भारतीय विज्ञान संस्थान

Smart City
MISSION TRANSFORM-NATION

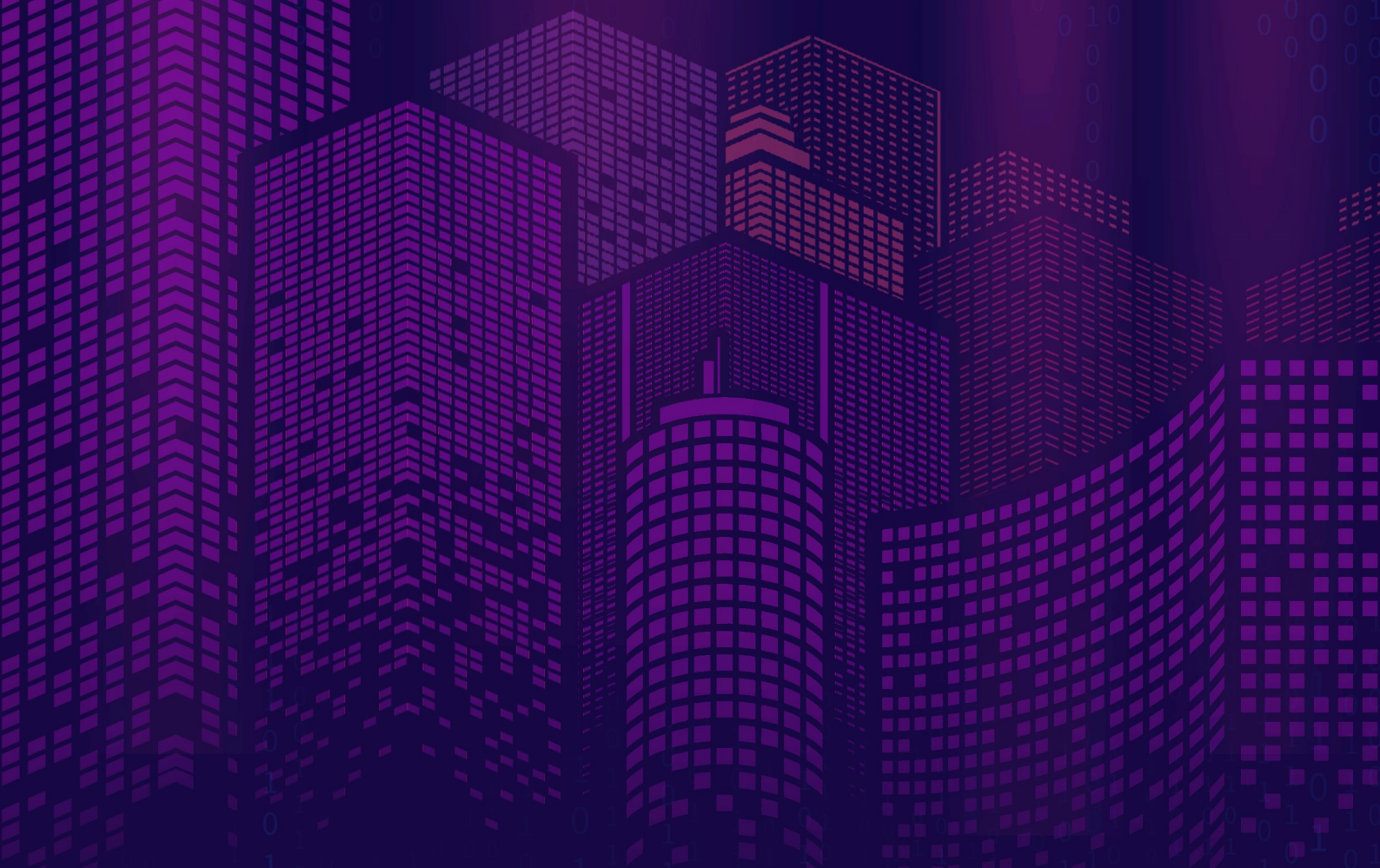
सत्यमेव जयते
Ministry of Housing and Urban Affairs
Government of India

आत्मनिर्भर भारत

स्वच्छ भारत
एक कदम स्वच्छता की ओर

IUDX

INDIA URBAN DATA EXCHANGE



High-Value Datasets

Abbreviations

Abbreviation	Definition
AI/ML	Artificial Intelligence / Machine Learning
BRTS	Bus Rapid Transit System
ETA	Estimated Time of Arrival
IMD	India Meteorological Department
RLVD	Red Light Violation Detection
GPS	Global Positioning System
SWM	Solid Waste Management

Scope

This document is intended to apprise the reader about the characteristics of a High-Value Dataset (HVD). The document also covers examples of HVDs across various domains, some of the data valuation frameworks being used across industries, various techniques for monetizing the data, and some example solutions/applications using HVDs.

The document also suggests how to assess the quality and relative score of an HVD. The document will conclude by highlighting the complete journey of an HVD from identification to value-creation and monetization.

