UNIVERSITY OF CALIFORNIA

Santa Barbara

MATDAN : Informing the Electorate

A Master's Project submitted in partial satisfaction of the requirements for the degree

MASTER OF SCIENCE

in

MEDIA ARTS AND TECHNOLOGY

by **Ambika Yadav**

Committee in charge:

George Legrady, Chair

Marko Peljhan, Co-Chair

Dr. Tobias Höllerer

December 2017

The Master's Project of Ambika Yadav is approved.

George Legrady, Professor, Chair

Marko Peljhan, Professor, Co-Chair

Dr.Tobias Hollerer, Professor

December 2017

ABSTRACT

MATDAN : Informing the Electorate by Ambika Yadav

Republic of India is regarded as the largest democracy in the world, due to the sheer number of its inhabitants. At the same time according to the Economist Intelligence Unit's democracy index, as of 2016, India ranks 32nd and is categorized as a flawed democracy. Matdan is an interdisciplinary project integrating information visualization methodologies with social sciences to build a data visualization platform to educate the electorate in India. It aims to act as a catalyst in India's transformation from a flawed democracy to a wholesome one.

The project is at present developed for the 2017 Legislative Elections for the state of Gujarat in collaboration with TCPD (Trivedi Center of Political Data), Ashoka University. The deliverable is an interactive data visualization built utilizing the javascript D3 library. The application has three visualization segments each derived from a different dataset. The first segment visualizes the historical Indian election dataset of TCPD. It is a time-series dataset of parameters attached with every constituency in Gujarat and is visualized by a Grid-Graticule Radial Visualization. The second section visualizes multivariate data profiling for candidates contesting in the upcoming elections. This information is extracted from the affidavit data collected by the Election Commision of India. The multivariate data is visualized using Radial Dendrogram Visualization. The final section visualizes aggregated socio-economic parameters accumulated from the Open Government Data (OGD) platform created by the Indian Government and is visualized using Streamgraphs.

The project aims to visualize this data to educate the electorate of the background and historical performance record of the competing candidates and political parties, and enable them to be able to make an informed voting decision. The outreach and insights of the project are evaluated by conducting a user study.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

1.Introduction

Matdan is a web based interactive data visualization application developed with the primary aim to educate the Indian electorate. The project has been done in collaboration with Trivedi Center of Political Data, Ashoka University [15] . The application educates and informs the electorate of the historical data of past governments and candidates of the government's contesting in upcoming elections. Matdan in particular focuses on the 2017 Gujarat Assembly State elections and data associated with it.

The information architecture of the application divides it into three major data visualization based sections which are Historical Election Data Visualization, Socio Economic Data Visualization and the Current Contesting Candidate Data Visualization. The data associated with each section is disparate, resulting in a unique data visualization for every section. The current candidate data visualization utilizes a Radial Dendrogram layout to visualize multivariate data. The socio economic data which is a collection of time series data is visualized using Stream graphs. Finally the historical election data is visualized with a Grid Graticule Radial Visualization. A streamlined five step process is put in place for the development and evaluation of each of these visualizations. The five steps are Problem Identification and Data Collection, Data Wrangling, Data Visualization Prototyping, Visualization Development and Deployment, and the User Study. The process is documented further in greater detail in the methodology section of this report.

The data resources associated with each section is available in the public domain, but at the same time is not very accessible to the public. Matdan aims to make this data more accessible and usable. The project presents data visualizations for this data on the web, empowering the

electorate to make a more informed vote. The development stack of the application utilizes the necessary web development toolkit, wherein the front end user interface is rendered using HTML, CSS, Javascript and D3 and the back end is supported by Python. Finally, the project is deployed on Google App Engine. The tools used to develop the application are further discussed in the Visualization Development and Deployment section of the report. A very crucial part of the project is evaluating the achievements of the goals of the project which is achieved by conducting a qualitative user study and is documented further in the report.

1.1 Problem Statement

The government of India is establishing organizations like Open Government Data and Election Commision of India to make the governance systems transparent. Using this data, media houses in India are able to present data driven stories to the public. However, there is a dearth of neutral resources both in terms of existing mediums and upcoming digital spaces, wherein the users can educate themselves about complete political landscape and performance in a user friendly manner. Voters' participation and awareness in the democratic and electoral processes is integral to a successful democratic process. It is the very basis of a wholesome democratic elections [3]. I believe Matdan is a stepping stone towards improvement of voters' participation and education.

