111972/2020/CL&ES

Ministry of Electronics and
Information Technology
Government of India

Report by the Committee of Experts on Non-Personal Data Governance Framework

Contents

INTRODUCTION

1. Brief of the Committee

1.1. The Ministry of Electronics & Information Technology (MeitY) constituted a Committee of Experts to deliberate on a Data Governance Framework. Office Memorandum No. 24(4)/2019-CLES dated 13.09.2019 was issued to create the 8 member committee. Stated goals for the committee were

i. To study various issues relating to Non-Personal Data.

ii. To make specific suggestions for consideration of the Central Government on regulation of Non-Personal Data.

1.2. The list of the Committee members is provided in **Appendix 1**.

2. Methodology

2.1. Consultations with stakeholders

i. As part of the deliberations the Committee met with representatives from various sectors of business (Indian and global companies) to get their views - Health, e-Commerce, Internet, Enterprise Subject matter experts, Not for Profit / think-tanks, technology service providers, etc.

ii. Several experts too presented their ideas / views and discussed with the Committee over meetings / video conference calls / mails.

2.2. In order to understand the current status of this topic across the world, the Committee did a literature review on this topic, and the relevant reports are referred across this document.

2.3. In this document –

i. Chapter 3 on 'Data – Trends and Socio-Economic Impact' presents trends in data availability, its socio-economic impact leading to imbalances in the market, under-optimal use of data for economic, social and public purposes, and makes a case for regulating data.

ii. Chapter 4 on ' Definition of Non-Personal Data and Key Roles' provides a definition of Non-Personal Data and defines three categories – Public,

Community andPrivate; and also defines three key roles, namely data principal, data custodian, and data trustee; and an institutional form of data infrastructures, namely a data trust.

iii. Chapter 5 on ‘Ownership of Data' articulates a legal basis for establishing rights over Non-Personal Data.

iv. Chapter 6 on ‘Undertakinga Data Business' defines a Data Business and articulates requirements for its registration and data disclosure.

v. Chapter 7 on 'Data Sharing' recommends mechanisms for data sharing while defining three purposes for data sharing.

vi. Chapter 8 on 'Non Personal Data Regulatory Authority' recommends a Non-Personal Data Authority and articulates its two roles (enabling and enforcing).It also proposes a separate legislation to govern and regulate Non Personal Data.

vii. Chapter 9 on 'Technology Architecture for Data Sharing' provides technology-related guidelines for digitally implementing the recommended rules and regulations around data sharing.

viii. Appendixes 1 to 7 provide supporting material to the various Chapters.

COMMITTEE DISCUSSIONS AND RECOMMENDATIONS

3. Data – Trends and Socio-Economic Impact

Key Take-Aways

- The world is awash with data.
- The proliferation of big data, analytics and Artificial Intelligence (AI) has led to the creation of many new information intensive services and also the transformation of existing businesses.
- Data inter alia contributes to economic value and wealth. Frameworks are being created to better understand the uses and benefits of data.
- Organizations have been discovering ways to generate value from data. The digital economy is witnessing the emergence of a few dominant players and a certain imbalance in the market.
- Given the increasing importance and value generation capacity of the data economy, governments around the world realise the need to enable and regulate all aspects of data, both Personal and Non-Personal Data.

Data availability and value generation from data

3.1. The world is awash with data. Planet scale adoption of theInternet, smartphones, and cloud driven apps, followed by increasing use of AI-systems are the main reasons why we are generating and consuming data at a scorching pace.

i. There are over 3 billion smartphone users in the world[1]. Instagram had over 277,000 stories posted, Google had over 4.4 million searches and Uber had over 9,700 rides every minute of the day in 2019[2].

ii. Estimates suggest that the world will generate about 90 zettabytes (approximately a billion terabytes) of data in this year (2020) and the next, more than all the data produced since the advent of computers[3]. By 2025, worldwide data is expected to grow to 175 zettabytes, with much of the data residing in the cloud[4].

1https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/, accessed on 15/03/2020
2https://www.forbes.com/sites/nicolemartin1/2019/08/07/how-much-data-is-collected-every-minute-of-the-day/#1dd7255b3d66
3https://www.economist.com/special-report/2020/02/20/a-deluge-of-data-is-giving-rise-to-a-new-economy
4https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html

iii. AI techniques like Machine Learning (ML)and deep learning require large data sets to provide accurate prediction models. A public databaseused to build deep learning models like Imagenet has more than 14 million hand-annotated images[5].

3.2. Digital transformations are happening all around the world. A proliferation of big data, analytics and AI has led to the creation of many new data intensive services and the transformation of existing services into data intensive services.

i. It is estimated that the global AI-derived business value in 2020 is likely to be about USD 2.65 trillion[6].Between 2018 and 2019, organizations that have deployed AI grew from 4% to 14%[7].

ii. Abundant availability of data is a primary driver for AI. We are witnessing increased traction for AI solutions in India. This AI powered economic growth in India has not only created new services but has also improved the quality of existing services. NASSCOM forecasts that India's analytics revenue in 2025 will be around USD 16 billion USD, about 32% of the global market[8].

iii. The demand for AI has in turn created a demand for AI talent in India. According to NASSCOM, the total demand for AI and big data, analytics talent in India is likely to grow from around 510,000 in 2018 to about 800,000 in 2021[9].

3.3. Traditionally there was value in selling processed data. Today, the typical process of value creation from data is as follows:

i. Data collection
ii. Cleansing and curating raw / factual data
iii. Populating databases in standardized formats
iv. Doing data mining and data analysis using various tools and techniques
v. Using curated data to train AI/ML systems
vi. Converting information into insights that help in prediction and decision making for revenue / profit generationas well as for social and public interest activities.

3.4. There are three ways in which organizations realize the value of their data – 1) Direct monetization, 2) Internal investments, and 3) Mergers and acquisitions. There

5https://www.newscientist.com/article/2127131-new-computer-vision-challenge-wants-to-teach-robots-to-see-in-3d/
6https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/
7https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/
8https://community.nasscom.in/wp-content/uploads/attachment/nasscom-indian-analytics-data-to-decisions-june-2016-sec.pdf
9https://www.nasscom.in/knowledge-center/publications/talent-demand-supply-report-ai-big-data-analytics

are a number of approaches developed to measuring the value of data and this is an evolving field[10, 11, 12, 13, 14, 15].

3.5. Frameworks are being developed to better understand the uses and benefits of data and its value[16] including 1) Treating data as an asset 2) Activity or usage value of data 3) Future value of data and 4) Prudent value of data.

i. Data is treated as an asset and monetized directly by trading it or building a service on top of the data.

ii. Data's value is based on the number of users and frequency of data access. Unlike a physical asset, the more a data is used, the more valuable it is likely to become.

iii. Data is treated as an intangible asset whose value may be discoverable at a future date, say during a Mergers & Acquisition activity.

iv. The prudent value approach values data sets based on the extent to which they could advance key business initiatives that support a company's overall business strategy.

Imbalance in data and digitalindustry

3.6. Some examples of the data based businesses include – social media, search, map-based services, online retail, ride-hailing platforms, digital healthcare, credit rating, etc.

i. User data and user generated content are collected and analysed often with AI to make better decisions for businesses and organizations. Our society experiences such data-enabled services in the form of platforms like Google Maps, Uber, Amazon, etc.

ii. It is reported that Google and Facebook together control about 60% of the Internet advertising market in the USA[17]. It is also estimated that Amazon had a

10 Chiehyeon Lim et al., "From data to value: A nine-factor framework for data-based value creation in information-intensive services", International Journal of Information Management, Volume 39, April 2018, Pages 121-135
11 Michael Chui et al., "Notes from the AI frontier: Applications and value of deep learning", McKinsey Global Institute, April 2018
12 Asha Saxena, "What is Data Value and Should it be Viewed as a Corporate Asset?", Dataversity, March2019
13 John Akred and Anjali Samani, "Your Data Is Worth More Than You Think" MITSloan Management Review, January 2018
14 Hanna KozlowskaJuly, "How much is your data worth?", Quartz, July 2019
15 AmirataGhorbani and James Y. Zou, "What is your data worth? Equitable Valuation of Data", https://arxiv.org/pdf/1904.02868.pdf
16 John Akred and Anjali Samani, Your Data Is Worth More Than You Think, MIT Sloan Management Review, January 2018
17https://www.reuters.com/article/us-alphabet-facebook-advertising/google-facebook-have-tight-grip-on-growing-u-s-online-ad-market-report-idUSKCN1T61IV

37% share of the online ecommerce market in the USA in 2019[18]. This is reflected in the very large market capitalization of these corporations.

3.7. For a few companies that dominate the digital and data business, the network effects lead to outsized benefits and creates a certain imbalance in the data/digital industry.

i. So far, a few startups from the 1990s and 2000s have gone on to become USD 1 trillion market capitalisation multinational corporations. One of the primary drivers of value of these companies is their ability to collect and analyse data of users which often leads to network effects that help them grow and become very dominant actors in the economy. These companies have also been in the forefront of adopting AI to analyse this data.

ii. In the list of the worlds' 70 largest platforms with respect to market capitalisation – America has 73%, China has 18% and Europe has 4% of the platforms[19].

iii. In a data economy, companies with the largest data pools have outsized, unbeatable techno-economic advantages. For example, studies[20] have shown that increasing a speech corpus size by 5 times reduces word-error-rate (i.e. errors in speech to text translation) by 10% or more, while cutting cost by significantly reducing the need for manual rating. Such a 10% reduction in error-rate used to take a generation of research earlier. But now, access to exponentially increasing data set sizes, large R&D budgets and unprecedented computing power are making it possible in much shorter time periods.

iv. A combination of a "first mover advantage" for these large data-driven platforms and businesses, with the sizable network effect and enormous data that they have collected over the years, has left many new entrants and start-ups being squeezed and faced with significant entry barriers. This may be the right time to set out rules to regulate the data ecosystem (which includes data collection, analysis, sharing, distribution of gains, destruction etc.) to provide certainty for existing businesses and provide incentives for new business creation, as well as to release enormous untapped social and public value from data.

18https://www.bloomberg.com/news/articles/2019-06-13/emarketer-cuts-estimate-of-amazon-s-u-s-online-market-share
19https://www.economist.com/business/2020/02/20/the-eu-wants-to-set-the-rules-for-the-world-of-technology
20https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43230.pdf

v. India is second most populous country in the world[21]. India also has the second highest number of smartphone users in the world[22]. Given this, and the current levels of Internet penetration in India, India can arguably be projected as being one of the top consumer markets, and by extension data markets in the world in the foreseeable future. Allowing the possibility of data monopolies, in a large consumer market such as India, could lead to the creation of imbalances in bargaining power vis-à-vis few companies with access to large data sets accumulated in a largely unregulated environment, on one side, and Indian citizens, Indian businesses including startups, MSMEs and even the Government, on the other. Therefore, the Government's role is to catalyse the data businesses in a manner that maximizes overall welfare.

vi. At the same time, the requirement for providing certainty and incentives for new business creation cannot be understated. It is because of robust IP rights, various data related privileges, that a lot of data-driven innovation has occurred. Therefore, while ensuring that markets function properly, sufficient and adequate incentives for new business creation must therefore be safeguarded.

vii. Lastly, potential harms could arise in terms of privacy violations arising from re-identification of anonymized data, or from the derivation of personally identifiable insights from non-personal data. Adequate measures would have to be developed in order to ensure that any data sharing framework does not dilute the protections afforded by the Personal Data Protection Bill, 2019 (PDP Bill)[23]. Accordingly, any eventual regulation will have to mitigate against the risks of privacy harms.

viii. Not only economic, but most key social, political and cultural activities will depend upon data, and suitable access to it. For instance, governments would need wide access to data in all sectors for public policy developmentand delivery of public services. While public agencies produce a lot of data, much of the required data will be collected by and be in the hands of private companies.Besides data philanthropy, some systematic mechanisms need to be developed to tap the social and public value of data.

21https://www.worldometers.info/world-population/india-population/
22https://www.statista.com/statistics/748053/worldwide-top-countries-smartphone-users/
[23]http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf

The Case for Regulating Data

3.8. Data creates economic value and wealth, apart from enormous social and public value. Data therefore is increasingly taking the centre-stage in core-technological businesses, all economic sectors around the world and in addressing various social and public administration issues. It is in this context, that the Committee has sought to set out the case for regulation of data. As a starting point therefore, one needs to understand the nature of data as an economic good, as also its social and public value. In this regard, data can be viewed through two lenses[24]– economic and informational.

i. Data as an economic resource has huge externalities: From an economic lens, data is non-rivalrous, yet excludable, and its use could have both positive and negative externalities.

ii. Data offers intrusive information about its subject: From an informational lens, one needs to recognise and understand the subject, content and use of data, and understand how any content and use of data, could give rise to harms. For instance, sensitive or personal data could lead to privacy harms. Even Non-Personal Data, including anonymised Personal Data, could provide collective insights that could open the way for collective harms (exploitative or discriminatory harms) against communities.

iii. Collective information / data is needed for social and public interest use. An instance of a collective harm is when such data is closed for public use and leads to welfare losses.

iv. Collective privacy refers to possibilities of collective harm related to Non-Personal Data about a group or community that may arise from inappropriate exposure or handling of such data. There remain concerns about safety of all such data in relation to the interests of the group or community about which the data is, whereby the term collective privacy is employed.For example,

- Data emerges about people of certain sexual orientation frequenting certain pubs / restaurants. And certain other groups of people, who are opposed to this sexual orientation, take adverse action on these pubs / restaurants. The group of people with such sexual orientation can exercise their collective privacy and ensure that such information is protected.
- Data emerges about people who suffer from a certain disease, which in a particular society has certain social stigma attached, and that they are

[24] Bennett Institute for Public Policy and Open Data Institute, "The Value of Data – Policy Implications", 2020

centred in a particular locality in the city. In response to such data, the residents of that locality are ostracized, certain services (like delivery etc.) are denied to them. The residents of the society can take recourse to protection under collective privacy.

