Project 2 Report :

Data Visualization for Mumbai

Akshay Kore
Interaction Design
M.Des (2014-16)

Guide: **Prof. Venkatesh Rajamanickam**

Industrial Design Centre
Indian Institute of Technology, Bombay

Approval Sheet

The Interaction Design Project II entitled “Data Visualization for Mumbai” by Akshay Kore, Roll Number 146330007 is approved, in partial fulfillment of the Master in Design Degree in Interaction Design at the Industrial Design Centre, Indian Institute of Technology Bombay

Internal :

External :

Guide :

Chairperson :

Declaration

I declare that this written document represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/ source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Akshay Kore
146330007

Industrial Design Centre,
Indian Institute of Technology Bombay

November 2016

Acknowledgements

I would like to thank Prof. Venkatesh Rajamanickam for his guidance and support through the tenure of this project. I am grateful to Prof. Anirudha Joshi, Prof. Girish Dalvi, Prof. Ravi Poovaiah, Prof. Uday Athavankar at IDC for their valuable feedback and suggestions. I would also like to thank Prof. Arnab Jana from C-USE IIT Bombay for his feedback during the initial phases of the project. I am grateful to the staff and authorities at the Municipal Corporation of Greater Mumbai, The Urban Design Research Institute in Mumbai, MMRDA, MIDC Mumbai and the Development Commissioner's Office, SEEPZ SEZ for their help and cooperation.

Feedback and guidance from Ar. Namrata Kapoor, Ar. Anuj Gudekar, Ar. Mugdha Deshpande, Ar. Mridula Pillai, Ar. Shreyas Khemalapure, Ar. Maitreyi Gulawani, Mr. Ramesh Palan Dr. Ms. Aparajita Bakshi and Prof. Himanshu Burte was of paramount importance. Most importantly I would like to thank Shweta Deo, Ruchita Chandsarkar, Indrajeet Roy, Prasad Ghone, Jayati Bandyopadhyay, Dileep Mohanan, Sitara Shah and Sagar Yende for their help, motivation and support.

Thank you to The Library. The Internet. Quora.

Abstract

Our governments create and release vast quantities of data about us and our surroundings by means of the census, budgets, human development reports, lan use studies , etc. This data is released in the public domain and is regularly accessed and used by policy and decision makers in both the public and private domain.

Visualization of enables the analysis of data and especially when the data is available in vas quantities. These decision and policy makers make use of data visualizations regularly. Visualization of data done by these individuals or groups is most often done with a purpose in mind. However, insights can come from anywhere, in seemingly disparate correlations of two or more datasets. Access to these datasets is often difficult due to the scattered nature of our government bodies and even after the access is gained, the datasets come in different tangible and intangible file formats making direct use of these datasets for visualization a difficult and often a laborious task.

With the city of Mumbai as an example, this project aims to enable an easy way to visualize the city's spatial data. This is done by means of a web based application called the MumbaiData Visualization Tool. The higher goal is to enable faster and better decision making by means of data visualization and set an example for all city governments in the country.

The project is open to use freely and is made available at http://www.mumbaidata.in/

Table of Contents

1. Introduction

The City of Mumbai has been a muse for writers, businessmen, economists, planners, filmmakers, artists for long time and the many of us have seen the city through the eyes of these 'experts'. Be it the prose and poetry, or the number crunching of the stock market, Bollywood and Hollywood movies or the paintings at art galleries or the ones on the compound walls.

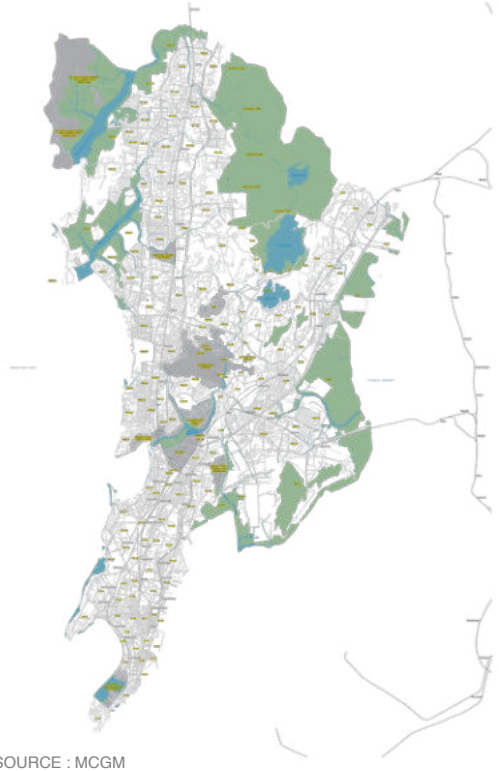
SOURCE : WIKI IMAGES

"The thing about Mumbai is you go five yards and all of human existence is revealed."
Julian Sands

We live in times where there is a ton of information available on the internet and off it. The problem we face now is not of an overload of information but that of disinformation. How do we make sense of what is going on in our cities, our country or even the world ? What do we base our decisions, our acts of policy making on? Whom do we believe in? What questions do we ask?

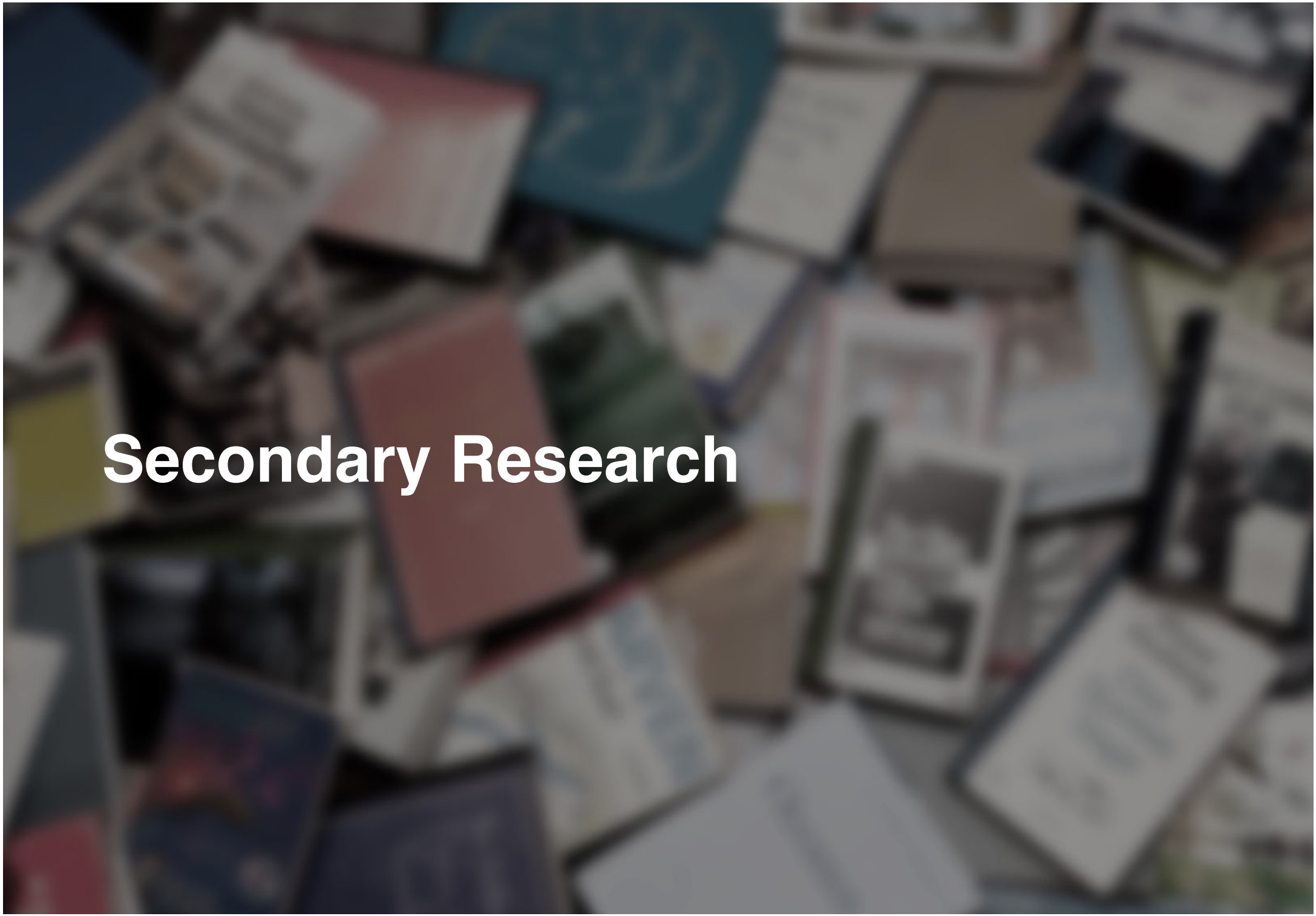
This begets a solution for a different way of looking at our cities. Not based on words and pictures alone. Not on moving images but an almost unbiased way of presenting the information.

Much of data of the city is available, but it is in a scattered format among different private and government bodies, NGOs, people, books , etc. It makes sense to organize this information, overlap datasets and find correlations. Thereby seeing the city in a different light and generate actionable insights.



SOURCE : MCGM

NOT TO SCALE


Secondary Research

2. Secondary Research

To attempt to solve to problem to visualizing vast quantities of data, it is important to look at what this data is and might be. The secondary research examines the concept of Open Data which is one one of the main types of data that the project attempts to use. The concept of Big Data has been explored since data about the city and its people is a large quantity of data and a large number of data points have been measured. Finally the data about the city has been looked into to imbibe a familiarity with the kinds of datasets currently available.

2.1. Open Data

Open data is the idea that certain data should be freely available for everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control[1]. The notion of Open Data and especially Open Government Data has been around for some years. In 2009 open data started to become visible in the mainstream, with various governments (such as the USA, UK, Canada and New Zealand) announcing new initiatives towards opening up their public information[2]. The Indian Government has also released some of its data as a part of the Open Government Data initiative on the website : https://data.gov.in/ . The data is available in a table format as well as APIs for some datasets.

"Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike."[3]

This large volume of data unlocks a new potential of using bureaucratic and other datasets and information to enable new policies, services; thereby improving the lives of citizen, making the government and society work better.

Increase in efficiency of processors and declining costs of storing data leads to a tremendous amount of resources to capture and store data. This has led to vast quantities of data about individuals, companies, climate, economy, health , etc. Every type of data that can be quantified, is being quantified. However, the potential of this vast dataset is largely untapped. This is mainly due to lack of access and a dearth of innovative ways to merge, overlap these seemingly disparate datasets to uncover what the data says, to find meaningful insights.

"The nature of innovation is that developments often comes from unlikely places."
Open Data Handbook

The value of these large datasets lies in how they can be reused, recombined in multiple ways. This can influence many areas of public and private life.

[1] http://en.wikipedia.org/wiki/Open_data

[2] http://opendatahandbook.org/guide/en/introduction/

[3] http://opendefinition.org/

Some of these areas include[1]:

- Transparency and democratic control
- Participation
- Self-empowerment
- Improved or new private products and services
- Innovation
- Improved efficiency of government services
- Improved effectiveness of government services
- Impact measurement of policies
- New knowledge from combined data sources and patterns in large data volumes

Governments in that sense, collect large quantities of data by means of the census, budgets, livelihood, etc. They are significant due to the quantity and the centrality of the data and it is binding on the government to provide this data to its citizen. Most of this data is public data and could be made open for individuals or institutions to freely analyze and generate new knowledge. There are numerous examples of individuals using Open Government Data in clever ways, thereby improving the lives of fellow citizens.

Examples

Open government data can also help you to make better decisions in your own life, or enable you to be more active in society. A woman in Denmark built findtoilet.dk, which showed all the Danish public toilets, so that people she knew with bladder problems can now trust themselves to go out more again. In the Netherlands a service, vervuilingsalarm.nl, is available which warns you with a message if the air-quality in your vicinity is going to reach a self-defined threshold tomorrow. In New York you can easily find out where you can walk your dog, as well as find other people who use the same parks. Services like 'mapumental' in the UK and 'mapnificent' in Germany allow you to find places to live, taking into account the duration of your commute to work, housing prices, and how beautiful an area is. All these examples use open government data.[2]

SOURCE : FINDTOILET.DK

Openness of Data

Openness of data means that individuals and institutions are free to use the data in whichever means they please without restrictions from copyright, patents or other mechanisms. The data provided should be accessible cheaply or for free through downloads over the internet.

This easy access is beneficial for others to reuse in novel ways.

[1] http://opendatahandbook.org/

[2] http://opendatahandbook.org/guide/en/why-open-data/

Here are a few policies in the openness of data that will be of great benefit[1]:

- **Keep it simple** : The data provided should be easily readable by a machine. So no PDFs and MS Word files, as it is difficult to parse these to access the data.
- **Move fast** : It is important to have a constant inflow of data and to digitize what is already available.
- **Be pragmatic** : It is impossible to have accurate data at all times and good enough is good enough. More data trumps accurate data. In particular it is better to give out raw data now than perfect data in six months' time. Note that the data available should not be incorrect and should only give room for inaccuracy.

Providing easy access to its data is of value to Government itself. If the data is already available, people will ask less questions and time and effort is saved in answering queries of each person facing a problem or requesting an access through RTI and the likes. This increases the efficiency of the Government's working and reduces costs associated with answering queries and completing these requests.
It must also be considered that not all of government's data can be made public and there will always be exceptions to providing data related to national security and the likes.

There are endless possibilities of using and reusing Open Data and overlapping different datasets. Like how Dr.John Snow used the map of England and location of cholera victims to find the cause of the outbreak at the dawn of the 19th century. This led to discovering that cholera was a waterborne disease. There are numerous opportunities now, more than ever to use data since data is so abundantly available.

Unexpected insights can be gained from combinations of different datasets and maybe the next big discovery or invention will probably come out of these methods.

[1] http://opendatahandbook.org/

2.2. Big Data

We live in a world with an abundance of data. Every aspect of our lives including our identity, our location and our relationships with other people, governments, companies, etc are being datafied. Datafication means conversion of information in a useful format either as numbers, visuals, etc. Our Governments collect and store vast amounts of data about us and our fellow citizens. They are vast storehouses of data. Data can be defined as something that allows it to be recorded, analyzed and reorganized[1].

Big Data can be defined as extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions[2]. In that sense, problems of population, education, health, transportation, governance are Big Data problems since they deal with a vast quantities of data. And as such they must be dealt with that mindset.

N = all

In many traditional institutions, for analyzing situations and data, random samples of data are used since it was difficult, time consuming and expensive to analyze such large datasets. With the increase in processing power and cheaper storage, it has become far easier to use in many cases, all of the data that is available.

Random sampling reduces Big Data problems to small data problems. However, this is the second best alternative. In the world of Big Data, the more data you have the better. Simple models of processing data accompanied by a lot of data are better than elaborate models with only samples of data. This data includes all the data along with the outliers. It is the job of the person who analyses the data to make note of these outliers and the frequency of their occurrences. The people handling Big Data problems must remember that all the data is better than samples of data.

Messiness

It is very rare that a particular data set, moreover a Big Data set would be completely accurate. As users of Big Data, we need to forgive the inaccuracies and be comfortable about the fact that large datasets are bound to be inaccurate till a certain extent. We need to embrace the messiness of the world and the accompanying data.

This is contrary to traditional wisdom for want of accurate data to do accurate analysis. Getting completely accurate data is easy and viable for small data problems. While dealing with Big Data problems, it is time consuming to procure completely accurate data and in many cases very expensive. This is besides the fact that often acquiring perfectly accurate data is downright impossible under the current system.

Take the example of population data. It is common knowledge that it is very difficult to know the exact population today, for children are born and people die every minute and it will require systems in place to count the exact population at any given time.

Population of India:

1,210,569,573

Population of India:

1.2 Billion

[1] Big Data: Victor Mayer-Schonberger and Kenneth Cukier

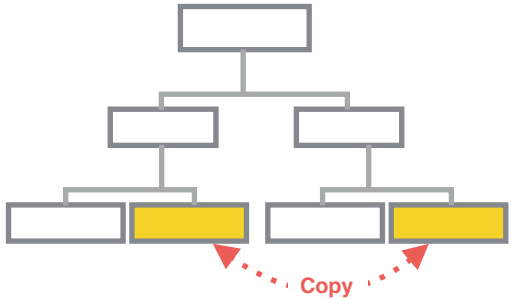
[2] Google definition

This would be difficult and expensive. Even so, we are comfortable with this knowledge and inaccuracy for there is not much difference between saying that India's population is 1,210,569,573 or 1.2 Billion[1]. We are fine with this estimate. As scale increases, showing exact numbers becomes less important.

"The world's we seek to understand are complex and intricate[2]"
Edward Tufte

To solve Big Data problems, we need to accept this inaccuracy. Sacrifice accuracy since it is expensive and time consuming. We need to realize that Big Data is always going to be messy and we need to solve these problems with messiness in mind.

"Treating data as something imperfect and imprecise let's us make superior forecasts[3]."



Hierarchical Systems often tend to create multiple copies of the same data points when handling Big Data

Hierarchy

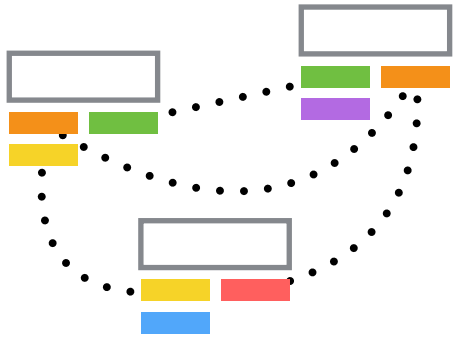
Hierarchical systems work well when the number of data points are small. In case of small data, the categorization of data is simpler. It is possible to SLIP (sort, label, integrate, prioritize)[4] the data.

What happens when data points overlap and when they cannot fit in a particular box ? Do we copy them again in another hierarchy element ? Imagine this one data point having affinity with a hundred different data points. It would be unwise to make a hundred copies.

Hierarchical systems fall apart in the face of Big Data problems. Although it is still possible to categorize, there is a much simpler method based on correlations rather than categorization.

It is much better to use a system of tagging data points rather than categorizing them in a hierarchy. Tagging enables the data points to have an affinity with other data points and leaves room for more correlations incase there are new insights or a new set of data is available.

Tagging data points can get messy, however the imprecision and messiness inherent in tagging is about accepting the natural chaos of the world. It makes the vastness of content more navigable[5].



Tagging of Data enables correlations without creating copies

[1] Census 2011

[2] Envisioning Information: Edward Tufte

[3] Big Data: Victor Mayer-Schonberger and Kenneth Cukier

[4] Laws of Simplicity: John Maeda

[5] Big Data: Victor Mayer-Schonberger and Kenneth Cukier

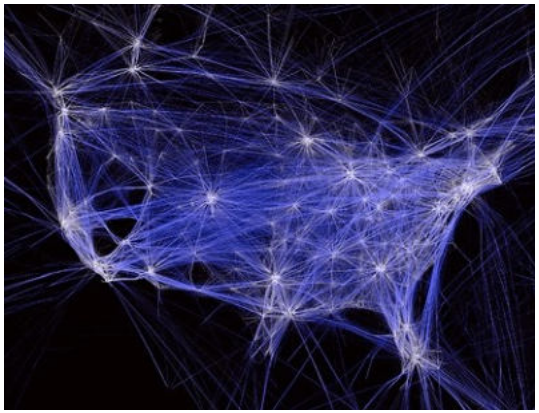
Correlations

When dealing with large datasets, as seen earlier it is better to tag and correlate than to categorize and bucket the data points in a hierarchy. Some type of data might fall in multiple categories. Take for example a book on the Gestalt Theory; a person from social sciences may place it in the Psychology bucket whereas it is also welcome in the Graphic Design bucket; it does not mean that you buy two books. If you were to tag these books, you could easily tag it as Psychology and Graphic Design, thereby solving the problem of categorization. This feeble example gives an idea of the powerful concept of tagging in the Big Data world. Tagging in this case may also be called Metadata. Metadata simply means data about data.

Correlations are extremely useful when trying to solve big data problems. Large problems contain large amounts of data and with enough data, the numbers speak for themselves. A correlation is enough.

Correlations shine in the Big Data world, like in the above visualization by Aaron Koblin on the flight patterns in the United States. One can see the busiest airports, density of arrivals and departures, airports with the maximum number of international flights and much more. All this from a single image.

We have since a long time depended on the method of starting out with a hypothesis and analyzing the data to prove or disprove the latter. But with so much data available, the data starts to speak for itself. We need not start with a hypothesis.



AARON KOBLIN'S FLIGHT PATTERNS
SOURCE : GOOGLE IMAGES

Humans have an innate need to find causal connections. It is easy to find numerous examples in mythology where one event is caused by another. When dealing with huge amounts of data, one needs to be aware of the subtle difference between correlation and causality. In simple terms, causality is when A causes B. Correlation on the other hand means that A and B happen at the time; so we need to watch out for B to predict that A will happen.

A study says that women are happier than men. Most women have longer hair. So the data suggest that there is a connection between long hair and happiness. This is a weak correlation. Correlation is a powerful tool, though we need to watch out for such weak correlations.

Mistaking correlation for causation almost led the state of Illinois, USA to send books to every child in the state because studies showed that books in the home correlated to higher test scores. Later studies showed that children from homes with many books did better even if they never read, leading researches to correct their assumptions with the realization that homes where parents buy books have an environment where learning is encouraged and rewarded[1]. Our Governments too don't have money to waste going in the wrong direction.

Big Data problems bring with them Big opportunities. We can solve large problems, make discoveries and even predict the state of world. It turns us into seers and wizards with a crystal ball, to look into the future.

[1] Freakonomics: Stephen J. Dubner and Steven D. Levitt

2.3.Mumbai's Data

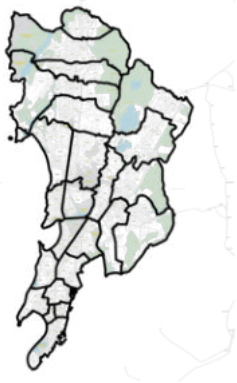
The Municipal Corporation of Greater Mumbai (MCGM) has released a vast amount of data to the public. This data includes the existing land use, population and other relevant data sets. Also data about governance, health, education, etc is made available by a number of other government bodies and NGOs. This project aims to bring these data sets together and visualize them to generate insights about the city and its people. This project tries to deal with the quantifiable information about the city's infrastructure, population, growth and environment. Other datasets may be used to find interesting insights about the city and its divisions. The project tries to intentionally do away with the social and emotional aspects of the data.

