

Chapter 16

Smart Urban Metabolism: A Big-Data and Machine Learning Perspective



Ruchira Ghosh and Dipankar Sengupta

Abstract Smart urban metabolism is a contemporary conception of urban metabolism which includes modern-day technologies dealing with the complex challenges of growing smart cities. Traditionally, urban metabolism deals with the influx-efflux of energy and flow of materials through urban space. However, with the growing needs of smart cities, these flow patterns are transiting as a complex network and are subject to interdisciplinary understanding. Furthermore, data availability is a major challenge faced by city planners due to the lack of data inventories and appropriate data management solutions to handle massive datasets, arising from these complex flow patterns. This is ensuing to inefficient adaptation of urban metabolism approaches, especially in developing economies. Thus, the situation remains grave when it comes to resource management of a smart city, and how urban areas may additionally deal with intricate issues like climate change when they are striving to understand their own material and energy cycling. In this chapter, we therefore, discuss how technologies like machine learning can equip urban metabolism, for its transition to “Smart Urban Metabolism.” The chapter presents use of technologies like big-data and machine learning, as effective methodologies to channelize and manage heterogeneous multidimensional datasets, adoption of practices, developing self-learning machine learning models, and gain novel insights via predictive analytics, in “Smart Urban Metabolism.” Precisely, for urban planners, the “Smart Urban Metabolism” can potentially be an effective approach for identifying complex issues in the flow patterns of energy and material in an urban space. This approach is a step toward sustainable city development.

Keywords Big-data analytics · Machine learning · Smart cities · Smart urban metabolism · Sustainable development · Urban metabolism

R. Ghosh (✉)

Centre for Sustainable Technologies, Belfast School of Architecture and the Built Environment, Ulster University, Jordanstown BT37 0QB, UK

e-mail: r.ghosh@ulster.ac.uk

D. Sengupta

Health Data Sciences Research Group, Centre for Optimal Health, School of Life Sciences, College of Liberal Arts and Science, University of Westminster, London W1W 6UW, UK

16.1 Introduction

Urban metabolism is a mandatory concept for growing cities to understand the complex network of urban needs and their resource allocation. However, with the expansion of cities, augmented urban needs, urban mining, and resource management comes huge multifaceted data to manage, along with its dynamism. The initial concept of urban metabolism can be traced back to the nineteenth century (Wolman 1965; Odum 1996). In the current era, it is best defined to be, “*a concept in which the city is using the biological notion referring to the internal processes by which living organisms maintain a continuous exchange of matter and energy with their environment to enable operation, growth, and reproduction*” (Céspedes Restrepo and Morales-Pinzón 2018). Globally, the cities cover < 2% of the Earth’s surface, however, consume ~78% of the energy (food, construction minerals, metals, etc.) including the energy consumption demands to generate these resources (Ulgiati and Zucaro 2019). Moreover, cities offer citizens with new opportunities for business, social security, education, and health services, that requires vast resource flows, within and outside the city boundaries. This exchange of energy and materials, to satisfy human settlement and sustenance, gets most often translated into environmental stress posing impacts as micro as locally in the city and as huge as at global scale, due to lack of resource channelization. The majority of the world’s economic activities are concentrated in urban areas, generating 80% of the global gross domestic product (GDP), which demands a greater part of this energy consumption to support these economic activities (Ferrão and Fernández 2013; UN-Habitat 2022).

With every fold increase in the population, migration, economic growth, and social changes, the resource metabolism pattern/trend will change, and therefore, with changing urban configurations, urban metabolism alone is not sufficient to understand and address the challenges of today’s cities or urban areas. Thus, cities need smart planning and a smart system leading to “Smart Urban Metabolism” (SUM) to grow smart in its functioning. SUM is a contemporary conception of urban metabolism which includes modern-day technologies like big-data machine learning, etc., that aids in dealing with the complex data challenges of growing smart cities. Although urbanization and globalization have been accelerated by technological advancements, however, to attain urban sustainability (UN Sustainable Developmental Goal (SDG) 11—sustainable cities and communities) (United Nations 2015), we also need to consider environmental, social, and economical challenges as precedence. This will help to assess the complexity of urban metabolic processes/systems/services for sustainable human settlement. Therefore, digitalization of urban material-energy flow patterns has become a priority for the development of urban settlements; and consequently, the concept of SUM will play a key role in achieving the objective (Caragliu et al. 2009; Dameri 2013; De Jong et al. 2015; Yu and Xu 2018).

The concept of SUM was proposed to deal with the constraints of urban metabolism (Bibri and Krogstie 2020). Furthermore, this advanced concept integrates both urban smart configuration and urban metabolism, emphasizing both the aspects in a holistic manner (Kitchin et al. 2015; Vinod Kumar and Dahiya 2017;

Macke et al. 2019). In recent years, the assessment of some city's (like Copenhagen, Singapore, Melbourne, etc.) utility services have been equipped with smart solutions, for example, real-time monitoring stations supported by big-data and participatory planning scheme—to augment urban service provisioning (i.e., with respect to water, waste, energy, emissions management, etc.) (Bettencourt 2014; Yigitcanlar et al. 2019). Within such a framework, local authorities, urban utility service providers, academics and research centers, manufacturers and companies, NGOs, and so on are required to work in synergy and actively (Paskaleva 2009; Longa 2011; Glazebrook and Newman 2018). Therefore, integrating natural and artificial intelligence into a unified and coherent multi-facet structure has become a priority and essential for policymakers globally. And thus, the urban data collected from real-time monitoring stations and analyzing them using big-data, machine learning, etc., would benefit in establishing an improved understanding of the city's key functions and performances (Bibri and Krogstie 2020).

Sustainable urban metabolism (SUM) study is a hybrid methodology and multi-dimensional in approach that involves economic and social perspectives of cities, environmental challenges, and technological options to deal with them together in harmony to develop smart and sustainable cities (Goal 11 of the United Nations 2030 Agenda for Sustainable Development encompasses targets to “make cities and human settlements safe, resilient and sustainable”) (United Nations 2015). This distinctive character makes SUM a strategic tool to assist urban policymakers, resource managers, and city planners. Aiding urban metabolism with big-data technologies and machine learning (Fig. 16.1) would therefore enable to adopt a data-driven methodology and build up knowledge from the systems investigation in an iterative way.