- The Committee believes that this is an emerging concept[25,26,27] that will need to be examined and defined in detail in the future.

v. Market transactions and market forces on their own will not bring about the maximum social and economic benefits from data for the society. Appropriate institutional and regulatory structures are essential for a thriving data economy and a well-functioning data society. The Committee's approach to regulating data, keeps such an understanding of data at its core.

Key Take-aways – Case for regulating data

The Committee believesthat rules and regulations are required to manage data in order to achieve the following enabling and enforcingbenefits:

i. Come up with a set of recommendations such that India can create a modern framework for creation of economic value from use of Data. To generate economic benefits for citizens and communities in India and unlock the immense potential for social / public / economic value data.
ii. To create certainty and incentives for innovation and new products / services creation in India. To encourage start-ups in India.
iii. To create a data sharing framework such that community data is available for social / public / economic value creation
iv. To address privacy concerns, including from re-identification of anonymised personal data, preventing collective harms arising from processing of Non-Personal Data, and to examine the concept of collective privacy.

3.9. In the context of this Committee, the case for regulating data is madein such a manner that the benefits accrue to India and its communities and businesses.For instance,

i. Sharing Non-Personal Data collected by both government and private organizations with citizens is likely to lead to increased transparency, better quality services, improved efficiencies, and more innovation[28].

[25] https://link.springer.com/chapter/10.1007/978-3-319-46608-8_8
[26] https://link.springer.com/article/10.1007/s13347-019-00351-0
[27] https://www.springer.com/gp/book/9783319466064
[28] http://opendatatoolkit.worldbank.org/en/starting.html

ii. The shared Non-Personal Data may be useful for Indian entrepreneurs to develop new and innovative services and products from which citizens may benefit.

iii. The Non-Personal Data may also be used by researchers, academia and governments for creating public goods and services like an Indian genome repository, data for training natural language translation systems on Indian languages, etc.

4. Definition of Non-Personal Data and Key Roles

It is thus understood that data is valuable and it must be regulated in an appropriate manner. For that to happen, a clear definition of Non-Personal Data and key roles in the Non-Personal Data ecosystemmust be articulated.

Definition of Non-Personal Data

4.1. Accordingly, the Committee considered various kinds and aspects of Non-Personal Data. Refer to **Appendix 2** for the information / examples of Non-Personal Data that the Committee considered.

i. Data may be categorised in many ways – arising from the subject of data (e.g. personal data); in relation to its purpose (e.g. AI training data, e-Commerce data); the sector to which it belongs (e.g. health data); the source of data (e.g. soil data); level of processing (raw / factual versus derived data); or the collector of data (e.g. public / Government or private data); or based the extent of involvement of stakeholders in the creation of data (provided, observed, derived, or inferred).

ii. Non-Personal Data – When the data is not 'Personal Data' (as defined under the PDP Bill), or the data is without any Personally Identifiable Information (PII), it is considered Non-Personal Data.

iii. A general definition of Non-Personal Data according to the data's origins[29] can be:

- Firstly, data thatnever related to an identified or identifiable natural person, such as data on weather conditions, data from sensors installed on industrial machines, data from public infrastructures, and so on.
- Secondly, data which were initially personal data, but were later made anonymous. Data which are aggregated and to which certain data-transformation techniques are applied, to the extent that individual-specific events are no longer identifiable, can be qualified as anonymous data.

iv. Given the importance of anonymisation (of Personal Data to make it Non-Personal Data) and to prevent the risk of re-identification, the Committee collated some of the basic anonymisation techniques in this report. Refer to **Appendix 3**.

29 European Commission, "Guidance on the Regulation on a framework for the free flow of Non-Personal Data in the European Union", 2019, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019DC0250&from=EN

v. The Committee realized that there are different terminologies used to describe Non-Personal Data and there was a need to provide a clear single definition in this report.

Recommendation 1: DefiningNon-Personal Data

Recommendation 1: Define Non-Personal Data

- The Committee has defined three categories of Non-Personal Data – 1) Public Non-Personal Data 2) Community Non-Personal Data & 3) Private Non-Personal Data.
- The Committee has also defined a new concept of 'sensitivity of Non-Personal Data', as even Non-Personal Data could be sensitive from the following perspectives – 1) It relates to national security or strategic interests; 2) It is business sensitive or confidential information; 3) It is anonymised data, that bears a risk of re-identification
- The Committee recommends that the data principal should also provide consent for anonymisation and usage of this anonymized data while providing consent for collection and usage of his/her personal data.

4.2. Public Non-Personal Data

i. Public Non-Personal Data means Non-Personal Data collected or generated by the governments, or by any agency of the governments, and includes data collected or generated in the course of execution of all publicly funded works.
   - All Non-Personal Data collected or generated by the Government where such data is explicitly afforded confidential treatment under a law, shall not constitute Public Non-Personal Data.

ii. Examples of Public Non-Personal Data
   - Anonymiseddata of land records, public health information, vehicle registration data etc.
   - A university collects pollution levels in the city based on a publicly funded project.

4.3. Community Non-Personal Data

i. A community is any group of people that are bound by common interests and purposes, and involved in social and/or economic interactions. It could be a

geographic community, a community by life, livelihood, economic interactions or other social interests and objectives, and/or an entirely virtual community.

ii. Community Non-Personal Data means Non-Personal Data, including anonymised personal data, and non-personal data about inanimate and animate things or phenomena – whether natural, social or artefactual, whose source or subject pertains to a community of natural persons. Provided that such data shall not include Private Non-Personal Data.
   - For instance, besides datasets collected by the municipal corporations and public electric utilities, datasets comprising user-information collected even by private players like telecom, e-commerce, ride-hailing companies, etc., should be considered Community Data.
   - Here, the 'raw / factual data', without any processing / derived insights, may be characterised as the Community Data.

4.4. Private Non-Personal Data

i. Private Non-Personal Data, means Non-Personal Data collected or produced by persons or entities other than the governments, the source or subject of which relates to assets and processes that are privately-owned by such person or entity, and includes those aspects of derived and observed data that result from private effort.
   - It includes inferred or derived data / insights involving application of algorithms, proprietary knowledge.
   - In the case of Generative Adversarial Networks[30], two AI engines contest against each other andcreate new data instances that resemble the AI engine's training data. Thisderived data is an example of private Non-Personal Data.
   - It may also include such data in a global dataset that pertains to non-Indians and which is collected in foreign jurisdictions (other than India).

4.5. Sensitivity of Non-Personal Data

i. In the case of Personal Data sensitivity spectrum, there exist three categories – General, Sensitive and Critical.

ii. Sensitivity of data is a concept defined in the context of Personal Data.
   - Clause 36 of the PDP Bill defines "sensitive personal data" as such personal data, which may, reveal, be related to, or constitute— (i) financial data; (ii) health data; (iii) official identifier; (iv) sex life; (v) sexual orientation; (vi) biometric data; (vii) genetic data; (viii) transgender status; (ix) intersex status;

[30] https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

(x) caste or tribe; (xi) religious or political belief or affiliation; or (xii) any other data categorised as sensitive personal data under section 15.

- The Government may notify the personal data which would be classified as Critical Personal Data.

iii. The Committee felt that it is important to bring in the concept of sensitivity to Non-Personal Data as well, from the following perspectives:
   - It relates to national security or strategic interests;
   - It bears risk of collective harm to a group (collective privacy etc.);
   - It is business sensitive or confidential information;
   - It is anonymised data, that bears a risk of re-identification

iv. Even after Personal Data is anonymised into Non-Personal Data, the possibilities of harm to the original data subject(s) is not totally gone, as it is being increasingly recognised that no anonymisation technique provides perfect irreversibility. Such harm may be individualor collectiveas a group or community(whereby harm can happen even without de-anonymisation). Any person and/or community therefore hasabiding concerns related to possible harm in anonymised or Non-Personal Data. Possibilities of such harm are obviously much higher if the original Personal Data is of a sensitive nature. Therefore, the Non-Personal Data arising from such sensitive Personal Data may be considered as sensitive Non-Personal Data.

v. The Committee recommends that Non-Personal Data inherits the sensitivity characteristic of the underlying Personal Data from which the Non-Personal Data is derived. Some examples to illustrate how the sensitivity characteristic may be inherited include:
   - If the Non-Personal Data is about health of people (even though it may be anonymised and aggregated), on the sensitivity spectrum it will be classified as Sensitive Non-Personal Data since the underlying data (on health) is classified as Sensitive Personal Data as per Clause 3 (36) of the PDP Bill.
   - Data collected about say, mobile penetration in a city (when aggregated and anonymised) may be treated as general Non-Personal Data since the underlying data collected (mobile phone ownership of a person) is treated as general Personal Data.
   - Data collected about pollution levels in a city may be classified as general Non-Personal Data.

vi. Other kinds of data whose underlying Personal Data may not be sensitive, or there may not be any underlying Personal Data at all, may still be sensitive with respect to collective harm - like Non-Personal Data related to vital infrastructure, which may be considered sensitive from a security perspective.

4.6. Consent for Anonymised Data

i. It is clear from industry feedback to the Committee and from its own research that large collections of anonymised data can be de-anonymised, especially when using multiple Non-Personal Data sets. This risk is considered by this Committee to be a valid one. Hence the individual (data principal) needs more protection.

ii. The guiding principle in this regard, should be that the Personal Data that is anonymized should continue to be treated as the Non-Personal Data of the data principal. In this manner, any subsequent harms arising from re-identification, or otherwise arising from processing, can be acted upon by the data principal.

iii. Under the PDP Bill, consent is necessary for the collection and processing of Personal Data. Since the conditions of 'specific' and 'capable of being withdrawn', as specified in PDP Bill Chapter II, 11 (2), do not apply to Non-Personal Data, we cannot assume that consent provided for Personal Data applies automatically to Non-Personal Data.

iv. Therefore, the Committee recommends that the data principal should also provide consent for anonymisation and usage of this anonymised data while providing consent forcollection and usage of his/her Personal Data.

v. The Committee also recommends that appropriate standards of anonymisation be defined to prevent / minimize the risks of re-identification.

Definition of Key Roles in the Non-Personal DataEcosystem

In order to develop and enable a robust Non Personal Data ecosystem, a set of roles / stake-holders and data infrastructures needs to be defined.

Recommendation 2: Defining Key Non-Personal Data Roles

Recommendation 2: Define Non-Personal Data Roles
1) Data Principal
2) Data Custodian
3) Data Trustees
4) Data Trusts

4.7. Data Principal

i. In case of Personal Data, data principal is the natural person to whom the personal data relates. However, in case of Non-Personal Data, the definition of a data principal is related to the type of Non-Personal Data - Public, Community and Private data, as well as based on different possible kinds of subjects of data.

ii. In case of Public Non-Personal Data:
   - Government may collect data pertaining to citizens (like census), companies (like company registration, financial filings) and communities.
   - The data principal will be the corresponding entities (individuals, companies, communities) to whom the data relates.

iii. In case of Private Non-personal Data:
   - Private sector may collect data pertaining to citizens (like customer surveys), companies (like vendor registration, vendor product information) and communities.
   - The data principal will be the corresponding entities (individuals, companies, communities) to whom the data relates.

iv. In case of Community Non-Personal Data:
   - A community, that is the source and/or subject of community data and as defined in Section 4.3 , may be treated as the data principal for such data, and should be able to exercise key rights, including economic rights, to this data.

4.8. Data Custodian

i. The data custodian undertakes collection, storage, processing, use, etc. of data in a manner that is in the best interest of the data principal.

ii. The data custodian may also be considered as data fiduciary, subject to certain directions and control and acting as per the interest of data principal/ group/community.
   - Such community 'best interest' will need to be channelled or communicated to data custodians by data trustees (as defined in Section 4.9) on behalf of the data principal community. It could be in the form of data advice, recommended data practices requirements/guidelines, etc. but must always meet the cannons of the best interests of the data subject community or the data principals.

iii. Data custodians have a 'duty of care' to the concerned community in relation to handling Non-Personal Data related to it. This concept of 'duty of care' is a general set of obligations, which can in time be specified better, by regulatory guidelines, practices, rules, legislations etc. This report does lay down some such duties in from of anonymisation standards and requirements, protocols and means for safe data sharing, etc.The duty of care should be operationalized through the “best interest” standard, and the prevention of harms to communities and individuals from the processing of Non-Personal Data.

iv. An appropriate Non-Personal Data framework legislation, while providing community data rights will also lay down principles and guidelines for various incentives for data custodians, respective data privileges, compensations, where needed, the nature of the required, well-regulated data markets, and so on.

- The framework law will also provide means to specifically protect and promote the interests of small Indian companies and startups. Symmetric data sharing obligations equally on all data businesses may not always work for small businesses, and may even be to their detriment. Provisions like threshold size for data sharing, and graduated sharing obligations, may be considered.

4.9. Data Trustees

i. The data principal group/community will exercise its data rights through an appropriate community data trustee. In the case of community data, unlike personal data where an individual can directly exercise control over her data, the concept of trustee for community data comes in, who would exercise such rights on the behalf of the community.

ii. Principles and guidelines about who can constitute the appropriate trustee in a given context of group/community data will be laid outby the mentioned framework legislation. In principle, it should be the closest and most appropriate representative body for the community concerned. For a lot of community data, the corresponding government entity or community body may act as the data trustee. Some examples are provided below.

- The Ministry of Health and Family Welfare, Government of India can be the trustee for data on diabetes among Indian citizens.
- The State Government of Manipur can be a trustee on data on Meitei language.
- Citizens groups (NGOs) registered in Whitefield locality in Bangalore can be the trustees on solid waste management data in Whitefield.

