Data "Of the People, By the People, For the People"

04
CHAPTER

"Cross the river by feeling the stones."
- Deng Xiaoping

Navigating in an uncertain, wobbly world requires constant monitoring of the path followed by the economy using real-time indicators. Thus, data can serve as the stones that enable one to cross the river. Concurrent with the data explosion of recent years, the marginal cost of data has declined exponentially while its marginal benefit to society has increased manifold. Therefore, society's optimal consumption of data is higher than ever. While private sector does a good job of harnessing data where it is profitable, government intervention is needed in social sectors of the country where private investment in data remains inadequate. Governments already hold a rich repository of administrative, survey, institutional and transactions data about citizens, but these data are scattered across numerous government bodies. Utilising the information embedded in these distinct datasets would inter alia enable government to enhance ease of living for citizens, enable truly evidence-based policy, improve targeting in welfare schemes, uncover unmet needs, integrate fragmented markets, bring greater accountability in public services, generate greater citizen participation in governance, etc. Given that sophisticated technologies already exist to protect privacy and share confidential information, governments can create data as a public good within the legal framework of data privacy. In the spirit of the Constitution of India, data should be "of the people, by the people, for the people."

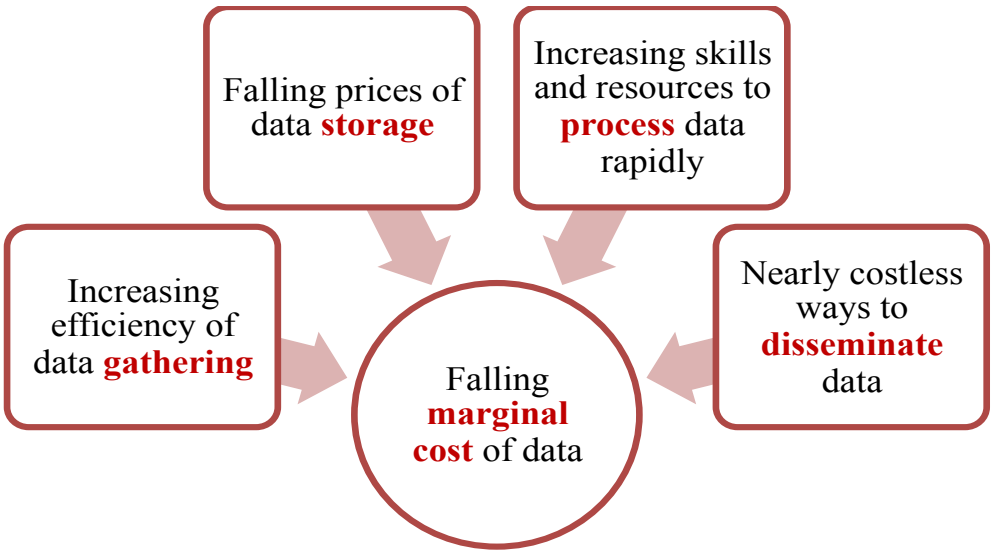
THE ECONOMICS OF DATA AND SOCIAL WELFARE

4.1 In recent years, the world has witnessed an information explosion – exponential increases in the amount of published data. As people increasingly use digital services to talk to each other, look up information, purchase goods and services, pay bills, transact in financial markets, file taxes, avail welfare services and engage with local leaders, data is being generated at an unprecedented scale. The global data infrastructure has largely proved reliable, fast and secure enough to handle this deluge of data (Mckinsey, 2011).

4.2 Concurrent with this data explosion, the marginal cost of data has declined exponentially and the marginal benefit to society of using this data is higher than ever.

4.3 As people shift their day-to-day activities online, they leave digital footprints

Figure 1: Falling Marginal Cost of Data



of these activities. Put differently, people produce data about themselves and store this data on public and private servers, every day, of their own accord. Data that would have involved a laborious survey to gather a few decades ago is today accumulating online at a near-zero cost, although it is scattered across sources.

4.4 Not everyone participates in the digital economy, of course. A majority of the poor still have no digital footprint. Among those who do, the range of activities undertaken online is quite limited. However, the cost of gathering data is still much lower than it was a few decades ago. Even if a door-to-door survey is the only way to gather a certain kind of data, we possess cheap technologies to log data online in real time, circumventing an otherwise laborious paper-based survey followed by a tedious data entry process.

4.5 Alongside the decreasing cost of gathering data, storage costs have decreased precipitously. The cost per gigabyte of storage has fallen from ₹61,050 in 1981 to less than ₹3.48 today. However, the surfeit of data and a limitless capacity to store it is of no use unless one can make sense of these colossal quantities of data in a reasonable time. Fortunately, human and technical capital to process data has evolved in parallel to the data inundation. Data science has evolved as a distinct, well-funded field of study that is constantly innovating ways to put data to efficient use. Courses in analytics have become ubiquitous. An increasing number of people are equipped with skills to handle large datasets. Data is still relatively expensive to process because it tends to be noisy, heterogeneous and inconsistent across sources, but technology is incessantly developing solutions to these problems.

4.6 Once processed, the cost of disseminating insights is negligible – it is nearly costless to transfer information through the internet. However, dissemination of data entails another cost – that of ensuring data privacy and security. Before data is disseminated, it needs to be stripped of personal identifiers and aggregated. While this is a direct cost, an indirect cost also exists – the cost of misuse of data. Accidental data leaks may bring forth legal consequences and substantial financial implications. However, technology has largely kept pace to mitigate

these risks.

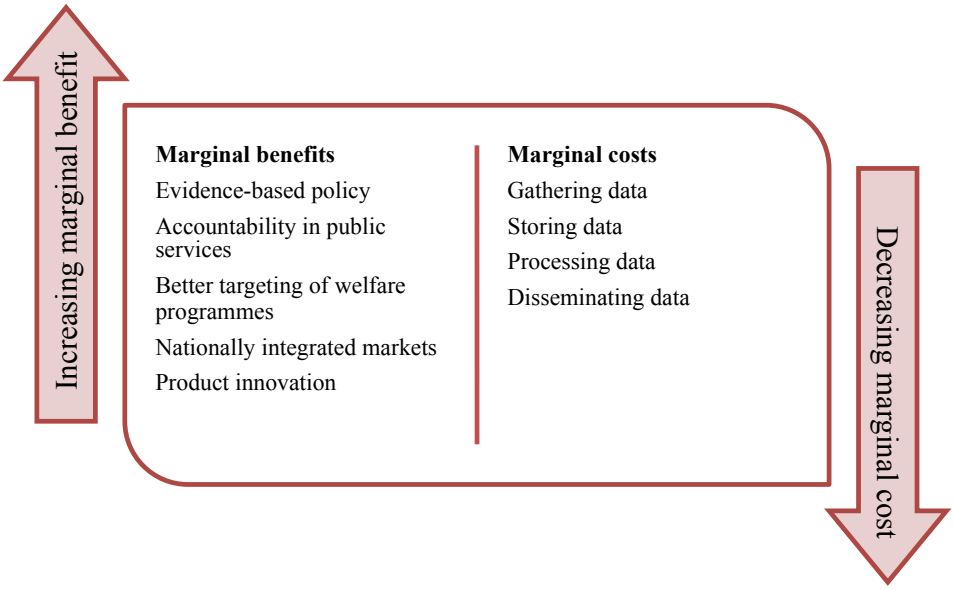
4.7 Together, the advancements in gathering, storing, processing and dissemination have lowered the marginal cost of data to unprecedented levels. Figure 1 summarises this phenomenon. Concurrently, the marginal benefit of data is higher than ever. A district education officer can make better decisions if he knows, for each school in his district, attendance rates of students and teachers, average test scores and status of school toilets. Similarly, parents can make better decisions about which school to send their children to if they know the average absenteeism rate of teachers in their village and can compare the rate to that in the neighbouring village. A multitude of scenarios exist in which harnessing the marginal unit of data can lead to sharp increases in public welfare.