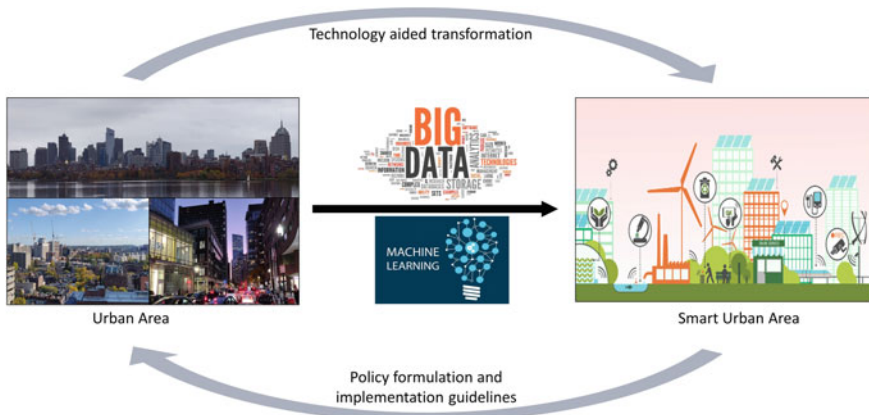


Fig. 16.1 Modern technology like big-data technologies and machine learning can aid landscaping urban metabolism into smart urban metabolism, and thus, assisting the city planners in policy formulation considering the future demands

Data are everywhere and growing in terms of volume, velocity, and variety, which cannot be handled with traditional data processing software, and therefore, is called as the big-data where the volume, velocity, and variety are referred to as the 3Vs of big-data. Estimates suggest, that currently, 2.5 quintillion bytes of data are generated globally every day, which will rise to 463 exabytes by 2025 (Vuleta 2021). And, therefore, big-data technologies have a huge role to play in managing as well as equipping us with the right tools to analyze these datasets for purposes like mining non-trivial patterns, classification, prediction or forecasting, etc. In SUM too, for transforming an “urban area” into a “smart urban area,” big-data and the associated technologies hold the key to planning, policy formulation, implementation, and governance (Kandt and Batty 2021). As it encapsulates the computing power to capture and process the real-time data (from components of an urban system like transportation, water consumption, electricity consumption, etc.), and store it historically, also provides with power to detect patterns or make predictions in real time, augmenting the quick decision-making capabilities.

On the other hand, machine learning (ML) provides the extension leading to establishing the analytics component of big-data technologies. ML has been at the forefront since the 1960s, with an aim of designing machines (i.e., computer programs) that can mimic human intelligence (Samuel 1959). The objective set for its development was to attain the computational capabilities for extracting patterns from datasets, making inferences, and using them for decision-making, which is difficult to be achieved from the prevailing statistical methods (Sengupta 2021). In the 1990s, Tom Mitchell defined, ML as “*a computer program learning from experience ‘E’ with respect to some class of tasks ‘T’ and performance measure ‘P’, if its performance at tasks in ‘T’ as measured by ‘P’, improves with experience ‘E’*” (Mitchell 1997). However, functionally in simplistic terms, ML is said to be enabling a computer program (compounding the powers of computer science, mathematics, and statistics) to learn from the data, for example, a computer program recommending products to a customer based on shopping preference along with additional parameters like purchase history, geographical location, ethnic origin, age, gender, occupation, etc. Similarly, in terms of urban metabolism, prediction of future electricity consumption, and water consumption, waste generation pattern can be acquired based on household parameters like socio-economic diversity, and goods’ consumption pattern. In both these situations, the computer program is learning from the available data, and its prediction performance improves as more data are made available to it.

Consequently, with the growing data in different realms, as well as the introduction of big-data technologies, the evolution of ML, and advancements in computing techniques (like, as parallel or distributed computing), their applicability has augmented across domains, ranging from art, health care, humanities, social science, to more philosophical and ethical studies. Similarly, they are also changing the way cities are evolving, bringing in historical and real-time data as well as multiple time scales to be considered, raising the prospects of making them smart and sustainable. And therefore, understanding the conventional urban metabolism with technologies like big-data technologies and ML to attain smartness in a city’s functioning, provides a greater opportunity for urban functions like the flow pattern of energy and materials

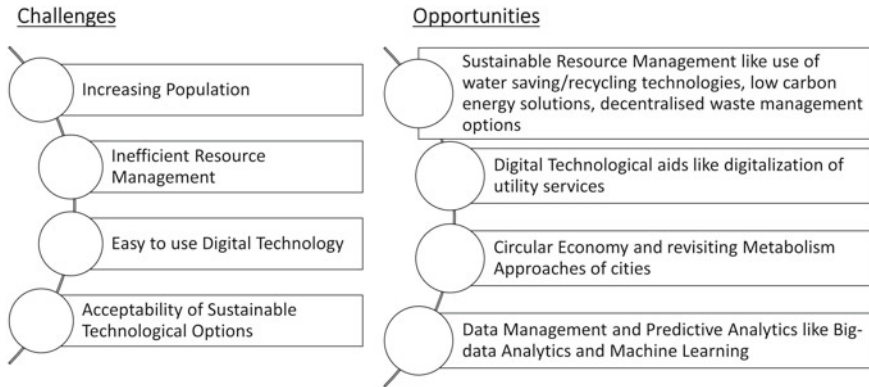


Fig. 16.2 Challenges and opportunities for smart urban metabolism (SUM)

to be assessed through a data-driven lens. Maintaining the urban ecosystem services is vital to sustainable urban development, emphasizing urban system resilience and ensuring public health and well-being. Thus, SUM is the answer to today's urban complexities by building the capacities of urban planners to sustainably design and plan for smart city infrastructure and services.

16.2 Barriers and Opportunities of Smart Urban Metabolism (SUM)

SUM faces several challenges to be adopted by the urban planners (Fig. 16.2); however, at the same time, the complex system in an urban area provides with opportunity to be adapted for future city planning.

16.3 Population and Smart Urban Metabolism Challenges

Better lifestyles and increased job opportunities are driving the population toward cities. And therefore, more and more migration is resulting in augmented resource demand followed by a fold increase in the waste production (with every unit increase in the resource demand and consumption, the waste production escalates, establishing a direct relationship between migration and waste production). This is usually an outcome of a lack of adequate knowledge about the pattern a particular city shows with respect to its relationship between the population influx rate and urban utilities demand and consecutively waste emerging out from the metabolism of these resources (Demaria and Schindler 2016; Facchini et al. 2017; Ipsen et al. 2019). However, the lack of comprehension of urban planners is mainly because they are not

well equipped to recognize and analyze the fluctuating urban complexities (energy and material flows). To address this, technologies like big-data, machine learning, etc., which have the capability to handle huge datasets, assess and predict are required to give an idea about a particular or several aspects of a city (Shahrokni et al. 2015). The implementation of SUM cannot achieve success unless the city planners enhance their capacity to accommodate the migration and its associated aspects into the urban utility service planning and provisioning (Lyons et al. 2018). This is vital because people are at the core of driving a city's fate by determining the metabolic flow of materials and energy and guiding the GDP growth, which has a direct impact on the city's future progression.

16.4 Resource Management

Urban metabolism is gaining popularity among urban policymakers, managers, and planners in one hand. However, because of the large range of indicators in this framework, on the other hand, often is cumbersome for them to choose the right ones for developing, monitoring, and evaluating the metabolic pathways. For this, the standardized framework of the urban indicators can provide a greater opportunity to assist the urban planners. ISO 37120 defines urban indicators as quantitative, qualitative, or descriptive measures that provide pattern and trends for a complex system (International Organization for Standardization 2018). Like, the city indicators can be used to estimate the forte and drawbacks of any city (Purnomo and Prabowo 2016). These performance indicators can assist in recognizing critical areas demanding consideration and also the ones performing good (McCool and Stankey 2004). And they can be combined to set a ranking system, that could be used as a tool to measure the competitiveness of any city system, communicate its marketing strategy, and set up an interactive and responsive city administration (Yu and Xu 2018). A few of these key indicators comes from: (a) The “*Key Performance Indicators for Sustainable Digital Multiservice Cities*” ETSI framework identifies 73 city indicators, (b) ISO 37122 provides a complete set of indicators and methodologies to support policymakers, and (c) ISO 37123 on resilient cities is a tool to monitor progress toward a resilient city (Attmsdmc 2017; International Organization for Standardization 2019a, b). Furthermore, the studies on Smart Cities and Communities add to the knowledge in the context of smart urban metabolism (International Telecommunication Union 2016a). Also, the UNECE-ITU “*Smart Sustainable indicators*” framework (Table 16.1) established by the UN in association with the International Telecommunication Union, and others evaluate the urban smartness and the metabolism pathways of urban energy and material flow to comply with the Sustainable Development Goals (United Nations Economic Commission for Europe (UNECE) 2015).