This document will help entities define and/or identify the HVDs under their domains, assess the value of those HVDs, and eventually identify the ways to monetize and/or utilize that data for creating value for society, environment, or economy.

Table of Contents

Power of Data

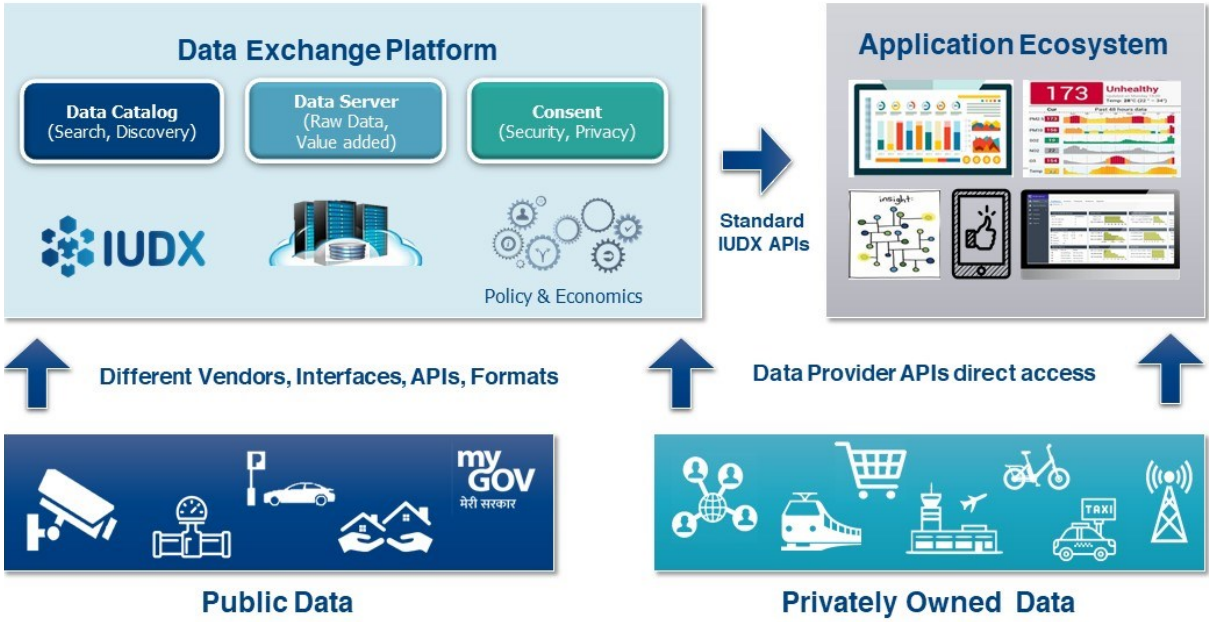
1. The Power of Data

The world has become increasingly digital and the applications in smart cities, townships, healthcare, agriculture, industry, e-commerce, etc are generating good quality electronic data. Data, until now, have mainly been used for deriving information, insights, trends, and managing the services. However, the power of data lies in its combinatorial possibilities when multiple datasets come together and create innovative applications for service delivery efficiency and end-user convenience, making the best use of AI/ML technologies, which makes data the “New Oil” and an economy.

2. India Urban Data Exchange (IUDX)

Data in most cases remains in silos in the respective application domains with different systems representing data in different ways, making sharing of data difficult and complicated. Lack of policy frameworks is also non-conducive for data sharing. Easy and efficient exchange of data among disparate urban data silos through a secure platform and policies to enable data sharing from multiple entities are important to facilitate open innovation.

India Urban Data Exchange (IUDX), initiated and funded by the Ministry of Housing and Urban Affairs (MoHUA) and supported by the Ministry of Electronics and Information Technology (MeitY) and NITI Aayog, is developed and deployed as a fully open-source cloud-based platform to enable easy and secure sharing of all types of data.