4.8 Consider data that the Government maintains about its citizens. Currently, much of the data is dispersed across different registries maintained by different ministries. This is why every time a citizen has to access a new service, they are asked to collect all the documents to prove their identity and prove their claim on the process. For example, to get subsidies or benefits due to a farmer, such as free electricity, an incomplete list of documents include the Ownership Certificate issued by Village Administrative Officer, Chitta Adangal (extract from land registry), a Patta (Record of Rights) or Sale Deed, No Objection Certificate from any Government Project nearby and other documents to prove that one is a farmer on that land. The effort involved in this application mostly doesn't involve any new audits or inspections from any government department. The citizen faces the inconvenience of having to retrieve data trapped in paper files within the government system to unlock a benefit she is entitled to. The government can thereby deliver a better experience to the citizen by bringing disparate datasets scattered across various ministries together.

4.9 If the information embedded in these datasets is utilised together, data offers potential to reduce targeting error in welfare schemes. For example, consider a hypothetical individual who is affluent enough to own a car but is able to avail BPL welfare schemes, though unwarranted. When datasets are unconnected, the vehicle registry does not speak to, say, the public distribution system registry. Consequently, the public distribution system continues to subsidise this individual erroneously. However, if the two datasets are integrated, such inclusion error can be minimised, saving valuable Government resources. In the same way, exclusion errors can be rectified.

4.10 In fact, the declining cost of data has spawned new benefits that did not exist a few years ago. Consider, for example, an app that informs farmers of the prices of produce across the country. There was a time when, even if the app existed, it would have been useless – knowing that prices are higher in the next district would not have mattered to a farmer as he would have had no way to access those prices without substantial costs in transport, storage and distribution (including the cost of a dozen middlemen). Today, with platforms such as e-NAM (the electronic National Agriculture Market, we possess the technology to allow farmers to make the sale online, locking in the higher price with delivery to follow seamlessly. In this way, data has the potential to integrate markets nationally, reduce the need for middlemen, reduce prices for end consumers and increase prices for farmers. Figure 2 summarises these costs and benefits.

Figure 2: Increasing Marginal Benefits and Decreasing Marginal Cost of Data



WHY MUST DATA BE TREATED AS A PUBLIC GOOD?

4.11 The decline in marginal cost of data clubbed with an increase in marginal benefit means that the optimal quantum of data that society should consume is much higher than before. If so, economic theory predicts that the economy should have, by now, seen a surge in efforts to harness and use data. This has indeed happened, but only partially.

4.12 Private sector investment in data-related endeavours is higher than ever before. A 2017 Forbes survey found that 53 per cent of companies actively use big data to make decisions (Dresner, 2017). The trend holds across industries as disparate as healthcare and financial services, and across geographies and company sizes. In fact, in the last two decades, the world has witnessed the emergence of companies, such as Facebook, Amazon, Instagram, etc., who earn revenue exclusively from people's data.

4.13 But, there are several areas where data is not as ubiquitously harnessed and used. Consider, for example, the agriculture market. If the marginal benefit to a farmer of acquiring price information is higher than the marginal cost of that information, he would pay for that information. Consequently, the private sector would cater to his need by gathering and selling him the information he wants. This would eventually lead to a nationally integrated agriculture market with one price. However, India does not yet have such a nationally integrated agriculture market, which should have happened if the marginal benefit of data today is indeed higher than the cost. Why has the corporate sector's data wave not found a parallel in the agriculture sector?

4.14 The economics of data we considered so far is as it pertains to society, not to an individual agent, be it a farmer or a firm. The economics facing a private company that contemplates providing data to farmers is different from that facing a social planner. To the private firm, the prospect of a nationally integrated agriculture market, and the resulting social welfare, is a positive externality. It is not a benefit that accrues to the firm as the firm cannot, in practice, charge the numerous agents in the economy who

would experience welfare gains. Therefore, the private firm's marginal benefit is not as high as society's marginal benefit. Because the firm does not internalize the benefit of social welfare, the optimal amount of data that the firm would gather and disseminate falls short of the social optimum.

4.15 Second, data comes in many forms with each form offering a different benefit. Data linked to an individual can range from extremely intimate – such as their biometric details, to the extremely public – such as their name. It includes data that is generated by human actions, and data that is derived through analysis, typically involving an algorithm. For example, whether an individual has paid his taxes is a generated data point. But, using the tax records and other data points, a credit bureau may assign the individual a credit score, which is a derived data point. Data that is not linked to a specific individual but is still available at an individual level of granularity, is called Anonymized data. Anonymized data is critical in some areas such as medical research. Data neither linked to an individual, nor at an individual level of granularity is known as public data, such as the census.

4.16 While the private sector has done an impressive job of harnessing some kinds of data – the kind that can be converted into a private profit – government intervention is required in other areas where private investment in data remains inadequate. The social sectors of the economy, such as education and healthcare, have lagged the commercial sectors in exploiting data. Because the private sector cannot internalize the social benefits of data in these sectors, the market for data in these sectors has so far not developed.

4.17 To ensure that the socially optimum amount of data is harvested and used, the government needs to step in, either by providing the data itself or correcting the incentive structure faced by the private sector, depending on the nature and sensitivity of data. Indeed, in the agriculture sector, the Government has done exactly this by creating the e-NAM, as it is unlikely that the private sector would come up with a solution like this on its own.

Ensuring data privacy while creating data as a public good

4.18 In the endeavour to create data as a public good, it is very important to consider the privacy implications and inherent fairness of data being used. Needless to say, the processes required for ensuring privacy of intimate data is very different from that required for anonymized or public data. The key difference in dealing with these different types of data is the knowledge and consent of the data principal. Even if not explicitly mentioned every time data is talked about in this chapter, it is assumed that *the processing of data will be in compliance with accepted privacy norms and the upcoming privacy law, currently tabled in Parliament.*

Economies of scale and scope in the data generation process

4.19 Apart from the obvious necessity that data must be accurate, the need for a Government-driven data revolution is motivated by three key characteristics that data must possess for the synergistic benefits to accrue. Specifically, the data generation process exhibits significant economies of scale and scope.