Information communication technology (ICT) indicators for Smart Cities are prerequisite in planning system by cities to adopt. With huge city expansion to support fluctuating socioeconomic variabilities, the information system needs to be revisited,

Table 16.1 List of resources for smart urban metabolism indicators

Urban standards	Functioning areas	Sectors targeted	Indicators	Urban standards
ISO/DIS 37123	Sustainable development indicators for resilient cities	Economic, Social, Environment	73	ISO/DIS 37123
ISO 37122:2019	Sustainable development indicators for smart cities	Economic, Social, Environment	80	ISO 37122:2019
ISO 37120:2018	Sustainable development indicators for city services and quality of life	Economic, Social, Environment	104	ISO 37120:2018
UNECE - ITU	Sustainable indicators for Smart Cities	Economic, Social, Environment	72	UNECE - ITU
ETSI TS 103 463	Indicators for Smart sustainable cities	Economic, Social, Environment	76	ETSI TS 103 463
ITU-T Y.4903/L.1603	Indicators for smart cities to assess the achievement of SDGs	Economic, Social, Environment	52	ITU-T Y.4903/L.1603
ITU-T Y.4902/L.1602	ICT sustainability impacts indicators in Smart Cities	Economic, Social, Environment	30	ITU-T Y.4902/L.1602
ITU-T Y.4901/L.1601	ICT indicators for Smart Cities	Economic, Social, Environment	48	ITU-T Y.4901/L.1601

Sources United Nations Economic Commission for Europe (UNECE) (2015), International Telecommunication Union (2016a, b, c, d), International Organization for Standardization (2018, 2019a, b)

and the performance indicators designed to give a perfect overview for city planners to adapt (International Telecommunication Union 2016a, b). These indicators primarily include information about smart water meters for water supply monitoring and smart electricity meters to track electricity supply and demand-response penetration for electricity customers, access to household sanitation, solid waste collection, wastewater sewer connections, Internet access at households, public transportation network, traffic monitoring, etc. (International Telecommunication Union 2016a, b) (ITU 2016a, b). However, the corresponding data captured in response to these indicators are massive and will further augment in the future with adding population and changing urban structure. Thus, here comes the vital role of technologies like big-data and machine learning to analyze the huge datasets and keep accessibility of the key indicator values and/or patterns handy for urban planners to plan and execute urban utility services.

The advancement in technologies like big-data and machine learning will be beneficial in increasing our understanding of how eminently smart city operates and how can an emerging smart city adapt to sustain in long run. Literature shows the relationship between these technologies, however, smart decision-making is considered indirect and is usually been facilitated through social innovation (Mahmoud et al. 2022). Although machine learning and/or big data have not been used much in the public sector, or urban utility service provisioning but have a promising aspect to investigate technical, managerial, and policy challenges faced by cities today (Kankanhalli et al. 2019). Megacities require a resilient administrative configuration to expediate synergistic working between the society and government to make better policy implementation and ease in smart decision-making (Torfing et al. 2012; Conway 2020). It necessitates integrating internal governance structures and establishing a public–private partnerships with external organizations (Meijer and Bolívar 2015). Analytics and/or data-driven governance will not only equip cities to create smart services but would also be able to provide smart data inventory through sensors and synthesizing data for urban safety governance (Meijer and Thaens 2018). Like, during the COVID-19 pandemic, the entire world did inventory and synthesized data to fight against the Coronavirus (SARS-CoV2). One such successful enactment was put up by the South Korean government, which utilized data-driven technology for encouraging proactive information exchange and implementation of safety protocols among its citizens (Park et al. 2020).

Moreover, the collective smart governance focusing on “Evolution hubs” by nurturing mutual information flow links between information centers, and research institutions can help in improving the social, economic, and ecological performances of smart cities. For example, Amsterdam Smart City, which has a distinctive bonding between municipal administration, academics and research institutes, start-up investors, entrepreneurs, private businesses, ordinary citizens, and other relative stakeholders knitted through a common thread of information flow. The data-driven technology plays a key role within this system in maintaining the “*urban sustenance lab*,” by smart data collecting, maintaining, storing, and synthesizing as per demand for future planning (Mora and Bolici 2017). One of the vital steps which Amsterdam adopts is engaging communities through the “Smart Citizen” program. It encourages the residents to participate as data representatives, demonstrating their knowledge of sustainability issues, whereas the data-driven technologies are the “Data Hub Manager.” This is one of the best examples exhibiting a collective contribution of humans with technology and ecology toward a city’s advancement as a smart city.

16.5 Big-Data Technology and Smart Urban Metabolism (SUM)

Big-data technology may be best defined as the software application(s) for extracting, processing, storing, and analyzing, massive datasets (structured, semi-structured, or

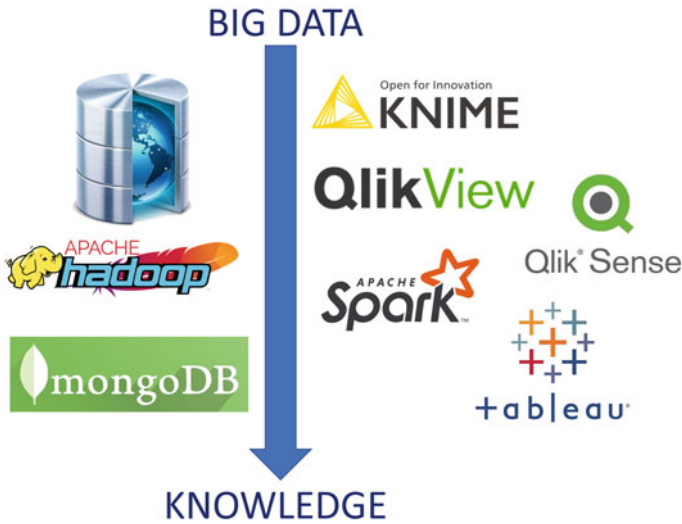


Fig. 16.3 State-of-art big-data technologies—Apache Hadoop and Mongo DB are leading solutions for storage purposes; KNIME and QlikView are among the best for data processing, while Qlik Sense provides cloud-based artificial intelligence (AI) integration platform; Apache Spark and Tableau software’s aid in analytics and visualization

unstructured) which cannot be handled with the conventional software’s, to obtain knowledge. The topmost solutions frequently used for storage, processing, and analysis (using machine learning) of big-data are (Fig. 16.3)—Apache Hadoop and Spark, KNIME, MongoDB, QlikView, RapidMiner, and Tableau, which have been discussed in the next sub-section.

