Data As Infrastructure

Introduction

In the 21st Century, data is infrastructure. This is because the managed and built environments increasingly depend upon data in real-time. Moreover, the sources of this data are potentially multiple, not necessarily arising from within the control of traditional institutions, and yet this data can be complex in form. It follows from this that data is a critical component and needs to be understood as a key part of 21st infrastructure but that it presents new challenges to those institutions concerned with the safe and effective management of infrastructure.

It is now widely accepted that the digitization of the economy has taken root in a way that means it is not confined to one sector. All sectors are affected in some common ways. Brynjolfsson and McAfee[1] are among those who have described this. The economic drivers behind digitization are successfully isolated and described by Goldfarb & Tucker[2], and there are important contributions to understanding given by economists including Levin[3] and Nordhaus[4]. It is proportionate to describe the economic and social ramifications within the frame of 'Creative Destruction', originally described by Schumpeter in 1942[5]. In this light, the importance of data can be expected to grow across most or all industry sectors. Its effective management will become ever more critical to the economy and to society more widely.

Data As A Public Good

The deep technological reformation of current society has meant that the potential of data is gathering attention alongside other celebrated advances in the hardware and infrastructure of society. New mechanisms for the assembly, management and processing of data provide a new impetus for rethinking how the data is best managed so that society can best utilize its resources, solve the most problems, and provide the most social good for the most people.

Data itself has become an important part of the infrastructure of the nation and must be managed for the best effect. For this reason, data can be widely described as constituting 'a public good.' Its availability and use is a necessary part of the public realm. It is more subtle and yet profound to identify that within this, data is taking on new roles in the real-time performance of infrastructure, and even in the design and maintenance of that infrastructure. Yet, this data often comes from diverse sources and must be managed in new ways. We reserve the term 'Data as Infrastructure' for these new circumstances where the incorporation of data into the performance, design and maintenance of infrastructures reaches a new level of sophistication.

---

[1] McAfee, A. and Brynjolfsson, E., 2017. Machine, Platform, Crowd: Harnessing Our Digital Future.
[2] Goldfarb, A. and Tucker, C., 2017. *Digital Economics* (No. w23684). National Bureau of Economic Research.
[3] Levin, J.D., 2011. *The economics of internet markets* (No. w16852). National Bureau of Economic Research.
[4] Nordhaus, W.D., 2015. *Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth* (No. w21547). National Bureau of Economic Research.
[5] Schumpeter, J., 1942. Creative destruction. *Capitalism, socialism and democracy*, *825*.

The realisation of the value of data across the public realm depends not only on the technological exploitation of massive amounts of data, but also upon the governance strategy of that data. A data governance strategy gains value when it is placed in the broader national context and when it is aligned with the overall vision of the nation. We synthesise these fragmented concerns into a more coherent framework. The framework provides a systematic way of defining the concept 'data as infrastructure' and of seeing its place alongside other important data governance initiatives.

Technological Drivers

The effective governance of data relies on the implementation of a number of 'Big Data' technologies. This term, 'Big Data' is now commonplace. It has become relevant because a cluster of innovations made it feasible to utilize larger and larger data sets, and that often these data sets are only semi-structured or even unstructured. These technologies that enable Big Data include the Cloud, ever faster chips, the Internet of Things, and Machine Learning. Together these innovations help in the collection, integration, validation, real-time analysis and reporting of massive amounts of data. Table 1 outlines an overview of these technologies and the type of problems that they are able to solve.

In practice, the technologies listed and described in Table 1 are utilized in bundles. To solve a specific problem, a combination of technologies are used together. These bundling effects amongst technologies makes prioritisation for investment very complex and the choices depend substantially on the data governance mode that we describe next.

