Addressing Data Challenges to Drive the Transformation of Smart Cities

EKATERINA GILMAN, University of Oulu, Oulu, Finland
FRANCESCA BUGIOTTI, Paris-Saclay University, CNRS, CentraleSupélec, LISN, Paris, France
AHMED KHALID, Dell Research, Dell Technologies, Cork, Ireland
HASSAN MEHMOOD, PANOS KOSTAKOS, and LAURI TUOVINEN, University of Oulu, Oulu, Finland
JOHANNA YLIPULLI, Aalto University, Espoo, Finland
XIANG SU, Norwegian University of Science and Technology, Trondheim, Norway
DENZIL FERREIRA, University of Oulu, Oulu, Finland

Cities serve as vital hubs of economic activity and knowledge generation and dissemination. As such, cities bear a significant responsibility to uphold environmental protection measures while promoting the welfare and living comfort of their residents. There are diverse views on the development of smart cities, from integrating Information and Communication Technologies into urban environments for better operational decisions to supporting sustainability, wealth, and comfort of people. However, for all these cases, data are the key ingredient and enabler for the vision and realization of smart cities. This article explores the challenges associated with smart city data. We start with gaining an understanding of the concept of a smart city, how to measure that the city is a smart one, and what architectures and platforms exist to develop one. Afterwards, we research the challenges associated with the data of the cities, including availability, heterogeneity, management, analysis, privacy, and security. Finally, we discuss ethical issues. This article aims to serve as a "one-stop shop" covering data-related issues of smart cities with references for diving deeper into particular topics of interest.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → **Data management systems**; • **Computing methodologies** → **Distributed computing methodologies**; • **Computer systems organization** → **Architectures**;

Additional Key Words and Phrases: Big data, smart city, urban computing, machine learning, data analysis

This work has been financially supported by the Academy of Finland UrBOT project (323630); the Academy of Finland 6G Flagship (346208); the Academy of Finland DISC project (332143); the CLEVER Project, funded by KDT JU under Grant agreement No. 101097560; the European Commission Grant IDUNN (101021911), and the Nordic University Cooperation on Edge Intelligence (168043).

Authors' Contact Information: Ekaterina Gilman (Corresponding author), University of Oulu, Oulu, Finland; e-mail: ekaterina.gilman@oulu.fi; Francesca Bugiotti, Paris-Saclay University, CNRS, CentraleSupélec, LISN, Paris, France; e-mail: francesca.bugiotti@lisn.upsaclay.fr; Ahmed Khalid, Dell Research, Dell Technologies, Cork, Ireland; e-mail: ahmed.khalid@dell.com; Hassan Mehmood, University of Oulu, Oulu, Finland; e-mail: hassan.mehmood@oulu.fi; Panos Kostakos, University of Oulu, Oulu, Finland; e-mail: panos.kostakos@oulu.fi; Lauri Tuovinen, University of Oulu, Oulu, Finland; e-mail: lauri.tuovinen@oulu.fi; Johanna Ylipulli, Aalto University, Espoo, Finland; e-mail: johanna.ylipulli@aalto.fi; Xiang Su, Norwegian University of Science and Technology, Trondheim, Norway; e-mail: xiang.su@ntnu.no; Denzil Ferreira, University of Oulu, Oulu, Finland; e-mail: denzil.ferreira@oulu.fi.

ACM 2157-6912/2024/11-ART88
https://doi.org/10.1145/3663482

ACM Reference format:
Ekaterina Gilman, Francesca Bugiotti, Ahmed Khalid, Hassan Mehmood, Panos Kostakos, Lauri Tuovinen, Johanna Ylipulli, Xiang Su, and Denzil Ferreira. 2024. Addressing Data Challenges to Drive the Transformation of Smart Cities. *ACM Trans. Intell. Syst. Technol.* 15, 5, Article 88 (November 2024), 65 pages.
https://doi.org/10.1145/3663482

1 Introduction

Cities play a crucial role as the engines of the economy and centers of connectivity, knowledge, and services [338]. Based on an estimation from the United Nations, 66% of the world's population will live in urban areas by 2050 [247]. Being the centers of growth and innovation, cities need to address significant challenges for the environment protection and citizens' prosperity and living comfort. These challenges become pronounced in large and rapidly growing cities, which concurrently struggle to establish a robust infrastructure to ensure clean air and water, energy supply, food, transportation, efficient waste management, and provisioning of public spaces—vital components for human well-being [213].

Cities are becoming commonly equipped with **Information and Communication Technologies (ICTs)** to improve their resourcing and the quality of life of their inhabitants, ultimately becoming smart cities. The term "smart sustainable city" is used to denote a city that is supported by the widespread adoption and extensive use of advanced ICT, which, coupled with various urban systems and domains and strategic coordination of their intricate interrelations, empowers the city to manage available resources sustainably and efficiently for improved economic and societal outcomes [82]. Cities are becoming smart and sustainable in ways that enable us to monitor, understand, analyze, and plan the city to improve the efficiency, equity, and quality of life for citizens in real time [75].

Smart cities are technologically modern urban areas leveraging networked systems to collect data and data analytics platforms to analyze data. The development of smart cities requires the integration of various subsystems to work together to achieve a common goal, which is a system of systems approach. A system of systems is a collection of independent but interrelated systems that have been developed and are operated to meet a common set of objectives. In the context of smart cities, a system of systems integrates multiple subsystems, such as transportation, energy, water, waste management, and public safety, into a single system. Such integration is crucial to achieve the common goal of improving the quality of life for citizens. To integrate these subsystems, smart cities rely on data [97].

Data are the key ingredient and enabler for the vision and realization of smart cities. A huge volume of data represents a large amount of information generated via and about people, objects, and interactions among them in smart cities. Such data produced in different sectors within a city can contribute to generating useful information for various stakeholders for decision-making, such as policy makers, citizens, domestic governance bodies, and industrial communities [47]. By analyzing data in smart cities, we can potentially understand activities and interactions and enhance the quality of the services offered to citizens, as well as provide benefits for city management, like contributing to lowering operational expenses. For example, in Seoul, the government has been collecting data related to healthcare, transportation, and residency to make it available to citizens and scientists [222]. From this data, various smart services can be developed leveraging ICT and big data solutions [55, 81, 104, 169, 239, 300]. However, there are many challenges that need to be tackled on the way from the "raw" data to a smart service, from the data and system perspectives.

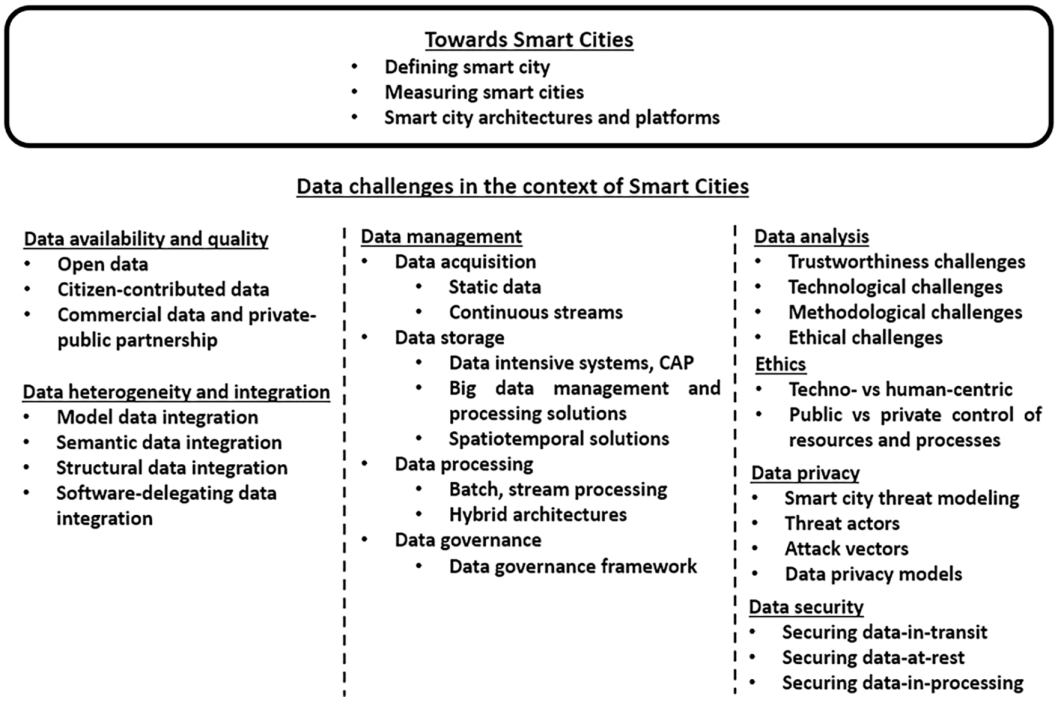

Fig. 1. Topics covered in the manuscript.

The smart city data integration and analytics platform is responsible for integrating data from various sources into a single system and performing analytics. The platform uses data integration tools and techniques to extract, transform, and load data from various sources, such as databases, sensing systems, and other monitoring devices. Once the data are integrated, it can be analyzed to provide insights into various aspects of city life. For example, data from traffic sensors can be used to optimize the traffic flow, reduce congestion, and improve public transportation. Similarly, data from energy consumption meters can be used to optimize energy usage, reduce costs, and improve energy efficiency. To manage all of this data, smart cities rely on a variety of data management systems, which are responsible for analyzing the integrated data to provide insights into various aspects of city life. The systems typically use a variety of data analytics techniques, such as **Machine Learning (ML)** and **Artificial Intelligence (AI)**, to analyze the data and identify patterns and trends. These insights can be used to optimize various aspects of city life, such as the traffic flow, energy consumption, and waste management.

This survey provides a holistic view covering data-related challenges of smart cities, see Figure 1. We start by defining and measuring smart cities and survey recent works of smart city architectures and platforms. Having this understanding, we then dive into exploring challenges and solutions for handling data in smart cities. First, we outline key issues and major research opportunities for data availability aspects and the topics and development actions toward open data, citizen-contributed data, as well as commercial data and private–public partnership. Given the diversity of smart city data, we also review challenges related to data heterogeneity and integration. Then, we examine data management issues, including data acquisition, storage, processing, and governance. After that, we explore challenges related to data analysis, ethics, data privacy, and security.

This article contributes with a careful investigation of the challenges associated with the data in smart cities. In a nutshell, our contributions are two-fold:

(1) We provide a comprehensive review of the latest development of the smart city concept. We also review existing solutions allowing for measuring smart cities as well as architectures and platforms for developing smart cities.
(2) We explore the challenges associated with the data of the smart cities, covering data availability and quality, heterogeneity, management, analysis, privacy, security, and ethical aspects, and a research agenda for addressing these challenges.

This article aims to serve as a "one-stop shop" covering data-related issues of smart cities with references to further explore particular topics of interest. The remainder of this article is structured as follows. Section 2 summarises the related work. We present a detailed discussion on defining and measuring smart city, and smart city architectures and platforms in Section 3. Afterwards, we discuss several significant research challenges, such as data availability, quality, heterogeneity, management, analysis, ethics, privacy, and security in Section 4, and conclude the article in Section 5.

2 Related Work

The urbanization and development of cities provide vibrant opportunities for academia and industry, which have inspired a number of significant related studies. For instance, Kitchin [200] provides a constructive view on the overall types of big data and smart urbanism. He also stresses the very relevant challenge of the corporatization of city governance and a technological lock-in when all the smart city-associated methods and technologies are available to large software and hardware companies, seeing this as a potential market for their products.

A number of research articles address the technological challenges for smart cities. Santana et al. [299] analyze requirements and software platforms for smart cities based on 23 projects. The authors placed these into four categories, including Cyber-Physical Systems, the **Internet of Things (IoT)**, Big Data, and Cloud Computing. Functional and non-functional requirements for smart city software platforms have been carefully investigated. Habibzadeh et al. [160] explore challenges, requirements, and solutions for sensing, communication, and security planes of smart cities. Similarly, Chamoso et al. [101] review technologies used for smart city development, as well as propose their own solution for global architecture for service management in smart city. Edge and fog computing paradigms offer promising solutions for smart cities. For instance, Perera et al. [270] explore the opportunities of fog computing for sustainable smart cities. Khan et al. [197] review edge computing applications in smart cities. The authors propose an edge computing taxonomy for edge computing-enabled smart cities, where the main blocks include security, edge analytics, edge intelligence, resources, caching, resource management, characteristics, and sustainability. Perera da Silva et al. [116] explore fog computing platforms published by the research community between 2015 and February 2021. They analyze the requirements for such systems, their architectural aspects, and how they support services provided to the users.

Technological issues of big data in smart cities are also covered in a few related works. Al Nuaimi et al. [56] review applications of big data in smart cities with the focus on opportunities and challenges for utilizing big data. Hasehem et al. [169] discuss the role of big data for sustainability and the improvement of living standards in cities with a focus on state-of-the-art technologies. Bibri and Krogstie [82] review the core enabling technologies of big data analytics and context-aware computing as ecosystems in relation to smart sustainable cities. Lim et al. [222] discuss diverse aspects of smart cities, reference models, and corresponding challenges.

A number of recent surveys address different emerging aspects of the data in smart cities. For instance, Gharaibeh et al. [150] provide an overview of data management issues, as well as discuss

privacy and security challenges. Usman et al. [339] explore the collection and analysis of multimedia data produced by smart cities. The authors focus on transportation, healthcare, and surveillance use cases and discuss various ML algorithms that could be utilized for such an analysis. Similarly, Habibzadeh et al. [159] focus on the application and data planes for smart city system design. The authors highlight cloud- and edge-based architectures to store and process the data, as well as describe various data analysis algorithms. Ma et al. [228] review the datasets being collected across 14 smart cities and the state-of-the-art in decision-making methodologies. This article further highlights both data and decision-making issues. Moustaka et al. [244] conduct a systematic review of the way urban data are produced, collected, stored, mined, and visualized in smart cities, covering the period 1996–2017. Based on this review, a set of taxonomies is proposed covering smart city data entities and methods. Some works focus more on data analysis and applications in smart cities. For instance, Chen et al. [103] explore the latest research on deep learning in smart cities. The authors study the problem from two perspectives, i.e., a technique-oriented perspective reviews deep learning models, while an application-oriented perspective studies representative application domains in smart cities. Finally, Deng et al. [126] examine how urban information can be visualized. The authors review urban visual analytics studies and specify 22 visualization types within spatial, temporal, and other property visualization categories.

Recently, more aspects related to data privacy and security have been covered. For example, Eckhoff and Wagner [132] provide a taxonomy of the application areas, enabling technologies, privacy types, attackers, and data sources for attacks in smart cities. Based on that, state-of-the-art privacy-enhancing technologies are reviewed and future research directions are discussed. Similarly, Sookhak et al. [312] look for the taxonomy of security and privacy issues of smart cities, highlight the security requirements, explore state-of-the-art security and privacy solutions, and present open research issues.

Finally, emerging concepts of digital twins, metaverses, and metacities have attracted research interests from academia. For instance, Mylonas et al. [245] explore the digital twins landscape in the context of smart cities. In addition to studying the domains where digital twins are presented, the authors also emphasize some challenges related to data from the digital twins perspective. Similarly, Bibri et al. [135] explore the emerging trends enabling data-driven smart cities for a digital and computing processes framework underlying the Metaverse as a virtual form of data-driven smart cities.

When compared to existing surveys, this review article focuses on the data aspects of smart cities, see Table 1. We provide an up-to-date state-of-the-art understanding of what a smart city is, how "smartness" can be measured, and what the data challenges are. In particular, this article focuses on data challenges related to availability and quality, data heterogeneity, data management, data analysis, privacy and security, and ethics. Therefore, this review provides a holistic view and aims to serve as a "one-stop shop" covering data-related issues of smart cities with references to further explore particular topics of interest.

3 Towards Smart Cities

3.1 Defining Smart City

The smart city concept is flexible and open, which is probably a central factor behind its popularity and global success. At the same time, it is also notoriously challenging to define [248]. The reasons are two-fold. On the one hand, scholars have mapped and categorized smart city development in different ways, depending on their background [200, 243]. On the other hand, different cities around the world have applied the agenda in their own terms, due to their specific economic, political,

Table 1. Existing Surveys About Smart City and Their Coverage of Topics Presented in This Article

Work	Focus	Architecture/ Platform	Data availability	Data heterogeneity	Data management	Data analysis	Privacy	Security	Ethics
[299]	requirements and software platforms	✓(ET, platforms, reference architecture)	○	○	○	○	○	○	×
[101]	technologies for SC development	✓(architecture, ET)	○	×	○ (storage)	○ (big data)	×	○	×
[270]	fog computing solutions for SC	✓(device management, commun. protocols)	✓(sensor data in fog computing)	○ (context, semantic annotation)	○ (general)	○ (fog computing)	○	✓(fog computing)	×
[197]	edge computing applications	✓(high-level edge-enabled SC, requirements, and open challenges)	×	○ (context-awareness)	○	○ (edge analytics and intelligence)	○ (edge computing)	✓(edge computing)	×
[56]	big data	×	○ (data sources, quality, and sharing)	×	○ (big data)	○ (big data processing platforms, algorithms)	○	○	○
[82]	big data, context-aware computing	×	✓(sensing)	×	○ (big data)	✓(big data, urban context)	○	○	×
[169]	big data	✓(big data)	×	×	○ (big data)	○	○	×	×
[222]	reference models	✓(big data)	○ (main sources of big data)	○	○	×	○	×	×
[160]	sensing, communication, and security	✓	✓(sensing, communication)	×	○	○	×	✓(crypto-, system-level)	×
[116]	fog computing platforms	✓(requirements, architecture, and services)	○	○	○ (ingestion, processing, storage, and query)	○	×	○	×
[150]	data management, security, and ET	×	×	×	✓(acquisition, coord. and management, quality and integrity, cloud vs fog, dissemination, ET)	✓(ML, DL, and real-time analytics)	○	✓	×
[339]	big multimedia data in SC	×	×	×	✓(multimedia data collection platforms)	✓(representation learning algorithms, DL, and data analytics)	×	×	×
[159]	data, applications planes of SC	✓	×	×	✓(requirements, architecture (cloud, and edge), storage and processing)	✓(data analytics, ML, DL, and visualization)	×	○	×
[132]	privacy in SC	×	○ (ET)	×	×	×	✓(types, protection, challeng., and solut.)	○	×

(Continued)

Table 1. Continued

Work	Focus	Architecture/ Platform	Data availability	Data heterogeneity	Data management	Data analysis	Privacy	Security	Ethics
[312]	security and privacy in SC	✓	×	×	×	×	✓(issues)	✓(requir., challeng., and solut.)	×
[103]	DL in SC	✓	×	✓(sensor, image/video, and text)	×	✓(DL, applcations, and challenges)	○	×	×
[228]	datasets, decision-making	×	✓	○	○	✓(modeling, decision-making)	○	○	×
[244]	data analytics, SLR	○ (SC as a data engine)	✓(urban data taxonomy)	×	×	×	✓(data analytics taxonomy)	×	×
[245]	digital twins	✓(digital twin)	○	○	○	○	○	○	○
This work	data challenges	✓(architectures and platforms)	✓(open, citizen-contributed, commercial, and private–public partnership)	✓(model, semantic, structural, and software-delegating)	✓(acquisition, storage, processing, and governance)	✓(trustworthiness, technological, methodological, and ethics)	✓	✓security (in-transit, at-rest, and in-proc.)	✓

challeng., challenges; commun. protocols, communication protocols; coord., coordination; DL, deep learning; ET, enabling technologies; in-proc., in-processing; requir., requirements; SC, smart city; SLR, systematic literature review; solut., solutions.
✓—comprehensive coverage, ○—some discussion, ×—not discussed or very light mention.

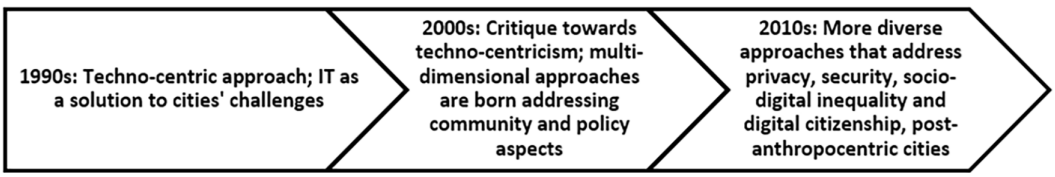

Fig. 2. Development of the smart city concept.

legal, social, and cultural arrangements [57]. Figure 2 presents a high-level evolution of smart city concept development.

In general, the smart city concept refers to optimizing city processes with ICT and, thus, creating better cities for all. The definitions of the early 2000s emphasized the streamlining of city operations and optimizing infrastructure through digital services. In addition, the idea of utilizing data in decision-making was already present in these early definitions [163]. The smart city was at first promoted especially by the private sector which saw urban ICT systems as an economic opportunity and as a way to work with the public sector [170, 309]. For example, IBM defined the agenda as follows: "Smarter Cities are urban areas that exploit operational data, such as that arising from traffic congestion, power consumption statistics, and public safety events, to optimize the operation of city services. The foundational concepts are instrumented, interconnected, and intelligent" [167].

As the popularity of the agenda increased, also a body of work presenting critique of the techno-centric approach of smart cities was born. Many articles suggested that the smart city

agenda could strengthen societal inequalities and lead to unjust cities [96, 174, 238, 292, 345]. Williamson summarized aptly [354], "urban research from geographical and sociological perspectives has sought to critique it [smart city development] in terms of being market-based, technocratic, surveillant, solutionist, militaristic, and reproductive of power asymmetries," see also, e.g., [64, 120, 233]. In other words, the critics argued that smart city development is often realized top-down, without paying attention to city inhabitants' specific needs, perspectives, and local life-words; it follows neoliberal logic; and it undermines ethical questions related to, e.g., free, open public space and privacy. Furthermore, the lack of environmental attributes was repeatedly criticized [243]; or, as Cugurullo puts it, the smart city "includes environmental ones as long as they can be monetized" [114].

Due to the increasing critical perspectives, the 2010's definition shows a shift in focus, i.e., policy and community aspects started to become more common in smart city development and related discussions. According to [131], the redefining of the term was arguably conducted to distance the concept from the technological determinism surrounding smart city. One of the central definitions offering a multi-dimensional perspective on smart cities has been formulated by Nam and Pardo [246]. Nam and Pardo have categorized key conceptual components of smart city into three areas, including technology (software and hardware infrastructure), human (creativity, diversity, and education), and institutional (governance and policy) aspects [246]. Here, the technology component emphasizes the necessity of well-functioning infrastructure and applications. Without this basis, and therefore, without engagement and cooperation between public institutions, private and educational sectors, and citizens, there is no smart cities [246]. The human factors category highlights the value of creativity, learning, and education for the city to become smart. That is, "a smart city is a humane city that has multiple opportunities to exploit its human potential and lead a creative life" [57]. Finally, the institutional dimension emphasizes the fundamental role of a supportive administrative environment (initiatives, structure, and engagement) and governance for the design and implementation of smart city [246]. Therefore, the connection of these factors implies that "a city is smart when investments in human/social capital and Information Technology (IT) infrastructure fuel sustainable growth and enhance a quality of life, through participatory governance" [246]. Several researchers have utilized and applied this multi-dimensional perspective on smart cities. For example, Yigitcanlar et al. [362, 363], have argued that by building on the drivers described by Nam and Pardo [246], i.e., focusing on technology, policy, and community, the limitations of earlier smart city model(s) could be tackled.

The current smart city literature has increasingly addressed the aspects relating to privacy, security, socio-digital inequality, and digital citizenship [170, 172, 364]. Further, there exists a strand of research that looks beyond human-centeredness and traces the possibility of a smart city model that takes into account non-human beings, i.e., animals and nature, in profound ways [226, 333, 363]. Nevertheless, there seems to be a constant tension between techno-centric visions and more holistic visions, and some authors fear that issues that have haunted smart city development already for decades will just be carried over to novel data and AI-focused urban visions [114]. Thus, social and environmental themes should always be carefully considered and plans on digitalization always embedded within broader urban policies to avoid one-sided, solutionist and fragmented approaches [114, 170].

Another view on smart cities is offered by standardization bodies, such as the International Telecommunication Union (ITU) [319] and the International Organization for Standardization (ISO) [28]. To understand the key components, the ITU conducted an analysis of smart cities and sustainable cities definitions [315]. In this analysis, 50 key words were extracted from 116 definitions found from various sources. Examples of the keywords which most occurred include quality of life, technology, people, systems, governance and administration, and economy. Therefore, common

themes and dimensions were formed from these keywords resembling the six characteristics from Giffinger et al. [151], including Quality of life and lifestyle; Infrastructure and services; ICT, communication, intelligence, and information; People, citizens, and society; Environment and sustainability; Governance, management and administration; Economy and finance; and Mobility [315]. This survey helped the ITU in identifying key essential terms for the definition of a Smart Sustainable City, defined by the ITU as "an innovative city that uses ICTs and other means to improve quality of life, efficiency of urban operation and services and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social, environmental, as well as cultural aspects." [315, 319]. ISO 37122 [49] provides another view from the perspective of standardization, underlining the role of sustainability. According to ISO, a smart city is "a city that increases the pace at which it provides social, economic and environmental sustainability outcomes and responds to challenges such as climate change, rapid population growth, and political and economic instability by fundamentally improving how it engages society, applies collaborative leadership methods, works across disciplines and city systems, and uses data information and modern technologies to deliver better services and quality of life to those in the city (residents, businesses, and visitors), now and for the foreseeable future, without unfair disadvantage of others or degradation of the natural environment" [49].

In theory, these standardization efforts could help to create a universal understanding of the smart city agenda. However, they should be used with caution because standards do not necessarily help to properly address local conditions such as differences in population, economic structures, city management, or social and cultural aspects that can affect smart city development drastically, as mentioned earlier.

3.2 Measuring Smart Cities

Given the diversity of interpretations, measuring the performance of smart cities is challenging [57]. Moreover, cities are very different in their history, culture, economy, and development goals. Therefore, to make the task approachable, quantified measures are suggested that can be tracked over time to give information about stasis and change of a particular phenomenon, i.e., indicators [201]. Kitchin et al. [201] distinguishes between single (measuring a single phenomenon) and composite (combining several measures) indicators. Also, indicators differ by their role, e.g., descriptive or contextual indicators provide key insights into phenomenon; diagnostic, performance, and target indicators serve as the means to diagnose a particular issue or assess performance; while predictive and conditional indicators are used to predict and simulate future situations and performance [201]. Here, we first briefly introduce some existing efforts toward measuring smart cities; and then highlight some data-related challenges for such indicators and indices.

A number of standardization and research efforts exist to suggest an approach for cities to monitor, analyze, and communicate the performance and progress toward achieving set goals [179, 216], see Table 2. For example, the International Communication Union has developed a number of ITU-T Recommendations on assessing different aspects of U4SSC, e.g., [318, 320, 321, 323]. For instance, ITU-T Y.4903/L.1603 [317, 323] proposes a set of **key performance indicators (KPIs)** for assessing cities in achieving smart sustainable goals. This recommendation formed the basis for the development of KPIs for smart sustainable cities by U4SSC initiative [108]. These KPIs establish criteria to evaluate ICT's contributions in making cities smart and sustainable and provide cities with the means to assess the achievements of sustainable development goals. U4SSC indicators form part of a holistic view of a city's performance in economy, environment, society, and culture dimensions. Over 100 cities worldwide already implement these KPIs, like Dubai, Valencia, and Moscow [314]. The ISO also puts effort into monitoring and developing sustainable and smart cities. For instance, a number of indicators for sustainable cities and communities were suggested in ISO 37120 [48],

Table 2. Some Standardization and Research Efforts Toward Measuring Smart Cities

Activity	Scope
CITYkeys H2020 EU project indicators [87]	Proposes indicators for assessing smart city projects and the corresponding city-level indicators. The indicators are categorized as: people, planet, prosperity, governance, and propagation themes, which are further split into subthemes. Altogether, 99 project and 76 city indicators have been presented.
ETSI, Key Performance Indicators for Sustainable Digital Multiservice Cities, ETSI TS 103 463 V 1.1.1 (2017-07) [181]	Proposes indicators based on the CITYkeys project [87]. Here, topics include: people, planet, prosperity, and governance.
ITU, Overview of key performance indicators in U4SSC, Recommendation ITU-T Y.4900/L.1600 [320]	Gives a general guidance to cities and suggests key performance indicators toward U4SSC, categorized as: ICT, environmental sustainability, productivity, quality of life, equity and social inclusion, and physical infrastructure.
ITU, Key performance indicators related to the use of ICT in U4SSC, Recommendation ITU-T Y.4901/L.1601 [318]	Focuses particularly on KPIs related to the use of ICT in U4SSC. Categorized into: ICT, environmental sustainability, productivity, quality of life, equity and social inclusion, and physical infrastructure.
ITU, Key performance indicators related to the sustainability impacts of ICT in U4SSC, Recommendation ITU-T Y.4902/L.1602 [321]	Focuses particularly on KPIs related to ICT impacts for U4SSC. Categorized into environmental sustainability, productivity, quality of life, equity and social inclusion, and physical infrastructure.
ITU, Recommendation ITU-T Y.4903/L.1603 [317] and its update Recommendation ITU-T Y.4903 [323]	Proposes KPIs to allow cities to monitor and assess the efforts in achieving sustainable development goals, becoming smarter and more sustainable cities. Indicators are categorized into: economy, environment, society, and culture groups.
ITU, U4SSC maturity model, Recommendation ITU-T Y.4904 [322]	Proposes a maturity model for sustainable smart cities, as well as methods to assess and plan future development strategies. Here, the focus is particularly on assessing the achievement of sustainable development goals toward ICT development of the cities. The proposed model has five layers and three dimensions: economic, environmental, and social. The KPIs are recommended for assessing maturity levels as well, like published in ITU-T Y.4901 [318], ITU-T Y.4902 [321], and ITU-T Y.4903 [323].
ISO, Sustainable cities and communities—Indicators for city services and quality of life, ISO 37120:2018 [48]	Proposes indicators to assess the performance of city services and quality of life. The indicators are grouped under economy, education, energy, environment and climate change, finance, governance, health, housing, population and social conditions, recreation, safety, solid waste, sport and culture, telecommunication, transportation, urban/local agriculture and food security, urban planning, wastewater, and water.
ISO, Sustainable cities and communities—Indicators for smart cities, ISO 37122:2019 [49]	Proposes indicators to assist cities in assessing the performance of city services and quality of life. Indicators are grouped under the same categories as in ISO 37120:2018 [48].

