 13–20, 2013.

[49] Y. Liu, Y. Wang, and Y. Jin, "Research on the improvement of MongoDB auto-sharding in cloud environment," in *Proc. 7th Int. Conf. Computer Science & Education (ICCSE '12)*, pp. 851–854, 2012.

[50] G. Holmes, A. Donkin, and I. H. Witten, "Weka: a machine learning workbench," in *Proc. 2nd Australian and New Zealand Conf. on Intelligent Information Systems*, pp. 357–361, 1994.

[51] K. Ericson and S. Pallickara, "On the performance of high dimensional data clustering and classification algorithms," *Future Generation Computer Systems*, vol. 29, no. 4, pp. 1024–1034, 2013.

[52] R. M. Esteves, R. Pais, and C. Rong, "K-means clustering in the cloud—a Mahout test," in *Proc. IEEE Workshops of Int. Conf. on Advanced Information Networking and Applications*, pp. 514–519, 2011.

[53] F. Liang, C. Feng, X. Lu, and Z. Xu, "Performance benefits of DataMPI: a case study with Big Data Bench," *Big Data Benchmarks, Performance Optimization, and Emerging Hardware*, *Lecture Notes in Computer Science*, pp. 111–123, 2014.

[54] M. Nemschoff, *Big Data: 5 Major Advantages of Hadoop*, http://www.itproportal.com/ [online].

[55] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, "Big Data challenge: a data management perspective," *Frontiers of Computer Science*, vol. 7, no. 2, pp. 151-164, 2013.

[56] L. Xia, R. Luo, B. Zhao, Y. Wang, and H. Yang, "An accurate and low-cost PM2.5 estimation method based on artificial neural network," in *Proc. 20th Asia and South Pacific Design Automation Conference*, pp. 190-195, 2015.

[57] R.K. Jain, J.M.F. Moura, and C.E. Kontokosta, "Big data+big cities: Graph signals of urban air pollution," *IEEE Signal Processing Magazine*, pp. 130-136, 2014.

[58] Y.L. Tan, V. Sehgal, and H.H. Shahri, "SensoClean: Handling noisy and incomplete data in sensor networks using modeling," *Technical Report: University of Maryland*, 2005.

[59] L. Xiang, J. Luo, and C. Rosenberg, "Compressed data aggregation: Energy-efficient and high-fidelity data collection," *IEEE/ACM Trans. Networking*, vol. 21, no. 6, pp. 1722-1735, 2013.

[60] S. Gonzalez-Valenzuela, M. Chen, and V.C.M. Leung, "Mobility support for health monitoring at home using wearable sensors," *IEEE Trans. Infor. Tech. in Biomedicine*, vol. 15, no. 4, pp. 539-549, 2011.

[61] S.J. Pan, D. Shen, Q. Yang, and J.T. Kwok, "Transferring localization models across space," in *Proc. Twenty-Third AAAI Conf. Artificial Intelligence*, pp. 1383-1388, 2008.

[62] V.W. Zheng, D.H. Hu, and Q. Yang, "Cross-domain activity recognition," in *Proc. 11th Int. Conf. Ubiquitous Computing*, pp. 61-70, 2009.

[63] P. Rashidi and D.J. Cook, "Activity recognition based on home to home transfer learning," in *Proc. 24th AAAI Conf. Artificial Intelligence*, pp. 45-52, 2010.

[64] F. Zhuang, P. Luo, H. Xiong, Y. Xiong, Q. He, and Z. Shi, "Cross-domain learning from multiple sources: A consensus regularization perspective," *IEEE Trans. Knowledge Data Engineering*, vol. 22, no. 12, pp. 1664-1678, 2010.

[65] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M.

Patel, R. Ramakrishnan, C. Shahabi, "Big data and its technical challenges," *Communications ACM*, vol. 57, no. 7, pp. 86-94, 2014.

[66] Y. Cheng, X. Li, Z. Li, S. Jiang, and X. Jiang, "Fine-grained air quality monitoring based on gaussian process regression," *LNCS 8835*, pp. 126-134, 2014.

[67] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, Jan Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive and Mobile Computing*, vol. 16, pp. 268-285, 2015.

