Using Big Data and its Analytical Techniques for Public Policy Design and Implementation in Latin America and the Caribbean

Patricio Rodríguez
Norma Palomino
Javier Mondaca

Knowledge and Learning Sector (KNL)

Felipe Herrera Library (FHL)

DISCUSSION
PAPER Nº
IDB-DP-514

September 2017

Using Big Data and its Analytical Techniques for Public Policy Design and Implementation in Latin America and the Caribbean

Patricio Rodríguez
Norma Palomino
Javier Mondaca

September 2017

IDB
Inter-American Development Bank

Norma Palomino (npalomino@iadb.org), Interamerican Development Bank, Knowledge and Learning Sector.

Patricio Rodriguez y Javier Mondaca, Center for Advanced Research in Education (CIAE), University of Chile.

Abstract

This paper discusses the constantly changing definition of big data and portrays the landscape of the most widely used analytical techniques in the context of public policy formulation in Latin America and the Caribbean. It also presents the conclusions drawn from three exploratory studies conducted by sector teams of the Inter-American Development Bank regarding firm-level productivity, sustainable urban mobility, and smart cities. Based on these studies, the paper addresses sensitive issues surrounding the use of big data in public policy, such as data security and ownership, privacy, ethical framework of data use, among others. The paper concludes by making recommendations for government agencies regarding the use of public value intelligence and suggesting a competency rubric for "smart consumers" of big data. The target audiences are primarily the decision makers in different sectors of the governments in the region, public sector professionals, and social and economic development specialists.

Keywords: big data, data analytics, data management, data science, decision making, public policy, public value intelligence

Table of Contents

1. Introduction

In today's modern economies, data are a key factor of production, just like fixed assets and human capital [1]. With the advent of information technologies, data that were once scarce have become overabundant [1]–[3]. Despite ever-growing data storage and computing capacity, it is estimated that the amount of data currently generated is greater than the physical storage capacity available: computing capacity grew by an annual rate of 58 percent in the period between 1986 and 2007 [4]. This clearly shows that we are in the era of big data.

The emergence of this trend has enabled the development of a series of technologies and expertise that depend on storage and computing capacities [5]. Big data can have positive impacts on many sectors, from retail and manufacturing to health and public administration [1].

Data use in many sectors has been compared to the advent of electricity: one simply cannot work without it [6]. Moreover, one can envisage a scenario in which knowledge production becomes radically different—that is, traditional scientific methods, theory, and professional expertise would no longer be required—as the data would "speak for themselves" [7].

The reality proved quite the contrary: big data failed to predict the spread of flu [8], [9] and the level of educational improvement [10] (see Section 4), which has called its effectiveness into question, with excessive optimism giving way to doom and gloom. Therefore, it is essential to understand big data to properly assess its potential, and even more importantly, its constraints. This paper will show that, contrary to popular belief, using advanced analytics requires expert decision making at all times: from choosing data sources to selecting analytical methodologies, and especially in interpreting and communicating the results. These decisions ultimately make or break a data analytics project.

Ultimately, even though big data analytics would be impossible without information and communications technologies (ICT), it is important to avoid the mistake of treating data analytics projects like any other ICT infrastructure implementation project. This is because data analytics projects are different in nature and potential impact, as they are strategic endeavors that support evidence-based decision making within organizations.

The objective of this paper is to revisit the concept of big data and its analytics techniques in the context of public policy formulation in Latin America and the Caribbean. The paper also examines three cases that will contribute to a better understanding of the challenges in the implementation of analytics projects in the public sector in the region, providing recommendations and suggestions for success. Therefore, the target audiences of this paper are primarily decision makers, public sector professionals, and government authorities at the micro, mezzo, and macro levels in the region.

The structure of the paper is as follows. Section 2 takes a brief look at the concept of big data and discusses data management and analytics methodologies, as well as the technologies used for said purposes. Section 3 presents three exploratory cases and examines how to apply big data analytics specifically to generate evidence that supports decision making in public policy in Latin America and the Caribbean. Section 4 draws conclusions and recommendations for government agencies regarding the use of public value intelligence and introduces a rubric for decision makers to assess their own competencies.

2. Conceptual Framework

It is necessary to establish a conceptual framework to define big data, as the term has several meanings. The general public usually associates big data with not only the data themselves, but also data processing, data analysis techniques and technologies, as well as the professionals involved and the skills required for carrying out these tasks.

2.1. Big Data: How It Has Been Defined Until Now

The term "big data" originated in the field of computer science, and typically refers to datasets whose sizes exceed the processing capacity of standard software and hardware available for data capture, storage, and analysis [1], [5], [11]–[15]. Initially, multiple authors used the so-called "**Three Vs**" to characterize **what** big data **is**. As shown in [5], [16]–[20], the "Three Vs" stand for:

(1) **Volume**: this refers to the enormous quantity of existing data. Volume has implications for the resources required for data storage and computing capacity. Although volume was the most notable characteristic of big data in the early years and hence the name, given the steady progress in hardware and software capacity, volume is no longer a defining feature of big data.

(2) The **velocity** of big data production and analysis, in other words, the speed at which big data are created, processed, analyzed, and stored [14], [19]. The current communication platforms and devices make it easy to create and/or share information, giving rise to large amounts of information that must be stored and processed in real time [18].

(3) The **variety** in the sources and types of data. Data type depends on its structure—namely, structured, semistructured, or unstructured [5], [16].

Although the "three Vs" definition of big data proved instrumental for modeling issues in the field of computer science and information technology, a definition based on technical characteristics has to be constantly revisited [21]. In response to this, other authors have introduced more qualitative dimensions into the definition of big data [16], [19], [22], such as:

(4) **Variability:** When data volume is low, anomalies (commonly known as outliers) that deviate from prominent patterns are present in observations due to the statistical effect of dispersion [23]. However, in big data, these anomalies exist in such abundance that the name "outlier" loses its meaning as they become an integral part of the big data to be analyzed. An example of this are the viral trends on the Internet [24].

(5) **Complexity**: This refers to the multiplicity and varying quantity of sources of existing data thanks to the proliferation of different online devices. Examples include satellite tracking devices (GPS), the sensors used to enable the Internet of Things (IoT), data spontaneously generated by citizens, and other phenomena of the digital society. Data sources come in two types: intersubject and intrasubject sources [25]. Intersubject sources refer to data collection from many subjects at the same point in time, while intrasubject sources refer to ongoing data collection from the same subject over time (for example, biometric data collected from a fitness tracker). Additionally, big data is exhaustive in its scope, as a dataset can contain all the observations of a particular sample, which allows analysis of the entire set [18]. Moreover, most big data is available in real time or as soon as it is generated, which makes it possible to predict the immediate future, or perform "nowcasting" [26]. In terms of depth, big data can provide the "highest resolution" picture, as it captures extensive details. Online transactions are an example in which big data records every minuscule detail of each operation [18]. Another characteristic of big data is

flexibility both in terms of their extensibility, or the capacity to readily aggregate new types of data, and scalability, or the ability to rapidly grow in size [18].

(6) **Veracity:** This is understood as quality, reliability, and certainty of data, especially in terms of origin and creation. For example, conducting big data analysis by using the messages on social media platforms can be fraught with false information or based on subjective perceptions that are inaccurate and misleading [27]. In the same vein, online transaction data is prone to interruptions or loss of segments due to technical issues. Using various datasets together can exacerbate this problem [28].

(7) **Representativeness**: This measures the extent to which big data provides an adequate representation of the populations analyzed, given the nature of the data or the means of data collection. For example, data generated by social media have problems of underrepresentation (due to lack of participation on or access to social media platforms), overrepresentation (accounts and profiles of deceased persons), and multiplicity (the same person having multiple accounts) [29].

2.2. Big Data Processing and Analysis

Unprocessed big data has little inherent value; it only becomes valuable once fully processed [16]. Big data can help predict investment returns, generate valuable insight, improve processes, and support decision making by reducing uncertainty, among other benefits.

The need to process and analyze big data has led to the creation of a discipline called **Data Science** [15], [19], which combines a broad range of multidisciplinary techniques such as Computer Science, Mathematics, Statistics, Econometrics, and Operational Research [30], [31]. The life cycle of data analysis consists of at least six stages, which are illustrated in Figure 1.

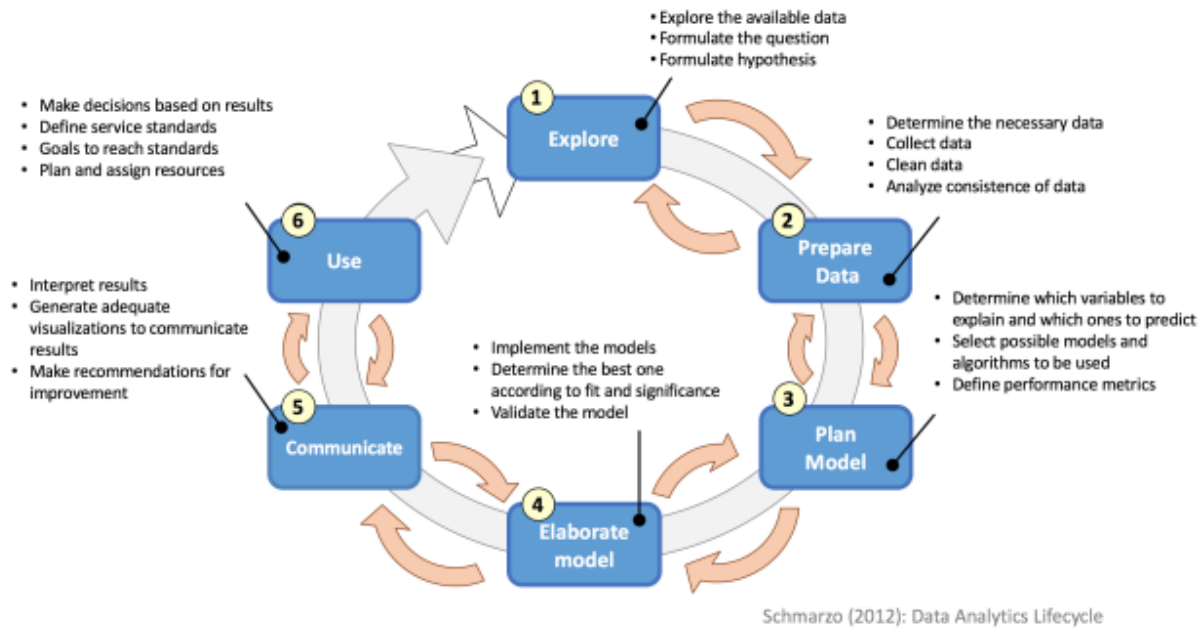

Figure 1: **Life Cycle of Data Analysis** [32].

