14

Data Valuation and Its Applications for Smart Cities

Mihnea Tufiş

Eurecat Technology Center of Catalunya, Carrer de Bilbao, Barcelona
Email: mihnea.tufis@eurecat.org

Abstract

The global economy is increasingly more reliant on data, with businesses adopting data-enabled decision-making practices in the form of analytics or machine learning. Views of data as an asset and the steady emergence of data markets depend on the capability of quantifying the value of data. We argue that the data-as-an-asset approach and focusing on assigning a price tag for data is complicated, due to the properties of data, the multiple value chains that it can generate, as well as legal and ethical implications. We introduce a data valuation process that recognises and integrates the contextual nature of data value, together with data quality and data utility assessments. The value of data is reported in a multi-faceted scorecard, which allows for an exploration of data value at different levels of aggregation. We explore how cities can benefit from the multitude of data they harvest in the process of digitalisation, and we argue that these benefits can be enhanced if cities were to have a more concrete understanding of the value of their data. We discuss their multiple roles with respect to big data processing, as producers, consumers, regulators, and educators for their citizens, and conclude with a list of data-centred actions that cities should implement as part of their smart city agenda.

14.1 Introduction

The Covid-19 pandemic and the current geopolitical changes are bringing a sense of urgency about the actions needed to tackle challenges raised by

older and deeper transformations – climate change, worldwide demographic dynamics, technological advancements, and their impact on the job market.

With increased pressure on central authorities, local governments may just have the opportunity to assume a more hands-on role in dealing with these challenges. Local administrations are best positioned to understand local challenges; they are able to design in terms of local specificities, they can develop a direct link to their citizens and businesses, and they can seek to create local and regional partnerships with other communities that share the same characteristics, challenges, or commitments.

The role of ICT solutions in achieving goals specific to smart city transformation is now indisputable. For example, ICT-based solutions could reduce commuting by 15%–20% and cut greenhouse gas emissions by 10%–15%,[1] and this at a rate of only 12% of city data currently being analysed and used for decision-making and management.[2] A special class is the big data solution, built around a backbone that typically involves massive data collection, processing, and analysis at scale; further downstream, these data are used by more advanced machine learning and artificial intelligence solutions to train autonomous systems for knowledge representation, reasoning, and decision-making. These solutions have the potential to impact every area in the life of cities and their citizens: infrastructure planning, mobility, land and district planning, energy, demographics, social inclusion, development of and engagement with local businesses, culture and entertainment, tourism, health services, and government.

Data scientists know too well that data is key to the success of every solution. Performing a significant analysis or training a reliable decision-making model has instant demands concerning data quantity and quality. And as soon as they step out of open and curated datasets, they quickly run into issues related to availability, interoperability, bias, and many more.

The discussion about data value cannot start without observing the change in the data production–consumption cycle, a consequence of the big data "revolution". Historically, data was produced when it was needed, tailored to the needs of those who would use it, and often consumed precisely by the same actors who enabled its generation. For example, a scientist who

---

[1] Woetzel, J., Remes, J., Boland, B., Lv, K., Sinha, S., Strube, G., Means, J., Law, J., Cadena, A., & von der Tann, V. (2018). Smart Cities: Digital Solutions for a More Liveable Future. McKinsey Global Institute.

[2] Gualtieri, M., & Yuhanna, N. (2014). *The Forrester Wave: Big Data Hadoop Solutions, Q1 2014*. Forrester. https://www.forrester.com/report/The-Forrester-Wave-Big-Data-Hadoop-Solutions-Q1-2014/RES112461

would need to measure the levels of a pollutant in a river would design the data requirements and model, organise, and perform the data collection and eventually process and analyse the collected data. Today, big data refers to massive, continuous, and often loosely structured data, a large portion of which is a by-product of activities, behaviours, or processes that are not always the primarily intended focus of data observations. We have shifted from recording data about a selected subset of activities, to generating data about nearly every aspect of our lives.

The success of a data-centred initiative lies in unlocking the value that data can generate in each context and with respect to the problem being solved. The current paradigm for data production and consumption leads to the perception of data as an asset, subject to exchange, the subsequent appearance of new stakeholders whose activity is based on the acquisition, re-packaging, and selling of datasets, and, finally, the steady emergence of data markets. In this context, a question is becoming increasingly pervasive: *what is the value of my data*?

Thus, the necessity to develop a process for establishing the value of data arises. Ideally, such a process should generalise to any kind of data, application domain, or economic sector. In the context of smart cities, its benefits would be manifold:

- Municipalities could understand the value chains generated by the data they are collecting.
- Consequently, they would be able to map the data value chains to practical outcomes and quantify their impact in the communities.
- Some of these data – depending on ethical implications – could even be exchanged in the emerging data markets and, thus, become a source of revenue for the communities and their individuals.
- A transparent methodology for data valuation could help develop a fair and responsible ecosystem around data markets.
- Participating in the creation of data spaces and digital federations would increase the cooperation between cities within the same region or dealing with similar challenges.
- Educate citizens on the real power of their personal data, its role in today's digital world, and empower them to have more control over this personal aspect of their lives.

In this chapter, we begin a discussion about the value of data: how to define it, what are its main drivers, and why is it difficult to establish it? We also

introduce the data valuation process and the data valuation component, the results of a three-year-long R&D project. Then, each section will try to connect to the area of smart cities, by exploring the bidirectional relationship between cities and data value – how each aspect of data valuation can help cities in their digital transformations as well as how cities, through the diversity of the data they collect and challenges that they present, can inform and improve data valuation methodologies.

14.2 Defining the Value of Data

14.2.1 Data through an economic lens – trading data

There is an interesting comparison that is usually made when illustrating the economic value of data and the difficulty of estimating it, especially when perceiving data as an economic asset: "*Facebook is now worth about $200 billion. United Airlines, a company that actually owns things like aeroplanes and has licenses to lucrative things like airport facilities and transoceanic routes between the U.S. and Asia, among other places, is worth $34 billion*".[3]

The view of data as a commodity gathered momentum with the advent of targeted online advertisement and its reliance on personal data. The "classic" model for (personal) data exchanges is for data-centric companies to offer a so-called "free" service in exchange for the users' personal data – the famous "if it's free, then you're the product". With the data deluge from the past decade and the gradual shift of businesses towards data-driven decision making, a gap appeared between their new aspirations and their data know-how. This has created the opportunity for a new group of stakeholders – data brokers – to join in an already unbalanced ecosystem. These intermediary enterprises exist "solely to collect personal data and sell it as a commodity to retailers, advertisers, marketers, even other data brokerages and government agencies".[4]

It is important to make the distinction between personal and non-personal data. According to Article 4 of the EU General Data Protection Regulation (GDPR),[5] personal data refers to information relating to natural

[3] Baldwin, H. (2015). Drilling Into The Value Of Data. Forbes. Retrieved from https://tinyurl.com/3jytwus9

[4] Madsbjerg, S. (2017). It's Time to Tax Companies for Using Our Personal Data. The New York Times. Retrieved from https://tinyurl.com/3dcpzrvt

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council – Of 27 April 2016 – On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection

persons who can be directly or indirectly identified from the data in question.[6] Personal data raises many additional challenges of social, legal, and ethical nature: what ownership model to adopt, should we even adopt one – since this would involve selling a form of identity,[7] and how do we adapt to different legal frameworks and different interpretations of privacy across cultures, to enumerate just a few.