[68] Y. Zheng, F. Liu, and H-P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1436-1444, 2013.

[69] Y. Zheng, X. Chen, Q. Jin, Y. Chen, X. Qu, X. Liu, E. Chang, W-Y. Ma, Y. Rui, and W. Sun, "A cloud-based knowledge discovery system for monitoring fine-grained air quality," *Microsoft Technical Report*.

[70] Y. Yu, D. Ganesan L. Girod, D. Estrin, and R. Govindan, "Synthetic data generation to support irregular sampling in sensor networks," *Geosensor Networks*, pp. 211–234, 2004, CRC Press.

[71] A. Jindal and K. Psounis, "Modeling spatially correlated data in sensor networks,"*ACM Transactions on Sensor Networks*, vol. 2, no. 4, pp. 466–499, 2006.

[72] S. Yu, L.-C. Tranchevent, B. De Moor, and Y. Moreau, *Kernel-based Data Fusion for Machine Learning*, Springer-Verlag, 2011.

[73] A. Overeem, H. Leijnse, and R. Uijlenhoet, "Country-wide rainfall maps from cellular communication networks," *PNAS*, vol. 110, no. 8, pp. 2741-2745, 2013.

[74] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PLoS ONE*, vol. 10, no. 3, pp. 1-17, 2015.

[75] Y. Lv, Y. Duan, W. Kang, Z. Li, and F-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865-873, 2015.

[76] D. Zhang, T. Hei, Y. Liu, S. Lin, and J. Stankovic, "A carpooling recommendation system for taxicab services," *IEEE Trans. Emerging Topics in Computing*, vol. 2, no. 3, pp. 254-266, 2014.

[77] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "#Earthquake: Twitter as a distributed sensor system," *Transactions in GIS*, vol. 17, pp. 124–147, 2013.

[78] X. Song, Q. Zhang, Y. Sekimoto, T. Horanont, S. Ueyama, and R. Shibasaki, "Intelligent system for human behavior analysis and reasoning following large-scale disasters", *IEEE Intelligent Systems*, vol.28, no. 4, pp. 35-42, 2013.

[79] T. Horanont, A. Witayangkurn, Y. Sekimoto, and R. Shibasaki, "Large-scale Auto-GPS analysis for discerning behavior change during crisis", *IEEE Intelligent Systems*, vol. 28, no. 4, pp. 26-34, 2013,

[80] A. Capozzoli, F. Lauro, and I. Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings", *Expert Systems with Applications*, vol. 42, no. 9, pp. 4324-4338, 2015.

[81] J-S. Chou and N-T. Ngo, "Smart grid data analytics framework for increasing energy savings in residential buildings", *Automation in Construction*, 2016.

[82] A. Khan, Sk. K. A. Imon, and S. K. Das, "A novel localization and coverage framework for real-time participatory urban monitoring", *Pervasive and Mobile Computing*, vol. 23, pp. 122-138, 2015.

[83] A. Weiler, M. Grossniklaus, M. H. Scholl, "Situation monitoring of urban areas using social media data streams", *Information Systems*, vol. 57, pp. 129-141, 2016.

[84] J. Froehlich, J. Neumann, and N. Oliver, "Measuring the pulse of the city through shared bicycle programs," in *Proc. Int Workshop Urban, Community, and Social Applications of Networked Sensing Systems*, pp. 16-20, 2008.

[85] http://sensingcity.org/. [Online].

[86] http://amsterdamsmartcity.com/projects/detail/id/58/slug/smart-traffic-management/. [Online]

[87] R. Becker, R. Caceres, K. Hanson, S. Isaacman, J.M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Communications of the ACM*, vol. 56, no. 1, pp. 74-82, Jan. 2013.

[88] C. Outram, C. Ratti, and A. Biderman, "The Copenhagen wheel: An innovative electric bicycle system that harnesses the power of real-time information and crowd sourcing," in *Proc. EVER Monaco Int. Exhibition & Conf. on Ecologic Vehicles & Renewable Energies*, Mar. 2010.