16.6 Big-Data Solutions—Storage, Processing, and Analysis

Apache Hadoop (<https://hadoop.apache.org/>) is open-source software, that provides a framework for large-scale distributed computing (most reliable storage and analysis) and has been implemented by companies like Facebook, LinkedIn, IBM, Microsoft, etc., to handle the massive data (Nandimath et al. 2013). In addition to, Spark (<https://spark.apache.org/>) is one more noteworthy solution from Apache, which was developed considering the benefits of MapReduce (Dean and Ghemawat 2008). It provides an engine supporting data science and machine learning on single-node machines or clusters or on the cloud environment (Zaharia et al. 2016). MongoDB (<https://www.mongodb.com/>) is another open-source platform that facilitates the management of unstructured or semi-structured or volatile data (i.e., changes frequently) and has been adopted by companies like eBay, MetLife, Google, etc. It is a NoSQL

document-oriented database with an easy-to-set-up environment (Bradshaw et al. 2019).

To address the challenges with the processing of a large set of data into any of the storage solutions, data processing operations are required, i.e., commonly referred to as ETL (Extraction, Transformation, and Loading processes) programming. KNIME (<https://www.knime.com/>) is a leading open-source software solution to provide these capabilities within the big-data structure (Jara et al. 2015). It provides an interactive framework that supports: merging and transforming data → modeling and visualizing → deploying and managing the model → interacting and adapting the model in the real world. Another leading solution available in the market for this objective is QlikView (<https://www.qlik.com/us/products/qlikview>). It is based on an associative analytics engine and centered on data modeling, that aids in deriving relationships between the data (Troyansky et al. 2015). Qlik has further boosted modern cloud-based analytics with Qlik Sense (<https://www.qlik.com/us/products/qlik-sense>), which augments and enhances human intuition with AI-powered insights (Troyansky et al. 2015). This in turn would boost the transition from passive to active analytics for real-time collaboration and action.

For analysis, besides Apache Spark, the most prominent software solutions for handling big-data are—RapidMiner and Tableau. RapidMiner (<https://rapidminer.com/>) is a powerful data tool for building predictive models with the support of machine learning (including deep learning) (Hofmann et al. 2016) and also supports the integration of the Apache Hadoop framework for storage purposes. Tableau (<https://www.tableau.com/>) is a leading data-driven business intelligence platform for forecasting, decision-making, and strategy implementation (Hoelscher and Mortimer 2018). The major USP of this software is an easy-to-use interface for transforming the raw data into knowledge, without the requirement of any prior programming experience. Additionally, it includes commending in-built options supporting enhanced data visualization in the form of interactive dashboards (Nair et al. 2016).

16.7 Role of Big-Data Solutions in Smart Urban Metabolism (SUM)

In terms of SUM, core essence of smart cities is the integration among utility systems (like electricity, solid waste, water, wastewater, etc.) and capturing of the data being generated. Big-data technologies are therefore becoming a necessity for the efficient functioning of smart as well as sustainable conurbations. The benefit of this would be the granularity (household → street → Particular Area → Urban Area) of the vast urban information that can be further analyzed via techniques like machine learning. Thus, big data equipped SUM can propel opportunities to address local as well as global urbanization challenges, and robust policy formulation to attain the SDG's.

Big-data can play a key role in transforming the way cities are planned considering multiple time scales (Bibri and Krogtie 2020). A ~98% increase in recycling

of food waste by South Korea is one of the landmark examples, which highlights how the use of a data-driven technology can make a greater impact. In Seoul alone, the implementation of “Smart Bins,” equipped with scales and radio frequency identification which could weigh the food waste as disposed, data in turn mapped to the citizen resident cards, have reduced food waste in the city by 47,000 tons (The World Economic Forum 2019). Another such example is the “city of sensors”—Singapore (Poon 2017). In 2014, ~1000 sensors were installed to capture data as part of the government’s initiative to transform Singapore to be the world’s first smart country under the “E3A” plan (Everyone, Everything, Everywhere, All the Time). The aim was to connect data from all aspects of urban life (traffic, infrastructure, etc.), with the data integrated among different departmental systems feeding into a central platform (Poon 2017). The benefit of this was evident during the Covid surge in 2020, as Singapore was one of the first countries to develop a tracing app safeguarding public health (Lee and Lee 2020).

So Bigdata Research Infrastructure, a pan-European initiative, provides services for obtaining, analyzing, and visualizing massive datasets for the ethically safe deployment of big-data analytics (Grossi et al. 2018). Furthermore, the emergence of a big-data platform like “Strategic Intelligence,” which brings in the latest global research and analysis from the leading research organizations, exploring and monitoring more than 100 global issues with tactical perceptions, is paving the way forward in this direction (The World Economic Forum). Symoto, an ongoing project of the Dutch firm—Except Integrated Sustainability, is another such encouraging prospective (Except Integrated Sustainability B.V.). It aims to build a software application that would enable the development, simulation, and monitoring of large-scale material and energy cycles, industrial symbiosis, and the circular economy. Thus, similar examples of system modelling and strategic decision making can be replicated worldwide to empower SUM and attain sustainable development globally.

To summarize, a few of the key benefits which big-data technologies can offer for SUM:

1. Federated and centralized data storage—for gathering time-centric data recorded by independent utilities, and then merging them at a granular level (each individual/customer) as a centralized storage system.
2. Predictive analysis—for smart planning for basic utilities like water, electricity, traffic, etc.
3. Temporal analytics—for future planning with every fold increase in population, the fold-change (increase/decrease) in demand for basic utilities.
4. Mapping resources—considering the historical data, current demands, and future trends.
5. Planning—devising energy-efficiency programs that aid in improving urban health and security.
6. Cost optimization—in quantitative terms while planning and designing new infrastructure, toward smart and sustainable utility planning.

7. Citizen engagement—promoting awareness, creativity, and innovation among the general populations to be the drivers of the change, as well as encouraging service providers with open data platforms.
8. Aid policymakers—understand the present and future need for policy support in smart city infrastructure and enforce policies accordingly.

16.8 Machine Learning (ML) and Its Role in the Smart Urban Metabolism (SUM)

Machine learning (ML) primarily comprises learning algorithms (computer programs), which are used to analyze datasets (Fig. 16.4). It is applied to the input data, for predicting the corresponding output values or identifying patterns, or for grouping the data based on closest properties (for example, Euclidean distance), with an acceptable performance value which can be improvised based on experience.

The “supervised learning” algorithms (for example, classification, regression, etc.) are used for the datasets in which the class labels (i.e., labeled output) are known, based on which a learning function is obtained by approximating the function for mapping set of input features with the output (Caruana and Niculescu-Mizil 2006). While the “unsupervised learning” algorithms (for example, clustering, association rule mining, etc.) are used for datasets which do not have any class labels (i.e., no labeled outputs), and therefore, the aim of the learning function is to infer the organic construct (like grouping or precedence of occurrences) within a set of closest properties (Ghahramani 2004). In the third category of “semi-supervised” algorithms (for example, reinforcement learning, grey-box modeling), the learning function is trained on a dataset with few class labels and then uses the knowledge to label the unlabeled data points of the dataset (Sinha 2014).