Note that the word infrastructure is used in its broad sense and includes residential, commercial, industrial infrastructure along with medical, educational, social amenities and utilities. Environment includes water bodies, rivers, forests, national parks, open spaces, mangroves, hills, mud flats, etc.

Divisions

The City of Mumbai has approximately 1 lakh 27 thousand land parcels[1]. For the purpose of ease of administration, the city is divided into 24 wards which are further divided into planning sectors. The City has a total of 151 Planning sectors.

The population of Greater Mumbai is 12.44 Million and is India's most populous city[2]. Each ward constitutes of a population between 1,00,000 to 5,00,000 persons[3].



24 Wards

SOURCE : MCGM
MAP NOT TO SCALE

151 Planning Sectors

SOURCE : MCGM
MAP NOT TO SCALE

[1] Talk on 22nd April in IIT Bombay by Mr. Vidyadhar Phatak, Urban Planner

[2] Census 2011

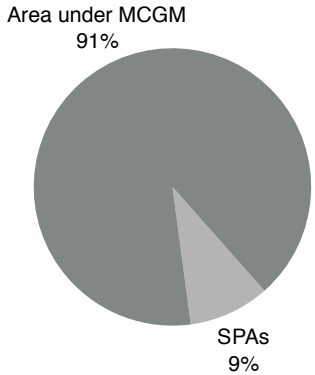
[3] Preparatory Studies DP 2014-34 : MCGM

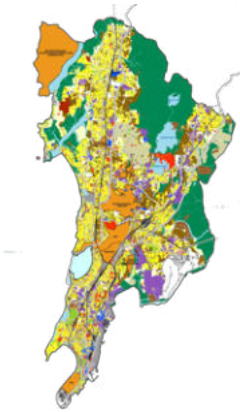
The Municipal Corporation of Greater Mumbai has provided data about the existing land use. This data is spread over multiple PDF files. It becomes laborious to extract data from so many files.

However, the MCGM has provided the data for 90.57% of the total area of Mumbai. The rest of the 9.43% area does not fall under the MCGMs jurisdiction. These areas are termed as Special Planning Areas (SPA).

The SPAs include

- The Gorai-Manori Tourism Zone
- Oshiwara District Centre (ODC)
- Airport
- Bandra Kurla Complex
- Wadala Truck Terminal
- Backbay Reclamation Area
- Marol MIDC
- SEEPZ SEZ
- Dharavi



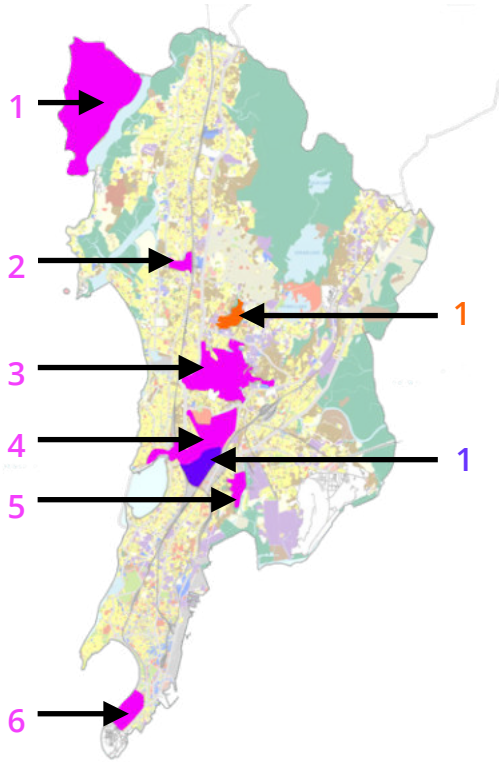


Existing Land Use Plan

SOURCE : ELU REPORT MCGM
MAP NOT TO SCALE

Area under MCGM

BASE MAP SOURCE : MCGM
MAP NOT TO SCALE



Area under SPAs

BASE MAP SOURCE : MCGM
MAP NOT TO SCALE

MMRDA
1. Gorai-Manori Tourism zone
2. Oshiwara District Centre (ODC)
3. Airport
4. Bandra Kurla Complex (BKC)
5. Wadala Truck Terminal
6. Backbay Reclamation Area

MIDC
1. Marol Industrial Area, SEEPZ SEZ

MHADA
1. Dharavi

List of Available Datasets :

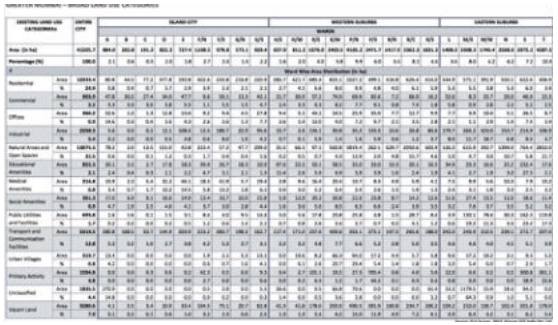
- MCGM
  - Existing Land Use of all Wards
- MMRDA (Land Use Data)
  - Gorai-Manori Tourism zone
  - Oshiwara District Centre (ODC)
  - Airport
  - Bandra Kurla Complex (BKC)
  - Wadala Truck Terminal
  - Backbay Reclamation Area
- MIDC and SEEPZ (Land Use Data)
  - Marol Industrial Area, SEEPZ SEZ
- MHADA (Land Use Data)
  - Dharavi
- Housing
  - Cessed Buildings
  - Heritage Sites and Buildings
  - Ready Reckoner Rates (Housing prices)
  - Slums
- Health
  - Hospitals in wards
  - Top 5 Diseases
  - Top 5 Sensitive Diseases
- Education
  - Gender Ratio in Primary Education
  - Dropout numbers in Primary Eductaion
  - No. of teachers
  - Distribution of schools and colleges
- Transportation
  - Existing transport networks
  - Proposed transport networks
  - Auto and Taxi stands
  - Peak time traffic roads
  - Heavy pedestrian congestion areas
  - Road Accident prone spots
- Water and Sanitation
  - Waste management facilities
  - Settlements not served by sewer lines
- Livelihood
  - Employment Rate
- Open Spaces
  - Wardwise Open space data
- Environment
  - Natural spaces and water bodies
  - Flood and Cyclone prone areas
  - Intertidal Zones
  - CRZ Boundaries
  - Wardwise Tree cover
  - Noise Pollution (Ambient noise)
  - Air Quality
- Governance
  - Police Stations and Jurisdiction
  - Ward wise FIRs : Murders
  - Ward wise FIRs : Crimes against women
  - Ward wise FIRs : Rapes
  - Ward wise FIRs 2011 - 2014
- Population
  - Ward wise population Data 2001-2011
  - Declining Growth rate of Population
  - Slum and Non-Slum population
- Budget

Problems with the City Datasets

Despite the availability of so much data, it is very difficult for a person looking at the data to make any sense of it.

Tabular Structure

Most of the data is in a tabular format. One can infer little by looking at tables and maps with a long legend next to it.



SOURCE : ELU 2012, SCE INDIA PVT LTD

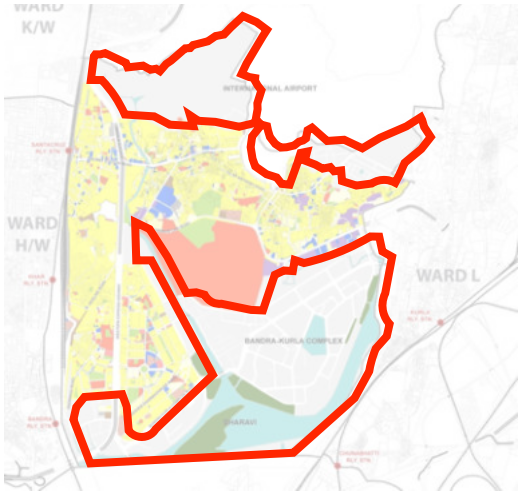
Multiple File Formats

Another problem with the data available is that it is spread across multiple PDF or Tabular format files. There is little inference that can be gained from these files and it becomes difficult to use the data in a constructive manner. Multiple file formats often prevents easy machine readability of the data. Converting these files into a machine readable format is often an expensive and laborious task and is an unnecessary step in the analysis of data.

Incomplete Datasets

Ward wise ELU data has been provided by the MCGM. However, this data is grossly incomplete for some wards, since parts of these wards fall under the SPAs, which is falls out of the MCGMs jurisdiction.

An example of this can be seen with ward H/E. A major part of the wards' area falls under the Airport and Bandra-Kurla Complex SPA. However, the ELU data that has been captured excludes these ares, which account for a sizable part of the ward's overall area. There are a number of examples like these in different wards.



WARD H/E : A LARGE PART OF THE WARD'S AREA FALLS UNDER SPA AND THEREFORE HAS BEEN EXCLUDED FROM THE WARD DATA.

SOURCE : ELU MAPS, MCGM
MAP NOT TO SCALE

Temporal Data types

The temporal data available can be either pointed out on a map or is a value that cannot be measured by its physical dimensions. There would also be data that falls into both these categories. It needs to be noted that all the data being used is quantifiable in nature i.e. we can define its position by means of a co-ordinate system or it is a number like the population of a ward. For this reason, it makes sense to divide the data type into spatial and non-spatial.

Spatial Data : It is data that can be pointed out on a map. It is situated in a space. The value of the data can be extracted from map, co-ordinates, etc in 2D space and 3D co-ordinates in a 3D space.

Non-Spatial Data : It is data that cannot be pointed out on a map, it exists as a number in time.

1 + 1 = 3 in the case of data. Much value can be gained from data sets if they are looked at in context with each other. An affinity can be identified between different and almost disparate data sets to gain new insights and generate new knowledge.

Spatial Data	Non-Spatial Data
Existing Land Use*	Ready Reckoner rates
Cessed Buildings	Top 5 diseases
Heritage sites and buildings + ASI sites	Top 5 sensitive diseases
Slums	Gender ratio in Primary education
Existing and proposed transport networks	Dropout numbers in Primary education
Road accident prone spots	Number of teachers
Auto-rickshaw and Taxi stands	Peak time traffic records
Heavy Pedestrian Congestion areas	Employment rate
Settlements not served by sewer lines	Number of vehicles
Flood and Cyclone prone areas	Noise pollution

Spatial Data	Non-Spatial Data
Intertidal Zones	Air quality
CRZ boundaries	Ward wise FIRs
Ward wise tree cover	Ward wise population - Slum and Non-Slum
	Budget 2015-16

*The Existing Land Use data includes data about :

- Medical Amenities
- Educational Amenities
- Open Space
- Natural Spaces and Water Bodies
- Residential areas
- Urban Villages
- Commercial Areas
- Offices
- Industrial Areas
- Social Amenities
- Public Utilities and Facilities
- Transport and Communication Facilities
- Primary Activity
- Vacant Land
- Unclassified Areas

The ELU data for all the wards has been taken from MCGM as well as from the SPA authorities which include MMRDA, MIDC and MHADA

The Existing Land Use (ELU) provided by the MCGM has more than 140 sub-categorizations for the land use. There are a number of instances where the subcategories in the same or different category have an affinity with each other.

The organizing diagram on the page shows different land use categories and sub categories in the ELU. The colored lines indicate affinity between different sub-categories.

The boxes around multiple categories indicate that these can be merged into a single category and its content can be branched out.

140+

Sub-Categories in the ELU

3. Preliminary Design Ideas

The project's aim is to look at the open data about the city of Mumbai and present it in an easily understandable manner for its citizen without dumbing down the data. The objective is to do a visualization of data to show it in ways that the users, observers of the data can gain new insights about the city. The ELU data provides a majority of the dataset needed to visualize the data. However, additional data has been collected from various other government organizations such as the MMRDA, MIDC, SEEPZ and others mainly for land use data for the SPAs as well as NGOs and other organizations like Praja, The Urban Design and Research Institute (UDRI), Mashal, etc for other useful data like crime rate, education among other data sets.

The visualizations for the project can be categorized into broad categories as follows :

- Housing
- Livelihood
- Education
- Health
- Open Spaces
- Environment
- Crime
- Transportation
- Water and Sanitation

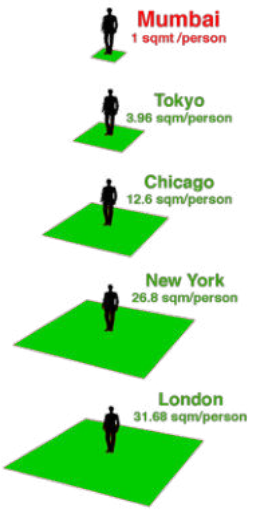
These categories are broad categories and mainly the available data from credible sources would be used to generate the visualizations. In some cases, only relevant data for a particular visualization may be used. A systemic design problem would be to create a system for visualization of the data to for reasons of scalability incase more diverse and relevant data becomes available.

"Re-describing the world is a necessary first step towards changing it[1]."
Salman Rushdie

Data Visualizations for the city can be categorized in three broad categories i.e. Comparative Data Visualization, Exploratory Data Visualization, Isolated Data Visualization.

Comparative Data Visualization

It is also important to consider that the data should not be looked at in silos. For this reason, it is better to compare the data of the city either with other cites that are similar in nature or if it is the case of wards, looking at adjoining wards could be a



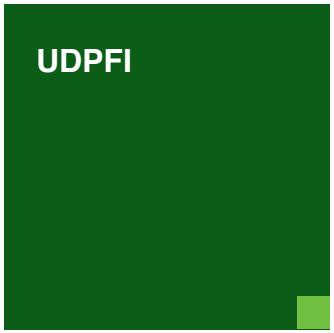
COMPARISON OF OPEN SPACE PER CAPITA BETWEEN DIFFERENT CITIES

SOURCE: "MUMBAI MUSEUM: A FACELIFT" THESIS BY AKSHAY KORE

solution. This helps in forming a judgement about the data in context with nearby areas. For example, the deficiency of Hospitals in a particular ward may be alarming fact. Although if the adjoining ward has a large hospital in its boundaries; it could mean an

[1] Imaginary Homelands: Salman Rushdie

excess of hospitals in the adjoining ward, thereby negating the deficiency in the ward being looked at.



COMPARISON OF OPEN SPACE PER CAPITA BETWEEN DIFFERENT URBAN DESIGN STANDARDS AND MUMBAI

Exploratory Data Visualization

It might not always be the case where the creator of the data visualization may know the intended use, or all the insights the visualizations shows. Too much data also is confusing. Giving the user a way to explore the datasets and find new insights is also a possible way to present the data. A system of layering datasets for multiple interpretations can be adopted. The purpose of an exploratory visualization is to reveal patterns and relationships not known so easily deduced without the aide of a visual representation. This type of visualization helps in producing new knowledge.

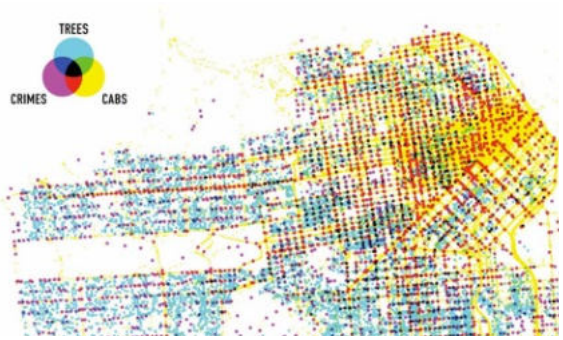
[1] Envisioning Information: Edward Tufte

AN EXPLORATORY VISUALIZATION OF THE AGE OF BUILDINGS IN NEW YORK.
SOURCE : HTTP://PUREINFORMATION.NET/BUILDING-AGE-NYC/

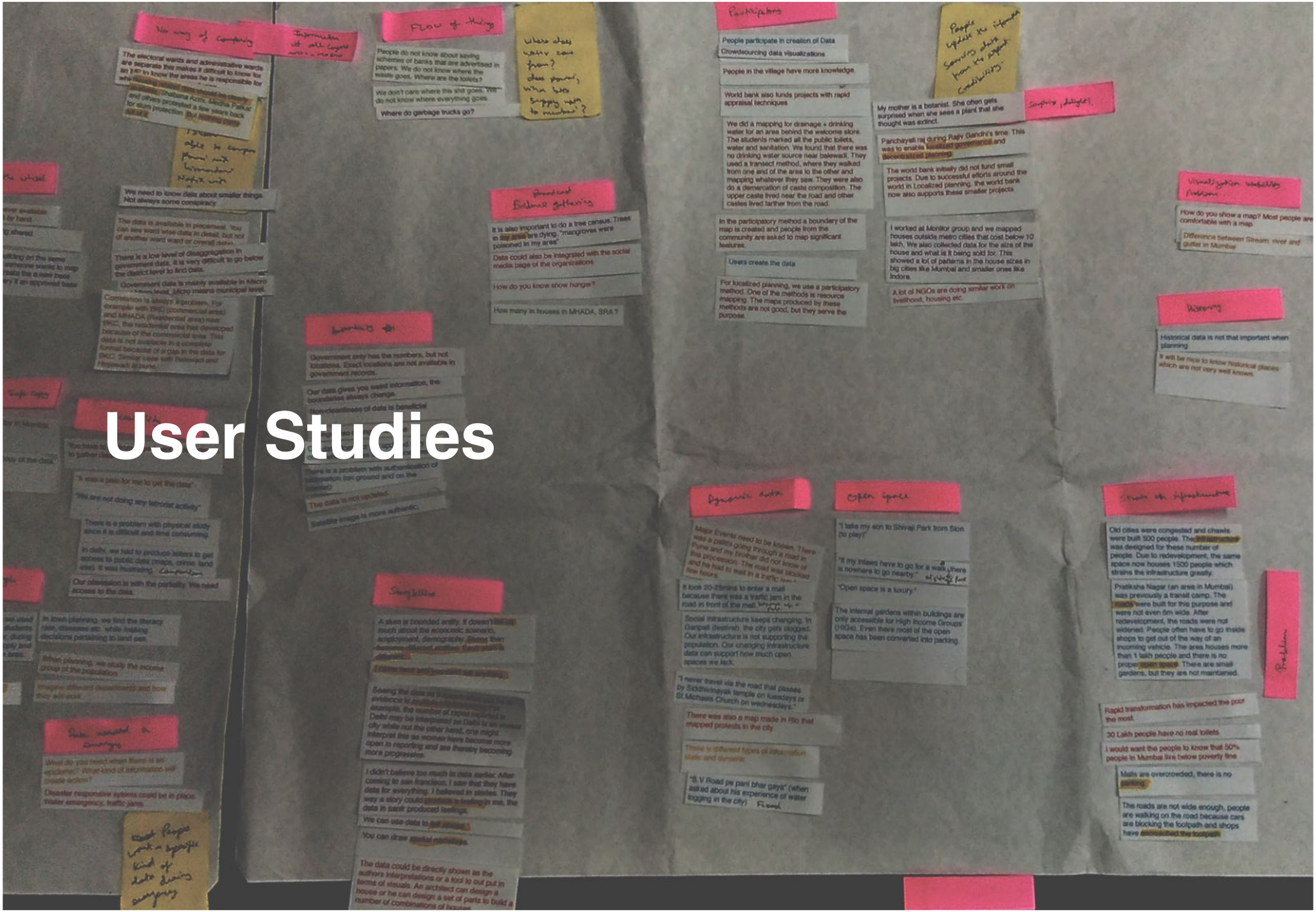
Isolated Data Visualization

Another strategy is to create a series of static and time based visualizations that highlight certain aspects of the quantitative data about the city. These could be in the form of a series of posters or interactive visualizations that are static or time based. This approach would require generation of multiple isolated visualizations.

"Simplicity of reading derives from the context of detailed and complex information, properly arranged. A most unconventional design strategy is revealed: to clarify, to detail[1]."
Edward Tufte



VISUALIZATION OF THE TREES, CRIMES AND TAXI CAB LOCATION IN NEW YORK.
SOURCE : HTTP://WWW.ARCHITECTMAGAZINE.COM



User Studies

4. User Studies

For the purpose of deciding on the approach to be taken while creating the data visualization for Mumbai, a user study was conducted with stakeholder groups.

4.1. Users

Users or consumers of data visualization are varied. The users may be students, common folk, teachers, architects, economists , etc. For the purpose of the project, it was decided that the visualization would be used by policy and decision makers in Government bodies, private organizations or NGOs. These users mainly include city and urban planners, academicians and architects.

Academicians	1
Architects	3
Policy	2
Urban Planners	6
Journalist	2
Total Number of Users Interviewed	**14**

For the purpose of the study, the names of the interviewers has been kept anonymous.

4.2. Contextual Enquiry

A method of contextual enquiry was employed to conduct the user study. The aim of the enquiry was to generate insights about the type of visualizations used by the user group and the needs and problems faced while making strategic decisions. A semi-structured interview was conducted for different types of users. For example, an academician would be asked a slightly different set of questions than an architect or a policy maker. Interviews were conducted with architects and urban planners working in both the public and private sector. Professors from the Tata Institute of Social Sciences, Mumbai were interviewed as policy makers since they are often involved in decisions concerning policy making and government. Two journalists were also interviewed since they work extensively with open and government data to gain insights about access of data and generation of visualization. One academician was interviewed to enquire about the usage of data visualizations in classroom scenarios for hypothetical decision making. Many of the Architects and Urban Designers interviewed play or have played a dual role of a teacher as well as a professional. The interviews were recorded and/or a transcript was created during the contextual enquiry.

Questionnaire

The users overall were asked the following questions during the semi-structured interview, although not necessarily in the order shown. Some questions might have been skipped for certain types of users and depending on the context. Each question is a type of a conversation starter to discover deeper problems and ideas in the users' experience. The aim was to empathize with the user and probe into the users' thought process while working on a problem.

- **How do you use these datasets?**
  This was asked to check the type of usage between different user groups, whether it was to generate an insight or prove a hypothesis. This would enable the discovery of a possible usage pattern of data and the type of visualizations used.

- **When was the last time you had to use one of the datasets?**
  This was asked to probe into the users' experience with using open datasets and gain insights about the user journey and touch-points.

- **What datasets did you use?**
  This was asked to find common data sources used and major overlaps between different user preferences.

- **Where did you have to look for ? Did you find what you were looking for?**
  This was asked to probe into the effort involved with accessing datasets.

- **What do you wish was easier in the process?**
  This was asked to find the immediate pain points that the users themselves could identify. Other pain points would be identified by generating an affinity map of the user statements.

- **How did you access the Information, what medium did you use? (mobile phone, desktop, books etc.)**
  This was asked to find a possible pattern or trend in the predominant medium of access of datasets.

Other questions that were asked included the following.

- What data should people know about? What matters?
- What are the important parts of the data? What is not so important?
- What is your stand on the data available?
- Any current problems with the data?