[89] F. Girardin, J. Blat, F. Calabrese, F.D. Fiore, and C. Ratti, "Digital footprinting: Uncovering tourists with user-generated content," *Pervasive Computing*, pp. 36-43, Oct.-Dec. 2008.

[90] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *IEEE Trans. Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141-151, Mar. 2011.

[91] F.E.A. Horita, J.P. de Albuquerque, L.C. Degrossi, E.M. Mendiondo, and J. Ueyama, "Development of a spatial decision support system for flood risk management in Brazil that combines volunteered geographic information with wireless sensor networks," *Computers & Geosciences*, vol. 80, pp. 84-94, 2015.

[92] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions", *Future Generation Computer Systems*, pp. 1645-1660, 2013.

[93] J. Miranda, N. Makitalo, J. Garcia-Alonso, J. Berrocal, T. Mikkonen, C. Canal, and J.M. Murillo, "From the Internet of Things to the Internet of People", *IEEE Internet Computing*, pp. 40-47, 2015.

[94] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of Things: A survey", *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 414-454, 2014.

[95] D. Gil, A. Ferrandez, H. Mora-Mora, and J. Peral, "Internet of Things: A review of surveys based on context aware intelligent services", *Sensors*, pp. 1-23, 2016.

[96] S. Sen, "Context-aware energy-efficient communication for IoT sensor nodes", in *Proc. ACM/EDAC/IEEE Design Automation Conference*, pp. 1-6, 2016.

[97] M. Donohoe, B. Jennings, and S. Balasubramaniam, "Context-awareness and the smart grid: Requirements and challenges", *Computer Networks*, vol. 79, pp. 263-282, 2015.

[98] A. Alamri, W.S. Ansari, M.M. Hassan, M.S. Hossain, A. Alelaiwi, and M.A. Hossain, "A survey on sensor-cloud: Architecture, applications, and approaches", *Int. Journal Distributed Sensor Networks*, pp. 1-18, 2013.

[99] S. Madria, V. Kumar, and R. Dalvi, "Sensor cloud: A cloud of virtual sensors", *IEEE Software*, pp. 70-77, 2014.

[100] A. Sen, and S. Madria, "Risk assessment in a sensor cloud framework using attack graphs", *IEEE Trans. Services Computing*, 2016. [in press]

Li-Minn Ang received his BEng and PhD degrees from Edith Cowan University in Australia. He is currently attached to the School of Computing & Mathematics at Charles Sturt University (CSU) in Australia. He is also a research leader of Intelligent Analytics & Sensing (IAS) group in CSU. He was previously an associate professor at Nottingham University. His research interests are in visual information processing, embedded systems & WSN, reconfigurable computing, the development of real-world computer systems, large-scale data gathering in sensor systems, big data analytics for sensor networks, and multimedia IoT. He has published over a hundred papers in journals and international refereed conferences. He is a senior member of the IEEE and a Fellow of the Higher Education Academy (UK).

Kah Phooi Seng received her BEng and PhD degrees from the Tasmania University in Australia. She is currently the Adjunct Professor in the School of Computing and Mathematics at CSU. Before returning to Australia, she was a Professor at Sunway University. Before joining Sunway, she was an Associate Professor in School of Electrical and Electronic Engineering at Nottingham University. Her research interests are in intelligent visual processing, multimodal signal processing, AI, multimedia WSN, affective computing and multimodal Big Data analytics. She has published over 230 papers in journals and international refereed conferences.

Adamu Murtala Zungeru received the PhD, MSc & BEng from Nottingham University, Ahmadu Bello University Zaria Nigeria and Federal University of Technology Minna Nigeria respectively. He was a Research Fellow in the Electrical Engineering and Computer Science Department at MIT USA. He is currently a senior lecturer at Botswana International University of Science & Technology (BIUST).

Gerald K. Ijemaru received his Bachelors of Engineering (B.Eng.) degree in Electrical & Electronics Engineering from Enugu State University of Science & Technology (ESUT), Nigeria in 2004, and Master of Science (M.Sc.) Degree from Coventry University, UK in 2010. He is a Lecturer in the Department of Engineering, Federal University Oye-Ekiti, Nigeria.