The goal of Matdan is to serve as a platform to bring maximum information required to the people. It will serve the purpose of informing the voters about relevant data, that is both explanatory and exploratory [4] for better comprehension. This goal of Matdan facilitates the emergence of the problem statement for the same which emphasizes on putting in place an Information Architecture, data collection and wrangling and finally development of effective and user centered visualizations to educate the electorate. The data associated with every section of

the project results in an distinct issue of visualizing a particular type of dataset. The three datasets visualized as a part of project Matdan are multivariate data, set of time series and multivariate time series. The design decisions of the visualizations for each of the three datasets is to be explored and deduced using a series of prototypes.

1.2 Relevance of Research

Republic of India is regarded as the largest democracy in the world. At the same time according to the Economist Intelligence [1] Unit's democracy index, as of 2016, India ranks 32nd and is categorized as a flawed democracy. India is also a developing nation challenged by plethora of governance issues, which are well defined by the United Nations Sustainable development goals [2].

This development of the nation is heavily dependent on the government. A well informed and educated Indian electorate is what would help transition the nation from a flawed democracy to a full democracy and also act as a catalyst for the development of the nation.

2. Literature Review

2.1 Computational Social Sciences

Social Sciences is an evolved science in today's age. It is highly driven by data insights and analysis. The visuals communicating information on social sciences have also gone through an evolutionary process from images to complex data visualizations. The pioneers of Data Visualization have developed numerous projects communicating data and information about the same. In my literature survey, I focused on similar visualizations to gain an understanding and find the right motivation.

Data visualization in political and social sciences, Andrei Zinovyev

Andrei Zinovyev puts across in his paper on Data Visualization in Political And Social Sciences [25], how social and political sciences have evolved over the years due to the increase in the amount of data collected. He says, " As a consequence, there exists a continuous trend for political science to gradually become a more quantitative scientific field and to use quantitative information in the analysis and reasoning. "

There are two main types of tables in social sciences which can be subjected to data visualization for better understanding and communication. The first is the "object-feature" table where every data point has various columns. The first column specifies the object we talk about and the other columns hold information about the object data value in the first column. An example, Andrei Zinovyev quotes here is the country feature dataset. The first column will hold the object property, which is the country's name and the other columns hold value of the

features like population, Economy etc. The second type of dataset is called the connection or distance tables where both the first column and first row are objects and the intersection of the two holds the value that connects the two. An example of this would an export-import dataset of many countries.

He mentions how data visualization plays an important role in the field and serves, both as a *powerful propagandistic medium and as an educational tool.* As humans have a better visual memory, when compared to auditory or verbal memory, these visualization also help the viewer retain the information with much more ease. He talks about four main categories of visualizing political data,

1. Statistical graphics and infographics with extensive use of color, form, size, shape and style to superimpose many quantitative variables in the same chart or diagram

2. Geographical information systems (GIS) to visualize geographically-linked data

3. Graph visualization or network maps for representing relations between objects

4. Projection of multidimensional data on low dimensional screens with further visualization, data cartography.

He concludes his paper by highlighting the importance of ethics and neutrality while developing visualizations. The fundamentals of these have been highlighted very well in books and records of Mr. Edward Tufte, a data visualization pioneer. Project Matdan drives great amount of inspiration from the work of Mr. Edward Tufte and Mr. Andrei Zinovyev .

Related Work

A Visual History of American Presidency, Graphicacy

Graphicacy is a data visualization studio in Washington DC, USA presided by Mr. Nathaniel Pearlman. Their work on visualizing history of American Presidency in A Visual History of

American Presidency visualizes various datasets together on a single graphic. It places each president in historical context, visualizing a remarkable range of political, social, and economic measures to succinctly tell the story of the presidency. Narratives are displayed within the larger context of American political history by aggregating and annotating hard data on population, presidential elections, Congress, the Supreme Court, the Cabinet, the U.S. economy, the federal budget and debt. The Time plot provides a new lens into American political history; it is not intended to be absorbed at a glance, but rather to be visited and revisited over time. [25]

Visualizing the Indian Elections, Gramener

Gramener is a data science company from India. Their work on Visualizing the Indian Elections has been one of the few visualization projects in India. The visualization developed are static image visualizations based on national level data. The data they visualize answers questions which range from the number of seats a party represents in the Lok Sabha to the wealth distribution of candidates contesting in the upcoming elections. They use stacked bar charts to visualize the party distribution in the houses. Many of their visualizations are also based of on India's geographic projection.