Data brokers add value to personal data which individuals generate or release during various online activities, by analysing it, aggregating it, generating user profiles, and enriching it with valuable (and often free) data compiled by the National Statistics Organisations (NSO). It is these bundles of repackaged data that are then sold back to different companies to power their data-centred use cases. The global data broker market size was around $246 billion in 2020 and is expected to grow to $365 billion by 2027.

The most recent evolution of data brokers comes in the shape of online platforms for monetising personal data. These platforms claim to be giving back to individuals the control over personal data and enable them to sell it themselves, ideally, by choosing what data and to whom. There does not seem to be much separating these platforms from large data brokers (and, in fact, there is nothing to prevent such a platform from growing into one), but where they do set themselves apart is that they acknowledge the value of personal data and are open to sharing a piece of the revenues with those who generate it.[8]

Finally, this complex landscape is completed by the presence of governments, trying to find their role within it, depending on their degree of understanding of today's digital transformations. First, governments are expected to assume a regulatory role with respect to data exchanges in general, and data brokers in particular. In Austria, a reported discussion about applying VAT on revenues resulting from big data transactions by social media companies was abandoned, citing difficulties in assigning a value to such a transaction.[32]

Regulation), no. Regulation (EU) 2016/679, European Parliament, 88 (2016). Available at https://tinyurl.com/mr3sxrm5

[6] Such information can refer to "*an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*".

[7] Renieris, E. M., and Greenwood, D. (2018). Do we really want to "sell" ourselves? The risks of a property law paradigm for personal data ownership. Medium. Retrieved from https://tinyurl.com/272uxk78

[8] Tufiş, M. and Boratto, L.. (2021). Toward a Complete Data Valuation Process. Challenges of Personal Data. J. Data and Information Quality 13, 4, Article 20 (December 2021), https://doi.org/10.1145/3447269

Similarly, the United States Senate started holding hearings with respect to the DASHBOARD Act (Designing Accounting Safeguards to Help Broaden Oversight and Regulation of Data), a piece of legislation designed to protect individuals' privacy by forcing companies to disclose to the users the "true value" of the data that concerns them.[9]

Beyond its role as a regulator, there are instances in which governments seek to act as a data broker itself. In 2014, citing the abundance of data it amasses, the UK ministers attempted to pass legislation that would allow HM Revenue & Customs (HMRC) to sell anonymised taxpayers' data to third parties. This came under harsh scrutiny since the British government's track record in terms of data security and data anonymisation practices is far from clean.[10] More worrisome is that despite restrictions and criticism, the HMRC went ahead and quietly released VAT registration data "for research purposes" to three private credit rating agencies (Experian, Equifax, and Dun & Bradstreet).

Establishing an equitable relationship with data brokers will be a challenge for cities in their quest to become smart through digitalisation. Cities may find themselves in a position similar to that of big data companies, in the sense that they are both able to generate as well as consume a large amount of data, some of which is behavioural and often personal. However, as opposed to big tech companies, cities are not primarily run for profit; a city's goal should be the wellbeing of its citizens, and following such principles, it can set it up as a new type of actor in the data exchange landscape – one that generates value through a responsible use of its citizens' data and redistributes this value back to them to improve the livelihood of the community on which it is built upon.

14.2.2 The price of personal data – a chaotic landscape

There is a wide range of personal data collected by data brokers: identification, demographic, location, behavioural, online activity, psychological, product, and political preferences. Most of the times, these data are sold in bundles, which prompts several questions: are all these equally important to a buyer, are they equally sensitive for a seller, and how do each of these stakeholders value them? A reward as low as one cent a month for sharing

[9] S.1951 - 116th Congress (2019-2020): Designing Accounting Safeguards To Help Broaden Oversight and Regulations on Data (2019/2020).

[10] Mason, R. (2014). HMRC to sell taxpayers' financial data. The Guardian. Retrieved from https://tinyurl.com/4x4xcptp

exclusively location data might not convince a user to give it away; however, a bundle of several data types that can amount to as high as $100/month could prompt individuals to invest time in building, managing, and selling personal data portfolios. A second observation concerns the wide price range at which the same type of data is sold. For example, Luth Research pays $100/month for a bundle containing location, social media activity, and browsing activity,[11] whereas Datacoup used to pay $8/month for a similar package.[12] We believe this discrepancy is due to the lack of established data markets, data trading rules, and, as we will see next, a significant gap between the monetary value expected by individuals and what is actually paid by data brokers.

There are also examples of good practices in terms of dealing with user data. Wibson Data Market[13] is trying to enforce transparency, by stating who the data is generated for and for what purpose. Spanish company Telefónica proposes the establishment of a data bank which allows their service users to log all their activities on the network; this is somewhat similar to AT&T's Gigabit, with the major difference being that the former would give users full control over their data,[14] whereas the latter would charge them an extra $29 for keeping their data private.

A study by another telecom giant, Orange, covering 2023 mobile phone users balanced across age categories and countries of origin (France, Poland, Spain, and UK),[15] suggests the existence of three factors that influence the perceived value of personal data:

1. the usefulness of the data to the beneficiary organisation;
2. the type of data;
3. the risk associated with sharing it.

The study also underlines that users are aware that their data is valuable to organisations, which can benefit from it and reveals an ordering relationship of how likely they are to share types of personal data (demographic $\succ$ activity

[11] Ross, W. (2014). Is Your Smartphone Privacy Worth $100 a Month? MIT Technology Review. Retrieved October 18, 2019, from https://tinyurl.com/mrv9czyp

[12] Simonite, T. (2013). Coming Soon: Take Your Own Personal Data to Market. MIT Technology Review. Retrieved from https://tinyurl.com/ksf68nwa

[13] Travizano, M., Sarraute, C., Ajzenman, G., & Minnoni, M. (2018). *Wibson: A Decentralized Data Marketplace* (arXiv:1812.09966).

[14] At the time of submitting this text, there are no mentions as to what the price of such a service would be and how would Telefónica benefit from it.