As shown in Figure 1, the life cycle of data analysis is not a linear process; often, questions need to be rephrased based on data availability or results reinterpreted in light of new evidence. Therefore, data analysis is an iterative process that may involve going back to earlier stages. In any case, big data processing can be broken down into two main components: data management and data analytics [16], [33].

Data management consists of three aspects: (1) collecting and storing the data, (2) cleaning or scrubbing the data, and (3) preparing the data for analysis. **Data analytics** refers to answering the questions and/or hypotheses formulated based on modeling or analytical techniques. As can be seen, this process is not notably different than that of scientific research in any discipline; the main difference lies in the overall characteristics of the data used, which were reviewed in the previous section, and in the challenges in data access and management.

There is a plethora of **data management** methodologies for carrying out the tasks discussed in the previous paragraph. Table 1 shows one way of classifying those methodologies.

Table 1: **Big Data Processing Methodologies. Prepared In-house Based on Gandomi and Haider [16]**

Data type	Examples of processing techniques by data type
Text	Information extraction: gathering structured data from a text, with entity recognition and relation extraction. Text summarization: summarizing one or multiple documents using natural language processing techniques. Question answering: responding to questions formulated in natural language using natural language processing techniques. Sentiment analysis: analyzing opinionated text to generate a negative or positive response.
Audio	Transcript-based approach: converting audio content into a textual transcript using automatic speech recognition with the help of large dictionaries, and then processing the output using text analytics techniques. Phonetic-based approach: converting the sounds and phonemes of a speech into a sequence, and searching for the phonetic representation of a given term within the sequence.
Video	Server-based architecture: dedicated server for performing video analytics. Edge-based architecture: video analytics performed locally and on raw, uncompressed data.
Social media	Content-based analytics: this approach focuses on data posted by users. The data are analyzed with the techniques described above, namely text, audio, or video analytics. Structure-based analytics: this approach synthesizes the structural attributes of social network and gleans intelligence from the relationships among participating entities. The techniques include community detection, social influence analysis, and link prediction.

There are several methodologies for **data analytics**, which are scientific analytical approaches (Table 2) that can have different technological implementations in the form of products and services.

Table 2: Examples of Methodologies for Big Data Modeling and Analysis

Methodology	Description	Applications / Examples
Spatial analysis	A set of techniques that analyze the geometric, topological, and geographical properties of a dataset [34]–[38].	• Spatial regressions (consumption vs. distance from shopping centers), simulations (supply chain performance with different warehouse locations).
Network analysis	A set of techniques that characterize relationships between discrete nodes in a graph or a network [39].	• Identifying opinion leaders to target marketing campaigns. • Identifying bottlenecks in the flow of information within a company. • Modeling transportation networks and predicting travel time from point A to point B.
Machine learning	A sub-discipline within Computer Science (historically known as Artificial Intelligence), which involves designing and developing algorithms that enable behaviors to be inferred from empirical data. There are two types of machine learning: supervised and unsupervised. Supervised learning is the task of inferring a function based on a set of training examples. These examples consist of a set of inputs (vectors) and a set of outputs that are correct answers (that satisfy the function). Having the correct answers make it possible to measure errors in the predictions[1]. In unsupervised learning, there are no known correct answers (or they are not necessary), and therefore there is no feedback for adjusting a function. The objective of the algorithm is to organize data or describe their structure.	• Predicting crime, school, and college dropout rates, post-operative life expectancy, or sales. • Making product suggestions and recommendations based on past purchases. • Natural language processing: voice and language recognition for human-computer interaction (for example, Siri, Cortana, and Alexa), and sentiment analysis of texts and social media inputs. • Pattern recognition: using handwritten text, image processing, and face recognition to search for crime suspects. • Anomaly detection: detecting bank fraud based on unusual purchase activities using credit cards.
Territorial intelligence	These are spatial analysis methodologies that use information technology to combine qualitative, quantitative and spatial approaches, while taking into consideration the participatory approach as well as global, multidisciplinary, and multisectoral approaches. Some examples are grouping analysis, atypical value analysis [40], [41] and multicriteria analyses that combine various spatial characteristics [42].	• Spatial indicators of the level of public and private services. • Gap analysis and spatial growth analysis of service demand and supply. • Urban and rural accessibility, territorial profiling in different geographical, sociodemographic, economic, and human development dimensions.

[1] Many traditional statistical analysis techniques such as multiple linear regression can be understood as machine learning models in which measures of error provide feedback for fine-tuning the models. In fact, there are models – such as multilevel models – that account for data nesting and require minimal measures of error to work.

Methodology	Description	Applications / Examples
Optimization	Numerical modeling techniques for redesigning and improving processes as well as complex, multidimensional systems.	• Optimizing resource allocation in hospitals, schools, production centers, and warehouses. • Production: programming production equipment and inventory management.
A/B testing	A technique that works with a control group and several test groups to determine what changes can lead to improvement in a certain target variable. Therefore, this technique is also known as split testing or bucket testing. The high data volume makes it possible to conduct and analyze a large number of tests, ensuring that the group sizes are large enough to detect statistically significant differences among control and test groups.	• Testing the effectiveness of different marketing campaigns. • Testing the effectiveness of a medical treatment or a type of education through natural experiments in which some subjects participate in an intervention whereas others do not, under different circumstances. The testing aims to work with subjects that are as similar as possible (matching) to control for as many variables as possible.
Simulation	Simulation models the behavior of complex systems to forecast, predict, and plan for scenarios.	• Predicting the financial performance of a company under uncertain circumstances. • Weather forecast.
Analytical visualization of data	A way to discover and understand patterns in large datasets through visual interpretation so that users can navigate and explore the data.	• Interactive visual analysis of the main components [43], [44].
Data visualization	Communicate information clearly and effectively using different methods of interactive graphic representation [45]–[48].	• Infographics. • Dashboards for tracking and synthesizing certain occurrences.

There are also technological products and services (and often software) available for data management and analysis. Table 3 lists some of these.

Table 3: Technological Services for Big Data Processing and Analysis

Service	Description	Examples
New analytical frameworks	Work environments that contain or can contain a series of packages and libraries that allow for code reuse to facilitate common tasks.	Hadoop (Google, Apache), Spark
Data warehouse and data lakes	Data repositories that only receive new data and allow databases to be created through data marts. The difference between a data warehouse and a data lake lies in the different structures of the data. Data warehouses store structured data (for example, tables with data organized in rows and columns), while data lakes can store structured, semi-structured, and unstructured data.	SQL Server, Azure SQL, NoSQL
Relational database	Database consisting of a collection of tables (relations). Relational database management systems (RDBMS) store a type of structured data. SQL is the most commonly used language for RDBMS (see below for more details).	MySQL, PostegreSQL, Oracle, SparkSQL
Non-relational database	Database that does not store data in table format (rows and columns).	MongoDB, Cassandra
Data visualization	Tools for visualizing data, of varying degrees of flexibility and versatility for customized results.	D3js, Google Charts, Tableau, Vega
Statistical tools/plugins	Statistical packages or extensions for conducting statistical analysis on data. Simpler tools/plugins use a graphic interface for tasks, while more complex ones require familiarity with a certain programming language.	SAS, Stata, SPSS, Matlab, R, Python, Pandas
Geographic information system	System designed for processing, storing, analyzing, and visualizing geographic data.	Leaflet, PostGIS, Esri ArcGIS, CartoDB
Cloud Computing	Enables on-demand access to a series of configurable computer resources, which can be used or unlocked without having to interact with the provider and does not require a significant amount of management resources [1]. Cloud computing has five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity, and capacity to automatically control and optimize resource use.	Amazon Web Services, Google Cloud Platform, Microsoft Azure, Digital Ocean: Cloud Services for Developers

The analysis cycle of data science requires the participation of specialists with a solid background in one or more fields such as computer science, application use and development, modeling, statistics, analytics, and mathematics. Theses specialists are known as **data scientists**, who explore, raise questions, conduct scenario analysis ("what happens if…"), and question existing assumptions and processes using multiple data sources from different origins [19], [49].

3. Using Advanced Analytics for Public Policy Decision Making, Design, Implementation, and Evaluation

When applied to public policy decision making, design, implementation, and evaluation, the objective of data science is to produce relevant, high-quality, and timely evidence to underpin and guide decisions. This involves identifying problems that go unnoticed and therefore are not actionable [43], and this process is known as data-driven decision making [15].

There is strong evidence showing that big data applications can play an important role in benefiting not only private firms, but also national economies and citizens [1], by creating value in the global economy to enhance private and public sector productivity and competitiveness and to generate economic surplus for consumers [1].

Private firms have been using data analysis to support decision making for a long time. Firms carry out complex calculations on their consumer data, using big data analysis techniques known as business intelligence to identify patterns and trends that predict future consumer behavior, assess the impact of segmentation in a marketing campaign, or recommend products and services based on past purchases, among others [1], [50].

Big data analysis can also help improve public administration by generating more and better solutions that meet the needs of healthcare, education, transportation, housing, assistance, and inclusion of certain socially, demographically, and geographically disadvantaged groups. This is particularly relevant for the public sector, whose prevailing culture is one of treating all citizens equally, regardless of their individual or collective characteristics [1].

Thus, with access to big data and the use of adequate analytical techniques, it has become possible to identify and measure previously invisible and therefore unsolvable issues. In light of this, a type of **"public value intelligence"** can be developed (a social equivalent of "business intelligence"), which can potentially become a strategic component of public policy decision making, design, implementation, and evaluation by the governments of Latin America and the Caribbean.