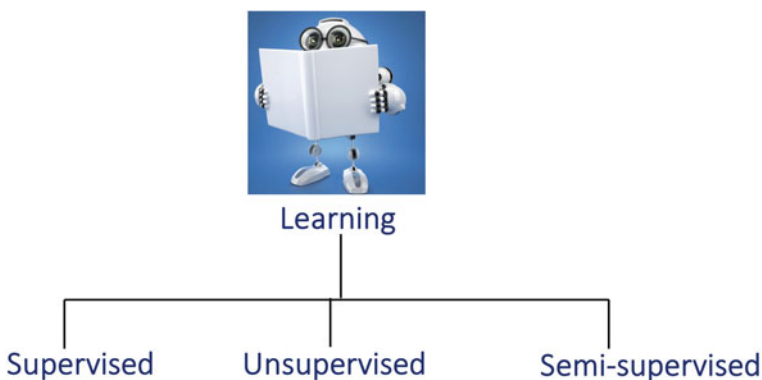


Fig. 16.4 There are three types of learning algorithms based on which ML can be categorized as: supervised, unsupervised, and semi-supervised

16.9 Building a Machine Learning (ML) Model for SUM

For building a robust ML model within a SUM environment (Fig. 16.5), the foremost step is to capture the data and store it historically in a big-data storage structure like Apache Spark or MongoDB (as discussed in the previous section). As the data sources are federated in nature (Architecture, Electricity, Traffic, Water, etc.), they can be stored in distinct data systems and/or maybe unified to be stored as a centralized system. In either case, the first step toward building an ML model would be preprocessing the data (i.e., preparing the data). This is an important step, as to ensure data are without any noise or bias.

Preprocessing the data may involve all or either of the following:

1. **Missing Values:** This is one of the commonest problems with datasets and needs to be handled, as it can cause bias and/or reduce the overall performance of the model. Typically, two approaches are used to handle this problem—removing records (instances and/or features) with missing values and the imputation method (Manly and Wells 2015; Hegde et al. 2019). There are different imputation methods (examples—mean/median/mode, interpolation, nearest neighbor, maximum likelihood estimate, etc.) that can be used to replace the missing values (Newgard and Lewis 2015). Another approach can be using a learning algorithm (like Naïve Bayes, k-nearest neighbor) that supports datasets with missing values.
2. **Data Imbalance:** This refers to the problem of class labels not represented equally in the dataset, which can introduce the problem of the model overfitting its predictive capabilities. Most often, this is a common problem, as datasets do not have an equal number of instances for each class label (Thabtah et al. 2020). This can be handled during sampling of instances or while model evaluation using

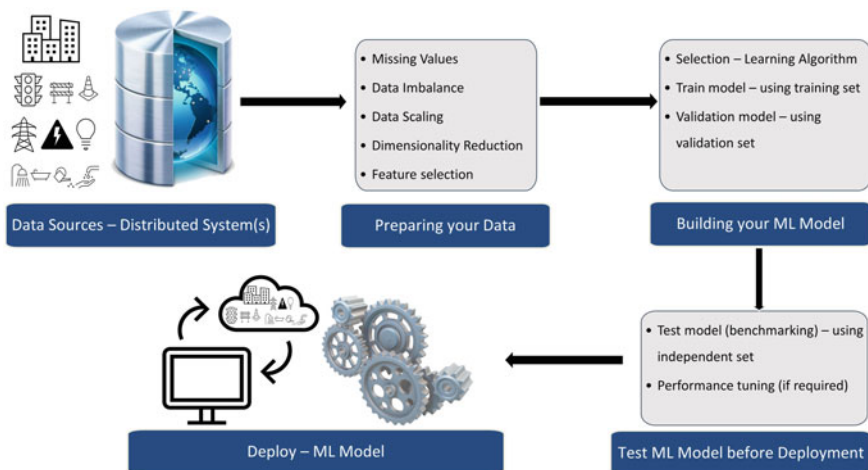


Fig. 16.5 Schematic representation for building a ML model, considering federated data sources in an urban area

performance metrics (like balanced accuracy, Kappa value, receiver operating characteristic (ROC) curve) which consider imbalances of classes in the data or via techniques like grey-box modeling (Grau et al. 2016).

3. **Data Scaling:** In some cases, the features in the dataset may occur in different ranges or scales (for example, traffic movement may be measured in km/hour, whereas the electricity consumption in KWH) and therefore are not ideal to build a model (using algorithms like regression based, support vector machine (SVM), etc.) as this may impact its overall performance. In such scenarios, scaling methods are used to generalize the data points and minimize the difference among these features. The following methods are commonly used and recommended for data scaling purposes—MaxAbs, MinMax, normalization, quantile transformation, robust scaling, and standardization (Ahsan et al. 2021).
4. **Feature Selection:** Big-data imply a large number of features to be considered in the data to build a model, which would trigger the complexity of learning to grow exponentially and impact the overall performance of the model. The feature selection step helps to reduce the number of features, i.e., creates a subset of relevant features to build a model from the original set of features. There is supervised and unsupervised feature selection measure that can be applied as a filter (selected before the ML algorithm) or wrapper approach (selected as part of the ML algorithm) (Marsland 2009; Cai et al. 2018).
5. **Dimensionality Reduction:** aims a similar focus as feature selection, however, the technique also helps to reduce the dimensional space of selected features (like generation of new synthetic features from a combination of the original features and then removing the least significant ones). This technique is also often used for data visualization for complex datasets having many features. Correlation, linear discriminant analysis (LDA), matrix factorization, and principal component analysis (PCA) are some of the commonest methods used for the aforesaid process (Marsland 2009; Reddy et al. 2020).

Once the data are prepared, it should be split into three parts—training, validation, and test sets (ideally, during the testing phase, the model should be benchmarked also with independent datasets) (Marsland 2009). To start with, 10% of this data should be split out as the “test set” (this sample of data is used to provide an unbiased evaluation of the final model) whereas the remaining dataset should be split-out into either 2:1 or k-cross folds, for “training” (sample of the data used to train the model function) and “validation” (this sample of data is used to provide an unbiased evaluation of the model fit from the training set while tuning the model hyperparameters) sets, respectively (Mitchell 1997; Marsland 2009).

Thereafter, an appropriate learning algorithm needs to be selected based on the task (like prediction, pattern identification, etc.). For example, neural network or SVM (both supervised learning) can be used for the prediction of energy consumption (Shapi et al. 2021). The “training” set should be used to train the model, followed by its “validation” with the validation set. Thereafter, the model should be tested upon the “test” set and/or benchmarked against independent datasets. Consequently, the optimized ML model can be implemented in the existing SUM ecosystem, for

monitoring real-time activities or decision-making purposes, or the outcomes may be used by urban city managers (or the government) for planning, formulation, and implementation of new guidelines.

16.10 Example of Machine Learning (ML) in Smart Urban Metabolism (SUM)

We have made a basic understanding of how to build an ML model for the SUM ecosystem. Next, we discuss with an example how an ML model designed over a historical dataset can influence in transforming a city into a smart city.

Water is one of the key urban utilities, and its responsible consumption is of utmost importance for smart cities in addressing sustainability challenges, like corresponding energy consumption (Ghosh et al. 2016), which in turn may regulate the health service demands. In this work, Sengupta and Ghosh focused on looking at patterns between residential water consumption, wastewater discharge, and their impact on electricity consumption, with New Delhi (megacity) as a case study (Sengupta and Ghosh 2022). They analyzed the available data for the period 2001–2011 (Kennedy et al. 2015) using ML techniques enabling learning from the data. The predictor models were designed on two best-performing algorithms—logistic regression (generalized linear model) and neural network (a state-of-the-art technique for deep learning applications), using the MLR R-package (Rumelhart et al. 1986; Bischl et al. 2016; Tolles and Meurer 2016).