4.20 First, when it comes to data, the whole is larger than the sum of its parts, i.e., it is more useful when it is married with other data. Consider, for example, the merging of disparate datasets maintained by different government agencies, such as transactions data extracted from the Jan Dhan accounts of the Department of Financial Services, Ministry of Finance, married to demand

for MGNREGA work from the Ministry of Rural Development. As MGNREGA can be a real-time indicator of rural distress (as discussed in Chapter 10 in this volume), the credit scoring done using the transactions data of Jan Dhan accounts can be used to provide credit in districts/panchayats that are experiencing distress. Such combining of disparate datasets can be extremely useful in obtaining the necessary richness required to design and implement welfare policies.

4.21 Second, data needs to cover a critical mass of individuals/firms so that comparisons and correlations can be assessed among individuals/firms to generate useful policy insights. Thus, to gather price data on trades across various product markets and across the country, a very large number of producers and buyers need to log their transactions on a platform in real time. To induce numerous agents to report transaction information regularly is a task that requires significant initial investment, which may prove prohibitive for the private sector.

4.22 Finally, data must have a long enough time-series so that dynamic effects can be studied and employed for policymaking. For instance, to undertake before-after evaluations to assess the effectiveness of policies, data that spans a long-enough time series is critical.

4.23 Data that contains all these three features is much more valuable than three different and disparate datasets that possess each of these criteria separately. Thus, the data generation process exhibits significant economies of scope. Also, the scale of effort required to create such data that exhibits all three features implies that the data generation process exhibits significant economies of scale as well because the (upfront) fixed costs involved in generating such data are significant. Private sector may not have the risk appetite or the capital to make the necessary investments required for generating data that possesses all three characteristics, viz., marries disparate datasets, covers a critical mass of individuals/firms, and spans a large time-series. Even if private sector were to put such rich data together, this would result in a monopoly that would reduce citizen welfare, on the one hand, and violate the principle of data by citizens, and, therefore, for citizens.

4.24 Most importantly, data carries some of the characteristics of public goods. It is non-rivalrous, i.e., consumption by one individual does not reduce the quantum available for others. In principle, data can be made excludable, i.e., it is possible to exclude people from accessing data, as many database firms do by erecting pay walls. However, there are some kinds of data – particularly data gathered by governments on issues of social interest – that should be democratised in the interest of social welfare. Such data should be made public goods. As the private sector would fail to provide an optimal amount of any non-rivalrous, non-excludable good, government intervention is required.

4.25 Data are generated by the people, of the people and should be used for the people. As a public good, data can be democratised and put to the best possible use. Box 1 describes the Open Government Data initiative taken by the Government, which is an illustration of the spirit of data as a public good. While this is an excellent start, the enormous benefits that can be reaped from treating data as a public good imply that Government must redouble its efforts in this direction.

BUILDING THE SYSTEM

4.26 The data system envisioned here involves predominantly data that people share with Government bodies with fully informed consent or is data that is legally sanctioned to be collected by the state for

Box 1: Open Government Data and citizen engagement

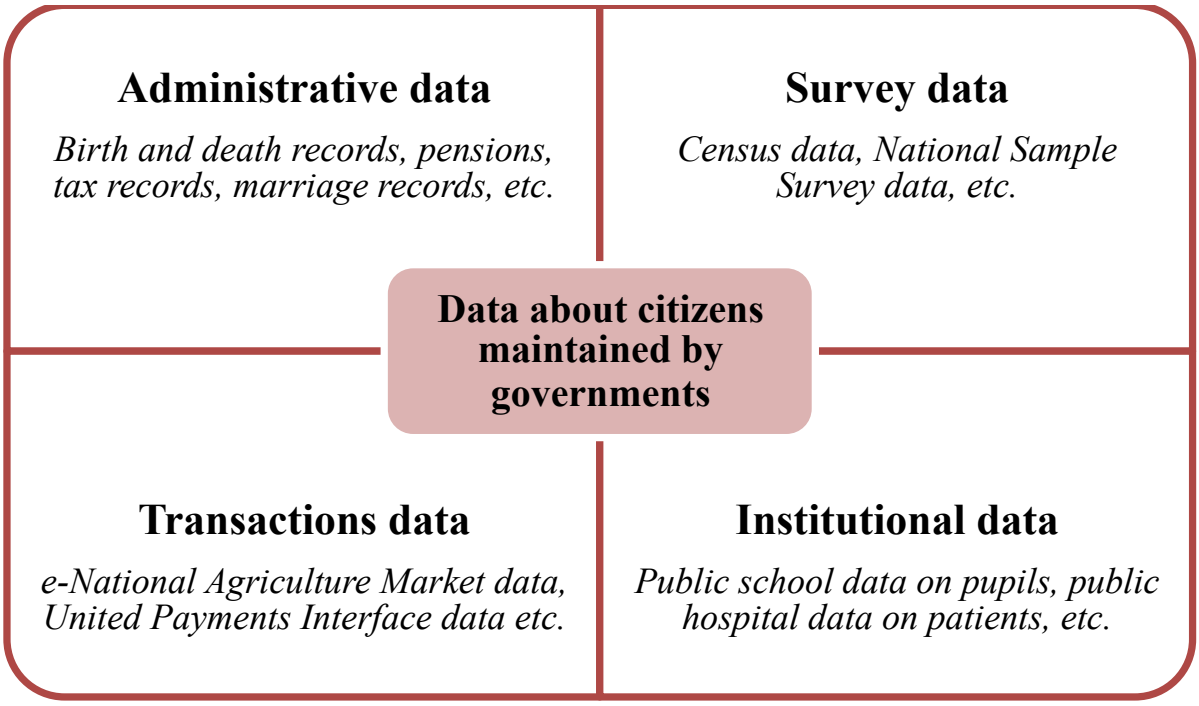
The Union government's Open Government Data platform allows citizens to access a range of government data in machine-readable form in one place. The portal allows union ministries and departments to publish datasets, documents, services, tools and applications collected by them for public use. Excluding datasets which contain confidential information, all other datasets are made available to the public, ranging from data on welfare schemes to surveys to macroeconomic indicators. The platform also includes citizen engagement tools like feedback forms, data visualisations, Application Programming Interface (APIs) etc.

Open data not only helps government officials make better decisions but also gets people involved in solving problems. Throwing open government data to the public multiplies the number of people analysing and deriving insights from data. Consequently, the usability of data itself increases.

To engage people meaningfully in solving problems, the Ministry of Human Resource Development recently initiated the Smart India Hackathon – an open innovation model to discover new, disruptive technologies that could solve India's most pressing problems. Smart India Hackathons are product development competitions in which participants get a problem statement and relevant data, using which they develop a prototype software or hardware. These competitions crowd-source solutions to improve governance and increase the efficacy of welfare schemes. None of this would be possible, of course, without reliable data.

an explicit purpose such as tax collection, or delivering welfare. The Government of India collects four distinct sets of data about people – administrative, survey, institutional and transactions data (see Figure 3 for summary and Box 2 for details). While the latter two databases are in a fledgling state, the first two are comprehensive and robustly maintained.