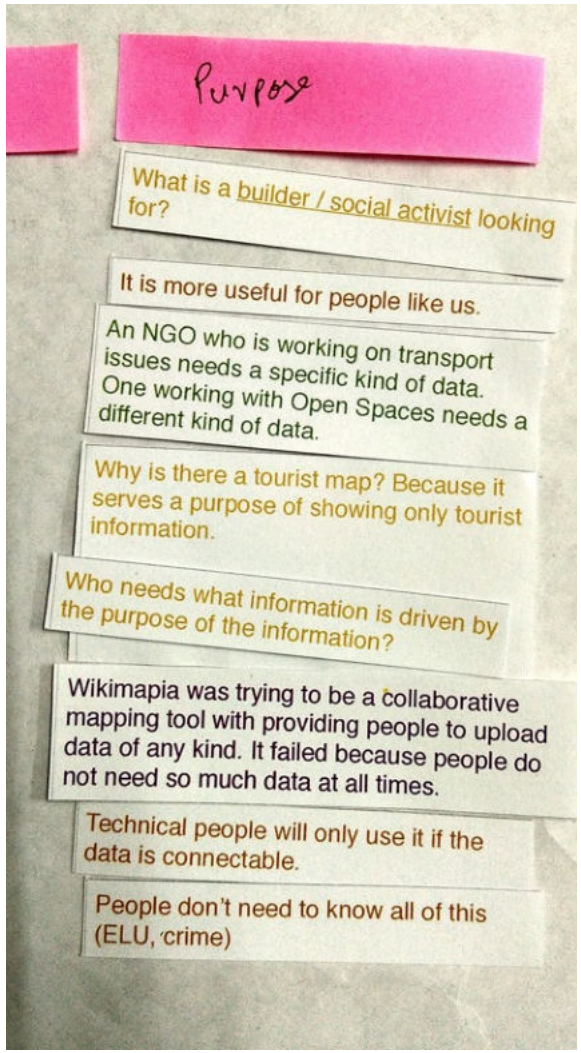
4.3.Affinity Mapping

An affinity map was created after the contextual enquiry. All the user statements were noted down and the most important statements were printed. The statements were grouped together depending on their affinities to each other. The users' needs, problems and pain points would be identified by generating an affinity map.

4.4.Insights

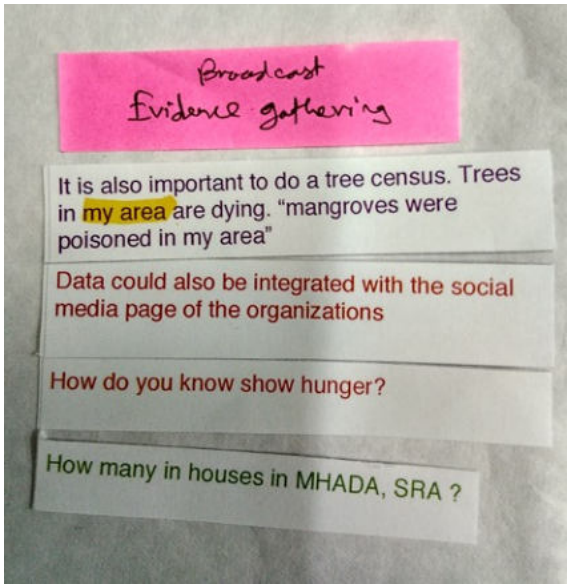
Purpose driven data visualization

The user group interviewed used or created data visualization with a particular purpose in mind and were not merely consumers of the data visualizations. The purpose of visualizing the data would vary from problem to problem and the users' would use different datasets or different types of correlations to probe into different problems.



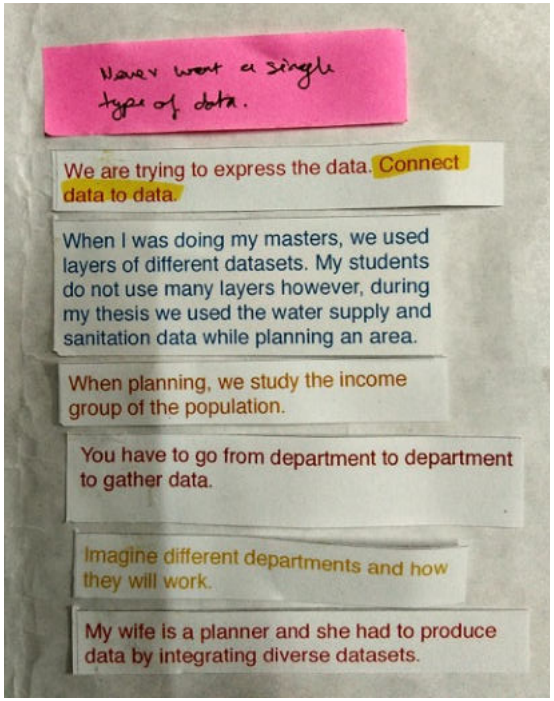
Evidence Gathering and Broadcasting

Many of the visualizations were used as a proof of evidence. In one such example, a user who was an urban designer said that the mangroves in the users' area were poisoned. If there was some evidence in the form of a map or a time series, the user would broadcast it on social media to gain traction and make the authorities notice the problem. Since, this was not possible for the user given the lack of technical expertise, the user decided to write a blog post instead, however it was not very successful.



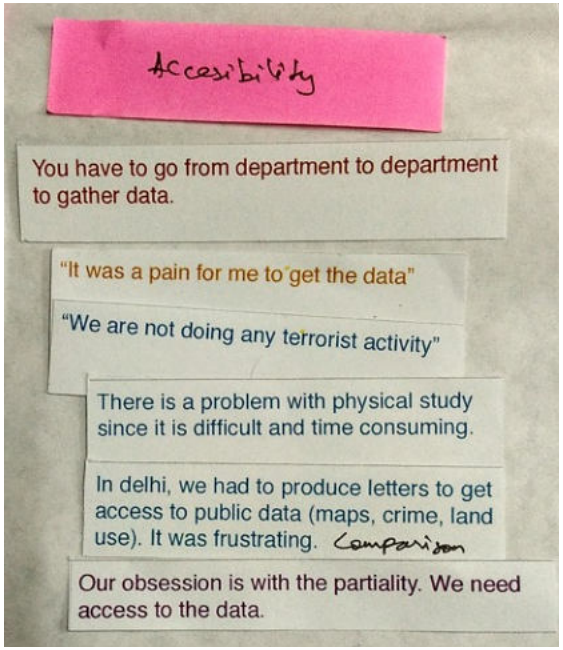
Users never want a single type of data

The affinity showed that different users used different types of data to generate visualizations that led to evidence and insights.



Accessibility of Datasets

From the affinity map, it was seen that there was a deep resentment with the process of access of data. Many of the users felt intimidated by Government office environments where they would to go to collect data or in case of availability of data on the internet, the datasets were in many cases not easily searchable and were often hidden deep within the websites of these authorities. One such user statement which showed the hostility was "we are not doing any terrorist activity".

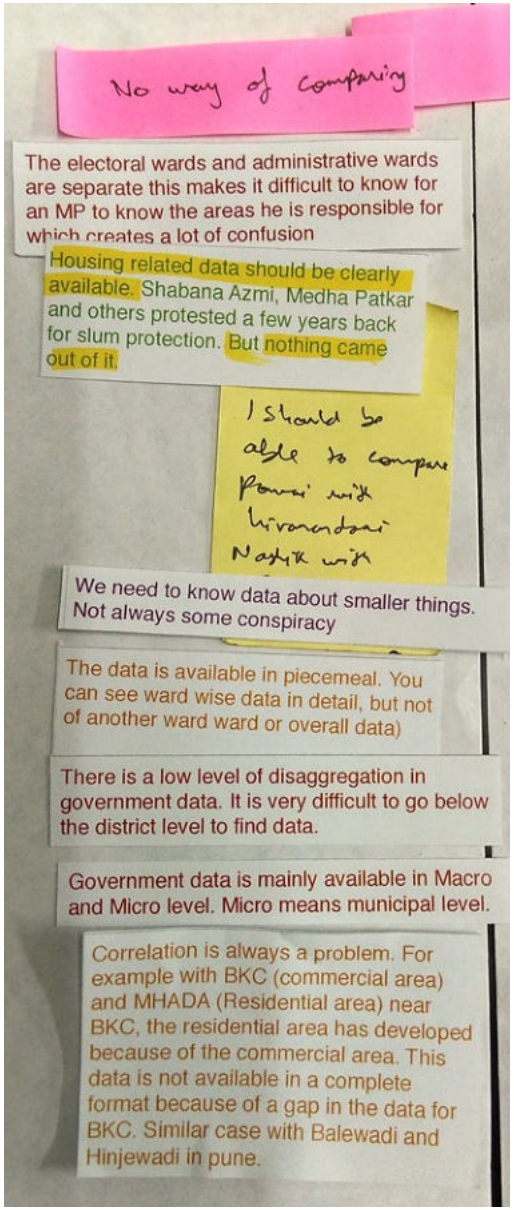


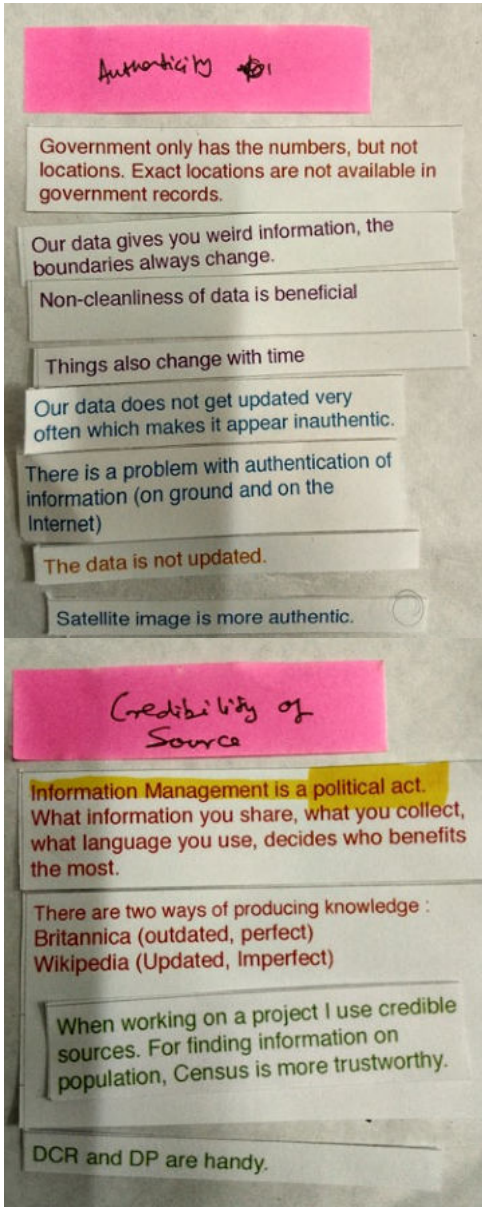
There is no (easy) way of comparing data

Many of the users felt that under the present scenario, it was very difficult to do a comparative study of different datasets or areas. The data is available in piecemeal and not often on a broad level. And yet when making data driven decisions, one needs to look at the overall scenario and not only smaller data silos.

Authenticity of data and Credibility of source

The affinity indicated that many of the users had a problem with the authenticity of the data. Many of the times the quantitative data was known but not the location. The datasets do not get updated very frequently and this makes it appear inauthentic. However, when asked about the credibility of the data, all users said that they found Government sources the most credible despite some errors because of their exhaustiveness. Authenticity and credibility of datasets was a major concern for journalists.



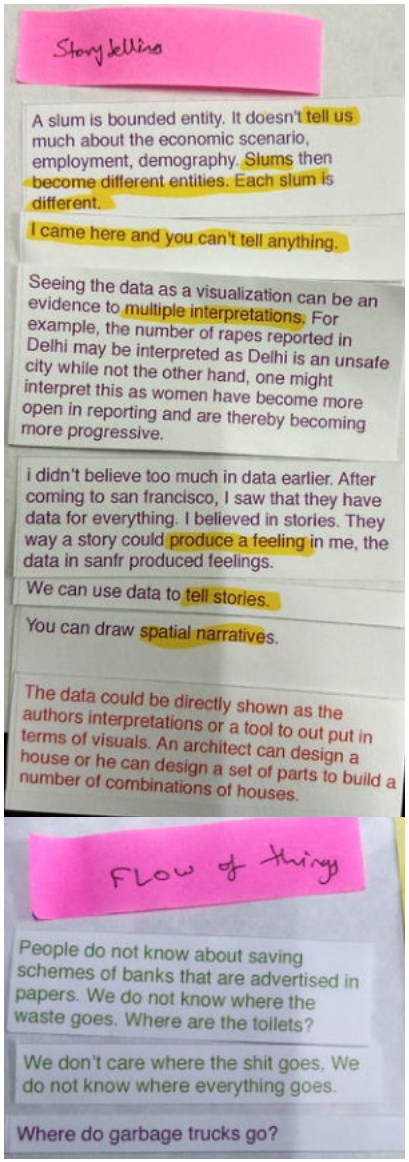


Dynamic Datasets

Many users referred to events that happened periodically and a decision was made by intuition or pre-knowledge. The temporal social infrastructure of the city keeps changing and the data should indicate the same to make decisions or generate insights about the city. This has been referred to as dynamic datasets.

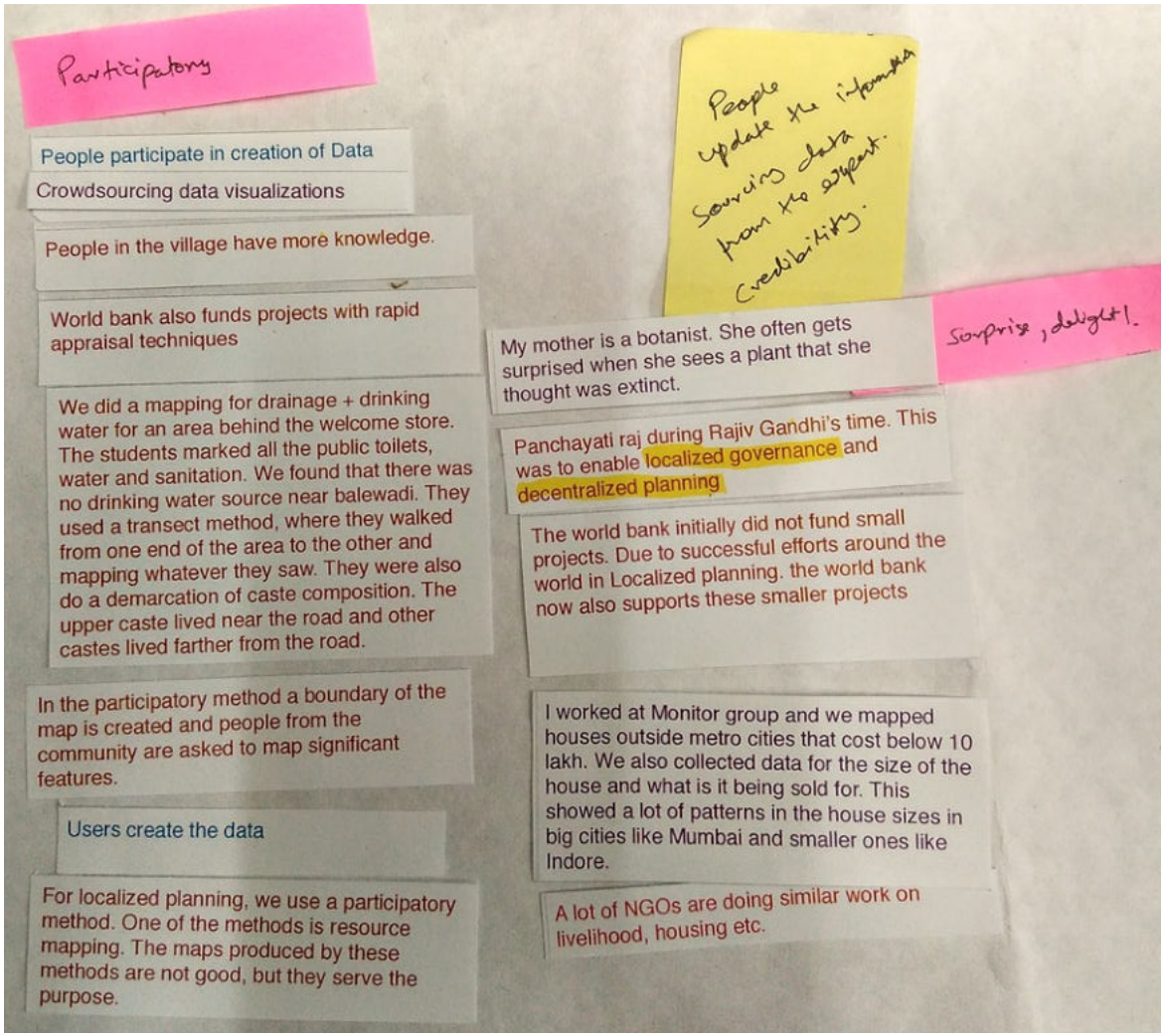
Multiple Interpretations of Data, Story Telling and the Flow of things

The affinity showed a high level of interest between different users about the story that the data told and how temporal things moved. One such user gave an example of a slum, where a slum is indicated as a bounded entity by government documents. However, on looking deeper into the data, a complex story is revealed and one finds that each slum is different and its problems cannot be solved in a one fits all solution. A single visualization may proclaim a fact or lead to multiple interpretations of the same data. Many users were intrigued by the way things moved.



Participatory Data Creation

Another recurrent insight was the idea that people participate in the creation of data. Many users felt that this was the need to conduct localized planning. Local people participating in data creation generates richer insights about the locality, enabling more thoughtful planning of policies and infrastructure.

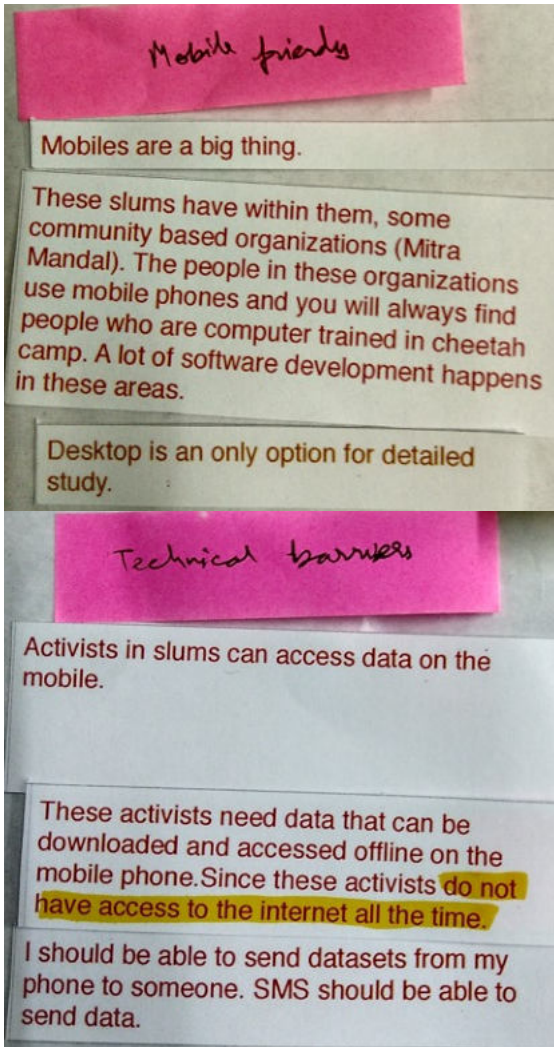


Desktop and Mobile usage

It was seen that for many issues of localized planning, a mobile phone was used to access broad level data on ground, while a desktop was used to conduct a detailed study.

In a nutshell, following were the key insights discovered after the user studies and the affinity mapping.

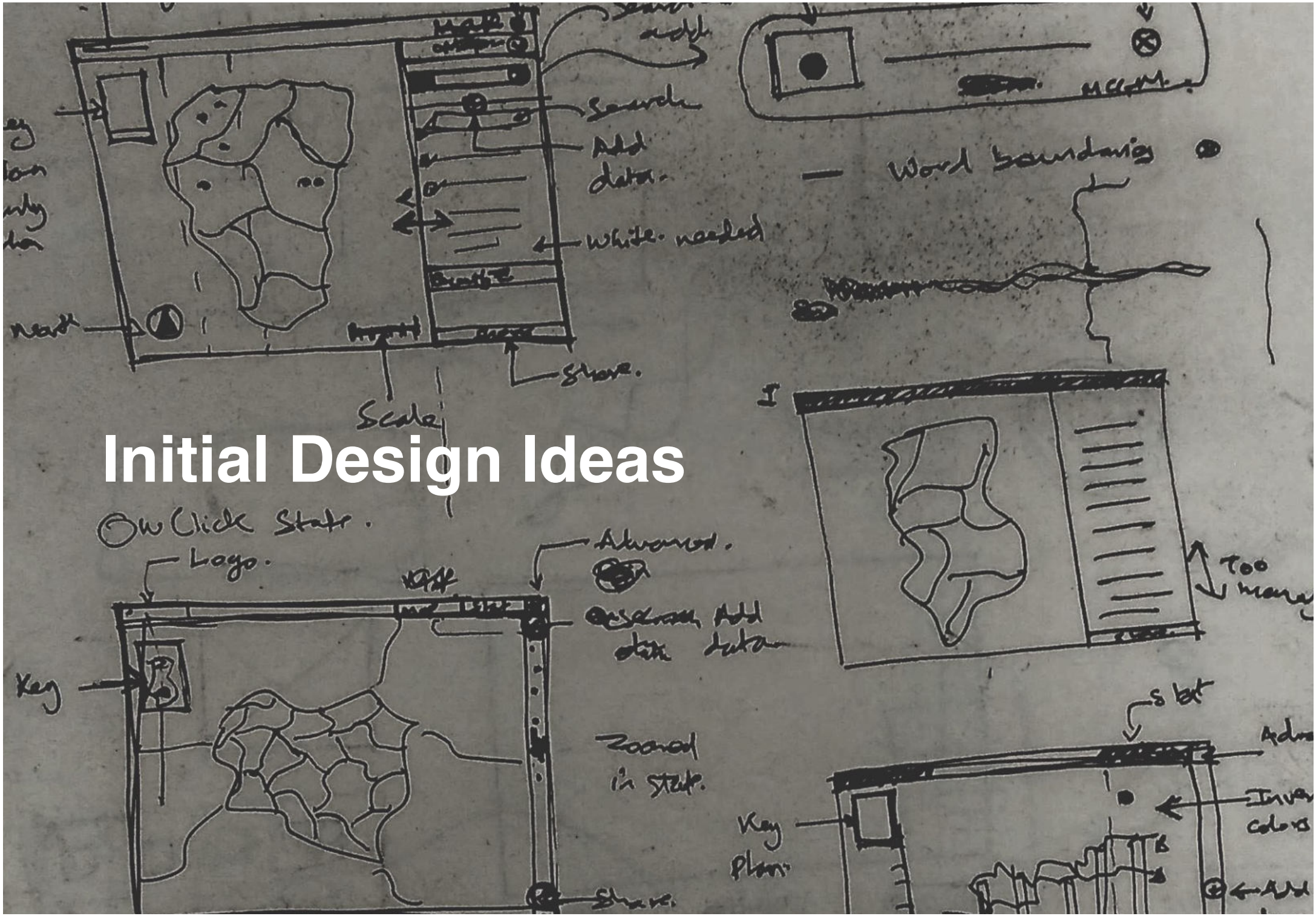
- Visualization of data for the user group interviewed is driven by **purpose**.
- Users do not need all data all the time but they do want to **compare** different types of data.
- Knowing the source of data is very important. **Authenticity** of data is an assurance.
- **Accessibility** of data is a major hurdle in working with data.
- Data needs to tell a **story**, need to know the flow of things.
- **Participation** of the citizen in collection of data for localized planning produces richer insights.
- **Correlation** of data is one of the main reasons for visualizing data. Data is almost never looked at in silos.
- Mobile phones are used as a **medium** to access data on ground, while a desktop is used to do a detailed study.
- Maps and map based visualizations are the most frequent visualizations used.