As the ROC curves (Fig. 16.6) and comparison of the performance measures (Table 16.2) between the two models demonstrate, model-A (based on logistic regression) outperformed model-B (based on the neural network). Using model-A, further analysis (Fig. 16.7) indicated, that there has been a two-fold increase in the electricity requirement with respect to water consumption and wastewater discharge. Such a model, when combined with socio-economic drivers and geographical details, has the potential to be used on a temporal scale to predict fold-change in the demands (water consumption, wastewater discharge, and corresponding electricity consumption). Therefore, this can be a potential interest of application for the smart city planners, as these predictions can be used or further optimized, while designing the water and wastewater management infrastructures, toward a smart and sustainable utility planning.

Thus, such ML models can help focus on enhancing and maintaining urban health by smart planning of basic utility services like water and electricity. The findings may also target mitigating augmentation in residential electricity consumption caused due to coping strategies to deal with poor water and wastewater services. Such studies may also benefit policymakers to understand the present and future need for policy support in smart city infrastructure and enforce policies accordingly.

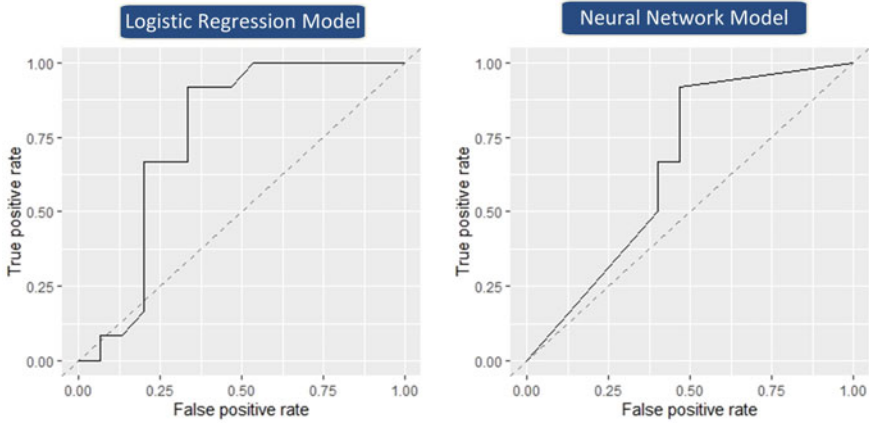


Fig. 16.6 ROC (receiver operating characteristic curve) comparing the performance of the logistic regression and neural network models

Table 16.2 Evaluation parameters for the two ML models

Model	Sensitivity (TPR)	Specificity (TNR)	False positive rate (FPR)	False negative rate (FNR)	Balanced accuracy	MCC
Logistic regression	0.92	0.53	0.46	0.08	0.73	0.47
Neural network	0.67	0.60	0.40	0.33	0.63	0.27

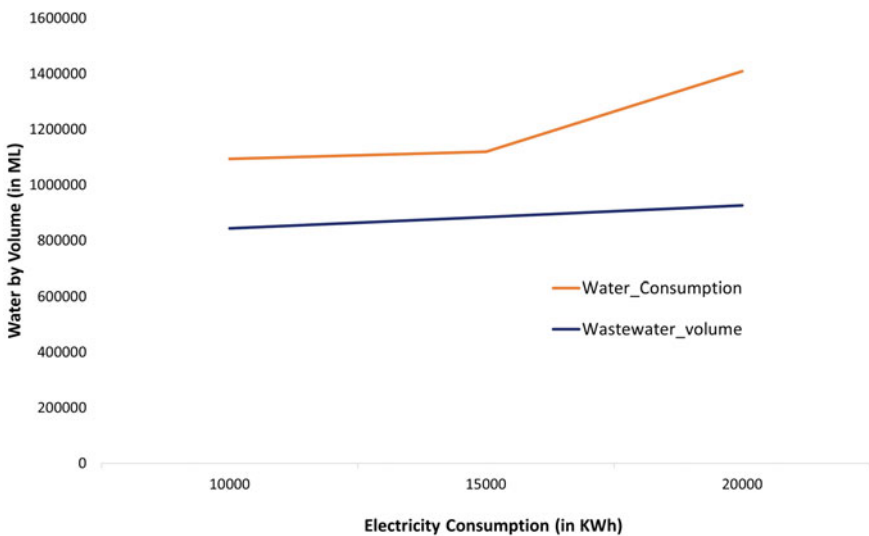


Fig. 16.7 Delhi—electricity consumption versus water consumption and wastewater volume (2001–2011)

16.11 Conclusions

The technologies discussed in this chapter are not the solitary solution to address the challenges in urban areas. However, they are certainly among the strongest tools to increase the efficacy of robust policy implementation and its effective maintenance by actively involving stakeholders through technology amalgamated with social innovation. In conclusion, the combined approach of urban metabolism with advanced technologies like big-data and machine learning is the way forward that will be principally beneficial in understanding the real-time city functions, as they can help urban city planners or governors to make smart decisions, support cities' socioeconomic factors, as well as smart implementation and accomplishment of policies.

References

- Ahsan MM, Mahmud MAP, Saha PK et al (2021) Effect of data scaling methods on machine learning algorithms and model performance. *Technologies* 9:52. <https://doi.org/10.3390/TECHNOLOGIES9030052>
- Attmsdmc (2017) TS 103,463—V1.1.1—access, terminals, transmission and multiplexing (ATTM); key performance indicators for sustainable digital multiservice cities
- Bettencourt LMA (2014) The uses of big data in cities. *Big Data* 2:12–22. <https://doi.org/10.1089/BIG.2013.0042>
- Bibri SE, Krogstie J (2020) Environmentally data-driven smart sustainable cities: applied innovative solutions for energy efficiency, pollution reduction, and urban metabolism. *Energy Inform* 3:1–59. <https://doi.org/10.1186/S42162-020-00130-8/TABLES/7>
- Bischl B, Lang M, Kotthoff L et al (2016) mlr: machine learning in R. *J Mach Learn Res* 17:1–5
- Bradshaw S, Brazil E, Chodorow K (2019) MongoDB: the definitive guide. O'Reilly Media, Inc.
- Cai J, Luo J, Wang S, Yang S (2018) Feature selection in machine learning: a new perspective. *Neurocomputing* 300:70–79. <https://doi.org/10.1016/J.NEUCOM.2017.11.077>
- Caragliu A, Del CB, Nijkamp P (2009) Smart cities in Europe. In: *Proceedings of the 3rd central European conference in regional science*, pp 49–59
- Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: *ACM international conference proceeding series*
- Céspedes Restrepo JD, Morales-Pinzón T (2018) Urban metabolism and sustainability: precedents, genesis and research perspectives. *Resour Conserv Recycl* 131:216–224. <https://doi.org/10.1016/J.RESCONREC.2017.12.023>
- Conway S (2020) Interactive governance: advancing the paradigm. *Administration* 68:63–68. <https://doi.org/10.2478/ADMIN-2020-0017>
- Dameri RP (2013) Searching for smart city definition: a comprehensive proposal. *Int J Comput Technol* 11:2544–2551. <https://doi.org/10.24297/IJCT.V11I5.1142>
- De Jong M, Joss S, Schraven D et al (2015) Sustainable-smart-resilient-low carbon-eco-knowledge cities; Making sense of a multitude of concepts promoting sustainable urbanization. *J Clean Prod* 109:25–38. <https://doi.org/10.1016/J.JCLEPRO.2015.02.004>
- Dean J, Ghemawat S (2008) MapReduce. *Commun ACM* 51:107–113. <https://doi.org/10.1145/1327452.1327492>
- Demaria F, Schindler S (2016) Contesting Urban metabolism: struggles over waste-to-energy in Delhi, India. *Antipode* 48:293–313. <https://doi.org/10.1111/ANTI.12191>
- Except Integrated Sustainability B.V. Except Integrated Sustainability B.V. <https://except.eco/>. Accessed 29 July 2022