ETSI, European Telecommunications Standards Institute; EU, Eurpoean Union; TS, technical specification.

which was further complemented with indicators for smart cities in ISO 37122 [49]. There, indicators are broken down into sectors, such as the economy, education, energy, the environment, and climate change. Also, indicators are complemented with meta-information about data sources, interpretation, and calculation methodologies. The World Council on City Data are involved in ISO indicators development and provides city certifications based on the implemented ISO 37120 indicators [44].

The CITYkeys EU Horizon 2020 project focused on the development and validation of key performance indicators and data collection procedures for monitoring and comparison of smart city solutions across European cities [274]. The CITYkeys indicators are based on an inventory of 43 existing indicator frameworks and categorized by people, planet, prosperity, governance, and

Table 3. Examples of Open Data Related KPIs

Indicator name	Assessment solution	Measurement mechanism	Description
Increase in online government services	CITYkeys project indicator [87]	Likert scale	Indicator analyzes the improvement in providing online government services, including open data platforms.
Quality of open data	CITYkeys project [87]	Likert scale	Indicator assesses the ease of use of datasets produced by the project and whether they are kept up-to-date.
Accessibility of open datasets	CITYkeys project [87], ETSI [181]	Average stars across all datasets according to the 5 star deployment scheme for Open Data defined by Tim Berners Lee (5stardata.info).	Indicator evaluates ease of use and the openness of city data.
Open datasets	CITYkeys project [87]	The number of open government datasets per 100,000 inhabitants.	Measures the number of open government datasets.
Open Data	ETSI [181]	Number of open government datasets per 100,000 inhabitants.	Measures the number of open government datasets.
Open data	ITU-T Y.4903 [323]	Total number of open datasets published divided by total number of datasets multiplied by 100.	Percentage and number of published inventoried open datasets.
Percentage of service contracts providing city services which contain an open data policy	ISO 37122:2019 [49]	Total number of service contracts providing city services which contain an open data policy divided by the total number of service contracts in the city, multiplied by 100.	The percentage of service contracts providing city services that have an open data policy.
Annual number of online visits to the municipal open data portal per 100,000 population	ISO 37122:2019 [49]	Total number of municipal open data portal visits divided by 1/100,000 of the city's total population.	Annual number of online visits to the municipal open data portal per 100,000 population.

propagation themes [87]. The themes are further broken down into sub-themes where 99 project (to assess single projects) and 76 city (to monitor evolution of the city) indicators have been selected and explained in detail with the mention of expected data sources [87]. What makes the CITYkeys project indicators different is that they are impact-oriented. They were also used by the **European Telecommunications Standards Institute (ETSI)** in their technical specification "Key Performance Indicators for Sustainable Digital Multiservice Cities" [181]. Table 3 presents some examples of indicators related to open data and their interpretation within different assessment suggestions.

Such assessment solutions also allow the creation of indices to enable the comparison and monitoring of city development progress. Indexes can be considered as "quantitative aggregation of many indicators and can provide a simplified, coherent, multi-dimensional view of a system" [236]. So, these are composite indicators, combining several indicators through weighting or statistics to create a new derived measure [201]. For instance, U4SSC KPIs form the basis for the U4SSC Smart Sustainable City Index that facilitates a comparative ranking of cities.

Although useful, the creation and usage of indicators must be done with care, since their validity is inbuilt in the process they are created. For instance, the indicators themselves describe the characteristics of the system state based on observed or estimated data [236]. This means that the diversity of data sources and measured quality challenges are inbuilt by definition, often making direct comparisons unfeasible. Moreover, Kitchin et al. [201] emphasize also that data do not exist

independently from the ideas, interests, technologies, practices, and systems involved. Therefore, they should be used and interpreted with caution. All these imply that assessment frameworks should provide a clear description, rationale, interpretation, benchmarking, and methodology for indicator calculation, as well as potential sources of possible data to use and links to other normative documents [108]. It is thus essential that the framework user is equipped with all the information regarding the data. Further, indicators can show that a problem exists, but they do not show its cause or tell what to do [201]. Therefore, they could be useful if monitored continuously, to see the progress if certain measures are taken. This also raises the question of whether a city index to rank cities is needed [86]. Given the fact that cities are very different from each other and have diverse histories, economics, and development goals, their ranking can be misleading and provide weak support for cities themselves in their development. Moreover, "indicators and measurements should not become a goal in themselves but support the fulfillment of individual cities' needs" [179]. From this perspective, indicators supporting continuous monitoring of important phenomena in the city could be valued more. Additionally, indicator visualization is important, since this may affect perception and interpretation [201, 250].

Indexes should also be used with caution. For instance, indices usually have a certain focus, that determines which indicators are included in it [86]. It is recommended to develop a solid theoretical framework to serve as the basis for the selection and combination of indicators into meaningful composite indicator [250]. Therefore, the developers of the index should understand and communicate the purpose and limitations of the index, as well as how different indicators relate, so that index interpretation is solid [236].

In addition, indexes also rely on a number of data processing techniques, such as aggregation, normalization, and weighting [250, 308]. Proper theoretical grounds should be followed, otherwise "'incompatible' or 'naive' choices (i.e., without knowing the actual consequences) in the steps of weighting and aggregation may result in a 'meaningless' synthetic measure" [155]. Moreover, it is recommended to test the aggregate measures for their robustness as a whole, to test how sensitive the index is to changes in the steps followed to construct it. In this regard, traditional techniques include uncertainty and sensitivity analysis [155, 250]. These imply that data, overall methodology, predetermined boundaries of the system, and comparability of results across the systems should be transparent and clearly communicated so that one is able to assess the performance and suitability of an index for a particular task [236]. Finally, aggregate indices may hide some information, e.g., it could be challenging to identify if few indicators have extreme values when the index aggregates hundreds of these into one number [236]. In such situations, it could be better to provide the indicators as frameworks and use visual tools to present these, for instance.

Indexes also require careful governance, because over time the data behind the indicators can change, therefore direct comparisons with previous versions may become unfeasible. The ability to compare various indicators and assessment frameworks provides means to ensure that the proper one is selected. Huovila et al. [179] provide a comparative analysis of standardized indicators for U4SSC, where seven sets of city indicators published by international standardization bodies are inspected in terms of their conceptual urban focus, city sectors, and types of indicators.

Acknowledging the limitations and challenges, indicators are still useful and provide the means to track the progress of certain phenomenon [36, 110]. The key message here is to enable as transparent and documented process as possible, ensuring that users of the indicators have a proper understanding and are able to make an informed judgment if the indicator is suitable for the task at hand.

3.3 Smart City Architectures and Platforms

Smart cities are very complex structures involving various stakeholders, technologies, and physical constraints; therefore, it is difficult to provide a unified reference architecture and a platform,

Table 4. Summary of Requirements for Smart City Architecture and Platform From Related Work

Functional requirements	Non-functional requirements
Summary from [23, 81, 159, 183, 222, 231, 268 300, 316, 335]	
—Handling big data characteristics, namely volume, velocity, variety, veracity, and value. —Definition of a city model, data models, and APIs —Data management —Data storage management —Data processing and analysis —External data access —Applications runtime management —WSN management —Service management, SLA —Software engineering tools, APIs —IoT device/resource discovery and management —IoT data marketplace —License management —Incorporation of feedback and monitoring	—Interoperability —Decoupled and distributed components —Openness —Legacy compatibility and heterogeneous landscape —Resilience to failure and robustness —Performance —Scalability —Security —Privacy —Context awareness —Adaptation —Extensibility —Configurability

SLA, service-level agreement; WSN, wireless sensor network.

since the development could be guided by local requirements [300]. In this section, we cover some existing efforts toward smart city architectures and platforms and summarize them into general architecture from the smart city data point of view.

The ITU defines architecture in general as "a definition of the structure, relationships, views, assumptions, and rationale of a system" [316]. There are many smart city architectures and their implementations presented by the research community, varying in their goals and details. Generally, smart city reference architectures should be technology-neutral and provide a clear set of capabilities and stages to be implemented to provide smart city services [142]. Moreover, such architectures aim to fulfill a certain set of requirements of the domain. Table 4 summarizes requirements for smart city architectures and platforms found in related work. As can be seen, in general, such requirements cover data and system management functionality, as well as non-functional requirements related to privacy, security, and system lifecycle management.

A number of architectural proposals exist with varying levels of detail. Some researchers provide quite a general perspective. For instance, Zygiaris [381] suggests seven layers, going from the layer covering essentials of the city (districts, inbuilt infrastructure, and so forth), to level aiming and promoting green and sustainable actions (like green transport practices, and planning), to technology and application covering layers (interconnection, instrumentation, open integration, and application layers), and, finally, to the innovation layer, focusing on the innovation ecosystem, which is vital for the prosperity of the cities and their inhabitants. Zheng et al. [371, 372] summarize the urban computing system framework, which is comprised of four general layers: urban sensing and data acquisition, urban data management, urban data analytics, and service provision. In contrast to other proposals, Zheng et al. [371] are more interested in methodological aspects, like

processing geo-spatial data at each layer (e.g., trajectory compression and map-matching in the urban data management layer).

Others focus more on the system development angle. For instance, Habibzadeh et al. [159] abstract smart city architecture into five generic planes: an application plane, sensing plane, communication plane, data plane, and security plane. There, each plane comprises a number of technologies, methods, and challenges. Santana et al. [299] provide reference architecture for the development of software platforms for smart cities based on analysis of 23 related projects. Compared to others, their architecture is more technology driven and is based on the cloud and networking layer, with IoT and Service middleware, user management, and social network gateway on top of that. The Big Data management component is responsible for all data aspects. In addition, the need for a toolkit and security support are presented in the architecture. The authors also emphasize that all components of the platform must support scalability, security, privacy, and interoperability. Santos et al. [301] focus on a sensing platform for smart cities. Their approach is to follow the data flow: sensing, data collection, data storage, processing, sharing, and hosting urban services. They integrate sensor data from mobile crowdsensing, environmental, and public transport vehicle sensing for analysis, data sharing, and smart city applications development. There, the importance of a unified spatio-temporal data model and the use of standard IoT data access methods are emphasized. Villanueva et al. [346] propose the Civitas platform to be seen as the core of a smart city IT infrastructure able to orchestrate different entities (like citizens, public institutions) connected to it via Civitas plugs. Middleware relies on core nodes which are servers hosting a variety of services. The authors emphasize the integration of intelligence, like common sense reasoning. When compared to others, this proposal is more broker-like. Bibri [81] provides an analytical framework for data-centric IoT applications for U4SSC. Their proposal provides a pipeline focused on IoT, Big Data, Cloud, and Fog programming paradigms. Its main components include urban systems and domains that should function and be managed by IoT and its underlying big data analytics; the urban big data sources, storage facilities, and data categories component is responsible for data collection, storage, and management; Cloud computing or fog/edge computing and Hadoop MapReduce architecture infrastructure for big data processing and management for knowledge discovery/data mining; Big data applications covers smart applications for diverse urban domains [81]. The CUTLER EU project proposes a data hub conceptual architecture to support data management and analysis for decision-making in municipalities [335]. In comparison to other proposals, they provide quite a general data-centric conceptual solution, which is then illustrated with concrete implementations for five pilot cases. Their main blocks in the architecture are: data collection, representing data acquisition functionality (like data sources, data crawlers, and data pre-processing); data integration platform supporting data ingestion, data storage, and access APIs to other components that will further manage and/or process the data; data analytics to support business logic of the smart city services; data governance to manage the data and data lifecycle; business model DevOps to bridge the gap between the big data technology and the business model of policy developments; and services and visualization responsible for smart city services and data visualization [335]. Similarly, Pereira et al. [268] suggest a platform for integrating heterogeneous data and aiding the development of smart city applications. In comparison to other proposals, their solution emphasizes a semantic-based data model. For example, in their proposal, information is grouped into layers that represent geographic or some particular domain information, like school or public safety. The information from different layers could be linked together to retrieve new information, e.g., information about safety close to schools. Architecture-wise, it is a distributed system consisting of SGeol middleware and middleware infrastructure, that includes components for managing users and data access security policies; managing data, its messaging, integration, and context; discovery of physical devices and their integration to the platform; real-time and batch analysis. The solution

also provides **representational state transfer (REST)** APIs for external data access and SGeoL Dashboard service offering edit, query, and visualization capabilities.

The SynchroniCity EU project (that also included partners with leading roles in standardization bodies) aimed to establish a reference architecture for the IoT-enabled city marketplace, ensuring interoperability and developing interfaces and data models for different verticals [231]. To achieve this, the SynchroniCity project analyzed available models and approaches for smart cities and summarized them with an architecture framework collecting the most common capabilities and technologies [231]. Their reference architecture consists of different logical modules, including Context Data Management to manage the context information coming from various data sources; an IoT Management module responsible for interaction with the devices using different standards or protocols to make them compatible with the framework; a Data Storage Management module responsible for data storage and access; an IoT Data Marketplace to facilitate business interactions between data suppliers and consumers by enabling digital data exchange; a Security, Privacy, and Governance module to handle security aspects related to data, IoT infrastructure and the platform services; Monitoring and Platform management services module which guards the platform configuration management and service activities monitoring; Southbound interfaces to connect the architecture to various data sources and IoT devices; and lastly, Northbound interfaces to provide platform functionalities to be used by the final smart city end-user applications [231].

Standardization bodies are also interested in providing architectural solutions enabling smart cities and they have similar views to the SynchroniCity project. For instance, ITU provides different angles on smart sustainable city reference architecture. Their ICT architecture from the communication view emphasizing a physical perspective relies on the top of the city physical infrastructure. This architecture consists of sensing, network, data and support, application, and operation, administration, maintenance and provisioning, and security layers. Architecture also demonstrates communication and exchange of information between layers [316]. ETSI puts context management and interoperability at the core of their platform [183]. They suggest a smart city platform that is based on the **Context Information Management API (NGSI-LD)** ecosystem [182, 184]. The main logical functions of their framework are as follows: Data ingestion and integration to collect data from different systems; an NGSI-LD Context Broker applying NGSI-LD API [184] for data interoperability; Semantics for construction and the use of semantic data and technologies; Analytics and AI to support analysis/prediction services for smart cities; Monitoring and management responsible for system operation monitoring and management; and Security and Access Control is responsible for authentication for smart city platform users and applications, access control policy management, and access control token management functions [183]. Their architecture also considers data spaces, through a Data space connector smart cities can connect to other data spaces and share data across other relevant systems [183]. Similarly, with a focus on context management, FIWARE suggests a reference architecture for smart cities. Their architecture is technology-oriented, where an Orion Context Broker is its core component. FIWARE provides data models, interfaces, and ready-made components for e.g., IoT, processing, analysis, and visualization of data [23]. For instance, the FIWARE platform was used to provide the main components for the underlying middleware infrastructure for SGeol middleware [268].

Figure 3 summarizes the essential functional blocks required for a smart city data platform. Logically, we divide the architecture into data sources, platform, and applications. The data sources represent possible data that can be used for the development of smart services. The platform incorporates a traditional data management pipeline. The important aspect here is interoperability and data models, as pointed out by some related work [23, 231]. Traditional blocks also include data storage and analysis. The data governance functional block ensures the overall usability of data assets in the platform. Data security and privacy are in the backbone of the platform. Finally,

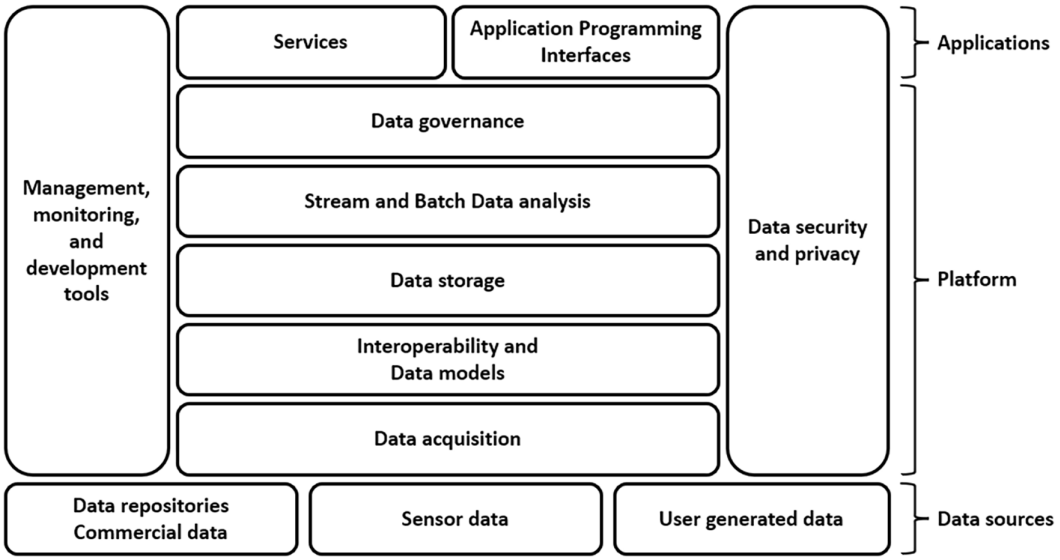

Fig. 3. Smart city data system architecture, a summary from related work.

management and development tools are needed to ensure that the platform is operational. On top of that, the development of APIs would facilitate accessing data/analysis results or performing certain actions. The services block represents numerous services that could be developed on top, like smart transportation services.

Concrete implementations of such functional blocks could vary greatly, from more centralized cloud-based solutions to more distributed ones, like edge-based solutions [197, 270]. Therefore, methods and tools could be selected accordingly. For example, architectures and platforms are proposed to support the development, deployment, and management of IoT systems across a number of devices with varying resources, e.g., Osmotic Computing Platform [347]. An in-depth review of methods and technologies for concrete implementations of smart city data architectures, as well as deployment and management frameworks, is out of the scope of this article. For such studies, refer to [159, 270, 299]. Instead, we focus on data challenges from a more conceptual standpoint, leaving their concrete implementation and selection of methods and tools to the developers.

4 Data Challenges in the Context of Smart Cities

The development of IoT and communication technologies has opened up numerous opportunities to assess a variety of phenomena in cities, like traffic, pollution, and economic wealth. City data are diverse in nature and has a variety of formats, availability, volume, spatio-temporal dependencies, and sensitivity concerns, to name a few. All this data should be processed and analyzed to derive comprehensive insights. Therefore, solutions are needed to work with such diverse data in a robust, efficient, secure, and ethical manner. This section reviews the main issues and approaches developed in the smart cities context in (1) data availability and quality, (2) data heterogeneity and integration, (3) data management, (4) data analysis, (5) ethics, (6) data privacy, and (7) data security.

4.1 Data Availability and Quality

Different taxonomies have been applied to classify urban data. For instance, Zheng et al. [372] suggest a division of urban data by the nature of the phenomena they present, like geographical, traffic, mobile phone signals, commuting, environment monitoring, social network, economy,

energy, and health care data. Another suggested taxonomy is based on data structures (point- and network-based types of data) and spatio-temporal properties (spatio-temporal static, spatial static but temporal dynamic, and spatio-temporal dynamic) [371]. Additionally, available urban data can be divided into five pools, including firewall (within the legacy systems of public agencies), open data, social, sensors/IoT, and commercial data [141]. Finally, urban data has also been divided based on including personal information, like non-personal data, aggregate data, de-identified data, and personal information [214]. In this subsection, we will highlight the urban data availability aspect, categorizing our exploration into open data, citizen-contributed data, and commercial data solutions. Also, we will discuss corresponding data quality considerations.

Open Data. Data are the key enabler for the vision and realization of smart cities. According to a European strategy for data, Big Data are considered as one of the key enablers to maximize the growth potential of the European digital economy and society [109]. Therefore, a large effort has been made to promote data suppliers and owners, even municipalities and governments to open their data for both research and business. To gain the benefits, an adaptation of municipal vision and governance strategies could be required to coordinate, enable, and support various forms of data-sharing initiatives [202]. Open data are the data that anyone can access, use, and share; it is available in machine-readable format, as well as licensed to permit data use in any way [35]. Governments and municipalities play a crucial role in the management of cities' data assets so that data-driven tools can be used to address challenges that cities face [73]. Therefore, there is also a strong recent trend to release much of public agencies' data as open data [141]. This is known as Open Government Data, which is defined as information collected, produced, or paid for by public bodies and licensed for free re-use for any purpose [35]. A number of open-source and commercial data portal platforms exist, providing the ability to publish data, enabling data access and visualizations like CKAN,[1] DKAN,[2] Socrata,[3] Opendatasoft,[4] PublishMyData.[5] Their availability, as well as the strong demand to share urban data, has resulted in a number of urban data platforms, containing both open and restricted in-use data. Barns [73] classifies these into data repositories—open data portals with the main goal to provide data sharing capabilities; data showcases that aim to visualize data, but the data itself is not always available or machine-readable; city scores—visualization of city performance in regard to a certain set of indicators; and data marketplaces enabling data access and reuse with performance monitoring. Examples of data repositories include, the New York City open data portal [119], which enables data access within a number of categories. Among the full information about the dataset, it is also possible to see the data snapshots and visualize the data in external services. Another example of data repositories is the Moscow City Government open data portal,[6] providing access to data classified into thematic topics, like healthcare, education, and culture. Datasets are equipped with basic information, including, among others, dates, formats, links to the source, and contact information of the persons responsible. Well-known city dashboards include the Dublin Dashboard,[7] which provides rich visualization opportunities as well as possibilities to get the data available. The London Datastore [32] also provides rich opportunities to visually explore the data, as well as gain access to it. However, when compared to other city dashboards, the London Datastore provides data-driven analytics based on their alignment to strategic planning and governance challenges for City Hall [73]. Table 5 provides

[1] https://ckan.org/
[2] https://getdkan.org/
[3] https://www.tylertech.com/products/socrata
[4] https://www.opendatasoft.com/
[5] http://www.swirrl.com/
[6] https://data.mos.ru/
[7] https://www.dublindashboard.ie/pages/index

Table 5. Summary of Selected Datasets

Dataset	Summary
City Pulse Aarhus City [58]	The dataset provides information related to traffic observations, weather situations, pollution, and cultural events from the city of Aarhus, Denmark. The dataset has been used, e.g., to forecast traffic situations, study privacy concerns, measure air pollution, and development of transfer learning algorithms [62, 175, 303].
Amsterdam [20]	The dataset measures traffic, accidents, crime statistics, economic activity, and pollution. It has been used, e.g., to estimate the effect of parking prices, to forecast traffic flows, for fast charging planning for vehicles, and for contextualization for sustainable development [77, 171, 188, 257].
Chicago Datasets [18]	Datasets include traffic congestion estimates, traffic counts, accident and emergency dispatches, energy usage, air and water pollution, and data-related to economic activity. The dataset has been used for, e.g., forecasting daily crime, traffic prediction, studying residential energy efficiency, and crime analysis surveys [54, 302, 373].
London [32]	The Greater London Authority provides a wide range of data-related to traffic counts, street crime, energy usage, data-related to for borough profiles, topsoil chemical data, wealth gap, and birth trends [22, 228]. The dataset has been used, e.g., to analyze crime patterns, forecast energy usage, and borough-level COVID-19 forecasting [128, 255, 283].
New York [119]	The portal provides data-related to vehicle collisions, crime data, energy and water data, air quality, water quality complaints, school districts, enrollment statistics, and others [119]. The data has been used in different studies to assess the needs after Hurricane Sandy, electricity estimation, crime prevention, study air pollution trends, and predicting burglaries [137, 193, 310, 374].
AirNow [1]	The AirNow platform provides air quality data about local areas in the United States, Canada, and Mexico from more than 500 locations [1]. The data has been used to study and forecast wildfire pollution, bias correction in air quality forecasting models, ozone forecasting, and the effect of ozone on children's health [60, 164, 232, 293].
Tokyo Open data [41]	The Tokyo Metropolitan Government has developed an open data portal to provide insights into different city segments. The platform provides case studies, data-related to bus stations, disaster prevention maps, and data-related to the environment (e.g., air pollution, landfill, sewerage, and so forth). The data portal has been used, e.g., to organize a hackathon to address administrative issues, analyze social trends related to COVID-19, investigate the crime harm index, and study issues related to the lack of educational data [173, 254, 332, 337].

a brief summary of selected available datasets. For deeper insights, an interested reader could refer to Ma et al. [228], who survey available city datasets.

There are a number of initiatives in the **European Union (EU)** advancing data sharing. For instance, the open data portal[8] provides access to data published by EU institutions and bodies.

[8]https://data.europa.eu/euodp/en/home

In addition, the portal provides opportunities for data visualizations and work with linked data. Furthermore, the European Data Portal harvests the metadata of public sector information available on public data portals across European countries.[9] Other data sharing activities include INSPIRE Geoportal[10] that collects data provided by EU Member States and several European Free Trade Association countries under the **Infrastructure for Spatial Information in Europe (INSPIRE)** Directive that focuses on creating an infrastructure for sharing environmental spatial information. Yet another known initiative is Copernicus, the Earth observation programme coordinated and managed by the European Commission and is implemented with the Member States, the European Space Agency, the European Organization for the Exploitation of Meteorological Satellites, the European Center for Medium-Range Weather Forecasts, EU Agencies and Mercator Océan.[11] Copernicus provides a number of services categorized under atmosphere, marine, land, climate change, security, and emergency themes, as well as access to satellites and *in situ* sensor data.

While acknowledging the power of such dashboards and portals, it is important to note that they require considerable effort to remain useful and provide utility for communities, municipalities/governments, and businesses. First, their purpose and interpretation should be as clear as possible, since the data itself, as well as data processing and analysis steps are known to be technology and methodology dependent, limited in time and location, and could be biased in interpretation [200, 201]. Second, such data platforms require active maintenance and support to ensure that they contain up-to-date information of the required quality. Support is also needed for both data providers and data consumers. For instance, proper effort is required to share the data. The data provider must ensure the content quality (completeness, cleanness, and accuracy), timeliness and consistency support, data representation model (use of standardized solutions, proper formats, and linked data), supply of proper metadata, as well as, addressing the legal aspects, i.e., to provide a license to use the data [35]. After the data are published, it should be properly maintained, i.e., checking data access and assessing and updating data itself and its metadata, as the data lineage and metadata allow users to assess the trustworthiness and data quality [201].

Legal issues regarding publishing and use of the data require careful treatment. For example, data ownership, legal grounds, and terms of use are often unclear for particular data sources within data repositories. Many data repositories have statements and references to legal documents in their terms and conditions on what kind of data are stored and how to use it, e.g., the Moscow City Government open data portal. However, e.g., including licence information in the data source description itself provides better transparency and eliminates confusion, check the London Datastore for an example.

Citizen-Contributed Data. The premise of citizen-contributed data are to facilitate and collect input for decision-making at large. Different approaches exist to harnessing citizens' data [211], including:

— *crowd markets*: to enable the aggregation of online individuals as collaborative input;
— *social media mining*: to retrieve publicly expressed opinions and content;
— *urban and in situ sensing platforms*: to unobtrusively collect data from citizens' daily dwellings.

Crowd Markets. Amazon's Mechanical Turk [2] and Figure-Eight [21] (previously Crowdflower) are today's largest platforms for aggregating online individuals' time to complete tasks that are computationally intensive but relatively trivial to a human. These platforms are purposefully generic, and a variety of tasks can be created. These tasks range from answering surveys, writing reviews,

[9] https://www.europeandataportal.eu/
[10] https://inspire-geoportal.ec.europa.eu/
[11] https://www.copernicus.eu/en

annotating images, transcribing audio, and others, i.e., tasks that are challenging to automate due to a high risk of error. The main challenge of crowd markets is to sustain the crowd size and quality. The literature shows that higher-paid tasks can attract workers at a higher rate. Emphasis on the importance of the work has a statistically significant and consistent positive effect on the quality of the work [290]. A practical example of leveraging crowd markets is Zensors [45], which enables sensing from any visual observable property. Zensors streams images where the crowd processes and labels them according to a well-defined set of instructions, enabling near-instant counting and another high-level sensing. Once sufficient human-based input is available, ML is applied to fully automate the process once the accuracy of the algorithms is high (>90%). This approach is also used by the Google Crowdsource initiative [19], where gamification and recognition given as badges are used to sustain and train ML classification algorithms.