2.2 Indian Political System

In India, the government elections are conducted at federal, state and local levels. The general elections are held once every five years, wherein the central/federal government or the members of the Lok Sabha are elected. Similarly, the assembly elections are also held once every five years in every state wherein the state government or the members of the state's Vidhan Sabha are elected. The Election Commision of India [16] headed by Dr. Nasim Zaidi is responsible for conducting fair elections and collecting the data associated with the elections.

The team at Trivedi Center of Political Data works with Election Commision of India to create databases holding historical Indian Elections data, which is utilised in Matdan to visualize the historical political landscape of Gujarat. This would inform the electorate about the past candidates from their area and provide relevant information .

3. Methodology

Matdan is an interdisciplinary project built at the intersection of data visualization, education, and social sciences. The project is an interactive web-based data visualization platform aiming to create a positive social impact by educating the Indian electorate. The project has lead to the development of an educational framework to facilitate informed voting decisions in the Indian context. There are three main visualization interfaces that the user interacts with. The prototyping, development and evaluation of each of the interface is streamlined and divided into five major sections.

3.1 Problem Statement and Data Collection

The problem statement is put together to understand the impact, need and requirements in terms of tools and data for every section. With a clear understanding of the metadata requirements and availability, required data sets are identified and collected as necessary.

Current Candidate Affidavit

Every member of the electorate must evaluate and compare candidates before casting a vote. Every candidate contesting in any election in India has to submit an affidavit [13] of their information to the Election Commision Of India. Once this is approved, the candidate can contest the election legally. This affidavit information is released by the Election Commision of India prior to the elections for democratic transparency, and help voters make a better decision. This affidavit data is processed and made available by the Association of Democratic Reforms.

The current candidate affidavit data is a multidimensional dataset. The overview and knowledge retrieval of the multidimensional dataset for a generic member of the electorate is minimal. The interface development of this data aims to present a solution to this problem by creating a user centered visualization of this multi dimensional data.

The data is collected by developing a Python based Web Scraper using the Beautiful Soup [14] library. Beautiful Soup is a Python package used for parsing XML and HTML documents. The python script accesses the MyNeta [15] web page to extract this data.

Socio Economic Parameters

The next section of the project aimed to provide the members of the electorate a visualization which is a visual transcript of the government performance over the years. The government's performance can be evaluated from the variance of socio economic parameters. The collected socio economic parameter data collected is categorized into four sections: Economy, Literacy, Health and Women Safety. Each dataset is a temporal dataset with quantitative parameter values. These parameters within each section are globally recognised parameters collected by different government organizations.

Digitalization of data in India started only in the past few years. This resulted is an incomplete set of data points, available only for recent years. The data for this visualization is made open source via various web based initiatives. The Open Gov Data [9] of the Digital India Initiative and India Stat [10] were the major resources used to collect this data.

Historical Election Results

Historic Election Data Visualization is a tool to understand the evolution of political demographics in the Gujarat Vidhan Sabha over the past 20 years. Gujarat Vidhan Sabha

comprises of 182 members called MLAs (Members of Legislative Assembly) , each elected from one of the 182 constituencies in Gujarat. Constituencies are geographic areas of a state which hold an equally divided electorate population and this categorisation is done by the Election Commision of India. The division of the state into constituencies is a procedure to ensure an equal electorate to MLA ratio in the state. By utilizing the Historic Election Data Visualization, the user can identify any constituency and the data associated with that constituency after every state elections in the past 20 years.

The Indian State Assembly Election and Candidates Data [6] is a part of the Lok Dhaba project initiated by Trivedi Center of Political Data, Ashoka University. The dataset for Gujarat state, from the year 1998 - 2012 is used in Matdan.

3.2 Data Wrangling

Data wrangling is the process of transforming and mapping **data** from one "raw" **data** form into another format with the intent of making it more appropriate for the application [20]. The data collected in the last step is transformed and mapped to the required data format for further prototyping and development.