- Facchini A, Kennedy C, Stewart I, Mele R (2017) The energy metabolism of megacities. *Appl Energy* 186:86–95. <https://doi.org/10.1016/J.APENERGY.2016.09.025>
- Ferrão P, Fernández JE (2013) Resource consumption of cities. *Sustain Urban Metab* 23–42
- Ghahramani Z (2004) Unsupervised learning. Lecture notes computer sciences (including Subseries lecture notes artificial intelligence lecture notes bioinformatics). https://doi.org/10.1007/978-3-540-28650-9_5
- Ghosh R, Kansal A, Aghi S (2016) Implications of end-user behaviour in response to deficiencies in water supply for electricity consumption—a case study of Delhi. *J Hydrol* 536:400–408. <https://doi.org/10.1016/J.JHYDROL.2016.03.012>
- Glazebrook G, Newman P (2018) The city of the future. *Urban Plan* 3:1–20. <https://doi.org/10.17645/UP.V3I2.1247>
- Grau I, Sengupta D, Matilde M et al (2016) Grey-box model: an ensemble approach for addressing semi-supervised classification problems. In: *Benelearn 2016: Belgian-Dutch conference on machine learning*
- Grossi V, Rapisarda B, Giannotti F, Pedreschi D (2018) Data science at SoBigData: the European research infrastructure for social mining and big data analytics. *Int J Data Sci Anal* 6:205–216. <https://doi.org/10.1007/S41060-018-0126-X/FIGURES/6>
- Hegde H, Shimpi N, Panny A et al (2019) MICE versus PPCA: missing data imputation in healthcare. *Inform Med Unlocked* 17:100275. <https://doi.org/10.1016/J.IMU.2019.100275>
- Hoelscher J, Mortimer A (2018) Using tableau to visualize data and drive decision-making. *J Account Educ* 44:49–59. <https://doi.org/10.1016/J.JACCEDU.2018.05.002>
- Hofmann M, Klinkenberg R, Safari an OMC (2016) *RapidMiner*. Chapman and Hall/CRC, p 525
- International Organization for Standardization (2018) ISO—ISO 37120:2018—sustainable cities and communities—indicators for city services and quality of life. In: ISO, Geneva. <https://www.iso.org/standard/68498.html>. Accessed 26 July 2022
- International Organization for Standardization (2019a) ISO—ISO 37122:2019—sustainable cities and communities—indicators for smart cities. In: ISO, Geneva. <https://www.iso.org/standard/69050.html>. Accessed 26 July 2022
- International Organization for Standardization (2019b) ISO—ISO 37123:2019—sustainable cities and communities—indicators for resilient cities. In: ISO, Geneva. <https://www.iso.org/standard/70428.html>. Accessed 26 July 2022
- International Telecommunication Union (2016a) L.1601: Key performance indicators related to the use of information and communication technology in smart sustainable cities. In: *Telecommun Stand Sect*. <https://www.itu.int/rec/T-REC-L.1601-201606-I>. Accessed 26 July 2022
- International Telecommunication Union (2016b) L.1602: key performance indicators related to the sustainability impacts of information and communication technology in smart sustainable cities. In: *Telecommun Stand Sect*. <https://www.itu.int/rec/T-REC-L.1602-201606-I>. Accessed 26 July 2022
- International Telecommunication Union (2016c) Y.4903: key performance indicators for smart sustainable cities to assess the achievement of sustainable development goals. In: *Telecommun Stand Sect*. <https://www.itu.int/rec/T-REC-Y.4903/en>. Accessed 26 July 2022
- International Telecommunication Union (2016d) ITU-T recommendation database. In: *Telecommun Stand Sect*. <https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=12884&lang=en>. Accessed 26 July 2022
- Ipsen KL, Zimmermann RK, Nielsen PS, Birkved M (2019) Environmental assessment of smart city solutions using a coupled urban metabolism—life cycle impact assessment approach. *Int J Life Cycle Assess* 24:1239–1253. <https://doi.org/10.1007/S11367-018-1453-9/FIGURES/8>
- Jara AJ, Genoud D, Bocchi Y (2015) Big data for smart cities with KNIME a real experience in the SmartSantander testbed. *Softw Pract Exp* 45:1145–1160. <https://doi.org/10.1002/SPE.2274>
- Kandt J, Batty M (2021) Smart cities, big data and urban policy: towards urban analytics for the long run. *Cities* 109:102992. <https://doi.org/10.1016/J.CITIES.2020.102992>
- Kankanhalli A, Charalabidis Y, Mellouli S (2019) IoT and AI for smart government: a research agenda. *Gov Inf Q* 36:304–309. <https://doi.org/10.1016/J.GIQ.2019.02.003>