[15] Loudhouse. (2014). The Future of Digital Trust. A European study on the nature of consumer trust and personal data (Industry No. 2; The Future of Digital Trust, p. 7). Orange.

and behavioural ≻ third-party or financial data).[16] The study also points to a paradox in consumers' understanding of sharing personal data: while a majority of respondents (77%) declare that privacy and transparency of data usage are important and identify the risk attached to sharing as an important factor influencing data value, they also indicate demographic data as the type they would most likely share – despite the clear risk of identity theft and online fraud attached to it.[17]

In a 2016 survey, credit comparison site Totally Money[18] asked 1000 UK consumers to estimate[19] the economic value of different categories of personal data. The results revealed interesting attitudes and different data-sharing practices, spread across demographic groups and types of data alike: young people (18–24 years old) value their data the most, while millennials value theirs the least (£1773); men value data about their online activity higher than women do (£1112 vs. £859 for email data, £1056 vs. £817 for browsing data, and £951 vs. £778 for location data). Perhaps the most surprising result of this study is the difference between the average self-estimate of respondents' personal data (£2031) and how much brokers are paying for it (£0.45). While the methodology of this study is not completely clear and it is difficult to assess the representativeness of the sample, the magnitude of this difference, together with the paradox observed by Smith,[17] points to the necessity of building "digital literacy", together with legal frameworks suitable to the consequences of the permeability of our digital traces and to the ease with which data companies can process and monetise them.

Interesting results are also coming from academia, with a recent increase in the study of methods for valuing user-generated data, particularly geolocation and online behaviour. In one of the most relevant experiments, Staiano *et al*.[20] simulated a data market for personal data transactions. Participants

---

[16] Third party data: email, personal preferences of other contacts; Behavioural data: location, mobile purchase history; Demographic data: name, date of birth, phone number.

[17] Smith, M. (2016). Proximus starts selling customer data reports for €700 a time. European Communications. Retrieved from https://tinyurl.com/muwm2ncd

[18] Davies, J. (2016). Consumers price their data at £2k – Companies pay 45p. Telecoms. Com. Retrieved from https://tinyurl.com/ysta7zwb

[19] TotallyMoney.com conducted research in June 2016 to identify the prices third-party companies pay for data to utilise in marketing campaigns: Financial Times, The Telegraph, McAffee, CostOwl.com, OnePoll.com.

[20] Staiano, J., Oliver, N., Lepri, B., de Oliveira, R., Caraviello, M., and Sebe, N. (2014). Money Walks: A Human-Centric Study on the Economics of Personal Mobile Data. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct, 583–594.

were equipped with devices gathering various types of data (calls, applications usage, location, and media usage) at three levels of aggregation (individuals, processed, and aggregated). They were then expected to sell the data to the Research Laboratory during auctions (reverse second price[21]), initially running weekly and then daily. In the nearly 600 auctions organised, participants received rewards totalling approximately €270, with a median price of €2 across categories. The auctions were also able to cast a light into the self-valuation of personal data, and just like in the study conducted by Orange,[22] it revealed an order of perceived value among the data types: location ≻ communication ≻ apps ≻ media; not surprisingly, processed data was held to a higher value than raw data. Two additional observations may be important take-aways when designing data valuation methods:

1. Increasing the frequency of auctions (from weekly to daily), decreased the value of the bids; an indication that the data market may play by the rules of supply and demand.

2. The value of data increased when unexpected situations arose (traffic jams caused by either a weather event or a local holiday); this suggests that the value of the same data is highly dependent on the context.

14.2.3 Challenges defining the value of data – beyond financial value

Until now, attempts at establishing the value of data were connected to large impact business events – mergers and acquisitions, bankruptcy, data transactions, and data breaches – and usually focused on the monetary value of data. This is perhaps why comparisons between data and other commodities (oil, gold, etc.) are usually making media headlines or are thrown in as a hook in conversations on the topic. We generally understand that there is value in data, mostly by connecting its applications to immediate outcomes and benefits, but it is very unclear what the source of this value is, the mechanisms through which it is created, and what it consists of (beyond the obvious monetary aspect).

Short and Todd[23] consider the value of data as the composite between the value of the asset itself, the value resulting from its use, and its expected or

[21] The lowest bidder wins, but the reward will equal the second-lowest bid.

[22] Loudhouse. (2014). The Future of Digital Trust. A European study on the nature of consumer trust and personal data (Industry No. 2; The Future of Digital Trust, p. 7). Orange.

[23] Short, J. E., and Todd, S. (n.d.). What's Your Data Worth? MIT Sloan Management Review. Retrieved November 8, 2020, from https://sloanreview.mit.edu/article/whats-your-data-worth/

future value. In her comprehensive inventory of academic papers and industry reports on the value of data, Slotin[24] observes how "striking [it is] that among [the] diverse perspectives, each author is grappling in their own way with the implications of data as a new economic asset, and yet there appears to be little consensus on how best to measure its value. One thing they can agree on is that measuring the value of data – and making [a] case for investing in data – is very difficult". Analogies with either tangible (oil) or intangible assets (patents, intellectual property, etc.) break at the point where the mapping between properties and assigned value becomes less clear (e.g., what is the difference in value between 32 and 35 GB of the same data? What is the difference between data that is 55% and 65% accurate? What is the value of a dataset that has already been used to train a machine learning model?). And perhaps this is normal since rules that apply to old commodities possibly do not even apply to this new kind of resource. To understand the difficulties of assessing the value of data, Mawer[25] follows the progression through each element of the data value chain, from raw data to action and potential value, and maps them to the sequential stages of the data lifecycle (discover, ingest, process, persist, integrate, analyse, and expose).

This follows on work by Porter[26] and Kaplinsky[27] – the first to describe value chains applied to the design, production, and delivery of products and services – subsequently adapted into knowledge value chains.[25,28] To explain the value of knowledge co-production, Peppard and Rylander[29] needed to break the linear model. They introduced the concept of network value, which allows its participants to function independently, within a framework of common principles.

Attard *et al*.[30] recognise the non-linearity of data value creation and refine the previously presented work into the data value networks (see Figure 14.1).

---

[24] Slotin, J. (2018). What Do We Know About the Value of Data? Global Partnership for Sustainable Development Data.

[25] Mawer, C. (2015). Valuing Data is Hard. Silicon Valley Data Science. Retrieved from https://tinyurl.com/495mt343

[26] Porter, M.E.. (1985). Competitive Advantage: Creating and sustaining superior performance. NY: Free Press.

[27] R. Kaplinsky, R. and Morris, M. (2002) A Handbook for Value Chain Research.

[28] Lee, C.C. and Yang, J. (2000). Knowledge value chain. Journal of Management Development, 19(9):783–794, 2000.

[29] J. Peppard and A. Rylander. (2006). From Value Chain to Value Network: European Management Journal, 24(2-3).