Figure 3: Data collected by the Government of India



Box 2: Rich data on citizens that Government can harness for the welfare of its citizens

Governments hold *administrative data* for mainly non-statistical purposes. Administrative datasets include birth and death records, crime reports, land and property registrations, vehicle registrations, movement of people across national borders, tax records etc. Governments also gather data to evaluate welfare schemes; for example, the Ministry of Drinking Water and Sanitation gathers data on toilet usage to assess the efficacy of the Swachh Bharat Mission.

Survey data, on the other hand, is data gathered predominantly for statistical purposes through systematic, periodic surveys. For example, the National Sample Survey Office conducts large-scale sample surveys across India on indicators of employment, education, nutrition, literacy etc. Because these data are gathered for statistical analyses, the identity of participants is irrelevant and unreported, although these identities may be securely stored at the back-end without violating any legal guidelines on privacy.

Institutional data refers to data held by public institutions about people. For example, a government-run district hospital maintains medical records of all its patients. A government-run school maintains personal information about all its pupils. State-run universities maintain records of students' educational attainment and degrees awarded to them. Most such data are held locally, predominantly in paper-based form. This data can be digitized to enable aggregation at the regional or national level.

Transactions data are data on an individual's transactions such as those executed on the United Payment Interface (UPI) or BHIM Aadhaar Pay. This is a nascent category of data but is likely to grow as more people transition to cashless payment services.

4.27 Data collection in India is highly decentralised. For each indicator of social welfare, responsibility to gather data lies with the corresponding union ministry and its state counterparts. Consequently, data gathered by one ministry is maintained separately from that gathered by another. Data on an individual's vehicle registrations, for example, is maintained by one ministry, whereas the same individual's property ownership records lie with another ministry. These datasets are further distinct from the individual's educational attainment records maintained by the state-run university he/she attended and from other demographic information gathered in the decennial Census survey. Because these datasets are unconnected (see below), each ministry only has a small piece of the jigsaw puzzle that is the individual/firm. However, if these different pieces could be put together, we would find that the whole is greater than the sum of parts. As learning from global best practices, boxes 3 and 4 highlight the case studies of Transport for London's data initiative and the data initiatives of the U.S. Government. Box 5 illustrates similar learning from closer home: the Samagra Vedika initiative by the Government of Telangana.

Box 3: Benefits of opening up Transport for London's data

Transport for London (TfL) releases a significant amount of data – such as timetables, service status and disruption information – in an open format for anyone to use, free of charge. Open travel data can support travel apps and real-time alerts to save time, reduce uncertainty and lower information costs, supporting growth in the tech economy and increasing the use of public transport. At last count, more than 600 apps were being powered specifically using the TfL open data feeds, used by 42 per cent of Londoners.

TfL has demonstrated that releasing data to the public can save users time to the economic value of between £15m and £58m per year. An analysis by Deloitte found that the provision of transport information through travel apps and real-time alerts is saving £70m-£95m per year in time, reduced uncertainty and lower information costs. Further, release of open data by TfL has supported the growth of London's tech economy to the value of £14m annually in gross value added (GVA) and over 700 jobs. It has also unlocked new revenue and savings opportunities and new ways of working at TfL, including a £20m increase in bus usage as customers are more aware of service opportunities. This data has been used by a range of apps, from early stage start-ups to global leading technology platforms, saving time and reducing stress. There are currently 13,700 registered users of TfL's open data.

The London Infrastructure Mapping Application is a new platform that allows utilities, boroughs, the Mayor of London and TfL to share information relevant to infrastructure investment and planning. By bringing together a range of data, the application facilitates improved collaboration between actors, joined-up approaches to construction and design, and better identification of future demand and capacity constraints. Information is visualised spatially through a bespoke mapping application that has been developed in consultation with users. Early evidence has found that the tool supports better alignment of investment – unlocking housing growth and reducing disruption throughout London – by allowing projects such as road works to be timed better, and saving costs through joined-up approaches to construction (e.g. joint ducting of utility cables and pipes).

TfL is currently developing a new tool that will analyse data feeds from Tube trains to provide maintenance staff with live information about the condition of a train. Using the tool, staff can analyse the data and identify where faults exist or might be developing and remedy them before they cause service issues. The tool has strong potential to make maintenance planning more efficient and prevent costly faults leading to service delays from occurring. The in-house capability will also help TfL save money by reducing third-party spend – currently around £46m over five years on an external maintenance support contract.

Source: National Infrastructure Commission Report on Data as a Public Good available at https://www.nic.org.uk/wp-content/uploads/Data-for-the-Public-Good-NIC-Report.pdf

Box 4: Data Initiatives taken by U.S. Government

On January 29, 2009, U.S. President Barack H. Obama issued a memorandum on open and transparent government and asked his administration to establish "a system of transparency, public participation, and collaboration". The open-government initiative mandates federal government and public agencies to publish their data online for public use in machine-readable format. In a bid to democratise its data, U.S. government made more than 138,000 data sets available to the public. In February 2015, the U.S. government appointed DJ Patil as its first ever Chief Data Scientist in the Office of Science and Technology Policy to unleash the power of data for the benefit of the American public.

Data.gov, the preeminent platform of U.S. Government shares data and meta-data from various public agencies and has 'value added' features like:

(i) Ability to filter by location, data-set type, tag, format, community etc.;

(ii) A tool to integrate data-sets and create visualizations;

(iii) Engagement with public to provide feedback and participation in various forums, blogs and communities to improve the quality of data-sets;

(iv) Resources that provide links to all federal agency API's;

(v) Developer hub with software development kits, open-source resources and repositories of code;

(vi) Enabling of data-driven decisions through education apps (search for college, search for public school districts), energy and environment apps (alternate fuel locator) and food and nutrition apps (fooducate, goodguide);

(vi) A platform of cities.data.gov to publish datasets from different cities across U.S.;

(vii) disaster.data.gov web portal for collaborative efforts in disaster management, healthcare, policing, and climate change;

(viii) Precision Medicine Initiative (PMI), a leading open data project that aims to create individualized treatments, expand genetically based clinical cancer trials and establish a national "Cancer Knowledge Network" to help with treatment decisions.