In one specific example, the World Bank launched the first Big Data Innovation Challenge [51] to recognize a series of initiatives in support of big-data-driven public value generation in various areas, among which are the following:

- **Poverty**: India used nighttime satellite images to analyze access to electricity in more than 600,000 villages to take stock of their needs. Sri Lanka and Pakistan used similar technology to measure other variables (number of cars, built area, shade, floor type, street type, among others), leading to more cost-effective indicators with comparable, if not greater, accuracy than traditional methods. Nigeria used nighttime satellite images (which show the sectors with access to electricity) to assess the relationship between poverty and market ineffectiveness by analyzing and aggregating the monthly prices of hundreds of commodities.
- **Crime and security**: In Bogota, a study examined the relationship between crime and urban infrastructure using route information from the bus rapid transit (BRT) system and risk terra in modeling for data analysis. The results found certain areas near hospitals, schools, pharmacies, and bus stations to be prone to assaults and murders, identified crime peak-hours, and predicted the parts of the city most vulnerable to potential criminal activity.

- **Transportation**: The Philippines saw the development of two applications, OpenRoad and Open Traffic. OpenRoad is an interactive portal that allows users to track the status of publicly funded road projects and provide feedback by project or by location. Open Traffic is an application that allows the user to visualize and analyze information on traffic speed by using the inputs from GPSs installed in taxis and data gathered from the cellphones of taxi drivers. In Belarus, a government-backed initiative led to the development of an application called RoadLab, which can assess the surface quality of streets and roads by using the accelerometer in cellphones and submitting location information through GPS. To do this, the application divides the streets into 100-meter segments, georeferencing the beginning and end points. An estimate value in reference to the International Roughness Index is calculated using the data collected.
- **Health**: In South Africa, an initiative has developed algorithms to consolidate the databases of different public institutions that manage information on AIDS patients and clinics, the amount of HIV lab tests, and other AIDS-related information. This makes it possible to identify the places that serve the highest proportions of AIDS patients at the national, provincial, district, and clinic-level [52].

Other relevant cases that did not enter the Big Data Innovation Challenge of the World Bank include two public health initiatives in Chicago and Indiana, United States [53]. Chicago improved its rodent eradication effort by carrying out a joint project with the pest control agency and using big data, while Indiana set up a high-level data center with top-notch professionals as well as investment in cutting-edge technology and information security to reduce infant mortality.

The following section presents an in-depth review of three cases of implementation of big data analytics in Latin America and the Caribbean. These cases are concrete examples of the use and impact of advanced analytics in the region, and they systematically reflect the main challenges, opportunities, and lessons learned from the projects. The first case generates data for public policy from the implementation of a specific intervention project in the region. The second case evaluates third party experiences to provide input for public policy. Different than the first two, the third case is a more academically oriented exercise, and yet it contributes valuable insight from a scholarly perspective and because of its methodology.

To make these cases more reader-friendly, each one is presented separately, and its respective review and analysis are structured as follows:

(1) Brief introduction of the case, identifying the needs/issues to be addressed, as well as the potential questions and hypotheses that may have been proposed.
(2) Analysis of what data had to be gathered or was available to answer the questions or evaluate the hypotheses proposed.
(3) Review of the methodologies used to collect and process data and to produce the results obtained.
(4) Summary of relevant lessons for public policy or for the case-specific topic.

3.1. Sustainable Urban Mobility, Big Data, and Public Policy: Cyclist Mobility Study in the City of Rosario, Argentina

Case Description, Needs and/or Issues Identified

The first study was carried out in Rosario, Argentina, to provide an account of the situation of cyclist mobility in the city using georeferencing devices. The research aimed at identifying (1) cyclist mobility patterns with respect to existing road infrastructure (whether cyclists use bike

paths or not), (2) the relationship between cyclist mobility patterns and road accidents, and (3) potential road infrastructure improvements [54].

The study consisted of two phases. In the first phase, 40 volunteers (bike owners) received and installed GPS tracking devices on their bikes. In the second phase, 150 bikes were provided through a public system called "*Mi bici tu bici"* ("my bike, your bike"). At the time of the study, there were 173 bikes in active use in the public system, which means that the GPS devices covered 85 percent of the fleet. The study included private bike users to assess alternative routes, as their mobility patterns are not geographically limited by the locations of public bike terminals.

Each device installed on publicly or privately owned bikes contained an accelerometer and a GPS, and could thus collect data on speed and routes. The data used to analyze accidents came from several government agencies: the Road Observatory of the Provincial Road Safety Agency of Santa Fe for information on accidents, injuries, and deaths; the Mobility Authority of Rosario for information on the relationship between bike path use and collisions in the city; as well as targeted interviews with private bike users to confirm the quantitative information obtained.

Methodologies for Data Analytics

It took two weeks to gather data from private bike users, and six weeks from public system users. In both cases, only workday data was collected (from Monday to Friday). Based on the data, the number of trips, travel time, average distance and speed, and the most used roads were determined. This data could be aggregated monthly, daily, hourly, or by individuals' age or gender.

In the first part of data processing, the data are visually represented by creating maps that showed the main corridors used by cyclists, the speed of travel, accident hotspots, and the severity of accidents. This step helped identify the roads, crossroads, and specific areas that require greater attention due to the frequency and severity of accidents. In the second part, the study relied on interviews to further examine the accident hotspots as well as overall issues.

Lessons for Public Policy

This case shows that big data analysis can help diagnose and define actions to improve citizen welfare. The study findings show that cyclists generally prefer roads with bike paths as they enable them to make faster and safer trips. With the data, it was also possible to locate the accident hotspots on roads without bike paths but with significant bike traffic. Additionally, the study identified several road hubs without bike paths that are frequently used by cyclists as they travel to and from government agencies, public utility companies, universities, and schools.

Based on this information, it is possible to make informed decisions to improve the services provided to citizens. In this particular case, the study sheds light on the situation on Oroño Boulevard, which acts as a hub of the city (and hosts several public utility companies, schools, health facilities, and cultural centers) and is widely used by cyclists despite having no ad hoc infrastructure or authorization for bike traffic. Once the reason for bike use is identified, various options for retrofitting the boulevard were proposed to accommodate pedestrian traffic and to maintain the neighborhood's appeal.

The use of public value intelligence presented one particular challenge. Data processing and analysis were conducted externally, which indicates a lack of infrastructure and human resources in the government agencies to undertake this type of project.

In the last phase, the same techniques for data collection and analysis could be used for project impact assessment. For example, observing changes in road infrastructure (old and new) use to determine if the objectives of improving public policy have been met.

3.2. Computing a New Path for Governance: Big Data Innovations in Latin America and the Caribbean

Case Description, Needs and/or Issues Identified

The second case is an analysis of four smart cities in Latin America. Smart cities have three governance-related objectives: enhance transparency, improve efficiency, and achieve continuous innovation [55].

The cases examined here are part of a study by the Inter-American Development Bank (IDB) [55] aimed at assessing the capacity of the cities to support innovative initiatives in big data analysis, understand the peculiarities of Latin America in these cities to enhance governance, and identify specific local challenges, solutions, and innovations. Below is a brief description of each case:

- **Bahia Blanca (Argentina)**: The Initiative *¿Qué pasa Bahía Blanca*? (QPBB) is a response to the tension between environmental activists and the local petrochemical industry. The government responded by installing sensors that can measure different environmental variables in strategic sectors and share the data captured through a platform and a cellphone application. This made it possible to track air and noise pollution produced by petrochemical plants in real time. The data gathered were made openly available on the platform.
- **Cordoba (Argentina)**: A public transportation fleet tracking system was developed, with special emphasis on the city center. Although one of the operators was already providing its own data on travel time, routes, and ticket revenues, a public transportation tracking system was deemed necessary for two reasons. First, there were concerns over the accuracy of the information provided—especially with respect to public transportation provider revenues—and the difficulty in consolidating operator-generated data. Second, the city was attempting to regain its trailblazing legacy in public transportation, which had become privately operated in recent years.
- **São Bernardo do Campo (Brazil)**: In response to the challenges posed by the city's growth, the federal government made a gradual effort to improve the infrastructure and logistics supporting several public services, creating *Você SBC*, a cellphone application that allows citizens to lodge complaints or make suggestions regarding a broad range of non-urgent services (such as trash collection problems, noise, potholes in the road surfaces, among many others). The application makes it possible to identify and track the needs of the city and its residents.
- **Fortaleza (Brazil)**: The *Fortaleza Inteligente* project was born out of two attempts to improve city governance that began in 2013. The first one was the creation of CITINOVA, a public foundation with a mission to promote the use of science, technology, and innovation in government to improve the services provided to citizens. The second attempt was the formulation of an "urgent transportation and transit action plan" (*Plano de Ações Imediatas de Transporte e Trânsito*, or PAITT), a master plan consisting of initiatives to upgrade public transportation and improve traffic in the city. *Fortaleza Inteligente* led to three pilot projects. The first project uses the data collected from GPSs installed in the buses of the public transportation system to avoid delays and overcrowding. The second one uses the data from the public bike system to analyze bike use in the city and make the case for a scale-up. The

third one created a dashboard that consolidates all the transportation system indicators and generates web-based visualizations.

Necessary and Available Data

Given the variety of initiatives, a plethora of data was needed. On one hand, the **Bahia Blanca** (Argentina) project required the installation of multiple sensors to provide ongoing tracking of air quality[2], sewage, and industrial noise in the vicinity of petrochemical plants, as well as information on the geolocation of these plants. To complement this information, cameras were installed to enable real-time monitoring of some of the plants.

The **Cordoba** (Argentina) project required the installation of GPS devices in all the buses of the public transportation system[3]. Additionally, payment data were collected from the terminals using a one-card payment system that covers the city's entire public transportation system; data on time, amount, and location (using GPS data) of payment in each one-card transaction were also collected.

The **São Bernardo do Campo (SBC)** project used the *Você SBC* platform to collect data sent from the city residents' own devices, which had been registered previously. The platform allowed users to submit various types of non-urgent requests, which are georeferenced to identify areas with overflowing trash or places where a fallen tree needs to be removed, as well as public spaces of interest that could host activities such as open-air markets.

The **Fortaleza** project gathered the greatest variety of data as it has three different components. Like the Cordoba project, the Fortaleza project installed GPS devices on public transportation to measure the speed and location of approximately 2,000 buses of the fleet. Fortaleza also implemented a single-card payment system that covers all public transportation (including public bikes) called *bilhete único* (single ticket). While the data collected in Cordoba and Fortaleza are similar, Fortaleza also gathered data from sensors installed at intersections, radars (to measure the level of vehicle congestion), and complementary sources such as statistics on crime and public transportation accidents.