Current Candidate Affidavit Data

The data collected from MyNeta web page is a multidimensional dataset with every data point profiling a unique candidate. The dimensionality of the dataset is reduced to 7, wherein every data point is characterized by 4 categorical, 2 continuous and 1 distinct dimension. Every data point has a unique distinct dimension which is the candidate's name. The categorical dimensions are the constituency from where the candidate contests, education level of the

candidate, number of criminal cases against the candidate and the party the candidate represents in the election. The continuous dimensions are the liabilities and assets owned by the candidate.

Socio Economic Parameter Data

The data collected from various sources were available in varying formats. The data wrangling process started with converting all the data files to csv format. The data pertaining to a single attribute was stored in multiple tables. These tables were then merged to contain all values of a attribute over different time indices in one file. Finally, attributes were aggregated into sections and this was the final dataset used in the visualization.
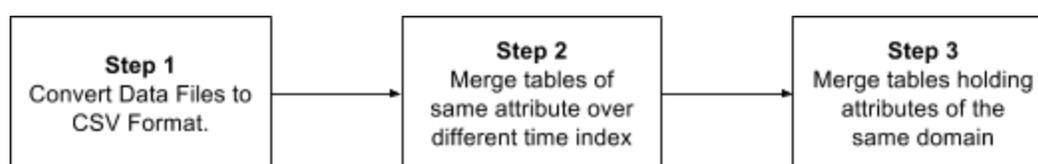

Figure 1 : Data Wrangling Process for Socio-Economic Data

Historical Election Result Data

The data collected from the Lok Dhaba project was reformatted to hold data points with the year attribute ranging from 1998 to 2010. The dataset was compressed in terms of the data attributes and points. Data points with candidate position equal to 1, implying the winning candidate was retained. Also data points with year of election falling within 1998 to 2017 were retained in the final dataset. Post data processing, every data point had the following attributes : Year of Election, Constituency Name, Constituency Type, Constituency Number, Candidate Name, Candidate Gender, Party Name and Voter Turnout.

There were discrepancies in constituency names and candidate names due to different spellings used over the years which were recognised in the process of data wrangling and processed to errors further on. The TCPD database had an error in the values of voter turnouts in Gujarat constituencies for the 2012 elections. This error was detected in the data wrangling process and reported back to the center, for correction.

3.3 Data Visualization Prototype

With the problem statement and the metadata of the available data in place, visualization prototypes are developed abiding by the basic principle of information visualization : overview first, zoom and filter and then details on demand. The prototype acknowledges five visualization principles. These principles are maximize complexity [4], minimise Lie Factor [5] , present comparisons [5] , add user Interaction and enable effective Data Point Annotation.

Current Candidate Affidavit Data Visualization Prototypes

The current candidate affidavit is multidimensional dataset wherein every data point is defined by a combination of categorical and continuous dimensions. As the data was a multidimensional dataset, the prototyping process began with the visualization of the dataset using the parallel coordinates visualization shown in figure 7. The data couldn't be effectively visualized using Parallel coordinates for two main reasons :

1. The dimensions of data visualized by parallel coordinates is generally continuous whereas in this case every data point was characterized by a combination of 4 categorical and 2 continuous dimensions. Using only parallel coordinates to visualize this dataset would confine datalines due to categories leading to visualization clutter.

2. The visualization would be dominated with the presence of overlapping lines due to the high number of datapoints again leading to visual clutter. This would restrict the users ability to comprehend different data points.
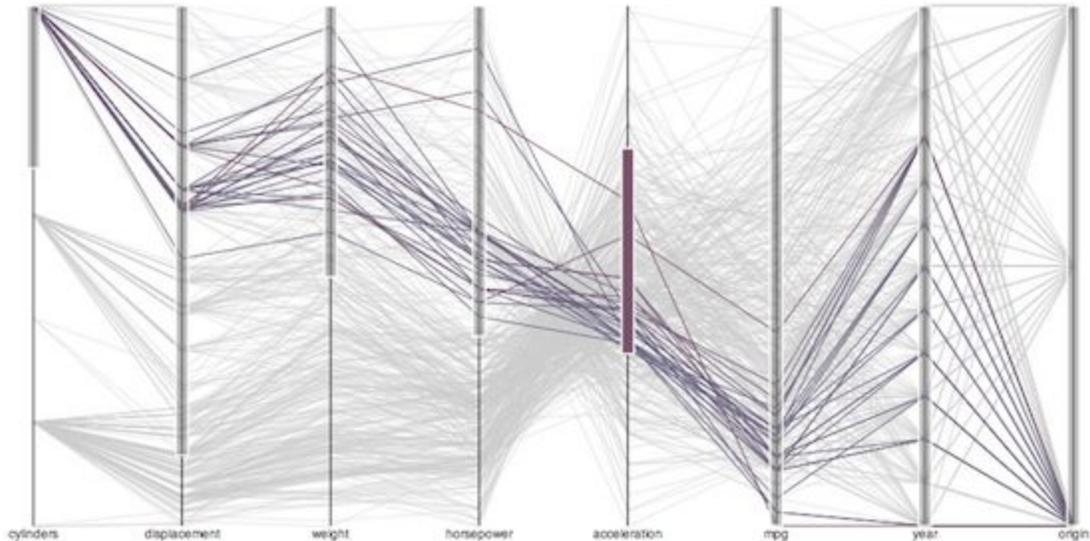