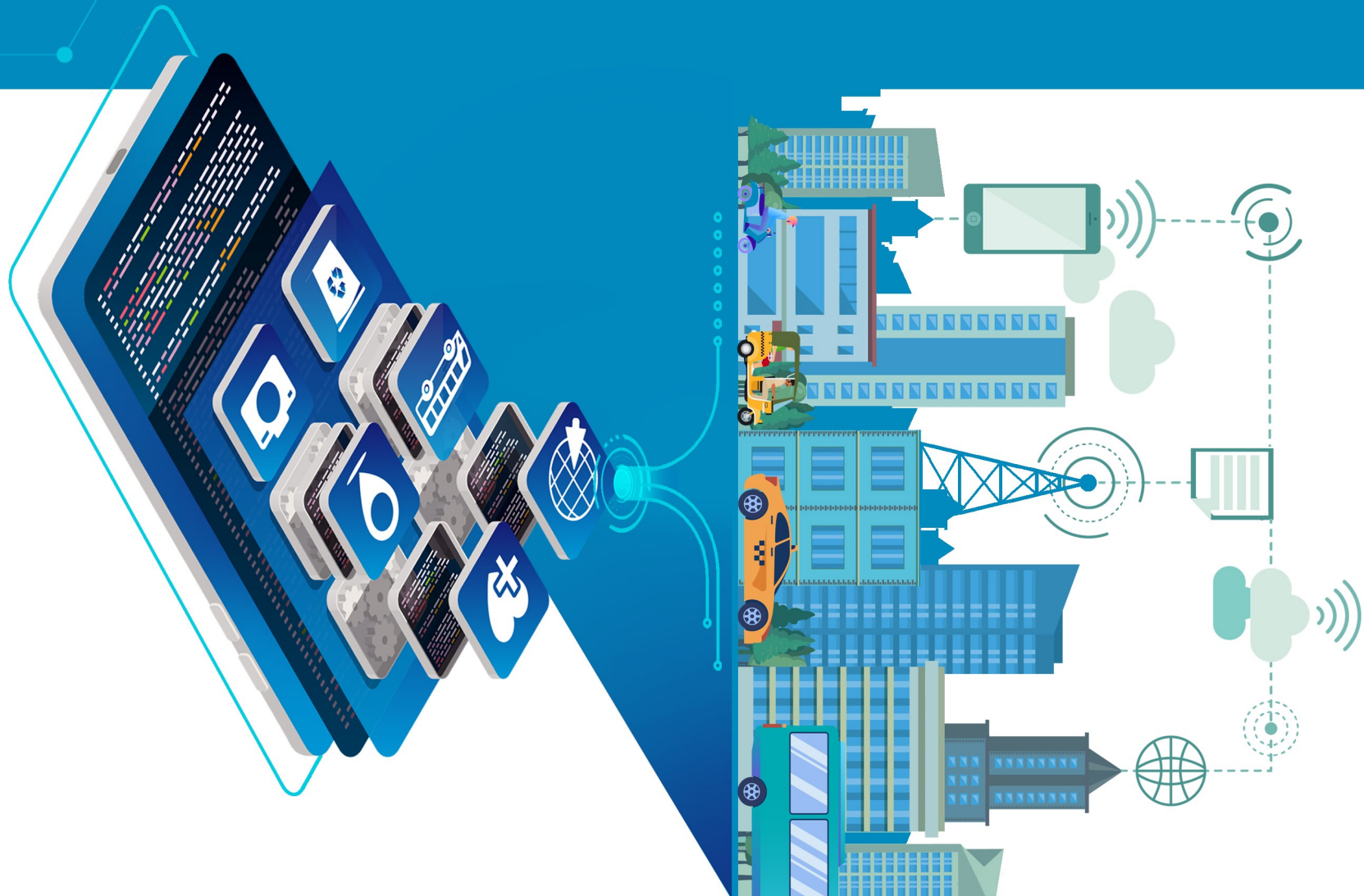


IUDX provides a way for accessing data in a unified, common format and enables data sharing and monetization between different entities. IUDX can be used for enabling data exchange within internal departments as well as external agencies to create innovative applications with new business/ revenue models aka data marketplace. The above figure shows an example of the IUDX ecosystem.

Public and privately-owned datasets of urban governance, mobility, healthcare, and citizen security can be exchanged through IUDX. These datasets are being taken by the industry/start-up ecosystem to build applications for traffic management, public transport, prevention of disease spread and healthcare infrastructure management, emergency assistance, solid waste optimizations, flood warning, citizen safety, etc.

The datasets which can derive important benefits for the society, the environment, and the economy aka High-Value Datasets (HVD), and various ways of valuing and monetizing HVD and the challenges involved in doing so are discussed in detail in the rest of the document.

High-Value Datasets



3. High-Value Datasets - Definition

The Datasets which could be instrumental in deriving important benefits for the society, the environment and the economy, in particular, because of their ability towards creating value-added services for efficiency and convenience by making the best use of AI/ML technologies are considered as High-Value Datasets. A HVD should be able to meet either of the following characteristics:

4. High-Value Datasets - Urban Governance

Based on the definition, some of the common HVDs across various domains along with their potential data owners and the value they can create have been described below:

S. No	Domain	Data Owners & Datasets	Value/Use-Case
1	**Mobility**	**Metro Rail Corporation:** Metro Schedules, Fare, Collection, Station locations, Route Network	• Multimodal Transport Application for citizens to plan their end to end journey.
		Public Bus Operator: Bus/BRTS Schedules, Fare, Collection, Stops, Routes, Live location, Direction	• Multimodal Transport Application • Bus occupancy and ETA information in Bus Application • Fleet optimisation based on commuter demand, convenience, occupancy levels, etc.
		Mobility operator (e-Bike / Taxi / Bicycle): Real-time availability, Fare, Location	• Multimodal Transport Application • Last mile connectivity for Bus & Metro
		Municipal Corporation or Private Parking/Charging operator: Location, Fare & Slot availability at Parking/Charging Infrastructure	• 'One Parking' or 'One Charging' App for the city by aggregating private and public parking/charging facilities.