4.5.Need Gaps

- It is often difficult to get access to all datasets and if a particular problem demands a specific data, it is a time consuming to get the data.
- Even after receiving the datasets, the data often requires some or extensive amount of processing before it is usable to do any analysis.
- Many a times, the users are not aware of the need of a particular dataset and need to use it on the fly.
- Most of the users interviewed, tend to use a graphical representation of data mainly in the form of maps to generate insights and make city or local level decisions.
- Majority of the users have a lack of technical expertise in map based data visualization and analysis tools like GIS.

"Information management is a political act. What information you share, what you collect, what language you use decides who benefits the most."
Prof. Himanshu Burte, TISS



Initial Design Ideas

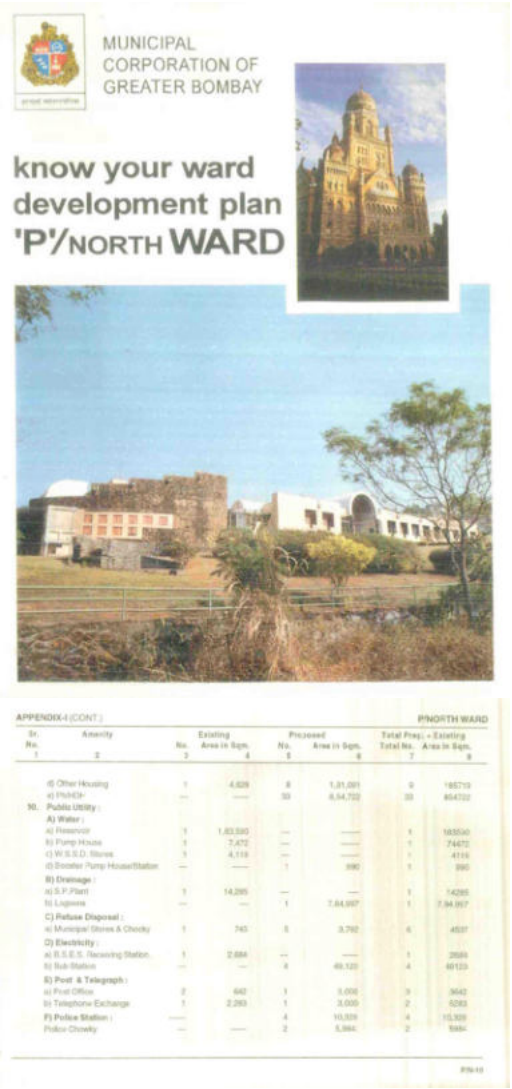
5. Initial Design Ideas

The insights and key findings from the user studies led to a few initial design ideas as shown below. An important finding was of users using purpose driven visualizations of data. This led to a key insight that it would be necessary to give the users the power to use multiple datasets easily without too much of post processing or technical skills. A series of visualizations would not work since it would require the creation of a large number of visualizations by hand and even that would not suffice the purpose of a particular visualization created. An exploratory visualization would be better suited for the current user group. These ideas also take some of their components from the preliminary design ideas mentioned earlier.

Comparison of Ward data

Comparison of data is an important task done by the users while making any decisions pertaining to the city. In case of Mumbai, the most frequent city division used is that of the Ward. The MCGM releases a KYW (Know Your Ward) document which is a booklet for each ward providing details about the ward like the population, the area of residential, commercial, industrial and other such land parcels, income levels, slum population, major transport networks , etc. The data from these booklets is used regularly by the user groups.

However, when comparing two or more wards, the users have to extract data from these documents since they are the only ones readily available and convert that into a suitable format and use them thereafter. The comparison of data of two or more wards may include the comparison of population, income levels of the population, area of the wards, flood prone areas, mangroves, forests, open spaces, medical and educational amenities, etc.

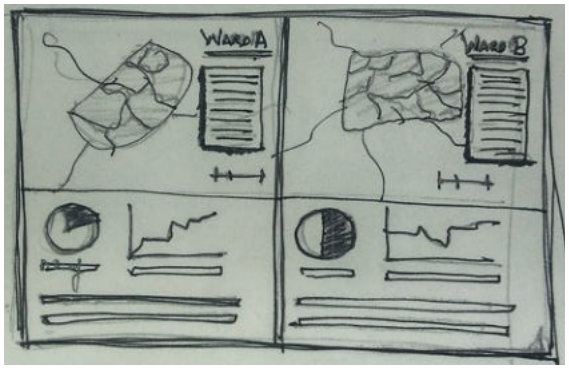


KNOW YOUR WARD BOOKLET FOR P/N WARD
SOURCE : MCGM

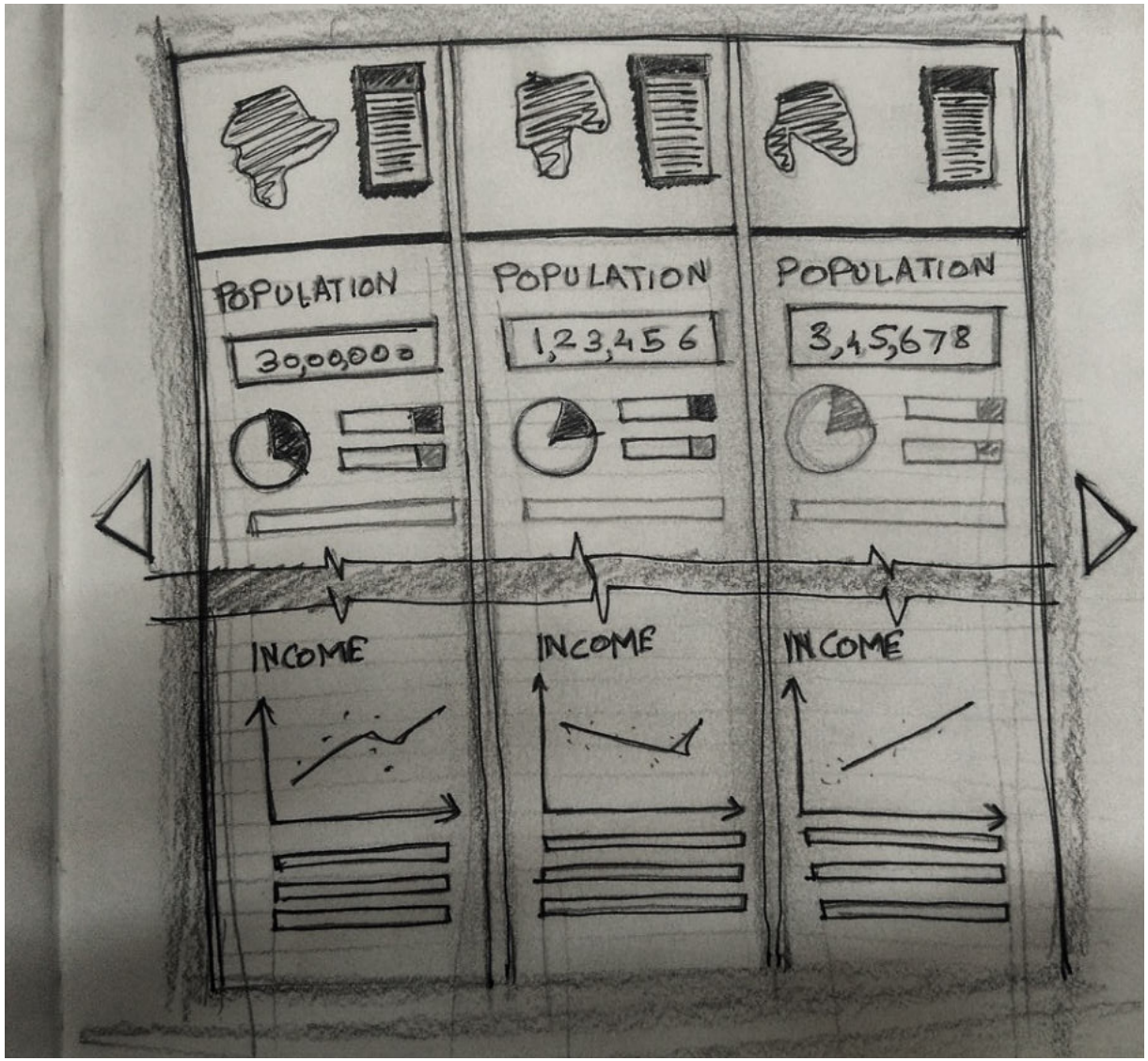
The design idea is of a tool that is a digital version of the Know Your Ward document which enables exploration and comparison of ward data. It would enable the user to compare individual data points as well as combine multiple data points of different wards.

Limitations

Ward boundaries are official boundaries, however they do not exist on ground. A comparison between wards may not lead to an actionable insight, since the problem may lie at the intersection of two wards or in certain parts in the wards.



SKETCH FOR THE COMPARISON OF WARD DATA TOOL



SKETCH FOR THE COMPARISON OF WARD DATA TOOL

Layering of Datasets

Comparison of areas can also be done by looking at the areas from an aerial point of view. A map based visualization can be created by segregating different elements of the map like roads, railways, tree cover, residential areas, etc and using these elements as layers on the map. Layering of datasets is an idea used since a long time by humans, may it be medieval cartographers or today's advanced Geographic Information Systems. It is an idea also employed by John Snow to discover the sources of Cholera deaths in England, thereby leading to the discovery that Cholera spread through water[1].

Tangible Layering

Datasets to be layered may be tangible in nature, like physical layers made out of transparent or translucent paper. These can be layered on top of each other to find correlations and patterns. The user group frequently uses tracing papers to layer or work on planning and is familiar with this approach of layering datasets.

LAYERS OF TRACING PAPER USED EXPLAIN AN ARCHITECTURAL SITE
SOURCE : HTTP://WWW.GARDENISTA.COM/

Limitations

Layering of datasets using physical and tangible layers to generate multiple data visualizations depending on the need and

A GROUP OF URBAN DESIGNERS DISCUSSING, NOTICE THE MULTIPLE TRACINGS ON PAPER
SOURCE : HTTP://RAVB.NL/

context is a favored approach, however it has some major shortcomings. It would be very difficult to scale the tangible system across a large number of decision makers. At the same time, updating, adding or deleting datasets across all users would be a laborious task. It would be extremely difficult to show time dependent visualizations with this approach.

1 Visual Display of Quantitative Information, Edward Tufte

Digital or Intangible Layering

A digital layering of the city's data is an equivalent of the tangible layering approach on a web based digital platform. The layering of datasets would be done on a digital medium with layers added and deleted to find correlations and patterns. This method is highly scalable as any change made on the application or to the datasets by the developer or the concerned authority would be reflected with all the users. This approach also enables visualization of time dependent visualizations, making it an extremely favorable approach.

Limitations

The approach would require the users to have a basic knowledge of using computers. The users would also lose the feeling of using a tangible layer of data.

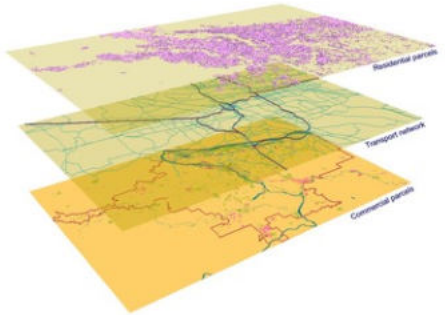


IMAGE REPRESENTATION OF LAYERING OF CITY DATA
SOURCE : HTTP://WWW.MEIPOKWAN.ORG/

MumbaiData

6. Final Design

It was found in the user studies that visualization is mainly a purpose driven activity for the user group. Users do not need all the data all the data, however they need quick access to any data without having to process the data beforehand. An analysis of the initial design ideas and their limitations along with the key insights from the user studies led to the formation of the MumbaiData. MumbaiData, as the name suggests is a map based visualization tool for Mumbai's Data. The idea of layering of datasets is a simple yet profound idea that has been employed in the tool. The data is visualized as layers overlaid on a base map. The data layers are preloaded and can be overlaid on the base map as and when the need arises.

6.1. Key Features

MumbaiData is a web based visualization tool and is made of two key components viz.

- The MumbaiData Visualization Tool
- Portal to access raw data used in the visualizations referred to as download datasets

Along with these key components, the web based interface consists of supplementary pages like the introduction to the tool and a guide to use the tool pages.

The MumbaiData Visualization Tool

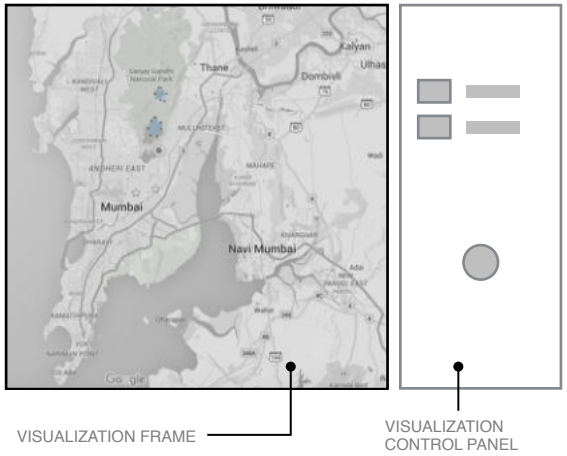
The MumbaiData Visualization tool has been divided into two parts, the Visualization Frame and the Visualization Control Panel.

Visualization Frame

This part is the frame where the user sees the visualization and any data the user chooses is visualized. The frame consists of a base google map layer. Google map has been used as a base layer due to the availability of other realtime data layers like traffic, terrain, local features like shops, clinics , etc., which are not easily captured by government datasets. The base map has been subdued by default for increased readability of the visualizations. However, this can be changed through the visualization control panel.

Visualization Control Panel

The Visualization Frame is controlled through the Visualization Control Panel (VCP). The VCP consists of a control to add overlays on the base map. These



overlays can be hidden or deleted from the visualization by means of the VCP. The VCP also enables the control of the color of the base map through the map view options. The VCP can be closed or opened. In the open state, the VCP also acts as a legend to the visualization in the Visualization Frame.

Download Datasets

The datasets used in the visualizations can be accessed as raw data though the Download Datasets page. The page consists of a simple list of links to all datasets used in the MumbaiData Visualization Tool and is updated when a new visualization layer is added to the tool. This enables the users to create their own visualizations or tools without having to worry about access or searching datasets on the world wide web.

6.2.Data Layer Types

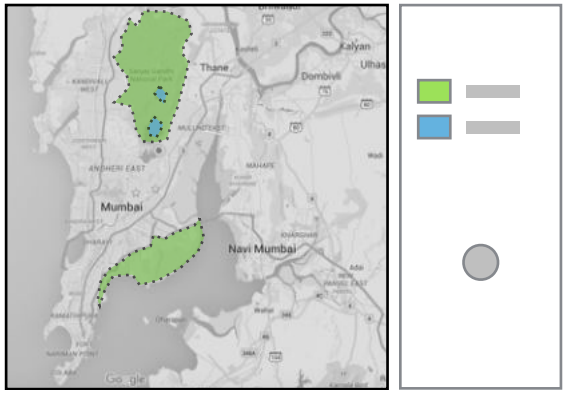
Overlays on the map can either be static or time based. As such, the visualization layers have been classified into two types viz. Static Data Layer and Dynamic Data Layer. The MumbaiData Tool handles these layer types differently.

Static Data Layer

These are the most recurrent types of the data layers. As the name suggests, these data type layers are static in nature. The inherent data does not change drastically for a long time. An example of this can be seen with residential areas, forests and natural areas type of data layers.

Dynamic or Time Varying Data Layer

Dynamic data layers are special types of layers that are time dependent. The inherent data changes with time and the time aspect can be controlled by the user. An example of this can be seen with cyclical events like festivals, traffic, climate , etc. This type of data layer becomes extremely useful in presentation of data stories or cause and effect scenarios.



AN EXAMPLE OF STATIC LAYERS

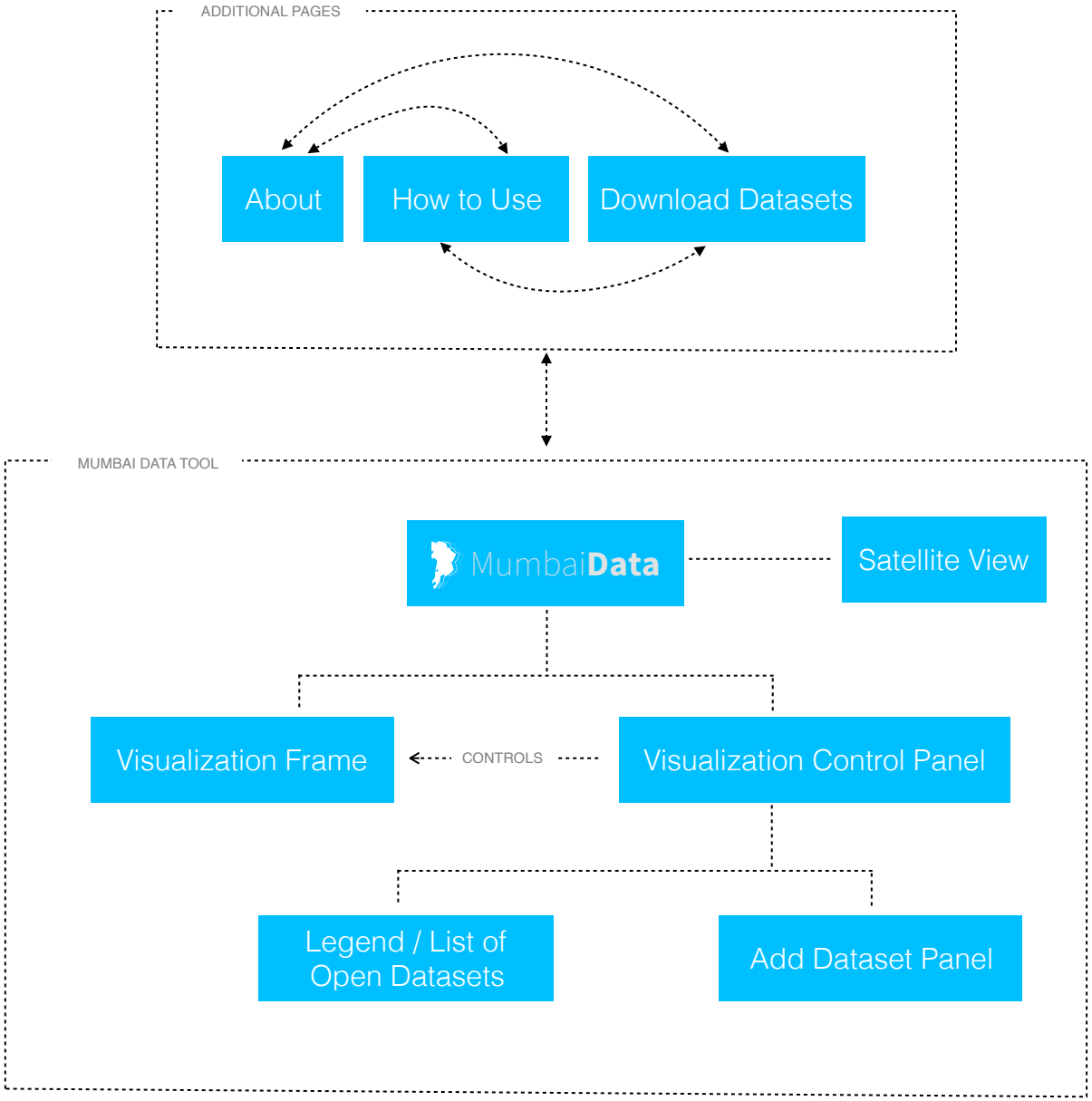
AN EXAMPLE OF THE USE OF A TIME VARYING DATA LAYER. THE YELLOW LINE INDICATES THE TIMELINE FOR INDICATION PURPOSE

6.3. Key user Tasks

Key user tasks may be defined as tasks which are imperative to the tool without which the tool loses its functionality to a great extent. The visualization tool enables the user to do certain key tasks which include adding a dataset, deleting a dataset from the legend, hiding and showing a dataset, zooming in and out and panning the visualization.

6.4. Information Architecture

The information architecture shows the journey the user takes to create visualization. A simple user journey would be Open MumbaiData Tool >> Click on Open Panel Button >> Click on the Add Layer Button on the Panel >> Select 2 datasets to Add >> Click on Add Selected Datasets. This example of a user journey enables the user to add layers to create a visualization.