Social Media Mining. Online social media mining on a large-scale allows us to consider users' posting of opinions and content in online social media to gain insights into unfolding events [290]. The widespread availability of smartphones and high-speed Internet has enabled a range of systems that collect a variety of different types of user contributions. For example, it is now possible to collect videos and photos in the field, e.g., YouTube, Instagram, Twitter, and Facebook. These platforms allow user-driven tagging with relevant keywords. The primary use of this media is for the platform, but researchers have found such user-generated content as sensor data, originating from end-users. Providing a system that allows users to easily contextualize and tag high-level data results in a valuable repository of knowledge. For example, Wheelmap[12] allows users to tag and search for wheelchair accessible places using a smartphone and a browser. Others share where they are [343] or whether a place is recommended [208], or reported the destruction aftermath of an earthquake [348]. Researchers keep exploring ways to use devices' sensors usage, as Citizen Science [266]. Citizen Science can be interpreted as individuals becoming active participants and stakeholders of data. Large-scale efforts, such as Wikipedia and OpenStreet Maps, allow users to publicly augment and annotate online information as text or geo-fenced markers. This wealth of everyday information about and around us creates numerous possibilities for new applications and research in general. Social media-enabled applications are primarily driven by smartphones for *in situ* context and are often deployed on application stores for ease of installation and updating the platform.

Urban and In-Situ Sensing Platforms. Urban and *in situ* systems pervasively collect data from citizens without the need to set up or install an app on someone's smartphone. These platforms often deploy sensors throughout a city. These can be invisible to citizens, e.g., underground traffic sensors, weather monitoring stations on top of a building, or they can be an integral part of the city, e.g., interactive public displays. A number of studies have investigated the use of public interactive displays for the purpose of data collection [63, 89, 176]. Opinionizer [89] is designed and placed in social gatherings (parties) to encourage socialization and interaction. Participants would add comments to a publicly visible and shared display. Due to the fear of "social embarrassment," the authors suggest public interactions to be purposeful.

The environment, both on and around the display, also affects its use and the data collected. The environment produces strong physical and social affordances, and such devices or solutions need to reveal their purpose regarding the social activity under study rapidly and to be able to seamlessly and comfortably encourage citizens to switch from being onlookers to becoming participants. TextTales [63] explored providing story authorship and civic discourse by installing a large, city-scale, and interactive public installation that would show a grid of text. A discussion on a certain photograph would start with text messages sent by citizens, displayed in a stream of comments.

[12] https://www.wheelmap.org

Beyond public display, citizens can also be involved in larger efforts to affect society at large. One such project is vTaiwan,[13] which is an online-offline consultation process that brings together government ministries, elected representatives, scholars, experts, business leaders, civil society organizations, and citizens. The platform allows lawmakers to implement decisions with a greater degree of legitimacy. It combines a website, meetings, hackathons, and consultation processes. For example, vTaiwan was crucial in the debate on Uber operations in Taiwan.[14] In a similar approach, Decidim[15] is a digital platform for citizen participation, helping citizens, organizations, and public institutions to self-organize democratically at scale. It provides a political network, citizen-driven initiatives and consultations and raises participatory budgets, thus allowing a democratic and flexible system where everyone can voice their opinion.

Overall, citizen-contributed data are a valuable source of information, and in some cases, it is the only way to understand the phenomenon of interest. However, such data collection initiatives and subsequent data analyzes should be planned well and performed with care. For instance, if citizens are asked to perform a measurement, they should be instructed on how to do it to get reliable value [90]. Some measurements may also require a calibration of the device [272]. In addition, one should have a strategy to deal with data gaps due to behavioral patterns of people taking measurements [284]. As in each study, one should ensure that a sample of users, contributing the data to the system, represents the population as fully as possible, and that no bias is introduced into the data collection strategy. Finally, privacy issues from such data collection initiatives should be checked and treated appropriately.

Commercial Data and Private–Public Partnership. A number of commercial organizations deploy infrastructures and utilize available urban data to provide and improve their services. Sharing these data with municipalities has been a subject of debate for a long time [46]. However, challenges with data have enabled various forms of commercial involvement, such as data markets and hubs. Such organizations facilitate connections between data providers and data consumers, especially if the data cannot be openly shared. One example of such a solution is the Platform of Trust[16] in Finland, that enables data movement between systems and organizations, taking care of trustworthiness and data harmonization issues. They also involve the community so that interested people can participate in creating harmonization models that are then published as open-source code.

Additionally, possibilities have been explored for data exchange between public and private organizations, e.g., the **City Data Exchange (CDE)** project created a marketplace for data exchange between public and private organizations [251]. This project was a collaborative effort of the Municipality of Copenhagen, the Capital Region of Denmark, and Hitachi. The CDE service provided collaboration between different partners on supply and demand of data and a platform for selling and purchasing the data for both public and private organizations. Based on the project, a number of challenges were identified, e.g., immature market as even though some companies buy data for their services, generally many are not yet ready to include data sharing in their core business or strategy; lack of use cases could affect the reluctance to invest resources in selling/buying the data; fragmented landscape; reluctance to share data on an open data portal, e.g., due to ethics or competitors' advantage reasons; lack of skills and competences to work with the data [251].

The development of such joint efforts requires trustworthy data stewardship. That is, "trustworthiness is the virtue of reliably meeting one's commitments, while trust is the belief of another that the trustee is trustworthy" [253]. Several models have been suggested to collaborate in data

[13] https://info.vtaiwan.tw/
[14] https://vtaiwan.tw/topic/uberx
[15] https://www.decidim.org
[16] https://www.platformoftrust.net/

use and share [185]. For example, data collaboratives[17] represent a form of partnership where a number of parties, like governments, companies, and others, collaborate to exchange and integrate data to help to solve societal problems or create public value [204]. Therefore, through such cross-sector and public–private collaboration initiatives it is possible to achieve much wider goals that may be difficult to perform by the parties by themselves only. One noteworthy example of data collaboratives in smart city context is 9,292[18] which is public–private collaborative, gathering and sharing public transportation data in the Netherlands. Obviously, data collaboratives possess all the challenges that data integration initiatives have, since the data comes from diverse providers, in different formats and has varying structures. However, as Klievink et al. [204] emphasize, data collaboratives are a collaboration and innovation phenomenon rather than a data phenomenon. Therefore, organizational, incentivization, and governance challenges should be considered as well. From this perspective, a number of additional challenges arise regarding vulnerabilities in opening the data, its possible misuse, and overall trust within the partnership. Coordination problems also include matching potential data providers and data users, maintaining data control and its unforeseen uses when shared, matching a problem with the data attributes, ensuring the shared data are useful and usable by the user, and aligning the incentives of providers to share proprietary data with the goals of the users [330]. Moreover, data collaboratives are not isolated constructs, therefore partners' incentives, goals and collaboration overall depend on the context, like institutional and governance frameworks, government interests, transparency/inclusiveness culture, and the means by which collaboration is legitimized [204]. Therefore, to achieve a successful collaborative, it could be helpful to organize the overall collaboration process and context in such a way that perceived vulnerabilities are dealt with [204].

Another initiative is data trust. The interest in data trusts started in 2017 where this model was proposed as a "set of relationships underpinned by a repeatable framework, compliant with parties' obligations, to share data in a fair, safe and equitable way" [162]. The Open Data Institute defines data trust as "a legal structure that provides independent stewardship of data" [166]. There are a number of interpretations of data trusts, e.g., it is assumed that a data trust could be simply an arrangement of governance or a legal agreement or such practices could be aggregated into architecture [253]. Hardinges places different interpretations and uses of data trust term into the following categories, including repeatable framework of terms and mechanisms; a mutual organization formed to manage data on behalf of its members; a legal structure; a store of data with restricted access; and public oversight of data access [165]. For instance, Sidewalk Labs proposes the establishment of an Urban Data Trust (that could evolve into a public sector agency over time) serving as an independent digital governing entity for their Sidewalk Toronto project, ensuring that responsible data handling is in place for digital innovation activities (Responsible Data Use) [214]. In addition to privacy laws, Sidewalk Labs suggests that all innovations aiming to collect/use urban data must go through Responsible Data Usage Assessment conducted by Urban Data Trust. This way, Sidewalk Labs aims to achieve the proper privacy and security practices, provide and use consistent and transparent guidelines for responsible use of data, and make urban data a public asset [214]. These goals align with O'Hara's emphasis on the purpose of a data trust, which is "to define trustworthy and ethical data stewardship, and disseminate best practice" [253].

Generally, successful engagement in any form of data-sharing partnership could require the adaptation of urban governance visions and strategies [202], as well as a transformation of the parties' institutional cultures and processes [124]. A certain level of data quality could be expected from commercial or private–public partnership data, since such data are often an asset for the

[17] http://datacollaboratives.org
[18] https://9292.nl/en

commercial success of the organizations. However, the technological and methodological biases should not be excluded, since the data could be generated for a particular purpose, but shared for potential other ones [200, 201]. Moreover, partnerships could suggest proper formalization of the responsibilities in data sharing (e.g., data representation models and metadata availability), usage (e.g., who, how, and for what purpose), and maintenance processes between collaborating parties, making sharing and usage of the data smoother.

4.2 Data Heterogeneity and Integration

During the last few years, a large amount of heterogeneous data has been available from various applications and tools. This is also true in the smart cities context, where the rapid adoption of intelligent applications has created new, different, and numerous data collections. These new sources have given new opportunities but also emerging challenges. An effective data analysis in the smart cities context has to consider the increasing amount of data coming from connected devices, multiple software solutions (developed by public and/or private institutions), and historical archives. However, since the systems producing and collecting data are heterogeneous, they provide data in multiple formats that must be integrated to be combined for running an effective analysis. The siloed and often incompatible nature of these sources has also made the interpretation and use of data more challenging [279]. We will explore the different strategies that, according to the literature, can be applied for integration of data for smart cities, summarized in Table 6:

—Model data integration
—Semantic data integration
—Structural data integration
—Software-delegating data integration

Model Data Integration. This approach to data integration has been developed in the previous decades starting from proposals focused on the integration of classical data models (such as Relational, XML, and Object-Oriented) [161, 237], and continuing with suggestions more focused on recent data formats (such as streams, NoSQL databases) [66, 217]. According to this methodology, all data, coming from different sources is collected in a central repository where an abstract model, grouping all the characteristics of the diverse sources, supports all the operations [78]. A major benefit of this methodology is that data collected and integrated (in theory) contains no redundancy, can be accessed uniformly, and can be trusted thanks to its integrity. Unfortunately, the definition of such a model is difficult since integrating concepts coming from different data models is not always easy. For example, it could be quite challenging to integrate two dissimilar concepts into the same model, such as a link from a graph data model and a column from a columnar data model. Moreover, the characteristics of Big Data make the maintenance of such a unified model tricky since the data model must be updated each time a new data source with a different data model is defined and needs to be integrated.

In the context of smart cities, the work by Ballari et al. [70] presents one of the first approaches in this direction. The authors focus on integrating sensor data and highlight the difficulties in finding a global scalable solution. Even though they introduce a global model (providing dynamic interoperability and considering the concepts of proximity, adjacency, and containment in different dynamic contexts), they still cannot manage to introduce a global schema that can be used to store data in a scalable manner. The CitySDK project [269] goes in the same direction, defining a global data model for integrating data concerning tourist information. Their global model designs structures for points of interest, events, itineraries, and categories/tags. The approach bases the data collection on a set of adapters that transfer the information from heterogeneous sources (mainly CVS, JSON, and XML files) to the global data model (implemented in document format and stored

Table 6. Data Integration Strategies in Smart Cities, with Their Benefits (+) and Challenges (–)

Model Data integration	Semantic Data integration	Structural Data integration	Software-delegating Data integration
All data belongs to a unified schema in a target meta-model	A general domain ontology represents all the concepts	Data integration occurs at the physical storage level	Off-the-shelf software is used for integration
+ Unified vision of data + Allows to identify and possibly eliminate data redundancy	+ Modularity and scalability + Easy and transparent integration of new data sources	+ Transparent to the high-level analysis + Unified and efficient data access patterns	+ Ready-made solutions + Modular solutions: new developments easily extend models
+ Algorithms defined in a general way on the global schema	+ Reasoning on objects and their relationships	+ Operations at data-fragment level that can scale-up easily	+ New analyzes can be included with new components
– Users must have a high capacity of abstraction	– Domain expert knowledge is required	– Security and privacy are fully delegated	– Data access depends on platform and its capabilities
– Usually, standard query languages are not available	– Already-available ontologies do not always fit the target scenario	– Access from external software and platforms is not easy	– Updates from vendors can affect the global design
– A new data source can impact the general model	– Poor support for stream analysis	– Data must have a uniform storage format and granularity	– Strong dependency on platform capabilities
Examples:	Examples:	Examples:	Examples:
[70, 92, 207, 237]	[37, 79, 84, 111, 117, 148, 156, 276]	[112, 128, 271, 279, 280, 285, 289]	[122, 278, 287, 307]

in MongoDB) using a REST API. This approach tries to solve the problem of the flexibility of the central data model by requiring the definition of a new adapter each time a specific data source is added to the system.

More recent approaches have managed to establish architectures based on the meta models provided by new technologies. This is the case of the data hub-like architecture, proposed by Koh et al. [207]. This approach integrates the technologies of stream processing, like Apache Kafka [10] with the support of Apache Spark [14] (also used for batch processing); the knowledge graph-structured base of Virtuoso for semantics, and the storage of Apache HBase [9] for quick real-time retrieval. Finally, they use Vert.x [42] a Java framework to provide scalability through its natively asynchronous task processing and abstraction of microservices. The design is still quite new and would have to be tested to evaluate its performance.

Cacho et al. [92] proposed viewing a smart city as a **System-of-Systems (SoS)** to help develop a framework upon which governments can benefit from the integration of public and private systems for planning, administrative, and operative purposes. They also identify challenges to the development of smart cities, namely: the escalation and complexity of the SoS to be developed, the multitude of stakeholders, the variety of domains, and emergent behaviors of the systems within.

In this context, they described the challenge of the unification of the information to handle the heterogeneity and the interoperability of the system under analysis using a global meta-layer.

Semantic Data Integration. One popular strategy for data integration is to use knowledge representation and ontologies. In computer science, an "ontology is an explicit specification of conceptualization. The term is borrowed from philosophy, where Ontology is a systematic account of Existence" [157]. To define an ontology on the top of a domain, in computer science, a representation of the knowledge with a set of concepts within a domain and the relationships between those concepts must be provided. This approach has been implemented and described in multiple cases, like [71, 84, 275, 325]. The benefits of semantic data integration are modularity, scalability, and the fast and easy integration of different formats of data while removing the need to have a centralized system to store all the data together. Bansal et al. [71] define a general Extract-Transform-Load framework, involving the creation of a semantic data model as a basis to integrate data from multiple sources. This is followed by the development of a distributed data collection that can be queried using the SPARQL query language. Psyllidis et al. [276] focus on the smart cities domain and present a similar approach. The data from multiple heterogeneous urban sources are integrated into a global ontology. On top of that, the authors define various interactive Web components (e.g., a Web ontology browser and interactive knowledge graph) to access the integrated ontology graph. Bianchi et al. [79] try to combine the definition of a semantic layer with a tool that provides to domain experts the possibility to perform in autonomy the integration of multiple and heterogeneous smart city data sources. Gaur et al. [148] propose a multi-level smart city architecture integrating data from wireless sensors for pressure, temperature, electricity, and others. Their architecture is composed of four layers and each layer has one responsibility. Layer 1 receives data in many different formats. Layer 2 is in charge of processing all the data into a single format, like **Resource Description Framework (RDF)** format. Layer 3 contains the inference engine for data integration and reasoning using semantic web technologies. Finally, Layer 4 is responsible for querying data. A different approach based on RDF-format data integration is presented by Consoli et al. [111]. There, the authors describe a platform implementing an ontology-integration approach that leverages the help of domain experts. For each data source, an ontology is created. The common conceptual layer allows to convert all the data in a target RDF data model. A similar solution to the RDF-format data integration from sensors is presented in [326].

Bischof et al. [84] share the consensus on the effectiveness of a semantic modeling strategy for smart cities and on the conceptual data model. The approach considers the data stream annotation with descriptions for data privacy and security, and data contextualization using hierarchies to categorize smart city data. In detail, the solution is based on the definition of a semantic description for smart city data, which is heterogeneous in nature, to facilitate discovery, indexing, querying, and so forth for future services. They consider data heterogeneity not just from the format point of view but also explore the nature of the data considering, for example, the different units of measurement that are provided. They propose to start collecting metadata and semantic descriptions and try to find a compromise with respect to the volume that this metadata might represent. The approach ends with the definition of the Semantic Sensor Network ontology developed by a World Wide Web Consortium incubator group which focuses on organizing and describing sensor capabilities and data processing. The HyperCat [117] project developed a standard knowledge representation using knowledge graphs to provide a uniform and machine-readable way to discover and query data distributed among many data hubs, where each data hub can provide inputs from different IoT components and networks. In this approach, applications can identify and use the data they need independently on the specific data hub they belong to. Finally, we can also cite the CityGML open data model based on XML format that is a standard for the storage and exchange of virtual 3D city models [156].

A semantic data integration approach is of interest to organization bodies as well. For example, it has been proposed by the Alliance for Internet of Things Innovation working group. Special attention must be devoted to the SAREF extension for smart cities [37] that provides a detailed model for some interesting use cases. The ISO [28] also works on smart city ontologies, for example, the foundation level concepts [31], the indicators [29] (populations, and so forth), and the city-level concepts [31]. These ontologies constitute a very interesting and rich source for developing standardized access tools and models and have been considered in multiple approaches that follow a semantic modeling strategy.

Structural Data Integration. Many efforts have recently considered data integration from a less abstract point of view and explore new possibilities offered by cloud platforms or data distribution tools. This kind of data integration looks at data as small pieces that must be integrated from the structural point of view. No generic data model is provided and no abstraction is defined at the application level. Structural data integration differs from model data integration because it does not strictly need a generic and abstract schema in a target model unifying the global vision of the data. This kind of data integration differs also from the software data integration that we will see below because it operates at the physical layer. The integration step is done in the storage layer of the platforms and frameworks. It is immediate to see that the data integration step is purely handled from a technology and a structural point of view. Petrolo et al. [271] tackle the challenge of creating a smart city from the sensor standpoint. That is, they approach the problem from the bottom-up and focus on the layers of data generation and consolidation. The authors propose a VITAL Platform combining the IoT and the Cloud of Things to help alleviate the heterogeneity of data generated from different systems on a pay-as-you-go scheme. This platform combines several protocols and communication technologies, including ontologies, semantic annotations, linked data, and semantic web services to promote system interoperability. However, they mention that the challenges that still remain to be tackled are big data and privacy and security issues. Both of these challenges have been approached by Rodrigues et al. [289] with their SMAFramework. Their framework promises to reduce the trouble of dealing with multiple heterogeneous sources (both historical and real-time generated) while allowing for multiple layers of access and security that can satisfy arising privacy and security norms. Furthermore, the SMAFramework can add additional data sources in a plug-and-play fashion. Their framework is based on a Multi Aspect Graph, which they have tested on geospatial and temporal data from New York City combining tweets with trips carried out by yellow taxis. Puiu et al. [279] propose a distributed framework called CityPulse to perform knowledge discovery and reasoning over real-time IoT data streams in cities. Their architecture includes a layer called "Sensor Connection," which is responsible for collecting the read data from the different sensors. Later, the data gathered is passed to another layer that parses it to extract relevant information. After the parsing, there is a module that performs semantic annotations by using an ontology created within the CityPulse framework. After the messages are annotated, the data are published in a message bus. Since data in the bus is already annotated with the Uniform Resource Identifiers from the framework ontologies, an RDF Stream Processing module is able to query the data over the streams. Moreover, the framework is able to discover certain events based on the analysis of the incoming annotated streams. Finally, they use a Service Oriented Architecture to allow consumers to query relevant streams of the different sources or events that were discovered in the message bus.

ML has also become a powerful methodology nowadays. According to research [128, 297], there is a synergy between ML and data integration and it becomes stronger over time. Modern ML models help to solve the schema-matching phase that can be considered one of the hardest problems in data integration [76]. For example, Deep learning allows the comparison of long text values by their embedding representations and starts to show promising results when matching texts

and dirty data. Recently, SLiMFast [285] has been proposed as a framework that expresses data fusion as a statistical learning problem over discriminatory probabilistic models and that can be adapted to explore the smart city data integration scenario. In the same context, Costa et al. [112] define a framework having a unified data warehouse that collects and stores all the available data in raw format. Their approach uses an internal model that exploits the characteristics of the Hadoop framework [8]. Unfortunately, their meta-model is not accessible from the outside and not many details about the conceptual data integration task are provided. Finally, Raghavan et al. [280] propose a prototype application based on a cloud-based API and architecture. Their solution defines specific layers providing (and restricting) simple but useful standard operations that hide the heterogeneity of the components. In these approaches, the tuning and optimization phases are critical steps that strictly depend on the characteristics of the input dataset. The challenges behind the generalization and optimization of these methodologies are just at the first exploration phase, and much interest has been shown by the database research community [146, 334].

Software-Delegating Data Integration. During the last few years, a new category of data integration approaches has been developed leveraging the power and the flexibility of the data access software layers available on cloud computing platforms and architectures. We classify these approaches under the name of software-delegating data integration. Specifically, this kind of data integration is performed by using the various services that are provided by the cloud platforms [129]. For example, Ribeiro et al. [287] propose an architecture based on microservices developed on the top of the Hadoop framework. Their proposal implements and improves the approach presented in InterSCity [122] with a more scalable objective. An approach also based on distributed architecture is described in [307]. In the proposed approach, data are collected from heterogeneous sources, converted internally in a target model according to a common protocol, and made available for the target analysis. This approach can be used in any context and can be exploited also by smart city applications. A similar scenario is also presented in [116] where a data integrator component is in charge of dispatch requests to data sources. Software-delegating data integration is very flexible and allows quick access and integration of data according to standard operations and patterns. On the other hand, the integration possibilities and the global maintenance become fully dependent on tools and operations offered by specific platforms and offered APIs. Any change and the evolution of the APIs can change the result and impact the data access.

4.3 Data Management

In recent years, data has gained significant momentum with the evolution of smart cities; therefore, data management at such a scale brings challenges [69, 239]. Big data tools and technologies now support data acquisition, storage, analysis, and governance [69]. However, given the volume, heterogeneity, and distributed environment nature of smart cities, it is still difficult to integrate and manage smart city data [258]. This section will explore the challenges and state-of-the-art solutions for data acquisition, integration, storage, analysis, and governance.

Data Acquisition. Data collection or acquisition means retrieval of the data from the data sources and feeding this data into the analytics platform for storage and further processing [336]. Data in smart cities is generated by diverse sources such as IoT, economic platforms, government offices, transportation, and social media [56, 228]. These data vary greatly in their nature (text/images/video/numeric), velocities, and formats. Some data sources are quite *static*, that is, they do not change often, like geospatial map data. Some data sources provide data at regular long-enough intervals, such as daily or monthly. Often, such static data sources have defined APIs to get the data, or data may be downloaded from other storage solutions. Since such data does not need to be processed and analyzed immediately, it can be loaded onto the data analysis platform, integrated

with other data sources, and made available for deeper offline analysis (so-called batch processing) [203, 336].

Many data sources generate data *continuously and at a high frequency*, like sensor readings. Often, such data needs to be processed as it becomes available, to react quickly or detect certain patterns or anomalies. Such incrementally available data are referred to as a stream, the data record as an event, and the near-real time processing of data as stream processing [203, 336]. In data stream terminology, we have producers (who generate events) and consumers (who process events) [203]. Collecting and processing streaming data requires dealing with delayed, missing, or out-of-order data; managing situations where producers send messages at a faster rate than consumers can process; and ensuring fault tolerance [203, 324, 336]. This also means that streaming data requires loosely coupled communication schemes. Common approaches here include messaging systems [203] that implement different communication patterns. For example, in a request-reply pattern, the client expects a reply from the server. In a publish-subscribe pattern, clients subscribe to certain messages published by the server they are interested in. In a pipeline pattern, producers push the results, assuming that consumers are pulling for them [336, 341]. Message-queuing systems facilitate communication between producers and consumers by inserting and reading the messages in the queues [336, 341]. Such an approach provides loose coupling in time, solving a number of challenges of streaming systems, such as the lag in capabilities to process events. Another issue is to handle the heterogeneity of producers and consumers. Message-queuing systems treat this via message brokers, namely application-level gateways that convert incoming messages into ones that recipients can understand [336, 341]. For example, in a publish-subscribe pattern, the brokers match the topics subscribed by the consumers to the topics published by producers [203, 336, 367]. Examples of such systems include Apache ActiveMQ [3] and Apache Kafka [10].

The recent developments in big data and smart cities have given birth to a number of reliable, fault tolerant and flexible data acquisition and ingestion solutions, like Apache Flume [6], Apache Spark [14], Apache Kafka [10], Apache Flink [5], and Apache NiFi [11]. Each of these frameworks is being widely used in academia and industry depending upon the requirements. In some cases, only one framework can meet the requirements, whereas the combination of these frameworks has also been observed [239, 258]. Therefore, while choosing any of such frameworks, one needs to be heedful of the final requirements. For example, if the data are being collected at its origin, it may require initial transformation and cleaning. In addition, as the data sources may have diverse acquisition frequencies and may require frameworks with capabilities for handling low-latency and batch-oriented data alongside data cleaning and data transformation functionalities.

Data Storage. The number of connected IoT devices worldwide is expected to reach 50 billion [296]. Since data are a key ingredient for smart city services, solutions and tools for efficient data storage and access are needed [93, 125].

Generally, smart city applications can be considered to be data-intensive. In addition to application-specific requirements, such applications should ensure that data are stored reliably and available for later use, search, and processing, and the results of expensive operations should be saved for speedy retrieval [203].

In recent years, a number of advanced SQL, document, graph, NoSQL, NewSQL, and Big Data data storage systems have been proposed and adopted by researchers and engineers. It is clear that some of them work better for certain tasks, provide certain guarantees, and the choice is always made based on the data model and system requirements [93, 168, 203]. Examples include MongoDB [33] which is a widely used document database, Apache Cassandra [4] as a representative of wide-column data storage solutions, or VoltDB [43] as a representative of NewSQL databases. Modern storage solutions enable distributed storage and processing by utilizing replication and sharding; they provide data querying capabilities and interfaces for most commonly used programming languages

and third-party systems, and cluster management functionality. Distributed implementation enables scalability, fault tolerance, and latency reduction. However, as CAP theorem says, "in a distributed database system, you can have at most only two of Consistency, Availability, and Partition tolerance" [168]. Here, Consistency refers to the property to deliver every user of the database an identical data view at any given instant; Availability promises an operational state in the event of failure; and Partition tolerance ensures the ability to maintain operations in the case of the network's failing between segments of the distributed system [168]. Therefore, in distributed implementations, usually, there is a tradeoff between consistency guarantees and other features.

Off-the-shelf big data management and processing platforms are available, such as Apache Hadoop [8] and **High Performance Computing Cluster (HPCC)** Systems platform [26]. Such platforms and the software ecosystem of applications developed around them provide complete solutions from data acquisition to data storage, analysis, and results delivery to the end user. *Apache Hadoop* is an open-source Java-based framework developed for data storage and processing in a distributed environment on commodity hardware. The main components of Apache Hadoop are: the **Hadoop Distributed File System (HDFS)**: a distributed file system facilitating storage and high-throughput access to massive-scale data; Hadoop YARN: a cluster resource management framework; Hadoop MapReduce: a system for parallel processing of data; and Hadoop Common: common utilities supporting other modules [8]. In addition, a number of tools have been developed for different purposes, e.g., to efficiently load the data to HDFS (like Apache Flume [6]), facilitate data storage and access (like Apache HBase [9]), process and analyze the data (like Apache Flink [5] and Apache Spark [14]), and to maintain configuration (Apache Zookeeper [17]).

The *HPCC System platform* is an open-source data lake platform supporting different data workflow capabilities [296]. Its main components are: Enterprise Control Language—a data-oriented declarative programming language; Thor—a bulk data processing cluster that cleans, standardizes, and indexes inbound data; and Roxie—a real-time API/Query cluster for querying data after refinement by Thor [40]. It also uses a distributed file system for storing data in the cluster following a record-oriented approach [241]. The indexed data available in Thor clusters can be used for low-latency querying by copying in Roxie clusters, which has been specifically designed for much faster results, unlike the Thor Cluster with batch orientation [241, 296]. In addition, as in Apache Hadoop, data are collected using different data acquisition frameworks such as Apache Flume [6], whereas in HPCC Thor, simply a web service can be used for uploading data to Thor clusters [267].

A number of big data storage solutions have also been proposed. For instance, *Apache Ozone* [12] is a scalable, robust, distributed object store for big data applications. It is designed to handle large amounts of data consistently, providing HTTP interfaces for integration with third-party applications. Ozone is built on top of existing Hadoop components, such as Hadoop YARN, HDFS, and Hadoop Key Management Server, and leverages their capabilities and integrations [262]. Ozone is also compatible with the existing Hadoop ecosystem, such as MapReduce, Spark, Hive, and Impala, and can be deployed alongside HDFS or as a standalone storage system. Apache Ozone in comparison with HDFS has several benefits. For example, HDFS has a single namespace which can become a major challenge for metadata operations. It does not support object-based protocols, such as **Simple Storage Service (S3)** [351], which are commonly used in cloud-native applications these days. Moreover, the fixed block size in HDFS can lead to inefficient storage space utilization and network overhead when it comes to small files. Apache Ozone supports S3 protocol and implements Hadoop Compatible File System to cater different application needs and preferences. Ozone also provides a rich set of features, such as security, replication, fault tolerance, and monitoring [351]. The fault tolerance of Ozone is ensured through its self-healing properties that allow it to

recover from sudden node failures, making the data highly available. In addition, it is capable of supporting a hierarchical namespace, enabling the maintenance of data in multiple buckets and directories [12].