2	**Traffic**	**Traffic Department:** Traffic density, Traffic violation, Surveillance/RLVD Camera Video feeds	• Make conventional Traffic lights into adaptive • Signal timing optimization for increased vehicle speed • Illegal parking detection • Flood warning • Video Analytics based Incident management Response system
		Private players: Road quality, Road Assets, Road signs	• Effective & efficient Road repairs • Making policies for improving road infrastructure • Autonomous driving
3	**Solid Waste Management**	**Municipal Corporation:** SWM Vehicle Info, Vehicle location, Waste weight, Employee info, Fuel Consumption	• Optimize Solid Waste management routes, pick ups and vehicle/personnel utilization thus creating cost & service delivery efficiency.
4	**Environment**	**Smart City Corporation:** Location of Environment Sensors, Air Quality	• Tackling societal challenges related to environmental pollution. • Air Quality data can indicate areas suitable for setting up factories
		Smart City Corporation: Flood Sensors Data & Location, Water level	• Create Urban flood models to predict and manage floods.
		IMD: Rainfall historical/forecast	
		Development Authorities / Water Resource Department / Water Board, Municipal Corporation: Hydrology datasets - Digital Elevation Maps, Drainage, Dam Discharge, Canals & River datasets	
5	**Water distribution**	**Water Resource department:** Water Quality, Water Pressure, Water Discharge, Water flow, Water level	• Optimize water distribution, Track and prevent waterborne diseases.

6	**Energy & Utilities**	**Electricity Board / Water Board / other Utility Boards:** Utility Distribution & Usage across houses, wards/zones	• Detect Revenue leakage in Utility bills by correlating power/water distributed, billed, type of property (commercial / residential) • Design incentives for reducing Utility wastage. • Can indicate areas where solar panels might be set up as an alternative source of energy.
7	**Healthcare**	**Health Department**: Diseases, Patients, Location, Medication, Recovery	• Healthcare Management Information system to monitor, track and predict diseases and prevent outbreaks and also for planning, enhancement & optimum utilization of health care infrastructure.
8	**Emergency services**	**Smart City Corporation/ Municipal Corporation/ Health Department / Fire Department/ Police Department:** Live-Location of Emergency vehicle (Ambulance, Fire vehicle, Police vehicle), On-duty information of Emergency vehicle	• Green corridor for Emergency vehicles • Incident Management Response system for emergencies.
9	**Tourism**	**State Tourism Development Corp:** Tourist attractions, Timings, Entry fee, Special markets, Stay options	• Tourist App for providing information, navigation, recommendation & payment option to tourists.
10	**Revenue Collection**	**Smart City Corporation / Revenue Department / Municipal Corporation / Development Authority:** Revenue generated from Property Tax, Professional Tax, Trade license, Vehicle registration, Facility booking	• Tax Revenue leakage detection correlating with type of property (commercial/ residential) • Utility revenue leakage detection correlating with power/water billed • Devising policies or solutions aimed at Revenue maximization.

11	**Citizen Grievances**	**Smart City Corporation:** Grievances regarding lack of Cleanliness, Garbage issues, Overflow of drains, Malfunctioning of amenities	• Solid Waste optimization • Devise efficient response mechanism for Grievances • Enhance public engagement
12	**Smart Elements**	**Smart City Corporation:** Location of Smart Elements like Public Addressing System, Emergency Call Box, Variable Messaging Display, WiFi Hotspot, Smart Kiosk, Public Toilet, Shelter Home	• Disaster or Incident management Response system • Improve citizen engagement • Safety Index of place • Tourist App or City App • Designing Walk paths
13	**Streetlights**	**Smart City Corporation/ Municipal Corporation:** Locations, Energy consumption, Feeder distribution	• Optimize energy consumption by switching off/dimming the lights based on people on the road. • Enhance Use Cases like Safety Index of places.
14	**Citizen Security**	**Law & order, Citizen security department:** Video Surveillance Feeds and analytics (Crime, Crowd)	• Illegal parking • Flood warning • Video Analytics based Incident management Response system
		Private players: Safety Index of places	• Safe Routes for travel • Tackling societal challenge of citizen safety
15	**GIS**	**Municipal Corporation:** City assets on a map	• GIS datasets will be useful for devising various Use Cases such as Revenue leakage detection, Flood Monitoring.

Similar datasets from other domains like agriculture, education, finance and contracts, crime and justice, earth observation, global development, government accountability, science and research, statistics, social welfare, etc. that are mentioned in the table below are also important and useful to create value for public good.

S. No	Data Category	Example High-Value Datasets
1	**Agriculture**	Crops, Harvest time, Yield, Storage, Market pricing
2	**Crime and Justice**	Crime statistics, Safety scores
3	**Earth observation**	Meteorological/weather, Forestry, Fishing
4	**Education**	List of schools, Performance of schools, Digital skills
5	**Finance & contracts**	Tenders, Local, State & National budget
6	**Government**	Election results, Salaries (pay scales) National Statistics, Census, Infrastructure, Wealth, Skills
7	**Science & Research**	Genome data, Research educational activity, Experiment result
8	**Social welfare**	Housing, Health insurance, Food security, Unemployment benefits

Beyond identifying HVDs, it is also important to identify the value of these HVDs. The next section discusses various data valuation frameworks being used across the industries to identify the value of the datasets.