Three types of data have been gathered for this assessment: (1) information on the specific aspects of each project (background, stages, key decision makers, design process, business model, among others); (2) information on the project status (with further details in the project analysis section), and (3) contextual variables beyond the control of the agents involved (such as political, technological, and macro-level social factors).

Methodologies for Data Analytics

To analyze the cases presented here, reports were written to provide information on the design and progress of the projects, as well as the obstacles, outcomes, and impacts. The main analytical tool used was the urban big data maturity model proposed by the IDB [55]. The model consists of a five-dimensional, five-tier rubric (see Annex A), which helps determine the development (or maturity) level of an urban big data initiative. For the purposes of case description and analysis, the dimensions of the rubric are defined as follows:

[2] Presence of particulate matter, sulfur dioxide, ozone, nitrogen oxides, and carbon monoxide, as well as wind speed and direction, and air pressure. Source: http://www.quepasabahiablanca.gov.ar.
[3] This includes the buses that already have other GPS devices installed by their respective operator.

1. **Open data**: providing the data required and creating demand for data-based governance
2. **Building data ecosystems**: creating communities and mechanisms for sharing data and building a culture of data use, particularly among decision makers and stakeholders
3. **Analytics**: the techniques used to analyze, summarize, and visualize information
4. **Data-based decision making**: including the individual skills as well as institutional and cultural practices necessary for using data to improve public policy
5. **Citizen participation and public services**: using data to create new types of government-citizen relationships

Each of these five dimensions has a rubric of adoption levels that range from solving specific, temporary issues to full ownership and significant ongoing improvement. Detailed information on each level of data adoption is found in the table in Annex A. The four cases analyzed here were assessed in reference to these dimensions. Table 4 below shows the final maturity levels of the cases.

Table 4: **Initial and Final Maturity Levels of Projects Based on the Model of Big Data Adoption for Solving Urban Problems in Zambrano [55].**

		Initial — final maturity level of each dimension in the rubric				
Case	Implementation period	Open data	Building data ecosystems	Analytics	Data-based decision making	Citizen participation and public services
Bahía Blanca	2012—2016	1.0 —2.0	1.0 —2.0	1.0 —1.0	1.0 —1.0	1.0—2.0
Córdoba	2012—2016	2.5—2.5	1.0 —2.0	1.5 —2.5	1.5 —2.5	2.0—2.0
São Bernardo do Campo	2014—2016	2.0—2.0	2.0—2.0	2.0—2.0	2.0—3.5	2.0—3.0
Fortaleza	2013—2015	2.5 —2.5	2.0 —3.0	2.0—3.0	2.0—3.0	2.0—2.0

Lessons for Public Policies

Based on the cases assessed and the analytical tools proposed, there are several things to consider when using big data to support public policy. The diverse range of objectives and needs and the different maturity levels spell out different challenges and tensions that provide lessons for the future.

Although the technical and analytical procedures are important, this section will focus on the characteristics, capacities, and potential of the institutions examined, as the latter is indispensable for success in implementing public value intelligence initiatives within governments. Generally speaking, success is measured against the following dimensions:

(1) **Building an institutional framework** that can generate, manage, and sustain the resources needed for infrastructure and the staff dedicated to public value intelligence. The framework can thus provide a common ground for establishing and discussing the objectives and scope of the abovementioned innovations such as the Fortaleza Initiative. Additionally, the framework can play a key role in defining data ownership or access, while the institutions should establish clear guidelines early on, and especially when they work with nongovernmental third parties.

(2) Achieving **transparent and seamless communication with other external entities** (public or private) and with the citizens. In the early stages of a public value intelligence project, the capacity to adapt the data to the needs of the citizens and not those of the data-processing entities is vital for strengthening ties with the citizens. The cases of Bahia Blanca (providing information to other entities) and of São Bernardo do Campo are good examples of this principle. In the latter, the *Você SBC* application—designed to reach out to the citizens—was developed in a hackathon. This can encourage citizen participation on two levels, by engaging them as agents of innovation and producers of data. Interaction with other public entities is also key: since data analysis can potentially be of use to other agencies, a multilateral perspective is necessary for achieving maximum impact. To facilitate this type of interaction, the implementation of open data policies and infrastructure is needed, as is a data ecosystem for various public agencies to share data with other key players. Academia can also contribute knowledge and experience, such as in the case of Cordoba, where a university and a public institution collaborated in data validation, fostering technological innovation in the private sector to achieve public objectives.

(3) **Availability of the necessary human capital** consisting of professional data analysts, decision makers, or key stakeholders who formulate questions and define objectives. Data scientists have the knowledge and skills required for improving open data infrastructure and ecosystems, and can suggest the right type of analytics needed. Decision makers and key stakeholders should have the capacity of guiding the work of the data scientists, making sure that it aligns with the needs of the citizens. The human capital required may be found outside of the public sector; for example, the Cordoba initiative relied on a private company to collect, clean, and analyze the data.[4] Still, the public sector must develop these skills in-house if they want to sustain the projects over the long run.

The abovementioned dimensions are interrelated. While these dimensions are consistent with the maturity model, they also share similarities with some of the institutional capacities required for big data adoption in the public sector [30], [43], [56]. Section 4 will provide an overview of these capacities.

Hence, the maturity model provides not only lessons for public policy, but also previously unexplored possibilities that can be worthwhile. This model allows for conducting ex ante, a posteriori, and self-assessments that help organizations identify their strengths and challenges in undertaking a public value intelligence project [55]. Two assessments were conducted on the projects at different points in time to gauge the projects' progress in terms of maturity level.

3.3. Using Firm-level Data to Analyze Growth and Dispersion in Total Factor Productivity

Case Description, Needs and/or Issues Identified

The third case is a study that uses firm-level data to estimate the growth and dispersion in total factor productivity (TFP). TFP is the portion of output not explained by the amounts of input needed for production, and is determined by the efficiency and intensity of the inputs used [57].

In the literature, there are few attempts to estimate the TFP of multiple countries simultaneously. This study, however, estimated the TFP of approximately 20 million firms in nearly 30 countries over a period of eight years, providing a much fuller picture of the status and evolution of TFP.

[4] ATOS SIEMENS was in charge of the design and implementation of the study as well as the data collection and analysis.

Two relevant background events provide context for this study: **(1)** TFP in all countries grew consistently in the years prior to the financial crisis of 2008, but began to decline steadily starting in 2011. **(2)** The dispersion in TFP among countries has widened since 2010. Firm-level behavior is expected to mirror the slowdown in TFP and the increase in dispersion.

Necessary or Available Data

Understanding the broader economic landscape requires gathering an enormous amount of firm-level data. The data were collected from the Orbis database [58], which contains all the information required to calculate TFPs, including financial reports and measures of productivity.[5]

Methodologies for Data Analytics

The project consisted of two main steps: preparing the data and determining the production functions. The more challenging step was cleaning and preparing the data, which was time-consuming and required the right expertise. Cleaning the data required imputing the missing data (which could be calculated based on the available data), and excluding data on firms that did not report key values. Preparing the data involved calculating new variables that describe firm workers, raw material, and machinery, based on assumptions regarding how inputs are used in different categories of firms.

Subsequently, production functions were calculated using regression models that determine output based on the available inputs. Four different methodologies were used in this step, all of which were based on ordinary least squares (OLS) to preserve the robustness of the results [59]. This was the most computing-intensive part of the process due to the amount of data and the variety of calculation methodologies used. Afterwards, the elasticities of the factors of production were calculated, which are the weights of each factor in its respective industry, country, and firm.

Lessons for Public Policy

The results show that there is no clear relationship between the dispersion observed and average TFP (accounting for different categories of firms) by country. When controlling for the TFP baseline levels, nearly all measures indicate that dispersion is negatively correlated with future growth and TFP.

This analysis employed a mix of big data techniques and traditional statistical methods. First, data are collected and analyzed using a server with computing capacity far superior to household devices (and to professional devices in some cases). Then, the analysis used OLS-based techniques drawn from traditional inferential statistics.

This case opens a debate on **representativeness** that is highly relevant to academia and to the public sector. This is because despite the existence of an enormous amount of data, European countries are still overrepresented in them.

Additionally, as the case uses traditional statistical methods in data samples that can be considered big data (due to their volume), the validity of the results can be compromised—a problem common to big data analytics. The problem can be framed in the following way: **What is the purpose of the analysis: to predict an occurrence or to understand causality**? Big data analytics is typically used to predict (even though it can be used to study causality), as it can

[5] 479 NAICS (North American Industry Classification System) codes were obtained from the Orbis database. NAICS codes is the standard system for classifying different types of business activities and is used by federal statistics agencies in the United States. (http://www.census.gov/eos/www/naics/).

provide external validation of results [60]. On the other hand, OLS methods are used to model casual explanations as they provide internal validation. In fact, given the objectives of the research—to examine the causes of variability and slowdown in TFP—the second alternative is needed. For this reason, representativeness-related challenges are not a significant issue. However, when researching firms in other countries or predicting their behavior using available data, it is necessary to assess the representativeness of the sample and of the methods used to ensure adequate external validity of the results.

4. Discussion

As shown in the previous section, big data analytics can help generate evidence for the design, development, and evaluation of public policies. This can potentially help improve decision making and support governments in providing better services to citizens.

Thus, advanced big data analytics can be a tool, and not an end in itself. Likewise, technology is a necessary but insufficient condition for conducting big data analytics. Developing public value intelligence at various levels of government (micro, mezzo, and macro level) is essential for building a culture of evidence-based decision making. As all new tools have certain methodological constraints and privacy issues, it is important to take into account the ethical, legal, intellectual, and safety implications when using big data analytics.

This section discusses the constraints of advanced analytics in light of the conceptual framework of big data (Section 2) and the case studies (Section 3). It also provides recommendations for implementation in government agencies, and considers several emerging opportunities such as sharing and disseminating data through various public entities.

4.1. Challenges and Constraints

Data Analysis, Methodologies, and Technologies

First, the overabundance of data has been matched with the representativeness of the data itself. Therefore, methodological issues such as source reliability are more relevant than ever [5], [28]. Specifically, data collected through digital channels only represent some of the more active users, and in the best-case scenario, only those with access to information and communication technologies (ICT), whose penetration rate in Latin America and the Caribbean is far below 100 percent [61]. Therefore, issues of underrepresentation, overrepresentation, and multiplicity make it harder to formulate generalized inferences, as they raise the question of whether or not big data represents the diversity of the population being studied [29].