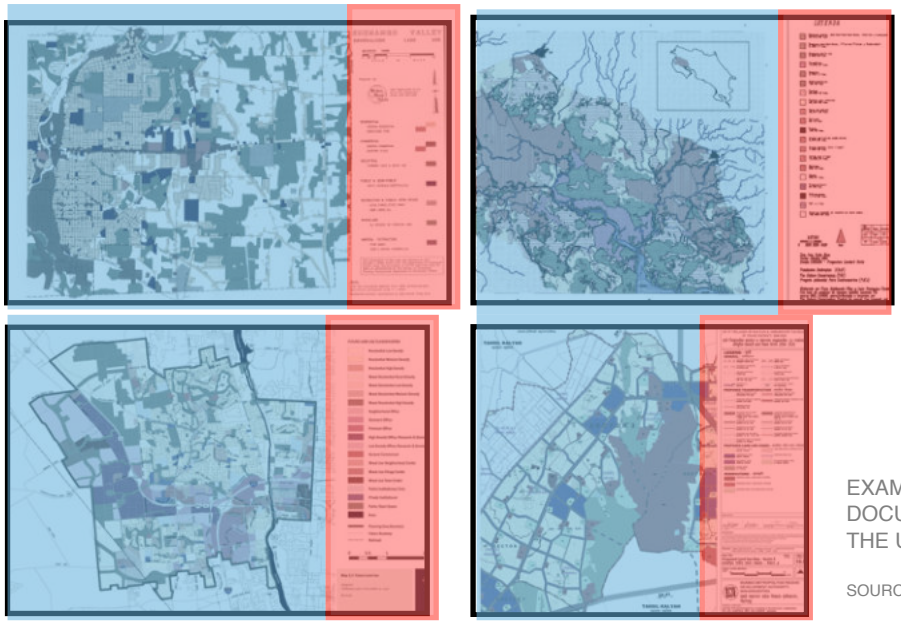


6.5.Metaphor

The user group uses certain types of government documents on a regular basis. It is an important consideration to use some of the users' pre-knowledge with using data documents as a User Interface metaphor for the visualization tool. It is assumed that most of the user population is familiar with using government map data. Since, the MumbaiData tool is a map based application, a number of maps released by the Government of India were analyzed and a user interface pattern emerged. It was of a legend on the right-side and the map visualization on the left occupying a dominant part of the document image. Along with this, the scale and style of the legend was an important UI consideration.

6.6.Initial Concept

The initial wireframe concept was an adaptation of the metaphor of the Government map based data documents for use in an interactive environment. It enabled the user to perform the key tasks of adding, deleting, hiding and showing datasets as well as zooming and panning of the user created visualization.



EXAMPLES OF DOCUMENTS USED BY THE USER GROUP

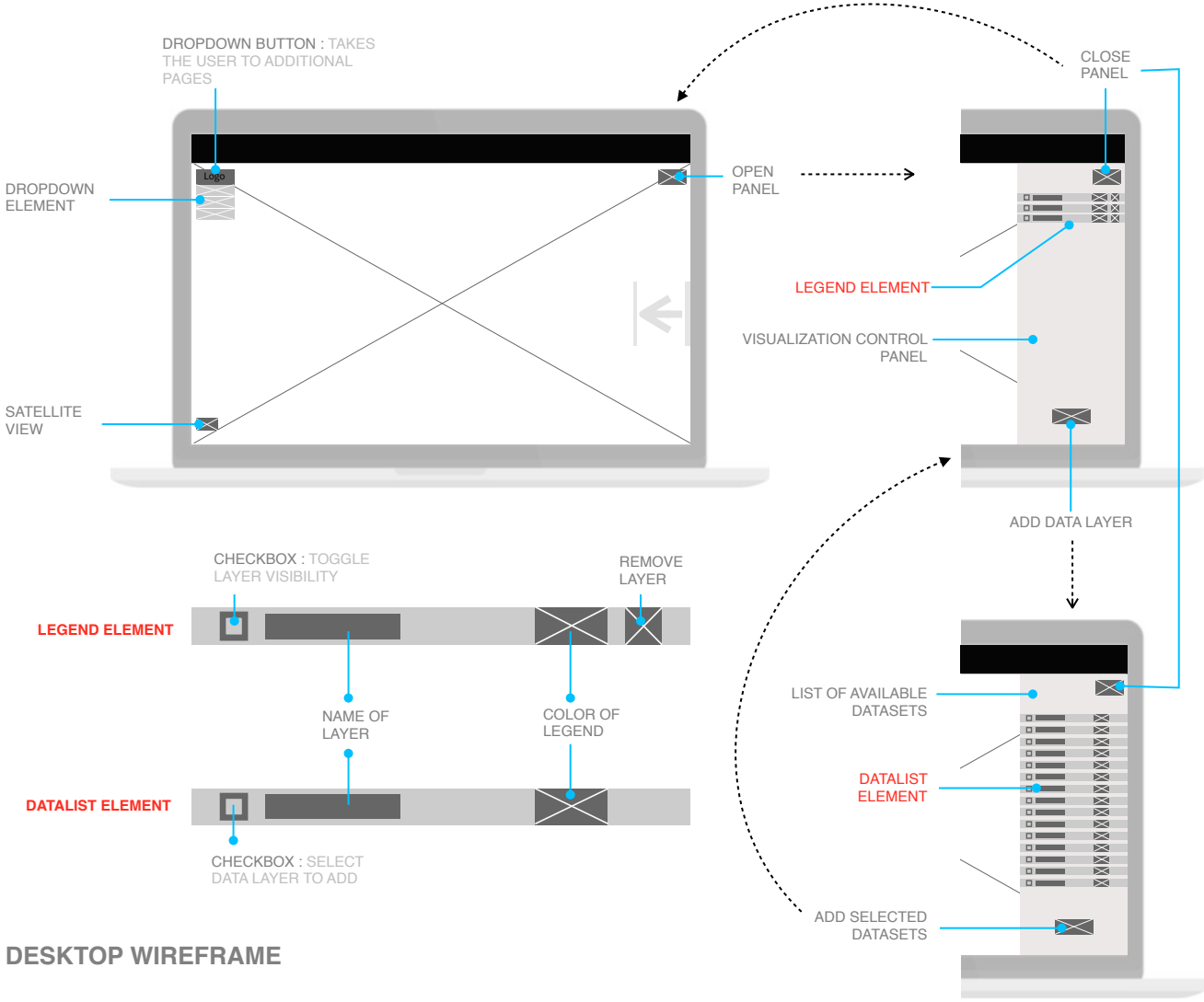
SOURCE: GOOGLE IMAGES

METAPHOR

Wireframe

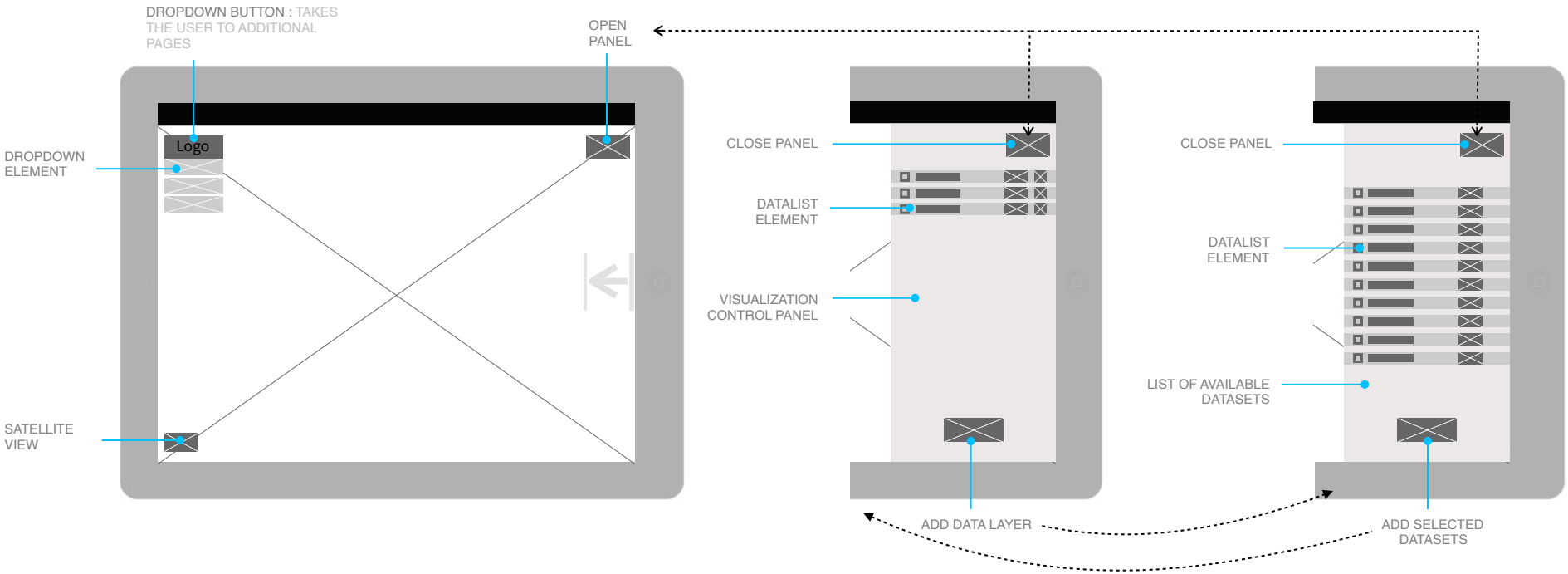
A low fidelity prototype of the tool for desktop, mobile and tablet form factor was created to evaluate it against the design decisions made. The following key design decisions were made while creating the prototype :

- The position of the Visualization Control Panel (VCP) should be on the right since most of the users are right handed[1].
- All buttons on the map are to be floating buttons to give more screen space to the visualization.
- The VCP should have an option of being hidden to give more screen space to the visualization. This is specifically important in case of form factors like the mobiles and tablets (vertical).
- All controls to create the visualization should be a part of the VCP.
- All buttons on the Visualization Frame should be as non intrusive as possible.

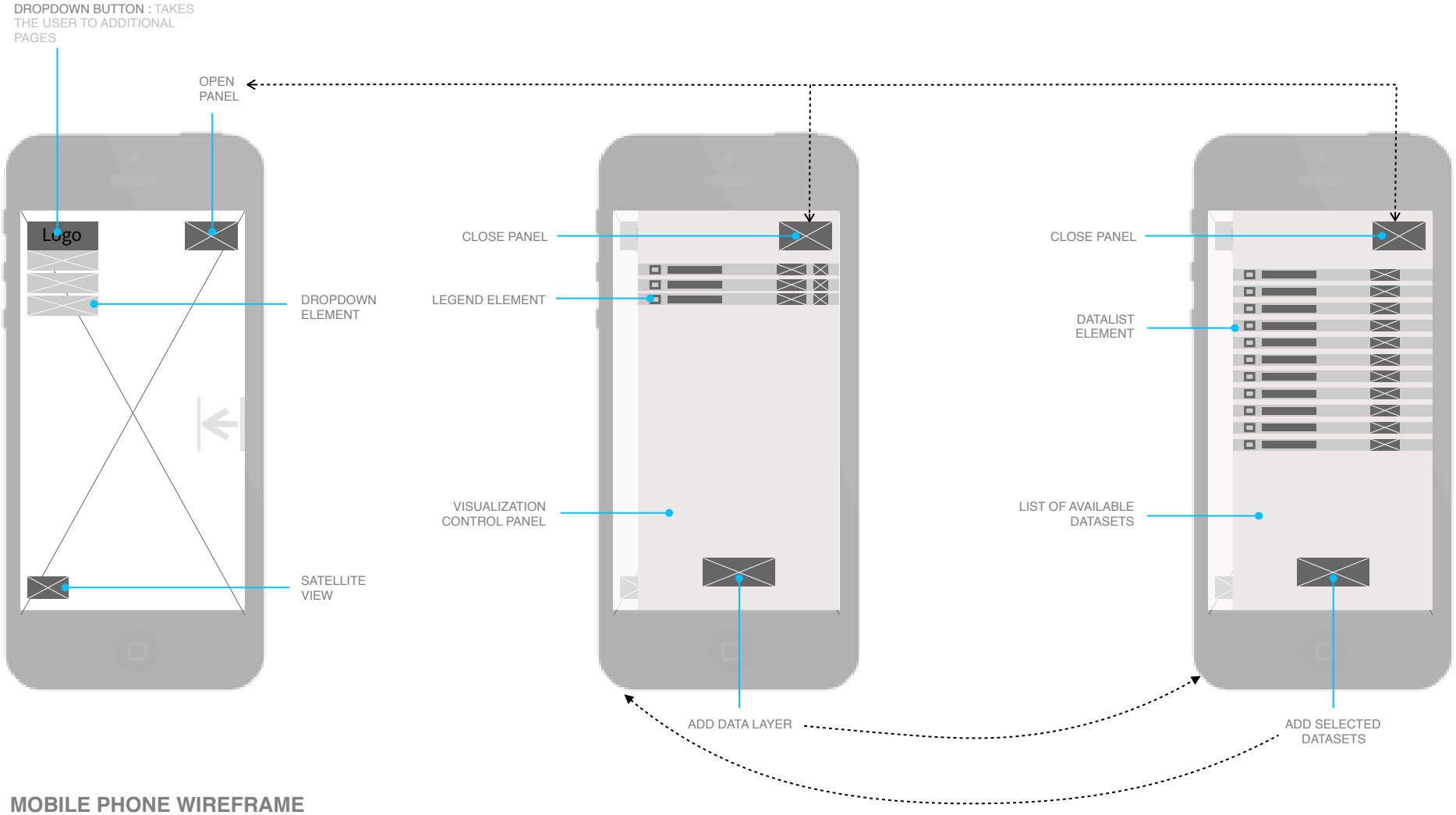


DESKTOP WIREFRAME

1 Most humans (say 70 percent to 95 percent) are right-handed, a minority (say 5 percent to 30 percent) are left-handed, and an indeterminate number of people are probably best described as ambidextrous. - www.scientificamerican.com/article/why-are-more-people-right/



TABLET WIREFRAME



MOBILE PHONE WIREFRAME

6.7.Heuristic Evaluation

In order to check the usability of the tool, a heuristic evaluation was conducted with a group of 4 students studying Interaction design at IDC IIT Bombay. The students were asked to use the wireframe tool for half an hour and then evaluate the tool on its usability.

Key Insights

The evaluation provided certain key insights into the usability of the tool. The following are the key insights which were imperative to the smooth functioning of the tool.

Center the Map

The MumbaiData Visualization Tool uses a Google Map as its base layer. After using the tool for a while, users may get lost in the map or pan too far away from the view of Mumbai. To go back to the original position, the users would have to find the location of Mumbai on the map or refresh the tool thereby losing any visualization already created. This could be easily solved by a 'Go Home' button that centers the map over Mumbai in its original position without the user having to lose any visualization created.

Colorblind Friendliness

Many of the data layers were found to be not colorblind friendly after conducting an Ishihara test on the data layers. It is important for the tool to be accessible to most of the users[1].

Sharing of Visualizations

The tool enables the creation of visualizations about Mumbai. However, it was found that there is a lack of completion to the act of creating a visualization without enabling the user to share it. In an interconnected world, sharing is of utmost importance.

Change appearance of the base map

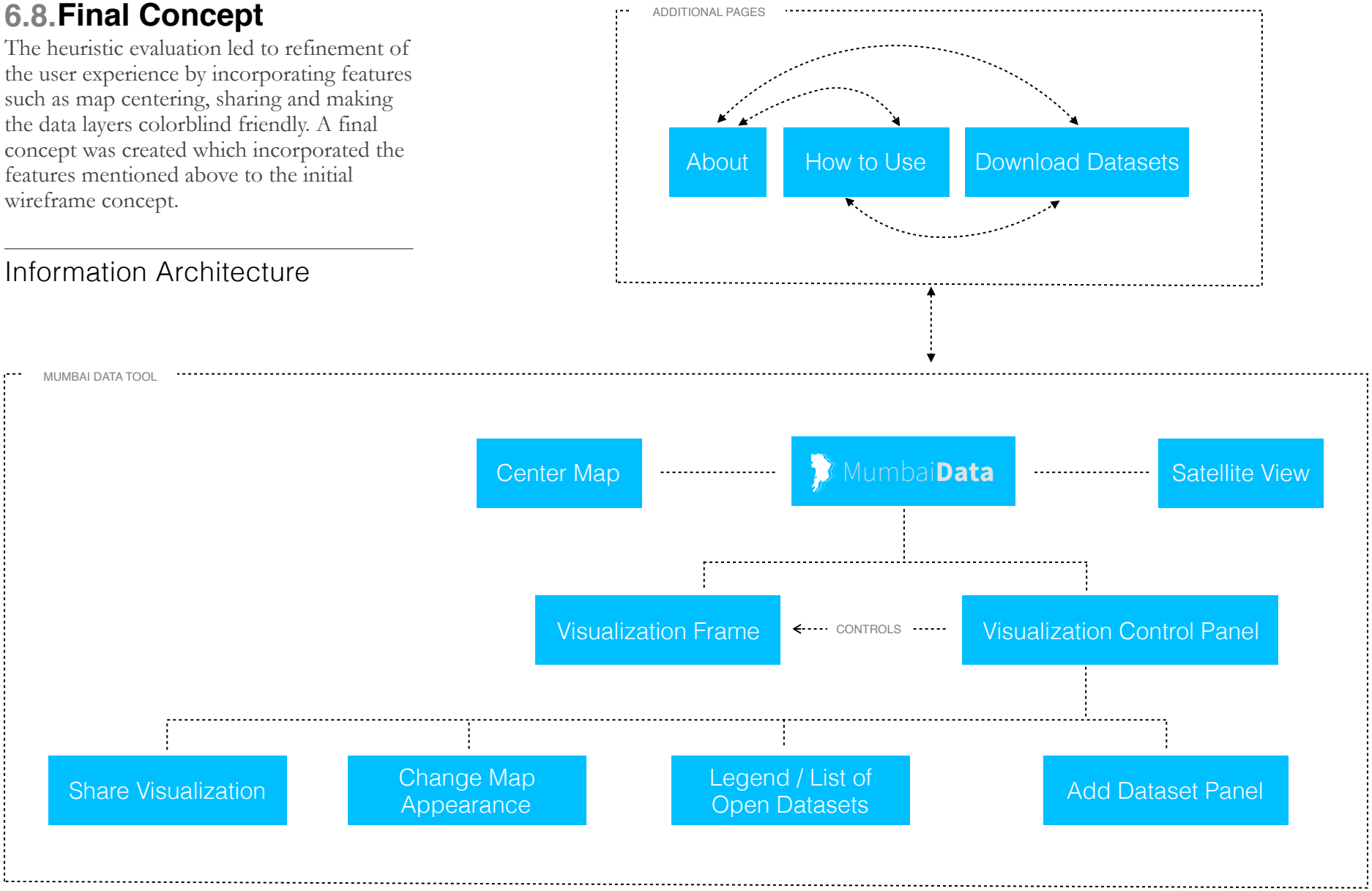
The MumbaiData Visualization Tool consists of a styled google map as the base map. Advanced users who are familiar with most of the tool's functionality might require to change the appearance of the base map in certain cases and for certain scenarios. A functionality could be provided to change the basic appearance of the base map.

1 About 8 percent of males, but only 0.5 percent of females, are color blind in some way or another, whether it is one color, a color combination, or another mutation. - Wikipedia

6.8.Final Concept

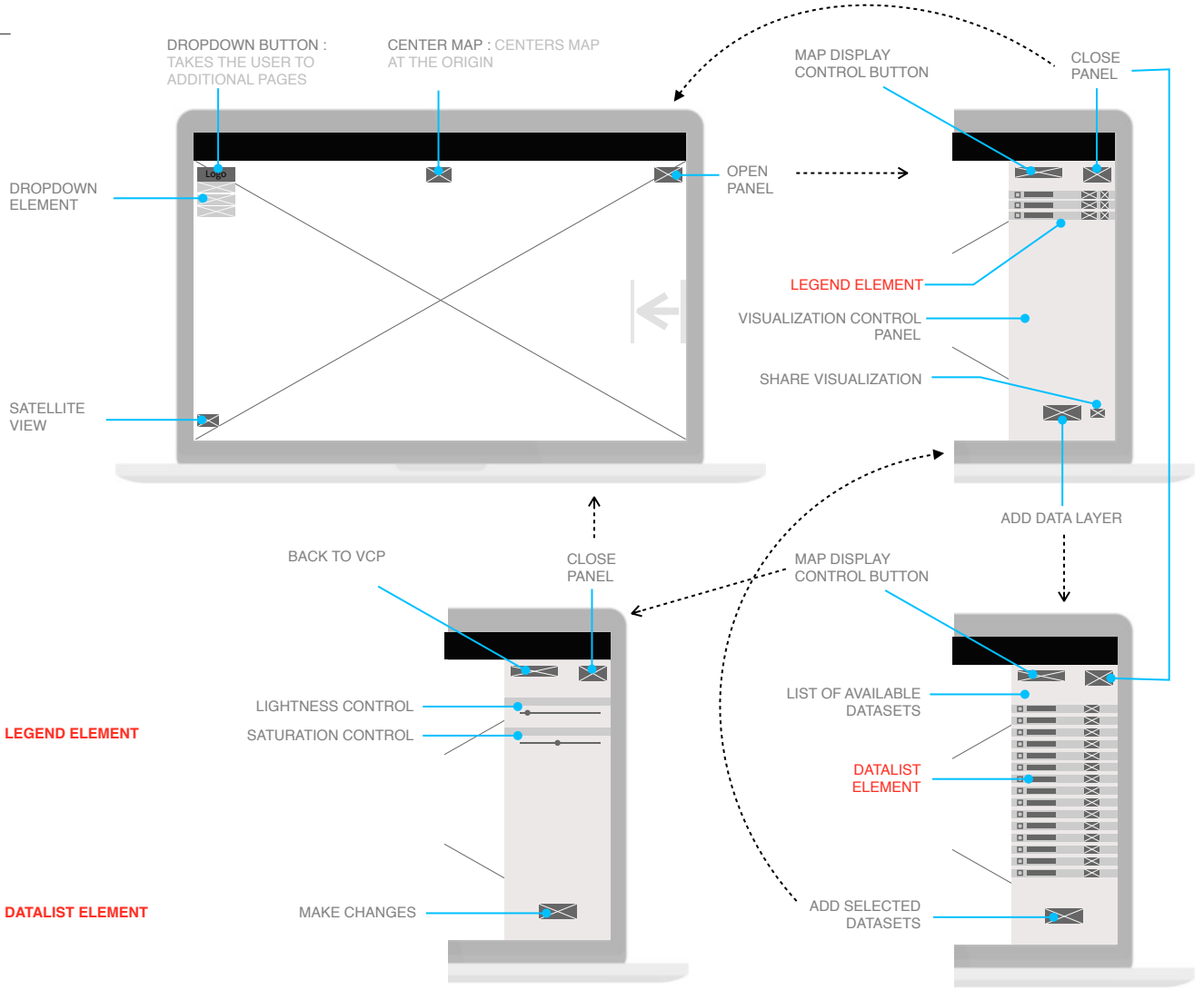
The heuristic evaluation led to refinement of the user experience by incorporating features such as map centering, sharing and making the data layers colorblind friendly. A final concept was created which incorporated the features mentioned above to the initial wireframe concept.