Smart city services often need to analyze patterns of moving entities changing their location in time (like vehicles or mobile phone users) or extent as well (like the spread of epidemic diseases) [158]. Such time-dependent geometries are called moving objects [158], therefore, storage solutions should be equipped with the opportunities to represent and query the dynamics of such data. Ilarri et al. [180] categorize state-of-the-art support for moving objects into two categories: **Moving Object Databases (MODs)** and data streams. However, they do emphasize that the boundary between these two groups is not always clear. MODs enhance database technologies with representation and management of moving objects [158, 180]. When compared to early spatio-temporal databases, MODs also allow for tracking continuous changes [158]. In particular, research has been conducted into models to track moving objects and corresponding query languages, handling uncertainty, indexing ensuring a low update overhead and efficient retrieval of the objects is conducted, please refer to [180] for details. Prominent examples of MODs that are in active development are MobilityDB [376], extending PostgreSQL and PostGIS with the moving object support, and SECONDO [249], an extensible database management system supporting various data models. The development of big data technologies has facilitated the storage and processing of traces of a large number of moving objects. A number of efforts exist nowadays to work with spatial and spatio-temporal big data [342]. Starting from equipping Apache Hadoop with support for spatial data, like data formats, spatial index structures, spatial operations (SpatialHadoop [38]), and spatio-temporal capabilities (ST-Hadoop [39]). To more recent proposals enriching Apache Spark [14] and distributed storage products with spatial or spatio-temporal capabilities. For instance, Apache Sedona [13] extends Apache Spark [14] and Apache Flink [5] with a set of tools for working with large-scale spatial data in cluster computing environments. Beast [134] is a Spark-based solution for exploratory data analysis on spatio-temporal data supporting a variety of data formats. GeoMesa [24] provides a set of tools for large geospatial data analytics. For instance, it adds spatio-temporal indexing on top of Accumulo, HBase[9], and Cassandra[4] databases to store spatial data types like points, lines, and polygons. Stream processing is enabled there by having spatial semantics on top of Apache Kafka [10].

Graph databases enable efficient storage and processing of graph data models, which is often met in the smart city domain, e.g., road network. A graph data model handles varying granularity and hierarchical differences in data well; and enables evolvability, meaning that the graph can be extended to reflect changes in the application domain [168]. Examples of solutions available to help store and work with graph data models in a largely distributed environment include Neo4j Graph Data Platform [34] and the Apache Giraph [7] processing system. Such solutions enable deploying graph data models on large clusters, if needed, and enable distributed graph processing by partitioning the data and processes between the nodes.

Data Processing. Most of the smart city applications rely on processing a large amount of data [339]. Depending on the application's requirements, this processing can be roughly divided into two groups: batch processing and stream processing.

Batch processing, often also called offline processing, takes a large amount of input data, runs a job to process it, and produces the output [203]. It is clear that jobs in batch processing could take a while. Therefore, they are often scheduled to run periodically, like once a day. If we consider the big data landscape of methods and technologies, then the MapReduce programming model [123], allowing processing of a large amount of data in a distributed manner, was the most popular approach, implemented also in the Apache Hadoop framework [8]. A MapReduce job consists of Map and Reduce tasks. First, the input data are split into portions that are processed by Map tasks

in a parallel manner. Then, the results of Map tasks are used by the Reduce tasks to compute the final output. It is also common for MapReduce jobs to be chained together into workflows so that the output of one job becomes the input to the next job [203]. However, the Hadoop MapReduce framework, e.g., does not have direct support for workflows, so the chaining occurs explicitly via storing intermediate results in the file system. This has certain downsides, such as a waste of storage space when intermediate results get replicated, redundancy of some programming code in map tasks, and the inability to start subsequent tasks before the previous ones are completed [203]. Dataflow engines have been developed that aim to solve these issues. They handle an entire workflow as one job rather than breaking it up into independent subjobs. Examples include Apache Flink [5], Apache Spark [14], and Apache Tez [16].

Stream processing, also often called near-real-time processing, processes events shortly after they happen. Therefore, stream processing has lower delays. There are a number of cases, when stream processing is required, such as anomaly detection, finding patterns, or simply streaming analytics. Basic terminology and technologies required to get stream data to processing engines were already presented in the previous Section 4.3. Here, we cover approaches for stream processing. Generally, there are two ways to process stream data: one-at-a-time and micro-batching [203]. For example, Apache Spark allows the use of a micro-batching approach [14]. In this approach, the processing engine splits the input data into small micro-batches, processes them, and produces the micro-batches of the results. The one-at-a-time approach is implemented by Apache Storm [15], for example.

Smart city applications are complex constructs fueled by diverse kinds of data. Therefore, hybrid approaches, combining both batch and stream processing, are often required. A number of architectural solutions to combine batch and stream processing have been suggested [121]. For instance, the Lambda architecture incorporates layers for batch processing, a speed layer for computation on recent data (real-time views), a serving layer which is a specialized distributed database allowing doing queries for batch analysis results (batch views). The query result is composed of both batch and real-time views [235]. Another approach is the Kappa architecture [212], which simplifies the Lambda architecture by removing the batch layer. This architecture relies on the use of a log-based system (e.g., Apache Kafka) able to retain all the data that may be reprocessed if needed. Then, we need to deal only with one type of system and making changes means just running a new instance of the job on the whole data, writing the results into a new table and redirecting the application to read the results from this new table. The old job and old results table can be stopped and removed. Liquid architecture [138] incorporates incremental processing, therefore reducing re-computation from scratch. Davoudian and Liu [121] discuss these and some other data system architectures (incorporating, e.g., Semantic Web technologies).

Data Governance. Data governance refers to the overall management of the availability, usability, integrity, and security of a platform's data assets. In the context of data management, governance covers aspects related to data access control, metadata, the data lifecycle, data usage, and regulation compliance [52]. It involves defining and implementing policies, standards, and procedures to ensure that the data are properly managed, compliant with regulations, e.g., the General Data Protection Regulation, and protected throughout its lifecycle. Data governance sits on top of other aspects of data management, i.e., acquisition, storage, processing, and analysis, and addresses the above-mentioned challenges.

A well-defined data governance framework is critical to ensuring compliance with existing data regulations and potential updates or modifications in real time [198]. A reliable governance framework can also enable evidence-based auditing and granular reporting to the data controllers

and data processors, especially in situations requiring legal examination. Additionally, data lifecycle management offers several advantages, including:

—Enhanced Agility and Efficiency: By ensuring that useful, accurate, and relevant data are readily available to recipients, data lifecycle management increases the agility and efficiency of data handling.
—Robust Data Protection Infrastructure: A well-implemented data lifecycle management system guarantees a strong data protection infrastructure, contributing to overall data security.
—Automation Feasibility: There is the potential to automate data management processes, leading to significant savings in terms of human resources and time.

Once data are created at source, it goes through various stages during its lifetime. These stages include collection, ingestion, storage, access, alteration, archival, and destruction [133]. Various challenges exist when handling data governance at each of this stage in a smart city environment. For example:

—Data Ownership and Sharing: Smart city platforms involve multiple stakeholders, including government agencies, private companies, and citizens. Clarifying ownership and sharing policies for data are crucial to avoid conflicts and ensure that data are shared in a transparent and fair manner [143, 187].
—Mismatch Between Organizational Structures: This may lead to data silos, duplications or lack of control as smart city platforms often involve multiple systems and data sources that may not be integrated [187]. To resolve such issues, organizations must have robust and standardized governance models across the entire data lifecycle, e.g., using the 4I framework [118].
—Interoperability and Data Quality: Smart city platforms rely on high-quality data to make informed decisions and enable intelligent automation. However, data quality can be affected by factors such as data entry errors, duplication, and data inconsistency. As discussed in Section 4.2, ensuring data quality can be challenging, particularly when the data are generated from multiple sources and is in multiple formats [72].
—Data Access Management: Ensuring that data access policies are enforced consistently and efficiently requires a comprehensive automated access management system that includes authentication, authorization, and audit trails. This is challenging in large-scale smart city platforms with multiple stakeholders, and smart solutions are needed to address the challenges, e.g., by using an automated smart-contract driven framework [366].

The dynamic and distributed nature of modern smart city platforms emphasizes the necessity of comprehensive data governance through the identification of each stage in the data lifecycle, and appropriate application of relevant controls, policies, and regulations. Identifying tags and metadata linked to each stage of the data lifecycle is also an essential requisite. This meticulous identification and tagging process from administrators of smart city platforms (see Section 3.3) will not only contribute to effective data management but will also ensure adherence to specific regulations governing each phase of the data's lifecycle [196]. Tools such as Apache Atlas[19] or DataHub[20] provide frameworks to manage metadata and tags and enable enterprises to effectively and efficiently meet their compliance obligations. As an example, some of the above lifecycle stages can consist of, but are not limited to, the following tags:

[19] https://atlas.apache.org/
[20] https://datahubproject.io/

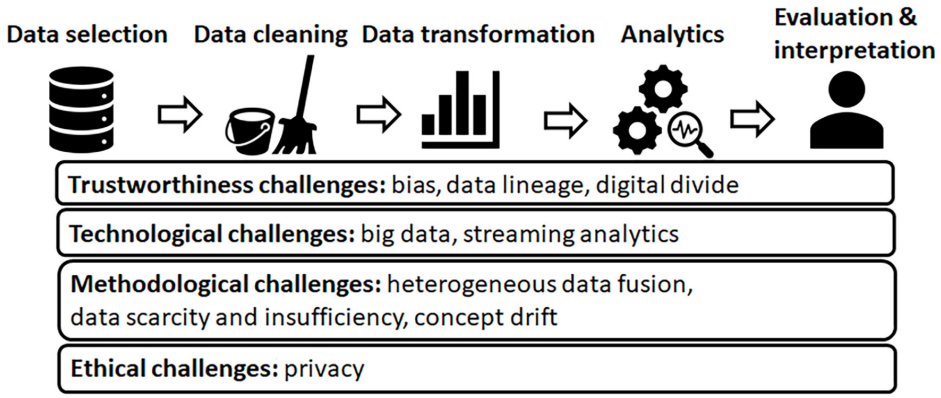

Fig. 4. The data analytics pipeline and some associated challenges within the smart city context.

—Data Collection: source, timestamp, collection region, owner, data format, unit of measurement, and description.
—Data Ingestion: whether the data are encrypted, anonymized and/or transformed, encryption algorithm, and quality status.
—Data Storage: timestamp, cloud provider, retention policies, storage format, storage locations, access point, and checksum.
—Data Access: duration or scope, user ID, role, access type, access, timestamp, and date of modification.
—Data Deletion: deletion method, expiry date, destruction timestamp, retention policy, confirmation, and reporting.

Once these stages and associated tags are identified, an efficient management mechanism can be developed for smart city platforms [281]. Policy engines, such as Apache Ranger,[21] can be employed to implement data lifecycle policies. Such engines and solutions should comprise the following essential components:

—Policy Manager: Maintains a comprehensive list of data regulations and policies that a smart city provider is required to comply with, when handling the user applications and data.
—Auditor: Records events occurring during the data lifecycle and maintains a track of these events for auditing by internal or external third-party auditors.
—Policy Enforcer: Applies the policies and regulations stored in the policy manager to user data stored or processed in the platform. Enforcers can be configured as plugins that run on top of data processing or storage components.

By combining these three elements, a solid foundation for trustworthy data governance can be established for smart city platforms.

4.4 Data Analysis

Data analysis is a key enabler when it comes to finding knowledge about how citizens and smart city operations function and interact, and for discovering unknown patterns and potential for optimization. Often, data, enabling the analysis, comes from the ICT infrastructure of the cities. Larger cities, as well as wealthier communities, are teeming with ICT technologies. However, diversity and inequality in sensing and communication infrastructures exist within and between

[21] https://ranger.apache.org/

cities. These issues further complicate ordinary data analytics pipelines in the smart city context, see Figure 4. For discussion purposes, we have organized the data analysis challenges in the context of smart cities into four categories: Trustworthiness challenges bring issues of reliability, confidence, and the truth of data analytics results; Technological challenges include tools and platforms enabling the analysis of smart city big data streams. Methodological challenges include the development of methods and algorithms to treat particular aspects of urban data, such as how heterogeneous data can be fused for analysis. Finally, Ethical challenges explore issues coming from the rapid equipment of the cities with the ICTs, like data privacy. However, we believe that this categorization of challenges is also relevant to other application domains.

Trustworthiness Challenges. Data analysis requires research to satisfy certain validity criteria, which, in turn, can be compromised by biases and challenges coming either from decisions/choices made over the data processing pipeline or some circumstances over which the researcher does not have control [256]. Table 7 provides comments for addressing trustworthiness challenges.

As it was discussed already in Section 4.1, data are not neutral and contains bias since there are decisions involved on what, how, when, by whom, and for what purpose the data has been measured/retrieved [200, 201]. When, e.g., social data are used for analysis, data platforms may even have embedded functional and normative biases coming from the possibilities to interact with the system and expectations of acceptable behavioural patterns, and so forth. [256]. Moreover, some data collection methods may favor certain kinds of communities over others, e.g., the use of mobile applications for reporting certain city issues, resulting in digital divide issues where not all the communities are equally presented [252]. Therefore, the data itself becomes biased, i.e., contains "systematic distortion in the sampled data that compromises its representativeness" [256]. In addition, often, in smart cities, data used for analysis was originally generated for some other purposes, like, e.g., usage of mobile phone data for identifying mobility patterns [350]. Therefore, a clear understanding is needed of the problem at hand and the data to be selected for analysis, as well as possible bias risks. One way to help in this direction is to support proper documentation of the data source and dataset itself, clearly stating the purpose, phenomena, means, and limitations of the data collection and subsequent use in the analysis. One example are datasheets, proposed by Gebru et al. [149], accompanying each dataset and documenting its motivation, creation, intended uses, and other relevant information.

Challenges also arise when moving the data through the data processing pipeline. For instance, data cleaning, enrichment, and aggregation procedures may significantly affect the dataset content, structure, or representation [256]. For example, decisions should be made on what to consider as an outlier and how to treat missing data. Manual data annotation is also prone to errors and subjectivity. Therefore, the quality of data should be assessed at each step [74].

The data analysis methodology should be adequate for the goals of the research. Moreover, expertise and thoroughness are required in both method selection and results evaluation and interpretation [256]. Algorithms, similarly to data, can be biased. For example, this may be due to the fact that biased data or too little good quality data are used for their construction, or due to design choices selected based on current or limited understanding of phenomena. Koene et al. [205] highlight four potential sources of unfairness in algorithmic systems: biased values in design (e.g., favoring one feature over another), biased training data, biased data (if the made algorithm works with problematic data), and inappropriate implementation of an algorithmic system. Since algorithms are becoming more and more integrated into human lives, appropriate measures must be in place to ensure their fairness, trustworthiness, and impartiality. However, how to assess the fairness of the algorithm is still an open research question [291, 378]. Research into discrimination-aware ML and data mining has emerged to discover and prevent possible discrimination ("adversary treatment of people based on belonging to some group rather than individual merits" [378]). For

Table 7. Notes on Trustworthiness Challenges

Problem	Comment
Data collection/selection	— A clear understanding of the problem at hand and the data to be selected for analysis, as well as possible bias risks, is needed. — Understanding and minimizing the possible effects of the data collection procedures (e.g., by whom and how the data were collected, whether the data collection was available for users of certain devices only, whether the data were collected during unusual circumstances (e.g., large festival), or whether the way the data collection procedure is organized affects the results (e.g., collecting opinions on public devices)). Refer to Olteanu et al. [256] for a deeper discussion of such issues for social data. — Ensure that the sample is as representative as possible [147]. — Proper documentation of the data source and dataset is needed, stating all the details, purpose, phenomena, means, and limitations of data collection and subsequent use in the analysis. An example instrument is datasheets [149].
Data pre-processing	— Understanding of the techniques to be used, their goals, and results. — Documenting the procedures and decisions made at each step (so, it is possible to trace back and forward and assess the made choices). — Data quality assessment at each step [74].
Data analysis	— Methodology should be adequate for the goals of the research; expertise and thoroughness are required in method selection, its proper implementation, results evaluation and interpretation [206, 256]. — Thorough documentation of the steps involved and decisions made. — Discrimination-aware ML [378]. — Algorithm fairness, transparency and accountability [205, 206]. — Explainable AI [224].
Evaluation, interpretation, and results delivery	— Proper selection of metrics and their interpretation, critical assessment [147, 201, 256]. — Proper visualization, interpretation, and explanation [99, 100, 201].

example, solutions have been proposed to prevent discrimination by either pre-processing the training data, model post-processing, or model regularization [378]. Furthermore, transparency and accountability are considered to be promising tools to achieve algorithmic fairness [205].

Efforts exist on different levels addressing particular algorithmic practices in legislation [205]. For example, the Automated Decision Systems Task Force was established in New York City to develop recommendations for the use and policy regarding automated decision systems helping agencies and offices in urban decision-making [50]. Additionally, expert groups and initiatives have been established to acknowledge the importance of dealing with ethical concerns of algorithmic systems. For instance, the Ethics Guidelines for Trustworthy AI have been published by the High-Level

Expert Group on AI, a panel established by the EU [25]. These guidelines present a comprehensive framework with an emphasis on ethics and robustness, aiming to attain reliable and trustworthy AI [25]. The **Institute of Electrical and Electronics Engineers (IEEE)** Global Initiative on Ethics of Autonomous and Intelligent Systems [27] aims to support stakeholders involved in the development of autonomous systems in the ethical implementation of intelligent technologies. This initiative also works on the IEEE P70XX series of standards to put the ethical principles discussed by the initiative into practical guidelines. For example, IEEE P7003 "Algorithmic Bias Considerations" aims to provide a framework that helps developers to identify and mitigate biases in the outcomes of the algorithmic system [206]. Working groups of the AI Subcommittee within ISO and International Electrotechnical Commission Joint Technical Committee (ISO/IEC JTC 1/SC 42) are examining the entire AI ecosystem, involving also aspects of AI trustworthiness [30].

Technological Challenges. A number of technological challenges and solutions to support both batch and stream data analysis were already discussed in Section 4.3. Therefore, we refer the reader to the Data Processing paragraphs of this section for details.

Methodological Challenges. A number of great surveys exist on the methods for urban data analysis, like heterogeneous data source fusion, methods to treat data sparsity issues, data analysis, and data visualization approaches [103, 244, 339, 369, 371, 372]. Therefore, we do not repeating such works here and provide a brief summary in Table 8 with selected methods. However, obviously, the actual landscape of methods used for the urban data analytics pipeline is much larger and readers are advised to go to the original publications for details.

Instead, here we would like to discuss a few important aspects that are usually less discussed in data surveys in the context of smart cities: knowledge transfer and adaptation to real-world changes.

As we have discussed at the beginning of this section, cities (and even regions of cities) vary in the data available. Therefore, due to unique data characteristics, scarcity or data insufficiency issues, the knowledge gained from one urban place cannot be directly applied to another one.

Humans can recognize and apply relevant knowledge and skills from experience to learn new tasks in new situations. For example, a person who can already play one musical instrument can learn to play another one much faster than a person who has never played any instrument before [353]. However, it is challenging to design a computer system able to apply the acquired knowledge and skills to a new, not seen before, task. Moreover, traditional ML and data mining technologies have the assumption that both training and future data come from the same input feature space and have the same distributions [353]. However, this is often not the case in the real-world, as it might be expensive, time-consuming, or difficult to obtain training data that matches the feature space and distribution of the test data. For instance, an activity recognition system may be developed for one person but used by another one with different sensors [353], or some sensing capabilities may simply be unavailable in an urban space [369, 371]. For such real-world examples, it is essential to utilize the already existing knowledge in new situations.

In urban computing, for instance, some knowledge could be received from one city and partially used in a city that does not possess that much data. It is clear that we cannot directly transfer the inference model learned based on the source city data, as the variables of interest in the target city could differ in their availability and characteristics. However, the relations discovered in one city could hold for the city of interest, and this information could be useful for the problem at hand [371].

To achieve knowledge sharing between urban spaces, *transfer learning* methods can be utilized. Transfer learning methodology transfers knowledge between domains [352, 353, 361, 372, 375]. Examples of implementing transfer learning in urban computing can be found for traffic and

Table 8. Selected Approaches for Smart City Data Analytics Pipelines, Often Encountered in Related Work

Problem	Examples of approaches	Comment
Data cleaning (missing data)	— Elimination — Imputation (e.g., by utilizing historical data (time series models, like ARMA, SARIMA), neighbors (e.g., inverse distance weighting, kriging for spatial data), for spatio-temporal models—collaborative filtering, and multi-view-based learning) [371]	Missing data can be kept as well if the method to be used can handle them.
Data cleaning (noisy values)	Domain knowledge boundaries, outlier detection, smoothing (Kalman filter, particle filter, and discrete wavelet transformation) [371]	Sometimes can be treated as missing data or with winsorizing. It could happen that an "invalid" observation indicates anomaly [102] or drift [88, 145].
Data normalization	Min-max, Z-score, and decimal scaling [147]	
Data transformation	Linear, quadratic, and Box-Cox [147]	
Data integration	— Stage-based methods (different datasets at different stages of data-mining task) — Feature level-based methods (direct concatenation to be used with regularization; feature learning with deep neur. netw.) — Semantic meaning-based methods (multi-view, similarity, probabilistic dependency, and transfer learning-based) [369, 371]	Zheng [369, 371] challenges the conventional data fusion in urban computing big data domain and categorizes promising methods for cross-domain data fusion.
Data sparsity	Collaborative filtering, matrix factorization, tensor decomposition, semisupervised learning, and transfer learning [369, 371]	
Data reduction	— Data reduction (feature selection (filter, wrapper and embedded), feature extraction/construction, PCA), — Sample numerosity reduction — Cardinality reduction (discretization) [147]	The idea is to get a reduced representation of the original dataset.
Spatial trajectories	— Trajectory pre-processing (noise filtering, segmentation, compression, map-matching, stay point detection, transformation if desired (e.g., to graph, matrix, and tensor)) [370, 371], — Indexing/retrieval [370, 371], — Analysis (pattern mining, classification, anomaly detection, uncertainty reduction, and privacy) [107, 190, 259, 370, 371]	Spatial trajectories data are often fused with other kinds of data [152, 153, 259, 370]. Application examples include transportation and mobility patterns, public safety, and health [259, 370, 371].
Data analytics (supervised learning)	Support vector machines (Class, Regr) K-nearest neighbor (Class, Regr) Random forests (Class, Regr) Decision trees (Class, Regr) Linear regression (Regr) Bayesian classifier (Class) Linear discriminant analysis (Class) Learning vector quantization (Class) [159, 339]	Application cases include blackout prediction for smart grid applications (SVM), human motion classification (Multi-class SVM), power management system (K-nearest neighbor), street lighting (SVM and Random Forests), air quality (Random Forests—linear regression), and smart grid system (Decision Trees) [159].
Data analytics (unsupervised learning)	k-means (Clust) DBSCAN (Clust) OPTICS (Clust) [159, 339]	Application cases include load profiles of smart meters (k-means), household power consumption from smart meters (DBSCAN) [159].

Continued

Table 8. Continued

Problem	Examples of approaches	Comment
Data analytics (reinforcement learning)	Q-Learning [159]	Application example is an adaptive traffic signal control (Q-Learning) [159].
Deep learning	Recurrent Neural Networks (strong temporal dependencies, NLP, and speech recognition) Convolutional Neural Networks (strong spatial dependencies, image, video, speech recognition, and NLP) Deep Belief Network (unsupervised feature learning) Stacked Autoencoder Network (unsupervised feature learning) Restricted Boltzmann Machine (feature learning, collaborative filtering, and dimensionality reduction) [103, 159, 339]	Application cases include traffic prediction, healthcare analysis, and air quality prediction (RNN) [103], public safety, transportation (CNN) [103, 159], transportation, healthcare (DBN) [103], traffic prediction (SAE) [103], transportation, healthcare (RBM) [159]. Hybrid approaches are used as well [103].
Visualization	Some examples from [99, 100, 126, 159]: — spatial analysis (heatmap, dot map, and flow map) — temporal analysis (timeline, streamgraph) — relationships exploration (parallel coordinate plot) — hierarchical analysis (tree diagram, sunburst chart) — indicators monitoring (icon, gauge chart) — distribution analysis (box plot, density plot, and histogram)	Selection of the technique depends on the objective of visualization and the phenomenon to be visualized and should be done thoughtfully [201].
Production model monitoring	— Data drift (various data validation techniques, starting from schema definitions/validations to continuous monitoring and logging of statistical properties [88]) — Concept drift (active/passive methods) [145, 240]	Monitoring model performance helps to address challenges of the real-world and respond accordingly by initiating retraining or full update of the model.

Based on related work review [88, 99, 100, 103, 107, 126, 145, 147, 159, 190, 244, 259, 339, 369–372]. CNN, convolutional neural networks; DBN, deep belief network; NLP, natural language processing. PCA, principal component analysis; RBM, restricted Boltzmann machine; RNN, recurrent neural network; SAE, stacked autoencoder; SVM, support vector machine.

human mobility prediction [189, 191], points of interest recommendation [127], and for optimizing locations [177, 223].

Pan and Yang [260] and Zheng et al. [372] provide a great introduction to the topic and taxonomy of transfer learning methods in general. When discussing what could be transferred in urban computing scenarios, Yang et al. [361] suggest three general categories: cross-modality transfer, cross-region/cross-city transfer, and cross-application transfer. *Cross-modality transfer* here refers to the situation when some data modality is missing from the region of interest, however, it is present in another one. Therefore, it would be useful if the information about the modality of interest could be inferred and used to improve the performance of the target application. *Cross-region/cross-city transfer* refers to the cases when the knowledge from data-rich cities is applied for the same or similar application in another city. *Cross-application transfer* refers to the cases when the knowledge is retrieved from some existing related application for which the data are available [361].

The research community suggests different methods to facilitate knowledge transfer in urban computing [361, 371, 372]. However, there are also certain challenges. First, urban computing has some unique characteristics, such as heterogeneous data modalities and spatio-temporal patterns and relationships [352]. Also, Yang et al. [361] emphasize the challenges of finding the appropriate source domain, application-specific linking of the source and target domains requiring expertise

with different methods, data privacy-preserving issues becoming more common in urban computing, and assessing transferability—that is quantitatively measuring the possible gain from applying transfer learning methods for particular resource and target domains.

Cities are also living constructs that constantly change, making previously gained knowledge obsolete. For example, developed ML models can decrease their performance due to evolved changes in data that occur for different reasons, e.g., because of changes in the physical environment where the sensor was deployed. This implies that systems developed for smart cities should be able to detect the changes and adapt themselves to provide adequate performance. Such issues are related to adaptive learning and concept drift adaptation [145, 225, 240]. Here, the *concept drift* refers to a phenomenon when the data distribution changes over time in a dynamically changing and non-stationary environment; and *adaptive learning* means updating the ML models on the fly in response to concept drift [145]. A large number of approaches have been suggested to deal with concept drift [225, 240]. However, there are still a number of challenges. For instance, when we deal with large data systems relying on a number of data streams, it could be a case that a drift could occur across multiple data streams [365]. One more challenge is to deal with multiple types of concept drift that can occur in the real-world [105]. Distributed ML, like federated learning, poses certain challenges for handling concept drift [95]. Concept drift detection research is not well presented for non-traditional data streams, such as when the data are represented as a graph [265]. Finally, there is still not much research on concept drift within unsupervised or semi-supervised settings [145, 225].

Ethical Challenges. It is clear that privacy is one of the major concerns in the smart city context [96, 200]. When we talk about data processing, one approach to preserve privacy is to reduce the amount of data to be transmitted and be able to carry out data analyzes on the nodes with the data themselves. For example, edge computing suggests the analysis of the data in proximity to where the data are collected, therefore supporting privacy [197, 270]. There are also ML approaches that allow learning the model in a privacy-preserving way, for example, federated ML [360]. However, there are certain challenges as well when we deal with distributed ML and edge intelligence, like data scarcity and consistency on edge devices and slow performance of collaborative learning tasks [357]. Here we would like to discuss ethical challenges beyond surveying methods and technologies for distributed data analysis; for such information, an interested reader could refer to, e.g., [344, 357, 360]. Therefore, we explore data-related ethical concerns in Section 4.5, data privacy in particular in Section 4.6, and measures to secure data in Section 4.7.

4.5 Ethics

At face value, the stated goals of smart cities—improved quality of life, ecological sustainability, and so forth—are highly ethical ones. However, concerns have been raised that these goals may be pursued at the cost of harmful side effects, and/or in such a way that some groups are excluded from enjoying the benefits. As noted above, there is no clear consensus on what exactly constitutes a smart city, and as a result of this, compiling a comprehensive and systematic presentation of the ethics of smart cities is a considerable challenge. However, in recent years there have been several attempts to conceptualize and categorize the various ethical concerns relevant to smart cities.

Calvo [94] identifies hyperconnectivity, algorithmization and datafication as key aspects of urban digital society, along with eight major ethical implications of these aspects: intrusion of privacy, social and economic exclusion, misuse of data, bias in decision-making algorithms, obsolescence of human skills and labor, dissolution of responsibility for decisions, objectification of human beings, and the imposition of technology on people. Goodman [154] names three challenges to the democratic governance of smart cities: privatization of functions (e.g., planning) and assets

(e.g., data) traditionally held to belong to the public sector, the conception of cities as platforms offering service providers access to public data, and a loss of autonomy through e.g., technology failure or vendor lock-in. Based on a review of the literature on smart city ethics, Ziosi et al. [377] establish four dimensions that are invariant across multiple smart city definitions and give rise to ethical concerns; these are the network infrastructure, post-political governance, social inclusion, and sustainability.