Second, there is a lack of proven and rigorous stochastic techniques to statistically compensate for errors, biases, or deviations [29]. Third, big data eludes the classical academic definition of data itself. In the traditional econometric analysis framework, for example, data are defined as the values of a variable that is part of a methodological model aimed at answering a research question[6]. For big data generated from transactions, the concept of data is not linked to any econometric model in particular, but to the record of transactions and their “footprints” on digital information platforms or systems [29], [62]. Therefore, big data is similar to administrative data in definition and type, as the latter is generated more organically and for purposes other than research[29].

Additionally, even though the bulk of data processing can be automated using a variety of existing technologies (see Table 1, Table 2 in Section 2.2), it does not mean that data scientists should

not make decisions, such as at which point in the life cycle of data analysis to mine and clean data (see Figure 1 in Section 2.2), what analytical methods to choose, and in the end, how to interpret the results, which are not self-explanatory [28]. In other words, big data analytics is not an entirely **objective** discipline since it has a significant **subjective** component [28]. For this reason, there are **10 critical points** [63] that help steer clear of the problems in big data analytics processes, which are to choose: **(1)** the right analytical problem, **(2)** the right subject population, **(3)** the right data sources, **(4)** the right data samples, **(5)** the correct versions of the model, **(6)** the right predictive variables, **(7)** the right modeling approach and algorithms, **(8)** the right model-validation frequency, **(9)** the right validations and adjustments to determine significance, and **(10)** the right types of visualizations.

Even with these precautions, prediction errors can still occur. For example, Google FluTrends, which sought to predict flu cases, failed spectacularly in early 2013. There is speculation that since the algorithm predicts the spread of flu based on users' Google searches, the widespread press coverage in late 2012 triggered flu-related searches by people who were not necessarily infected [8]. Other attempts to explain the failure point to the Google-generated diagnosis suggestions based on the users' symptom searches [9]. In either case, prediction algorithms must be constantly adjusted and validated using other data sources, since the abovementioned issue of population bias is inherent to online searches.

In other cases, the cause of prediction failure is not the data themselves, but the mistakes made in data analysis or interpretation [5]. It is highly probable that large data volumes can produce spurious correlations among variables and statistically significant results [5], [15], [64]. As models can be prone to overfitting[7], their potential extrapolation and hence interpretation are highly context-dependent [28].

A case in point is the analysis of PISA results, which show that the most successful education systems are those that focus on both quality and equity [65], but do not necessarily indicate a causal relationship. However, reforms were implemented based on the interpretation of major trends in national and international assessments without an in-depth understanding of the details of what makes a difference in schools [10].

In conclusion, one cannot expect the use of advanced big data analytics in itself to substitute more traditional methods of research analysis; on the contrary, it should and can complement other tools [66], especially qualitative ones.

Privacy, Ethical and Legal Issues, Security, and Ownership

The analysis of personal variables containing big data for public and private purposes inevitably raises ethical and legal issues regarding:

(1) **Protection of personal privacy**, in other words, anonymizing the individuals whose data are being analyzed.

(2) **Analysis of private information**, namely, concerns over the inequality and damage caused by intruding into people's personal lives[6] [67], [68].

[6] One example of this is Target, an American supermarket chain, which tried to predict who among its customers are getting pregnant and send promotional ads based on their purchase of certain items. The problem arose when one of these ads was sent to a teenager who had not disclosed her pregnancy to her parents [50].

(3) **Ownership of big data** and the rights and authorizations regarding their management, maintenance, use, and exploitation.

Regarding the **protection of personal privacy**, it is possible to identify individuals by combining various data sources. For example, Sweeney [69] estimates that 87 percent of Americans can be identified using only zip code, gender, and date of birth. Similarly, a study conducted by Bahamonde et al. [70] shows how easy it is to obtain and determine someone's home address using the information stored in prepaid public transportation cards (Bip!) in Santiago, Chile.

Regarding the **analysis of private information**, intrusion into people's personal lives can lead to discriminatory practices in employment eligibility or access to services. Related to the previous point, there are information safety issues surrounding privacy protection in the processes of data collection, management, and analysis. One tool used for this purpose is data encryption, both in data storage and distribution channels.

Another aspect that has raised concerns is **data ownership**: who owns a dataset—tech platform owners (such as Facebook), or the people who use these platforms (in other words, individuals who create profiles and flood their virtual walls with big data)? What rights and licenses of data use are associated with ownership? What is considered "fair" and "safe" use—use that respects the integrity of personal information? In essence, the fact that big data is publicly available does not mean that it is ethical to use it as one pleases. It is necessary to ensure a conscious use of data through mechanisms that hold the people involved in data analysis [28]—especially public servants—accountable. In light of this, when external entities (such as universities or firms) are the ones conducting the analyses, it is important to properly safeguard data ownership, establish protection mechanisms, and prohibit further use of the data for other purposes. These types of legal issues need to be clarified at the outset [1] and carefully weighed, particularly when the analysis is done by third parties or when it involves the use of products whose terms and conditions grant permission to the manufacturer or service provider to access the data or transfer ownership back to the service provider. For example, who owns the data generated from cellphone activities—the phone companies or the users? Is the degree of aggregation sufficient to guarantee complete anonymity of the individuals? These issues must be addressed when defining big data for public policy design.

Ultimately, **legal responsibility** lies with whoever is in charge of managing the potential negative consequences of big data analysis: issues with data ownership and protection, personal privacy and consumer protection, and data safety, among others [71].

4.2. Recommendations

The following is a series of recommendations based on the case studies and the discussion in the context of the conceptual framework proposed.

Recommendations on the Use of Public Value Intelligence by Government Agencies

Implementing public value intelligence projects requires a series of **institutional capacities** within the government. Some authors have identified at least three types of capacities: human capital, technology, and strategy formulation [30], [43], [56], which are specified below:

- **Human capital**: Necessary for tasks such as analyzing the available information; cleaning, preparing, formatting, and ensuring the reliability of the data; and providing training that focuses on data analysis and data-based solutions. On the other hand, there are few "smart

consumers" of data who can evaluate the information with a critical eye. Leadership is needed to raise data awareness, improve data use, and build a data-oriented organizational culture.

- **Technology**: Few technological resources and software services exist for using big datasets and data storage. There is also a lack of interoperability among the systems of different agencies and/or departments, and among tools for taking data-based actions.
- **Strategy formulation**: It is necessary to have a plan to determine the urgent questions to address, the data to collect, and the techniques to use for data analysis. The plan should also contemplate strategic alliances with organizations whose mission is to support the use, quality, and reliability of data.

It is important to establish an **institutional framework** to sustain the initiatives over time. Townsend and Zambrano-Barragan [55] showed how Bahía Blanca ended up phasing out its initiative, while in the case of Fortaleza, an institutional framework helped sustain its initiative, the resources involved, the working guidelines, and even access to data. The institutional framework should also promote **transparent and smooth communication with other external entities** (see Section 3.2), and thus tackling the fundamental challenge of organizing and sharing data to be used in the analysis. This involves sharing data among the different government agencies [20] and developing in-house leadership to define how big data will be used and for what purpose. It is also important for the decision makers themselves to get involved to ensure timely access to data (especially the data that are only available during certain periods of time), and move toward a culture of data-based decision making. The Cordoba initiative led to partnerships between a university and a public institution, and put in place the incentives that promote private investment in technology to help achieve public objectives.

Clear and transparent communication with the citizens is also necessary (see Section 3.2), of which Bahia Blanca and São Bernardo do Campo are two examples. The former achieved smooth communication with the citizens and adapted the data to their needs. The latter involved citizens in the development of the application (through a hackathon) and data production.

Regarding **human capital**, it is necessary to have professionals who conduct the data analysis and the **smart consumers** of the evidence produced. The task of smart consumers is to formulate questions, and critically analyze the information received, questioning the sources, assumptions, and methodology used to produce the information [43]. This paper proposes a competency rubric for smart consumers (**Annex B**).

For **data scientists**, it is necessary to have professionals with particular skillsets to conduct analyses that can produce valuable information and generate input for the data-oriented decision-making cycle. However, these professionals are not easy to find. It is estimated that by 2018, the United States alone will see a talent shortfall of between 140,000 and 190,000 data scientists, and about 1.5 million managers and analysts capable of posing the right questions and understanding the results [1]. Therefore, it is necessary to educate and train professionals to close this gap.

Unfortunately, potential scarcity is not the only challenge surrounding professional capacity. The growing use of data on human behavior means that there is less need for skills related to engineering, or the so-called hard sciences, and more need for multidisciplinary and multisectoral analysis in different social, demographic, and geographical contexts. In other words, professionals who can navigate diversity skillfully are in demand.

Enhancing Transparency in Evidence-generating Analytics

As mentioned above, data science requires making a series of decisions regarding data analyses, and working with debatable assumptions that can influence the evidence generated for decision making and the formulation of public policies. Therefore, it is essential to **document and ensure transparency in the analytical processes conducted**, so that they can be audited and accountability mechanisms can be applied. There are opportunities for constantly improving the analyses and the results, disseminating the methodologies in the public sector, and in particular, potentially correcting mistakes in a timely fashion. This is especially important in cases of leakage of personal information or inequalities that arise from biased recommendations based on an algorithm. In this regard, the experience of a new area of research called e-science [72], which focuses on the traceability and reproducibility of data-intensive research, can be useful for addressing this topic.

4.3. Opportunities

The remainder of the section summarizes the opportunities identified based on the challenges, recommendations, and analysis of the case studies.

Development (or maturity) Level of Big Data Projects and of "Smart Consumers" of Evidence from Big Data Analysis

The first opportunity lies in using the rubric designed by Townsend and Zambrano-Barragan [55] to assess the initiatives that use big data to address urban issues (**Annex A**). With minor adjustments, the rubric can be used to assess the overall degree of maturity of any big data analysis project in the public sector. In this case, the "citizen participation and public service" dimension can be swiftly adapted to the context of the project being evaluated. Likewise, the rubric should also account for the smart consumers (**Annex B**) to ensure that they meet the minimum competency levels required to interpret and use the evidence produced from big data analysis.