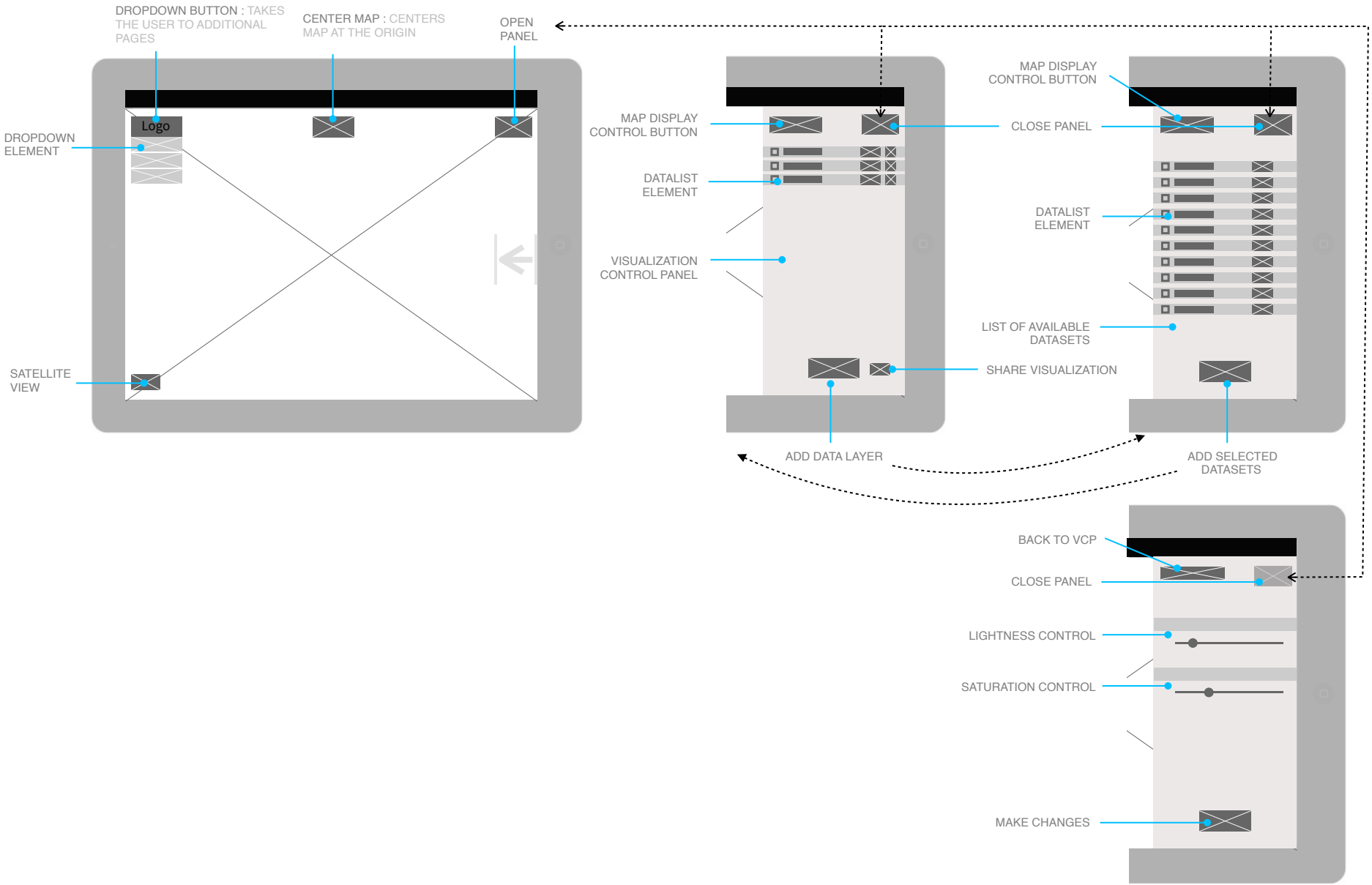
Information Architecture



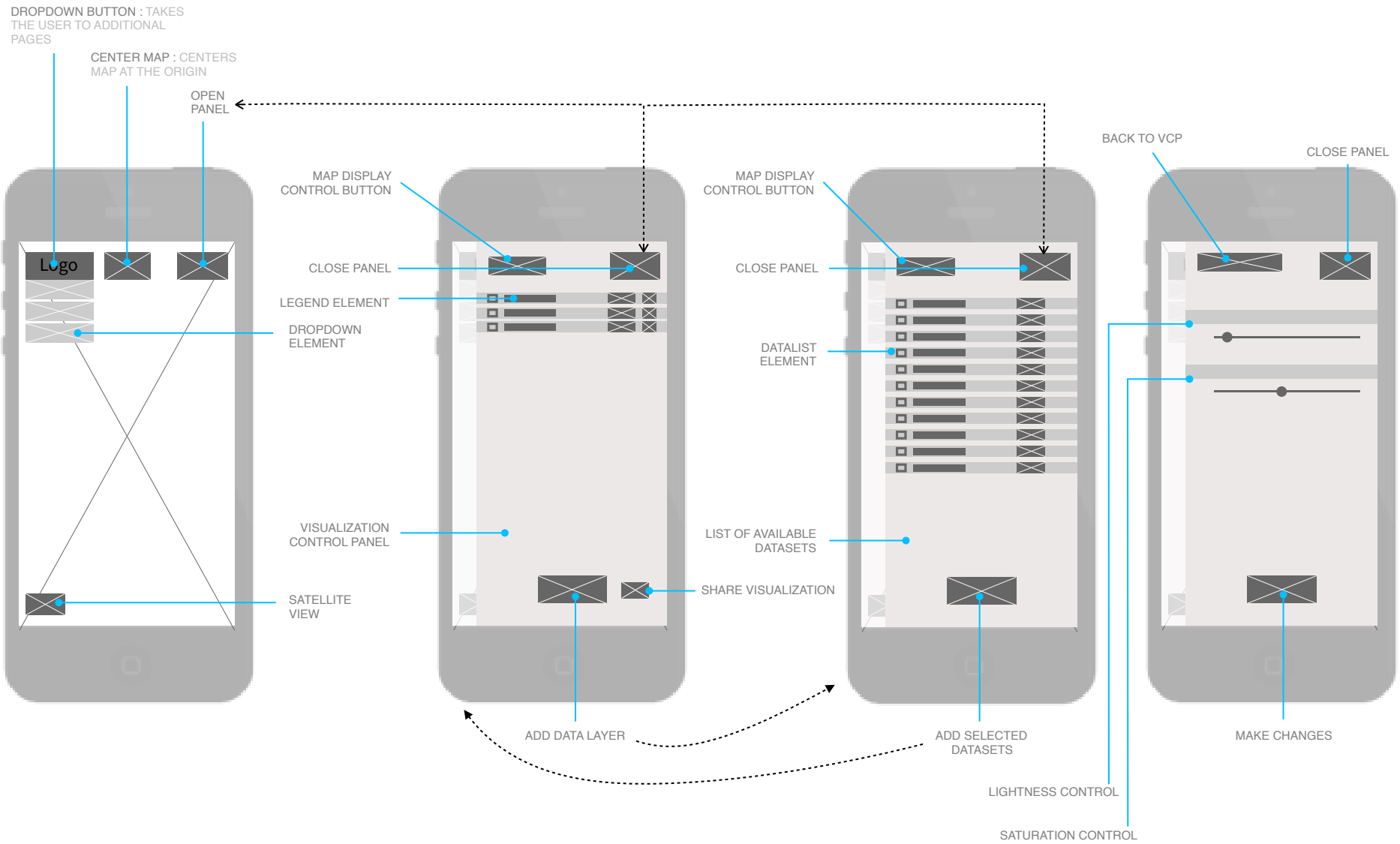
Wireframe Final Concept

DESKTOP WIREFRAME



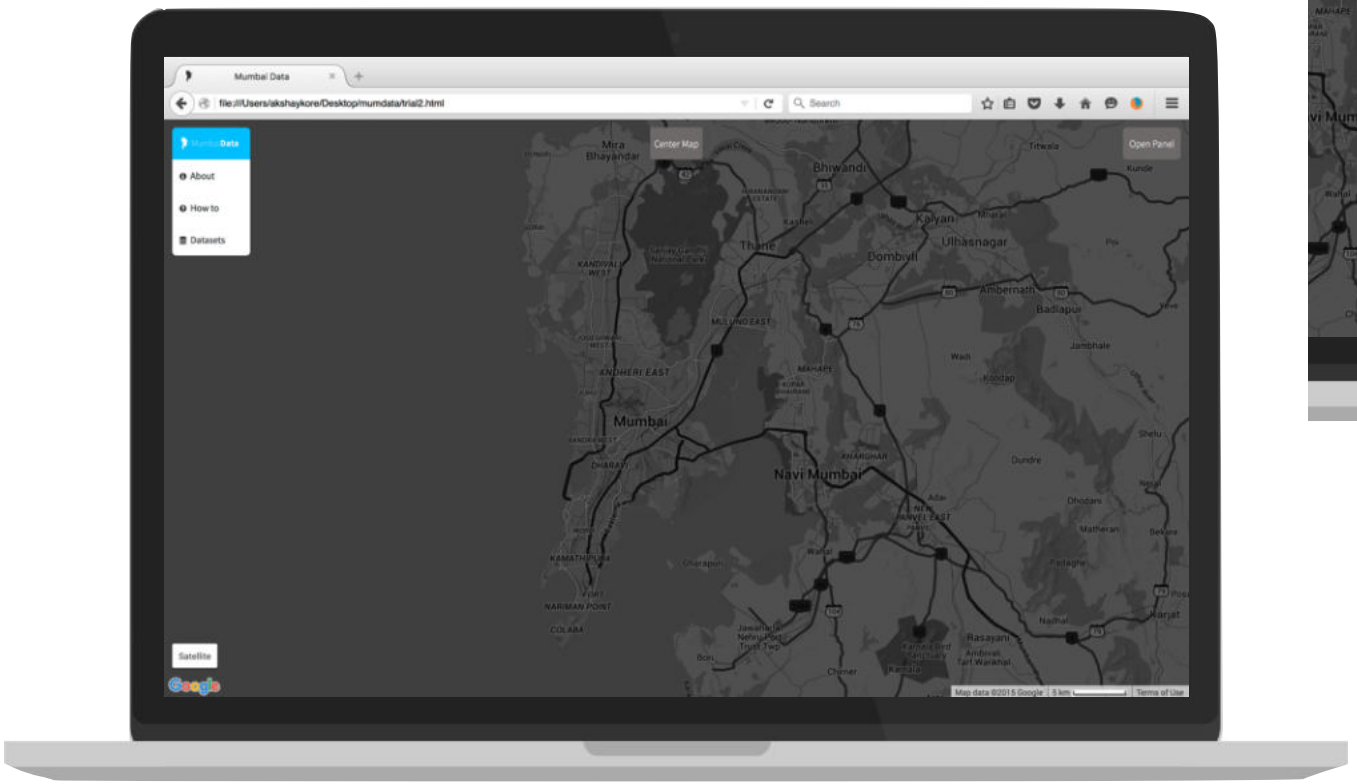


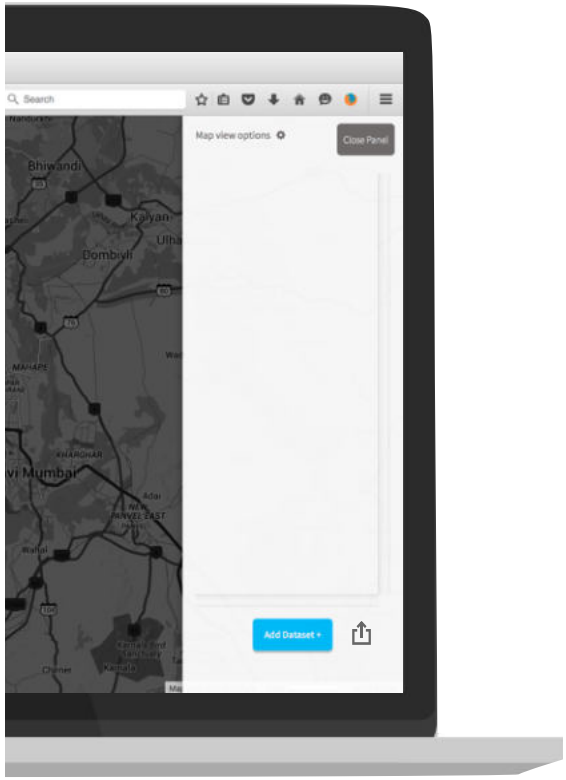
TABLET WIREFRAME



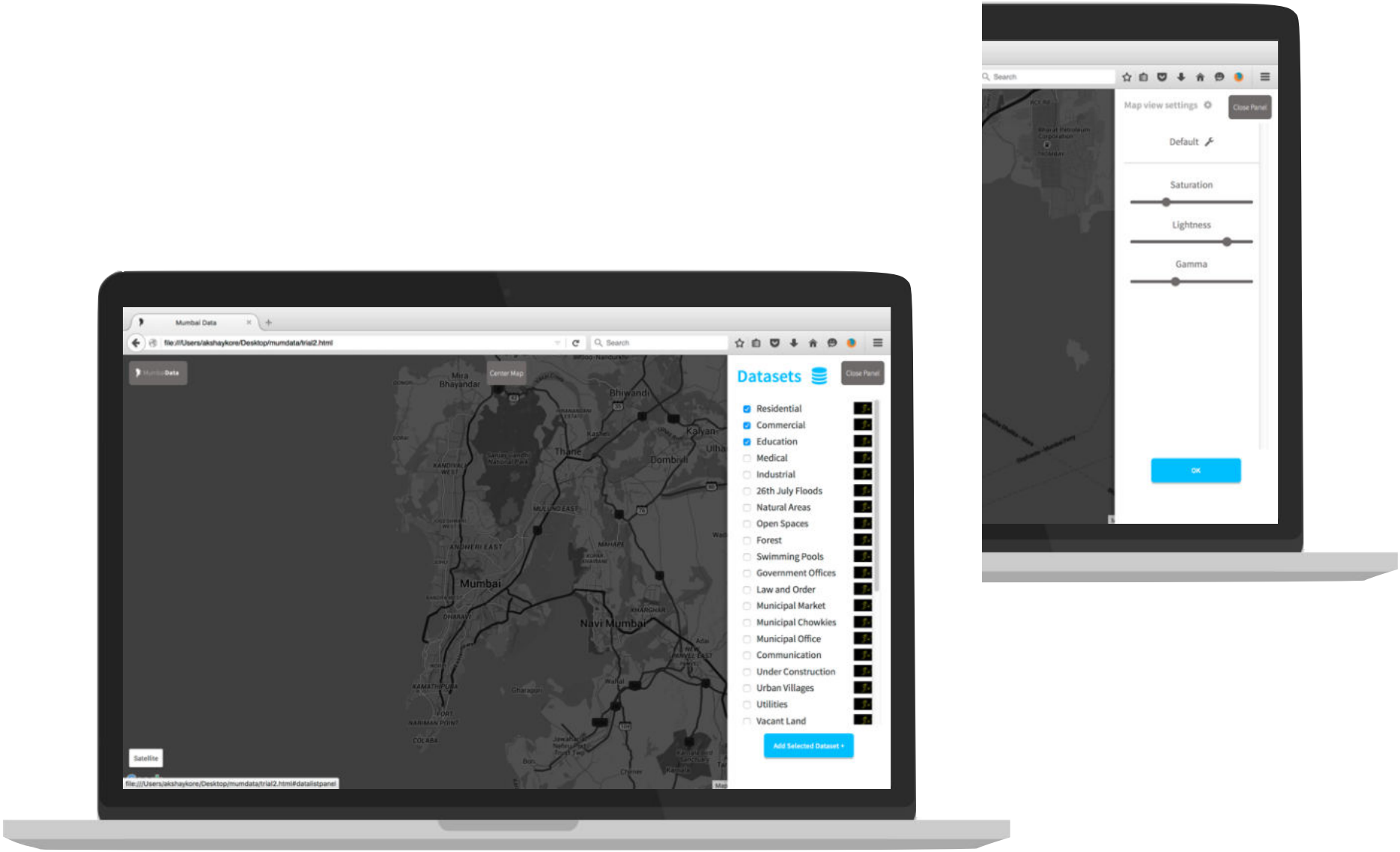
MOBILE PHONE WIREFRAME

User Interface Screens

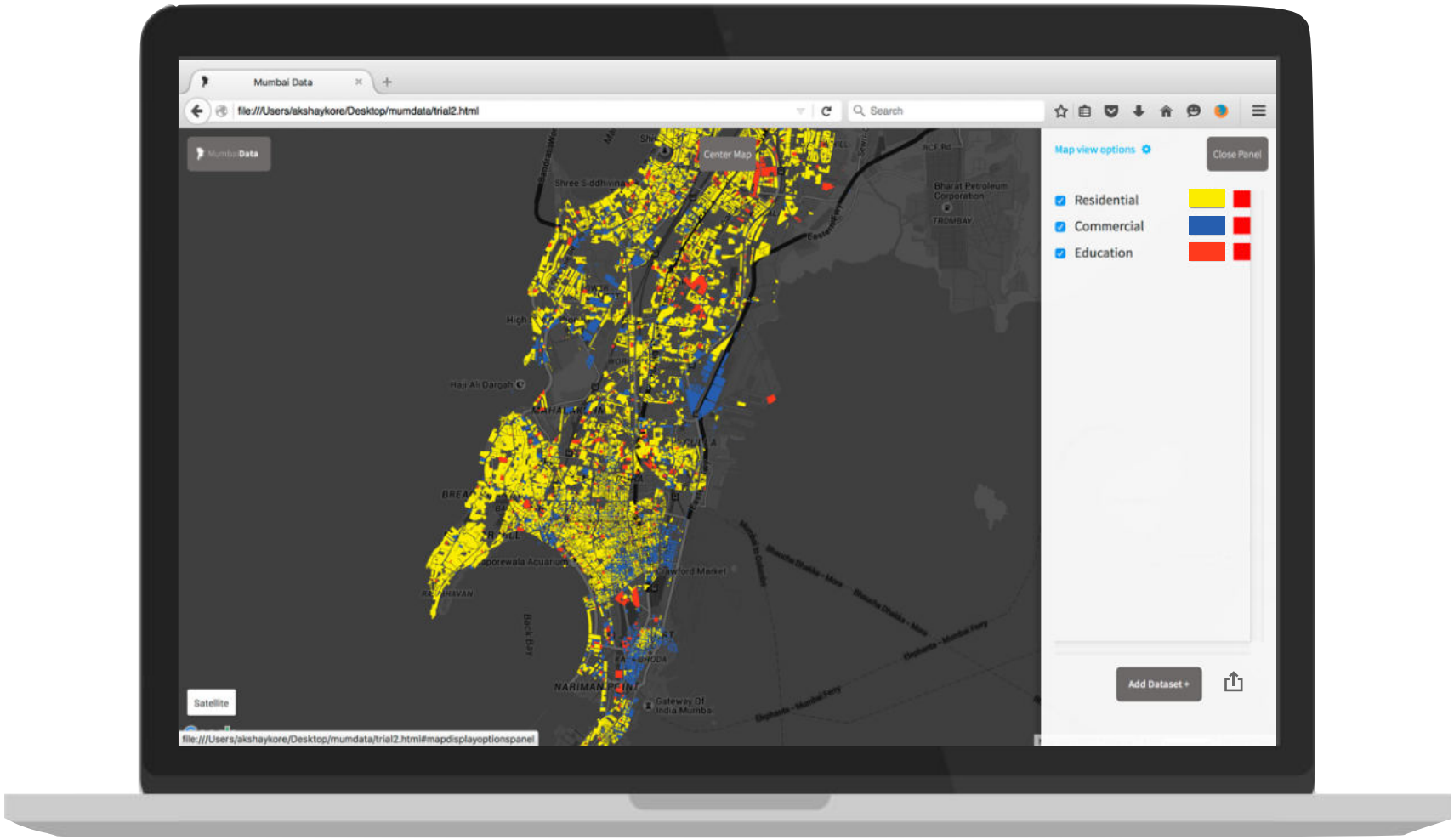




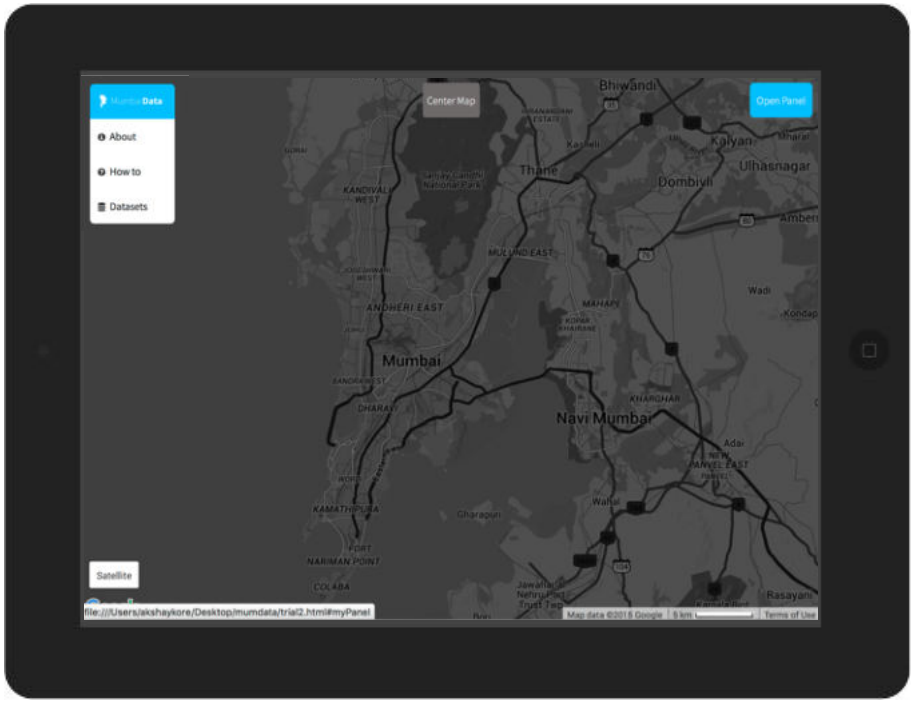
DESKTOP INTERFACE

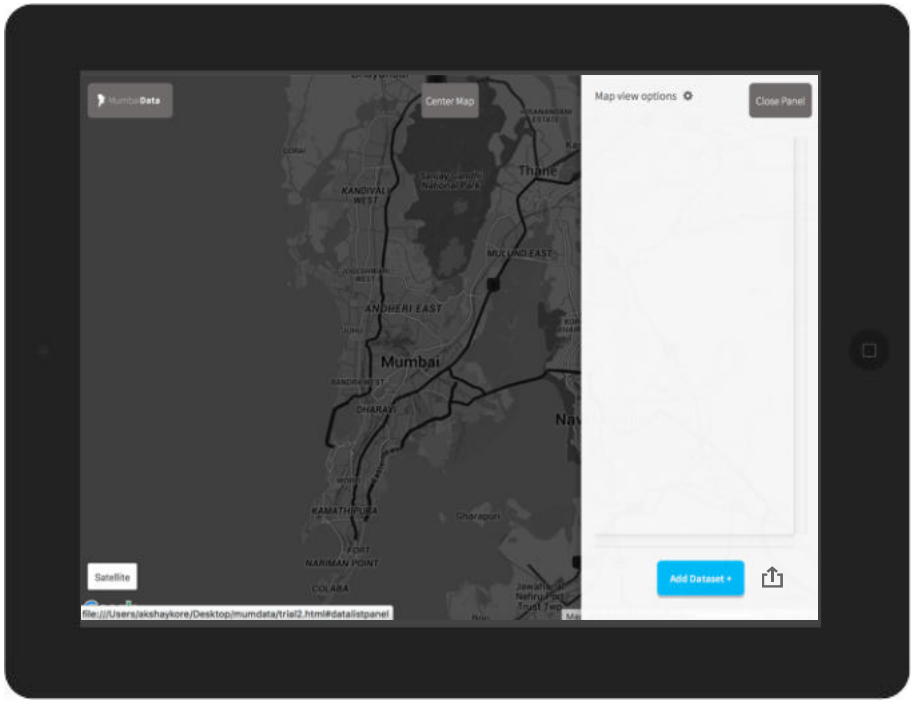


DESKTOP INTERFACE

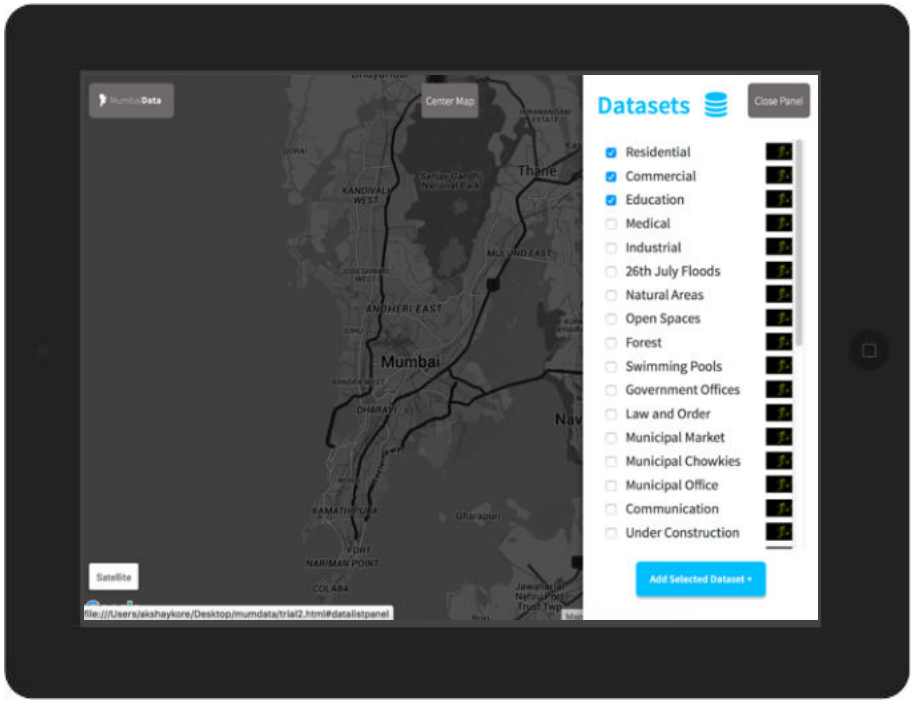


DESKTOP INTERFACE

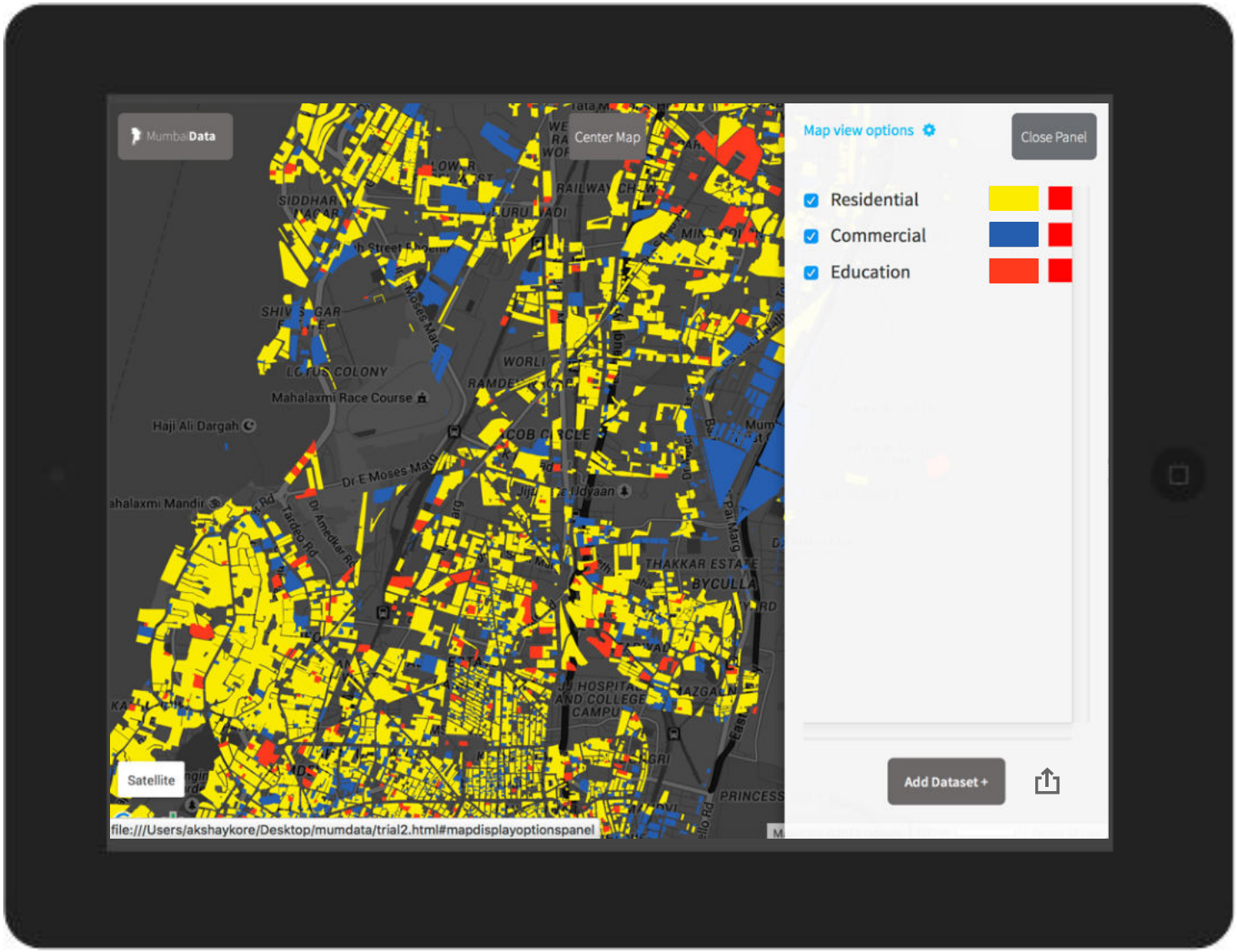




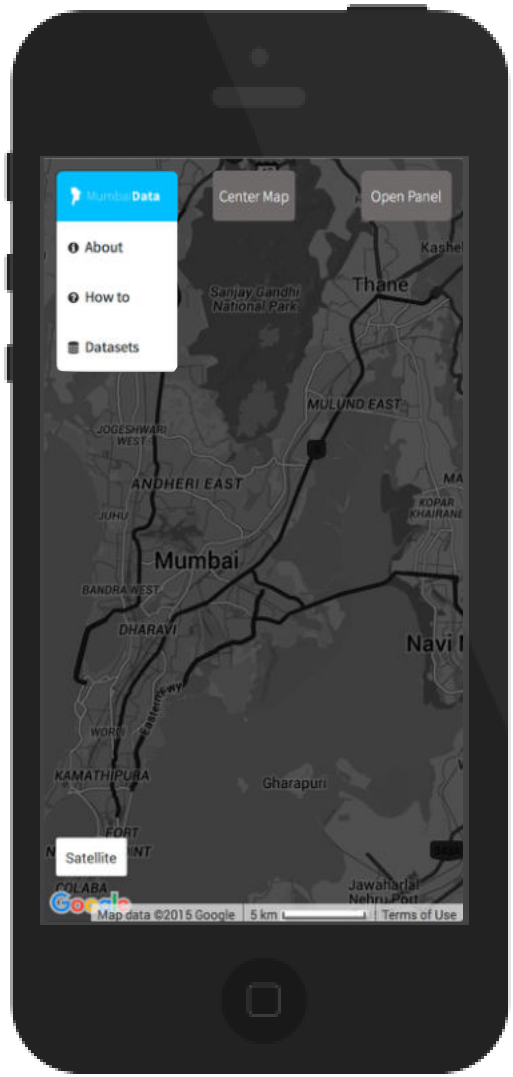
TABLET INTERFACE

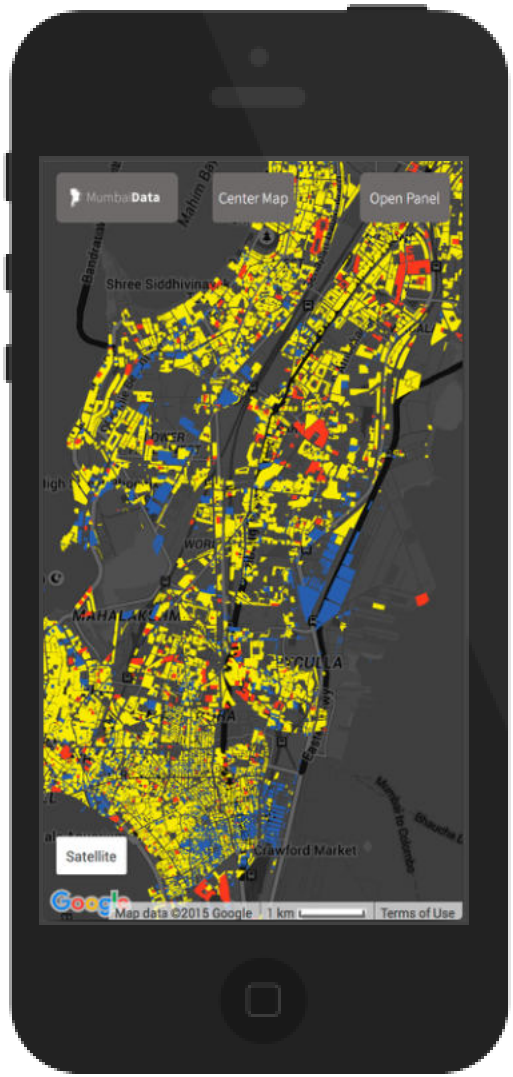


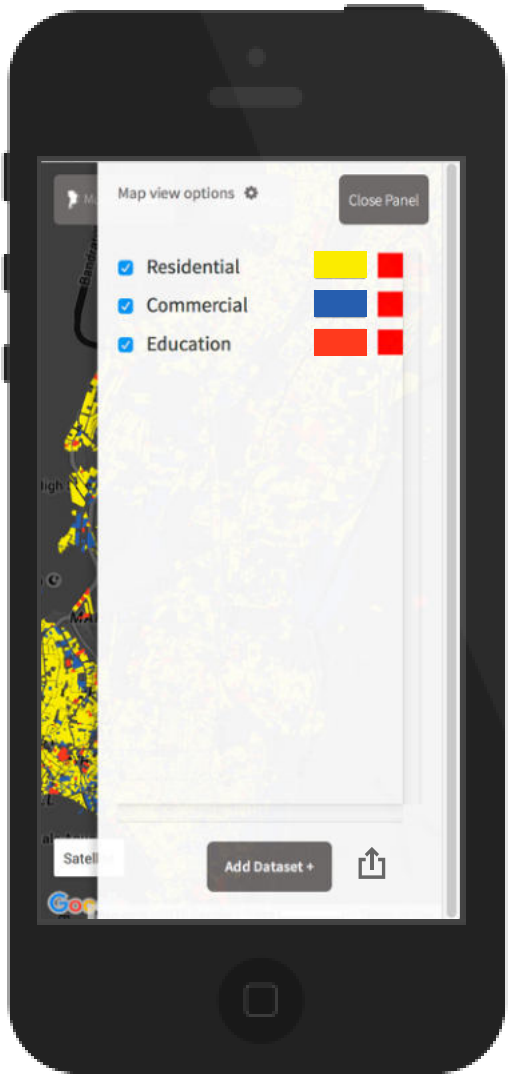
TABLET INTERFACE



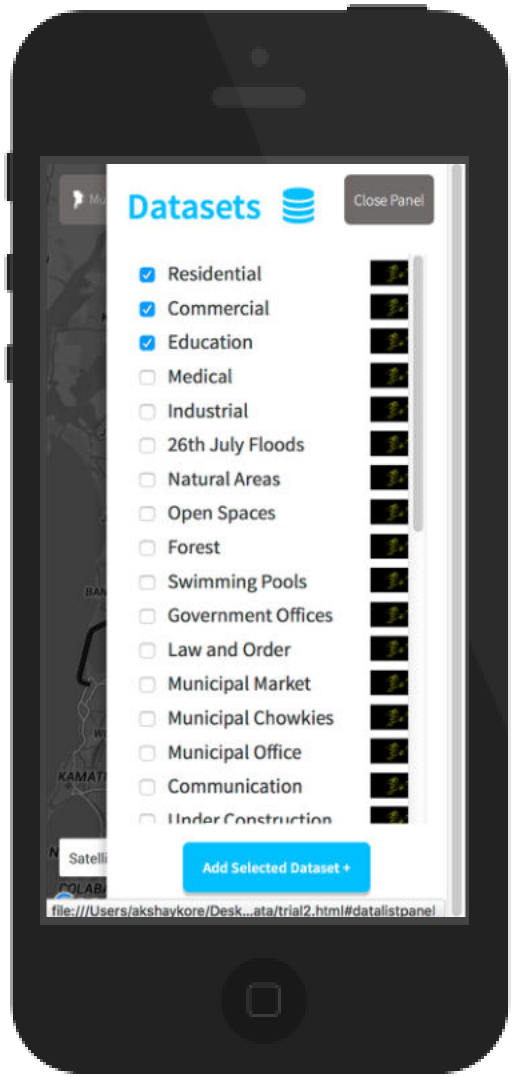
TABLET INTERFACE

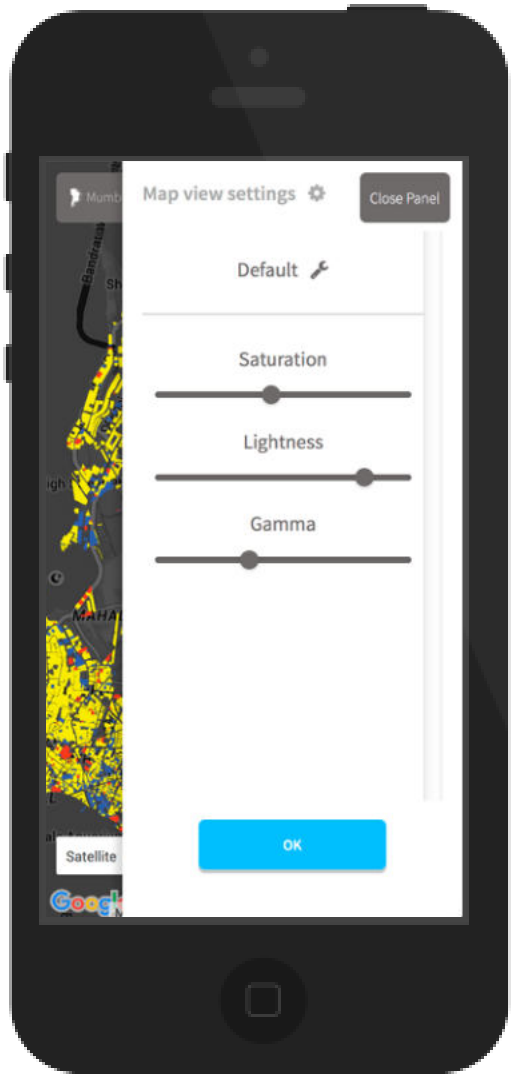






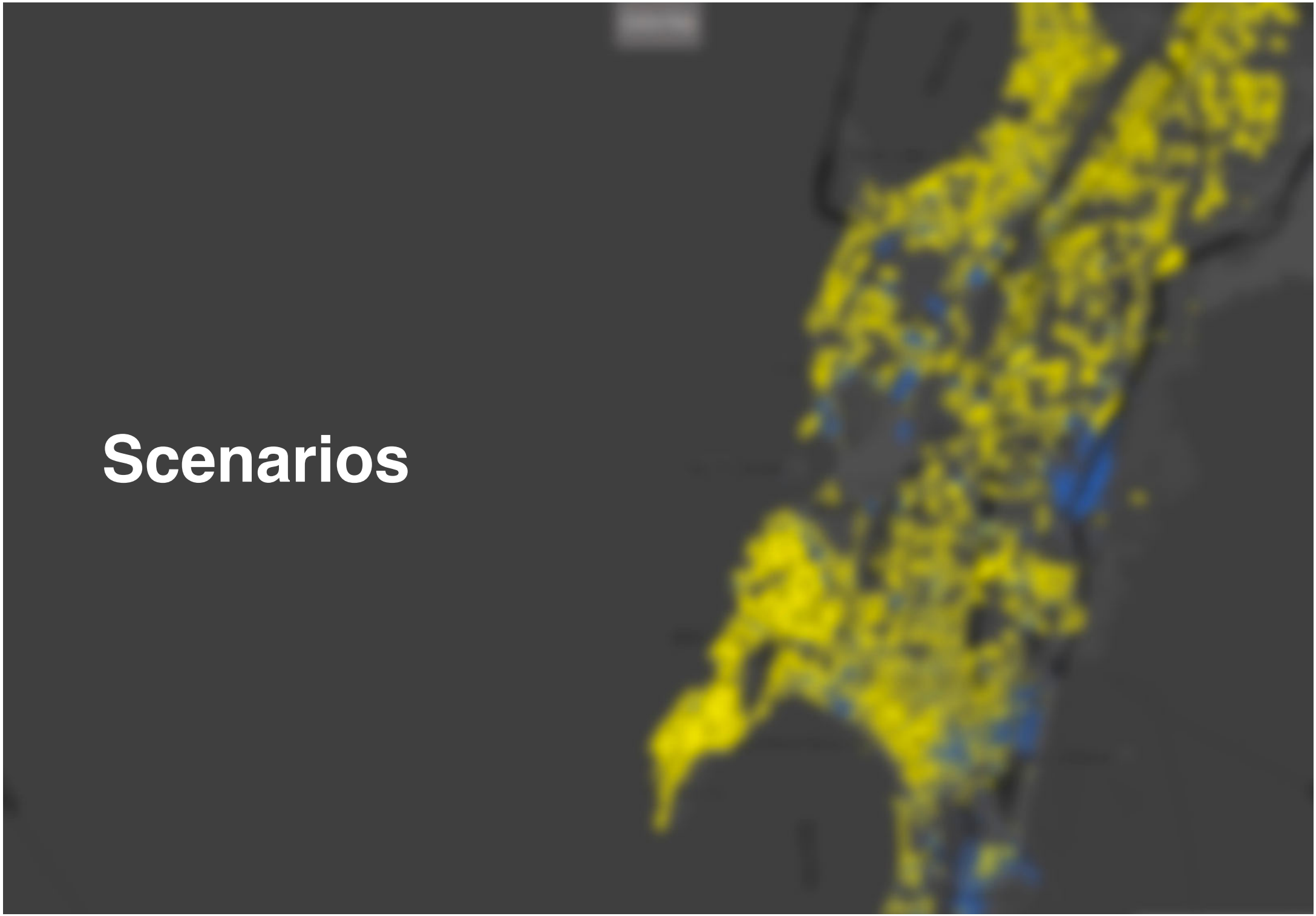
MOBILE PHONE INTERFACE





MOBILE PHONE INTERFACE

Scenarios



7. Scenarios