Figure 2 : Parallel Coordinates Data Visualization

The failures of the first prototype to accommodate discrete data dimensions led to the modification of the parallel coordinate system to accommodate the discrete parameter. The parallel axis of every parameter in the visualization was divided into segments, wherein each segment represented a distinct discrete value of that parameter. The length of every segment visualizes the marginal distribution of that parameter value. Instead of representing every data point as a distinct line, clusters of similar data points were visualized as polygons extending over the parallel axis. This second prototype took the shape of Sankey Flow Diagram.

A sankey diagram [21] is a visualization used to depict a flow from one set of values to another. The things being connected are called nodes and the connections are called links. In this prototype we use sankey diagram to visualize multivariate data with continuous and categorical

parameters. The sankey diagram prototype was not the right choice because of two main reasons :

1. The user is not able to locate every candidate with a distinct visual representation. Instead candidates are clustered and visualized as polygons.

2. Due to the kind of data we have as we move across the visualization the overlapping of the polygons will keep increasing, which would in turn increase the visual clutter.
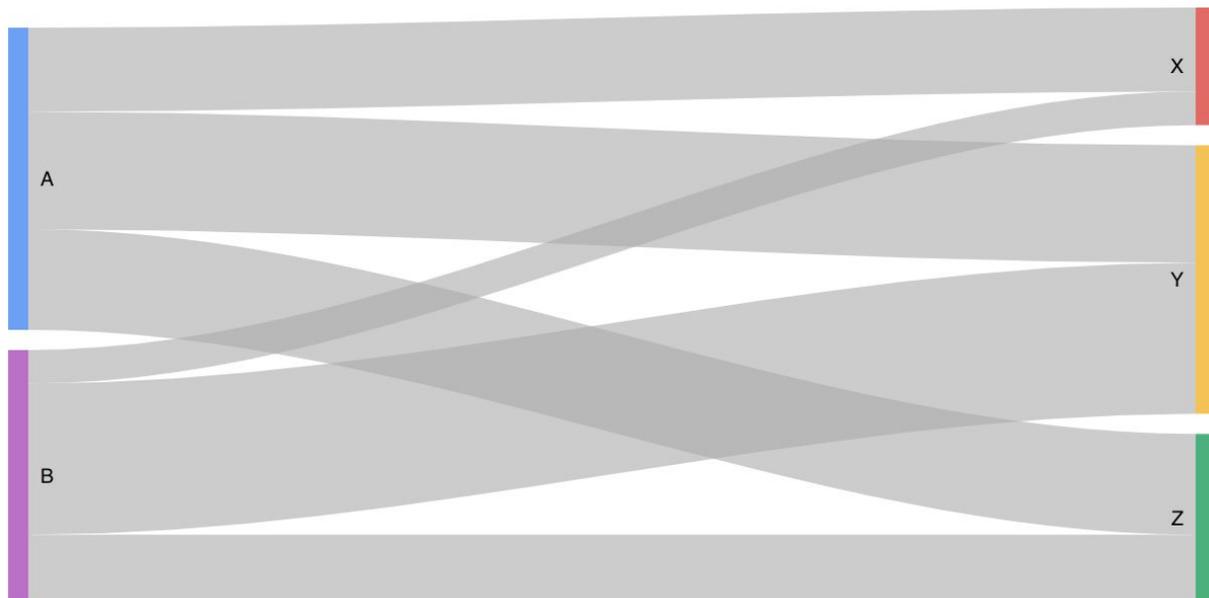