[30] Attard, J., Orlandi, F., & Auer, S. (2017). Exploiting the Value of Data through Data Value Networks. *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance*, 475–484. https://doi.org/10.1145/3047273.3047299

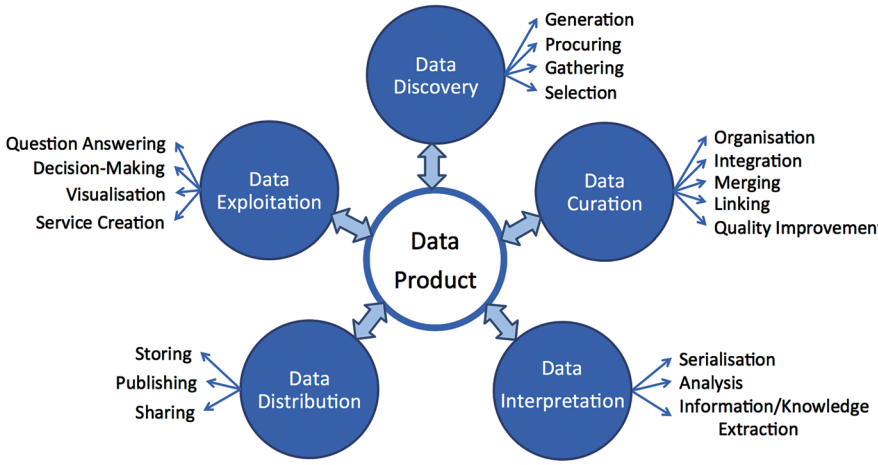

Figure 14.1 Data value network as illustrated by Attard *et al.*[30]

This is composed from activities (depicted by circles), each of them consisting of several actions (connected to the circles by simple arrows).

The strength of the network lies in the fact that it is built precisely around the characteristics that make the value of data difficult to assess:

1. Non-tangible product. It is infinitely shareable, and while some data may deprecate with time, it can be reused over and over, without losing its properties. It is also true that in some cases (e.g., social media big data, a company's financial data, industrial production data) part of these data's value relies on their uniqueness or on their sharing restrictions. However, excepting these cases, given the promotion of FAIR principles[31] and open science, accessibility (as opposed to closed silos) is expected to increase its value.

2. Non-sequential processing. While theoretically data processing pipelines have a relatively linear progression, in practice, things are a bit more complicated. Value generating activities can be skipped, executed in parallel, or in slightly different order than the theoretical one, or be a part of iterative loops.

3. Several actors can cooperate when realising the activities. Moreover, each of these actors may output a data product on their own or contribute to the co-creation of a data product.

[31] https://www.go-fair.org/fair-principles/

4. The network allows for the existence of nested value chains. These are formed by the actions that compose each activity.
5. The network allows for the existence of recurring value chains. This implies that value chains can be created as long as the data stays relevant. The data product resulting from a certain activity could be the "final product" or could form the input to a branching activity, thus perpetuating the chain.
6. The activities can be performed independently by any number of actors.

This model currently appears to be the best equipped to model and possibly quantify the high context-dependent nature of the data valuation process.

14.3 The Data Valuation Process

The data valuation process (DVP) and the data valuation component (DVC), implementing it, were developed as part of the Horizon 2020 Safe-DEED (Safe Data Enabled Economy Development) project.[32] It considers that the value of data is generated from two main areas: data quality and data usability, which are assessed through the lens of the context in which the data will be used. The context is set by the user, during a context definition procedure, based on which the relevant components of data quality and data usability are established (see Figure 14.2).

The tool is trying to maximise the automation degree of all these processes and proposes in-depth analyses to support the value of data and the reduction of the time dedicated to the data valuation process.

Since this is a complex problem, the presentation of the results avoids the generation of a single aggregate value. Instead, the platform generates a set of scores (for different perspectives and at different levels of detail), thus informing the user on the strengths and weaknesses of the dataset they are assessing.

Next, we look at each of the building blocks of the DVP and we try to understand how cities can use them (either as part of the DVP or as independent components) to understand and properly use the value of their data.

14.3.1 Data contexts

Let us consider a dataset containing GPS traces of taxis in a city. For a ride-hailing application, such data would provide a way into estimating the

[32] https://safe-deed.eu/

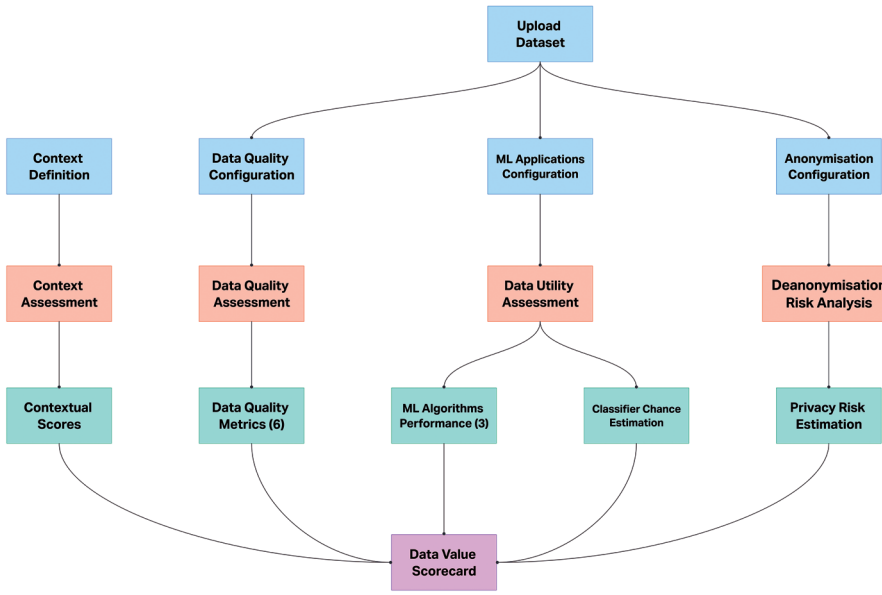

Figure 14.2 Overview of the components of the data valuation process. (Components requiring interaction with the user are in blue (rows 1 and 2 from above), automatised components in orange (row 3), scoring components in green (row 4), and the output component in purple (row 5). The arrows indicate the sequence of actions from processing the data to generating the result.)

customer needs in different areas of the city, at different times, allowing them to develop machine learning solutions for load balancing and trip planning and eventually maximising their revenues. The local administration could use this data to understand road congestion and travel times and plan infrastructure interventions (repairs, extensions, and restrictions), modify public policies (congestion taxes), or plan connected services (public transportation). A retailer could look at this data in conjunction with other sources and understand behavioural patterns of people living in different areas of the city and thus plan opening schedules, logistic operations, or decide to open new branches.

Data can have different values for different roles within the same organisation. Data containing the flow of passengers through a certain area might be enough for a planning manager and his team who decide to build a larger shelter or a new bus stop, but for the R&D department working on a new routing algorithm, such information might be overly aggregated and useless for their necessities. Even within the same department, different tasks might impose different requirements from the same dataset. The data science team

might be able to provide a good enough analysis of travel patterns from data which contains trips aggregated over 30-minute intervals, but such a dataset will not be useful if the task is to create an accurate traffic prediction model.

The contextual nature of data is often cited as one of the main reasons for which assigning value to it is difficult. We have seen how different value chains can be completed with the same raw data or how similar value chains can be completed with different raw data,[24,29] depending on the purpose of the data processing. Slotin[23] extends that observation and concludes that context-specific, impact-based methods might be the most suitable for communicating data value, despite this specificity also being their main drawback. In their data quality principles, the US National Institute of Statistical Science (NISS) cite contextual factors (purpose, user, and time) among those that influence data quality.