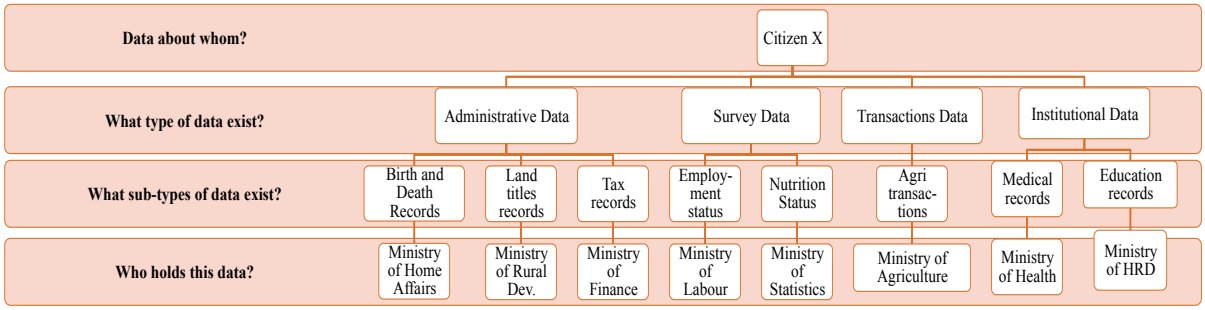
United States Department of Agriculture (USDA) releases regular forecasts for prices of various agricultural commodities that are used worldwide. The projections identify major forces and uncertainties affecting future agricultural markets; prospects for long-term global economic growth, agricultural production, consumption, and trade; and U.S. exports of major farm commodities and future price movements. The projections can also be used to analyse impacts of alternative policy scenarios. This easy availability of data projections, in effect, drives agricultural markets across the U.S.

Box 5: Federalism in learning among governments: Telangana Government's Samagra Vedika initiative

The Telangana Government's Samagra Vedika initiative gives a flavour of the potential benefits of integrating data sets. The initiative links around twenty-five existing government datasets using a common identifier – the name and address of an individual. Seven categories of information about each individual were linked in this aggregation exercise – crimes, assets, utilities, subsidies, education, taxes and identity information. Each individual was then further linked to relatives such as spouse, siblings, parents and other known associates. The initiative also puts in place all the necessary safeguards to preclude any tampering of data or violation of privacy. The right to add or edit data in the database varies by ministry or department. A given department can only write data for select fields – the motor vehicles department cannot, for instance, manipulate data relating to education, even though it can view the data.

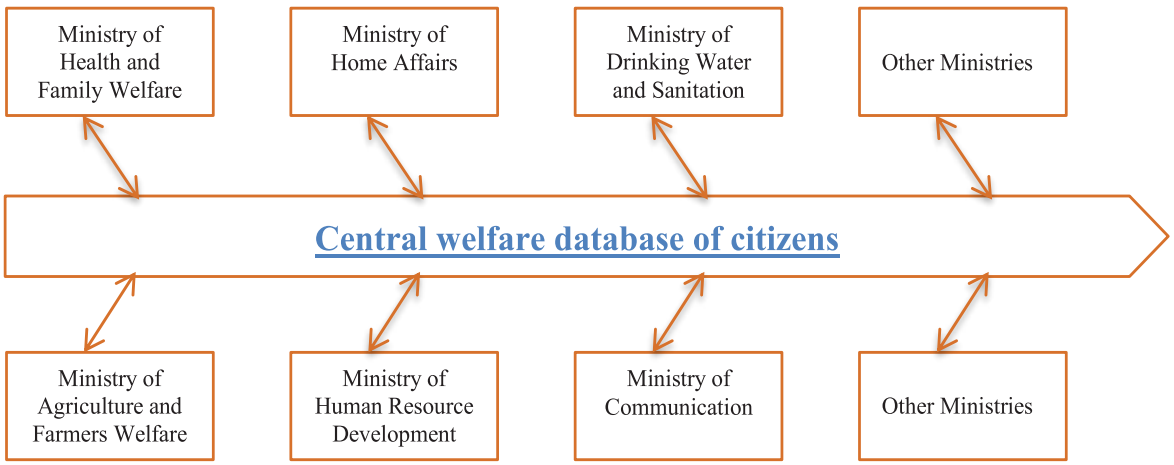
4.28 Of late, there have been some discussions around the "linking" of datasets – primarily through the seeding of an Aadhaar number across databases such as PAN database, bank accounts, mobile numbers, etc. A point of clarification is in order here. When one adds an Aadhaar number to an existing database such as a database of bank accounts, it is only one more column that is added to the table. The linking is so to speak "one-way". Banks can now use the tokenized Aadhaar Number (i.e., a proxy 64-digit number that is based on, but not equal to the individual's 12-digit Aadhaar Number) to combine duplicate records and weed out benaami accounts. But, this does not mean that the UIDAI or government can now read the bank account information or other data related to the individual.

Figure 4: Illustration: Data about a hypothetical citizen X is dispersed across multiple, unconnected datasets



Note: Names of some ministries abbreviated. Data types and sub-types are illustrative examples only and not exhaustive.

Figure 5: An enterprise architecture for Governance



4.29 Figure 4 illustrates the current system. Data about a hypothetical citizen X is spread across ministries. In fact, the illustration is an oversimplification. For example, data about X's medical records would lie with the Government-run district hospital where she was most recently treated. In all likelihood, these medical records are in paper-based form and do not get aggregated for analysis at the state level, let alone the central level. Nevertheless, the illustration depicts the union ministries ultimately responsible for each kind of data.

4.30 Figure 5 illustrates the Data Access Fiduciary Architecture. Each department of the government is responsible for making available the data they hold as a data provider. These departments must take care to appropriately treat private data and public data with the standards they require. This data is then made available through a Data Access Fiduciary to the Data Requestors. Data Requestors may be public or private institutions but can only access the data if they have appropriate user consent. The Data Access Fiduciary themselves have no visibility on the data due to end-to-end encryption. *Such a model puts user consent in the centre of the government's initiative to make Data a Public Good.*

4.31 A citizen is, of course, not the only possible unit of analysis. One may want to analyse schools, villages or hospitals. Different databases of villages, for example, may be utilised together if a unique identifier for every village exists. Unfortunately, in many cases, ministries and departments have their own codes for a given geographical area; for example each village is characterised by a pin code assigned by the Department of Post, a village code assigned by the Department of Rural Development and a health block assigned by the Ministry of Health and Family Welfare. The lack of a common identifier makes it difficult to consolidate information.

4.32 An initiative to address this issue is the Local Government Directory, an application developed by the Ministry of Panchayati Raj. A comprehensive directory of all local administrative units, the platform maps each land region entity to a local Government body (like villages with their respective gram panchayats) and assigns location codes compliant with Census 2011. The Local Government Directory is a great example of Data as a Public Good. The Ministry of Panchayati Raj has made important headway in solving a difficult problem, common to every government and private institution trying to serve rural India – the lack of formal addresses – by assigning a code to every place. Instead of simply sitting on that data, the Ministry has also published it for all to use. If all government databases requiring location codes are aligned with the codes in this directory, then all databases will share a common location field that can help in merging data, and reduce accuracy errors in the distribution of welfare.

Utilizing technological advances to eliminate all privacy concerns

4.33 The integrated system's efficacy relies on three critical features. First, while any ministry should be able to view the complete database, a given ministry can manipulate only those data fields for which it is responsible. Second, updating of data should happen in real time and in such a way that one ministry's engagement with the database does not affect other ministries' access. Third and most importantly, the database should be secure with absolutely no room for tampering.