Figure 3 : Sankey Diagram

In order to resolve the visual clutter and enable better data point accessibility for the user the prototype was further modified. The third prototype was developed by separating the parallel axis for every subset generated by the previous dimension. Visually, if the segregation is made as linear chart, the space occupied by the visualization is a lot more if the same segregation is performed radially. The resulting prototype was a radial dendrogram with the width of links derived from the number of elements in the subset and every node representing a different category at a particular granularity level.
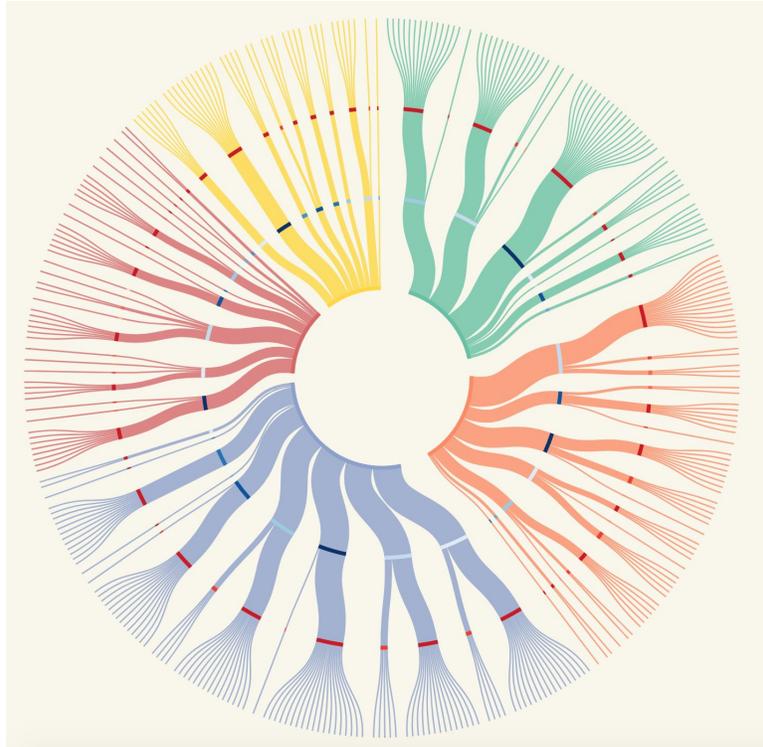
Figure 4 : Radial weighted dendrogram

The distribution of subsets into different categories in this prototype was not very intuitive and easy to comprehend from a user's perspective. Also, there is a restriction in terms of scaling this visualization with an increase in the number of data points in the dataset. This led to last prototype wherein the weight on the links was dropped and every node was visualized as a circle similar to the classic radial dendrogram.
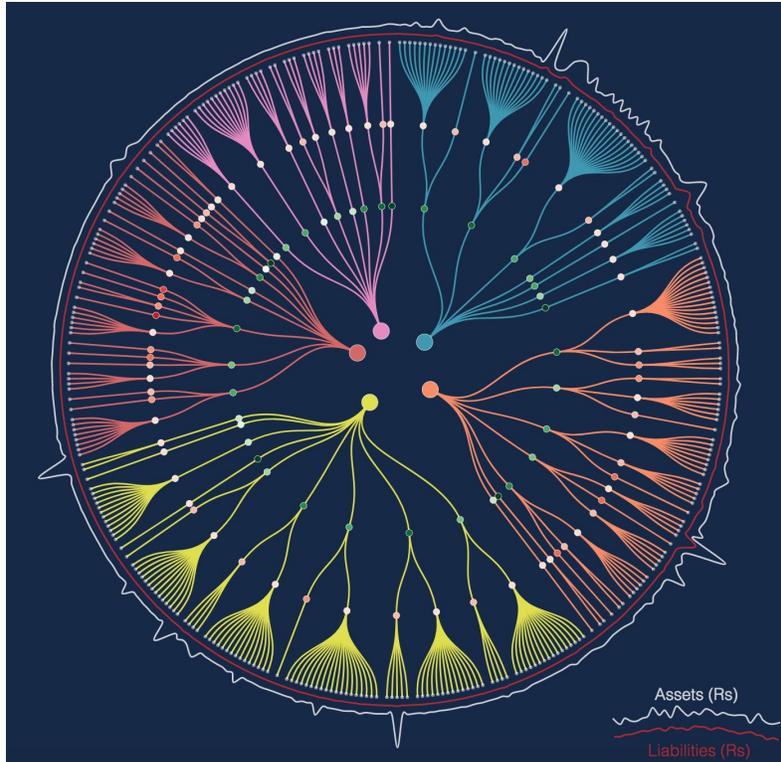
Figure 5 : Radial dendrogram