Data Sharing and Dissemination within the Public Sector

The second opportunity lies in the potential use of the data generated in public sector analytics projects for other purposes. For example, in the case of the *Você SB* cellphone application developed in São Bernardo do Campo (Brazil), the citizen-generated data can serve as valuable input for various government agencies responsible for security, the environment, and other areas.

There is also potential for synergy between different government agencies as they can conduct analysis together to shed light on the need for decision making and multisectoral policy formulation in the areas of transportation, environmental pollution, and the concentration of schools [73]. To tap into this potential, governments should move forward with policies regarding ongoing data generation and maturation, and put in place an institutional framework that **(1)** guides the use of big data analysis to create a culture of evidence-based decision making; **(2)** sustains the management and maintenance of big data with the necessary safeguards (see Section 4.1), and **(3)** promotes clear and smooth communication with other government agencies and external entities (such as universities and research centers).

Types of Issues to be Addressed

The third opportunity for using big data analytics lies in a specific type of pure prediction problems [60]. For these types of problems, it is not necessary to establish causality to make decisions in the typical assessment scenarios with or without the implementation of a public policy. Kleinberg

et al. [60] called these “umbrella problems,” which have to do with decision making. For example, is the chance of rain high enough to justify taking an umbrella? In this case, one does not need to know what causes rain; one just needs to estimate whether it will rain or not.

Therefore, to support decision making, one can use supervised machine learning techniques (see Table 2) and rely on past data to train the algorithm to produce a more accurate and timely forecast than a human expert can. Applications have been developed for a variety of purposes, such as:

- Socioeconomic profiling of a certain geographic area based on satellite information to determine area-specific social assistance policies [74].
- Estimating a student’s risk of dropping out [75], [76] and choosing the most cost-effective retention intervention [77].
- Improving audit policies by using predictive surveys based on online user reviews [78].

References

[1] J. Manyika *et al.*, “Big Data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, 2011.

[2] Staff Science, “Challenges and Opportunities,” *Science*, vol. 331, no. 6018, pp. 692–693, Nov. 2011.

[3] The Economist, “Data, data everywhere,” *The Economist*, Feb-2010.

[4] M. Hilbert and P. López, “The world’s technological capacity to store, communicate, and compute information,” *Science*, vol. 332, no. 6025, pp. 60–65, Apr. 2011.

[5] S. T. McAbee, R. S. Landis, and M. I. Burke, “Inductive reasoning: The promise of Big Data,” *Hum. Resour. Manag. Rev.*, 2016.

[6] J. Bertolucci, “Big Data’s new buzzword: datafication,” *InformationWeek*, 2013. [Online]. Available: http://www.informationweek.com/big-data/big-data-analytics/big-datas-new-buzzword-datafication/d/d-id/1108797?print=yes. [Accessed: 19-Dec-2016].

[7] C. Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,” *WIRED*, 23-Jun-2008. [Online]. Available: https://www.wired.com/2008/06/pb-theory/. [Accessed: 18-Dec-2016].

[8] D. Butler, “When Google got flu wrong,” *Nat. News*, vol. 494, no. 7436, p. 155, Feb. 2013.

[9] T. Harford, “Big data: are we making a big mistake?,” *Financial Times*, 28-Mar-2014. [Online]. Available: https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0. [Accessed: 19-Dec-2016].

[10] V. Strauss, “‘Big data’ was supposed to fix education. It didn’t. It’s time for ‘small data.,’” *The Washington Post*, 09-May-2016. [Online]. Available: https://www.washingtonpost.com/news/answer-sheet/wp/2016/05/09/big-data-was-supposed-to-fix-education-it-didnt-its-time-for-small-data/. [Accessed: 19-Dec-2016].

[11] J. Chen *et al.*, “Big Data challenge: a data management perspective,” *Front. Comput. Sci.*, vol. 7, no. 2, pp. 157–164, Apr. 2013.

[12] L. Manovich, “Trending: The promises and the challenges of big social data,” *Debates Digit. Humanit.*, vol. 2, pp. 460–475, 2011.

[13] M. R. Parks, “Big Data in communication research: Its contents and discontents,” *J. Commun.*, vol. 64, no. 2, pp. 355–360, Abril 2014.

[14] D. J. Power, “Using ‘Big Data’ for analytics and decision support,” *J. Decis. Syst.*, vol. 23, no. 2, pp. 222–228, April 2014.

[15] F. Provost and T. Fawcett, “Data science and its relationship to big data and data-driven decision making,” *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[16] A. Gandomi and M. Haider, “Beyond the hype: Big Data concepts, methods, and analytics,” *Int. J. Inf. Manag.*, vol. 35, no. 2, pp. 137–144, April 2015.

[17] R. Kitchin, “Big data and human geography: Opportunities, challenges and risks,” *Dialogues Hum. Geogr.*, vol. 3, no. 3, pp. 262–267, 2013.

[18] R. Kitchin, “Big Data, new epistemologies and paradigm shifts,” *Big Data Soc.*, vol. 1, no. 1, pp. 1–12, 2014.

[19] I.-Y. Song and Y. Zhu, “Big Data and data science: what should we teach?,” *Expert Syst.*, vol. 33, no. 4, pp. 364–373, August 2016.

[20] L. Tomar, W. Guicheney, H. Kyarisiima, and T. Zimani, “Big Data in the public sector: Selected applications and lessons learned,” Inter-American Development Bank, 2016.

[21] L. Hill, F. Levy, V. Kundra, B. Laki, and J. Smith, *Data-Driven Innovation for Growth and Well-being*. OECD, 2014.

[22] L. Rodríguez-Mazahua, C.-A. Rodríguez-Enríquez, J. L. Sánchez-Cervantes, J. Cervantes, J. L. García-Alcaraz, and G. Alor-Hernández, “A general perspective of Big Data: applications, tools, challenges and trends,” *J. Supercomput.*, vol. 72, pp. 3073–3113, 2015.

[23] B. F. Welles, “On minorities and outliers: The case for making Big Data small,” *Big Data Soc.*, vol. 1, no. 1, pp. 1–2, 2014.

[24] K. Nahon and J. Hemsley, *Going viral*, 1st ed. Polity Press, 2013.

[25] J. D. Morrison and J. D. Abraham, “Reasons for enthusiasm and caution regarding Big Data in applied selection research,” *Ind. Psychol.*, vol. 52, no. 3, pp. 134–139, 2015.

[26] L. Taylor, R. Schroeder, and E. Meyer, “Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?,” *Big Data Soc.*, vol. 1, no. 2, p. 2053951714536877, 2014.

[27] United Nations Global Pulse, “Big Data for development: opportunities & challenges”: A Global Pulse White Paper,” United Nations Global Pulse, 2012.

[28] D. Boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Inf. Commun. Soc.*, vol. 15, no. 5, pp. 662–679, 2012.

[29] F. Kreuter and R. D. Peng, “Extracting Information from Big Data: Issues of Measurement, Inference and Linkage,” in *Privacy, Big Data, and the Public Good*, J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, Eds. Cambridge University Press, 2014, pp. 257–275.

[30] J. A. Marsh, J. F. Pane, and L. S. Hamilton, “Making Sense of Data-Driven Decision Making in Education,” 2006. [Online]. Available: http://www.rand.org/pubs/occasional_papers/OP170.html. [Accessed: 28-Jan-2017].

[31] UNESCO, “Policy brief - Learning Analytics.” UNESCO Institute for Information Technologies in Education, 2012.

[32] B. Schmarzo, *Big Data: Understanding How Data Powers Big Business*. Wiley, 2013.

[33] A. Labrinidis and H. V. Jagadish, “Challenges and Opportunities with Big Data,” *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 2032–2033, 2012.

[34] A. E. Joseph and P. R. Bantock, “Measuring potential physical accessibility to general practitioners in rural areas: a method and case study,” *Soc. Sci. Med.*, vol. 16, pp. 85–90, 1982.

[35] W. Luo and Y. Qi, “An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians,” *Health Place*, vol. 15, no. 4, pp. 1100–1107, December 2009.

[36] W. Luo and F. Wang, “Measures of spatial accessibility to health care in a GIS environment: synthesis and a case study in the Chicago region,” *Environ. Plan. B Plan. Des.*, vol. 30, no. 6, pp. 865 – 884, 2003.

[37] J. Radke and L. Mu, “Spatial Decompositions, Modeling and Mapping Service Regions to Predict Access to Social Programs,” *Geogr. Inf. Sci.*, vol. 6, no. 2, pp. 105–112, December 2000.

[38] Z. Wei, “A study of accessibility to health facilities for elderly population in metro Atlanta using a categorical multi-step floating catchment area method,” Thesis, uga, 2013.

[39] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[40] L. Anselin, “Local Indicators of Spatial Association—LISA,” *Geogr. Anal.*, vol. 27, no. 2, pp. 93–115, 1995.

[41] H. J. Miller, “Tobler’s First Law and Spatial Analysis,” *Ann. Assoc. Am. Geogr.*, vol. 94, no. 2, pp. 284–289, Jun. 2004.

[42] J. I. Barredo, M. Kasanko, N. McCormick, and C. Lavalle, “Modelling dynamic spatial processes: simulation of urban future scenarios through cellular automata,” *Landsc. Urban Plan.*, vol. 64, no. 3, pp. 145–160, Jul. 2003.

[43] M. Bienkowski, M. Feng, and B. Means, “Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics.” U.S. Department of Education, Office of Educational Technology, 2012.

[44] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002.

[45] V. Friedman, “Data Visualization and Infographics | Smashing Magazine,” 2008. [Online]. Available: http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics/. [Accessed: 18-Jun-2014].

[46] M. Lima, *Visual Complexity: Mapping Patterns of Information*. Princeton Architectural Press, 2011.

[47] J. Steele and N. Iliinsky, Eds., *Beautiful Visualization: Looking at Data through the Eyes of Experts*, 1st ed. O’Reilly Media, 2010.

[48] N. Yau, *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*, 1st ed. Wiley, 2011.

[49] IBM, “What is a Data Scientist? – Bringing big data to the enterprise,” 2015. [Online]. Available: http://www-01.ibm.com/software/data/infosphere/data-scientist/?cm_mc_uid=23497733502814413056500&cm_mc_sid_50200000=1441305650. [Accessed: 07-Sep-2015].

[50] L. Floridi, “Big data and their epistemological challenge,” *Philos. Technol.*, pp. 1–3, 2012.