The tool may be used by decision and policy makers in public institutions like government departments and public sector banks, private organizations like private sector banks and private companies as well as NGOs and individuals. The scenarios assume certain level of knowledge about working with data. Many of scenarios assume availability of datasets even though the tool may not contain these datasets in the present.

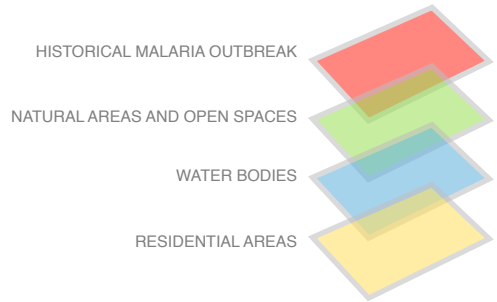
7.1. Scenario 1 - Malaria Breeding Grounds

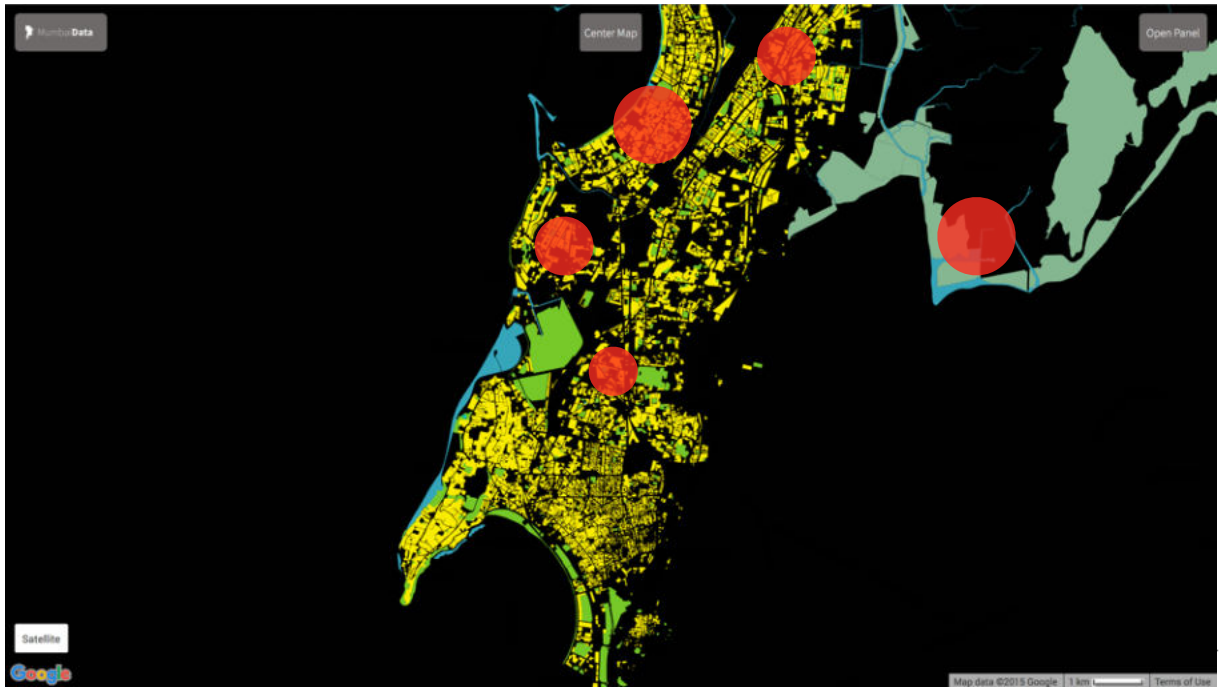
There is an outbreak of malaria in certain parts of the city and the health department wants to find the possible breeding grounds for mosquitos to take proper action. An officer in the health department decides to map possible locations in the city where mosquitos could be breeding. Mosquitos breed near stagnant water sources. He opens the MumbaiData tool and visualizes all the water bodies in Mumbai. He looks at the areas most affected by malaria in the past. He also overlays a forest, mangroves and open spaces areas in the city since these are also possible breeding grounds for mosquitos.