Data Valuation Frameworks

5. Data Valuation Frameworks

There is no exact formula for assessing or placing a value on data. However, below are some of the data valuation frameworks, widely used across industries and/or geographies:

a. Treating data as an asset

In this approach, data is treated as an asset and monetized directly by trading it or by building a service around it and selling that service. The value of such datasets is largely market-driven. A case in point is Bloomberg terminal's real-time market & financial data which it provides as a service & monetizes by charging a subscription fee for the service.

b. Contribution based value of data

This approach involves valuing the data based on the number of users contributing to it. For example, for any navigation application, the more the number of users contributing to it, the more accurate and valuable it will be. Therefore, the value for such datasets will be proportional to the number of users contributing to it.

c. Prudent value of data

This approach values the data based on its potential to drive key initiatives aimed at cost reduction and/or revenue enhancement. For example, data like 'Bus occupancy and ETA' will increase efficiency for bus operations and is expected to increase the revenue from bus operations by 1-2%. For such datasets, the value can be set by converting expected future benefits (cash flows or earnings) into a single discounted monetary value.

d. Market value of data

In this approach, data can be competitively priced based on the value at which identical datasets are being valued in the market. This approach can be used when any identical data has been involved in an observable market-based transaction. It should also factor in a price multiple for any significant time difference from the observable transaction.

This approach considers how much it would cost to replace or develop the same data. This approach ignores the amount, timing, and duration of future economic benefits. It involves assigning a value based on either the historical cost of developing the same data or by estimating the current cost for producing identical data.

e. Cost-based value of data

This approach considers how much it would cost to replace or develop the same data. This approach ignores the amount, timing, and duration of future economic benefits. It involves assigning a value based on either the historical cost of developing the same data or by estimating the current cost for producing identical data.

f. Download statistics based value of data

This approach uses the download statistics of a data to determine its need & hence value, however, it can only be used when data is listed on a data exchange/sharing platform and where the download statistics are being monitored. Though this approach captures the demand for the data, however, it does not necessarily reflect the potential impact a dataset can have and hence should be used when other valuation methods aren't possible.

g. Game-theory based value of data

This approach involves valuing the data based on the number of players in the market who can supply similar data. A data owner can charge a reasonable premium for a unique data of which there are no other players in the market. However, in case of multiple players having similar data, the value for the data can be set using the preferred game theory strategy based on the competitive or cooperative landscape between the players.

h. Relative value of data

This approach can be used for analyzing where a particular dataset ranks in comparison to others by measuring it against a set of impact-based characteristics and by assessing its quality and volume. The relative rank can be used either to prioritize a dataset from a group or to assign it a value by multiplying the relative rank with a monetary weight.

Now, based on the nature of the data and also based on the availability or ease of estimation of the required information for respective frameworks, an entity can choose the most relevant framework to assess the value of its data. The above-mentioned frameworks provide a good starting point for valuation but its effectiveness will also depend on how accurate are the data owner's assumptions & calculations.

Business Models for Data Monetization

6. Business Models for Data Monetization

There are various business models for Data monetization and depending on the type/nature of data the relevant business model can be selected. The selection of a business model will also depend on how valuable the data is for one's business in a strategic context. Some of the business models for data monetization are listed below:

a. Direct data sale

In this model, a data provider can directly sell/provide his data to any data consumer in exchange for money. The data provider needs to decide whether they would like to share the raw data, aggregate data or insights derived. The data provider then has to finalize whether to share data through its own channel or through a data exchange platform like IUDX. The price will be market-driven in the long run, however, for setting up the price initially, the relevant data valuation framework can be used.

An example of Direct Data Sale can be a city corporation selling the city GIS data by setting a fixed price for it. Data consumers can buy this data by paying a fixed price and can then use it for any purpose.

b. Value based pricing

In this model, the data is priced based on the value-created using the data. The value created will be in the form of enhanced service delivery or end user convenience.

Service delivery efficiency is expected to reduce the cost. For example, a Solid Waste Management (SWM) company can effectively utilize SWM data to achieve 25-30% reduction in SWM operational expenditure by optimising routes, pick-ups, vehicle & personnel utilization.

The increased end user convenience normally results in increased revenue through increase in customer base. For example, a Multimodal App or providing information like Bus Occupancy with ETA to the commuters can increase customer satisfaction resulting in an increased ridership & thus an increase in revenue of 1-2%.

c. Additional revenue options

The data can also be used for creating new businesses or revenue streams as given below:

- Context aware advertisements in the applications, public displays
- Subscription/usage based fee from the users
- Direct monetization of data being generated from new service/ application

Use Cases Possible with High-Value Datasets

7. Use Cases possible with HVDs

The power of data lies in its combinatorial possibilities when multiple HVDs come together to create an AI/ML driven application/solution that opens possibilities for new business/ revenue models. Some example use cases in the area of urban governance are as follows:

a. Solid Waste Pick-up & Route Optimization

SWM optimization solution will use GPS location of waste collection vehicles, weight of collected waste, citizen grievances regarding waste and other SWM related data for efficient routing, pick-up & vehicle/personnel utilization. Apart from citizens' satisfaction, a clean city, and reduced traffic/pollution, it will also help in reducing the city's SWM operational expenditure by 25-30%. For example, for a city like Varanasi having annual SWM expenditure of around Rs. 30 Cr, it can result in a cost-saving of Rs. 7.5 to 9 Cr.