Building a solution that takes contexts into account has, first, to surmount the challenging aspects of defining, formalising, and encoding them. With research focused specifically on contexts for data value being almost inexistent, we turned to the literature on metadata for datasets and data quality assessments to seek for meaningful parallels. This confirms the context-dependent nature of data value and brings a first level of clarity concerning the layers that compose a valuation context:

- organisational profile;
- business user profile;[33,34,35]
- a specific task, personal preferences;[34]
- business rules/processes;[32]
- organisational and government regulations.[32,33]

In a sense, defining contexts is akin to understanding users, identifying use case scenarios, and deriving user requirements.

Recent work focusing on data profiling and valuation of metadata offers valuable leads into how data valuation contexts could be established and

[33] Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218.

[34] Cai, L., and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(2), 1–10.

[35] Even, A., and Shankaranarayanan, G. (2006, November 10). Value-Driven Data Quality Assessment. Proceedings of the 2005 International Conference on Information Quality. MIT IQ Conference, MIT, Cambridge, MA, USA.

quantified. Among them, we distinguish a questionnaire-based method for mapping data properties to data value,[36] the creation of datasheets for datasets,[37] and the Dataset Nutrition Label,[38] a diagnosis framework providing critical information at the point of data analysis.

As for the DVP, it requires that a user provides as much information as possible about the context in which a dataset will be used. This is done through a questionnaire with clear answers, which are then mapped to values, yielding the contextual value of the dataset. The questionnaire is structured in the following layers:

1. Systems and economics: availability and access; purpose
2. Legal and obligations: data protection; legal-terms-obligations
3. Data science: tools; format
4. Data properties: data velocity; data transformations; data quality; data age
5. Business: frequency of use; benefits

These early methodologies for context formalisation underline the importance of metadata that accompany a dataset. Metadata give a generic view into the origins of a dataset, the methods for generating it, the purpose for which it was generated, its format and access to it, the licenses that may apply to it, and the methods and tools used to process it up to its current form. Many of the data-driven stakeholders, whether producing or consuming data, are far from insuring the bare minimum with respect to metadata. This results in data being difficult to index, find, and (re-)use, essentially decreasing or even cancelling its value.

Through their technology departments, we recommend that cities step forward and assume the responsibility for creating such metadata. For those cities that already have IoT and e-government infrastructures in place, documenting such metadata (either by filling the DVC context valuation forms or

[36] Kannan, K., Ananthanarayanan, R., and Mehta, S. (2018). What is my data worth? From data properties to data value. http://arxiv.org/abs/1811.04665

[37] Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2020). Datasheets for Datasets. http://arxiv.org/abs/1803.09010

[38] Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2020). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. In D. Hallinan, R. Leenes, S. Gutwirth, and P. De Hert (Eds.), Data Protection and Privacy: Data Protection and Democracy (pp. 1–26). Oxford: Hart Publishing.

creating datasheets[36]) would add value to their current efforts and allow them to explore further avenues for their data.

For those cities that are still at the beginning of their digital transformation journey, this may be a good opportunity to do things right from the onset, by creating the data collection and management infrastructure, processing pipelines and metadata in a coherent way.

At this point, it is crucial that cities reach out to data and information practitioners, since these communities have the know-how needed to build fair and efficient data and metadata infrastructures. The ideal team should include:

1. Technical experts: data scientists, data engineers, library, and information science professionals.
2. Legal experts, preferably specialised in technology, intellectual property, or consumers law, able to advocate for the citizens who will be impacted by data processing, and foresee or react to future ethical and legal issues that will arise.
3. Experts from within city administration, preferably project managers who are able to map the requirements derived from urban challenges to technical solutions. These specialists should be the glue connecting the needs of the city and its citizens, the technical solution, within the legal and ethical boundaries.

While building an in-house team would be the ideal setup, budget limitations might require cities to seek for partnerships with universities, technology centres, or private companies. In these cases, cities should proceed with care, as lack of budget, technical knowledge, or legal safeguards may lead to their data being used for other purposes than those intended. Losing the trust of the citizens (either as data subjects or users of data products) can lead to the failure of digitalisation efforts altogether.

The creation of metadata and their evolution towards a standard for context formalisation is crucial for the success of data valuation methodologies. They provide the building pieces for describing a variety of contexts and discovering those dimensions that are important for the value of data in these contexts. Generalising the different contexts would enable for both context-dependent and context-independent analyses of the value of data. Finally, connecting these contexts to quantifiable values can lead to the establishment of transparent and fair data markets.

In this respect, cities have the advantage of processing a wide variety of data (both personal and non-personal), which makes them the ideal partners for pushing the R&D of methods for data valuation.

14.3.2 Data quality assessment

The earliest preoccupations towards a formal understanding of quality date back to its application to assembly-line production and manufacturing in the beginning of the 20th century and accelerated later in the 1950s and 1970s with its adoption to business practices. Along the years, various definitions have been put forth, referring to quality as "conformance to requirements",[39] Joseph Juran's famous "fitness for use",[40] or the "degree to which a set of inherent characteristics fulfils requirements".[41] One definition refers to quality as the "value to some person",[33] recognising the intrinsic value derivable from data quality as well as its contextual nature.

With the development of ICTs, interest in quality of data has sparked during the 1990s. The democratisation of the internet and the advent of big data and data-centred solutions generated more interest in the topic and laid the ground for a currently mature and dynamic research field. In 1996, the Total Data Quality Management Group at MIT adopted the "fitness for use" definition and acknowledged its dependency on the consumers. The principles of data quality by the US National Institute of Statistical Sciences (NISS) adopt the view of data as a product and, as such, consider that its quality results from the process that generates them. Later, data quality was enacted at the governmental level, as was the case of the US Data Quality Act[42] or the Welsh Data Quality Initiative Framework.[43] In Europe, Bergdahl *et al*. report on the successful integration of data quality assessment in the activities of several National Statistics Organisations: Statistics Sweden, Statistics Norway, CBS in the Netherlands, the Austrian Quality Concept (an in-house quality reporting system), the ONS Guidelines for Measuring Statistics Quality (a grading scheme for statistical products), and the Slovenian Statistical Office (data quality measurement for short-term statistics).[44]

---

[39] Crosby, P. B. (1988). Quality is Free: The Art of Making Quality Certain, New York: McGraw-Hill.

[40] Juran, J.M. (1951). Quality Control Handbook. 4th ed.

[41] International Organisation for Standardisation. (2015). ISO 9000 Family for Quality Management Systems.

[42] Office of Management and Budget. (2006). Information quality guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by agencies. Available at: https://tinyurl.com/ychxhmsd

[43] NHS Wales. (2004). Data Quality Initiative Framework. Project Report.

[44] Bergdahl, M., Elvers, E., Földesi, E., Kron, A., Lohauß, P., Mag, K., Morais, V., Nimmergut, A., Viggo Sæbø, H., Timm, U., and Zilhão, M. J. (2007). Handbook on Data Quality Assessment Methods and Tools. European Commission - Eurostat.