Purpose	**Technologies**	**Solution to**[6]
Data Collection. Integrating and Unifying Different Sources of data	Sensing (including radar, lidar, sonar, satellite imaging, thermal imaging, quantum sensing and the use of drones), Cloud Technologies and the Internet of Things	Selection Problem
Dimensionality Reduction of Massive Datasets and Real-Time Predictive Modelling	Machine Learning (e.g. Deep Learning)	Prediction Problem
Transaction Verification, Data Accuracy	Blockchain and Distributed Ledger Technologies	Verification Problem
Prototyping, Design Diagnostic and Operation Monitoring	Virtual and Augmented Reality, Digital Twinning	Replication Problem

Table 1. Technological Drivers

Data Governance Strategies

We formalise the data governance modes around four major themes based on the role that the government can take. This role of government is very important in all the governance modes but can

[6] The 'Selection Problem' refers to the problem of making the most relevant data accessible. The 'Prediction Problem' denotes a range of problems where the occurrence of an outcome is predicted using highly dimensional data. The 'Verification Problem' highlights a type of problem where it is hard to examine the validity of records through accurate tracking of a large number of prior transactions. The 'Replication Problem' relates to a range of problems where learning is vital to the performance of a system but it is costly or difficult.

vary greatly among them. 'Data as Infrastructure' comprises the ultimate position where Government is wholly or partially 'a Smart System.'

In practice, all government institutions will seek hybrid arrangements across these four themes according to their unique, current and expected challenges.

1- Government as a Provider

The first and most common form of provision is the designation and release of government data as public data. Part of this is data that is generated in relation to public infrastructure. The rationale behind such initiatives (commonly referred to as Open Data Initiatives) is the idea that releasing such data brings about accountability and transparency, and that it empowers a form of participatory governance. Data is considered as a public good and access to this data is a potential right for every member of society. It is argued that such initiatives promote participation, increase innovation and facilitate evidence-based decision making.

In this conception, government is only the provider of the data. The data itself can be used by citizens, bodies such as non-governmental organizations (NGOs), and the private sector; all for different purposes. Although the provision of data is subject to public request and scrutiny, a hierarchical approach is used to maintain control over the type of content that is released and to manage issues relating to data privacy. This unidirectional system of control is limited in scope. In order to ensure the maximum potential value is developed from the data, an open data initiative will seek to impose only a minimal critical set of controls over the usage of the data. The 'government as a provider' category can be considered as analogous to the provision of a park wherein the government decides on its location and content but has a minimal framework of control over how the park is used. Such a 'park' is encapsulated by government acting as a provider in Open Data UK (https://data.gov.uk/).

2- Government as an Enabler

An alternative approach is for the government to provide a unified data marketplace. This approach has been less common than the simpler role of 'Government as a Provider', but has recently gained momentum. 'Government as an Enabler' is developed in recognition that government is not the only entity that has data that is important to the needs of society. Much of the data about infrastructure and its use is in the control of the private sector (e.g. telecommunications companies, IT companies, logistics companies), semi-state organizations (e.g. public transport franchisees), universities and research organizations, and individuals themselves. This approach of 'Government as an Enabler' also potentially satisfies the common call that data held by the government should not be used by businesses free-of-charge. This is justified on the basis that the data has been gathered at taxpayers' expense and that is has economic value. A governance strategy based on enablement helps governments to control access and use. Enablement implies that the role of government is to design a data marketplace so that there is exchange among suppliers and users, and the optimal value of the data is realized by participants in the market.

In this category, data is considered as a commodity and exchanged via the medium of the market therein. This facilitates more efficient use of the data by a greater multiplicity of interested parties and potentially leads to more data-driven innovations and to economic growth. Examples of this approach include Data For London (https://data.london.gov.uk/data-for-london/) and the Copenhagen City Data Exchange (https://www.citydataexchange.com/#/home).

3- Government as a Lab

In this approach government will seek to develop a network of different providers and users through a specific initiative. It does so in order to manage a research agenda within the network. The research questions and their answers might be attained through a kind of closed or semi-closed environment that relies upon only certain parties and data. Typically, this will be known as a 'lab.' A lab allows government to bring data providers together in order to answer specific policy issues or questions and hence is pro-active and managed through formal research governance.

This formal research governance is concerned with issues of quality in terms of inputs, outputs and process, and is also concerned with the interface to policy mechanisms themselves. Designing an appropriate ecosystem and providing the right incentives demands a high level of government involvement (at least in the design and maintenance phase) and enriched collaborations and partnerships between public sector, private sectors and citizens. In this, the data itself can be considered as embodying a concept. The co-creation processes turn the concept into valuable innovations, enabling more efficient use of infrastructure and helping in the delivery of public services. A well-known example of a lab is that of the Bristol Living Lab, (http://www.openlivinglabs.eu/livinglab/bristol-living-lab).