b. Multimodal Transport Application

Multimodal Transport Application will consume data from multiple transport services, walk/cycle paths, safety index of places and suggest all possible travel options based on commuters' preferences. Beyond benefits like reduced congestion, pollution, carbon footprint and safe travel, such added convenience for citizens could increase revenue of transport services by 1.5 to 2%. Apart from cost savings, data providers can also ask application partners to share revenue, once they start monetizing the App by charging commuters a fee and/or from advertisements and/or by monetizing App data that have travel patterns.

c. Green Corridor for Emergency Vehicles

Green Corridor Application will consume the real-time location of on-duty emergency vehicles and interact with traffic signals to establish a congestion-free movement of emergency vehicles at traffic junctions. This is expected to reduce the deaths due to delay in timely delivery of emergency services by 30%. This application not only reduces the travel time of emergency vehicles, it also reduces manual efforts to facilitate this movement and traffic jam ripples.

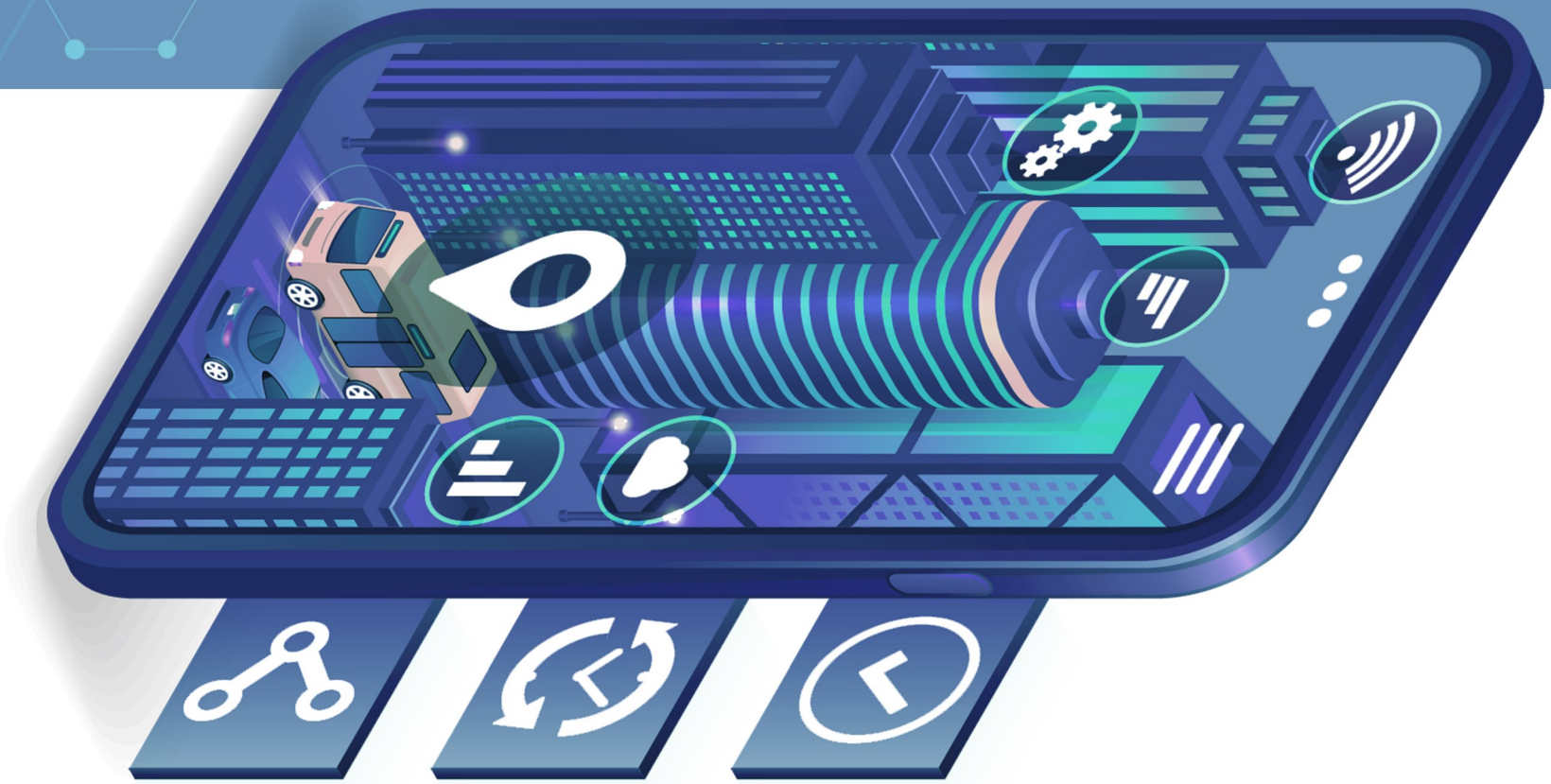
d. Flood Analytics and Management System