- A public university in Hyderabad that is collecting data on the state of roads in Hyderabad as part of research project can be a trustee of the data it has collected.
- The Directorate of Urban Land Transport may become a trustee of traffic data collected by multiple ride-sharing platforms, besides traffic data from the city police department, in order to develop a city traffic solution.

iii. In certain cases, mandatory data sharing will be required to open up competition in any concerned sectorenabling startups, or for other community/ public interest purposes discussed in this report. The data trustee may seek enforcement of safeguardson the sharing of community Non-Personal Data of which it is the trustee, before the data regulator.

iv. Data trustees can also recommend to the data regulator the enforcementof soft obligations on data custodians, like transparency and reporting mechanisms, or stronger ones involving regulation of data practices, within the framework to be specified by legislation. This will depend on many different factors including nature of data, kind of data practices, context of data use, nature and sensitivity of the involved sector, nature of expected outcomes, etc.

v. In seeking and enforcing data sharing with regard to various community data on specific data requests, the data regulator (defined inChapter 8) will work in collaboration with the data trustee of community data sought to be shared. For example, the data regulator may work with the government transport department(playing the role of a data trustee), on whether, how and with whom their community data related to commuting through various modes of transportation, is shared.

4.10. Data Trusts

i. Data trusts are the institutional structures, comprising specific rules and protocols for containing and sharing a given set of data.

ii. Data trusts can contain data from multiple sources, custodians, etc. that is relevant to a particular sector, and required for providing a set of digital or data services.

iii. Data custodians may voluntarily share data in these data trusts, as many private organizations may come forward to share data held by them. Another important source of data pooled into these common data trusts will be from public organizations producing and holding various public data.

iv. Governments / data trustees may also seek mandatory sharing of important data for a sectorfor specific purposes, which would also be managed and provided by such data trusts. It may also consist of both mandatorily and voluntarily shared data.

v. Such data trusts and infrastructures can be managed by public authorities – on the parallel of public infrastructure underpinning much of industrial economy, or these can be managed by new, neutral bodies, cooperatives, or industry associations, and so on. Different forms may be found fit for different kinds of data and different sharing needs.

vi. The term data infrastructure further brings in also the corresponding technical-material elements required for data sharing, like actual databases, APIs, organisational systems, etc.

All these roles and the data sharing requirements and arrangements will be governed by the Non-Personal Data rules and regulations, pulled together under a new legislation, and by the Non-Personal Data Authority, the data regulator for Non-Personal Data, which will be discussed in Chapter 8.

5. Ownership of data

The Committee discussed the ownership of Non-Personal Data and articulated a legal basis for establishing rights over it.

Legal basis for establishing rights over Non-Personal Data

5.1. The Committee developed certain guiding principles for establishing legal rights over data.

i. Data sovereignty: The ownership of the Non-Personal Data collected about people in India and collected in India should be defined. The laws, regulations and rules of the Indian State apply to all the data collected in/from India or by Indian entities.

ii. The term "ownership" holds full meaning only in terms of physical assets. Regarding intangible assets like knowledge and data, the term 'ownership' is relatively loosely employed to meana set of primary economic and other statutory rights. For such intangible assets, many actors may have simultaneous overlapping rights and privileges. At times, such rights and privilege of different actors may not even interfere with one another, but this is not always so. It is therefore important that such rights and privileges related to Non-Personal Data are clearly defined and ascribed. Accordingly, the notion of "beneficial ownership/interest" has been adopted by the Committee, in ensuring that a Community's interests are safeguarded regarding non-personal data over which there is an expectation of benefits being accrued to itself.

iii. Accordingly, the Committee recommends that:

- In case of Non-Personal Data derived from personal data of an individual, the data principal for personal data will continue to be the data principal for the Non-Personal Data, which should be utilized inthe best interest ofthat individual.
- The rights over community Non-Personal Data collected in India should vest with the trustee of that community, with the community being the beneficial owner, and such data should be utilized in the best interest of that community.

iv. Benefits accrue to relevant Indian communities: The Committee agreed that the benefits accruing from the processing of community Non-Personal Data, should accrue not only to the organizations that collects such data, but also equally to the community that typically produces the raw / factual data that is being captured. Accordingly, such data may be shared in instances where there are defined grounds or purposes for sharing of Non-Personal Data (refer to Recommendation 5)with citizens, Indian start-ups, Indian companies, Indian

public and private universities, Indian public and private research labs, Indian Non-Government organizations, and the Indian Central and State Governments.

v. State and community bodies as data protector and enabler for citizens / community: Citizens and communities have an important role to play in how much Non-Personal Data gets generated and used. They also should have a fair share of the benefit created from such Non-Personal Data.
   - However, communities need an adequate system of decision implementation on their behalf, which has to be fully in their interest. This will be done by a data trustee.
   - The legitimate trustee for any set of community data will be the closest and most appropriate representative body for that community, which will, in many cases, be an appropriate community body or Central/ State/ Local government agency.
   - This should however be undertaken in a strict rules-based manner, with adequate checks against abuse of power by government or other representative agencies, which requires an elaborate institutional structure for this purpose.

Recommendation 3: Articulating a legal basis for establishing rights over Non-Personal Data

5.2. Public Non-Personal Data

i. The Committee believes that since Public Non-Personal Data, as defined at Section 4.2 above, is derived from public efforts, the datasets created partake of the characteristics of a national resource.

5.3. Community Non-Personal Data

i. Non-Personal Data relating to individuals is often not just a proprietary or personal asset, but may be considered a collective or shared asset because many parties have overlapping legitimate contributions to and interests in it. Communities are collective subjects and significant stakeholders, and as such the legitimate societal or economic beneficiaries of any community's data. This has been operationalized through the definition of “Community Non-Personal Data”, as provided at Section 4.3 above.

ii. The case can therefore be made for a legal basis of community’s rights over data about the community that may be collected by private data custodians or public organizations. It establishes why, for instance,

- Raw / factualdatasets comprising anonymised user-information data collected by private data custodians (such as telecom, e-commerce, ride-hailing companies, etc.), may be considered Community Data.
- The principle of raw data is standards compliant, machine readable and fidelity as collected. The raw data will be made available in usable formats, and only an open, reviewed license-free standard can be used.
- Private data custodian's drones taking pictures of agriculture farms of local farmers, with or without standing crops, and using it to analyse soil types, health of crops etc. may be considered as community data.

- **'Data source' logic of Community Non-Personal Data**– Just like the economic rights to natural resources arising from a community are considered to primarily belong to it, the value of social resources of Community Non-Personal Data should primarily accrue to it (instead of the default whereby data custodians take up the entire value of such data).

- **'Data subject' logic of Community Non-Personal Data** – it is data about the concerned group/community and provides systemic intelligence about it. Such systemic intelligence, in every sector, and every walk of social, economic, political and cultural life, is a great power over the community concerned. The group/community should be able to determine and control how such data and intelligence is used – maximising data's benefits for itself and eliminating or minimizing harms.

iii. Data being non-rivalrous makes it different from natural resources in the sense that there can be multiple data custodians for the same/ similar data and the value of data may also be consumed by the relevant community as well as others, without degrading its value to the relevant community. Therefore, a legal framework based on source must consider these factors so that data custodians have appropriate incentives to collect the data and the community's rights do not result in undue restriction of use of the data by others.

iv. Allocation of primary economic and other statutory rights over the data, should therefore be operationalised through the concept of "beneficial ownership/ interest", such that the Community's interests are safeguarded, and the data may be used to further economic good, well-being, rights and dignity of the community.

v. The legitimate trustee for any set of community data will be the closest and most appropriate representative body for that community, which will, in many cases, be an appropriate community body or Central/ State/ Local government agency.

5.4. Private Data

i. In the "Private Non-Personal Data", as defined in Section 4.4., only such raw / factual data pertaining to a community, that is collected by a private organization may need to be shared, subject to the well-defined grounds (refer to Recommendation 5)at no remuneration.

ii. As the processing value-add over the raw data increases, appropriatemechanisms may be leveraged for data sharing. Refer to Section7.4. (iii).

iii. Algorithms / proprietary knowledge may not be considered for data sharing.

6. Undertaking a Data Business

In order to ensure greater access to Non-Personal Datain a systematised manner, the Committeeproposes the idea of creating a new category of businessin India called 'Data Business'.Refer to **Appendix 4**for emerging global frameworksin this domain that the Committee considered.

Key Takeaways – Data Business

- Create a new category / taxonomy of business called 'Data Business' that collects, process, store, or otherwise manages data, and meets certain threshold criteria.
- Data Business is a horizontal classification and not an independent industry sector.Many existing businesses in various sectors, collecting data beyond a threshold level, will get categorized as a Data Business.
- Data Businesses will provide, within India, open access to meta-data and regulated access to the underlying data.
- The compliance process will be light-weight and fully digital.

Recommendation 4: Defining a Data Business

Definition of a Data Business

6.1. Create a new category of business called Data Business

i. Organizations are deriving new or additional economic value from data, by collecting, storing, processing, and managing data. For instance, a hospital derives economic value not only from providing medical services, it may derive additional value by harnessing the medical data and offering value-added services (such as personalized treatment plans, medicines etc.).

ii. Hence create a new category / taxonomy of business called 'Data Business' that meets certain data threshold criteria.

iii. Data business is not an independent industry sector. It is a horizontal classification cutting across different industry sectors.
   - For example, companies in banking / finance, telecom, Internet-enabled services, transportation, consumer goods, travel, universities, private research labs, non-government organisations etc. may be classified as 'Data Businesses' based on a certain threshold of data collected / processed that will be defined by the regulatory authority (as defined in Chapter 8).

iv. Data Business discovery process – As Data Businesses are a horizontal category, whereby any business could be a data business, it is important for the process to be discoverable - where the business knows they need to register and what to do without being aware of the regulation apriori. And further the design principle for registration of and disclosures byData Businesses is to be purely digital, lightweight and self-certified with a transparent framework written in code.

6.2. Data Business registration process &management

i. It is important for such a business to register as a 'Data Business' once it reaches a certain data-related threshold. This will be applicable to not only commercial organizations, but also Governments and other non-government organizations that collect, process or otherwise manage data.
   - Below the threshold, registration as a Data Business may be voluntary.
   - This is a one-time activity and there is no necessity to obtain a license to be a Data Business.

ii. The Data Business registration system will be an Open API which will also be reflected in a web application and a smartphone App. The registration process will take an officer of the company a few minutes to register, eSign and delegate a data officer for periodic disclosure. Threshold requirements may vary with time, context and need and will be fixed and intimated by Non-Personal Data Authority (as defined in Chapter 8), if needed in consultation with sector regulators.

iii. Initial registration would require a business ID (or country code and country business ID), digital platform/business name(s), associated brand names, rough data traffic and cumulative data collected in terms of number of users, records and data. Also needs to be stated is, the nature of data business, and kinds of data collection, aggregation, processing, uses, selling, data-based services developed etc.

iv. Once the data traffic/ collection exceeds set limits, the Data Business would be required to submit meta-data about data user and community from which data is collected, with details such as classification, closest schema, volume, etc. This will be as per a directory of data classification and schema published by the Non-Personal Data Authority (as defined in Chapter 8).

v. Businesses engaging in new types of data are encouraged to propose both improvements and extensions of the directory and schema which would go

through a peer review, academic review process as per IETF framework, guided by a Technical Advisory body created as per Open API guidelines.

vi. It is suggested the Non-Personal Data Authority (as defined in Chapter 8) define appropriate time period(s) within which Data Businesses will integrate their raw data pipes with the authority for submission of raw data upon request as per the regulatory guidelines.

Data Disclosure and Compliance Requirements of a Data Business

6.3. Data disclosure requirements from Data Business

i. Require Data Business organizations (companies, governments, non-government organizations, etc.) to disclose data elements collected, stored and processed,and data-based services offered.The report can be made in digital format.

ii. Every Data Business must declare what they do and what data they collect, process and use, in which manner, and for what purposes (like disclosure of data elements collected, where data is stored, standards adopted to store and secure data, nature of data processing and data services provided). This is similar to disclosures required by pharma industry and in food products.

iii. There should be a harmonisation of data-related directories and disclosures required for Personal Data and Non-Personal Data, so that businesses supply the same information only once.

iv. The meta-data about data being collected, stored and processed by the Data Business is stored digitally in meta-data directories in India. Open access is provided within India to these meta-data directories.

v. Access to meta-data of Data Businesses – Indian citizens and India-based organizations will have open access to the meta-data about data collected by different Data Businessesincluding governments. By looking at the meta-data, potential users may identify opportunities for combining data from multiple Data Businesses and/or governments to develop innovative solutions, products and services. Subsequently, data requests may be made for the detailed underlying data.

Key Takeaways – Access to meta-data

- The Committee strongly believes that meta-data sharing by Data Business will spur innovation at an unprecedented scale in the country.
- One of the associated key objectives is to promote and encourage the development of domestic industry and startups that can scale their data-based businesses.
- For example, automobile companies may collect data about roads through various sensors. A startup, will know that this data is available based on the meta-data provided by automobile companies. The startup can request for access for this data and can combine this data with public traffic data to create a solution for safest road routes for senior citizens.
- For example, a government funded research lab may collect and publish data on air pollution across different locations in the city. The traffic department and a real time navigation app may publish road traffic data. A smart-city startup, looking at the pollution and traffic meta-data, and may decide to create a solution for identifying safe and least polluted routes.

6.4. Compliance requirements of a Data Business

i. These compliance requirements of a Data Business are irrespective of whether the business is regulated or not by another sectoral regulator. The sectoral regulators can ride on top of the Data Business compliance requirements i.e. use these compliance requirements as a base and add any sector specific data disclosure requirements.

ii. All entities that collect / process Non-Personal Data, above a threshold level, will be subject to an institutional authority (or institutional authorities) that will both enable and regulate various aspects of data. Such an authority must be competent, trained and must include people with industry experience (as defined in Chapter 8).

iii. The compliance requirements may be voluntary when the Data Business is small butcompulsory beyond a data-related threshold.

iv. The process of oversight will be transparent, light-weight and fully digital.