4.34 The prospect of empowering the government with such comprehensive, exhaustive information about every citizen may sound alarming at first. However, this is far from the truth. First, large quantities of data already exist in government records, and the objective is only to use this data in a more efficient way. The proposal envisioned here does not gather any new information; rather, it seeks to make available all data within the government for citizens, government, private and public institutions to utilize the data subject to user consent and appropriate privacy and fairness related constraints.

4.35 Second, people can always opt out of divulging data to the government, where possible. For example, one can choose not to participate in a survey or use government-run payments services. There are exceptions, of course. People cannot buy and drive vehicles without a license and registration certificate; but not requiring these data would threaten the enforcement of property rights, road safety and national security, which cannot be compromised. But for the remaining categories of data – institutional, survey and transactions – people have the choice.

4.36 Third, even if there is no viable private market choice of certain public services, the choice to share the individually linked data from such services will always be with the citizen under the Data Access Fiduciary Architecture. Further, immutable access logs

for all data would be available so that citizens know who has seen their data and why.

4.37 The principle is that most data are generated by the people, of the people and should be used for the people. Enabling the sharing of information across datasets would improve the delivery of social welfare, empower people to make better decisions, and democratize an important public good.

TRANSFORMING INDIA'S DATA INFRASTRUCTURE

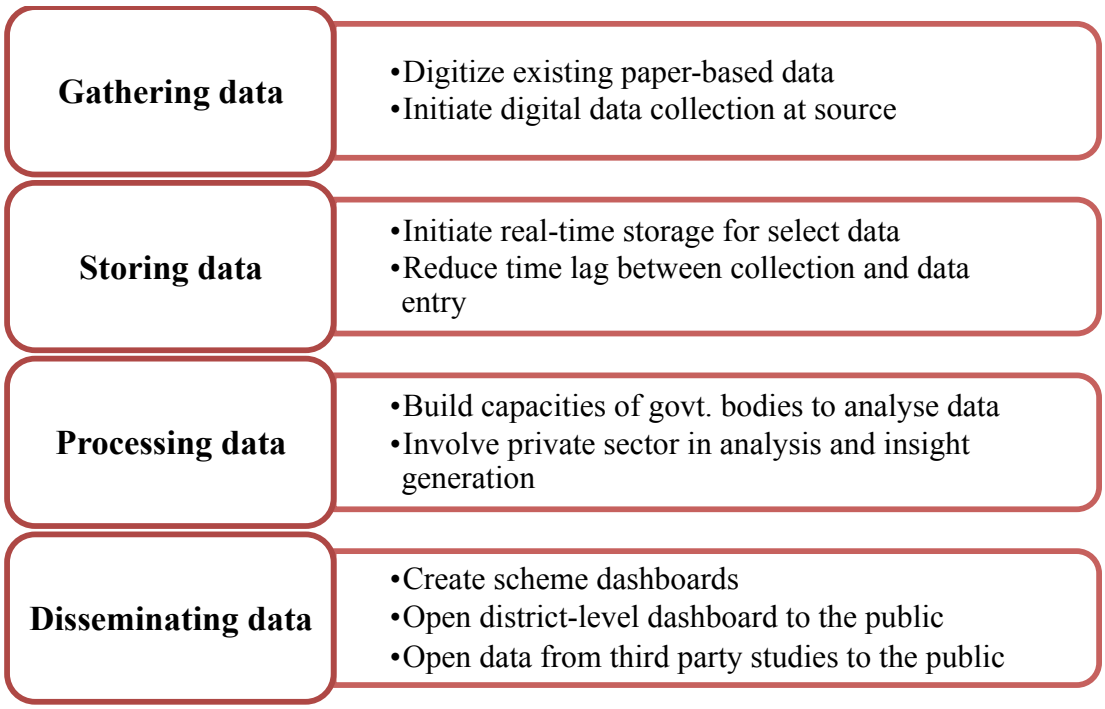
4.38 Undoubtedly, the system outlined above requires a robust data infrastructure. While combining datasets will itself reap rewards, the benefit is limited if data is of an uncertain quality, not amenable to easy processing etc. Harnessing data consists of four steps – gathering, storing, processing and disseminating data, each of which offers room for improvement in India (see Figure 6).

Gathering data: Directly digital rather than paper to digital

4.39 Unless data is in a digital, machine readable format, its utility is limited. Paper records of data cannot be accessed by all stakeholders. Even if paper records of schemes are scanned and stored in a digital repository, they do not lend themselves to easy analysis. For data to serve its purpose, it needs to be transmuted into a digital, machine readable form that can be downloaded and analysed. While most state and national level data is already in this form, digital records are not yet ubiquitous at the more granular administrative levels. For example, a citizen cannot view his village's sanitation status on a digital database. At best, he can look up his district. For many schemes, data is available only at the state level.

4.40 The recently launched Digitize India initiative is an ingenious solution to the

Figure 6: Transforming data gathering, storage, processing and dissemination



tedious task of converting paper-based data into digital form. The initiative crowd-sources the data entry effort by presenting volunteering citizens with snippets of scanned documents, which they type into a data entry portal. Correct entries earn cash rewards for the contributor (see Box 6 for details).

4.41 While the scheme goes a long way to digitize existing paper records, a parallel initiative is needed to convert the very process of data collection into a digital one, as opposed to collecting on paper first and converting to a digital format later. Undoubtedly, the exercise of supplying every hospital, gram panchayat, school or block office with an electronic data-entry device is quite ambitious, but achievable. Digital data collection at the source ensures that data is logged exactly as observed, obviates the redundancy of data entry, and eliminates data entry errors. A truly digital ecosystem is achievable only when all layers of information that constitute the ecosystem are digitized.

Storing data

4.42 Public service delivery can benefit from real-time storage of data. The city of St. Petersburg, Florida, recently created a citizen dashboard called St. Pete Stat, where citizens can view dynamically updated information about the performance of city departments, status of development projects. The dashboard includes various interactive tools such as an interactive map of all police calls in the city.

4.43 Not all types of data are amenable to real-time storage, of course. Further, the incremental cost of a real-time database may not always justify its marginal benefit. However, there are a few sectors where, if not real-time, at least a high frequency of logging data on the server can yield substantial benefits.

4.44 Education is one such example. Even a seemingly short period of six months is a long time in the life of a child – it amounts to half an academic year. Say a school does not have a functional toilet for girls, leading girls to remain absent from school. The problem needs to be rectified as soon as possible; otherwise these girls would lose out on valuable schooldays. The block or district education officer should receive a daily, or at least weekly, report of the status of toilets so that in case a problem is not rectified within a week or two, they can take the required action.