This use case aims to integrate the hydrology datasets and sensor measurements for rain, water level and stream flow to create AI/ML enabled urban flood models to predict floods and water levels. The prediction can be converted to action items for city administration, disaster management authorities, and citizens. Over the last 6 decades, floods on average have resulted into a monetary loss of Rs. 8000 Cr/ annum. Predicting floods in advance could reduce these losses significantly in addition to saving lives. The data about the flood-prone areas and the patterns could also be monetized in the real estate business.

e. Revenue Leakage Detection

This use case suggests a data-driven approach to spot revenue leakage by correlating data like property classification (commercial/residential), tax, trade licenses, utility bills and GIS based land-use mapping for detecting anomalies & Revenue leakage and eventually making recommendations to fix it. This can also help in highlighting critical loopholes in the current property tax collection system. The city stakeholders can either share a part of revenue saved or a fixed fee with the industry partner developing this solution.

Quality of Datasets



8. Quality of Datasets

The Use Cases described in the previous section reflects on the potential benefits of using a HVD. However, to materialize on the potential benefits it is essential that the HVD is of good quality. The quality of a HVD can be assessed through IUDX's 3-C Data Quality Assessment Model. The 3-C's of IUDX data quality assessment are described below:

Fig: 3-C Data Quality Assessment Model

a. Complete

The completeness refers to the comprehensiveness of data, i.e., the data should contain information for all the expected parameters relevant to that dataset and there is no missing information. For example, the live locations of a bus may not individually be of much value, but when it establishes association with other transit parameters like Trip_ID, Route_ID etc. it becomes a comprehensive transit dataset.

b. Consistent

Data can be referred to as consistent if it arrives as per the defined interval and measurement units. For example, for transit data the location of a vehicle should arrive every 10 seconds and the speed should be represented in km/h. The consistency of a data can be challenging when the data is aggregated from numerous sources. For example, aggregating transit data from multiple vendors for bus unit tracking can be challenging if they are publishing data at different intervals and/or in different units. In such cases, it is important to iron out differences to maintain uniformity & make data consistent.

c. Consumable

Data is referred to as consumable if it is available in an electronically sharable format. Consumability refers to the ease by which data of variable size and formats is allowed to be consumed. For example, in case of the transit data even excel sheets with the static locations of bus stops can be made available as a programmatic pull to ensure data integrity. Consumability is important because if the data availability is dependent on any manual intervention then we cannot build any Mission critical application over such data.

Therefore, it is important that the HVDs should also be of good quality, i.e, it should be able to pass the '3-C Data Quality Assessment' check, for it to be really useful for data-driven use cases.

Framework for Relative Scoring of an HVD

9. Framework for relative scoring of an HVD

This section will provide a framework to identify the relative scope of an HVD. To devise a framework, we have considered 3 important parameters - Characteristics of HVD, Volume of HVD, and Quality of HVD. The process of classifying an HVD as High or Low across these parameters is as described below:

Characteristics of an HVD:

An HVD can be classified under High Characteristics category if it meets 2 or more characteristics out of the 5-point characteristics of HVD given below:

1. Facilitate the creation of new services or help in cost reduction and/or revenue enhancement for existing services and/or improve service delivery access, efficiency, and quality of existing services.
2. Assist in tackling societal challenges such as climate change, health, poverty, mobility, financial exclusion, skill-gap, etc .
3. Facilitate the creation of new businesses, revenue streams, and/or new jobs .
4. Improves transparency, accountability, and openness of government/organizations towards the people.
5. Useful for policy making, devising public programs, and/or improving citizen engagement.

For example, a SWM dataset will be classified under the high category as not only it helps in reducing SWM operational cost but also tackles societal challenge of Waste management .

Volume of an HVD:

The number of data points in an HVD can be indicative of its volume. For example, a transit dataset having information of over 1000 buses can be categorised as high volume and a transit dataset having information of 100 buses can be categorised as low volume.

Quality of an HVD:

The quality of a dataset can be identified using the IUDX's 3-C Data Quality Assessment Model. The datasets meeting all the 3 Quality parameters can be considered as high-quality HVDs while others can be classified as low-quality HVDs.