In a synthesis of the above three articles, focusing specifically on ethical concerns related to the collection and use of data in smart cities, two major themes emerge:

— *Techno-Centric vs Human-Centric Smart Cities:* adopting a techno-centric and techno-optimistic approach to smart city building can lead to emphasizing technological capabilities over human needs and optimizing relatively easy-to-quantify metrics such as economic efficiency over more elusive ones such as livability.
— *Public vs Private Control of Resources and Processes:* as decisions related to city planning and governance are increasingly determined by data and algorithms, power over these decisions is increasingly being transferred from elected representatives and public authorities to private businesses that control the data and provide the algorithms.

From the perspective of data, perhaps the most obvious ethical issue involving smart cities is privacy. Regardless of the definition, the collection and processing of digital data in large quantities is one of the characteristic features of a smart city, and a significant portion of this will be the personal data of the citizens. As pointed out by König [209], collecting ever more data are in contradiction with the principle of data minimization, and even if the data are rendered impersonal through anonymization, it may still become a threat to individuals' informational autonomy through re-identification or re-purposing. A smart city is thus effectively a vast surveillance system with no feasible informed consent or opt-out mechanism available, and while the intention of the surveillance may be benign, the data flows involved may be so complex that the technical and legal safeguards in place are not enough to guarantee the security and privacy of the data. If some of the data are controlled by companies, these may have an interest in exploiting it commercially, exacerbating the risk to privacy; by partnering with such companies, the city is effectively acting as an enabler for what Zuboff [380] has termed surveillance capitalism.

An archetypal example of surveillance technology is the surveillance camera. Combined with modern AI techniques, the footage captured by such a camera is no longer merely something to be viewed by a human authority after some kind of incident has occurred, but acts as input data to ML algorithms for purposes such as facial recognition. **Facial Recognition Technology (FRT)** has a variety of security-related applications that can be argued to enhance safety in the city, but their civil rights implications cannot be ignored; in addition to the privacy issues, facial recognition systems have been observed to be prone to racial bias where people belonging to certain groups are more likely to be misidentified than others [98], leading to concerns about the social justice impact of using biased algorithms for policing. Similar concerns have been raised about predictive policing systems, although there have been few independent empirical evaluations of the fairness of such systems and these have not produced clear evidence linking them to increased discrimination [59]. In contrast, the use of FRT in policing has been found to contribute to greater racial disparity in arrests, although this cannot be simply attributed to algorithmic bias as the sole explanatory factor [192].

Altogether, seven different types of harmful bias (or "sources of downstream harm") in ML are identified in [328]: historical, representation, measurement, aggregation, learning, evaluation and deployment bias. If these are not identified and eliminated, increased reliance on data and algorithms in smart city decision-making will result in decisions whose fairness is questionable. There is an

issue with the transparency of the decisions as well, since the explainability of some popular ML techniques is poor [80, 91] and the implementations may be guarded as trade secrets by their vendors, making it difficult, if not impossible, to subject them to thorough external auditing. Furthermore, there are accountability implications if the transition from traditional to smart city means that governance decisions are increasingly determined by data through opaque computational processes, since there is then a risk that responsibility for the decisions will become detached from traditional democratic processes.

Ziosi et al. [377] use the term "post-political" to describe the increasing role of private organizations and automated decision-making in smart city governance. Besides the issues identified above, another problematic aspect of this is that underneath the ostensible objectivity and rationality of post-political governance through data and algorithms, the selection and prioritization of optimization targets is inherently political, since these reflect the values of the smart city. Goodman [154] captures this by conceptualizing smart cities as digital platforms where the pursuit of efficiency may sideline other important values. Furthermore, they point out that decisions regarding what data to collect and how are also political, and if some groups of citizens are not adequately represented by the data, the members of such groups are at risk of being excluded from the benefits of the smart city. The people most likely to be excluded are those who are affected by existing digital divides and, therefore, are already at a disadvantage [94, 154, 377].

Various authors have criticized the focus on technology and efficiency in smart cities and have advocated a more human-centric approach. Biloria [83] puts this idea succinctly by introducing the concept of an empathic city. Human-centric models for smart city data governance are discussed in [209] and [264]. The MyData Global Network is advocating more human-centric governance of personal data in general; its guiding principles are codified in the MyData Declaration [273], which calls for a transition from formal rights to actionable ones, from data protection to data empowerment, and from closed ecosystems to open ones. Several examples of cities pioneering initiatives aligned with the MyData principles are given by Lähteenoja and Sepp [215].

In the literature, the concept of a smart city is frequently paired with that of a smart citizen. In terms of having an established definition, the latter is even more elusive than the former, but insofar as a smart city is one that emphasizes human values and needs over technological capabilities, a key aspect of smart citizenship is empowerment. From a data perspective, a human-centric smart city is thus one that not merely protects the data of its citizens but empowers them to use data to advance their personal values and goals and to participate in the definition of new data-based services. Technological innovation is a necessary enabler for this, but it is also necessary to ensure that the citizens have a sufficient level of data literacy to take advantage of the opportunities presented by smart city technology, lest this become another divide where some people are excluded from enjoying the benefits of the smart city. Proposed solutions are scarce in the literature, but the Urban Data School initiative described in [355] is aimed at this exact purpose in the context of the Milton Keynes smart city project in the UK.

Table 9 presents a summary of the ethical considerations involved in addressing the data challenges of smart cities. Three relevant smart city aspects are identified here: *smart city goals*, referring to the determination of the objectives and underlying values of the smart city; *smart city governance*, referring to how decisions are made in the planning and operation of the smart city; and *smart city life*, referring to how the everyday life of the individual citizen is transformed in the smart city. Associated with each of these aspects is an opportunity for betterment, and associated with each opportunity are risks arising from the central dichotomies identified above, techno-centric vs human-centric and public vs private.

Table 9. Ethical Opportunities and Risks Related to Three Smart City Aspects

Aspect	Opportunity	Techno-centric risk	Privatization risk
Smart city goals	Sustainability goals and human values as principal drivers	Emphasizing "rational" values, overlooking less readily quantifiable ones	Emphasizing business prospects of private actors, overlooking public good
Smart city gover-nance	Better-informed decisions through judicious use of data and software systems	Fairness/transparency/ accountability issues in algorithmic decision-making	Transfer of political power to entities not subject to democratic control
Smart city life	Empowerment of citizens to co-create and enjoy quality-of-life enhancements	Exclusion of subsets of the population through digital divides	Deterioration of privacy through commercial exploitation of personal data

4.6 Data Privacy

New and emerging technologies are promoting the development of an ecosystem for connected places within smart cities, but at the expense of a rapidly widening threat landscape. Attacks against smart infrastructure and privacy have made it clear in recent years that the demands of the smart city transformation, including data collection and processing needs, face significant multi-level governance requirements, such as the need for more transparency, accountability, and security and privacy [171, 349]. Prior work has focused on establishing comprehensive threat modeling tools and conceptual frameworks to better protect smart cities, as well as describing threat actors, their **tactics, techniques, and procedures (TTPs)**, and how to mitigate attacks against connected things and places [227].

Smart City Threat Modeling. Threat modeling is a method for systematically identifying various types of threat actors, attack vectors, and mitigation actions against malicious activities that may harm applications, networks, or other computer systems [305]. Smart cities have unique cyber risks that span many vertical sectors and industries such as energy, transportation, healthcare, education, and public services. Particularly, the increased interconnection of devices and systems generates new challenges for city security management that go beyond conventional security issues. The four innovations listed below are expected to have, or already have, a significant impact on cyber risks in connected cities [261]: (1) convergence of IT and Operational Technology, (2) the interoperability of new and old systems, (3) the integration and fusion of services, and (4) the proliferation of AI and automation [195].

Against this backdrop, more research on smart city threat modeling is now available, with the goal of developing approaches and tools for better assessing system vulnerability and adopting cyber-security analytics [106, 115, 144, 194, 356, 379]. Similarly, municipal, regional, and national governments are becoming more proactive in their legislative approaches to smart city threat modeling, allowing for a more focused and concentrated approach to smart and connected city cyber security [171, 227]. An indicative example of this trend is the threat model for future smart cities developed by the **European Union Agency for Network and Information Security (ENISA)** covering the healthcare and public transport sectors [136, 220]. Similar efforts have been noticed in the respective national cyber-defense authorities across the globe [113, 171, 288]. The subsections below introduce the high-level components of the threat modeling approaches for smart cities.

Threat Actors. Cyber Threat Actors (CTA) are responsible for a considerable number of threats to smart cities [219]. These are groups or individuals who engage in malicious activities that intentionally aim to harm infrastructure for monetary of other gains. CTA groups are frequently divided into the following categories according to their underlying motives, goals, and known affiliations: (1) cybercriminals, (2) insiders, (3) nation-states, (4) hacktivists, (5) terrorist organizations, and (6) script kiddies [298]. Among these threat actors, nation-state actors, also known as **Advanced Persistent Threats (APTs)**, are regarded as the most dangerous and stealthy operators [219]. The MITRE corporation, which curates one of the most widely accessible knowledge bases of adversary tactics and techniques, currently lists about 135 APT groups and associates (i.e., threat groups, activity groups, and threat actors) that share similar methodologies (i.e., TTPs) and operate in different geographical regions [242].

Attack Vectors. While the classification of threat actors can help analysts determine the magnitude of a threat, smart city threat modeling also requires prior knowledge *vis a vis* the initial origin (i.e., tail) of the attack vectors. Namely, the approach developed by ENISA considers two broad conditions that pivot around the perceived intentionality of a threat [220]. These are distinguished between threats from *intentional attacks* and threats from *accidents*. In the context of public transport systems, intentional attacks can include the following: (1) eavesdropping and sniffing, (2) theft, (3) tampering and alteration, (4) unauthorized use and access, (5) distributed denial of service, (6) loss of reputation, and (7) ransomware. In addition to the threats that might be caused by certain individuals or groups, there is also the possibility of threats being caused by accidents, including: (1) hardware failure and/or malfunctioning, (2) operator or user error, (3) end of support or obsolescence, (4) electrical and frequency disturbance or interruption, (5) acts of nature, and (6) environmental incidents. Evidently, in the context of smart and connected people, places, and things, attacks against data (intentional or accidental) are the most common security threat that can inevitably erode privacy.

Data Privacy Models. Data privacy and confidentiality in the smart city are a diverse problem due to the usage of data aggregation to form links, which makes anonymity difficult to accomplish [349]. In addition to the attack vectors outlined above, research in recent years has concentrated on numerous risks against data privacy. Cyber threats against privacy often take aim at: (1) personally identifiable information comprised of personal attributes such as Social Security Numbers that uniquely identify a person; and (2) quasi-identifying attributes comprised of a combination of attributes, such as name, age, and address that, when combined with external information, may be used to re-identify all or part of the respondents to whom the information pertains. In particular, prior works have looked at various types of information disclosure that can lead to a privacy breach (i.e., to reidentification) [368].

- *Identity disclosure* happens when an adversary achieves the correct mapping of microdata (i.e., individual population unit records files) from a database to an existing real-life entity [229].
- *Attribute disclosure* occurs when the adversary is able to deduce more accurately any additional features of a person from the information accessible in the disclosed data [331].
- *Inferential disclosure* occurs when the attacker can infer or more accurately determine the confidential value of a variable in a dataset by comparing the statistical properties of the released data to the information available [304].
- *Social link disclosure* occurs when an attacker can re-identify a hidden relationship between two users that may lead to identity, attribute or inferential disclosure [368].
- *Affiliation link disclosure* happens when the adversary can determine that a person is affiliated to a specific group, resulting in a higher risk that may lead to identity, attribute or social link disclosure [368].

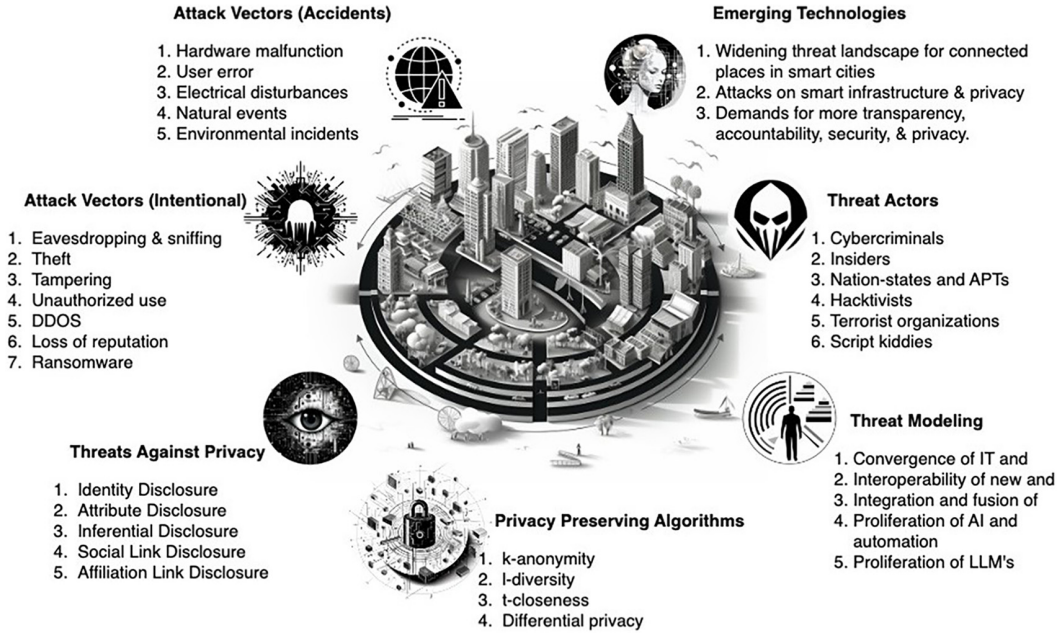

Fig. 5. Comprehensive overview of smart city cybersecurity challenges and threat landscape. DDOS, distributed denial of service; LLM, large language model.

With the advent of data-driven sectors in smart cities (e.g., healthcare, transport, and governance), protecting data privacy without compromising the utility of the collected data has become a conundrum. A number of privacy-preserving algorithms and models have been proposed to address the various information disclosure risks, including k-anonymity [331], l-diversity [229], t-closeness [221], and differential privacy [130]. Similarly, solutions exist toward achieving trajectory privacy [107, 190]. These algorithms leverage anonymization methods including *generalization, suppression, anatomization, bucketization, permutation, and perturbation* [329]. Figure 5 presents an overview of the cybersecurity challenges, threat landscape, and privacy issues encountered by smart cities.

4.7 Data Security

Data security measures include protecting and securing data from unauthorized access or usage at various stages. Generally, the data in a smart city platform may go through three stages. The data can be at rest i.e., stored in the storage medium. The data can also be in the state of transit or motion during internal or external communication. Finally, the data may be in the state of processing, which is when the data are in CPUs, GPUs, or other processing units. Data security implies protecting the data in each of these states [196].

A smart city infrastructure must aim to protect the data stored or processed in the platform through appropriate procedures and encryption methodologies. Similarly, to ensure privacy and protection for data-in-motion, the communications between users and services must also be encrypted. Existing security standards are generally strong and resistant to most attacks, and a significant portion of vulnerabilities in platforms arise from misconfigured devices or lack of implementation of the correct security protocols [313].

Another challenge is the decentralized nature of some smart city ecosystems where numerous devices and applications owned by multiple entities or stakeholders are onboarded to the ecosystem.

The decentralization can create vulnerabilities, as each entity may have different security protocols and risk profiles, potentially leading to weak links in the overall security chain [67]. Blockchains have been used to develop security frameworks to enable secure data communication in smart cities [85]. This is due to various key characteristics of blockchain, or Distributed Ledger Technologies in general, such as immutability, audit trails, resilient consensus mechanisms, and cryptography.

Projects such as Alvarium [61] go beyond by building **Data Confidence Fabrics (DCFs)** that provide measurable confidence scores for data moving from applications and devices. A DCF collects various trust insertion technologies into a single platform and binds them together through standardized APIs and an open framework. Example technologies and tools include the silicon-based Root of Trust [282], open authentication and data ingestion APIs, metadata handling, immutable storage and DLTs [85].

Authentication Authorization and Accounting frameworks are also used to control and track access to resources within a network. For example, SAML [178] or OAuth [218] standards can be used to enable single sign on for smart city networks or user interfaces. However, these solutions face several limitations, especially if applied in isolation. For example:

—Complex implementations that may not be feasible or scalable in distributed smart city environments [311].
—Limited access control features once the users are authorized [327].
—Limited scope with a focus on APIs and lack of support for other security aspects, e.g., data protection [139].

Therefore, to ensure data protection in smart city platforms, it is essential to consider and apply appropriate and comprehensive security measures at each data stage, as discussed below.

4.7.1 Securing Data-in-Transit. Applications running in a smart city platform may interact with other internal system components or user applications running inside the platform, as well as other applications external to the platform. While both these communication types may have varying security requirements, it is essential to implement robust security measures for both. To secure data in transit, it is recommended to use strong and stable cryptographic protocols such as **Transport Layer Security (TLS)**. TLS allows various cipher suites for combining a set of different encryption schemes, key exchange mechanisms, and authentication choices. TLS can also encrypt the data independent of the application-layer protocol being used, making it flexible, reliable, and widely popular [286].

Encryption. Encryption schemes are generally divided into symmetric and asymmetric cryptography. In a symmetric key algorithm, the same key is used for the encryption and decryption of data. Algorithms for symmetric encryption are generally lightweight and fast, but they require pre-sharing the key with all parties involved in the transaction. A prominent example of a symmetric key algorithm is the **Advanced Encryption Standard (AES)** [277]. An asymmetric encryption scheme uses a pair of public–private keys to protect messages between two parties, where the keys are shared using a secure key exchange protocol. Asymmetric algorithms are slower with equivalent security levels and require larger keys. Examples of asymmetric key algorithms include RSA and ECC [68].

In both categories of encryption schemes, the fundamental strength of secure, key-based encryption algorithms is the computational difficulty involved in recovering a plaintext from the ciphertext without the key. A stronger encryption is more difficult to attack but may also be more compute intensive. A smart city platform can select a scheme based on the resource availability and the sensitivity of the data in transit. While selecting a lightweight algorithm, a smart city platform can rely on the lightweight cryptography project by the National Institute of

Standards and Technology that standardizes cryptography algorithms for resource-constrained devices [230].

While encryption algorithms are generally secure, their implementations can be vulnerable to side-channel attacks, deliberate or accidental backdoors, and key exchange in asymmetric cryptography. Homomorphic encryption is an emerging field, which allows computations on encrypted data without decryption. However, efficiency and practicality remain a challenge in this field [234]. Legal and regulatory compliance also require further research to design algorithms that can balance privacy and security [295].

Certificate Authorities (CAs). An alternative to individually sharing key pairs between entities for above-mentioned encryption mechanisms is the use of **Public Key Infrastructures (PKIs)** and CAs. A CA is a trusted third-party organization that issues and manages digital certificates required for secure communication and authentication. The digital certificates contain public keys and other information about the identity of the entity that holds the corresponding private key.

This centralized management of key pairs improves security, scalability and simplifies the management process, e.g., the revocation of all user certificates can be accomplished by modifying the CA. Drawbacks of using certificates may include the initial deployment cost of the PKI and complex configurations. However, given the potential benefits to the higher level of security, a PKI is recommended for smart city platforms for identity verification and encrypted communication. A single point of failure is another drawback due to the centralized nature of the PKIs. Highly available [51] or decentralized [359] CAs can be implemented that achieve the above-mentioned benefits while removing the single point of failure.

Keystores. A keystore is a local storage location in a device for cryptographic keys, digital certificates, and other sensitive information used for encryption and authentication. A software-based keystore is generally a central repository on a device, that securely stores keys and sensitive information of various applications running on the device. Hardware rooted keystores are a special type of stores that are implemented in hardware, such as a smartcard or a trusted platform module, instead of software. This provides a higher level of security, as the cryptographic keys are stored in a tamper-resistant hardware device [199]. Furthermore, the keys are isolated from the main operating system and other software, reducing the attack surface, and making it more difficult for attackers to access the keys. Both software and hardware based keystores offer a secure way to manage and store private keys that are used by different communication channels.

It is important to note that some keystore implementations or libraries may have vulnerabilities or weak security measures, prone to brute force attacks [140]. Therefore, when choosing a software-based keystore, compliant keystores must be selected from reputable and trustworthy organizations. Another open challenge is the scalability of keystores in smart cities with large numbers of devices and entities, especially where computing resources are limited, e.g., at the Edge. As the number of keys and the volume of data increases, managing keys in a keystore can become more complex and may require advanced infrastructure. Hardware keystores, while safer than software, can still be vulnerable to supply chain attacks, where malicious actors compromise the hardware during manufacturing or distribution. Secure onboarding using the DCFs mentioned above [61] can be investigated to address such vulnerabilities.

Virtual Private Networks (VPNs). VPNs provide an additional layer of security to communication by encrypting the data in transit and using secure tunnels to connect devices and entities. A VPN establishes a secure private network connection, often referred to as a tunnel, over an unsecured channel like the Internet. This allows endpoints to communicate securely where the encrypted data are unintelligible to malicious entities or eavesdroppers. From the perspective of the end user, this encryption process is seamless, enabling them to carry out tasks as though they were operating on a local network. Two common tunneling protocols used for VPNs are the

Point-to-Point Tunneling Protocol (PPTP) and the **Layer 2 Tunneling Protocol (L2TP)**. PPTP is simple and fast, but it may not be as secure as other protocols. L2TP provides stronger security than PPTP, but it is also slower and requires more processing power. Smart city platforms can employ different VPN tunneling protocols based on the infrastructure capabilities and application requirements [186]. Managing the overhead caused by L2TP and VPNs, in general, is another open challenge [65], especially in resource-constrained smart city environments.

4.7.2 Securing Data-at-Rest. Encrypting communications through TLS protects the data-in-motion. However, the data stored and resting in storage devices remains susceptible to security breaches or attacks. Firewalls and port blockers provide some protection to the stored data against attacks by restricting access. However, completely securing the data-at-rest can be achieved by encrypting the data-at-rest, which provides an additional layer of defense.

Transparent Encryption Zones. Transparent Encryption Zones in storage systems, such as HDFS [306], are a feature that automatically encrypts and decrypts data as it is written to and read from storage disks in a device. When set up, special directories are created by the system, called encryption zones. Write operations to these zones are encrypted and read operations from the zones are decrypted. Encryption and decryption occur transparently for end-users and do not require modifications to their applications [263]. Moreover, due to the end-to-end encryption, only the client possesses the capability to encrypt and decrypt the data. The storage system or any external application never gets access to decrypted data. Hence, data stored in encryption zones is secure against insider attacks as well.

For encrypting data-at-rest, the commonly used mechanism is AES [277]. However, other mechanisms, such as homomorphic encryption, can be used to achieve more advanced functionality and further improve security and privacy [53]. As storing data in encryption zones requires more computational cost, data owners may consider the sensitivity of information when deciding between plain text storage or encryption zones [210]. Meeting regulatory and compliance requirements can be challenging with transparent encryption, as auditors may require more fine-grained control and visibility into data access and usage, prompting research into novel encryption mechanisms.

Centralized Encryption Service. To enable a system-wide security feature for all applications running in the platform, an encryption service can be deployed that operates in the same way as the encryption zones in the storage system [358]. This centralized encryption service is required to provide fundamental security features such as key generation and management, data signing and verification, and various symmetric or asymmetric encryption algorithms. HTTP or REST APIs may be developed and provided to clients for accessing various security functions as a service and interacting with encrypted data. [196].

An example usage of such a service will include an application initiating a request to the encryption service, asking it to encrypt the data before storing it. To decrypt the data later, the application will send another request to the service, seeking the decryption of the previously encrypted data. The key management component of the encryption service, built with tools like HashiCorp Vault,[22] will be responsible for storing and handling all the keys involved in these encryption and decryption processes. While such a centralized encryption service has the potential to provide data-at-rest encryption for all services on the platform, it is essential to consider and analyze the associated overhead costs and time required for encrypting and decrypting data in comparison to other performance indicators. For example, a real-time application where latency is of critical importance may suffer from system-wide data-at-rest encryption, and the constraints may need to be relaxed for such an application.

[22] https://www.vaultproject.io/

4.7.3 Securing Data-in-Processing. Encrypted data-at-rest needs to be decrypted before a processor can successfully and meaningfully process it. This creates a potential attack opportunity for a malicious entity that may have gained access to the processor, memory, or kernel. **Trusted Execution Environments (TEEs)** can be used to protect the data in such scenarios. A TEE is a tamper-resistant processing environment that runs on a separation kernel. It guarantees the authenticity of the executed code, the integrity of the runtime states (e.g., CPU registers, memory, and sensitive I/O), and the confidentiality of its code, data, and runtime states stored in a persistent memory [294].

The TEE creates an isolated environment that resists against all software attacks as well as the physical attacks performed on the main memory of the system. Attacks performed by exploiting backdoor security flaws are not possible. Because TEE is a relatively new technology, it is expensive and limited in availability [340]. TEEs also require further research into the performance overhead, especially in terms of context switching between secure and non-secure environments. The TEE landscape also lacks uniform standards, as different hardware vendors implement TEEs with varying features and interfaces, raising challenges for software developers. As such, TEEs may only be used for extremely confidential or sensitive data and in environments with very low or zero trust. The security methods and solutions presented in previous sections will suffice for most smart city platforms and use cases.

5 Conclusions

This review article has covered multiple and sometimes overlapping aspects related to smart cities, starting from the understanding of what a smart city is, how it can be measured, and what kind of architectures and platforms could technically facilitate it. It is clear that the concept of smart cities continues to evolve, changing from a very technology-oriented one to a more solid and united concept, entailing societal needs and human potential. It is also interesting to see how the concept gradually combines the views from different research disciplines. Further research is needed to understand how to measure the smartness of a city since it is not so simple. Indicators, if any, should be considered carefully, namely what kind, how to measure and assess the quality of measurements, and how to interpret them. Moreover, cities should be evaluated individually, considering their own cultural and historical circumstances, development goals, and progress.

A number of architectures have been proposed to equip cities with smart services. Obviously, every such solution should rely on the city's own facilities, requirements, and goals. It is clear that there is a strong need for standardized reference architectures that could guide the development of smart city solutions. Standardization bodies have worked on such proposals, and they consider many pitfalls, like suggesting loosely coupled architectures, multi-tenancy, and security. Such reference architectures are highly abstract, and that makes them technology-neutral. In addition, current progress in software development has provided a great number of tools and instruments for the implementation of basic communication pipelines. However, the key challenge is still in data. How the data can be used securely, how the data can be shared, how it can be ensured that the data are used according to the claimed specifications, how to ensure the data quality, how to ensure proper data representations, and there are many more questions. These issues are easy to address when dealing with an individual single system. However, it is challenging to achieve this kind of proper data pipeline in a large-scale ecosystem comprising of a number of data providers, data processors, and services.

This survey delved deeply into the data issues associated with smart cities. We started by exploring the data availability aspects. Here, the topics and development actions toward Open data, citizen-contributed data, as well as commercial data and private–public partnership were studied. Each category has certain challenges. For example, ensuring privacy, guaranteeing quality and usability,

and data lifecycle management are some general open questions. In addition, understanding that data could have a bias is a must, e.g., if citizen-contributed data are collected from some restricted area or from owners of a particular device then it is clear that such data does not present the situation of the whole city. When considering private–public partnerships, trustworthy data stewardship is required.

The smart city domain is quite unique in the variety of data used for the services provided. Therefore, addressing data heterogeneity issues is of utmost importance. We have examined related research, which we have categorized into model, semantic, structural, and software-delegating data integration. Each approach has its own advantages and drawbacks, discussed in the corresponding Section 4.2.

Moreover, data sources in smart cities could generate lots of data and be highly distributed geographically. Smart city services may have different requirements for data processing delays. Therefore, proper data management must be accomplished. In this review, we have explored data acquisition, data storage, data processing, and data governance management issues. For instance, in a smart city, we need various approaches for acquiring the information from the data sources, storing it reliably, searching and accessing it quickly, and both batch and stream processing. Finally, data lifecycle management is very important in the context of a smart city. Therefore, we reviewed the state of the art in data governance, as well as challenges, and possible solutions.

Data analysis is a key enabler for smart city services. Here, we do not review data analysis methods, this can be found in textbooks in general and other related work [159, 339]. Instead, we look more generally at the challenges that the smart city domain brings into the traditional data processing pipeline [372]. Here, we shaped our analysis into: trustworthiness, technological, methodological, and ethical challenges. In addition, we explore in more detail Ethics (Section 4.5), Data Privacy (Section 4.6), and Data Security (Section 4.7) aspects since these are fundamental stones to achieving trust in smart city services.

This article aims to serve as a "one-stop shop" comprehensively reviewing data-related issues of smart cities with references for diving deeper into particular topics of interest. We hope that this work will inspire future research and development on urban computing and related fields.

Acknowledgments

We would like to thank D.Sc.(Tech.) Aapo Huovila for fruitful discussions on standardized indicators for smart cities.