4- Government as a Smart System

A fourth approach is to develop highly automated and intelligent closed systems that support both the real-time working of the environment and an ongoing process of analysis/learning about the optimisation of this environment. This 'smart' environmental model is most obviously illustrated in the 'Smart City' concept but is more general. Effectively the concept applies whenever the Internet of Things and other monitoring systems are brought into the wholescale and real-time management of a facility or a geographical area. Given that such a system is controlled by or on behalf of the government, it can be described as 'Government as a Smart System.' Regulated algorithms will instrument space and will determine many things that potentially have political or economic ramifications, e.g. who has access to physical space, road-space, natural environments or services. Smart systems will learn through Artificial Intelligence about any issue within their scope, e.g. the best movement of emergency vehicles, crop interventions, patterns of lighting or refuse collection. As data generates the behaviour of infrastructure, it can be said that data is in a sense also a hard infrastructure and that it needs to be maintained and managed through a formal approach, analogous to the way that physical infrastructure itself is managed. This kind of system is necessarily closed for the reason that data quality is key, but the system will also support learning and can be integrated into an overall governance framework alongside other roles of government (1-3 above).

The management of infrastructural data is possible through centralised silos drawing from each aspect of physical infrastructure. New algorithmic and storage advances support the collecting, merging, visualising and analysing massive amounts of data. The Smart City architecture promoted by IT corporations normally relies upon this type of hierarchical arrangement. The closed governance structure is able to provide real-time monitoring of the all infrastructure to which it is linked. This security comes at the expense of issues such as privacy, ownership and flexibility. An example is The Dubai Smart City (http://www.smartdubai.ae/) but the meteorological project Radar Meteorológico of the Rio de Janeiro Centre of Operations exhibits the same characteristics (http://centrodeoperacoes.rio/). There are significant existing initiatives in rising economies including India (http://smartcities.gov.in/content/). The Kingdom of Saudi Arabia's plans to develop Neom will

constitute a new global benchmark in the scale of use of intelligent, algorithmically-driven, infrastructural data (https://www.bloomberg.com/graphics/2017-neom-saudi-mega-city/). The redevelopment of Toronto Quayside is a further benchmark example. Its plans incorporate closed, hierarchical management of data alongside open data in hybrid governance arrangements (https://www.sidewalklabs.com/). The ambient intelligence that will characterise such environments will increasingly support ecological management of urban areas, whilst also facilitating the same sort of 'smart' management of non-urban environments. Hence, whether the setting is urban or countryside, an ecological paradigm will be dominant.

'Smart System' characteristics are growing through the same combinatorial innovation effects as described earlier as 'Technological Drivers.' Digital Twinning, for example, is an important initiative that allows a key infrastructure, or a key part of an infrastructure, to be managed through a twin. This twin is effectively a data representation of the infrastructure that takes real-time and other data into the management processes of that real-infrastructural component. To illustrate, a gas turbine, a jet engine or a sluice might have a twin that supports its real-time monitoring and management. In turn, this implies that a higher level of automation will follow as many of the decisions will not need human intervention. The twin can take care of an increasing percentage of decisions. Moreover, it also implies the greater use of data through the lifecycle of the infrastructural item, as the digital twin can be created at the design stage of the component and then used in the governance of its manufacture, installation, maintenance and decommissioning. As stated, there are examples of this 'digital twinning' approach available for complex manufactured artefacts like jet engines, but it is clear that the concept applies much more widely, e.g. to buildings or to roads. Once different elements of an infrastructure are combined together then it becomes possible to conceive of a range of infrastructural components all collaborating through their digital twins (e.g. a stadium with a road system and footbridge). This begets a sophisticated level of process automation across components of the environment[7]. As well as advantages in efficiency across the lifecycle of components, issues also arise, for example in the example given, the twin of a stadium might, take "algorithmic authority" over when people can leave a football game or concert because that twin is responsive to traffic flows across a nearby footbridge. Again, such scenarios imply the need for proper governance.

[7] This is the Software Engineering concept of an 'active model' e.g. Snowdon, B. and Kawalek, P., 2003. Active meta-process models: a conceptual exposition. *Information and software Technology*, *45*(15), pp.1021-1029.