Now, once an HVD's category across all 3 parameters (Characteristics, Volume & Quality) is identified, one can then use the below framework to estimate HVD's relative score:

S.No.	Characteristics Category	Volume Category	Quality Category	Relative Score
1	High	High	High	8
2	Low	High	High	7
3	High	Low	High	6
4	Low	Low	High	5
5	High	High	Low	4
6	Low	High	Low	3
7	High	Low	Low	2
8	Low	Low	Low	1

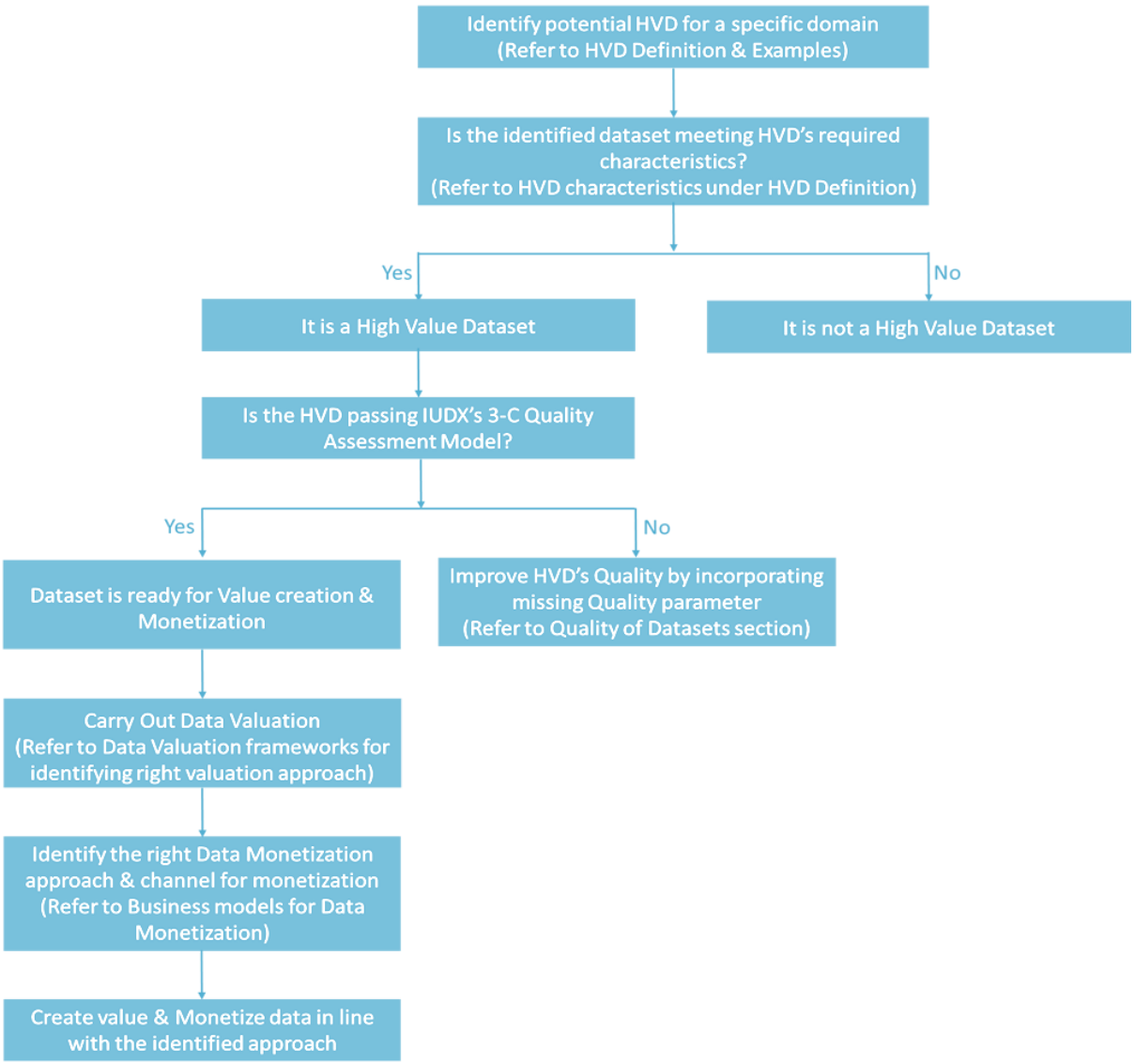
Based on the categorical value (High or Low) of Characteristics Score, Volume and Quality, an HVD can be assigned a relative score between 1 & 8. The more the score, the better the HVD & the better it's scope for monetization. The above framework also shows that a low score on quality has the most detrimental effect on an HVD and at the same time it also presents an opportunity for HVD's owner to improve its quality for improving its scope for monetization. Depending on the missing quality parameter(s), the quality can be improved by incorporating missing resource(s), if any, or by publishing the data at a defined interval and measurement units or by making it available in an electronically sharable format.

Though all HVDs can create significant benefits for society, understanding its relative score will further help in knowing where the HVD stands and what needs to be done to improve its scope of monetization.

Summarizing HVD Exercise

10. Summarizing HVD Exercise

HVD exercise refers to the complete process of identifying the HVDs, carrying out its valuation, assessing its quality, creating value, identifying the right monetization approach, and eventually monetizing it. The below pipeline provides a high-level overview of the complete HVD exercise:



The above pipeline, apart from describing the key stages of an HVD exercise, also suggests how issues such as lack of clarity around HVD, its quality and the ways to value it and build a solution over it and/or ways to monetize it can be effectively tackled by referring to various sections of this document.

It's time for the Public/Private organizations to start working on the HVD exercise, so as to identify HVDs in the domain of their interest and eventually work towards maximizing its potential by building an application/solution over it and/or monetizing it.