6.5. Rigorous and Lightweight Process

i. Economic value to the country is unlocked only when a number of qualified Indian companies and innovators participate. A key metric we suggest that the regulator measures, monitors and publishes for feedback regularly is the cost of compliance.

ii. Technology and digital tools including for digital adjudication and compliance processing may be used to smoothen and make these processes friction free. It is recommended that all the rigour and due diligence be moved to code to the extent possible while generating sufficient dashboards to bring to bear human judgement. Academic organisations and innovators may be funded or challenged with incentives and contracts to continuously improve these systems.

7. Data Sharing

Data sharing refers to the provision of controlled access to private sector data, public sector data and community data to individuals and organisations for defined purposes and with appropriate safeguards in place. The Committee strongly believes that open-access to meta-data and regulated access to the underlying data of Data Businesses will spur innovation and digital economy growth at an unprecedented scale in the country.This will also necessitate establishment of mechanisms to support data requests and data sharing.

Data Sharing Purpose

Why, and under what conditions, should data be requested and shared? Various stakeholders, including the governments, citizens, startups, companies, universities, research labs, non-government organisations etc., may request Data Businesses for underlying data for defined purposes.

Recommendation 5 – Data Sharing Purpose

- Sovereign purpose – Data may be requested for purposes of national security, legal purposes, etc.
- Core Public Interest purpose – Data may be requested for community benefits or public goods, research and innovation, policy making, for better delivery of public-services etc.
- Economic purpose – Data may be requested in order to encourage competition and provide a level playing field or encourage innovation through startup activities (economic welfare purpose), or for a fair monetary consideration as part of a well-regulated data market.

Recommendation 5: Defining Data-Sharing Purpose

7.1. Data Sharing for Sovereign Purposes

i. Data may be requested for national security, law enforcement, legal or regulatory purposes. Some non-exhaustive examples of these are:
   - Data requested for mapping security vulnerabilities and challenges, including people's security, physical infrastructure security and cyber security.
   - Data required for crime mapping, devising anticipation and preventive measures, and for investigations and law enforcement.
   - Data required for pandemic mapping, prediction and prevention, and also subsequent interventions.
   - Data required by a regulator to understand and keep abreast of developments in a sector with regard to need for regulatory interventions

- Legislations in other countries too have allowed access to data (both Personal Data and Non-Personal Data) for safeguarding national security[31,32].
  - Data typically used in context of national security include telecommunications metadata, geospatial or financial data etc.

7.2. Data Sharing for Core Public Interest Purposes

i. Data may be requested for community uses / benefits or public goods, research and innovation, for policy development, better delivery of public-services, etc.
   - Certain data held by private sector when combined with public-sector data or otherwise may be useful for policy making, improving public service, devising public programs, infrastructures, etc. and, in general, supporting a wide range of societal objectives including science, healthcare, urban planning etc.

ii. India to specify some high-value datasets
   - India should specify a new class of data at a national level (through relevant government departments acting as data trustees of the dataset) – data of special public interest or high-value dataset, like health, geospatial and/or transportation data.
   - Progressively identify other priority sectors for harnessing the economic and societal benefits from leveraging Non-Personal Data. For example, agriculture, education, skills development, MSMEs support, logistics etc.

iii. Utilize data for research purposes
   - Create data spaces (environments which brings together government agencies, startups, universities, research labs, companies, Non-Government Organizations, citizens, etc.) to promote intensive data-based research.
   - We can envisage these data spaces to be sectoral and creation of sector-specific clouds for strategic sectors and other domains.
   - For example, Non-Personal Data can also be used by Indian researchers and government agencies for creating public goods and services like an Indian genome repositoryetc. which can then be leveraged by both public and private organisations.

31 Paul F Scott, “National Security, Data Protection, and Data Sharing after the Data Protection Act 2018”, University of Glasgow

32 Louis de Koker et al., “Big Data Technology and National Security”, Data to Decisions Cooperative Research Centre, 2018

iv. Consider health sector as a pilot use-case for Non-Personal Data Governance Framework.
   - Health data of an individual is considered to be sensitive personal data under the PDP Bill 2019. Yet at the same time, large anonymised data sets of health data, could lend community level insights into diseases, epidemics, and community genetics – leading to better tailored health solutions for the community. Accordingly, the Committee considered health as a pilot use-case for the Non-Personal Data governance framework:
     - Large anonymised datasets of health data, in as much as they would relate to a defined community of natural persons, would constitute community Non-Personal Data.
     - Accordingly, such data could be required to be shared for either regulatory purpose (public health purposes, disease control and prevention) orcore public purpose (better healthcare, accuracy, increased specificity health care models, treatment protocols and diagnostic bots), or economic purpose (supporting digital start-ups and domestic digital industry in health sector).
   - By providing appropriate access to health data, algorithms may be run onsuch data to develop new diagnostic bots / AI systems for healthcare diagnosis, delivery and patient care, to benefit the community which has beneficial ownership over the community health Non-Personal Data.

7.3. Data Sharing for Economic Purposes

Data may be requested in order to encourage competition and provide a level playing field or encourage innovation through start-up activities (economic welfare purpose), or for a fair monetary consideration as part of a well-regulated data market, etc.

i. Data request by startups / businesses
   - Startups / businesses would have access to the meta-data about data collected by different Data Businesses and governments. By looking at the meta-data, these startups / businesses may identify opportunities for combining data from multiple Data Businesses and/or governments to develop innovative solutions, products and services. Such an open access to meta-data information of Data Businesses, leading to subsequent requests for and access to detailed underlying data, will spur innovation in the country.
   - For example, transportation companies may collect data about roads through various sensors. A startup, will know that this data is available based on the meta data provided by these companies. The startup can request for access to this data and can combine this data

with public traffic data to create a solution for the safest road routes for senior citizens.

- For example, both India's startups and established companies can benefit when transport related Non-Personal Data is shared. There are opportunities for new products and improved efficiencies in existing transportation services when Non-Personal Data are combined from public road transport run by governments, traditional taxi and rental cab services, on-demand transportation services, Indian Railways, metro services in Indian cities, parking contractors, traffic management systems, etc.
- A request from a startup/business is a private request to the data custodian for sharing data. If there is a dispute arising from such requests, the data regulatory authority(as defined in Chapter 8) will evaluate the genuineness of such requests based on social/ public/ economic good and mandate that the appropriate raw/factual data be shared. In such cases, a public shared database is typically created so that this can be accessed by all.

ii. Data request by data trustee / governments
    - For important community data for different sectors that may be pre-identified by the data trustee / governments in consultation with sector regulators/ authorities, the data trustee / governmentsmay themselvesdirectly seek access to such community data from private actors holding it, and place such data in appropriate data infrastructures or data trusts, and make it available to all relevant parties.

iii. Setting up data &cloud innovation labs and research centres to develop, test and implement new digital solutions
    - These innovation labs are practical physical environments or field validation centres in which organizations develop, test and implement effective digital solutions.
    - The innovation labs facilitate collaboration among stakeholders in industry, research, education, government & policy on specific data-enabled themes and applications like Interoperability, 5G, Internet of Things, and Artificial Intelligence, addressing specific issues/ problems in different sectors.

iv. Leverage data as training data for AI/ML systems
    - Without data, there cannot be AI systems. And without its own world class AI systems in key sectors, no country can be a contender among

top global economies in the digital era. If India is to have its own large scale AI systems in domains like health, agriculture, urban mobility, education etc. it needs to create large databases of high-quality datasets specific to Indian conditions / context. For example, A computer vision algorithm in autonomous vehicle context working at about 80% efficiency in Western conditions, may work only at about 40% efficiency on Indian roads. Hence the need for India road conditions data to be captured and used for training autonomous vehicles [33] and enabling wide access to such data for Indian companies, including startups.

- Organizations (public, private, startups, research etc.) may be eligible to run their respective algorithms on centralized anonymised data systems (even without necessarily giving them access to download the underlying data) and thus train their AI systems and develop their potentially market-disrupting solutions or otherwise generally useful and/or competitive, solutions, products and services. This will need development of necessary data infrastructures or data trusts, which may require a proactive approach by the data regulator concerned.
- Raw / factual data on its own cannot be used as training data for AI systems. Raw / factual data needs to be labelled properly, and input data and expected results provided to act as training data. Incentive mechanisms need to be developed to allow data collectors to provide AI training datasets or for specialised data service providers (for example an Indian startup) to do the required labelling. A role for third party data infrastructures or trusts again gets underlined here to meet such an imperative, which can integrate the services of such specialised data service providers for labelling data and so on.

Data Sharing Mechanisms

Recommendation 6: Defining Data-Sharing Mechanisms and Checks and Balances

7.4. Appropriate data sharing mechanisms for sharing public, community and private data need to be established.

i. The Government should improve on existing Open Government Data initiatives, and should ensure that high-quality Public Non-Personal Datasets are available.

33 itihaasa Research and Digital, "Landscape of Artificial Intelligence / Machine Learning Research in India", 2018-19, http://www.itihaasa.com/pdf/itihaasa_AI_Research_Report.pdf

ii. Ensure horizontal applicability of data sharing principles to all Non-Personal Data – Public, Community as well as Private (i.e. raw / factual data that pertains to a community). This will enable greater data sharing and lead to an overall increase in quality of shared data.

iii. With respect to sharing private data, the following mechanisms may be developed:
   - Only the raw / factual datapertaining to community data that is collected by a private organization needto be shared, subject to well-defined grounds (refer to Recommendation 5)atno remuneration.
   - At points or levels where processing value-add is non-trivial with respect to the value or collective contribution of the original community data and collective community resources used, (or otherwise for reasons of over-riding public interest) data sharing may still be mandated but on FRAND (fair, reasonable and non-discriminatory) based remuneration.
   - Subsequently, with increasing value-add it may just be required that the concerned data is brought to a well-regulated data market and price be allowed to be determined by market forces, within general frameworks of openness, fairness etc.
   - And, at a certain level of high value-add it may indeed largely be left to the private organisation that collects the data as to how it wishes to use the data, whereby economic privileges – even if only de facto – are mostly considered now to appropriately inhere in it.

7.5. Data sharing mechanisms should consider the following:

i. As we have discussed, Indian citizens and organizations would have access to the meta-data about data collected by different Data Businesses. By looking at the meta-data, different stakeholders may identify opportunities for combining data from multiple Data Businesses and/or governments to develop innovative solutions, products and services.

ii. The process of data sharing starts with a data request being made to the relevant Data Business. The data requests may be made for the detailed underlying data.

iii. A business including start-up may raise a data sharing request to a data custodian based on the meta-data of the data custodian. If the data custodian services the request, the transaction is complete.

iv. If the data custodian refuses to share the request, the request is made to the Non-Personal Data Authority (refer to Chapter 8). The authority evaluates the

request from social/ public/ economic benefit perspective. If the request is genuine and can result in such benefits, the authority will request the data custodian to share the raw/factual data. If the authority determines that the benefits are not real, the request is denied.

v. When the data is to be shared under this request, the data trustee may decide to make the data available for encouraging domestic startup-up activity, based on a determination of the best way this data can spur more innovation and Indian economic benefit. Data trustee may create a data trust to manage this public use database to ensure de-anonymisation concerns are fully addressed.

vi. Similarly, the data custodian may decide that the data insights and derivates they have are valuable and may decide to create value-added services beyond the raw data and may market data services with such value-added data.

vii. Refer to **Appendix 5**for the background information that the Committee considered on data sharing mechanisms and approaches in other countries.

7.6. Checks and balances – There are a few checks and balances emplaced to ensure appropriate implementation of the rules and regulations with respect to data sharing.

i. Location – The directories / databases contain data from multiple facets of people's lives that are prone to deanonymisation and if exposed would constitute a critical loss of privacy. Hence, the location of these Non-Personal Data may follow guidelines derived from the corresponding Personal Data related provisions of Clause 33 of PDP Bill.
   - Sensitive Non-Personal Data may be transferred outside India, but shall continue to be stored within India.
   - Critical Non-Personal Data (which will follow the definition of Critical Personal Data which is to be notified by the Central Government) can only be stored and processed in India.
   - General Non-Personal Data may be stored and processed anywhere in the world.
   - For all Indian community Non-Personal Data or public Non-Personal Data taken outside India, Indian law and regulation will continue to primarily apply on such data, in precedence over any other jurisdiction's law or regulation. This will include data sharing requirements as needed and legitimately called for. The safeguards may be in the form of obligations, bilateral arrangements etc. based on the potential risks considering protections available in the foreign jurisdictions. Those who take Indian community or public Non-Personal Data outside India will bear full legal responsibility for complying

with any such immediate or future data sharing or other regulatory requirements.

ii. Contract – Both the cloud provider and data business agree contractually to comply with the terms of storage, processing and usage of this dataas specified by the data regulator.

iii. Tools – Testing and probing tools are continuously run on the data in these secure clouds and reports generated, auto-submitted by cloud providers and registered organisations to check compliance.

iv. Expert Probing – Registered experts, registered academic labs and registered Indian organisation, so registered through a self-serve peer review process, are encouraged to probe the released /sharedaggregate data, the cloud defences and cloud internals (via interfaces given to registered organisations) for vulnerabilities including the risk of reidentification, report them via the regulator's APIs to the authority as well as the relevant entity in real-time with guaranteed SLAs for acceptance, mitigation and public notification post mitigation.

v. Academic-IndustryAdvisory Body – A joint Indian advisory body headed by a globally recognised technical expert can suo motu suggest changes to the standards, algorithms and fund improvements of these tools and systems.

vi. Liability – One reason for standards driven approach is that organisations, that comply thoroughly with the laid-down standards via annual light weight self-reported, self-audited digital compliance reports, exhibit good faith and havebest-effort internal processes in-line with the best of industry standards, are to be indemnified against any vulnerability found as long as they swiftly remedy it.