Box 6: How Digitize India works

Government departments upload scanned copies of paper records on the Digitize India platform. These scanned documents are shredded into snippets with meaningful data. These snippets are randomly served to digital contributors. Digital contributors are citizens who volunteer their time on the portal. Upon receiving snippets, the contributor reads the information and types it into a data entry portal. All converted data are verified against the corresponding snippet. Correct entries earn their contributors reward points, which can either be redeemed for cash or donated to the Digital India initiative. Once all snippets corresponding to a particular document are converted into digital data, the platform reassembles the document in digital form and supplies it back to the government department. Any Indian citizen with an Aadhaar number can participate as a digital contributor. Citizens are incentivised through reward points and recognition as 'digital contributors' and may even earn certificates as 'Data Entry Operators'. Ingeniously, the program also features a mobile app so that even citizens without a laptop or desktop computer can participate.

If the problem takes months to come to the notice of the block or district officer, the girl children in the school will have lost months' worth of learning! Digital dashboards updated in real time or at least weekly would avert such harsh consequences. Similarly, consider agriculture, where farmers need data about weather conditions, expected rainfall, input and output prices in real time to make daily decisions. Facilitating real-time access to such data is essential to improve agricultural productivity and farmer incomes.

4.45 With the widespread adoption of ICT approaches in public service delivery, real time data collection and storage is no longer an ambitious and distant dream but very much realisable, at least in select sectors and contexts.

Processing data

4.46 A deluge of data will be created when data is collected digitally, stored in real-time, and utilised with existing data. While this deluge has tremendous potential to transform governance, unleashing this potential requires skill. A district government official, for example, should have the analytical skills to make use of the data in that district. In the absence of such skill, investing in the data infrastructure is of limited use.

4.47 Governments at all administrative levels should invest in building their internal capacities to exploit data in real time, perform analyses and translate data into meaningful information. While every government department may have a dedicated analytics or data insights division, Ministry of Statistics and Program Implementation and Ministry of Electronics & Information Technology can act as nodal departments to steer such efforts at the national level.

4.48 The government may also consider opening certain kinds of data to private players with all the necessary security safeguards. Data shared with the public on the Open Government Data portal is completely anonymised and aggregated over a large number of individuals. Such data precludes careful statistical analyses, where individual-level observations are required. However, after obfuscating all personally identifiable information, if this data is shared with the private sector, government can harness the skills and enthusiasm of data analytics professionals to gain the maximum possible insights from the data.

Disseminating data

4.49 The Open Government Data portal is an effective tool to disseminate data to the public. But, as most citizens do not have the time or skills to employ analytical tools to dissect databases, easy visualisation tools are critical. The portal has enabled several visualisation tools already, and these may be augmented with the following.

4.50 First, the Government may initiate scheme dashboards for every major Government scheme with granular data, all the way to the village level. Dashboards should be ready recipients of data from the concerned government body and ready displayers of the same data in real-time to the public. The Swachh Bharat Mission is an exemplar dashboard that may be replicated for other schemes to allow citizens to track the physical and financial performance of welfare schemes. Currently, these dashboards exist only for a small number of schemes. For example, ICDS Anganwadi Services scheme, the largest scheme for women and child development in India, does not have a publicly available dashboard.

4.51 In existing dashboards, the metrics showcased may be augmented, especially to show progress at the more granular levels of administration. For example, the Swachh Bharat dashboard can include information on the number of swachhta doots (sanitation helpers) deployed at the district level or the number of information, education and communication (IEC) campaigns in that district. In all likelihood, the data on these metrics is available with the government but is not open to the public, perhaps because the data is not in digital form.

4.52 Second, many state governments have instituted district-level dashboards based on Management Information Systems (MIS) for various programmes. These states include Andhra Pradesh, West Bengal, Madhya Pradesh and Chhattisgarh. Except for the Andhra Pradesh and West Bengal ones, these dashboards are not easily accessible to citizens. They require a password, which is only available to the local administration. Some dashboards are not operational anymore (broken links), others display outdated data. While the investment to initiate such a dashboard is commendable, the benefits will be greater if the state governments continually maintain these dashboards to disseminate information.

4.53 Finally, many central ministries and state departments commission data collection initiatives to conduct needs assessments or impact evaluations of schemes. Although most of these studies are made public, the underlying data are not. As these studies are carried out in partnership with government bodies, they should be made available to the public so that independent analyses may be carried out to validate the findings of these studies.

APPLICATIONS

4.54 Once the infrastructure is in place, the applications are innumerable. A robust data backbone can empower every stakeholder in society, from the Central Government to a local government body, from citizens to

Box 7: NREGAsoft and e-governance in MGNREGA

NREGAsoft is a comprehensive e-governance system for the MGNREGA scheme. Accessible by a range of stakeholders, it captures the complete flow of all MGNREGA work at every level – from the centre all the way to the panchayat. In the spirit of citizens' right to information, the system makes available documents like muster rolls, registration application register, job card/employment register/ muster roll issue register and muster roll receipt register, which are otherwise inaccessible to the public.

The system has no language barriers to usage; it is accessible in a number of local languages. In fact, even the illiterate can use the interface as it leverages sounds and icons in a touch-screen kiosk model. It is designed to be used by a range of stakeholders, from workers who are beneficiaries of the scheme to gram panchayats to district programme coordinators to banks and post offices. Even citizens who are not beneficiaries of the scheme may view information on the portal.

The software consists of several modules, which together comprehensively span all activities and all stakeholders. For example, while the worker management module forms the backbone of all worker-related services, a fund management module tracks the movement of funds from the central ministry all the way to the workers' pockets, a grievance redressal module helps stakeholders including the illiterate to lodge complaints and track responses, and a bank/post office module allows financial institutions to get wage information and enter details of money credited in accounts. Other modules assist with cost estimation, social audit, knowledge network etc.

the private sector. A few (inexhaustive and illustrative) potential benefits are described below.

Governments themselves as beneficiaries

4.55 Being able to retrieve authentic data and documents instantly, governments can improve targeting in welfare schemes and subsidies by reducing both inclusion and exclusion errors. Datasets that utilise information across various datasets can also improve public service delivery. For example, cross-verification of the income-tax return with the GST return can highlight possible tax evasion.

Private sector firms as beneficiaries

4.56 The private sector may be granted access to select databases for commercial use. Consistent with the notion of data as a public good, there is no reason to preclude commercial use of this data for profit. Undoubtedly, the data revolution envisioned here is going to cost funds. Although the social benefits would far exceed the cost to the government, at least a part of the generated data should be monetised to ease the pressure on government finances. Given that the private sector has the potential to reap massive dividends from this data, it is only fair to charge them for its use.

4.57 Consider, for example, allowing the private sector access to data about students' test scores across districts (with all personal information completely obfuscated). Using test scores of students, demographic characteristics of each district and publicly available data on the efficacy of public education schemes, a private firm may be able to uncover unmet needs in education and cater to these needs by developing innovative tutoring products tailored to the specific needs of specific districts. These products would not only create profits for the private sector, but also monetize data and generate revenues for the government, in addition to improving education levels and social welfare.