Data quality can be regarded as the ability of data to serve its purpose – generally seen as the needs of an organisation in terms of operations, planning, and decision-making.[45] Therefore, in order to evaluate the quality of data, a plethora of data quality assessment methodologies have been developed over the recent years, adopting different perspectives and covering an even larger spectrum of quality dimensions in their attempt to encompass the multitude of assessments that gather under the data quality umbrella.

To clarify, "a Data Quality Dimension (DQD) is a recognised term used by data management professionals to describe a [property] of data that can be measured or assessed against defined standards in order to determine the quality of data".[46] Dimensions focus on measuring and communicating the quality of data, as opposed to describing what the data represents.

14.3.3 Data quality metrics and dimensions

Historically, there is a correlation between the development of ICTs and that of data quality assessment methods. The early systems were monolithic, usually consisting of a single data source and simple data flows, and the only source of errors would come from data entry. Data quality would, therefore, involve accuracy, consistency, completeness, and time-related metrics. The evolution towards network-based systems involved a re-adaptation of these dimensions; with the later advent of the web, data sources have become more numerous and more varied and, as a consequence, new dimensions such as accessibility and reputation had to be considered. Currently, peer-to-peer systems require a new rethinking of these dimensions and, more importantly, the consideration of privacy issues. This evolution of ICT systems is itself one of the causes for the number of methodologies, some of which specialised on subsets of data quality issues.

An overview of all dimensions and subsumed metrics[47] allows us to confirm the complexity and multi-dimensionality of the concept of data quality. The Total Data Quality Management Group at the MIT defines 15 quality

[45] Lebied, M. (2018). Guide To Data Quality Management & Metrics for Effective Data Control. Datapine. Retrieved from https://tinyurl.com/2p8fmxrx

[46] Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G., and Schwarzenbach, J. (2013). The Six Primary Dimensions for Data Quality Assessment—Defining data quality dimensions. DAMA UK.

[47] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys, 41(3).

dimensions,[48] the Data Management Association for the UK focuses on 6 primary dimensions,[46] and Statistics Netherlands mentions 49 factors that influence the quality of secondary data and groups them into 5 focus areas.[49]

Due to the contextual nature of data quality assessment, there is little to no consensus as to what might be a subset of necessary data quality dimensions to consider. But is there a subset of "basic" dimensions and metrics that should always be considered when assessing data quality?

A review of DQA methodologies points towards a set of four such DQDs, namely: completeness, validity, accuracy, and timeliness.[33,34,44,45,50,51,52,53] ISO/IEC 25012 confirms these as well as "credibility" as inherent characteristics of data quality.[54]

Data quality assessment is key to unlocking the value of data and if they are to embrace digital transformations, cities should place it at the centre of their technical activities. Thus, cities should seek to ensure the assessment along the five DQDs previously mentioned, by using appropriate metrics (see Table IX in the work of Batini *et al*.[47] for a complete list of DQMs). The next natural step would be to invest in the technical expertise necessary to address the shortcomings identified along each of these dimensions. Again, while data quality requirements might differ on a case-by-case basis, there are minimum data quality requirements that are expected from data and seeking to achieve these already increases its general usability.

---

[48] Warner, M.R., and Hawley, J. (2019). Designing Accounting Safeguards To Help Broaden Oversight and Regulations on Data. Retrieved from https://tinyurl.com/3fujstf3

[49] van Nederpelt, P., and Daas, P. (2012). 49 Factors that Influence the Quality of Secondary Data Sources. In: Quality and Risk Management (12). Statistics Netherlands. The Hague.

[50] Behkamal, B., Kahani, M., Bagheri, E., and Jeremic, Z. (2014). A Metrics-Driven Approach for Quality Assessment of Linked Open Data. Journal of Theoretical and Applied Electronic Commerce Research, 9(2), 11–12.

[51] Görz, Q., and Kaiser, M. (2012). An Indicator Function for Insufficient Data Quality – A Contribution to Data Accuracy. In H. Rahman, A. Mesquita, I. Ramos, and B. Pernici (Eds.), Knowledge and Technologies in Innovative Information Systems (Vol. 129, pp. 169–184). Springer Berlin Heidelberg.

[52] Piprani, B., and Ernst, D. (2008). A Model for Data Quality Assessment. In R. Meersman, Z. Tari, and P. Herrero (Eds.), On the Move to Meaningful Internet Systems: OTM 2008 Workshops (Vol. 5333, pp. 750–759). Springer Berlin Heidelberg.

[53] Sebastian-Coleman, L. (2010). Data Quality Assessment Framework. The Fourth MIT Information Quality Industry Symposium.

[54] International Organisation for Standardisation. (2008). ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model.

To take on this challenge, cities would need to employ the same type of profiles as in the case of metadata generation. In fact, data quality assessment is an even higher technical endeavour, requiring a variety of profiles: analysts, data scientists, data engineers, domain experts, and legal experts. Therefore, the discussion in Section 14.3.1 about how such a team would work and the comments with respect to outsourcing these activities apply here as well.

Using a solution like the DVC can speed up the process, due to its configurable data quality assessment module. This would not require an entire data science team; it would, however, require the collaboration between a domain expert and a data specialist, who would be able to understand the quality requirements of a dataset and their connection to the problem. Once such experts become trained in using the DVC, they can apply it to any available dataset and, thus, get in-depth knowledge of the quality of data and eventually quantify its value.

14.4 Aggregating and Reporting the Value of Data

The success of the data valuation process depends on its adoption by data practitioners which, given the multi-dimensional nature of data valuation, depends on:

- the capacity of the platform to promote the transparency of the assessment processes;
- the interpretability and replicability of results;
- the degree to which such results can be used by practitioners.

The last item refers to the necessity to aggregate the results of the subprocesses that compose data valuation into a single measure that can be easily understood at different levels of organisations and based on which, ultimately, decisions of economic nature can be made. Thus, notions such as "energy label" for data or "price tag" for data are appealing, especially to those operating at commercial or executive levels of organisations. However, such aggregate measures are both difficult to construct (at least for now) and may lead to confusing or inaccurate interpretations, which could undermine the whole data valuation effort.

Interest in developing a single measure to characterise data first appeared in the community of data quality practitioners. Pipino *et al*. point to the fact that a single-value aggregated measure – a quality index – could be subject to the same deficiencies that affect other commonly used indices (Dow Jones Industrial Average, Consumer Price Index, etc.).[33] These derive

from the statistical methods used for estimations, the interpretation of different components, their contribution weight to the final index, the methodologies used for choosing these components, etc. Similar shortcomings are noticed by Bronselaer *et al*.[55] who warn about the difficulty in interpreting an aggregation of DQMs, each referring to very different quality facets. Even when choosing a reporting scale, both groups of researchers point to relevant challenges, whether it is the difficulty of aggregating DQMs operating on different scales[33] or the loss of interpretability of a result that standardises all DQMs in the [0,1] interval.[54] Bergdahl *et al*. mention that previous attempts to compile composite indicators for data quality by NSOs have failed and refer to the contextual nature of DQA as the main constraint for selecting the right subset of indicators and assigning them suitable weights.[43]

Reporting is paramount in promoting the adoption of innovative platforms, especially if they involve complex evaluation processes, like the DVC.