[51] K. M. Kelm *et al.*, “Big data innovation challenge : pioneering approaches to data-driven development,” The World Bank, 107751, Jan. 2016.

[52] W. MacLeod, J. Bor, K. Crawford, and S. Carmona, “Analysis of Big Data for better targeting of ART adherence strategies: spatial clustering analysis of viral load suppression by South African province, district, sub-district and facility (April 2014-March 2015),” The World Bank, 2015.

[53] S. Goldsmith, S. Crawford, and B. Weinryb Grohsgal, “Innovations in public service delivery - Issue No. 4: predictive analytics: driving improvements using data,” IDB Discussion Paper No. IDB-DP-440, Jul. 2016.

[54] H. Terraza, P. Deregibus, C. Galeota, and M. Ponce de León, “Movilidad urbana sostenible, datos masivos y políticas públicas: estudio de la movilidad de los ciclistas en la ciudad de Rosario (Argentina) a través del uso de dispositivos de geo-referenciación,” Inter-American Development Bank, 2016.

[55] A. Townsend and P. Zambrano-Barragan, “Computing a new trajectory for urban gobernance: Urban Big Data innovation in Latin America and the Caribbean,” Inter-American Development Bank, 2016.

[56] B. Means, C. Padilla, A. DeBarger, and M. Bakia, *Implementing Data-Informed Decision Making in Schools: Teacher Access, Supports and Use*. US Department of Education, 2009.

[57] D. Comin, “Economic Growth,” in *Economic Growth*, S. N. Durlauf and L. E. Blume, Eds. Palgrave Macmillan UK, 2010, pp. 260–263.

[58] Bureau van Dijk, “Orbis | Detailed global private company information,” 2017. [Online]. Available: http://www.bvdinfo.com/en-gb/our-products/company-information/international-products/orbis. [Accessed: 12-Jan-2017].

[59] D. Bahar, “Using firm-level data to study growth and dispersion in total factor productivity,” The Brookings Institution Harvard Center for International Development, 2016.

[60] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, “Prediction policy problems,” *Am. Econ. Rev.*, vol. 105, no. 5, pp. 491–495, 2015.

[61] ECLAC “Estado de la banda ancha en América Latina y el Caribe 2016,” ECLAC, Sep. 2016.

[62] J. Howison, A. Wiggins, and K. Crowston, “Validity Issues in the Use of Social Network Analysis with Digital Trace Data,” *J. Assoc. Inf. Syst.*, vol. 12, no. 12, Dec. 2011.

[63] J. Kobielus, “Data Scientist: Master the Basics, Avoid The Most Common Mistakes,” *IBM Data&Analytics Hub*, 02-Jul-2013. .

[64] K. Crawford, “The hidden biases in Big Data,” *Harvard Business Review*, 01-Apr-2013.

[65] OECD, *Equity and Quality in Education*. Paris: Organisation for Economic Co-operation and Development, 2012.

[66] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” *Science*, vol. 343, no. 6176, pp. 1203–1205, Mar. 2014.

[67] C. Miller, “When Algorithms Discriminate,” *The New York Times*, 09-Jul-2015.

[68] Nature, “More accountability for big-data algorithms,” *Nat. News*, vol. 537, no. 7621, p. 449, Sep. 2016.

[69] L. Sweeney, “Simple Demographics Often Identify People Uniquely,” Pittsburgh, 2000.

[70] J. Bahamonde, A. Hevia, G. Font, J. Bustos-Jiménez, and C. Montero, “Mining Private Information from Public Data: The Transantiago Case,” *IEEE Pervasive Comput.*, vol. 13, no. 2, pp. 37–43, Abril 2014.

[71] European Big Data Value Partnership, “European Big Data value strategic research & innovation agenda: Version 0.99,” Jul. 2014.

[72] T. Hey, S. Tansley, and K. Tolle, “Jim Gray on eScience: A transformed scientific method,” *Fourth Paradigm Data-Intensive Sci. Discov.*, vol. 1, 2009.

[73] P. Rodríguez *et al.*, “Apoyando la formulación de políticas públicas y toma de decisiones en educación utilizando técnicas de análisis de datos masivos: el caso de Chile,” 2016.

[74] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, “Combining satellite imagery and machine learning to predict poverty,” *Science*, vol. 353, no. 6301, pp. 790–794, Aug. 2016.

[75] C. Escobar and F. Lolas, “Desarrollo de un sistema prototipo para la detección temprana de la deserción escolar en escuelas públicas chilenas",” Memoria de Título, Universidad Adolfo Ibáñez, Santiago de Chile, 2015.

[76] Mineduc, “Informe de Piloto de Modelo Predictivo, Seguimiento de Estrategias de Apoyo (Sistema de Alerta Temprana).” Jul-2015.

[77] Microsoft, “Predicting student dropout risks, increasing graduation rates with cloud analytics,” Aug-2016. [Online]. Available: https://customers.microsoft.com/en-us/story/tacomapublicschoolsstory. [Accessed: 14-Dec-2016].

[78] J. S. Kang, P. Kuznetsova, M. Luca, and Y. Choi, “Where not to eat? Improving public policy by predicting hygiene inspections using online reviews.,” in *EMNLP*, 2013, pp. 1443–1448.

[79] P. Rodríguez and J. Mondaca, “Rúbrica de competencias de los consumidores inteligentes de inteligencia de valor público.” 2016.

[80] G. Bellinger, D. Castro, and A. Mills, “Data, information, knowledge, and wisdom,” 2004.

[81] M. O. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications*, 1 edition. Natick, Mass: A K Peters/CRC Press, 2010.

[82] P. R. Keller and M. M. Keller, *Visual cues: practical data visualization*, vol. 2. IEEE Computer Society Press Los Alamitos, CA, 1993.

Annex A: Maturity Model Rubric for Big Data Use in Addressing Urban Issues

Maturity level	Open data	Cultivating data ecosystems	Analytics	Data-based decision making	Participation and public services
Level 5 — Optimizing (Smart urban collaboration)	Read-write platforms empower user community curation and extension of data; governance protocols embedded in software enable responsible sharing.	Industry, academia, government, and citizens sharing trusted data; data marketplaces create safe, secure platform for many-to-many exchange of big urban data.	Open analytics platforms enable rapid innovation in algorithms; self-optimization of operations through extensive automation of analytics.	The organization and its operations continuously adapt and improve using analytical insight in line with strategic policy objectives; processes that require modest human judgement are subject to automation.	Citizen-driven vision and governance innovation agenda; open innovation platforms for data-driven public services, shared data-enabled governance.
Level 4 — Advanced (Smart administration)	All non-sensitive data published openly, with robust data user community support and dataset request processes exist.	Most useful data are 'big'; crowd-sourcing data collection widespread; external data exchange with private sector; incentives for data sharing are commonplace.	Predictive analytics is used widely to optimize the organization's decision-making so that the best actions are taken to maximize operational effectiveness and achieve policy outcomes.	Decision makers are well informed with insight from analytics and the organization is capable of acting to maximize key performance indicators; processes that require little human judgement are automated.	Thorough citywide service integration with pockets of citizen prosumers driving service innovation; robust cross-department innovation management.
Level 3 — Intermediate (Smart decision-making	Open data policy and regulation mandates timetable for comprehensive data disclosure, subject to security and privacy review; real-time data published when feasible.	Integrated sensor networks supporting multiple users; data platforms enable automated sharing; mash-ups from diverse sources.	Predictive analytics provide insight on the likelihood of important changes in activity patterns affecting the organization's operations or policies; accelerating improvements through machine learning and other techniques.	The organization is able to make limited business decisions using analytical insight to improve operational efficiency and generate more value; data dashboards support a data-driven culture.	City-initiated vision, strategy, and implementation for data-driven participation and public services; integrated delivery platform incorporates citizen feedback loops.
Level 2 — Basic (Government of a smart city)	Open data portal aggregates published government datasets.	Application-specific sensor networks collect relevant data; policies for data privacy, security and sharing established; data quality is poor; cross-linking requires time-consuming manual integration.	Analytics are used to inform decisionmakers about the causes and contributing factors for key processes and events in the organization's operations.	The organization understands the causes behind what they observe, buts its culture is largely resistant to adapting to take advantage of the insight.	Pockets of public services innovation, with some integration and cross-department data sharing; limited citizen engagement.
Level 1 — Ad Hoc	Data sharing enabled through scattershot regulations and departmental policy.	Agencies rely on historical data exhaust from operations; data is in silos with little sharing.	Analytics are limited to describing what has happened.	The application of analytical insight is the choice of the individual and has little effect on how the organization operates.	Little data-enabled engagement or use of data in governance or service delivery; digital public services strategies do not exist or are isolated.

Source: Townsend and Zambrano-Barragan [55].

Annex B: Competency Rubric for Smart Consumers of Public Value Intelligence [79]

Objectives

The rubric aims to present the general criteria for the end user of the outcomes of public value intelligence projects to assess their own capacity to approach, understand, and ultimately make decisions based on the information available.

Conceptual Framework

Although the literature differentiates between different types of literacies, "literacy" for the purposes of this rubric refers to information literacy. Information is defined from the perspective of computer science [80], which differentiates between data and information. In this regard, "data" refers to the "raw" values (that may or may not be organized into a matrix or a database) that have not gone through any type of processing or analysis. Therefore, having data is not enough for drawing conclusions or making decisions. Information, on the other hand, has been processed in some manner, and can be used to support decisions, judgment, or conclusions. Since the end user will have access to information and not data, the rubric will focus on information literacy.

The design of the rubric contemplated two types of elements: first, the components, concepts, and skills described by Means et al. [56] for assessing data-based decision-making capacity; and second, the characteristics and tasks that should be clarified for users of visualized information as discussed by Ward et al. [81].

The instrument designed by Means et al. [56] has five components[7]:

1. **Framing the question**: the question should be answerable based on the available data. There has to be a semantic relationship between the question raised and the structure of the data.
2. **Finding the data**: identifying the relevant data for answering the question.
3. **Understanding the format**: getting familiarized with the way in which the data are presented (for example, table or graph), which helps structure a response that addresses the question raised.
4. **Interpreting the data**: knowing the statistical concepts—at least qualitatively—to understand the data behavior. For example, one or more atypical values can skew the average value of a certain measurement.
5. **Using the data**: applying the data to specific, appropriate contexts, and understanding how data reflect or attempt to capture a given occurrence.