8. Non-Personal Data Regulatory Authority

The Committee discussed a set of policies, regulations, rules and systems that needs to be put in place to govern Data Businesses and data sharing. It also studied the emerging rules and regulations in other countries governing Non-Personal Data and data sharing. For the latter, refer to **Appendix 6.**

The Government of India has tabled in Parliament the PDP Bill. In as much as the regulation of personal data is driven by the need to protect data principals who provide their information from a violation of their personal privacy, the regulation of Non-Personal Data would be driven by the need to unlock the value inherent in this form of data, as well as to protect from collective harms.The regulation of Non-Personal Data should also be driven by the need to ensure that its permitted use does not result in the unauthorised re-identification of the individuals contributing to that aggregate data.

Non-Personal Data Authority– Roles and Responsibilities

Recommendation 7: Establishinga Non-Personal Data Authority

8.1. The Committee discussed the creation of a Non-Personal Data Authority in considerable detail. It considered options like

i. Can the sharing of Non-Personal Data be self-regulated by business and other stakeholders?

ii. Can various sectoral regulators address issues that are related to Non-Personal Data?

iii. Can the Data Protection Authority(DPA), proposed in the PDP Bill, address Non-Personal Data too in coordination with the Competition Commission of India (CCI) and other sectoral regulators?

iv. Can a department within the Government coordinate the roles of various regulators such as the DPA, the CCI, and other sector regulators to regulate Non-Personal Data?

8.2. Ultimately, the Committee felt that the best option is to create a separate Non-Personal Data Authority.

i. This is a new and emerging area of regulation. The regulatory authority will need specialized knowledge (of data governance, technology, latest research and innovation in the space of Non-Personal Data, etc.) and will have to keep pace with the rapidly evolving technological landscape.

ii. The nature of tasks and focus required of this authority are quite different from those of existing ones.
   - Unlike the DPA which is focussed on prevention of personal harm, this authority will focus on unlocking value in Non Personal Data for India.
   - Unlike CCI, this authority will be a proactive actor providing early and continued support for Indian digital industry and startups, and ensuring that necessary data is available for all the needed social, public and economic purposes. This authority must evaluate the nature of data sharing requests to avoid unfair or spurious requests which don't serve social, public or economic purposes.
   - Unlike sector regulators, this authority will have the expertise and a cross-cutting view and role for ensuring data sharing (which requirement often crosses sectoral boundaries), and sectoral regulators can build additional data regulations etc. if required, over those developed by this authority in a horizontal fashion.
   - This authority should work in consultation with the DPA, CCI and other sector regulators, as appropriate, so that issues around data sharing, competition, re-identification or collective privacy are harmoniously dealt with.

iii. Such a new authority will have two roles to play – 1) Enabling role and 2) Enforcing role

Key Takeaways – The Non-Personal Data Authority has two roles to play

- Enabling role: ensuring that data is shared for sovereign, social welfare, economic welfare and regulatory and competition purposes and thus spurring innovation in the country
- Enforcing role: ensuring all stakeholders follow the rules and regulations laid, provide data appropriately when data requests are made, undertaking ex-ante evaluations of the risk of re-identification of anonymised personal dataand so on.

iv. Thus, the Non-Personal Data Authority should be taskedwith enabling legitimate sharing requests and requirements, and with regulating and supervising corresponding data sharing arrangementsinvolving Data Businesses, data trustees and data trusts.

v. The Non-Personal Data Authority should also be tasked with addressing market failures and supervising the market for Non-Personal Data. The harms that such a regulatory agency should be addressing include:

- Lack of information in terms of Non-Personal Data usage, the quantum and nature of actual Non-Personal Data assets held by an enterprise, and the consequential potential harms that could result from such Non-Personal Data collection and processing activities.
- Linked to market failure, addressing any potential negative externalities caused by Non-Personal Data collection and processing activities, including re-identification, deanonymisation, and potential discriminatory harms to customers and communities.
- Lack of sufficient levels of competition, and access to Non-Personal Data, resulting in exploitative (discriminatory terms of transactions vis a vis other businesses or customers) or exclusionary (directed at restricting competition, and raising market entry barriers) harms.

vi. The Authority will administer the Non-Personal Data legislation and its various specific provisions.
- This includes, exercising various powers for regulating 'Data Businesses', defining and updating threshold values for registration as a Data Business, supervising data porting and sharing mandates and requests, managing the meta-data directories, adjudicating on data-sharing disputes, and so on.
- Any mandatory sharing for instance, if not acceded to directly by data holders / custodians, would require determination by the Non-Personal Data Authority.It will base such a decision on the guiding principles laid down in the Non-Personal Data legislation and in consultation with the appropriate data trustee, in cases involving community Non-Personal Data.These will be further clarified in codes of conducts brought out by the Authority.
- Among other tasks, the Authority will also certify rules and technology frameworks for various kinds of data sharing, data safety, anonymisation etc. and set standards in this regard.

vii. The Non-Personal Data Authority will ensure a level playing field for all Indian actors to fulfil the objective of maximising Indian data's value to the Indian economy. Network effects can amplify the benefits for a few mega technology companies that dominate the digital and data business today. This is to ensure fair and effective competition in digital and data marketsand industry, in a proactive manner.

viii. Privacy and Non-Personal Data protection follow the PDP Bill.
- Under Clause 91 of this bill, it is stated that "*Nothing in this Act shall prevent the Central Government from framing of any policy for the digital*

economy, including measures for its growth, security, integrity, prevention of misuse, insofar as such policy do not govern personal data." This clause seeks to enjoin and empower the Central Government to frame policies and regulations for digital economy in respect of Non-Personal Data.

- However, at the same time, Non-Personal Data which can be anonymized Personal Data is not completely immune from risks of re-identification, or accidental identification, and in a manner could lead to privacy harms. For instance, Non-Personal Data relatable to a community could lead to the identification of certain constituent persons of that community. It is for these very reasons that Clause 82(1) of the PDP Bill 2019, provides for offences relating to the re-identification of anonymised data, despite largely keeping anonymised information outside the scope of the Bill.
- Accordingly, The Non-Personal Data Authority must seek to work within the frameworks of the expected privacy legislation and in consultation with the Data Protection Authority, and mitigate such risks with an ex-ante evaluation of the risk of re-identification of anonymised data, prior to approving requests for data-sharing.

ix. The Non-Personal Data Authority will recognize Non-Personal Data ownership rights and privileges and incentives to innovate.

x. The Non-Personal Data Authority will inter alia have some members with relevant industry experience.

8.3. Harmonisation and enabling a Non-Personal Data Act

i. The roles of the proposed Personal Data Authority (from PDP Bill 2019), the Competition Commission of India (under the Competition Act, 2002),and the proposed Non-Personal Data Authority, should be harmonised.

ii. The regulations proposedfor Non-Personal Data can be enforced effectively and at a national scale only if they are incorporated as part of a new national law. The Committee strongly recommends that the proposed Non-Personal Data Governance Framework becomes the basis of a new legislation for regulating Non-Personal Data.

9. TechnologyArchitecture

The Committee consideredsome technology related guiding principles that can be used for creating and functioning of shared data directories / data bases, and for digitally implementing the rules and regulations related to data sharing.

Key Takeaways – Technology Architecture

- API mechanisms for accessing data
- Data security – storage in distributed format
- Creating a standardized data exchange approach (regardless of data type, exchange method or platform)
- Prevent de-anonymization – Best of breed Differential Privacy algorithms

9.1. The guiding principles for such a technology architecture include:

i. Mechanisms for accessing data – A number of different mechanisms exist for accessing data including downloads, Application Programming Interfaces (APIs), and data sandboxes.
   - All sharable Non-Personal Data and datasets created or maintained by government agencies, companies, startups, universities, research labs, non-government organisations, etc. should have a REST (Representational State Transfer) API for accessing the data.
   - Data sandboxes can be created where experiments can be run, algorithms can be deployed and only output being shared, without sharing the data.

ii. Distributed for data security –data storage in a distributed format so that there is no single point of leakage; sharing to be undertaken using APIs only, such that all requests can be tracked and logged; all requests for data must be operated after registering with the company for data access etc.Even when data is stored in a distributed or federated form, as appropriate, there could be coordinated management of them like would be required for data trusts and data infrastructures for important Non-Personal Data in different sectors.

iii. Creating a standardized data exchange approach (regardless of data type, exchange method or platform)
   - Data that is collated should be available appropriately on a data exchange for stakeholders to use and make inferences.
   - Exchange should be able to take-in any form of data and produce output that is standardized and usable to all stakeholders.

iv. Prevent de-anonymization – Best of breed Differential Privacy algorithmsmay be used to create anonymised data. Mechanisms must be put in place to ensure that re-identification of anonymised data does not occur.

- A number of other technologies can come into play in managing data like,differential privacy replaces one data set with another that includes different information, but has the same statistical patterns; Homomorphic encryption allows algorithms to crunch data without decrypting them; and blockchains enable one to manage data access rights[34].

9.2. The Committee has encapsulated these technicalguiding principles into an illustrative three-tiered system architecture spanning legal safeguards, technology and compliance.**Refer to Appendix 7**. There may also be other appropriate ways to technically implement the recommendations of this Committee.

34 https://www.economist.com/special-report/2020/02/20/are-data-more-like-oil-or-sunlight

10. Summary

Data – Trends and Economic Impact

i. The world is awash with data.The proliferation of big data, analytics and Artificial Intelligence (AI) has led to the creation of many new information intensive services and also the transformation of existing businesses.

ii. Data inter alia contributes to economic value and wealth. Frameworks are being created to better understand the uses and benefits of data.

iii. Organizations have been discovering ways to generate value from data. The digital economy is witnessing the emergence of a few dominant players and a certain imbalance in the market.

iv. Given the increasing importance and value generation capacity of the data economy, governments around the world realise the need to enable and regulate all aspects of data, both Personal and Non-Personal Data.

Case for regulating data

i. The Committee believes that rules and regulations are required to manage data in order to achieve the following enabling and enforcing benefits:
   - Come up with a set of recommendations such that India can create a modern framework for creation of economic value from use of Data. To generate economic benefits for citizens and communities in India and unlock the immense potential for social / public / economic value data.
   - To create certainty and incentives for innovation and new products / services creation in India. To encourage start-ups in India.
   - To create a data sharing framework such that community data is available for social / public / economic value creation
   - To address privacy concerns, including from re-identification of anonymised personal data, preventing collective harms arising from processing of Non-Personal Data, and to examine the concept of collective privacy.

ii. The case made for regulating data is made in such a manner that the benefits accrue to India and its communities and businesses.

Recommendation 1: Defining Non-Personal Data

i. The Committee has defined three categories of Non-Personal Data – 1) Public Non-Personal Data 2) Community Non-Personal Data & 3) Private Non-Personal Data.

ii. The Committee has also defined a new concept of 'sensitivity of Non-Personal Data', as even Non-Personal Data could be sensitive from the following perspectives – 1) It relates to national security or strategic interests; 2) It is business sensitive or confidential information; 3) It is anonymised data, that bears a risk of re-identification
    - The Committee recommends that Non-Personal Data inherits the sensitivity characteristic of the underlying Personal Data from which the Non-Personal Data is derived.

iii. The Committee recommends that the data principal should also provide consent for anonymisation and usage of this anonymized data while providing consent for collection and usage of his/her personal data. Also, the Committee recommends that appropriate standards of anonymisation be defined to prevent / minimize the risks of re-identification.

Recommendation 2: Defining Key Non-Personal Data Roles

i. There are three key Non-Personal Data roles, namely data principal, data custodian, and data trustee; and an institutional form of data infrastructures, namely a data trust.
    - In case of Government and Private Non-Personal Data, the data principal will be the corresponding entities (individuals, companies, communities) to whom the data relates.In case of Community Data, a community is the data principal.
    - The data custodian undertakes collection, storage, processing, use, etc. of data in a manner that is in the best interest of the data principal.
    - The data principal group/community will exercise its data rights through an appropriate data trustee.
    - Data trusts are the institutional structures, comprising specific rules and protocols for containing and sharing a given set of data.The term data infrastructure further brings in also the corresponding technical-material elements required for data sharing, like actual databases, APIs, organisational systems, etc.

Recommendation 3: Articulating a legal basis for establishing rights over Non-Personal Data

i. The laws, regulations and rules of the Indian State apply to all the data collected in/from India or by Indian entities.

ii. The term "ownership" holds full meaning only in terms of physical assets. For intangible assets like data, many actors may have simultaneous overlapping rights and privileges.Accordingly, the notion of "beneficial ownership/interest"

has been adopted by the Committee, in ensuring that a Community's interests are safeguarded.

iii. In the case of Community Non-Personal Data
   - 'Data source' logic of Community Non-Personal Data ownership – Just like the economic rights to natural resources arising from a community are considered to primarily belong to it, the value of social resources of Community Non-Personal Data should primarily accrue to it (instead of the default whereby data custodians take up the entire value of such data).
   - 'Data subject' logic of Community Non-Personal Data – it is data about the concerned group/community and provides systemic intelligence about it. The group/community should be able to determine and control how such data and intelligence is used – maximising data's benefits for itself, and eliminating or minimizing harms.
   - The Community Non-Personal Data collected in or from India belongs to the community concerned. The rights over this data collected in India should vest with the trustee of that community, with the community being the beneficial owner.

iv. In case of Public Non-Personal Data
   - Since this data is derived from public efforts, the datasets created partake of the characteristics of a national resource.

v. In the case of Private Non-Personal Data
   - Only the raw / factual data (the principle of raw data is standards compliant, machine readable and fidelity as collected) pertaining to community data that is collected by a private organization may need to be shared subject to the well-defined grounds.
   - As the processing value-add over the raw data increases, appropriate mechanisms may be leveraged for data sharing.
   - Algorithms / proprietary knowledge may not be considered for data sharing.