An overlay of the above features reveals the possible locations and breeding grounds for mosquitos for the affected areas. The health department officer can also issue preventive measures in other areas in the city with similar characteristics. An overlay for all residential areas may be created to inform citizens living in these localities about the outbreak and preventive measures may be taken.

Malaria Breeding Grounds

Data Layers



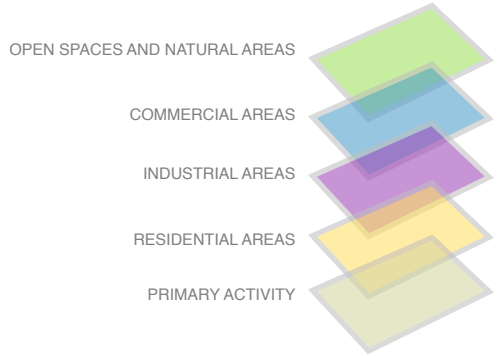


7.2. Scenario 2 - High Pollution Areas

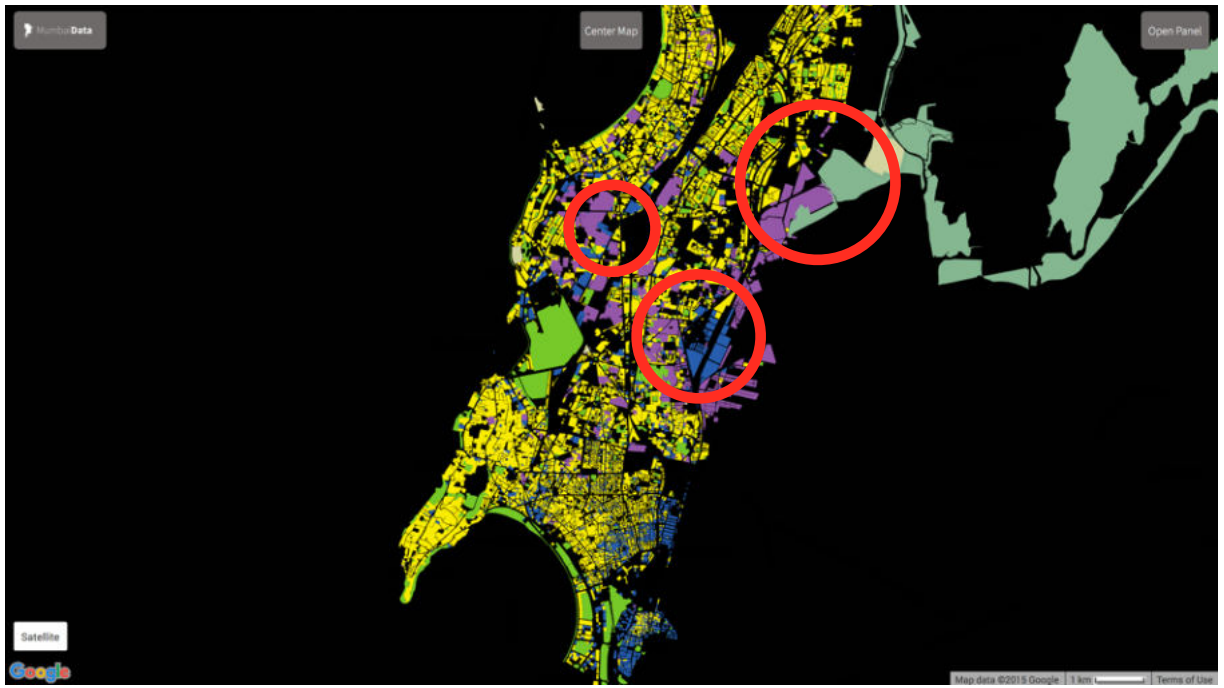
Pollution levels are on the rise in the city. The Pollution Control Board needs to find out the possible areas where there is an alarming level of pollution. The board has limited amount of resources and a survey of the entire city would be highly time consuming and expensive. A senior official in the department decides to find out possible areas where it makes the most sense to conduct a survey and enable efficient allocation of the limited resources.

He uses the MumbaiData Tool to overlay areas with higher density of residential, commercial, primary activity like cowsheds , etc., and industrial areas along with high traffic density roads. Open spaces and natural areas are overlaid on the map. These overlays generate a pattern of density and areas near natural areas and open spaces and near roads with a low density of traffic are discarded from the areas to be surveyed. Certain pockets in the city are found to be most susceptible to high levels of pollution. After shortlisting these pockets, the official deploys people and resources to survey these areas.

Data Layers



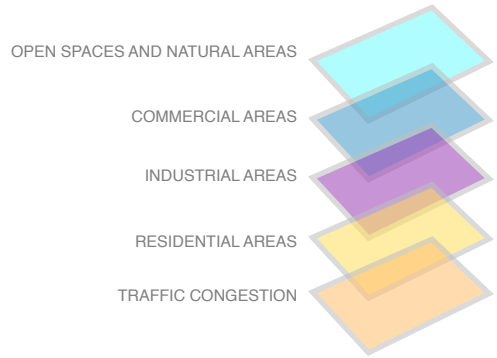
High Pollution Areas



7.3. Scenario 3 - Traffic Management in Floods

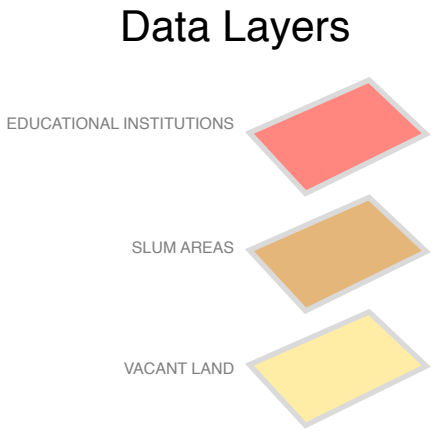
Every year during monsoon, certain areas in the city get flooded and cause havoc with the traffic infrastructure. This was identified as a recurring event and new infrastructure would take time to build. It is the month of April and monsoons are just around the corner. Building infrastructure solutions is a much longer effort. The Municipal corporation is tasked with the massive effort of managing the traffic during the floods scenario. A planner working with the Municipal corporation decides to use the MumbaiData tool to do a preventive planning for the traffic in the city during floods.

The planner visualizes the flood prone areas in the city along with major centers of employment like commercial and industrial areas along with all the transport infrastructure like roads, railways and airport. The planner also overlays the residential areas, slums. The knowledge of the location of areas where people live and areas where people work may help in predicting the density of flow of traffic during different times of the day. Traffic densities are also overlaid. This enables the planner to find the hotspots for high traffic congestion in flood prone areas and helps the Municipal corporation to divert the traffic during such scenarios. It would also help the Municipal corporation to decide the location of pumps for removing the flood water.



7.4. Scenario 4 - Building a School for Underprivileged

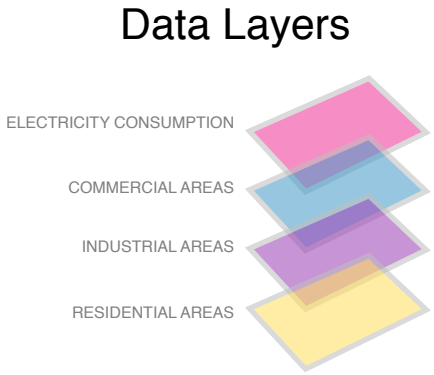
The CSR Wing of a multinational corporation decides to build a school for underprivileged children. They have recently expanded to Mumbai and are unaware of the on ground realities. This corporation employs a consultant to find the most appropriate location to build the school. The consultant uses the MumbaiData tool to find out where most of the underprivileged children are located. He overlays the housing prices in the city on the map along with location of slums. With real estate prices skyrocketing in the city, it makes sense to find out the informal settlements in the city rather than the most expensive real estate. Location of vacant plots near the slums is identified through the tool. He then overlays the number of schools and other educational institutions in and around these slums and discovers that there are an ample amount of schools near slums. Along with this data for the number of dropouts in the locality is checked and serious gaps are found between the dropout rates and number of school seats. It is concluded that it would not be beneficial for the underprivileged children to have one more school in their locality, rather the funds and resources could be diverted towards initiatives like employing more teachers or renovating the existing schools or providing incentives to students to attend school.



Building a School for Underprivileged

7.5. Scenario 5 - Anomalies in Electricity Consumption

With the advent of Internet of Things devices, static data is of little use when real time data can be made available. Governments can use realtime data from buildings on parameters like water and electricity usage to monitor their consumption in the city. Anomalies can be found out in the consumption by correlating consumption with building type. The average electricity and water consumption of a residential building will be lesser than an industrial structure. An unnatural spike in the consumption may be looked into by the government authority to investigate leakage or theft of water and electricity.



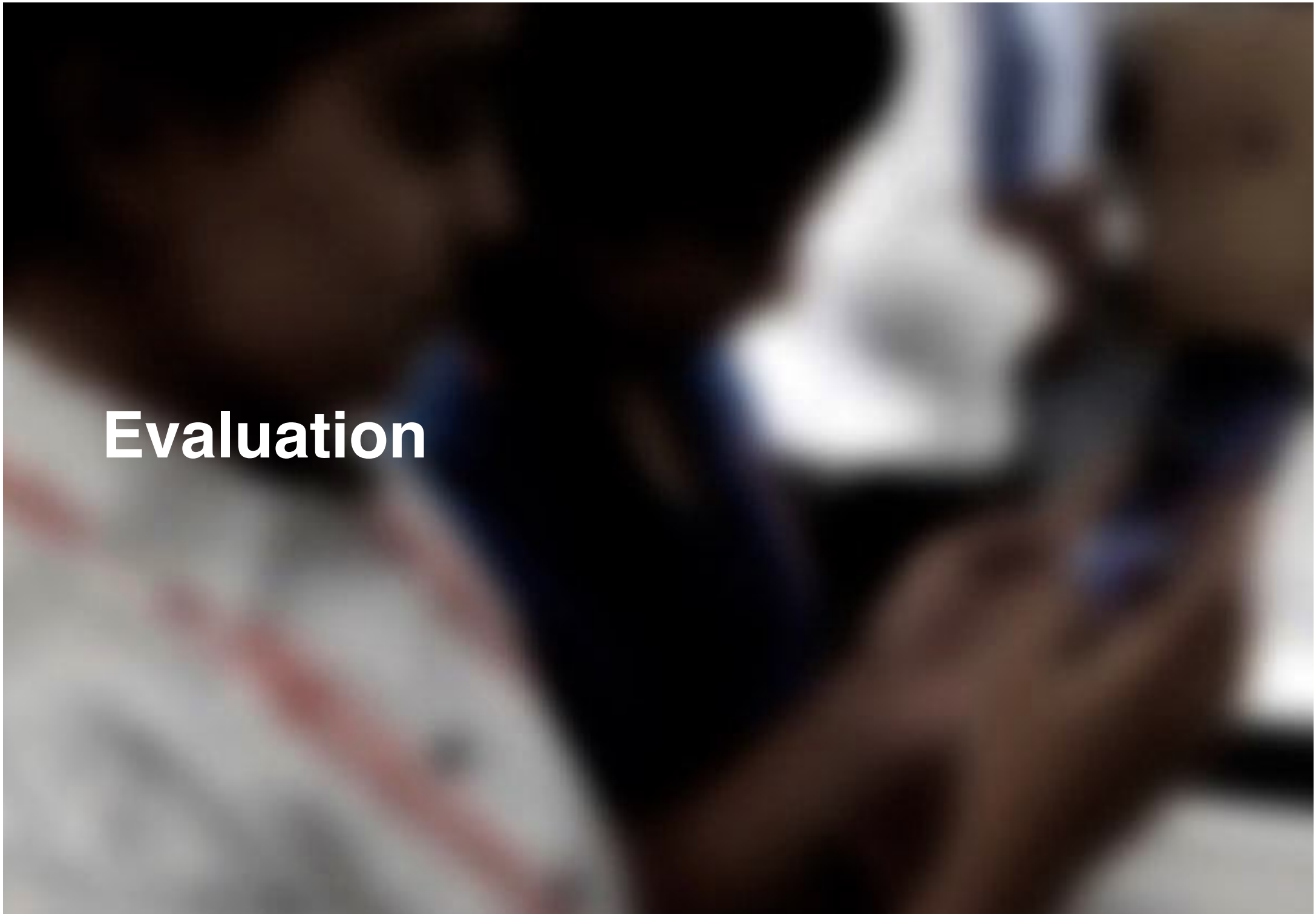
Anomalies in Electricity Consumption

8. Limitations

No design is perfect, and as such the MumbaiData Tool also has its limitations.

- There is a need for the users to be familiar with certain data terminologies before using the tool.
- The tool can only be used if the user has access to the internet and cannot be used offline at present.
- The user cannot upload his own datasets easily and as such there is a need of technical knowledge in programming for the user to do so.
- There could be a slight error with the positioning of the layers on the map, however the tool does not aim at accuracy rather it embraces the messiness inherent in the data sets.

Also limitations are not permanent and many of the limitations mentioned can be solved in the future updates made to the tool.


Evaluation

9. Evaluation

The purpose of evaluation for the project is to understand the usability, feasibility and conceptual issues with the MumbaiData Visualization Tool on multiple form factors.

9.1. Objectives

- To evaluate the usability of the add and remove dataset functionality on a mobile, desktop and tablet form factor.
- To evaluate the usability of the user interface.
- To check the find-ability of the datasets.
- To examine the utility of the MumbaiData Tool.

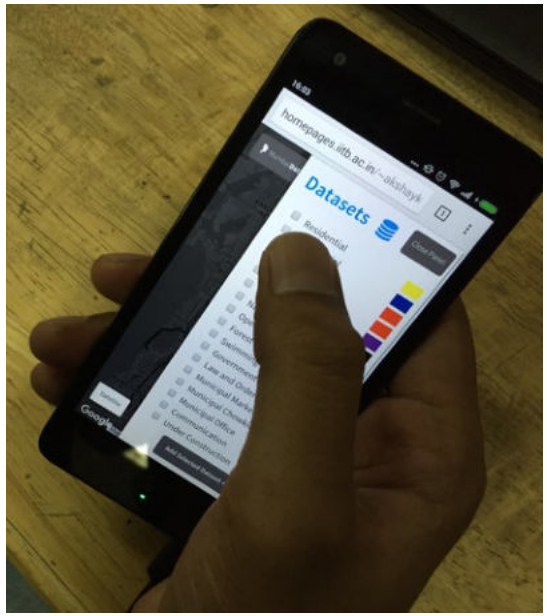
9.2. User Testing Protocol

A think aloud test was conducted for the user testing. As the name suggests the participants have to use the tool while continuously thinking out loud, or verbalizing their thoughts as they move through the interface or perform the tasks given.The usability test is conducted with multiple users. Each user is given a single type of form factor device to use the tool. In case of mobile phones, the tool is tested with the mobile phone held vertically. In case of tablets, a horizontal orientation is used to test the tool. It is assumed that the users have a knowledge of english and know a bit about the datasets beforehand.

Type of form factor	No. of Users
Desktop	4
Mobile	4
Tablet	4
Total	**12**

Following is the protocol for the Think Aloud Test:

- The user was asked to Open and Close the Side Panel.
  This is to familiarize the user with the Visualization Control Panel.
- The user was asked to add any dataset of his choice.
  This is an early success task to instill confidence in the user.
- To check the find-ability of the tool, the user is given a slightly complex task.
  - The user is told to Adding Residential, Commercial and Water Bodies dataset. This ensures find-ability of the datasets.
  - The user is then asked to hide the Residential and remove Commercial layer from the visualization.
    This is done to check the usability of the show/hide and remove buttons.
- The user was then asked to convert the base map to its satellite view.
  This is done to check the find-ability of the satellite view button.
- The user was asked to Zoom into his place of residence and go back to the default view.
  This is done to check the usability of the center map button. If the user manually zooms out, then the usability of the functionality is in question.



9.3. Insights : Think Aloud Test

- Users found it easy to use the application for simple visualization tasks once a pattern was found.
- Most users found it difficult to find and use advanced features like center map and map view settings.
- It was found that the tablet form factor tested the best with users followed by desktop and mobile.

Task	Success value out of 4		
	Desktop	Mobile	Tablet
Add any dataset	3	1	2
Add specific datasets.	4	4	4
Hide one dataset and remove another from the control panel	3	3	4
Convert the map into a satellite view	4	4	4
Go back to the default view	0	0	1
Zoom in to powai lake and go back to default zoom and center.	3	2	3

THE ABOVE TABLE SHOWS THE SUCCESS VALUE OUT OF 4 FOR DIFFERENT TASKS THE USERS WERE ASKED TO PERFORM IN THE THINK ALOUD TEST WITH DIFFERENT DEVICE FORM FACTORS. THE HIGHEST VALUE BEING 4. A TASK WAS CLAIMED TO BE UNSUCCESSFUL IF THE USER TOOK MORE THAN 30 SECONDS TO COMPLETE THE TASK.

9.4. Insights : Feedback Survey

Apart from the Think Aloud Test, a contextual feedback was taken from the users. The users would be asked to identify scenarios in which they would use the tool. Whether there were any major conceptual flaws with the tool. What would be a preferred medium to use the tool? A total of 19 users were interviewed.

- All the users found the idea very interesting.
- 14 out of 19 users found the tool to be easy to use.
- 15 out of 19 users said that they would be willing to use the tool.
- 12 out of 19 users found the absence of the sharing / download visualization feature to be a major problem.
- One user said that the tool is good for doing quick research.
- There was a problem of accuracy that was noticed by 9 out of 19 users.
- All users suggested additional datasets to be included in the tool.

10.Future Work

No project is complete, as such MumbaiData too has a wide scope for improvement. Following is the future work that is proposed for the tool.

- Multiple Language Support
- Participatory Data Creation
- Sharing and exporting of the visualization
- Personalization of the visualization by creating user profiles.
- Include Non-Spatial Datasets like numerical data.
- Add user profiles and saved states.
- Combining Spatial and non-spatial data analysis

Bibliography

Boy, Jeremy; Detienne, Francoise; Fekete, Jean-Daniel. (2015). Storytelling in Information Visualizations: Does it Engage Users to Explore Data?. CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems

Censusindia.gov.in. (2015). Census of India Website : Office of the Registrar General & Census Commissioner, India . [online] Available at: http://www.censusindia.gov.in/

Chandola, Varun; Vatsavai, Ranga Raju; Bhaduri, Budhendra. (2011). iGlobe: an interactive visualization and analysis framework for geospatial data. COM.Geo '11: Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications

Dubner, Stephen; Levitt, Steven. (2005). Freakonomics

Gao, Tong; Hullman, Jessica; Adar, Eytan; Hecht, Brent; Diakopoulos, Nicholas. (2014). NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News. CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

Graves, Alvaro; Handler, James. (2013). Visualization tools for open government data. dg.o '13: Proceedings of the 14th Annual International Conference on Digital Government Research

Lu, Yun; Zhang, Mingjin; Li, Tao; Guang, Pudong; Rishe, Naphtali. (2013). Online spatial data analysis and visualization system. IDEA '13: Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics

Maeda, John, (2006). Laws of Simplicity

Mayer-Schonberger, Victor; Cukier, Kenneth. (2013). Big Data

MCGM. (2010). Mumbai Human Development Report 2009

MCGM. (2015). Preparatory Studies DP 2014-34

Mcgm.gov.in. (2015). Welcome to The Municipal Corporation of Greater Mumbai. [online] Available at: http://www.mcgm.gov.in/irj/servlet/prt/portal/prteventname/HtmlbEvent/prtroot/pcd!3aportal_content!2fcom.mcgm.fcontent_MCGM!2fcom.mcgm.faboutus!2fcom.mcgm.rHome!2fcom.mcgm.pAboutUsHome!2fdev_plan/documents/Draft%20Development%20Plan/?QuickLink=qlddevplan

Meirelles, Isabel. (2013). Design of Information

Ministry of Urban Development India. (2014). Urban and Regional Development Plans Formulation & Implementation Guidelines

Myers, Risa B.; Lomax, James W. III; Manion, Frank J.; Wood, Nancy M.; Johnson, Todd R. (2010). Data visualization of teen birth rate data using freely available rapid prototyping tools. IHI '10: Proceedings of the 1st ACM International Health Informatics Symposium

Opendatahandbook.org. (2015). The Open Data Handbook. [online] Available at: http://opendatahandbook.org/

Perer, Adam; Shneiderman, Ben. (2008). Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces

PRAJA; UDRI. (2014). Planning For Mumbai : The Development Plan For Greater Mumbai 2014-2034

Scientificamerican.Com. (2015). Why Are More People Right-Handed?. [Online] Available At: http://www.scientificamerican.com/article/why-are-more-people-right/

Shen-Hsieh, Angela; Schindl, Mark. (2002). Data Visualization For Strategic Decision Making. Chi '02: Case Studies Of The Chi2002|Aiga Experience Design Forum

Tufte, Edward. (1983). Visual Display Of Quantitative Information

Tufte, Edward. (1990). Envisioning Information

Tufte, Edward. (1997). Visual Explanations

Wikipedia. (2015). Open data. [online] Available at: https://en.wikipedia.org/wiki/Open_data