Even though Means et al. [56] use the term "data" when they refer to what is measured by the instrument, the concept encompasses data processing and information production from the data.

Generally speaking, the purpose of the rubric is to evaluate information and not data, as the deliverables are processed and hence limited. On one hand, this means that the end user has no or little involvement in data processing (except to make the format more aesthetically pleasing at best). On the other hand, the framing of the question is constrained by the limits of the information provided.

[7] To keep it consistent with the definitions used by Means et al. [56], these components will also focus on data and not information.

Ward et al. [81] have defined the tasks for and the characteristics of the users of visualization. The tasks refer to the actions required to make a valid judgment based on the visualization presented. The following list draws from the actions outlined in [82]:

1. **Identify**: recognizing the relevant elements in the tool presented
2. **Locate**: finding the position of a given item in the tool presented
3. **Distinguish**: determining if an element is different than another
4. **Categorize**: classifying different types of elements
5. **Group**: dividing elements into groups based on relationships or common features.
6. **Prioritize:** organizing a set of elements in a particular order.
7. **Compare**: examining the similarities and differences between two or more elements.
8. **Associate**: building a relationship between two or more elements.
9. **Correlate**: establishing a two-way relationship between two or more elements.

These tasks help define the different levels of expertise required of the user in order to work with visualizations. These tasks will be implemented through using the descriptors in the visualization rubric.

According to Ward et al [81], there are five characteristics related to data knowledge and skills[8] that users must possess in order to adequately understand the information. These are:

1. **Familiarity with the field**: the user's expertise in and knowledge of the field or the context of the data in which they operate.
2. **Familiarity with the task**: the user's experience with the task assigned.
3. **Familiarity with the data**: the user's experience with the data from which the information was obtained, and whether or not the user has come up with a mental framework to make sense of the data.
4. **Familiarity with the techniques of visualization**: the user's experience with and expertise in specific visualization techniques.
5. **Familiarity with the context of visualization**: the user's knowledge of and expertise in the visualization tools used to present information.

Components

Some of the components proposed in [56] would overlap if used in the specific context of the rubric. For example, the component of "framing the question" would overlap with "using the data," as the user should know and understand the context of the data use to frame a question that is relevant and can be answered with the data. The components of "finding the data" and "understanding the format" also have elements in common—if the user cannot work with the data, they cannot adjust the formats in which to present the information. In other words, "finding the data" would depend on familiarity with the format in which the information is presented. Therefore, based on the recommendations made in [56], the paper proposes three components for the rubric:

1. **Identifying the information**: finding the information regardless of how it is presented requires recognizing the type of information presented and its applicable contexts.
2. **Understanding the information**: familiarity with the format in which the information is presented and the skills required to understand it; this requires understanding the ways in

[8] Same as Means et al. [56], Ward et al. [81] also focus on data.

which information is presented, finding the relevant information, and comparing its elements, among others.

3. **Interpreting the information**: using information from sources other than the ones presented to better understand the available information; this requires knowledge of the concepts, methodologies, and statistics to go beyond a superficial understanding and to verify the validity of the conclusions drawn from the information.

Descriptors, Skills, and Knowledge

This section defines two elements: first, the range of descriptors that capture the skills and knowledge needed for each component; second, the skills and knowledge themselves. The definitions of both draw from the points made by Ward et al. (2010).

1. **Identifying the information**: the descriptors capture different levels ranging from having no knowledge to having an in-depth knowledge of the information, with the latter meaning that the user can establish logical connections between the pieces of information and with other relevant knowledge. Ward et al. (2010) have defined a select range of user characteristics as regards the knowledge required. Not all the characteristics are mentioned here, since some overlap with the elements related to the component of "understanding the information," so the characteristics are best kept separate to keep things organized:
    a. **Familiarity with the field**: the user's expertise in and knowledge of the field or the context of the information received.
    b. **Familiarity with the data**: the user's experience with the data from which the information was obtained, and whether or not the user has come up with a mental framework to make sense out of the data.
    c. **Familiarity with the context of visualization**: the user's knowledge of and expertise in the tool(s) of visualization with which information is presented.
2. **Understanding the information**: the descriptors related to the comprehension of information are defined based on the tasks outlined by Ward et al. (2010). For the purposes of the rubric, these descriptors are divided into three groups. The first group addresses the capacity to identify, find, and distinguish between pieces of information. The second group addresses the capacity to categorize, group, and prioritize information. The third group addresses the capacity to compare, relate, and correlate information.
    a. **Basic visualization**: the most common type of visualization—for example, contingency tables, bar graphs, among others.
    b. **Multidimensional visualization**: similar to basic visualization but with more than three dimensions. Therefore, one should take certain special considerations into account when designing this type of visualization to facilitate comprehension.
    c. **Geospatial visualization**: visualizations that reference geographic areas; this requires familiarity with the maps and with the areas to be represented.
3. **Interpreting the information**: as in the component of "identifying the information," the descriptors for "interpreting the information" range from superficial knowledge or literal comprehension of the concepts to the capacity to process and critically evaluate information as well as to understand the information presented through visualization in light of external information.

a. **Concepts**: all the notions that the user should be familiar with and have a certain degree of expertise in, to take the correct approach to the information presented.

b. **Statistics**: the basic notions of descriptive statistics, such as measures of central tendency, measures of position, and measures of spread, which the user should be familiar with in order to understand information and check its validity.

Rubric

The following is a rubric for the dimension of **identifying the information**. The expected level – in terms of skills or knowledge – of the end user is highlighted in **grey.**

Table B1: Rubric for the Dimension of "Identifying the Information"

Skill or knowledge	Descriptors			
	None	**Basic**	**Intermediate**	**Advanced**
Familiarity with the field	The user's understanding is below the threshold required for the field of analysis.	The user has a general but superficial understanding of the field, without detailed knowledge of the functional and structural aspects that can help them critically assess the information in visualizations.	The user has adequate knowledge of the field. The user has sufficient knowledge of the functional and structural aspects to understand the information in visualizations.	The user has expert knowledge of the field. The user understands the functional and structural aspects in extensive detail, and can thus critically assess the information presented in different visualizations.
Familiarity with the data	The user has no data-related knowledge, does not know the data source or the valid context for application.	The user has minimal knowledge of the data source, the context of application, and the implications. The user is not able to assess the validity or the relevance of data in the visualizations.	The user is familiar with some of the data and their sources, knows the objectives for which the data were produced, their contexts of application, and implications for the system. The user can assess the relevance of some of the data in the visualizations.	The user has expert knowledge of the different types of data and their sources, knows and understands the objectives for which the data were produced, their contexts of application, and their implications for the educational system. The user is capable of critically assessing the relevance of the data in the visualizations.
Familiarity with the context of visualization	The user does not know the context of visualization used nor similar contexts.	The user knows the visualization contexts similar to the one used, and can conduct minimal processing of the information presented.	The user knows the context of visualization, but has difficulty in processing the information presented therein.	The user is familiar with the context of the visualization used, and can process the information effectively.

The following is a rubric for the dimension of **understanding the information.** The expected level—in terms of skills or knowledge—of the end user is highlighted in **grey.**

Table B2: Rubric for the Dimension of "Understanding the Information"

Skill or knowledge	Descriptors			
	None	**Basic**	**Intermediate**	**Advanced**
Basic visualization	The user cannot differentiate between different types of basic visualization, nor between basic visualization and other types of information visualization.	The user can identify some basic visualizations, recognize and find the relevant elements in each type, and distinguish between these elements.	The user can categorize the elements of a basic visualization into different types, and organize or sort them based on common criteria. The user also knows when it is appropriate to use the abovementioned techniques.	The user has advanced knowledge and can thus compare, relate, and/or correlate the various elements within and between basic visualizations of similar characteristics.
Multidimensional visualization	The user cannot differentiate between different types of multidimensional visualizations or between multidimensional visualizations and other types of information visualization.	The user can identify some multidimensional visualizations and distinguish them from basic visualizations. The user can also recognize and find the relevant elements in each visualization and differentiate between these elements.	The user can categorize the elements of a multidimensional visualization into different types and organize or sort them based on common criteria. The user also knows when it is appropriate to use the abovementioned techniques.	The user has advanced knowledge and can thus compare, relate, and/or correlate the various elements within and between multidimensional visualizations of similar characteristics. The user can propose adjustments to visualizations based on need or requirement for decision making.
Geospatial visualization	The user cannot differentiate between different types of geospatial visualizations or between geospatial visualizations and other types of information visualization.	The user can identify some geospatial visualizations and their elements (such as "street blocks"). The user can also recognize and find the relevant elements in each visualization and differentiate between these elements.	The user can categorize the elements of a geospatial visualization into different types and organize and/or sort them based on common criteria. The user can also understand geographical indicators visually and knows when it is appropriate to use	The user has advanced knowledge (for example, of how to use mathematical-geospatial models) and can thus compare, relate, and/or correlate the various elements of a geospatial visualization, and relate it to other visualizations. The user can also

Skill or knowledge	Descriptors			
	None	Basic	Intermediate	Advanced
			the abovementioned techniques.	propose adjustments to visualizations based on need or requirement for decision making.

The following is a rubric for the dimension of **interpreting the information.** The expected level – in terms of skills or knowledge – of the end user is highlighted in **grey.**

Table B3: Rubric for the Dimension of "Interpreting the Information"

Skill or knowledge	Descriptors			
	None	Basic	Intermediate	Advanced
Concepts	The user does not understand the most basic concepts needed to make sense of the information presented in visualizations.	The user has minimal conceptual knowledge and can understand the information in visualizations.	The user has sufficient conceptual knowledge, which complements their understanding of the visualizations.	The user knows about topics that are not necessarily related to the information but help enrich their understanding from the perspective of fields other than their specialization.
Statistics	The user does not understand the most basic notions of descriptive statistics, and therefore cannot interpret the information presented in visualizations.	The user knows very basic notions such as average, which is not enough for understanding the information in visualizations.	The user understands the notions of descriptive statistics, such as measures of central tendency, spread, and position, as well as basic notions of probability, and can thus understand and expand on the information in visualizations.	The user understands the notions and techniques of advanced statistics and data mining, and can thus not only understand the information but also assess it critically from a statistical perspective.