	Government as a Provider	*Government as an Enabler*	*Government as a Lab*	*Government as a Smart System*
Data Considered as	Public Good	Commodity	Concept	Feedback
Government Involvement	Low	Medium	High	High
Structural Mode	Hierarchy	Market	Network	Hierarchy
Motivation	Transparency, Participatory governance	Monetising the value of the data	Co-Creation	Closed Governance with highly efficient execution.
Examples of Initiatives	Open Data	Data Marketplace	Living Lab	City Dashboard

Table 2. Data Governance Strategies.

Conclusion: A Conceptual Model of "Data as Infrastructure"

It is the public infrastructure of a nation that enables the delivery of public services and which provides the platform for markets and culture. Nations vary in terms of their institutional arrangements over the management of their different infrastructures. As these nations then head towards the exponential changes of a digital era, the use of data in conjunction with physical infrastructure will lead to an environment that is ever more efficient and evermore intelligent.

Ultimately, the strategy of any given nation is shaped both by its historical context and its vision toward the future. What kinds of infrastructure does it have, what will it have, and how will these fit into the broadest socio-economic context? The efficient use of public data starts with the evaluation of the existing physical infrastructure and how the different parts of it will be impacted by the vision of the future. Identification of the problems and complexities associated with such infrastructure will help government decide upon its position and pursue the most appropriate form of data governance strategy (a hybrid data governance mode). Upon selecting the right form of governance, the government is then able to invest on the right portfolio of technology bundles.

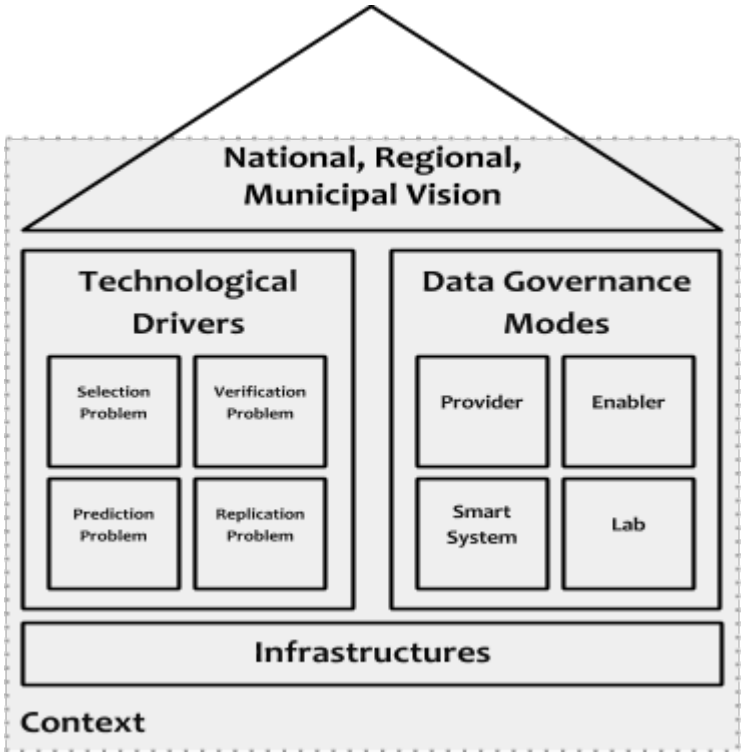

Figure 1. Data As Infrastructure Conceptual Model

Returning to the grand scheme of digitization, there is no doubt that the management of infrastructure is subject to the same pressures of digitization as are seen in other sectors. An ambient intelligence will be available across all kinds of environment to ensure the best utility and care of those different environments. Benefits and issues await. Within this trajectory of progress, data and its management become increasingly critical, partially because its effective use enables improved learning and improved policy, and partially because in increasingly complex ways it becomes part of the actual performance of that infrastructure itself. It ultimately follows that data *is infrastructure* and has to be managed *as infrastructure*.

Authorship.

Peter Kawalek, Centre for Information Management, School of Business and Economics, Loughborough University, Loughborough LE11 3TU. p.kawalek@lboro.ac.uk

Ali Bayat, Alliance Manchester Business School, Manchester University, M13 9SS.
Ali.Bayat@postgrad.mbs.ac.uk