References

[1] AirNow.gov. Retrieved from https://www.airnow.gov/(accessed June 2024).
[2] Amazon Mechanical Turk. Retrieved from https://www.mturk.com/ (accessed June 2024).
[3] Apache ActiveMQ. Retrieved from http://activemq.apache.org/ (accessed June 2024).
[4] Apache Cassandra. Retrieved from https://cassandra.apache.org/ (accessed June 2024).
[5] Apache Flink. Retrieved from https://flink.apache.org/ (accessed June 2024).
[6] Apache Flume. Retrieved from https://flume.apache.org/ (accessed June 2024).
[7] Apache Giraph. Retrieved from https://giraph.apache.org/ (accessed June 2024).
[8] Apache Hadoop. Retrieved from https://hadoop.apache.org/ (accessed June 2024).
[9] Apache HBase. Retrieved from https://hbase.apache.org/ (accessed June 2024).
[10] Apache Kafka. Retrieved from https://kafka.apache.org/ (accessed June 2024).
[11] Apache NiFi. Retrieved from https://nifi.apache.org/ (accessed June 2024).
[12] Apache Ozone. Retrieved from https://ozone.apache.org/ (accessed June 2024).
[13] Apache Sedona. Retrieved from https://sedona.apache.org/1.5.0/ (accessed June 2024).
[14] Apache Spark™ - Unified Analytics Engine for Big Data. Retrieved from https://spark.apache.org/ (accessed June 2024).
[15] Apache Storm. Retrieved from https://storm.apache.org/ (accessed June 2024).

[16] Apache Tez. Retrieved from https://tez.apache.org/ (accessed June 2024).
[17] Apache Zookeeper. Retrieved from https://zookeeper.apache.org/ (accessed June 2024).
[18] City of Chicago. Data Portal. Retrieved from https://data.cityofchicago.org/ (accessed June 2024).
[19] Crowdsource by Google. Retrieved from https://crowdsource.google.com/ (accessed June 2024).
[20] Data Portal. Amsterdam. Retrieved from https://data.amsterdam.nl/ (accessed June 2024).
[21] Figure Eight. The Essential High-Quality Data Annotation Platform. Retrieved from https://www.figure-eight.com/ (accessed June 2024).
[22] Find Open Data. Data Portal. Retrieved from https://www.data.gov.uk/ (accessed June 2024).
[23] FIWARE Smart Cities. Retrieved from https://www.fiware.org/about-us/smart-cities/ (accessed June 2024).
[24] GeoMesa. Retrieved from https://www.geomesa.org/ (accessed June 2024).
[25] High-Level Expert Group on Artificial Intelligence. Retrieved from https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence (accessed June 2024).
[26] HPCC Systems. Retrieved from https://hpccsystems.com/ (accessed June 2024).
[27] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Retrieved from https://standardsieee.org/industry-connections/ec/autonomous-systems.html (accessed June 2024).
[28] International Organization for Standardization. Retrieved from https://www.iso.org/home.html (accessed June 2024).
[29] ISO/IEC 21972:2020 Information Technology - Upper Level Ontology for Smart City Indicators. Retrieved from https://www.iso.org/standard/72325.html (accessed June 2024).
[30] ISO/IEC JTC 1/SC 42, Artificial Intelligence. Retrieved from https://www.iso.org/committee/6794475.html (accessed June 2024).
[31] ISO/IEC PRF 5087-1 Information Technology - City Data Model - Part 1: Foundation Level Concepts. Retrieved from https://www.iso.org/standard/80718.html (accessed June 2024).
[32] London Datastore – Greater London Authority. Retrieved from https://data.london.gov.uk/ (accessed June 2024).
[33] MongoDB. Retrieved from https://www.mongodb.com/ (accessed June 2024).
[34] Neo4j Graph Data Platform. Retrieved from https://neo4j.com/ (accessed June 2024).
[35] Open Data Goldbook for Data Managers and Data Holders. Practical Guidebook for Organizations Wanting to Publish Open Data. Retrieved from https://data.europa.eu/sites/default/files/european_data_portal_-_open_data_goldbook.pdf (accessed June 2024).
[36] The Reference Framework for Sustainable Cities. Retrieved from http://rfsc.eu/ (accessed June 2024).
[37] SAREF Semantic Model for Smart Cities. Retrieved from https://saref.etsi.org/saref4city/v1.1.2/ (accessed June 2024).
[38] SpatialHadoop. Retrieved from http://spatialhadoop.cs.umn.edu/ (accessed June 2024).
[39] ST-Hadoop. Retrieved from https://st-hadoop.cs.umn.edu/ (accessed June 2024).
[40] Taming the Data Lake:The HPCC Systems Open Source Big Data Platform. Retrieved from https://cdn.hpccsystems.com/whitepapers/wp_introduction_HPCC.pdf (accessed June 2024).
[41] Tokyo Metropolitan Government Open Data Catalogue (translated from Japanese). Retrieved from https://portal.data.metro.tokyo.lg.jp (accessed June 2024).
[42] Vert.x. Retrieved from https://vertx.io/ (accessed June 2024).
[43] VoltDB. Retrieved from https://www.voltactivedata.com/ (accessed June 2024).
[44] World Council on City Data. Retrieved from https://www.dataforcities.org/ (accessed June 2024).
[45] Zensors: Smart Video Analytics. Retrieved from https://www.zensors.com/ (accessed June 2024).
[46] Uber Blog. 2015. Driving Solutions To Build Smarter Cities. Retrieved October 22, 2019 from https://www.uber.com/blog/boston/driving-solutions-to-build-smarter-cities/
[47] Gartner. 2015. Market Guide for Smart City Operations Management Platforms and Ecosystems. Retrieved November 13, 2019 from https://www.gartner.com/en/documents/3089931/market-guide-for-smart-city-operations-management-platfo
[48] International Standard ISO 37120. 2018. Sustainable Cities and Communities — Indicators for City Services and Quality of Life.
[49] International Standard ISO 37122. 2019. Sustainable Cities and Communities - Indicators for Smart Cities.
[50] New York City. 2019. Automated Decision Systems Task Force Report. Retrieved February 13, 2020 from https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf
[51] Josh Aas, Richard Barnes, Benton Case, Zakir Durumeric, Peter Eckersley, Alan Flores-López, J. Alex Halderman, Jacob Hoffman-Andrews, James Kasten, Eric Rescorla, Seth Schoen, and Brad Warren. 2019. Let's Encrypt: An automated certificate authority to encrypt the entire web. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2473–2487.
[52] Rene Abraham, Johannes Schneider, and Jan Vom Brocke. 2019. Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management* 49 (2019), 424–438.

[53] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)* 51, 4 (2018), 1–35. DOI: https://doi.org/10.1145/3214303
[54] Tanzina Afrin and Nita Yodo. 2022. A long short-term memory-based correlated traffic data prediction framework. *Knowledge-Based Systems* 237 (2022), 107755.
[55] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi. 2015. Applications of big data to smart cities. *Journal of Internet Services and Applications* 6, 1 (2015), 25.
[56] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi. 2015. Applications of big data to smart cities. *Journal of Internet Services and Applications* 6, 1 (Dec. 2015), 25. DOI: https://doi.org/10.1186/s13174-015-0041-5
[57] Vito Albino, Umberto Berardi, and Rosa M. Dangelico. 2015. Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology* 22, 1 (2015), 3–21. DOI: https://doi.org/10.1080/10630732.2014.942092
[58] Muhammad I. Ali, Feng Gao, and Alessandra Mileo. 2015. CityBench: A configurable benchmark to evaluate RSP engines using smart city datasets. In *Proceedings of the 14th International Semantic Web Conference (ISWC '15)*. W3C, Bethlehem, PA, 374–389.
[59] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E. Gilbert. 2022. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law* 30, 1 (Mar. 2022), 1–17. DOI: https://doi.org/10.1007/s10506-021-09286-4
[60] Ahmed Alnuaim, Ziheng Sun, and Didarul Islam. 2023. AI for improving ozone forecasting. In Ziheng Sun, Nicoleta Cristea and Pablo Rivas (Eds.), *Artificial Intelligence in Earth Science*. Elsevier, 247–269.
[61] Project Alvarium. Linux Foundation Edge. Retrieved from https://alvarium.org (accessed June 2024).
[62] N. G. Nageswari Amma and F. Ramesh Dhanaseelan. 2018. Privacy preserving data mining classifier for smart city applications. In *Proceedings of the 3rd International Conference on Communication and Electronics Systems (ICCES '18)*. IEEE, 645–648. DOI: 10.1109/CESYS.2018.8724081
[63] Mike Ananny and Strohecker Carol. 2009. TexTales: Creating interactive forums with urban publics. In *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. Marcus Foth (Ed.), IGI Global, Boston, 68–86. DOI: https://doi.org/10.4018/978-1-60566-152-0.ch005
[64] Margarita Angelidou. 2017. The role of smart city characteristics in the plans of fifteen cities. *Journal of Urban Technology* 24, 4 (2017), 3–28. DOI: https://doi.org/10.1080/10630732.2017.1348880
[65] Raymond Angelo. 2019. Secure protocols and virtual private networks: An evaluation. *Issues in Information Systems* 20, 3 (2019).
[66] Paolo Atzeni, Francesca Bugiotti, and Luca Rossi. 2014. Uniform access to NoSQL systems. *Information Systems* 43 (2014), 117–133.
[67] Gbadebo Ayoade, Vishal Karande, Latifur Khan, and Kevin Hamlen. 2018. Decentralized IoT data management using blockchain and trusted execution environment. In *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI '18)*. IEEE, 15–22.
[68] Saiful Azad and Al-Sakib K. Pathan. 2014. *Practical Cryptography: Algorithms and Implementations Using C++*. CRC Press.
[69] Muhammad Babar, Fahim Arif, Mian A. Jan, Zhiyuan Tan, and Fazlullah Khan. 2019. Urban data management system: Towards big data analytics for internet of things based smart urban environment using customized Hadoop. *Future Generation Computer Systems* 96 (2019), 398–409.
[70] Daniela Ballari, Monica Wachowicz, and Miguel Á. M. Callejo. 2009. Metadata behind the interoperability of wireless sensor network. *Sensors (Basel, Switzerland)* 9 (May 2009), 3635–51. DOI: https://doi.org/10.3390/s90503635
[71] Srividya K. Bansal. 2014. Towards a semantic extract-transform-load (ETL) framework for big data integration. In *Proceedings of the 2014 IEEE International Congress on Big Data*. 522–529. DOI: https://doi.org/10.1109/BigData.Congress.2014.82
[72] Payam Barnaghi, Maria Bermudez-Edo, and Ralf Tönjes. 2015. Challenges for quality of data in smart cities. *Journal of Data and Information Quality (JDIQ)* 6, 2–3 (2015), 1–4.
[73] Sarah Barns. 2018. Smart cities and urban data platforms: Designing interfaces for smart governance. *City, Culture and Society* 12 (2018), 5–12. DOI: https://doi.org/10.1016/j.ccs.2017.09.006.
[74] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)* 41, 3, Article 16 (Jul. 2009), 52 pages. DOI: https://doi.org/10.1145/1541880.1541883
[75] Michael Batty, Kay W. Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, and Yuval Portugali. 2012. Smart cities of the future. *The European Physical Journal Special Topics* 214, 1 (Nov. 2012), 481–518. DOI: https://doi.org/10.1140/epjst/e2012-01703-3
[76] Zohra Bellahsene, Angela Bonifati, and Erhard Rahm. 2011. *Schema Matching and Mapping* (1st ed.). Springer Publishing Company, Incorporated.

[77] Eline A. Belt, Thomas Koch, and Elenna R. Dugundji. 2023. Hourly forecasting of traffic flow rates using spatial temporal graph neural networks. *Procedia Computer Science* 220 (2023), 102–109.

[78] Philip A. Bernstein and Howard Ho. 2007. Model management and schema mappings: Theory and practice. In *Proceedings of the 33rd International Conference on Very Large Data Bases*. University of Vienna, Austria, ACM, 1439–1440.

[79] Devis Bianchini, Valeria De Antonellis, Massimiliano Garda, and Michele Melchiori. 2021. Smart city data modelling using semantic web technologies. In *Proceedings of the IEEE International Smart Cities Conference (ISC2 '21)*. 1–7.

[80] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoit Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29, 2 (Jun. 2021), 149–169. DOI: https://doi.org/10.1007/s10506-020-09270-4

[81] Simon E. Bibri. 2018. The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. *Sustainable Cities and Society* 38 (2018), 230–253. DOI: https://doi.org/10.1016/j.scs.2017.12.034

[82] Simon E. Bibri and John Krogstie. 2017. The core enabling technologies of big data analytics and context-aware computing for smart sustainable cities: A review and synthesis. *Journal of Big Data* 4, 1 (Nov. 2017), 38. DOI: https://doi.org/10.1186/s40537-017-0091-6

[83] Nimish Biloria. 2021. From smart to empathic cities. *Frontiers of Architectural Research* 10, 1 (2021), 3–16. DOI: https://doi.org/10.1016/j.foar.2020.10.001

[84] Stefan Bischof, Athanasios Karapantelakis, Cosmin-Septimiu Nechifor, Amit P. Sheth, Alessandra Mileo, and Payam M. Barnaghi. 2014. Semantic modelling of smart city data. In *W3C Workshop on the Web of Things*.

[85] Kamanashis Biswas and Vallipuram Muthukkumarasamy. 2016. Securing smart cities using blockchain technology. In *Proceedings of the IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS '16)*. IEEE, 1392–1393.

[86] Peter Bosch, Sophie Jongeneel, Hans-Martin Neumann, Iglar Branislav, and Aapo Huovila. 2016. *Recommendations for a Smart City Index*. Project Deliverable, D3.3.

[87] Peter Bosch, Sophie Jongeneel, Vera Rovers, Hans-Martin Neumann, Miimu Airaksinen, and Aapo Huovila. 2017. *CITYkeys Indicators for Smart City Projects and Smart Cities*. Report.

[88] Eric Breck, Marty Zinkevich, Neoklis Polyzotis, Steven Whang, and Sudip Roy. 2019. Data validation for machine learning. In *Proceedings of the SysML*. Retrieved from https://mlsys.org/Conferences/2019/doc/2019/167.pdf

[89] Harry Brignull and Yvonne Rogers. 2003. Enticing People to Interact with Large Public Displays in Public Spaces. In *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction*.

[90] Matthias Budde, Andrea Schankin, Julien Hoffmann, Marcel Danz, Till Riedel, and Michael Beigl. 2017. Participatory sensing or participatory nonsense? Mitigating the effect of human error on data quality in citizen science. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3, Article 39 (Sep. 2017), 23 pages. DOI: https://doi.org/10.1145/3131900

[91] Nadia Burkart and Marco F. Huber. 2021. A Survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (Jan. 2021), 245–317. DOI: https://doi.org/10.1613/jair.1.12228

[92] Nélio Cacho, Frederico Lopes, and Thaís Batista. 2017. Challenges to the development of smart city systems: A system-of-systems view. In *Proceedings of the XXXI Brazilian Symposium on Software Engineering*. 244–249. DOI: https://doi.org/10.1145/3131151.3131189

[93] Hongming Cai, Boyi Xu, Lihong Jiang, and Athanasios V. Vasilakos. 2016. IoT-based big data storage systems in cloud computing: Perspectives and challenges. *IEEE Internet of Things Journal* 4, 1 (2016), 75–87.

[94] Patrici Calvo. 2020. The ethics of smart city (EoSC): Moral implications of hyperconnectivity, algorithmization and the datafication of urban digital society. *Ethics and Information Technology* 22, 2 (Jun. 2020), 141–149. DOI: https://doi.org/10.1007/s10676-019-09523-0

[95] Giuseppe Canonaco, Alex Bergamasco, Alessio Mongelluzzo, and Manuel Roveri. 2021. Adaptive federated learning in presence of concept drift. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN '21)*. 1–7. DOI: https://doi.org/10.1109/IJCNN52387.2021.9533710

[96] Paolo Cardullo and Rob Kitchin. 2019. Smart urbanism and smart citizenship: The neoliberal logic of 'citizen-focused' smart cities in Europe. *Environment and Planning C: Politics and Space* 37, 5 (2019), 813–830. DOI: https://doi.org/10.1177/0263774X18806508

[97] Everton Cavalcante, Nélio Cacho, Frederico Lopes, Thais Batista, and Flavio Oquendo. 2016. Thinking smart cities as systems-of-systems: A perspective study. In *Proceedings of the 2nd International Workshop on Smart (SmartCities '16)*. Association for Computing Machinery, New York, NY, Article 9, 4 pages. DOI: https://doi.org/10.1145/3009912.3009918

[98] Jacqueline G. Cavazos, P. Jonathon Phillips, Carlos D. Castillo, and Alice J. O'Toole. 2021. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 1 (Jan. 2021), 101–111. DOI: https://doi.org/10.1109/TBIOM.2020.3027269

[99] Teresa Cepero, Luis G. Montané-Jiménez, Edgard Benítez-Guerrero, and Carmen Mezura-Godoy. 2022. Visualization in smart city technologies. In *Smart Cities*. Sergio Nesmachnow and Luis Hernández Callejo (Eds.), Springer International Publishing, Cham, 86–100.

[100] Teresa Cepero, Luis G. Montané-Jiménez, and Gina Paola Maestre-Góngora. 2022. Data visualization guide for smart city Technologies. In *Electronic Governance with Emerging Technologies*, Fernando Ortiz-Rodríguez, Sanju Tiwari, Miguel-Angel Sicilia, and Anastasija Nikiforova (Eds.), Springer Nature, Switzerland, Cham, 176–191.

[101] González-Briones Alfonso Rodríguez-Sara Chamoso, Pablo and Juan M. Corchado. 2018. Tendencies of technologies and platforms in smart cities: A state-of-the-art review. *Wireless Communications and Mobile Computing* 2018 (2018), 17. DOI: https://doi.org/10.1155/2018/3086854

[102] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41, 3, Article 15 (Jul. 2009), 58 pages. DOI: https://doi.org/10.1145/1541880.1541882

[103] Qi Chen, Wei Wang, Fangyu Wu, Suparna De, Ruili Wang, Bailing Zhang, and Xin Huang. 2019. A survey on an emerging area: Deep learning for smart city data. *IEEE Transactions on Emerging Topics in Computational Intelligence* 3, 5 (2019), 392–410. DOI: https://doi.org/10.1109/TETCI.2019.2907718

[104] Yang Chen, Arturo Ardila-Gomez, and Gladys Frame. 2017. Achieving energy savings by intelligent transportation systems investments in the context of smart cities. *Transportation Research Part D: Transport and Environment* 54 (2017), 381–396.

[105] Chun Wai Chiu and Leandro L. Minku. 2022. A diversity framework for dealing with multiple types of concept drift based on clustering in the model space. *IEEE Transactions on Neural Networks and Learning Systems* 33, 3 (2022), 1299–1309. DOI: https://doi.org/10.1109/TNNLS.2020.3041684

[106] Sabarathinam Chockalingam, Wolter Pieters, André Teixeira, and Pieter van Gelder. 2017. Bayesian network models in cyber security: A systematic review. In *Secure IT Systems*. Helger Lipmaa, Aikaterini Mitrokotsa, and Raimundas Matulevicius (Eds.), Springer International Publishing, Cham, 105–122.

[107] Chi-Yin Chow and Mohemad F. Mokbel. 2011. *Privacy of Spatial Trajectories*. Springer, New York, NY, 109–141. DOI: https://doi.org/10.1007/978-1-4614-1629-6_4

[108] United 4 Smart Sustainable Cities. 2017. Collection Methodology for Key Performance Indicators for Smart Sustainable Cities. Retrieved from https://u4ssc.itu.int/u4ssc-methodology/ (accessed June 2024).

[109] European Comission. 2020. Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European Strategy for Data. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

[110] European Commission and Directorate-General for Environment. 2018. *Indicators for Sustainable Cities*. Publications Office. DOI: https://doi.org/doi/10.2779/121865

[111] Sergio Consoli, Misael Mongiovic, Andrea G. Nuzzolese, Silvio Peroni, Valentina Presutti, Diego Reforgiato Recupero, and Daria Spampinato. 2015. A smart city data model based on semantics best practice and principles. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. Association for Computing Machinery, New York, NY, 1395–1400.

[112] Carlos Costa and Maribel Yasmina Santos. 2017. The SusCity big data warehousing approach for smart cities. In *Proceedings of the 21st International Database Engineering & Applications Symposium (IDEAS '17)*. Association for Computing Machinery, New York, NY, 264–273.

[113] National Protective Security Authority. 2023. Security-Minded Approach to Developing Connected Places. Retrieved from https://www.npsa.gov.uk/security-minded-approach-developing-connected-places (accessed June 2024).

[114] Federico Cugurullo. 2021. *Frankenstein Urbanism: Eco, Smart and Autonomous Cities, Artificial Intelligence and the End of the City*. Routledge.

[115] Corinne Curt and Jean-Marc Tacnet. 2018. Resilience of critical infrastructures: Review and analysis of current approaches. *Risk Analysis* 38, 11 (2018), 2441–2458. DOI: https://doi.org/10.1111/risa.13166

[116] Thiago Pereira da Silva, Thais Batista, Frederico Lopes, Aluizio R. Neto, Flávia C. Delicato, Paulo F. Pires, and Atslands R. da Rocha. 2022. Fog computing platforms for smart city applications—A survey. *ACM Transactions on Internet Technology* 22, 4 (Feb. 2022), 1–32. DOI: https://doi.org/10.1145/3488585

[117] Mathieu d'Aquin, John Davies, and Enrico Motta. 2015. Smart cities' data: Challenges and opportunities for semantic technologies. *IEEE Internet Computing* 19 (Nov. 2015), 66–70. DOI: https://doi.org/10.1109/MIC.2015.130

[118] Avirup Dasgupta, Asif Gill, and Farookh Hussain. 2019. A conceptual framework for data governance in IoT-enabled digital IS ecosystems. In *Proceedings of the 8th International Conference on Data Science, Technology and Applications*. SCITEPRESS–Science and Technology Publications.

[119] City of New York Data, NYC Open. NYC Open Data. Retrieved from https://opendata.cityofnewyork.us/ (accessed June 2024).

[120] Ayona Datta. 2015. New urban utopias of postcolonial India: 'Entrepreneurial urbanization' in Dholera smart city, Gujarat. *Dialogues in Human Geography* 5, 1 (2015), 3–22. DOI: https://doi.org/10.1177/2043820614565748

[121] Ali Davoudian and Mengchi Liu. 2020. Big data systems: A software engineering perspective. 53, 5 (2020), 1–39. DOI: https://doi.org/10.1145/3408314

[122] Arthur de M. Del Esposte, Eduardo F. Z. Santana, Lucas Kanashiro, Fabio M. Costa, Kelly R. Braghetto, Nelson Lago, and Fabio Kon. 2019. Design and evaluation of a scalable smart city software platform with large-scale simulations. *Future Generation Computer Systems* 93, C (Apr. 2019), 427–441.

[123] Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI '04)*. San Francisco, CA, 137–150.

[124] Aoife Delaney and Rob Kitchin. 2023. Progress and prospects for data-driven coordinated management and emergency response: The case of Ireland. *Territory, Politics, Governance* 11, 1 (2023), 174–189. DOI: https://doi.org/10.1080/21622671.2020.1805355

[125] Yuri Demchenko, Paola Grosso, Cees De Laat, and Peter Membrey. 2013. Addressing big data issues in scientific data infrastructure. In *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13)*. IEEE, 48–55.

[126] Zikun Deng, Di Weng, Shuhan Liu, Yuan Tian, Mingliang Xu, and Yingcai Wu. 2023. A survey of urban visual analytics: Advances and future directions. *Computational Visual Media* 9, 1 (2023), 3–39. DOI: https://doi.org/10.1007/s41095-022-0275-7

[127] Jingtao Ding, Guanghui Yu, Yong Li, Depeng Jin, and Hui Gao. 2020. Learning from hometown and current city: Cross-city POI recommendation via Interest drift and transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4, Article 131 (Sep. 2020), 28 pages. DOI: https://doi.org/10.1145/3369822

[128] Xin Luna Dong and Theodoros Rekatsinas. 2018. Data integration and machine learning: A natural synergy. *Proceedings of the VLDB Endowment* 11, 12 (Aug. 2018), 2094–2097. DOI: https://doi.org/10.14778/3229863.3229876

[129] Nicola Dragoni, Saverio Giallorenzo, Alberto Lluch Lafuente, Manuel Mazzara, Fabrizio Montesi, Ruslan Mustafin, and Larisa Safina. 2017. *Microservices: Yesterday, Today, and Tomorrow*. Springer International Publishing, Cham, 195–216.

[130] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Proceedings of the International Conference on Theory and Applications of Models of Computation*. Springer, 1–19.

[131] Carmen Echebarria, Jose M. Barrutia, and Itziar Aguado-Moralejo. 2021. The smart city journey: A systematic review and future research agenda. *Innovation: The European Journal of Social Science Research* 34, 2 (2021), 159–201. DOI: https://doi.org/10.1080/13511610.2020.1785277

[132] David Eckhoff and Isabel Wagner. 2018. Privacy in the smart city-applications, technologies, challenges, and solutions. *IEEE Communications Surveys & Tutorials* 20, 1 (2018), 489–516. DOI: https://doi.org/10.1109/COMST.2017.2748998

[133] Mohammed El Arass and Nissrine Souissi. 2018. Data lifecycle: From big data to smartdata. In *Proceedings of the IEEE 5th international congress on information science and technology (CiSt '18)*. IEEE, 80–87.

[134] Ahmed Eldawy, Vagelis Hristidis, Saheli Ghosh, Majid Saeedan, Akil Sevim, A. B. Siddique, Samriddhi Singla, Ganesh Sivaram, Tin Vu, and Yaming Zhang. 2021. Beast: Scalable exploratory analytics on spatio-temporal data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. Association for Computing Machinery, New York, NY, 3796–3807. DOI: https://doi.org/10.1145/3459637.3481897

[135] Bibri Simon Elias, Allam Zaheer, and Krogstie John. 2022. The Metaverse as a virtual form of data-driven smart urbanism: Platformization and its underlying processes, institutional dimensions, and disruptive impacts. *Computational Urban Science* 2, 1 (2022), 24. DOI: https://doi.org/10.1007/s43762-022-00051-0

[136] European Union Agency for Network and Information Security (ENISA). 2016. Smart Hospitals: Security and Resilience for Smart Health Service and Infrastructures. DOI: https://data.europa.eu/doi/10.2824/28801

[137] Adriana Eugene, Naomi Alpert, Wil Lieberman-Cribbin, and Emanuela Taioli. 2022. Using NYC 311 call center data to assess short-and long-term needs following Hurricane Sandy. *Disaster Medicine and Public Health Preparedness* 16, 4 (2022), 1447–1451.

[138] Raul Castro Fernandez, Peter R. Pietzuch, Jay Kreps, Neha Narkhede, Jun Rao, Joel Koshy, Dong Lin, Chris Riccomini, and Guozhang Wang. 2015. Liquid: Unifying nearline and offline big data integration. In *Proceedings of the Conference on Innovative Data Systems Research*.

[139] Daniel Fett, Ralf Küsters, and Guido Schmitz. 2016. A comprehensive formal security analysis of OAuth 2.0. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. Association for Computing Machinery, New York, NY, 1204–1215. DOI: https://doi.org/10.1145/2976749.2978385

[140] Riccardo Focardi, Francesco Palmarini, Marco Squarcina, Graham Steel, and Mauro Tempesta. 2018. Mind your keys? A security evaluation of java keystores. In *Proceedings of the Network and Distributed System Security Symposium (NDSS '18)*. 1–15.

[141] European Innovation Partnership for Smart Cities & Communities (EIP-SCC). 2016. EIP-SCC Urban Platform Management Framework, Enabling Cities to Maximize Value From City Data. Retrieved from https://eu-smartcities.eu/sites/default/files/2017-09/EIP_Mgnt_Framework.pdf

[142] The United for Smart Sustainable Cities. 2022. Redefining Smart City Platforms: Setting the Stage for Minimal Interoperability Mechanisms. A U4SSC Deliverable on City Platforms. Retrieved from https://www.itu.int/en/publications/Documents/tsb/2022-U4SSC-Redefining-smart-cityplatforms/index.html#p=1

[143] Johannes Franke and Peter Gailhofer. 2021. Data governance and regulation for sustainable smart cities. *Frontiers in Sustainable Cities* 3 (2021), 148.

[144] Ulrik Franke and Joel Brynielsson. 2014. Cyber situational awareness—A systematic review of the literature. *Computers & Security* 46 (2014), 18–31. DOI: https://doi.org/10.1016/j.cose.2014.06.008

[145] João Gama, Indrundefined Zliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* 4, Article 44 (Mar. 2014), 37 pages. DOI: https://doi.org/10.1145/2523813

[146] Apurva Gandhi, Yuki Asada, Victor Fu, Advitya Gemawat, Lihao Zhang, Rathijit Sen, Carlo Curino, Jesús Camacho-Rodríguez, and Matteo Interlandi. 2023. The tensor data platform: Towards an AI-centric database system. In *Proceedings of the 23th Conference on Innovative Data Systems Research (CIDR '23)*. Retrieved from www.cidrdb.org.

[147] Salvador García, Julián Luengo, and Francisco Herrera. 2015. *Data Preprocessing in Data Mining*. Springer International Publishing, Cham. DOI: https://doi.org/10.1007/978-3-319-10247-4

[148] Aditya Gaur, Bryan Scotney, Gerard Parr, and Sally McClean. 2015. Smart city architecture and its applications based on IoT. *Procedia Computer Science* 52 (2015), 1089–1094.

[149] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer W. Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. https://arxiv.org/abs/1803.09010

[150] Ammar Gharaibeh, Mohammad A. Salahuddin, Sayed Jahed Hussini, Abdallah Khreishah, Issa Khalil, Mohsen Guizani, and Ala Al-Fuqaha. 2017. Smart cities: A survey on data management, security, and enabling technologies. *IEEE Communications Surveys & Tutorials* 19, 4 (2017), 2456–2501. DOI: https://doi.org/10.1109/COMST.2017.2736886

[151] Rudolf Giffinger, Christian Fertner, Hans Kramar, Robert Kalasek, Natasa Pichler-Milanović, and Evert Meijers. Smart Cities: Ranking of European Medium-Sized Cities. Retrieved from http://www.smart-cities.eu/download/smart_cities_final_report.pdf (accessed June 2024).