A first component of reporting is data profiling, which is usually performed as an entry point to data quality management,[44] right before data analysis. This gives an initial insight into the data (ranges, distributions of attributes, pair-wise correlations, etc.) and supports the definition of data quality requirements.[56]

Once DQA is performed, there are several approaches to reporting an often-multi-dimensional result and eventual aggregates.

- Report the cost associated with poor quality of data and summarise it in a data quality scorecard.
- Issue a certificate of data quality or a quality alert, depending on whether quality requirements are satisfied or not. It is recommended that only a small number of self-explanatory labels (e.g., "sufficient quality", "experimental data", etc.) are created and that, once introduced, they stay in circulation for some time. Labels should also include "expiration dates" and allow for constant recertification of datasets, reactive to changes in content or requirements.
- Create data narratives that highlight the impact of good or bad data. Impact-based approaches for data valuation[52] tell compelling stories and connect data to contexts and clear outcomes.[57]

---

[55] Bronselaer, A., De Mol, R., and De Tre, G. (2018). A Measure-Theoretic Foundation for Data Quality. IEEE Transactions on Fuzzy Systems, 26(2), 627–639.

[56] Jones, D. (2016). Data Profiling vs Data Quality Assessment – Let's Explain The Difference. Data Quality Pro. Retrieved from https://tinyurl.com/2swe9ww3

[57] Hammond, K. J. (2013). The Value of Big Data Isn't the Data. Harvard Business Review.

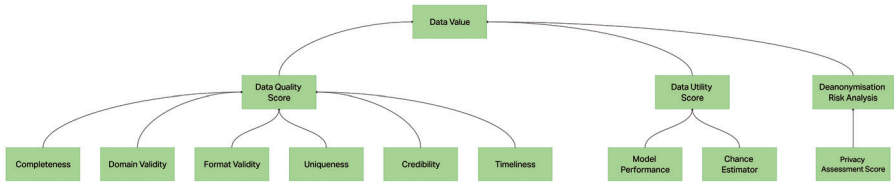

Figure 14.3 The aggregation of the scores generated from the different sub-components of the DVC into the data value scorecard.

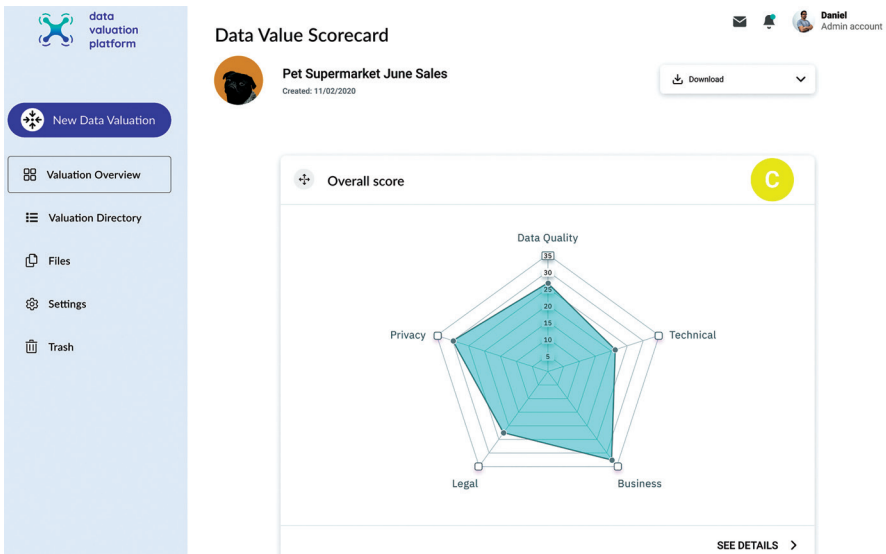

Figure 14.4 Data value scorecard. A combined view over the general data quality score and its composing data quality metrics.

It is important to underline that the DVP does not generate a price tag for data. We have already discussed the technical difficulties for achieving that (e.g., the properties of data as an asset, the properties of data value chains, and lack of adapted economic methodologies). Beyond these, data pricing (and, in particular, private data) raises important legal issues (ownership, intellectual property, etc.), as well as moral ones – should we even engage in transactions and monetary exchanges involving digital extensions of human identities?

The DVP thus focuses on giving a multi-faceted quantification of the components of data value (data quality, data utility, and privacy), within the defined context (see Figure 14.3, Figure 14.4 and Figure 14.5). This

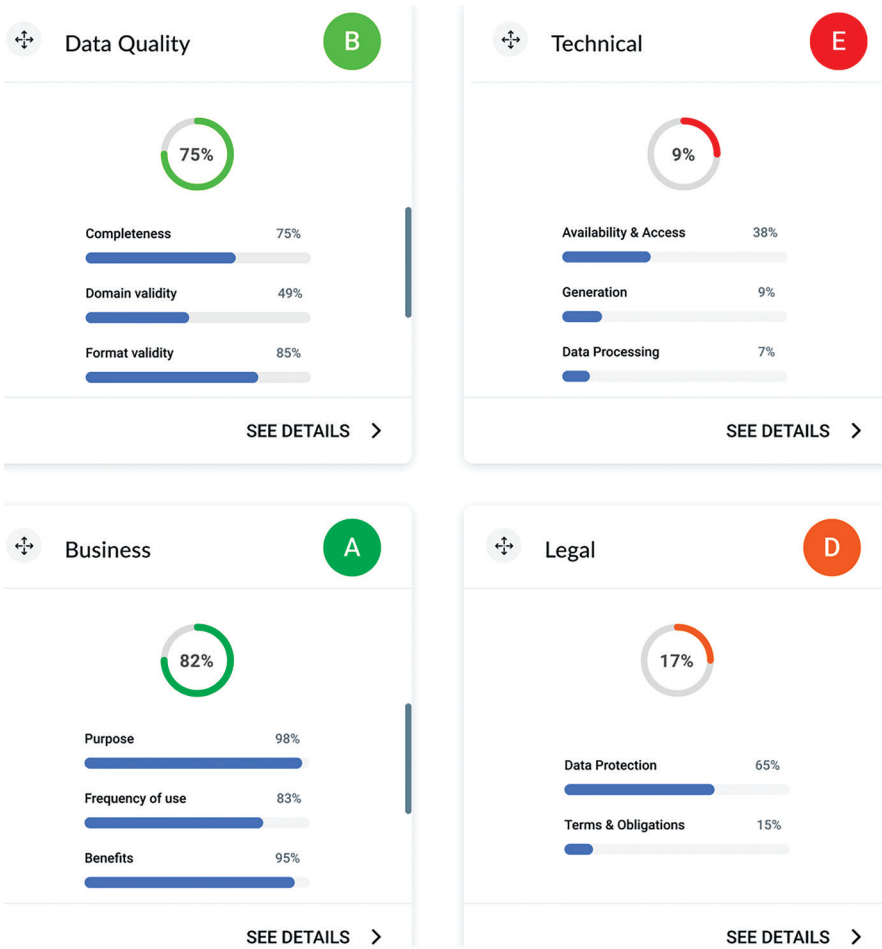

Figure 14.5 Data value scorecard. A multi-dimensional view of the contextual score.

multi-dimensional reporting system allows for stakeholders to grasp the value of data from a variety of perspectives and at different aggregation levels; it does not exclude a one-value score, but it invites practitioners to explore the reasons behind it.