Recommendation 4: Defining a Data Business

i. Organizations are deriving new or additional economic value from data, by collecting, storing, processing, and managing data. For instance, a hospital derives economic value not only from providing medical services, it may derive additional value by harnessing the medical data and offering value-added services (such as personalized treatment plans, medicines etc.). Hence create a new category / taxonomy of business called 'Data Business' that meets certain data threshold criteria.

ii. Data Business is a horizontal classification and not an independent industry sector.Many existing businesses in various sectors, collecting data beyond a threshold level, will get categorized as a Data Business.

iii. It is important for such Data Business to register once they reach a certain data-related threshold. This will be applicable to not only commercial organizations, but also Governments and other non-government organizations that collect, process or otherwise manage data.
   - Below the threshold, the registration as a Data Business may be voluntary.
   - This is a one-time activity and there is no necessity to obtain a license to be a Data Business.

iv. Every Data Business must declare what they do and what data they collect, process and use, in which manner, and for what purposes (like disclosure of data elements collected, where data is stored, standards adopted to store and secure data, nature of data processing and data services provided). This is similar to disclosures required by pharma industry and in food products.

v. The compliance process will be light-weight and fully digital.

vi. The meta-data about data being collected, stored and processed by Data Businesses will be stored digitally in meta-data directories in India. There will be open access within India to meta-data and regulated access to the underlying data.

vii. Indian citizens and India-based organizations will have open access to the meta-data of the data collected by different Data Business and Government. By looking at the meta-data, potential users may identify opportunities for combining data from multiple Data Businesses and/or Government to develop innovative solutions. Subsequently, data requests may be made for the detailed underlying data.

viii. The Committee strongly believes that meta-data sharing by Data Business will spur innovation at an unprecedented scale in the country.
   - One of the associated key objectives is to promote and encourage the development of domestic industry and startups that can scale their data businesses.
   - For example, automobile companies may collect data about roads through various sensors. A startup, will know that this data is available based on the meta-data provided by automobile companies. The startup can request for access for this data and can combine this data with public traffic data to create a solution for safest road routes for senior citizens.

Recommendation 5: Define Data-Sharing Purpose

i. Sovereign purposes – Data may be requested security, legal, law enforcement and regulatory purposes. For instance,
   - Data requested for mapping security vulnerabilities and challenges, including people's security, physical infrastructure security and cyber security.
   - Data required for crime mapping, devising anticipation and preventive measures, and for investigations and law enforcement.
   - Data required for pandemic mapping, prediction and prevention, and also subsequent interventions.

ii. Core Public Interest purposes – Data may be requested for community uses / benefits or public goods, research and innovation, for better delivery of public-services, policy development, etc.
   - Certain data held by private sector when combined with public-sector data or otherwise may be useful for policy making, improving public service, devising public programs, infrastructures, etc. and, in general, supporting a wide range of societal objectives including science, healthcare, urban planning etc.
   - India to specify some high-value datasets
   - Utilize data for research purposes
   - Consider health sector as a pilot use-case for Non-Personal Data Governance Framework.

iii. Economic purposes – Data may be requested for economic welfare purposes – in order to encourage competition and provide a level playing field in any sector, including, very importantly, for enabling domestic startup activities, or for a fair monetary consideration as part of a well-regulated data market, etc.
   - Data request by startups
   - Data request by data trustees / governments
   - Setting up data &cloud innovation labs and research centres to develop, test and implement new digital solutions
   - Leverage data as training data for AI/ML systems

Recommendation 6: Defining Data-Sharing Mechanisms and Checks and Balances

i. Appropriate data sharing mechanisms for sharing public, community and private data need to be established.

ii. The Government should improve on existing Open Government Data initiatives, and should ensure that high-quality Public Non-Personal Datasets are available.

iii. With respect to sharing private data, the following mechanisms may be developed:

- Only the raw / factual datapertaining to community data that is collected by a private organization may needto be shared, subject to well-defined grounds and not based on any remuneration.
- At points or levels where processing value-add is still moderate with respect to the value or collective contribution of the original community data and collective community resources used, (or otherwise for reasons of over-riding public interest) data sharing may still be mandated but on FRAND (fair, reasonable and non-discriminatory) based remuneration.
- Subsequently, with increasing value-add it may just be required that the concerned data is brought to a well-regulated data market and price be allowed to be determined by market forces, within general frameworks of openness, fairness etc.
- And, at a certain level of high value-add it may indeed largely be left to the private organisation that collects the data as to how it wishes to use the data, whereby economic privileges – even if only de facto – are mostly considered now to appropriately inhere in it. Thus, algorithms / proprietary knowledge may not be considered for data sharing.

iv. The process of data sharing starts with a data request being made to the relevant Data Business. The data requests may be made for the detailed underlying data.

v. A business including start-up may raise a data sharing request to a data custodian based on the meta-data of the data custodian. If the data custodian services the request, the transaction is complete.

vi. If the data custodian refuses to share the request, the request is made to the Non-Personal Data Authority (refer to Chapter 8). The authority evaluates the request from social / public / economic benefit perspective. If the request is genuine and can result in such benefits, the authority will request the data custodian to share the raw/factual data. If the authority determines that the benefits are not real, the request is denied.

vii. When the data is to be shared under this request, the data trustee may decide to make the data available for encouraging domestic startup-up activity, based on a determination of the best way this data can spur more innovation and Indian economic benefit. Data trustee may create a data trust to manage this public use database to ensure de-anonymisation concerns are fully addressed.

viii. Similarly, the data custodian may decide that the data insights and derivates they have are valuable and may decide to create value-added services beyond the raw data and may market data services with such value-added data.

ix. There are a few checks and balances (location, contract, tools, liability etc.) emplaced to ensure appropriate implementation of the rules and regulations with respect to data sharing.
   - Location – The directories / databases contain data from multiple facets of people's lives that are prone to deanonymisation and if exposed would constitute a critical loss of privacy. Hence, the location of these Non-Personal Data may follow guidelines derived from the corresponding Personal Data related provisions of Clause 33 of PDP Bill. For example, Sensitive Non-Personal Data may be transferred outside India, but shall continue to be stored within India; Critical Non-Personal Data can only be stored and processed in India.
   - Contract – Both the cloud provider and data business must agree contractually to comply with the terms of storage, processing and usage of this data as specified by the data regulator.
   - Tools – Testing and probing tools are to be continuously run on the data in secure clouds and reports generated, auto-submitted by cloud providers and registered organisations to check compliance.
   - Liability – One reason for standards driven approach is that organisations, that comply thoroughly with the laid-down standards via annual light weight self-reported, self-audited digital compliance reports, exhibit good faith and have best-effort internal processes in-line with the best of industry standards, are to be indemnified against any vulnerability found as long as they swiftly remedy it.

Recommendation 7: Defining a Non-Personal Data Authority

i. The Committee felt that the best option is to create a separate Non-Personal Data Authority.
   - This is a new and emerging area of regulation. The regulatory authority will need specialized knowledge (of data governance, technology etc.) and will have to keep pace with the rapidly evolving technological landscape.
   - The nature of tasks and focus required of this authority are quite different from those of existing ones like the Data Protection Authority (DPA), Competition Commission of India (CCI) and sector regulators.
   - This authority should work in consultation with the DPA, CCI and other sector regulators, as appropriate, so that issues around data sharing, competition, re-identification or collective privacy are harmoniously dealt with.

ii. The Non-Personal Data Authority is to be tasked with enabling legitimate sharing requests and requirements; with regulating and supervising corresponding data sharing arrangements involving Data Businesses, data trustees and data trusts;

with addressing market failures and supervising the market for Non-Personal Data.

iii. The Non-Personal Data Authority has two roles to play
   - Enabling role: ensuring that data is shared for sovereign, social welfare, economic welfare and regulatory and competition purposes and thus spurring innovation, economic growth and social well-being in the country.
   - Enforcing role: ensuring all stakeholders follow the rules and regulations laid, provide data appropriately when legitimate data requests are made, undertaking ex-ante evaluations of the risk of re-identification of anonymised personal dataand so on.

iv. The roles of the proposed Personal Data Authority (from PDP Bill 2019), the Competition Commission of India (under the Competition Act, 2002), and the proposed Non-Personal Data Authority, should be harmonised.

v. The regulations proposed for Non-Personal Data can be enforced effectively and at a national scale only if they are incorporated as part of a new national law. The Committee strongly recommends that the proposed Non-Personal Data Governance Framework becomes the basis of a new legislation for regulating Non-Personal Data.

Technology Architecture

i. The Committee considered some technology related guiding principles that can be used for creating and functioning of shared data directories / data bases, and for digitally implementing the rules and regulations related to data sharing.
   - API mechanisms for accessing data
   - Data security – storage in distributed format
   - Creating a standardized data exchange approach (regardless of data type, exchange method or platform)
   - Prevent de-anonymization – Best of breed Differential Privacy algorithms

Appendix 1: List of Committee Members

Members of the Committee

i)	Shri Kris Gopalakrishnan, Co-Founder Infosys	Chairman
ii)	Additional Secretary / Joint Secretary, DPIIT	Member
iii)	Ms. Debjani Ghosh, President NASSCOM	Member
iv)	Dr. Neeta Verma, DG, National Informatics Centre	Member
v)	Shri Lalitesh Katragadda, Founder Indihood	Member
vi)	Prof. Ponnurangam Kumaraguru, IIIT Delhi	Member
vii)	Shri. Parminder Jeet Singh, IT for Change	Member
viii)	Additional Secretary, MeitY	Member Convenor
ix)	Krishnan Narayanan, N. Dayasindhu, itihaasa Research and Digital	ReportPreparation

Appendix 2: Examples of Non-Personal Data

1. Data can be categorised in many ways; the subject of data (e.g. personal data); in relation to its purpose (e.g. AI training data, e-Commerce data); the sector to which it belongs (e.g. health data); the source of data (e.g. soil data); level of processing (raw / factual versus derived data); or the collector of data (e.g. public / Government or private data); or based the extent of involvement of stakeholders in the creation of data (provided, observed, derived, or inferred).

2. A mixed dataset, which represent a majority of datasets used in the data economy, consists of both personal and Non-Personal Data.
    i. In the European Union context, the Non-Personal Data Regulation applies to the Non-Personal Data of mixed datasets; if the Non-Personal Data part and the personal data parts are 'inextricably linked', General Data Protection Regulation apply to the whole mixed dataset.

3. Categorisation of data based on its creation[35] – A categorisation of data can help assess the extent to which different stakeholders are involved in the creation of data, including cases where users (consumers and businesses) interact with a data product (good or service) such as an e-government service, a social networking service, etc.
    i. One approach includes four categories of data: i) provided (applications registrations, survey responses, social network postings etc.); ii) observed (cookies on a website, data from sensors etc.); iii) derived (computational scores, classification based on common attributes etc.); and iv) inferred data (scores developed using statistical, advanced analytical techniques, or AI/ML).
    i. Such a categorization helps in framing regulation & policy. For example, in the European Union, the right to data portability under the GDPR would include 'provided' as well as 'observed' data. It would however exclude data 'derived' (& 'inferred') from additional processing – data that are often considered proprietary.

4. Anonymised Data
    i. Anonymisation allows data to be shared, whilst preserving privacy. The process of anonymising data requires that identifiers (both direct identifiers like names and indirect identifiers like age or occupation) are changed in some way such as being removed, substituted, distorted, generalised or aggregated[36].

35 OECD, "Enhancing Access to and Sharing of Data : Reconciling Risks and Benefits for Data Re-use across Societies", 2019, https://www.oecd-ilibrary.org/science-and-technology/enhancing-access-to-and-sharing-of-data_276aaca8-en
36 https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation.aspx

ii. However, new research[37] shows that current methods for anonymizing data still leave individuals at risk of being re-identified. So, policymakers should be careful about what constitutes anonymised data. Also, the technical specifications and architecture should ensure that the chances of re-identifying anonymised data are minimised significantly.

iii. The Committee has collated some of the basic anonymisation techniques in this report in Appendix 3: Primer on Anonymity.

5. AI Training Data

i. During the development of an AI system, three different sets of data are required to train, fine-tune and test the machine learning models. They include the training dataset, the validation dataset, and the testing dataset. The training data set will include input data and expected results and is used to train a machine learning algorithm. These are typically mixed data sets.

ii. Training data for autonomous vehicles in India would include data on Indian roads and vehicles. Training data on fashion purchases in India would include data on purchases of clothes and the buyers on an e-Commerce platform.

iii. In the case of Generative Adversarial Networks (GANs), two AI engines compete against each other to produce data for reinforced learning for the underlying AI system. This may be considered an example of a derived Non-Personal Data.

6. E-Commerce Data

i. e-Commerce data relate to customers' orders, needs, preferences, interests, shopping patterns, feedback, customer satisfaction level, delivery times etc. It also includes insights related to products on the store, competitors' data, and technical data as well. These are typically mixed data sets.

ii. Typical e-Commerce Data attributes[38]

- Customer demographics like age, gender, location
- Product Discovery KPI - the factors that help understand how and when customers find the product
- Onsite traffic metrics - the factors that reveal the amount and time of traffic to a web store
- Email / social media engagement
- Conversion attributes - conversion of visitors into customers on a particular e-Commerce store

7. Government or Public Data

[37] https://www.sciencedaily.com/releases/2019/07/190723110523.htm
38 https://datarade.ai/data-categories/ecommerce-data/guide

i. The Open Government Data (OGD) Platform of India or data.gov.in is a platform supporting the open data initiative of Government of India. It provides access to datasets and documents published by ministries / departments of the Government of India. India may build on its OGD initiative and expand on its national data strategy.

ii. Some countries have started to specify a new class of data at a national level – data of public interest or high-value dataset, like geospatial and/or transportation data. The Governments are combining both public and private sector data as well as personal and Non-Personal Data to create such data of public interest[39].