[152] Ekaterina Gilman, Anja Keskinarkaus, Satu Tamminen, Susanna Pirttikangas, Juha Röning, and Jukka Riekki. 2015. Personalised assistance for fuel-efficient driving. *Transportation Research Part C: Emerging Technologies* 58 (2015), 681–705. DOI: https://doi.org/10.1016/j.trc.2015.02.007

[153] Ekaterina Gilman, Satu Tamminen, Anja Keskinarkaus, Theodoros Anagnostopoulos, Xiang Su, Susanna Pirttikangas, and Jukka Riekki. 2020. Fuel consumption analysis of driven trips with respect to route choice. In *Proceedings of the IEEE 36th International Conference on Data Engineering Workshops (ICDEW '20)*. 40–47. DOI: https://doi.org/10.1109/ICDEW49219.2020.000-9

[154] Ellen P. Goodman. 2020. Smart city ethics: How "smart" challenges democratic governance. In Markus D. Dubber, Frank Pasquale, and Sunit Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780190067397.013.53

[155] Salvatore Greco, Alessio Ishizaka, Menelaos Tasiou, and Gianpiero Torrisi. 2019. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research* 141, 1 (Jan. 2019), 61–94. DOI: https://doi.org/10.1007/s11205-017-1832-9

[156] Gerhard Gröger, Thomas H. Kolbe, Claus Nagel, and Karl-Heinz Häfele. 2012. OGC City Geography Markup Language (CityGML) Encoding Standard. OGC Standard OGC 12-019 Open Geospatial Consortium. Retrieved from http://www.opengeospatial.org/standards/is 35.01.01; LK 01.

[157] Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 2 (1993), 199–220.

[158] Ralf Hartmut Güting and Markus Schneider. 2005. *Moving Object Databases*. Morgan Kaufmann Publishers.

[159] Hadi Habibzadeh, Cem Kaptan, Tolga Soyata, Burak Kantarci, and Azzedine Boukerche. 2019. Smart city system design: A comprehensive study of the application and data planes. *ACM Computing Surveys (CSUR)* 52, 2, Article 41 (May 2019), 38 pages. DOI: https://doi.org/10.1145/3309545

[160] Hadi Habibzadeh, Tolga Soyata, Burak Kantarci, Azzedine Boukerche, and Cem Kaptan. 2018. Sensing, communication and security planes: A new challenge for a smart city system design. *Computer Networks* 144 (2018), 163–200. DOI: https://doi.org/10.1016/j.comnet.2018.08.001

[161] Jean-Luc Hainaut. 2006. The transformational approach to database engineering. In *Proceedings of the International Summer School on Generative and Transformational Techniques in Software Engineering (GTTSE '06)*. LNCS, Vol. 4143. Springer, 95–143.

[162] Dame Wendy Hall and Jérome Pesenti. 2017. Growing the Artificial Intelligence Industry in the UK. Retrieved from https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

[163] Robert E. Hall, B. Bowerman, J. Braverman, J. Taylor, H. Todosow, and U. Von Wimmersperg. The Vision of a Smart City. Retrieved from https://www.osti.gov/biblio/773961 (accessed June 2024).
[164] Sophia Hamer, Jennifer Sleeman, and Ivanka Stajner. 2023. Forecast-aware model driven LSTM. https://arxiv.org/abs/2303.12963
[165] Jack Hardinges. 2018. What is a Data Trust? Retrieved November 5, 2019 from https://theodi.org/article/what-is-a-data-trust/
[166] Jack Hardinges and Peter Wells. 2018. Defining a 'Data Trust'. Retrieved November 5, 2019 from https://theodi.org/article/defining-a-data-trust/
[167] Colin Harrison, Barbara Eckman, Rick Hamilton, Perry Hartswick, Jayant Kalagnanam, Jurij Paraszczak, and Peter Williams. 2010. Foundations for smarter cities. *IBM Journal of Research and Development* 54, 4 (2010), 1–16. DOI: https://doi.org/10.1147/JRD.2010.2048257
[168] Guy Harrison. 2016. *Next Generation Databases: NoSQL, NewSQL, and Big Data*. Apress.
[169] Ibrahim A. Targio Hashem, Victor Chang, Nor B. Anuar, Kayode Adewole, Ibrar Yaqoob, Abdullah Gani, Ejaz Ahmed, and Haruna Chiroma. 2016. The role of big data in smart city. *International Journal of Information Management* 36, 5 (2016), 748–758. DOI: https://doi.org/10.1016/j.ijinfomgt.2016.05.002
[170] Tali Hatuka, Toch Eran, Birnhack Michael, and Hadas Zur. 2020. The Digital City: Critical Dimensions in Implementing the Smart City. Available at SSRN. DOI: https://dx.doi.org/10.2139/ssrn.3766782
[171] Wei He, Wanqiang Li, and Peidong Deng. 2022. Legal governance in the smart cities of China: Functions, problems, and solutions. *Sustainability* 14, 15 (2022), 9738.
[172] Arne Hintz, Lina Dencik, and Karin Wahl-Jorgensen. 2017. Digital citizenship and surveillance—Digital citizenship and surveillance society—Introduction. *International Journal of Communication* 11, 0 (2017), 731–739. DOI: https://ijoc.org/index.php/ijoc/article/view/5521
[173] Taisei Hirakawa, Keisuke Maeda, Takahiro Ogawa, Satoshi Asamizu, and Miki Haseyama. 2021. Analysis of social trends related to COVID-19 pandemic utilizing social media data. In *Proceedings of the IEEE 10th Global Conference on Consumer Electronics (GCCE '21)*. IEEE, 43–44.
[174] Robert G. Hollands. 2014. Critical interventions into the corporate smart city. *Cambridge Journal of Regions, Economy and Society* 8, 1 (Aug. 2014), 61–77. DOI: https://doi.org/10.1093/cjres/rsu011
[175] Ali Reza Honarvar and Ashkan Sami. 2019. Towards sustainable smart city by particulate matter prediction using urban big data, excluding expensive air pollution infrastructures. *Big Data Research* 17 (2019), 56–65.
[176] Simo Hosio, Vassilis Kostakos, Hannu Kukka, Marko Jurmu, Jukka Riekki, and Timo Ojala. 2012. From school food to skate parks in a few clicks: Using public displays to bootstrap civic engagement of the young. In *Pervasive Computing*. Judy Kay, Paul Lukowicz, Hideyuki Tokuda, Patrick Olivier, and Antonio Krüger (Eds.), Springer, Berlin, 425–442.
[177] Chenyu Hou, Bin Cao, Sijie Ruan, and Jing Fan. 2021. TLDS: A transfer-learning-based delivery station location selection pipeline. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 4, Article 50 (Aug. 2021), 24 pages. DOI: https://doi.org/10.1145/3469084
[178] John Hughes and Eve Maler. 2005. Security assertion markup language (SAML) v2. 0 technical overview. *OASIS SSTC Working Draft sstc-saml-tech-overview-2.0-draft-08* 13 (2005), 12.
[179] Aapo Huovila, Peter Bosch, and Miimu Airaksinen. 2019. Comparative analysis of standardized indicators for Smart sustainable cities: What indicators and standards to use and when? *Cities* 89 (2019), 141–153. DOI: https://doi.org/10.1016/j.cities.2019.01.029
[180] Sergio Ilarri, Eduardo Mena, and Arantza Illarramendi. 2010. Location-dependent query processing: Where we are and where we are heading. *ACM Computing Surveys (CSUR)* 42, 3, Article 12 (Mar. 2010), 73 pages. DOI: https://doi.org/10.1145/1670679.1670682
[181] European Telecommunication Standards Institute. 2017. ETSI TS103463 Key Performance Indicators for Sustainable Digital Multiservice Cities. Technical Specification V1.1.1 (July 2017). Retrieved September 30, 2019 from https://www.etsi.org/deliver/etsi_ts/103400_103499/103463/01.01.01_60/ts_103463v010101p.pdf
[182] European Telecommunication Standards Institute. 2019. Context Information Management (CIM); Information Model (MOD0), ETSI GS CIM 006 V1.1.1 (July 2019), Group specification. Retrieved from https://www.etsi.org/deliver/etsi_gs/CIM/001_099/006/01.01.01_60/gs_CIM006v010101p.pdf
[183] European Telecommunication Standards Institute. 2022. Context Information Management (CIM);NGSI-LD; Guidelines for the Deployment of Smart City and Communities Data Platforms. ETSI GR CIM 020 V1.1.1 (December 2022), Group Report. Retrieved from https://www.etsi.org/deliver/etsi_gr/CIM/001_099/020/01.01.01_60/gr_CIM020v010101p.pdf
[184] European Telecommunication Standards Institute. 2022. Cross-cutting Context Information Management (CIM); NGSI-LD API, ETSI GS CIM 009 V1.6.1 (August 2022), Group specification. Retrieved from https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.06.01_60/gs_CIM009v010601p.pdf

[185] Open Data Institute. Mapping the Wide World of Data Sharing. Retrieved from https://theodi.org/project/the-data-access-map/ (accessed June 2024).
[186] Sohely Jahan, Md S. Rahman, and Sajeeb Saha. 2017. Application specific tunneling protocol selection for virtual private networks. In *Proceedings of the International Conference on Networking, Systems and Security (NSYSS '17)*. IEEE, 39–44.
[187] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly* 37, 3 (2020), 101493.
[188] Huaxiong Jiang, Stan Geertman, and Patrick Witte. 2023. The contextualization of smart city technologies: An international comparison. *Journal of Urban Management* 12, 1 (2023), 33–43.
[189] Renhe Jiang, Xuan Song, Zipei Fan, Tianqi Xia, Zhaonan Wang, Quanjun Chen, Zekun Cai, and Ryosuke Shibasaki. 2021. Transfer urban human mobility via POI embedding over multiple cities. *ACM/IMS Transactions on Data Science* 2, 1, Article 4 (Jan. 2021), 26 pages. DOI: https://doi.org/10.1145/3416914
[190] Fengmei Jin, Wen Hua, Matteo Francia, Pingfu Chao, Maria E. Orlowska, and Xiaofang Zhou. 2023. A survey and experimental study on privacy-preserving trajectory data publishing. *IEEE Transactions on Knowledge and Data Engineering* 35, 6 (2023), 5577–5596. DOI: https://doi.org/10.1109/TKDE.2022.3174204
[191] Yilun Jin, Kai Chen, and Qiang Yang. 2022. Selective cross-city transfer learning for traffic prediction via source city region re-weighting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. Association for Computing Machinery, New York, NY, 731–741. DOI: https://doi.org/10.1145/3534678.3539250
[192] Thaddeus L. Johnson, Natasha N. Johnson, Denise McCurdy, and Michael S. Olajide. 2022. Facial recognition systems in policing and racial disparities in arrests. *Government Information Quarterly* (Aug. 2022), 101753. DOI: https://doi.org/10.1016/j.giq.2022.101753
[193] Kyung H. Jung, Zachary Pitkowsky, Kira Argenio, James W. Quinn, Jean-Marie Bruzzese, Rachel L. Miller, Steven N. Chillrud, Matthew Perzanowski, Jeanette A. Stingone, and Stephanie Lovinsky-Desir. 2022. The effects of the historical practice of residential redlining in the United States on recent temporal trends of air pollution near New York City schools. *Environment International* 169 (2022), 107551.
[194] Maxim Kalinin, Vasiliy Krundyshev, and Peter Zegzhda. 2021. Cybersecurity risk assessment in smart city infrastructures. *Machines* 9, 4 (2021), 78. DOI: https://doi.org/10.3390/machines9040078
[195] Nektaria Kaloudi and Jingyue Li. 2020. The AI-based cyber threat landscape: A survey. *ACM Computing Surveys* 53, 1, Article 20 (Feb. 2020), 34 pages. DOI: https://doi.org/10.1145/3372823
[196] Ahmed Khalid and Hassan Mehmood. 2020. Guidelines and Recommendations for Data Privacy in Public Administration. CUTLER D3.4. Retrieved from https://www.cutler-h2020.eu/download/1798
[197] Latif U. Khan, Ibrar Yaqoob, Nguyen H. Tran, S. M. Ahsan Kazmi, Tri N. Dang, and Choong Seon Hong. 2020. Edge-computing-enabled smart cities: A comprehensive survey. *IEEE Internet of Things Journal* 7, 10 (2020), 10200–10232. DOI: https://doi.org/10.1109/JIOT.2020.2987070
[198] Vijay Khatri and Carol V. Brown. 2010. Designing data governance. *Communications of the ACM* 53, 1 (2010), 148–152.
[199] Steven L. Kinney. 2006. *Trusted Platform Module Basics: Using TPM in Embedded Systems*. Elsevier.
[200] Rob Kitchin. 2014. The real-time city? Big data and smart urbanism. *GeoJournal* 79, 1 (Feb. 2014), 1–14. DOI: https://doi.org/10.1007/s10708-013-9516-8
[201] Rob Kitchin, Tracey P. Lauriault, and Gavin McArdle. 2015. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science* 2, 1 (2015), 6–28. DOI: https://doi.org/10.1080/21681376.2014.983149
[202] Rob Kitchin and Niamh Moore-Cherry. 2021. Fragmented governance, the urban data ecosystem and smart city-regions: The case of Metropolitan Boston. *Regional Studies* 55, 12 (2021), 1913–1923. DOI: https://doi.org/10.1080/00343404.2020.1735627
[203] Martin Kleppmann. 2017. *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. O'Reilly Media.
[204] Bram Klievink, Haiko van der Voort, and Wijnand Veeneman. 2018. Creating value through data collaboratives. *Information Polity* 23, 4 (2018), 379–397. DOI: https://doi.org/10.3233/IP-180070
[205] 205] Ansgar Koene, Chris Clifton, Yohko Hatada, Helena Webb, Menisha Patel, Caio Machado, Jack LaViolette, Rashida Richardson, and Dillon Reisman. 2019. A Governance Framework for Algorithmic Accountability and Transparency. https://data.europa.eu/doi/10.2861/59990
[206] Ansgar Koene, Liz Dowthwaite, and Suchana Seth. 2018. IEEE P7003TM standard for algorithmic bias considerations. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness (FairWare '18)*. 38–41. DOI: https://doi.org/10.23919/FAIRWARE.2018.8452919
[207] Jason Koh, Sandeep Sandha, Bharathan Balaji, Daniel Crawl, Ilkay Altintas, Rajesh E. Gupta, and Mani B. Srivastava. 2017. Data hub architecture for smart cities. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems (SenSys '17)*. Delft, Netherlands, 77:1–77:2.

[208] Andreas Komninos, Jeries Besharat, Denzil Ferreira, John Garofalakis, and Vassilis Kostakos. 2017. Where's everybody? Comparing the use of heatmaps to uncover cities' tacit social context in smartphones and pervasive displays. *Information Technology & Tourism* 17 (2017), 399–427.

[209] Pascal D. König. 2021. Citizen-centered data governance in the smart city: From ethics to accountability. *Sustainable Cities and Society* 75 (Dec. 2021), 103308. DOI: https://doi.org/10.1016/j.scs.2021.103308

[210] Panos Kostakos, Hassan Mehmood, Marta Cortés, Thuy Truong, and Eleni Ntzioni. 2020. Update of the Final Version of the Data Collection, Management & Protection Framework Integrated within CUTLER Architecture. CUTLER D3.5. Retrieved from https://www.cutler-h2020.eu/download/1184

[211] Vassilis Kostakos, Jakob Rogstadius, Denzil Ferreira, Simo Hosio, and Jorge Goncalves. 2017. *Human Sensors*. Springer International Publishing, 69–92. DOI: https://doi.org/10.1007/978-3-319-25658-0_4

[212] Jay Kreps. 2014. Questioning the Lambda Architecture. Retrieved from https://www.oreilly.com/radar/questioning-the-lambda-architecture/

[213] Markku Kulmala, Tom V. Kokkonen, Juha Pekkanen, Sami Paatero, Tuukka Petäjä, Veli-Matti Kerminen, and Aijun Ding. 2021. Opinion: Gigacity—A source of problems or the new way to sustainable development. *Atmospheric Chemistry and Physics* 21, 10 (2021), 8313–8322. DOI: https://doi.org/10.5194/acp-21-8313-2021

[214] Sidewalk Labs. 2019. The Digital Innovation Plan. Retrieved from https://storage.googleapis.com/sidewalk-toronto-ca/wp-content/uploads/2019/06/23135715/MIDP_Volume2.pdf

[215] Viivi Lähteenoja and Sille Sepp. 2021. State of MyData 2021. Retrieved from https://www.mydata.org/publication/state-of-mydata-2021/

[216] Chun S. Lai, Youwei Jia, Zhekang Dong, Dongxiao Wang, Yingshan Tao, Qi H. Lai, Richard T. K. Wong, Ahmed F. Zobaa, Ruiheng Wu, and Loi Lei Lai. 2020. A review of technical standards for smart cities. *Clean Technologies* 2, 3 (2020), 290–310. DOI: https://doi.org/10.3390/cleantechnol2030019

[217] Jeff LeFevre, Jagan Sankaranarayanan, Hakan Hacigümüs, Junichi Tatemura, Neoklis Polyzotis, and Michael J. Carey. 2014. MISO: Souping up big data query processing with a multistore system. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. 1591–1692.

[218] Barry Leiba. 2012. Oauth web authorization protocol. *IEEE Internet Computing* 16, 1 (2012), 74–77.

[219] Antoine Lemay, Joan Calvet, Francois Menet, and José M. Fernandez. 2018. Survey of publicly available reports on advanced persistent threat actors. *Computers & Security* 72 (2018), 26–59. DOI: https://doi.org/10.1016/j.cose.2017.08.005

[220] Cédric Lévy-Bencheton, Eleni Darra, Daniel Bachlechner, and Michael Friedewald. 2015. *Cyber Security for Smart Cities—An Architecture Model for Public Transport*. The European Union Agency for Network and Information Security (ENISA), Technical Report.

[221] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2006. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 106–115.

[222] Chiehyeon Lim, Kwang-Jae Kim, and Paul P. Maglio. 2018. Smart cities with big data: Reference models, challenges, and considerations. *Cities* 82 (2018), 86–99. DOI: https://doi.org/10.1016/j.cities.2018.04.011

[223] Yan Liu, Bin Guo, Daqing Zhang, Djamal Zeghlache, Jingmin Chen, Sizhe Zhang, Dan Zhou, Xinlei Shi, and Zhiwen Yu. 2021. MetaStore: A task-adaptative meta-learning model for optimal store placement with multi-city knowledge transfer. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 3, Article 28 (apr 2021), 23 pages. DOI: https://doi.org/10.1145/3447271

[224] Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. 2020. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In *Machine Learning and Knowledge Extraction*. Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl (Eds.), Springer International Publishing, Cham, 1–16.

[225] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. 2019. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2019), 2346–2363. DOI: https://doi.org/10.1109/TKDE.2018.2876857

[226] Anna Luusua, Johanna Ylipulli, and Emilia Rönkkö. 2017. Nonanthropocentric design and smart cities in the anthropocene. *IT-Information Technology* 59, 6 (2017), 295–304. DOI: https://doi.org/doi:10.1515/itit-2017-0007

[227] Chen Ma. 2021. Smart city and cyber-security; technologies used, leading challenges and future recommendations. *Energy Reports* 7 (2021), 7999–8012. DOI: https://doi.org/10.1016/j.egyr.2021.08.124

[228] Meiyi Ma, Sarah M. Preum, Mohsin Y. Ahmed, William Tärneberg, Abdeltawab Hendawi, and John A. Stankovic. 2019. Data sets, modeling, and decision making in smart cities: A survey. *ACM Transactions on Cyber-Physical Systems* 4, 2, Article 14 (Nov. 2019), 28 pages. DOI: https://doi.org/10.1145/3355283

[229] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3–es.

[230] Hasindu Madushan, Iftekhar Salam, and Janaka Alawatugoda. 2022. A review of the NIST lightweight cryptography finalists and their fault analyses. *Electronics* 11, 24 (2022), 4199.

[231] Martino Maggio, Francesco Arigliano, Ömer Özdemir, José Manuel Cantera, Eunah Kim, Ignacio Elicegui Maestro, Andrea Gaglione, and Angelo Capossele. 2018. Reference Architecture for IoT Enabled Smart Cities, Update. Retrieved from https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bd4a731f & appId=PPGMS

[232] P. Makkaroon, D. Q. Tong, Y. Li, E. J. Hyer, P. Xian, S. Kondragunta, P. C. Campbell, Y. Tang, B. D. Baker, M. D. Cohen, A. Darmenov, A. Lyapustin, R. D. Saylor, Y. Wang, and I. Stajner 2023. Development and evaluation of a North America ensemble wildfire air quality forecast: Initial application to the 2020 Western United States "Gigafire". *Journal of Geophysical Research: Atmospheres* 128, 22 (2023). DOI: https://doi.org/10.1029/2022JD037298

[233] Hug March and Ramon Ribera-Fumaz. 2016. Smart contradictions: The politics of making Barcelona a self-sufficient city. *European Urban and Regional Studies* 23, 4 (2016), 816–830. DOI: https://doi.org/10.1177/0969776414554488

[234] Chiara Marcolla, Victor Sucasas, Marc Manzano, Riccardo Bassoli, Frank H. P. Fitzek, and Najwa Aaraj. 2022. Survey on fully homomorphic encryption, theory, and applications. *Proceedings of the IEEE* 110, 10 (2022), 1572–1609.

[235] Nathan Marz and James Warren. 2015. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems* (1st ed.). Manning Publications Co.

[236] Audrey L. Mayer. 2008. Strengths and weaknesses of common sustainability indices for multidimensional systems. *Environment International* 34, 2 (2008), 277–291. DOI: https://doi.org/10.1016/j.envint.2007.09.004

[237] Peter McBrien and Alexandra Poulovassilis. 1999. A uniform approach to inter-model transformations. In *Proceedings of the 11th International Conference on Advanced Information Systems Engineering (CAiSE '99)*. LNCS, Vol. 1626. 333–348.

[238] Colin McFarlane and Ola Söderström. 2017. On alternative smart cities. *City* 21, 3–4 (2017), 312–328. DOI: https://doi.org/10.1080/13604813.2017.1327166

[239] Hassan Mehmood, Ekaterina Gilman, Marta Cortes, Panos Kostakos, Andrew Byrne, Katerina Valta, Stavros Tekes, and Jukka Riekki. 2019. Implementing big data lake for heterogeneous data sources. In *Proceedings of the IEEE 35th International Conference on Data Engineering Workshops (ICDEW '19)*. IEEE, 37–44.

[240] Hassan Mehmood, Panos Kostakos, Marta Cortes, Theodoros Anagnostopoulos, Susanna Pirttikangas, and Ekaterina Gilman. 2021. Concept drift adaptation techniques in distributed environment for real-world data streams. *Smart Cities* 4, 1 (2021), 349–371. DOI: https://doi.org/10.3390/smartcities4010021

[241] A. Middleton and P. D. L. R. Solutions. 2011. *HPCC Systems: Introduction to HPCC (High-Performance Computing Cluster)*. White Paper, LexisNexis Risk Solutions.

[242] MITRE 2022. *Groups/MITRE ATT & CK®*. Retrieved Sep 27, 2022 from https://attack.mitre.org/groups/

[243] Luca Mora, Roberto Bolici, and Mark Deakin. 2017. The first two decades of smart-city research: A bibliometric analysis. *Journal of Urban Technology* 24, 1 (2017), 3–27. DOI: https://doi.org/10.1080/10630732.2017.1285123

[244] Vaia Moustaka, Athena Vakali, and Leonidas G. Anthopoulos. 2018. A systematic review for smart city data analytics. *ACM Computing Surveys (cSuR)* 51, 5 (2018), 1–41. DOI: https://doi.org/10.1145/3239566

[245] Georgios Mylonas, Athanasios Kalogeras, Georgios Kalogeras, Christos Anagnostopoulos, Christos Alexakos, and Luis Muñoz. 2021. Digital twins from smart manufacturing to smart cities: A survey. *IEEE Access* 9 (2021), 143222–143249. DOI: https://doi.org/10.1109/ACCESS.2021.3120843

[246] Taewoo Nam and Theresa A. Pardo. 2011. Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (dg.o '11)*. ACM, New York, NY, 282–291. DOI: https://doi.org/10.1145/2037556.2037602

[247] United Nations. 2015. World Urbanisation Prospects. The 2014 Revision. Retrieved June 16, 2019 from https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Report.pdf

[248] Paolo Neirotti, Alberto De Marco, Anna Corinna Cagliano, Giulio Mangano, and Francesco Scorrano. 2014. Current trends in smart city initiatives: Some stylised facts. *Cities* 38 (2014), 25–36. DOI: https://doi.org/10.1016/j.cities.2013.12.010

[249] Jan K. Nidzwetzki and Ralf H. Güting. 2015. Distributed SECONDO: A highly available and scalable system for spatial data processing. In *Advances in Spatial and Temporal Databases*. Christophe Claramunt, Markus Schneider, Raymond Chi-Wing Wong, Li Xiong, Woong-Kee Loh, Cyrus Shahabi, and Ki-Joune Li (Eds.), Springer International Publishing, Cham, 491–496.

[250] OECD/European Union/EC-JRC. 2008. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing, Paris, https://doi.org/10.1787/9789264043466-en

[251] Municipality of Copenhagen and Capital Region of Denmark. 2018. City Data Exchange - Lessons Learned From a Public/private Data Collaboration.

[252] Dietmar Offenhuber. 2014. Infrastructure legibility—A comparative analysis of open311-based citizen feedback systems. *Cambridge Journal of Regions, Economy and Society* 8, 1 (Mar. 2014), 93–112. DOI: https://doi.org/10.1093/cjres/rsu001

[253] Kieron O'Hara. 2019. *Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship*. Web Science Institute White Papers.

[254] Tomoya Ohyama, Kazunori Hanyu, Masayuki Tani, and Momoka Nakae. 2022. Investigating crime harm index in the low and downward crime contexts: A spatio-temporal analysis of the Japanese crime harm index. *Cities* 130 (2022), 103922.

[255] Frederik Olsen, Calogero Schillaci, Mohamed Ibrahim, and Aldo Lipani. 2022. Borough-level COVID-19 forecasting in London using deep learning techniques and a novel MSE-Moran's I loss function. *Results in Physics* 35 (2022), 105374.

[256] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13. DOI: https://doi.org/10.3389/fdata.2019.00013

[257] Francis Ostermeijer, Hans Koster, Leonardo Nunes, and Jos van Ommeren. 2022. Citywide parking policy and traffic: Evidence from Amsterdam. *Journal of Urban Economics* 128 (2022), 103418.

[258] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. 2018. Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences* 30, 4 (2018), 431–448.

[259] Gang Pan, Guande Qi, Wangsheng Zhang, Shijian Li, Zhaohui Wu, and Laurence Tianruo Yang. 2013. Trace analysis and mining for smart cities: Issues, methods, and applications. *IEEE Communications Magazine* 51, 6 (2013), 120–126. DOI: https://doi.org/10.1109/MCOM.2013.6525604

[260] Sinno J. Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (Oct. 2010), 1345–1359. DOI: https://doi.org/10.1109/TKDE.2009.191

[261] Piyush Pandey, Sean Peasley, Deborah Golden, and Mahesh Kelkar. 2019. Making smart cities cybersecure: Ways to address distinct risks in an increasingly connected urban future. Deloitte insights. Retrieved from https://www2.deloitte.com/us/en/insights/focus/smart-city/making-smart-cities-cyber-secure.html (accessed June 2024).

[262] Laura P. Prieto. 2022. *Introducing Object Storage in Hadoop Ecosystem*. Technical Report.

[263] Raj R. Parmar, Sudipta Roy, Debnath Bhattacharyya, Samir K. Bandyopadhyay, and Tai-Hoon Kim. 2017. Large-scale encryption in the Hadoop environment: Challenges and solutions. *IEEE Access* 5 (2017), 7156–7163.

[264] Krassimira Paskaleva, James Evans, Christopher Martin, Trond Linjordet, Dujuan Yang, and Andrew Karvonen. 2017. Data governance in the sustainable smart city. *Informatics* 4, 4 (2017), 41. DOI: https://doi.org/10.3390/informatics4040041

[265] Ramesh Paudel and William Eberle. 2020. An approach for concept drift detection in a graph stream using discriminative subgraphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14, 6, Article 70 (Sep. 2020), 25 pages. DOI: https://doi.org/10.1145/3406243

[266] Eric Paulos, Ian Smith, and R. Honicky. 2008. Participatory Urbanism. Urbanatmospheres. net (Accessed May 18, 2010).

[267] Michael E. Payne, Linh B. Ngo, Flavio Villanustre, and Amy W. Apon. 2014. Managing the academic data lifecycle: A case study of HPCC. In *Proceedings of the IEEE International Conference on Big Data (Big Data '14)*. IEEE, 22–30.