4.58 Alternatively, datasets may be sold to analytics agencies that process the data, generate insights, and sell the insights further to the corporate sector, which may in turn use these insights to predict demand, discover untapped markets or innovate new products. Either way, there is tremendous scope for the private sector to benefit from the data and they should be allowed to do so, at a charge. Fortunately, stringent technological mechanisms exist to safeguard data privacy and confidentiality even while allowing the private sector to benefit from the data.

Citizens as beneficiaries

4.59 Citizens are the largest group of beneficiaries of the proposed data revolution. Consider the case of Digital Locker. It is in many ways similar to the plan we have outlined above. But it is restricted to certain

Box 8: The National Scholarship Portal

The Government of India has already made headway in integrating data on scholarships. The National Scholarship Portal was initiated to harmonize all scholarship schemes implemented by various ministries at the central and state levels. The portal serves as an umbrella platform for all scholarship related services ranging from student application, application receipt, processing and sanction to disbursal of funds. In addition to creating a transparent database of all beneficiaries of all government scholarship schemes at various levels, the portal reduces hassles in discovering scholarships and facilitates direct benefit transfer (DBT).

documents that the state issues. Citizens no longer need to run from pillar to post to get "original" documents from the state such as their driving licence, Aadhaar card, PAN card, CBSE results, etc. These documents are critical in the life of every resident of India. These documents are most needed by those who depend on the state for welfare. They are also often the hardest to secure for the same vulnerable group.

4.60 DigiLocker makes all their documents available, in a verified format, in one place on the cloud. Citizens only need an internet connected device, smartphone or computer to access the locker. It helps digitize downstream processes such as college admissions. For anyone who has had to retrieve a lost or missing document from the government or have to get photocopies "attested", the DigiLocker experience is immensely more time-saving and user-friendly. They do not have to live in the fear of their precious documents being lost to the elements or other misfortune.

4.61 In a similar vein, the Reserve Bank of India has announced the Non-Banking Financial Company-Account Aggregator (NBFC-AA). Even in the finance industry, an individual leads a complex financial life with their data spread across multiple providers. One may hold a bank account with State Bank of India, but take a loan from HDFC. Their mutual fund investments may be through Axis Bank, but their insurance is through LIC. They also have a credit card from Standard Chartered and use Motilal Oswal to invest in some stocks directly. In such a case, the data needed by the individual to piece together their own financial life is distributed across many data providers. The NBFC-AA allows users to pull that data together, for any purpose the citizen requires. This may be for personal finance management, or maybe to apply digitally for a new housing loan. The NBFC-AA neither reads data nor creates an invasive 360-degree dataset. It simply enables citizens to demand their data from these institutions in a machine-readable format, so that it can be used by them meaningfully.

WAY FORWARD

Through Aadhaar, India has been at the forefront of the data and technology revolution that is unfolding. As data for social welfare may not be generated by the private sector in optimal quantity, government needs to view data as a public good and make the necessary investments. The benefits of creating data as a public good can be generated within the legal framework of data privacy. Going forward, the data and information highway must be

Box 9: Idea of a National Health Registry

For Swachh Bharat to transform into Swasth Bharat and eventually Sundar Bharat, citizens' health is paramount. Prevention is far more important in this endeavour than cure. A national health register, that maintains health records of citizens with all the necessary privacy safeguards can go a long way in enabling health analytics for predictive and prescriptive purposes. Such a national health register would be identified using a citizen's Aadhar. As a doctor can access the medical history of a patient from this national health register, this facility would be especially useful in emergency/trauma cases and can potentially save several lives. The various components of this register can include databases for (i) hospitals and public health centres, (ii) surveillance of syndromes, (iii) immunization information systems, (iv) electronic laboratory reporting, and (v) sub-registries for key diseases requiring intervention such as diabetes, hypertension, cancer, AIDS, etc. Anonymized data from the register can be sold to private parties for analytics, which would then enhance prevention by offering predictive and prescriptive knowledge.

viewed as equally important infrastructure as the physical highways. Such a stance can help India leapfrog to utilise the benefits of technological advances for the welfare of its people. In the spirit of the Constitution of India, data "of the people, by the people, for the people" must therefore become the mantra for the government.

CHAPTER AT A GLANCE

- Given technological advances in gathering and storage of data, society's optimal consumption of data is higher than ever.
- As private sector may not invest in harnessing data where it is profitable, government must intervene is creating data as a public good, especially of the poor and in social sectors of the country.
- Governments already hold a rich repository of administrative, survey, institutional and transactions data about citizens, but these data are scattered across numerous government bodies. Merging these distinct datasets would generate multiple benefits with the applications being limitless.
- Given that sophisticated technologies already exist to protect and share confidential information, data can be created as a public good within the legal framework of data privacy. In thinking about data as a public good, care must also be taken to not impose the elite's preference of privacy on the poor, who care for a better quality of living the most.
- As data of societal interest is generated by the people, it should be "of the people, by the people, for the people."

REFERENCES

Brook, E., Rosman, D., Holman, C. 2008. "Public good through data linkage: measuring research outputs from the Western Australian data linkage system." *Australian and New Zealand journal of public health* 32(1): 19-23.

Dresner Advisory Services, LLC. 2017. *Big Data Analytics Market Study.* https://www.microstrategy.com/getmedia/cd052225-be60-49fd-ab1c-4984ebc3cde9/Dresner-Report-Big_Data_Analytic_Market_Study-WisdomofCrowdsSeries-2017.pdf

Hilbert, Martin. 2016. "Big Data for Development: A Review of Promises and Challenges." *Development Policy Review* 34:135-174.

James M., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., and Angela Hung Byers. 2011. *Big Data: The next frontier for innovation, competition, and productivity.* Mckinsey Global Institute.

https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx

Lane, J., Stodden, V., Bender, S. and H. Nissenbaum. 2014. *Privacy, Big Data, and the Public Good: Frameworks for Engagement.* New York: Cambridge University Press. http://assets.cambridge.org/97811070/67356/frontmatter/9781107067356_frontmatter.pdf

Ministry of Electronics & Information Technology. 2019. "Open Government Data Portal: An Overview." Accessed June 20. https://www.meity.gov.in/writereaddata/files/OGD_Overview%20v_2.pdf.

Nilekani, Nandan. 2018. "Data to the People. India's Inclusive Internet." *Foreign Affairs* 97(5).

Ritchie F, and Richard Welpton. 2011. "Sharing risks, sharing benefits: Data as a public good." Presentation at Joint UNECE/ Eurostat work session on statistical data confidentiality, Tarragona, Spain, October 26-28. http://eprints.uwe.ac.uk/22460/1/21_ Ritchie-Welpton.pdf

Rodwin MA, and JD. Abramson. 2012. "Clinical trial data as a public good." *JAMA* 308: 871-2.

Taylor, L. 2016. "The ethics of big data as a public good: Which public? Whose good?" *Philosophical Transactions of the Royal Society* 374: 1–13.