- Australia has classified its Geocoded National Address File (G-NAF) as one of its most high-valued data sets.
- In Germany, the government has established the research initiative mFUND, to support the development of data-based business models for smart mobility (Mobility 4.0). A central aspect for the programme is the provision of mobility and geo-data (e.g. transport and traffic data, hydrological data, climate and weather data). For this purpose, data access and sharing are promoted according to open data principles and technically supported by the creation of a central, open data access point for mobility-related data (mCLOUD). This initiative is funded by the German Federal Ministry of Transport and Digital Infrastructure with EUR 150 million to be invested between 2016 and 2020.
- National governments have started to specify a new class of data at a national level – data of public interest or high-value dataset, like geospatial and/or transportation data[40].
- The European Commission is proposing to create nine common European data spaces - industrial (manufacturing), green deal, mobility, health, financial, energy, agriculture, public administration, and skills[41].
- The European Commission has identified six data types that appear to have the most value: geospatial, earth observation and environmental, meteorological, statistics, company data, and transport data[42].

39 OECD, "Enhancing Access to and Sharing of Data : Reconciling Risks and Benefits for Data Re-use across Societies", 2019, https://www.oecd-ilibrary.org/science-and-technology/enhancing-access-to-and-sharing-of-data_276aaca8-en

40 OECD, "Enhancing Access to and Sharing of Data : Reconciling Risks and Benefits for Data Re-use across Societies", 2019,

41 https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

42 Open Knowledge Foundation, 'What data counts in Europe? Towards a public debate on Europe's high value data and the PSI Directive', 2019

8. Community Data
    i. Community Data is data about a community (a community is a collection of people bound together by common purpose, objective or geography) and is an example of non-personal data. Both factual data and wisdom-of-the-crowd constitute community data.
        - Factual data of the community, of its habitat or resources.
        - And, non-factual, non-personal, creative data when contributed by a community with deemed value to that or other community collectively constitutes wisdom of the crowd.
    ii. Examples of community data include, climate conditions or weather data, aggregate data of how many cabs are there on the road in an Indian city, etc.
    iii. A refined definition of community and community Non-Personal Data is provided in Section 4.3 of the report.

9. Private Data
    i. Private Non-Personal Data is data collected by private players from and about things, processes, etc that are entirely private to them, or owned by them, or those aspects of 'derived data' which arise from private effort
        - It includes inferred or derived data / insights involving application of algorithms, propriety knowledge.
        - The example of two AI engines competing against each other to produce derived data for reinforced learning for the underlying AI system is an example of private Non-Personal Data.
        - It may also include a global dataset that contains information about non-residents collected in foreign jurisdictions (other than India).

10. Sensitivity of Non-Personal Data
    i. In the case of Personal Data sensitivity spectrum, there exist three categories – General, Sensitive and Critical. Sensitivity of Non-Personal Data also needs to be considered.
    ii. There are other frameworks[43] that have categorised data along its sensitivity spectrum, based on the data's connection with a natural person, their propensity for being kept private, whether they are ascriptive or not, whether they are legally protected from discrimination or not, and their connection with sensitive issues such as beliefs or health.

43 John MM Rumbold et al., "What are data? A categorization of the data sensitivity spectrum", School of Science and Technology, Nottingham Trent University, 2014

Appendix 3: Primer on Anonymity

A primer on anonymisation techniques is provided here. Some of these techniques are academic pursuits and some of them are methods already used in industry tools.

1. K-anonymity[44]
   i. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release[45]. This is one of the most popular and old techniques for structured data.

2. L-diversity[46]
   i. The l-diversity model is an extension of the k-anonymity model which reduces the granularity of data representation using techniques including generalization and suppression such that any given record maps onto at least k-1 other records in the data. The l-diversity model handles some of the weaknesses in the k-anonymity model where protected identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group exhibit homogeneity. The l-diversity model adds the promotion of intra-group diversity for sensitive values in the anonymization mechanism[47].

3. T-closeness[48]
   i. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness[49].

4. Diffix (High-Utility Database Anonymization)[50]
   i. Diffix acts as an SQL proxy between the analyst and an unmodified live database. Diffix adds a minimal amount of noise to answers—Gaussian with a standard deviation of only two for counting queries—and places no limit on the number of

[44] http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf
[45] https://en.wikipedia.org/wiki/K-anonymity
[46] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and MuthuramakrishnanVenkitasubramaniam. 2007. L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1, Article 3 (March 2007). DOI=http://dx.doi.org/10.1145/1217299.1217302https://personal.utdallas.edu/~muratk/courses/privacy08f_files/ldiversity.pdf
[47] https://en.wikipedia.org/wiki/L-diversity
[48] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, 2007, pp. 106-115. doi: 10.1109/ICDE.2007.367856 https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf
[49] https://en.wikipedia.org/wiki/T-closeness
[50] https://aircloak.com/wp-content/uploads/Diffix-High-Utility-Database-Anonymization.pdf

queries an analyst may make. Diffix works with any type of data and configuration is simple and data-independent: the administrator does not need to consider the identifiability or sensitivity of the data itself.

5. ARX[51]
   i. ARX is another tool to anonymize data. ARX is divided into four perspectives, which model different aspects of the anonymization process. As is shown below, these perspectives support 1) configuring privacy models, utility measures and transformation methods, 2) exploring the solution space, 3) analysing data utility and 4) analysing privacy risks. ARX is built on many research publications, they have a team to maintain the code, bug fixes, etc.

6. Amnesia[52]
   i. Amnesia is a flexible data anonymization tool that transforms relational and transactional databases to dataset where formal privacy guaranties hold.
   ii. Amnesia is a data anonymization tool, that allows to remove identifying information from data. Amnesia not only removes direct identifiers like names, SSNs etc but also transforms secondary identifiers like birth date and zip code so that individuals cannot be identified in the data. Amnesia supports k-anonymity and km-anonymity.
   iii. km-anonymity requires that each combination of up to m quasi identifiers must appear at least k times in the published data. The intuition behind km-anonymity is that there is little privacy gain from protecting against adversaries who already know most of the terms of one record, and significant information loss in the effort to do so.
   iv. There is an online GUI based system of Amnesia[53].

7. μ-ARGUS & τ-ARGUS[54]
   i. μ-ARGUS to be used to protect microdata and τ-ARGUS to be used to protect tabular data.
   ii. These tools are available in both Windows and other platforms[55].

8. Anonimatron[56]
   i. There are also publicly available open source projects on anonymization, including GDPR compliant testing. Some of the features of Anonimatron are:

[51] https://arx.deidentifier.org/
[52] https://amnesia.openaire.eu/
[53] https://amnesia.openaire.eu/amnesia/
[54] http://neon.vb.cbs.nl/casc/mu.htm
[55] https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_7_PPdeWolf.pdf
[56] https://realrolfje.github.io/anonimatron/

- Anonymize data in databases and files.
- Generates fake email addresses, fake Roman names, and UUID's out of the box.
- Easy to configure, automatically generates example config file.
- Anonymized data is consistent between runs. No need to re-write your tests to handle random data.
- Extendable, easily implement and add your own anonymization handlers
- 100% Java 1.8, multi-platform, runs on Windows, Mac OSX, Linux derivatives.
- Multi database, uses SQL92 standards and supports Oracle, PostgreSQL and MySQL out of the box. Anonimatron will autodetect the following JDBC drivers: DB2, MsSQL, Cloudscape, Pointbase, Firebird, IDS, Informix, Enhydra, Interbase, Hypersonic, jTurbo, SQLServer and Sybase.

9. Differential Privacy
   i. Goal is to perform aggregative analysis (statistics about the data) without compromising the privacy of an individual data point[57]. Differential privacy offers strong and robust guarantees that facilitate modular design and analysis of differentially private mechanisms due to its composability, robustness to post-processing, and graceful degradation in the presence of correlated data[58]. This method is prominently used in technological implementations now, etc. Apply uses differential privacy in its iPhone.

[57] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Third conference on Theory of Cryptography (TCC'06), Shai Halevi and Tal Rabin (Eds.). Springer-Verlag, Berlin, Heidelberg, 265–284. https://link.springer.com/chapter/10.1007%2F11681878_14

[58] https://en.wikipedia.org/wiki/Differential_privacy

Appendix 4: Emerging Global Frameworks related to Data Business

In the data economy, the proliferation of big data, analytics and AI has led to the creation of information intensive services where information interactions exert the greatest effect on value creation. Thus, a new category of business, 'Data Business', may be envisaged that collects / manages / or otherwise manages data, and meets certain threshold criteria.

i. One study[59] developed a nine-factor framework for data-based value creation in information-intensive services. The factors include (1) data source, (2) data collection, (3) data, (4) data analysis, (5) information on the data source, (6) information delivery, (7) customer (information user), (8) value in information use, and (9) provider network.

Globally, such a concept of defining a new category of 'Data Business' is only emerging. Here are a few examples of related global taxonomies.

1. Bureau of Economic Analysis (BEA), USA definition of Digital Economy[60] – BEA in a 2018 working paper includes the following categories under Digital Economy:
   i. Digital-enabling infrastructure needed for a computer network to exist and operate – computer hardware, software, telecommunications equipment and services, structures like data centres, IoT, and support services
   ii. e-Commerce – digital transactions that take place using that system – Business-to-business (B2B) e-commerce, Business-to-consumer (B2C) e-commerce, Peer-to-peer (P2P) e-commerce
   iii. Digital media – the content that digital economy users create and access

2. OECD classification of data-enabled services[61] – In a 2018 paper on recording and measuring data, OECD categorizes data-enabled services as follows:
   i. Providing services for free or at very low prices to gather data of users which are subsequently used to detect behavioural patterns to provide other producers with targeted advertising services (like Google Ads, Facebook, etc.), or to offer other services (e.g. using information from payment systems etc.)
   ii. Using data generated as part of the primary production process, to improve the efficiency of the internal operations and/or to detect behavioural pattern to

59 Chiehyeon Lim et al., "From data to value: A nine-factor framework for data-based value creation in information-intensive services", International Journal of Information Management, Volume 39, April 2018, Pages 121-135
60https://www.bea.gov/sites/default/files/papers/defining-and-measuring-the-digital-economy.pdf
61http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=SDD/CSSP/WPNA(2018)5&docLanguage=En

support own sales. (like Amazon using dynamically generated recommendations, Walmart using analytics to optimise supply chain and pricing models.)

iii. Creation of new types of services by using and analysing big data.

iv. Provision of data-related services by collecting data from a vast number of different, mostly free, available data sources, normalising formats and providing access, with revenues from subscription or usage fees.

v. Data facilitators, providers of data tools such as providing storage media, servers and workstations, data collection, analysis and visualisation software, database management software, encryption technology and software, data protection technology, etc.

vi. Creation of freely available information or knowledge by communities of people, providing their contributions for free. (like Wikipedia, ResearchGate)

3. A framework[62] for establishing the 'data-drivenness' of a market:

   i. Market definition (user centric) – an index of data-drivenness applied at the industry level would indicate, for instance, industry A has a high degree of data-drivenness and therefore mandatory data sharing is warranted, whereas industry B is only mildly data-driven such that there should be no mandatory data sharing.

   ii. Study the demand side of the market: what drives users' consumption utility?

   iii. Study the supply side of the market: what drives objective measures of product quality?

62 Jens Prüfer, Friedrich-Ebert-Stiftung, "Competition Policy and Data Sharing on Data-driven Markets", 2020, library.fes.de/pdf-files/fes/15999.pdf

Appendix 5: Data Sharing Mechanisms and Frameworks

Data sharing refers to the provision of controlled access to private sector, public sector and community data to individuals and organisations for defined purposes and with appropriate safeguards in place.

1. Types of data sharing
   i. We could classify data sharing by the nature / 'ownership' of data (e.g. Government, private, community etc.) or the manner of its sharing (e.g. closed, open, semi-private etc.)
   ii. There is also likely to be different obligations in sharing provided, observed, derived and inferred data.
   iii. Government data sharing (G2B and G2C): Sharing of public information by the Government for the purposes for re-use by organisations (including companies and startups) and individuals alike.
      - Regulatory examples: Directive (EU) 2019/1024, on open data and the re-use of public sector information
   iv. Community data sharing refers to community datawhich, may be with private actors too, is needed to be shared under community data sharing guidelines and requirements.
   v. Private / Industrial data sharing (B2B): Sharing of industrial data between organisations involved in the same commercial or non-commercial point of the value chain.
      - Examples: International Data Spaces (IDS) Association, Industrial Internet Consortium(IIC), Data Market Austria, Ocean Protocol and the IOTA Foundation[63,64,65,66]
   vi. Open data sharing: Sharing of industrial data inside or outside of a value network, Government public information and the data of willing participants shared in their individual or collective capacity through sharing mechanisms/instrument.
      - With respect to industrial data, when such data is legally open, it means that the data is published under an open license and that the conditions for re-use are limited to attribution. Second, the data is technically open, which means that the file is machine readable and non-proprietary.
      - Regulatory examples: Australian Data Sharing and Release bill, 2018

63 https://www.internationaldataspaces.org/
64 https://www.iiconsortium.org/
65 https://datamarket.at/
66 https://oceanprotocol.com/

vii. Anonymised data sharing: Sharing of anonymised personal data is important to develop new business or innovation, especially in the context of AI and big data systems.
   - Clause 2 of the Personal Data Protection Bill 2019 in India clarifies that the Act with regard to personal data of Indians (and save for clause 91) would not be applicable to the processing of anonymised data. Even under GDPR, after this process of anonymisation, the data is no longer subject to personal data protection regulations.
   - However, new research shows that current methods for anonymizing data still leave individuals at risk of being re-identified. So, care should be taken about what constitutes anonymised data.
   - Technical specifications and architecture should ensure that the chances of re-identifying anonymised data are minimised significantly.