The reporting of data value is a challenging problem, as it tries to connect various dimensions of data value, to valuation contexts, the needs of various stakeholders, as well as human factors. It would greatly benefit from applying user experience design principles to identify the best way to interact with the user and present the outcome of data valuation.

Once again, cities and their communities can prove very useful in helping refine this aspect of data valuation. The variety of challenges, data, and data-centred use cases that cities generate places them as an important contributor to data valuation R&D. Cities can become a facilitator for communities and individuals to have an active role in the processes of digital transformations. Administrations can reach out to their citizens and businesses and organise focus groups, in which they can connect with data valuation professionals. We believe that such exchanges will benefit all parties.

- Data value professionals can educate the large public on the value of data, challenges, and how they can benefit from it. This should promote the use and further development of methodologies and tools for data valuation.
- Cities can discuss with communities their different data-centred projects and try to understand future lines of development and shape public policies. This could be a part of the constant dialogue required between city administrations and citizens.
- Citizens and businesses will be able to communicate their needs and their expectations and help shape the next generation of technologies, which have to become more than harvesting tools of digital identities.

14.5 Takeaways for Cities

Cities are a diverse concentration of people and activities giving rise to a multitude of daily challenges in their everyday functioning, as well as in their quest to serve their communities. Their response to today's societal challenges places them at the centre of digital transformations. The adoption of IoT technologies and the permeability of social media mean that we now have a better view than ever into the lives of these macro-organisms. Obviously, this turns them into ideal testbeds for new, impactful technologies, but it depends on each city (both community and management!) to leverage these characteristics and place cities in the driving seat of these transformations.

Educate citizens about digital transformations and its value for the community. Understanding the potential of the data that is generated in a city must begin with its managers. They are in charge in setting a city's policies and digital education cannot be ignored anymore. From here, this should be diffused towards their communities, by means of communication and education projects aiming to make everyday citizens aware of the capacity of the city to record human behaviour, the variety of this behaviour, the fact that this

behaviour generates value and can lead to progress in theirs and other's communities, and the potential misuses of this data and how it can be prevented.

Involve citizens when designing solutions. Citizens need to feel that they are more than just "users" or "data points" and it is up to the cities and their democratic, participative processes to give citizens a voice. Involving citizens in designing technical solutions has multiple advantages.

- First, it will lead to higher adoption of these solutions, especially if these involve advanced technologies. Public funds must go in carefully designed citizen experiences, as opposed to expensive apps with stale designs that nobody uses.
- It will promote trust in these technologies, also improving adoption rates and building an honest relationship between citizens and managers. Cities should avoid becoming yet another big data processor or data broker.
- Finally, it will create a sense of community, which gives more responsibility to the citizens: they care more about their neighbourhood or city, which has benefits in terms of creating safe and resilient communities.

Design fair, explainable, and privacy preserving technology. We are at a point at which artificial intelligence solutions are being deployed at an accelerated rhythm. Nevertheless, the research community is revealing increasingly more cases of bias present both in AI systems and the data that is powering them. Moreover, the collateral effects of some of these systems have been challenging ethical and societal principles, as well as our current legal frameworks. Cities can respond to these by promoting partnerships with researchers from these areas. The trove of data they are holding would do the following:

- Promote partnerships with researchers from these areas. The trove of data that they hold contains a variety of sensitive information (demographic, behavioural, financial, etc.) and could allow specialists to advance the state-of-the-art in data privacy and algorithmic fairness, while cities would benefit from fair, explainable, and privacy enhancing technologies.
- Lead the discussions about the adaptation of legal frameworks. Lawmakers need to work with technical and legal experts on understanding the consequences of deploying data-centred technologies in cities, their potential conflicts with human and citizen rights, and how to transpose this into law. This is particularly interesting in the case of

cities becoming active as data brokers or managing personal data on behalf of citizens.

Get involved in data markets. One possible avenue is for cities to try to monetise the data they generate, by participating in data markets. An interesting discussion concerns the way in which the revenue obtained from such data would be distributed: to every individual as a form of pay-off for the contribution of their data, redirected towards the city's budget and reinvested in local projects or a combination of the two. Obviously, such an avenue would need to address issues such as data ownership or data rights management and would absolutely need to involve the use of privacy and fairness checkups, as described earlier.

Open the data. Don't silo the data! We have discussed how different value chains, executed by different actors, can generate value from data. Opening the data (e.g., by implementing FAIR principles) will tap into the creativity of other stakeholders and support innovations that were not initially considered. Like before, this should be done while previously deploying privacy preserving and fairness promoting mechanisms. Opening personal or any kind of sensitive data has to lead to scientific advancements and the creation of fair, responsible data products, without the cost of exploiting the lives of those who helped generate this data.

Build data teams and know-how. As we mentioned in Sections 14.3.1 and 14.3.3, cities should consider the creation of data-focused technical teams. These teams should be dedicated to the creation of metadata (to facilitate the creation and quantification of data contexts) and the highly technical data quality analysis. Alternatively, cities can opt for using advanced analyses tools, like the DVC, in which case they should invest in preparing technical staff that is able to understand use cases, translate them to technical requirements, and interpret multi-dimensional output. If these activities end up being performed in partnerships with external parties (tech companies, consultancy companies, and universities), they should make sure that they put in place the right legal and contractual mechanisms for protecting the intellectual property of the data and the privacy of the data subjects. Finally, it is highly recommended that cities allocate resources for two important roles:

- technical facilitator, to bridge the gap between city-specific requirements and teams working on technical solutions;
- legal and ethical experts, dedicated to identifying the challenges put forth by data-centred solutions, their impact on citizens, and how they could lead to an evolution of the current legal framework.

Communicate. Data value is a complex, multi-dimensional concept. While everyone is aware of the value of data, this is still difficult to quantify and report. Promoting data valuation methodologies relies on the capacity of the target audience to grasp the intended message. Practitioners insist on the power of impact-based data valuation methods, able to convey value by creating data narratives. Data wrapped in stories are 22 times more memorable than bare facts[58] and this is where cities can tap into their communication capabilities, by making citizens understand how the data that they contribute has the capacity to drive change in their communities.

[58] http://chicagoanalyticsgroup.com/blog/archives/01-2017