[268] Jorge Pereira, Thais Batista, Everton Cavalcante, Arthur Souza, Frederico Lopes, and Nelio Cacho. 2022. A platform for integrating heterogeneous data and developing smart city applications. *Future Generation Computer Systems* 128 (2022), 552–566. DOI: https://doi.org/10.1016/j.future.2021.10.030

[269] Ricardo L. Pereira, Pedro C. Sousa, Ricardo Barata, André Oliveira, and Geert Monsieur. 2015. CitySDK tourism API-building value around open data. *Journal of Internet Services and Applications* 6, 1 (2015), 24:1–24:13. DOI: https://doi.org/10.1186/s13174-015-0039-z

[270] Charith Perera, Yongrui Qin, Julio C. Estrella, Stephan Reiff-Marganiec, and Athanasios V. Vasilakos. 2017. Fog computing for sustainable smart cities: A survey. *ACM Computing Surveys (CSUR)* 50, 3, Article 32 (Jun. 2017), 43 pages. DOI: https://doi.org/10.1145/3057266

[271] Riccardo Petrolo, Valeria Loscrí, and Nathalie Mitton. 2014. Towards a smart city based on cloud of things. In *Proceedings of the ACM International Workshop on Wireless and Mobile Technologies for Smart Cities (WiMobCity '14)*. Association for Computing Machinery, New York, NY, 61–66. DOI: https://doi.org/10.1145/2633661.2633667

[272] Judicaël Picaut, Nicolas Fortin, Erwan Bocher, Gwendall Petit, Pierre Aumond, and Gwenaël Guillaume. 2019. An open-science crowdsourcing approach for producing community noise maps using smartphones. *Building and Environment* 148 (2019), 20–33. DOI: https://doi.org/10.1016/j.buildenv.2018.10.049

[273] Antti Poikola, Daniel Kaplan, and Tanel Mällo. 2017. Declaration of MyData Principles. Retrieved from https://www.mydata.org/participate/declaration/

[274] CITYkeys EU Horizon 2020 Project. Retrieved from http://http//www.citykeys-project.eu (accessed June 2024).

[275] Achilleas Psyllidis. 2015. Ontology-Based Data Integration from Heterogeneous Urban Systems: A Knowledge Representation Framework for Smart Cities. In *14th International Conference on Computers in Urban Planning and Urban Management* (CUPUM2015).

[276] Achilleas Psyllidis, Alessandro Bozzon, Stefano Bocconi, and Christiaan Bolivar. 2015. A Platform for Urban Analytics and Semantic Data Integration in City Planning. In: Celani, G., Sperling, D., Franco, J. (eds) Computer-Aided Architectural Design Futures. The Next City - New Technologies and the Future of the Built Environment. CAAD Futures 2015. Communications in Computer and Information Science, vol 527. Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-662-47386-3_2

[277] NIST FIPS Pub. 2001. 197: Advanced encryption standard (AES). *Federal Information Processing Standards Publication* 197, 441 (2001), 0311.

[278] Dan Puiu, Payam Barnaghi, Ralf Tönjes, Daniel Kumper, Muhammad I. Ali, Alessandra Mileo, Josiane Parreira, Marten Fischer, Sefki Kolozali, Nazli Farajidavar, Feng Gao, Thorben Iggena, Thu-Le Pham, Cosmin-Septimiu Nechifor, Daniel Puschmann, and Joao Fernandes. 2016. CityPulse: Large scale data analytics framework for smart cities. *IEEE Access* 4 (Jan. 2016), 1086–1108. DOI: https://doi.org/10.1109/ACCESS.2016.2541999

[279] Dan Puiu, Payam M. Barnaghi, Ralf R Tönjes, Daniel Kümper, Muhammad I. Ali, Alessandra Mileo, Josiane X. Parreira, Marten Fischer, Sefki Kolozali, Nazli Farajidavar, Feng Gao, Thorben Iggena, Thu-Le Pham, Cosmin-Septimiu Nechifor, Daniel Puschmann, and João Fernandes. 2016a. CityPulse: Large scale data analytics framework for smart cities. *IEEE Access* 4 (2016), 1086–1108.

[280] Subashini Raghavan, Boung-Yew Lau Simon, Ying Loong Lee, Wei L. Tan, and Keh K. Kee. 2020. Data Integration for Smart Cities: Opportunities and Challenges. In: Alfred, R., Lim, Y., Haviluddin, H., On, C. (eds) Computational Science and Technology. Lecture Notes in Electrical Engineering, vol 603. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-15-0058-9_38

[281] Kumar Rahul and Rohitash K. Banyal. 2020. Data life cycle management in big data analytics. *Procedia Computer Science* 173 (2020), 364–371. DOI: https://doi.org/10.1016/j.procs.2020.06.042 International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020.

[282] Janusz Rajski, Maciej Trawka, Jerzy Tyszer, and Bartosz Włodarczak. 2022. Hardware root of trust for SSN-basedDFT ecosystems. In *Proceedings of the IEEE International Test Conference (ITC '22)*. IEEE, 479–483.

[283] Aryan Ratra, Aryan Agarwal, Satvik Vats, Vikrant Sharma, Vinay Kukreja, and Satya Prakash Yadav. 2023. A comprehensive review on crime patterns and trends analysis using machine learning. In *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS '23)*. IEEE, 732–736.

[284] Andrew Rebeiro-Hargrave, Pak Lun Fung, Samu Varjonen, Andres Huertas, Salla Sillanpää, Krista Luoma, Tareq Hussein, Tuukka Petäjä, Hilkka Timonen, Jukka Limo, Ville Nousiainen, and Sasu Tarkoma. 2021. City wide participatory sensing of air quality. *Frontiers in Environmental Science* 9 (2021), 773778. DOI: https://doi.org/10.3389/fenvs.2021.773778

[285] Theodoros Rekatsinas, Manas Joglekar, Hector Garcia-Molina, Aditya Parameswaran, and Christopher Ré. 2017. SLiMFast: Guaranteed results for data fusion and source reliability. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD '17)*. Association for Computing Machinery, New York, NY, 1399–1414.

[286] Eric Rescorla. 2018. The Transport Layer Security (TLS) Protocol Version 1.3. RFC 8446. DOI: https://doi.org/10.17487/RFC8446

[287] Murilo B. Ribeiro and Kelly R. Braghetto. 2022. A scalable data integration architecture for smart cities: Implementation and evaluation. *Journal of Information and Data Management* 13, 2 (2022), 207–223. DOI: https://doi.org/10.5753/jidm.2022.2485

[288] National protection and programs directorate, Office of cyber and infrastructure analysis. 2015. The Future of Smart Cities: Cyber-Physical Infrastructure Risk. Report. Retrieved from https://cyberir.mit.edu/?q=future-smart-cities-cyber-physical-infrastructure-risk

[289] Diego O. Rodrigues, Azzedine Boukerche, Thiago H. Silva, Antonio A. F. Loureiro, and Leandro A. Villas. 2017. SMAFramework: Urban data integration framework for mobility analysis in smart cities. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '17)*. Association for Computing Machinery, New York, NY, 227–236. DOI: https://doi.org/10.1145/3127540.3127569

[290] Jakob Rogstadius, Maja Vukovic, Claudio A. Teixeira, Vassilis Kostakos, Evangelos Karapanos, and Jim A. Laredo. 2013. CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development* 57, 5 (Sep. 2013), 4:1–4:13. DOI: https://doi.org/10.1147/JRD.2013.2260692

[291] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638. DOI: https://doi.org/10.1017/S0269888913000039

[292] Marit Rosol, Gwendolyn Blue, and Victoria Fast. 2019. *Social Justice in the Digital Age: Re-Thinking the Smart City with Nancy Fraser*. UCCities - Global Urban Research at the University of Calgary Working Paper #1. DOI: https://doi.org/10.31235/osf.io/wkqy2

[293] Franziska Rosser and John Balmes. 2023. Ozone and childhood respiratory health: A primer for US pediatric providers and a call for a more protective standard. *Pediatric Pulmonology* 58, 5 (2023), 1355–1366.
[294] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. 2015. Trusted execution environment: What it is, and what it is not. In *Proceedings of the 2015 IEEE Trustcom/BigDataSE/Ispa*, Vol. 1. IEEE, 57–64.
[295] Pintu K. Sadhu, Venkata P. Yanambaka, Ahmed Abdelgawad, and Kumar Yelamarthi. 2022. Prospect of internet of medical things: A review on security requirements and solutions. *Sensors* 22, 15 (2022), 5517.
[296] Seref Sagiroglu and Duygu Sinanc. 2013. Big data: A review. In *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13)*. IEEE, 42–47.
[297] Tanvi Sahay, Ankita Mehta, and Shruti Jadon. 2020. Schema matching using machine learning. In *Proceedings of the 7th International Conference on Signal Processing and Integrated Networks (SPIN '20)*. 359–366. DOI: https://doi.org/10.1109/SPIN48934.2020.9071272
[298] Mirko Sailio, Outi-Marja Latvala, and Alexander Szanto. 2020. Cyber threat actors for the factory of the future. *Applied Sciences* 10, 12 (2020), 4334. DOI: https://doi.org/10.3390/app10124334
[299] Eduardo F. Z. Santana, Ana P. Chaves, Marco A. Gerosa, Fabio Kon, and Dejan S. Milojicic. 2017. Software platforms for smart cities: Concepts, requirements, challenges, and a unified reference architecture. *ACM Computing Surveys (Csur)* 50, 6, Article 78 (Nov. 2017), 37 pages. DOI: https://doi.org/10.1145/3124391
[300] Eduardo F. Z. Santana, Ana P. Chaves, Marco A. Gerosa, Fabio Kon, and Dejan S Milojicic. 2018. Software platforms for smart cities: Concepts, requirements, challenges, and a unified reference architecture. *ACM Computing Surveys (CSUR)* 50, 6 (2018), 78.
[301] Pedro M. Santos, João G. P. Rodrigues, Susana B. Cruz, Tiago Lourenco, Pedro M. d'Orey, Yunior Luis, Cecília Rocha, Sofia Sousa, Sérgio Crisóstomo, Cristina Queirós, Susana Sargento, Ana Aguiar, and João Barros. 2018. PortoLivingLab: An IoT-based sensing platform for smart cities. *IEEE Internet of Things Journal* 5, 2 (2018), 523–532. DOI: https://doi.org/10.1109/JIOT.2018.2791522
[302] Parthasarathy Saravanan, Jeganathan Selvaprabu, L. Arun Raj, A. Abdul Azeez Khan, and K. Javubar Sathick. 2021. Survey on crime analysis and prediction using data mining and machine learning techniques. In *Advances in Smart Grid Technology: Select Proceedings of PECCON 2019—Volume II*. Springer, 435–448.
[303] Nimra Shahid, Munam A. Shah, Abid Khan, Carsten Maple, and Gwanggil Jeon. 2021. Towards greener smart cities and road traffic forecasting using air pollution data. *Sustainable Cities and Society* 72 (2021), 103062.
[304] Natalie Shlomo. 2020. Integrating differential privacy in the statistical disclosure control tool-kit for synthetic data production. In *Proceedings of the International Conference on Privacy in Statistical Databases*. Springer, 271–280.
[305] Adam Shostack. 2014. *Threat Modeling: Designing for Security*. John Wiley & Sons.
[306] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. The hadoop distributed file system. In *Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. IEEE, 1–10.
[307] Jorge Silva, João Gabriel Almeida, Thaís Batista, and Everton Cavalcante. 2021. Aquedücte: A Data Integration Service for Smart Cities. In *Brazilian Symposium on Multimedia and the Web (WebMedia '21), Belo Horizonte, Minas Gerais, Brazil, November 5–12, 2021*. Adriano César Machado Pereira and Leonardo Chaves Dutra da Rocha (Eds.), ACM, 177–180.
[308] Rajesh Kumar Singh, H. R. Murty, S. K. Gupta, and A. K. Dikshit. 2012. An overview of sustainability assessment methodologies. *Ecological Indicators* 15, 1 (2012), 281–299. DOI: https://doi.org/10.1016/j.ecolind.2011.01.007
[309] Ola Söderström, Till Paasche, and Francisco Klauser. 2014. Smart cities as corporate storytelling. *City* 18, 3 (2014), 307–320. DOI: https://doi.org/10.1080/13604813.2014.906716
[310] Adir Solomon, Mor Kertis, Bracha Shapira, and Lior Rokach. 2022. A deep learning framework for predicting burglaries based on multiple contextual factors. *Expert Systems with Applications* 199 (2022), 117042.
[311] Juraj Somorovsky, Andreas Mayer, Jörg Schwenk, Marco Kampmann, and Meiko Jensen. 2012. On breaking SAML: Be whoever you want to be. In *Proceedings of the 21st USENIX Security Symposium (Security '12)*. USENIX Association, 21.
[312] Mehdi Sookhak, Helen Tang, Ying He, and F. Richard Yu. 2019. Security and privacy of smart cities: A survey, research issues and challenges. *IEEE Communications Surveys & Tutorials* 21, 2 (2019), 1718–1743. DOI: https://doi.org/10.1109/COMST.2018.2867288
[313] Shreyas Srinivasa, Jens M. Pedersen, and Emmanouil Vasilomanolakis. 2021. Open for hire: Attack trends and misconfiguration pitfalls of IoT devices. In *Proceedings of the 21st ACM Internet Measurement Conference (IMC '21)*. Association for Computing Machinery, New York, NY, 195–215. DOI: https://doi.org/10.1145/3487552.3487833
[314] Telecommunication Standardization Sector of ITU (ITU-T). KPIs on Smart Sustainable Cities. Retrieved from https://www.itu.int/en/ITU-T/ssc/Pages/KPIs-on-SSC.aspx (accessed June 2024).
[315] Telecommunication Standardization Sector of ITU (ITU-T). 2014. *Smart Sustainable Cities: An Analysis of Definitions*. Focus Group Technical Report.

[316] Telecommunication Standardization Sector of ITU (ITU-T). 2016. ITU-T Y.4400 series—Smart sustainable cities—Setting the framework for an ICT architecture. *ITU-T Y-Series Recommendations* Supplement 27.
[317] Telecommunication Standardization Sector of ITU (ITU-T). 2016. Key performance indicators for smart sustainable cities to assess the achievement of sustainable development goals. *ITU-T Recommendation* Y.4903/L.1603.
[318] Telecommunication Standardization Sector of ITU (ITU-T). 2016. Key performance indicators related to the use of information and communication technology in smart sustainable cities. *ITU-T Recommendation* Y.4901/L.1601.
[319] Telecommunication Standardization Sector of ITU (ITU-T). 2016. Overview of key performance indicators in smart sustainable cities. *ITU-T Recommendation* Y.4900/L.1600.
[320] Telecommunication Standardization Sector of ITU (ITU-T). 2016. Overview of key performance indicators in smart sustainable cities. *ITU-T Recommendation* Y.4900/L.1600 (2016).
[321] Telecommunication Standardization Sector of ITU (ITU-T). 2016. Overview of key performance indicators in smart sustainable cities. *ITU-T Recommendation* Y.4902/L.1602.
[322] Telecommunication Standardization Sector of ITU (ITU-T). 2019. Smart sustainable cities maturity model. *ITU-T Recommendation* Y.4904.
[323] Telecommunication Standardization Sector of ITU (ITU-T). 2022. Key performance indicators for smart sustainable cities to assess the achievement of sustainable development goals. *ITU-T Recommendation* Y.4903.
[324] Michael Stonebraker, Uundefinedur Cetintemel, and Stan Zdonik. 2005. The 8 requirements of real-time stream processing. *ACM Sigmod Record* 34, 4 (Dec. 2005), 42–47. DOI: https://doi.org/10.1145/1107499.1107504
[325] Xiang Su, Jukka Riekki, Jukka K. Nurminen, Johanna Nieminen, and Markus Koskimies. 2015. Adding semantics to internet of things. *Concurrency and Computation: Practice and Experience* 27, 8 (2015), 1844–1860. DOI: https://doi.org/10.1002/cpe.3203
[326] Xiang Su, Hao Zhang, Jukka Riekki, Ari Keränen, Jukka K. Nurminen, and Libin Du. 2014. Connecting IoT sensors to knowledge-based systems by transforming SenML to RDF. *Procedia Computer Science* 32 (2014), 215–222. DOI: https://doi.org/10.1016/j.procs.2014.05.417 *The 5th International Conference on Ambient Systems, Networks and Technologies (ANT '14)*, the 4th International Conference on Sustainable Energy Information Technology (SEIT '14).
[327] San-Tsai Sun and Konstantin Beznosov. 2012. The devil is in the (implementation) details: An empirical analysis of OAuth SSO systems. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS '12)*. Association for Computing Machinery, New York, NY, 378–390. DOI: https://doi.org/10.1145/2382196.2382238
[328] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9. DOI: https://doi.org/10.1145/3465416.3483305
[329] V. Shyamala Susan and T. Christopher. 2016. Anatomisation with slicing: A new privacy preservation approach for multiple sensitive attributes. *SpringerPlus* 5, 1 (2016), 1–21.
[330] Iryna Susha, Marijn Janssen, and Stefaan Verhulst. 2017. Data collaboratives as "bazaars"? A review of coordination problems and mechanisms to match demand for data with supply. *Transforming Government: People, Process and Policy* 11, 1 (2017), 157–172. DOI: https://doi.org/10.1108/TG-01-2017-0007
[331] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.
[332] Yusuke Takamori, Junya Sato, Masahiro Fujimoto, Masaki Endo, Shigeyoshi Ohno, Daiju Kato, and Hiroshi Ishikawa. 2022. Current Status of Examples of Initiatives Using Open Data in Government. In Proceedings of the Fourteenth International Conference on Advances in Multimedia (MMEDIA 2022). 14–17.
[333] Martin Tomitsch, Joel Fredericks, Dan Vo, Jessica Frawley, and Marcus Foth. 2021. Non-human personas: Including nature in the participatory design of smart cities. *Interaction Design and Architecture(s)* 50, 50 (2021), 102–130. DOI: https://doi.org/10.55612/s-5002-050-006
[334] Sandhya Tripathi, Bradley Fritz, Mohamed Abdelhack, Michael Avidan, Yixin Chen, and Christopher King. 2022. Deep Learning to Jointly Schema Match, Impute, and Transform Databases. https://arxiv.org/abs/2207.03536
[335] Thuy Truong, Ahmed Khalid, and Philip Leroux. 2020. Optimized Version of Reference Architecture Including Update to IoT Interfaces. CUTLER D2.5. Retrieved from https://www.cutler-h2020.eu/download/1181
[336] Filareti Tsalakanidou, Ekaterina Gilman, Panos Kostakos, and Andrew Byrne. 2018. Requirements for Data Crawling, Integration and Anonymization. CUTLER D3.1. Retrieved from https://www.cutler-h2020.eu/download/517
[337] Mio Tsubakimoto. 2022. Current status and issues of university education-related data in TOKYO OPEN DATA. *IIAI Letters on Institutional Research* 1 (2022). DOI: https://doi.org/10.52731/lir.v001.041
[338] European Union. 2011. Cities of Tomorrow. Challenges, Visions, Ways Forward. Retrieved June 16, 2019 from http://ec.europa.eu/regional_policy/sources/docgener/studies/pdf/citiesoftomorrow/citiesoftomorrow_final.pdf
[339] Muhammad Usman, Mian Ahmad Jan, Xiangjian He, and Jinjun Chen. 2019. A survey on big multimedia data processing and management in smart cities. *ACM Computing Surveys (CSUR)* 52, 3, Article 54 (Jun. 2019), 29 pages. DOI: https://doi.org/10.1145/3323334

[340] Dalton Cézane Gomes Valadares, Newton Carlos Will, Jean Caminha, Mirko Barbosa Perkusich, Angelo Perkusich, and Kyller Costa Gorgonio. 2021. Systematic literature review on the use of trusted execution environments to protect cloud/fog-based Internet of Things applications. *IEEE Access* 9 (2021), 80953–80969.
[341] Maarten van Steen and Andrew S. Tanenbaum. 2017. *Distributed Systems*. Maarten van Steen.
[342] Polychronis Velentzas, Antonio Corral, and Michael Vassilakopoulos. 2021. *Big Spatial and Spatio-Temporal Data Analytics Systems*. Springer, Berlin, 155–180. DOI: https://doi.org/10.1007/978-3-662-62919-2_7
[343] Jayant Venkatanathan, Denzil Ferreira, Michael Benisch, Jialiu Lin, Evangelos Karapanos, Vassilis Kostakos, Norman Sadeh, and Eran Toch. 2011. Improving users' consistency when recalling location sharing preferences. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*. Springer, 380–387.
[344] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S. Rellermeyer. 2020. A survey on distributed machine learning. *ACM Computing Surveys (CSUR)* 53, 2, Article 30 (Mar. 2020), 33 pages. DOI: https://doi.org/10.1145/3377454
[345] Jenni Viitanen and Richard Kingston. 2014. Smart cities and green growth: Outsourcing democratic and environmental resilience to the global technology sector. *Environment and Planning A: Economy and Space* 46, 4 (2014), 803–819. DOI: https://doi.org/10.1068/a46242
[346] Félix J. Villanueva, Maria J. Santofimia, David Villa, Jesús Barba, and Juan Carlos López. 2013. Civitas: The smart city middleware, from sensors to big data. In *Proceedings of the 2013 7h International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. 445–450. DOI: https://doi.org/10.1109/IMIS.2013.80
[347] Massimo Villari, Maria Fazio, Schahram Dustdar, Omer Rana, Devki N. Jha, and Rajiv Ranjan. 2019. Osmosis: The osmotic computing platform for microelements in the cloud, edge, and internet of things. *Computer* 52, 8 (2019), 14–26. DOI: https://doi.org/10.1109/MC.2018.2888767
[348] Aku Visuri, Zeyun Zhu, Denzil Ferreira, Shin'ichi Konomi, and Vassilis Kostakos. 2017. Smartphone detection of collapsed buildings during earthquakes. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*. Association for Computing Machinery, New York, NY, 557–562. DOI: https://doi.org/10.1145/3123024.3124402
[349] Morta Vitunskaite, Ying He, Thomas Brandstetter, and Helge Janicke. 2019. Smart cities and cyber security: Are we there yet? A comparative study on the role of standards, third party risk management and security ownership. *Computers & Security* 83 (2019), 313–331. DOI: https://doi.org/10.1016/j.cose.2019.02.009
[350] Christian Voigt and Jonathan Bright. 2016. The lightweight smart city and biases in repurposed big data. In *Proceedings of the 2nd International Conference on Human and Social Analytics (HUSO '16)*.
[351] Shiraz Ali Wagan, Muhammad Junaid, Nawab Muhammad Faseeh Qureshi, Dong Ryeol Shin, and Keehyun Choi. 2020. Comparative survey on big data security applications, A blink on interactive security mechanism in apache ozone. In *Proceedings of the Global Conference on Wireless and Optical Technologies (GCWOT '20)*. IEEE, 1–6.
[352] Leye Wang, Bin Guo, and Qiang Yang. 2018. Smart city development with urban transfer learning. *Computer* 51, 12 (2018), 32–41. DOI: https://doi.org/10.1109/MC.2018.2880015
[353] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data* 3, 1 (May 2016), 9. DOI: https://doi.org/10.1186/s40537-016-0043-6
[354] Ben Williamson. 2015. Educating the smart city: Schooling smart citizens through computational urbanism. *Big Data & Society* 2, 2 (2015), 2053951715617783. DOI: https://doi.org/10.1177/2053951715617783
[355] Annika Wolff, Daniel Gooch, Jose Cavero, Umar Rashid, and Gerd Kortuem. 2019. Removing barriers for citizen participation to urban innovation. In *The Hackable City: Digital Media and Collaborative City-Making in the Network Society*. Michiel de Lange and Martijn de Waal (Eds.), Springer, 153–168. DOI: https://doi.org/10.1007/978-981-13-2694-3_8
[356] Marc Wright, Hassan Chizari, and Thiago Viana. 2022. A systematic review of smart city infrastructure threat modelling methodologies: A Bayesian focused review. *Sustainability* 14, 16 (2022), 10368. DOI: https://doi.org/10.3390/su141610368
[357] Dianlei Xu, Tong Li, Yong Li, Xiang Su, Sasu Tarkoma, Tao Jiang, Jon Crowcroft, and Pan Hui. 2021. Edge intelligence: Empowering intelligence to the edge of network. *Proceedings of the IEEE* 109, 11 (2021), 1778–1837. DOI: https://doi.org/10.1109/JPROC.2021.3119950
[358] Tarun K. Yadav, Justin Hales, and Kent Seamons. 2022. Poster: User-controlled system-level encryption for all applications. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 3499–3501.
[359] Alexander Yakubov, Wazen Shbair, Anders Wallbom, David Sanda, and Radu State. 2018. A blockchain-based PKI management framework. In *Proceedings of the First IEEE/IFIP International Workshop on Managing and Managed by Blockchain (Man2Block) Colocated with IEEE/IFIP NOMS 2018, Tapei, Tawain 23–27 April 2018*.
[360] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2, Article 12 (Jan. 2019), 19 pages. DOI: https://doi.org/10.1145/3298981

[361] Qiang Yang, Yu Zhang, Wenyuan Dai, and Sinno Jialin Pan. 2020. *Transfer Learning*. Cambridge University Press. DOI: https://doi.org/10.1017/9781139061773

[362] Tan Yigitcanlar. 2021. Smart city beyond efficiency: Technology-policy-community at play for sustainable urban futures. *Housing Policy Debate* 31, 1 (2021), 88–92. DOI: https://doi.org/10.1080/10511482.2020.1846885

[363] Tan Yigitcanlar, Marcus Foth, and Md. Kamruzzaman. 2019. Towards post-anthropocentric cities: Reconceptualizing smart cities to evade urban ecocide. *Journal of Urban Technology* 26, 2 (2019), 147–152. DOI: https://doi.org/10.1080/10630732.2018.1524249

[364] Johanna Ylipulli and Aale Luusua. 2019. Without libraries what have we? Public libraries as nodes for technological empowerment in the era of smart cities, AI and big data. In *Proceedings of the 9th International Conference on Communities & Technologies - Transforming Communities (C & T '19)*. Association for Computing Machinery, New York, NY, 92–101. DOI: https://doi.org/10.1145/3328320.3328387

[365] Hang Yu, Weixu Liu, Jie Lu, Yimin Wen, Xiangfeng Luo, and Guangquan Zhang. 2023. Detecting group concept drift from multiple data streams. *Pattern Recognition* 134, C (Feb. 2023), 11 pages. DOI: https://doi.org/10.1016/j.patcog.2022.109113

[366] Yuanyu Zhang, Mirei Yutaka, Masahiro Sasabe, and Shoji Kasahara. 2020. Attribute-based access control for smart cities: A smart-contract-driven framework. *IEEE Internet of Things Journal* 8, 8 (2020), 6372–6384.

[367] Xinwei Zhao, Saurabh Garg, Carlos Queiroz, and Rajkumar Buyya. 2017. Chapter 11—A taxonomy and survey of stream processing systems. In *Software Architecture for Big Data and the Cloud*. Ivan Mistrik, Rami Bahsoon, Nour Ali, Maritta Heisel, and Bruce Maxim (Eds.), Morgan Kaufmann, Boston, 183–206. DOI: https://doi.org/10.1016/B978-0-12-805467-3.00011-9

[368] Elena Zheleva and Lise Getoor. 2011. Privacy in social networks: A survey. In *Social Network Data Analytics*. Springer, 277–306.

[369] Yu Zheng. 2015. Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data* 1, 1 (Mar. 2015), 16–34. DOI: https://doi.org/10.1109/TBDATA.2015.2465959

[370] Yu Zheng. 2015. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 3, Article 29 (May 2015), 41 pages. DOI: https://doi.org/10.1145/2743025

[371] Yu Zheng. 2018. *Urban Computing*. The MIT Press.

[372] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3, Article 38 (Sep. 2014), 55 pages. DOI: https://doi.org/10.1145/2629592

[373] Qing Zhu, Fan Zhang, Shan Liu, and Yuze Li. 2022. An anticrime information support system design: Application of K-means-VMD-BiGRU in the city of Chicago. *Information & Management* 59, 5 (2022), 103247.

[374] Rui Zhu, Man Sing Wong, Mei-Po Kwan, Min Chen, Paolo Santi, and Carlo Ratti. 2022. An economically feasible optimization of photovoltaic provision using real electricity demand: A case study in New York city. *Sustainable Cities and Society* 78 (2022), 103614.

[375] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109, 1 (2021), 43–76. DOI: https://doi.org/10.1109/JPROC.2020.3004555

[376] Esteban Zimányi, Mahmoud Sakr, and Arthur Lesuisse. 2020. MobilityDB: A mobility database based on PostgreSQL and PostGIS. *ACM Transactions on Database Systems (TODS)* 45, 4, Article 19 (Dec. 2020), 42 pages. DOI: https://doi.org/10.1145/3406534

[377] Marta Ziosi, Benjamin Hewitt, Prathm Juneja, Mariarosaria Taddeo, and Luciano Floridi. 2022. Smart Cities: Reviewing the Debate about their Ethical Implications. Available at SSRN. DOI: https://doi.org/10.2139/ssrn.4001761

[378] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. https://arxiv.org/abs/1511.00148

[379] Ioannis Zografopoulos, Juan Ospina, Xiaorui Liu, and Charalambos Konstantinou. 2021. Cyber-physical energy systems security: Threat modeling, risk assessment, resources, metrics, and case studies. *IEEE Access* 9 (2021), 29775–29818. DOI: https://doi.org/10.1109/ACCESS.2021.3058403

[380] Shoshana Zuboff. 2019. Surveillance capitalism and the challenge of collective action. *New Labor Forum* 28, 1 (Jan. 2019), 10–29. DOI: https://doi.org/10.1177/1095796018819461

[381] Sotiris Zygiaris. 2013. Smart city reference model: Assisting planners to conceptualize the building of smart city innovation ecosystems. *Journal of the Knowledge Economy* 4, 2 (2013), 217–231.

Received 19 May 2023; revised 13 February 2024; accepted 25 March 2024