2. Data sharing mechanisms
   i. Government data sharing
      - Data sharing framework: Building on the framework created by National Data Sharing & Accessibility Policy (NDSAP)[67], the default practice should be proactive release of data upon request generated through the Open Data Portal.
   ii. Community data sharing
      - Data originating from the community and belonging to it, but existing with private parties, may be required to be shared when appropriate as per different sharing mechanisms, mediated by data trusts, etc.
   iii. Private / Industrial data sharing (B2B)
      - Data monetisation: unilateral approach under which companies make additional revenues from the data they share with other companies. Data can also be monetised through the provision of services.
      - Data marketplaces[68]: trusted intermediaries that bring data suppliers and data users together to exchange data in a secure online platform. These businesses make revenue from the data transactions occurring in the platform.
      - Industrial data platforms: collaborative and strategic approach to exchange data among a restricted group of companies and/or startups. They voluntarily join these closed, secure and exclusive environments to foster the development of new products/services and/or to improve their internal efficiency. Data may be shared for free, but fees may also be considered.

67https://data.gov.in/sites/default/files/NDSAP.pdf
68http://www.bdva.eu/sites/default/files/BDVA%20DataSharingSpace%20PositionPaper_April2019_V1.pdf

3. The Government may have to play a role in incentivising and orchestrating data partnerships, either by acting as independent trusted third parties or by engaging with the private sector in Public-Private-Partnership (PPP) mode. This is achieved through appropriate rules and regulations.
    i. For example, the European Commission is examining data sharing between the private and public sector in order to guide policy making and improve public services.
        - Manufacturers of IoT [Internet of Things] objects usually determine access to the non-personal and automatically generated data from IoT objects, which have been triggered by the data-users.

4. In Europe we can see examples like the Finnish Health and Social Data PermitAuthority[69], French Health Data Hub[70], European Open Science Cloud[71] that allows Europe's 1.7 million researchers and 70 million science and technology professionals a virtual environment to store, share and re-use the large volumes of information generated by the big data revolution.

5. There exist frameworks[72] that examine the opportunities of enhancing access to and sharing of data. They highlight the factors that need to be considered including data typologies, key data-access mechanisms and the main types of actors and their roles.

69https://www.findata.fi/en/
70https://www.health-data-hub.fr/
71 EOSC Strategic Implementation Roadmap 2018-2020, https://ec.europa.eu/research/openscience/pdf/eosc_strategic_implementation_roadmap_short.pdf#view=fit&pagemode=none
72https://www.oecd-ilibrary.org/sites/b4d546a9-en/index.html?itemId=/content/component/b4d546a9-en&mimeType=text/html

Appendix 6: Rules and Regulationsaround Data Sharing

To facilitate data sharing, rules and regulations need to be established. These rules and regulations may address aspects like data regulator, user registration, data disclosure requirements, audit requirements, data usage context and others.

1. Different countries are adopting different strategies and experimenting with regulations to govern data.
   i. The European Commission has published a slew of communications on 'A European strategy for data[73]', and 'Shaping Europe's digital future[74]' and a white-paper on 'On Artificial Intelligence - A European approach to excellence and trust[75]'.
   ii. The last G20 meeting launched the 'Osaka Track', a proposed plurilateral agreement on digital trade, that provided global rules for "data governance" based on "free flow of data with trust". India and a few other developing countries have refused to sign up to the Osaka Track[76].
   iii. In Germany, the Federal Ministry for Economic Affairs announced a federated data infrastructure called "Gaia-x", a legal-cum-software layer to implement granular national data policy, that would allow firms to move data and computing workloads between rival clouds more easily.
   iv. Some Western countries may soon discuss a "Data Freedom Act" which would create a new regulated entity for that purpose[77]. The ideas discussed in the proposed Act include[78]:
      - Data about people is not just a personal asset, because many parties have shared, overlapping legitimate interests in it.
      - Enhancing individual user's powers to negotiate over their data may in fact handover greater control to bigger companies. Increased collective bargaining power should be the basis of any new data policy.
      - Stronger privacy laws are only a first step, but not enough. 'Financial interests' (economic value of the data pertaining to an individual / community) and 'Control interests' (purposes for which the data of the individual / community may be used) exist beyond privacy interests.
   v. Several jurisdictions such as the EU and US have already initiated investigations into the impact of virtual data monopolies on competition in the market. For example, recognizing these very imbalances, the German Competition Law was

73 https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf
74 https://ec.europa.eu/info/sites/info/files/communication-shaping-europes-digital-future-feb2020_en_3.pdf
75 https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
76 https://www.economist.com/special-report/2020/02/20/governments-are-erecting-borders-for-data
77 https://www.economist.com/special-report/2020/02/20/who-will-benefit-most-from-the-data-economy
78 https://issuu.com/radicalxchange/docs/data_legislation_paper_--_20191031

amended in 2019, empowering the German Bundeskartellamt with wider powers of monitoring and enforcing competition rules in the Digital Economy. These include amendments, that bring into the ambit of the German Competition Law, non-price offerings (such as search engines). In particular, the German law now clarifies that transactions where no monetary consideration is paid also constitute a market and can fall within the scope of competition law. Moreover, aspects that are critical for the market power of platforms and networks (such as network effects and access to data) have been introduced into the law as new criteria for market power.

2. Other countries have put in place systems and mechanisms for data sharing. An example is the Japan's Certification System for data-sharing platforms that support companies that want to share their data (on energy, industrial machine and logistics to solve social problems like accident prevention, energy management etc.).
   i. This system includes a data request system, i.e. a system that allows data-sharing companies to request data that have been provided to relevant ministries and agencies.
   ii. The government also provides support through tax incentives and administrative guidance. It can also revoke accreditation in some cases.

3. Another example is that of the Government of Victoria in Australia, which has put in place a Data Exchange Framework for Government and third party data exchange[79]. The data exchange model consists of the following steps:
   i. Manage data requests, assess readiness and authority to exchange – ensuring the exchange (or transfer) happens in a secure, transparent and compliant manner and sufficiently describing the data and its quality to enable the data recipient to assess fitness for their intended purpose.
   ii. Apply business rules to ensure reliable, consistent and sustainable data exchange and decision making
   iii. Identity mechanisms and tools – which tools and templates to use will support streamlined, safe and authorised data exchange
   iv. Exchange data

4. Finland's (2018) Act on Transport Services through deregulation gives more room to develop innovative, digitally enabled services. It obliges all service providers to open certain essential data to all as well as ticketing and payment APIs for single

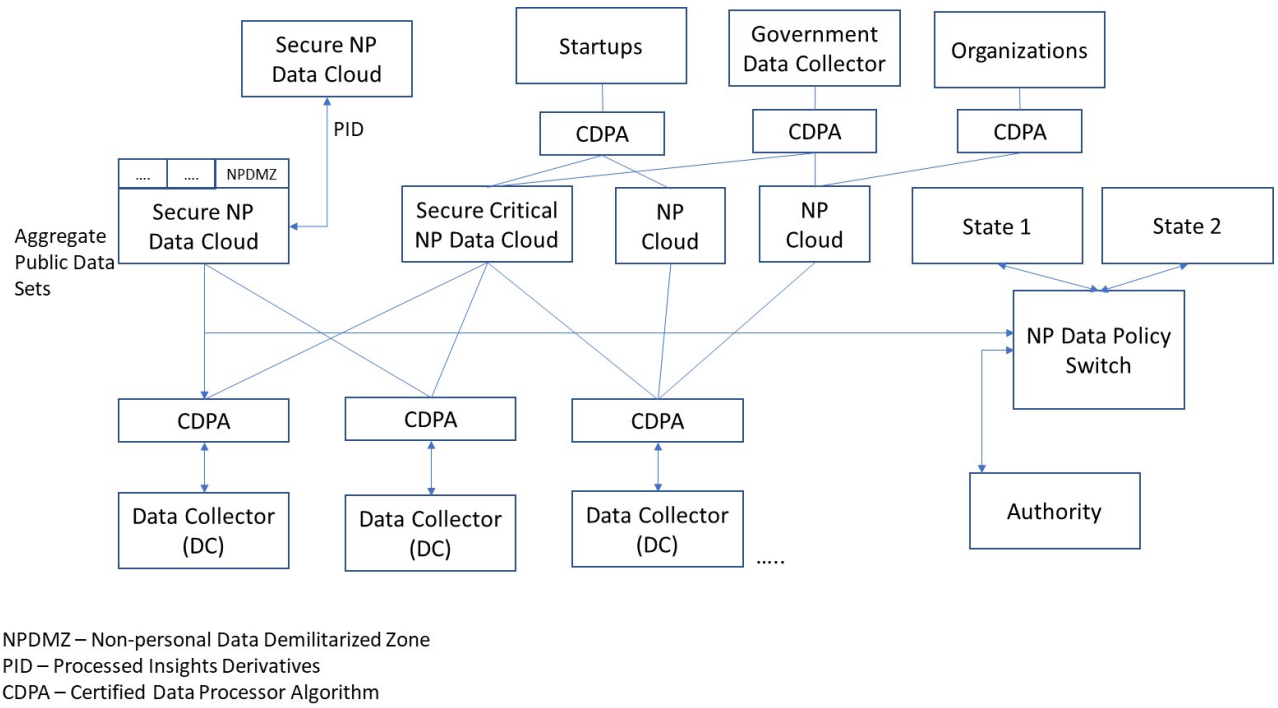
79https://www.vic.gov.au/sites/default/files/2019-07/Data-Exchange-Framework_0.pdf

trip/ticket to third parties. The Act makes it possible to examine transport as a whole, i.e. as one service[80].

80https://www.oecd-ilibrary.org/sites/276aaca8-en/1/2/5/index.html?itemId=/content/publication/276aaca8-en&_csp_=a1e9fa54d39998ecc1d83f19b8b0fc34&itemIGO=oecd&itemContentType=book

Appendix 7: Illustrative Technology Architecture for Data Sharing

The Committee presents an illustrative three-tiered system architecture spanning legal safeguards, technology and compliance to enable data sharing.



NPDMZ – Non-personal Data Demilitarized Zone
PID – Processed Insights Derivatives
CDPA – Certified Data Processor Algorithm

Figure: Architecture of different stakeholders, data and control flow.

1. A technology architecture that enables data sharing.
   i. Data Business / Data trustees may implement this architecture when they are faced with a data request.
   ii. Best of breed Differential Privacy algorithms [Refer to 'Appendix 3: Primer on Anonymity' in this report for different algorithms] are used to create anonymised data to best effort by the Data Custodians and in compliance with rules set by the Non-Personal Data Authority.
      - The Non-Personal Data Authority will specify the minimum threshold of anonymity and base it on acceptable standards.
      - These best of breed Differential Privacy algorithms should be jointly evolved by Indian academia and industry, continuously improved using a combination of global open source improvements and with funding to Indian research organisations.

- These algorithms along with their open-source implementations are made available to Indian organizations along with minimum recommendations for each major type of data.
- These recommendations may be cemented and continuously evolved by leading technical experts using an open standards-based IETF process, perhaps making these global standards as well through IEEE and WWW.

iii. The data sets so anonymised are then submitted or when real-time, streamed into a new construct called "Secure Non-Personal Data Clouds" (see details below).

iv. Due to the risk of this data being deanonymized, we cannot let raw / factual user data even after differential privacy clean-up into the public. The post Differential Privacy-Clean data is then made available to the Indian community in two forms:

- Aggregated data: Again, using standards and algorithms evolved by academia and Indian industry, aggregated data either aggregated to sufficient levels post Differential Privacy-Clean or raw / factual is made available as public data sets to all Indian organisations.
- Raw / factual post Differential Privacy-Clean data is made available in the confines of the Secure Non-Personal Data Cloud, where the cloud provider provides APIs for registered, verified Indian organizations to submit code packages that run in the cloud generating gross aggregate derivative products like ML models, statistical insights and so on.

2. Policy Switch

i. Each data trustee may want to exercise its authority to govern data deemed in their respective domains. However, the best innovation happens in the boundaries and interconnections between datasets - traditionally separated by such governance functions. This can significantly reduce economic value realisation and stifle innovation if each data trustee creates a separate window of clearance and rules for using data under their regulation. A new approach is suggested to address this aspect – of a digital Non-Personal Data Policy Switch ("Policy Switch") as defined below.

ii. Using the Policy Switch, even though regulations can emerge from various institutions and regulatory bodies, the encoding, rationalisation (to ensure no contradiction), implementation and clearance/ compliance enforcement may be with a single authority - who is subject to the regulatory guidelines issued by various data trustees.

- And since handling data subject to multiple regulatory bodies can get complex exponentially, a way to efficiently and rapidly realise economic benefit and large scale public good of Non-Personal Data without

sacrificing regulatory granularity or diluting individual authorities is to bring these together digitally.

iii. The central idea of the Policy Switch is a single digital clearing house for regulatory management of Non-Personal Data. The Policy Switch is defined by a set of APIs and a Policy Markup Language spanning all aspects of managing Non-Personal Data publicly and privately. The Policy Markup Language encodes all interactions and transactions relevant to Non-Personal Data spanning:
    - Policies: e.g. access rules, anonymisation standards, aggregation standards, business rules, security standards
    - Adjudication workflows: e.g. verification, exception adjudication, certification
    - Compliance: e.g. registration, compliance submissions, that are applicable to Non-Personal Data such that Non-Personal Data Custodians, both public and private, only have to interface with and comply with the Policy Switch digitally, no matter the types or sources of data with which they are engaged.
    - To reduce the burden on various governance authorities, the Non-Personal Data Authority will create a base set of minimum set of policies, workflows and compliance rules that all Non-Personal Data must comply with – mostly to safeguard privacy of people and ensure economic benefit will go to India.
    - In addition, it is recommended that the Non-Personal Data Authority manage a stream of academic research and grand challenges to create reference policies, evolve the markup language and reusable tools to simplify the management of Non-Personal Data by regulators, Data Collectors and Data Processors.
    - A further suggestion is to design this policy markup language to be evolutionary. For example, rules, often stated as principles and guidelines, rarely spell out every corner case. A well-implemented policy switch will continuously capture corner cases that emerge via built-in adjudication workflows and after verification, update the marked-up policies so that corner cases are captured in definitions as whitelists or blacklists; and as conditional exceptions in the rule hierarchy.