- Kennedy CA, Stewart I, Facchini A et al (2015) Energy and material flows of megacities. *Proc Natl Acad Sci U S A* 112:5985–5990. https://doi.org/10.1073/PNAS.1504315112/SUPPL_FILE/PNAS.1504315112.SD01.XLSX
- Kitchin R, Lauriault TP, McArdle G (2015) Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Reg Stud Reg Sci* 2:6–28. https://doi.org/10.1080/21681376.2014.983149/SUPPL_FILE/RSRS_A_983149_SM0001.JPG
- Lee T, Lee H (2020) Tracing surveillance and auto-regulation in Singapore: ‘smart’ responses to COVID-19. *Media Int Aust* 177:47–60. <https://doi.org/10.1177/1329878X20949545>
- Longa RD (2011) *Urban models and public-private partnership*. Springer, Berlin Heidelberg
- Lyon G, Mokhtarian P, Dijst M, Böcker L (2018) The dynamics of urban metabolism in the face of digitalization and changing lifestyles: understanding and influencing our cities. *Resour Conserv Recycl* 132:246–257. <https://doi.org/10.1016/J.RESCONREC.2017.07.032>
- Macke J, Rubim Sarate JA, de Atayde Moschen S (2019) Smart sustainable cities evaluation and sense of community. *J Clean Prod* 239. <https://doi.org/10.1016/J.JCLEPRO.2019.118103>
- Mahmoud M, Perez-Ortiz M, Asad Bokhari SA, Myeong S (2022) Use of artificial intelligence in smart cities for smart decision-making: a social innovation perspective. *Sustainability* 14:620. <https://doi.org/10.3390/SU14020620>
- Manly CA, Wells RS (2015) Reporting the use of multiple imputation for missing data in higher education research. *Res High Educ* 56:397–409. <https://doi.org/10.1007/S11162-014-9344-9/TABLES/1>
- Marsland S (2009) *Machine learning: an algorithmic perspective*, 2nd edn. CRC Press
- McCool SF, Stankey GH (2004) Indicators of sustainability: challenges and opportunities at the interface of science and policy. *Environ Manage* 33:294–305. <https://doi.org/10.1007/S00267-003-0084-4>
- Meijer A, Bolívar MPR (2015) Governing the smart city: a review of the literature on smart urban governance. *Int Rev Admin Sci* 82:392–408. <https://doi.org/10.1177/0020852314564308>
- Meijer A, Thaens M (2018) Quantified street: smart governance of urban safety. *Inf Polity* 23:29–41. <https://doi.org/10.3233/IP-170422>
- Mitchell TM (1997) *Machine learning*. McGraw-Hill Science/Engineering/Math
- Mora L, Bolici R (2017) How to become a smart city: learning from Amsterdam. *Smart Sustain Plan Cities Reg* 251–266. https://doi.org/10.1007/978-3-319-44899-2_15/COVER
- Nair L, Nair L, Shetty S, Shetty S (2016) Interactive visual analytics on big data: tableau vs D3.js. *J e-Learning Knowl Soc* 12
- Nandimath J, Banerjee E, Patil A et al (2013) Big data analysis using Apache Hadoop. *Proc 2013 IEEE 14th Int Conf Inf Reuse Integr IEEE IRI* 700–703. <https://doi.org/10.1109/IRI.2013.6642536>
- Newgard CD, Lewis RJ (2015) Missing data: how to best account for what is not known. *JAMA* 314:940–941. <https://doi.org/10.1001/JAMA.2015.10516>
- Odum HT (Howard T (1996) *Environmental accounting: EMERGY and environmental decision making* 370
- Park Y, Choe Y, Park O et al (2020) Contact tracing during coronavirus disease outbreak, South Korea, 2020. *Emerg Infect Dis* 26:2465. <https://doi.org/10.3201/EID2610.201315>
- Paskaleva KA (2009) Enabling the smart city: the progress of city e-governance in Europe. *Int J Innov Reg Dev* 1:405. <https://doi.org/10.1504/IJIRD.2009.022730>
- Poon L (2017) Singapore, City of Sensors. <https://www.bloomberg.com/news/articles/2017-04-21/privacy-takes-a-backseat-in-singapore-s-vision-of-a-smart-nation>. Accessed 29 July 2022
- Purnomo F, Prabowo H (2016) Smart city indicators: a systematic literature review. *J Telecommun Electron Comput Eng* 8:161–164
- Reddy GT, Reddy MPK, Lakshmana K et al (2020) Analysis of dimensionality reduction techniques on big data. *IEEE Access* 8:54776–54788. <https://doi.org/10.1109/ACCESS.2020.2980942>
- Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323:533–536. <https://doi.org/10.1038/323533a0>

- Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM J Res Dev*
- Sengupta D (2021) Machine learning in precision medicine. *Mach Learn Big Data, IoT Med Inform* 405–419. <https://doi.org/10.1016/B978-0-12-821777-1.00013-6>
- Sengupta D, Ghosh R (2022) Artificial intelligence for addressing smart cities poor urban health. *Second Int Conf Water, Megacities Glob Chang*
- Shahrokni H, Lazarevic D, Brandt N (2015) Smart urban metabolism: towards a real-time understanding of the energy and material flows of a city and its citizens. *J Urb Technol* 22:65–86. <https://doi.org/10.1080/10630732.2014.954899>
- Shapi MKM, Ramli NA, Awaln LJ (2021) Energy consumption prediction by using machine learning for smart building: case study in Malaysia. *Dev Built Environ* 5:100037. <https://doi.org/10.1016/J.DIBE.2020.100037>
- Sinha K (2014) Semi-supervised learning. In: *Data classification: algorithms and applications*
- Thabtah F, Hammoud S, Kamalov F, Gonsalves A (2020) Data imbalance in classification: experimental evaluation. *Inf Sci (ny)* 513:429–441. <https://doi.org/10.1016/J.INS.2019.11.004>
- The World Economic Forum (2019) South Korea once recycled 2% of its food waste. Now it recycles 95% | World Economic Forum. In: *WeForum*. <https://www.weforum.org/agenda/2019/04/south-korea-recycling-food-waste/>. Accessed 27 July 2022
- The World Economic Forum Strategic Intelligence. <https://intelligence.weforum.org/>. Accessed 29 July 2022
- Tolles J, Meurer WJ (2016) Logistic regression: relating patient characteristics to outcomes. *JAMA* 316:533–534. <https://doi.org/10.1001/JAMA.2016.7653>
- Torring J, Peters BG, Pierre J, Sørensen E (2012) Interactive governance: advancing the paradigm. *Interact Gov Adv Paradig* 1–288. <https://doi.org/10.1093/ACPROF:OSO/9780199596751.001.0001>
- Troyansky O, Gibson T, Leichtweis C (2015) QlikView your business : an expert guide to business discovery with QlikView and Qlik Sense
- Ulgiasi S, Zucaro A (2019) Challenges in urban metabolism: sustainability and well-being in cities. *Front Sustain Cities* 1:1. <https://doi.org/10.3389/FRSC.2019.00001>
- UN-Habitat (2022) World Cities Report 2022: envisaging the future of cities | UN-Habitat. <https://unhabitat.org/world-cities-report-2022-envisaging-the-future-of-cities>. Accessed 28 July 2022
- United Nations (2015) THE 17 GOALS | Sustainable Development. In: *United Nations—Dep Econ Soc Aff*. <https://sdgs.un.org/goals>. Accessed 25 July 2022
- United Nations Economic Commission for Europe (UNECE) (2015) UNECE endorses smart sustainable cities indicators to improve performance of cities | UNECE. In: *UNECE*. <https://unece.org/housing-and-land-management/press/unece-endorses-smart-sustainable-cities-indicators-improve>. Accessed 26 July 2022
- Vinod Kumar TM, Dahiya B (2017) Smart economy in smart cities 3–76. https://doi.org/10.1007/978-981-10-1610-3_1
- Vuleta B (2021) How much data is created every day? In: *Seedscientific*. <https://seedscientific.com/how-much-data-is-created-every-day/>. Accessed 29 Jul 2022
- Wolman A (1965) The metabolism of cities. *Sci Am* 213:179–190. <https://doi.org/10.1038/SCIENTIFICAMERICAN0965-178>
- Yigitcanlar T, Kamruzzaman M, Foth M et al (2019) Can cities become smart without being sustainable? A systematic review of the literature. *Sustain Cities Soc* 45:348–365. <https://doi.org/10.1016/J.SCS.2018.11.033>
- Yu W, Xu C (2018) Developing smart cities in China: an empirical analysis. *Int J Public Adm Digit Age* 5:76–91. <https://doi.org/10.4018/IJPADA.2018070106>
- Zaharia M, Xin RS, Wendell P et al (2016) Apache spark: a unified engine for big data processing. *Commun ACM* 59:56–65. <https://doi.org/10.1145/2934664>