Socio Economic Parameter Data Visualization Prototypes

The socio-economic data collected were grouped into four sections : Economy, Healthcare, Education and Women Safety. Each of these had many parameters quantifying the improvements and depreciations. The dataset within every section is a collection of time series.

The first prototype to visualize this collection of time series data was the conventional stacked bar chart. The bar charts failed at highlighting the change in the value of a components over time.
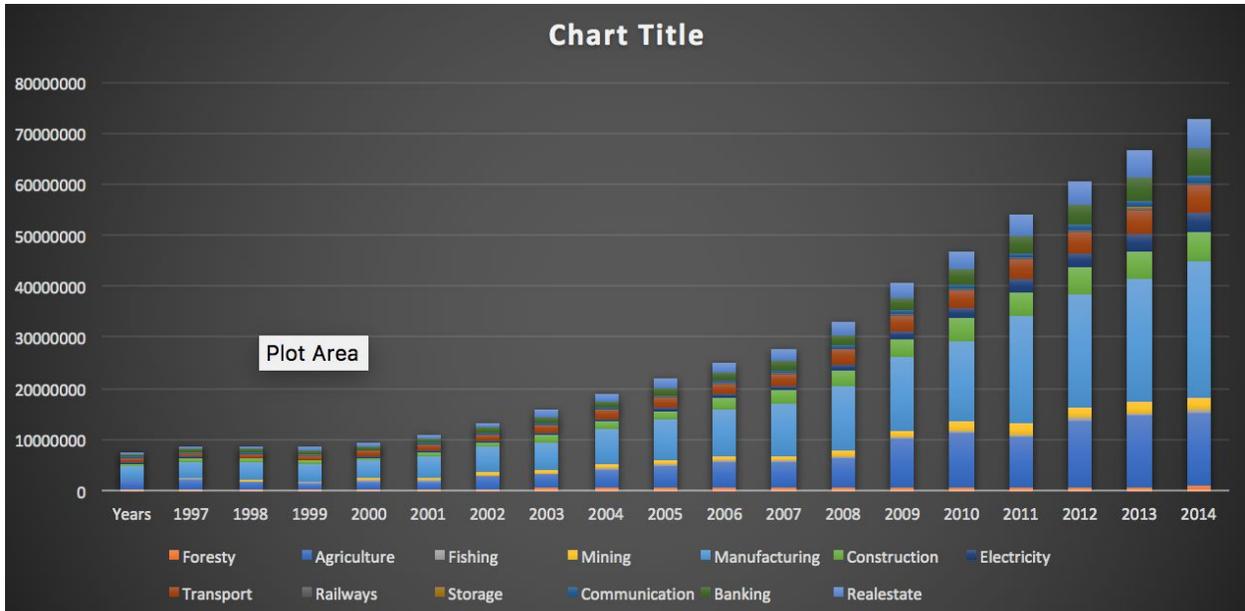
Figure 6 : Stacked Bar Chart

The second prototype aimed to highlight the change in values in every sector without any visual clutter. This was achieved by using Streamgraph [17] to visualize this dataset. Streamgraph is a stacked area graph which is displaced around a an axis.

There are four main visual cues of a streamgraph are :

1. The x axis, which represents the time.

2. Every stripe of the streamgraph represents each one of the time series in the collection.

3. The width of the stripe visualizes the values of the time series at a particular point.

4. The color of the stripe can be used to represent an attribute of individual time series.

Byron and Wattenberg [23] describe the geometry of a streamgraph precisely, by using the following notation. They model the time series as a set of n real-valued non-negative functions, $f_1, \ldots, f_n$. The baseline function that defines the bottom of the stacked graph is referred by g_0. In

22

the initial   setup the baseline is considered to be the x = 0. Then the top of the layer

corresponding to the ith time series f_i is therefore given by the function g_i , where

$$g_i = g_0 + \sum_{j=1}^{i} f_j$$

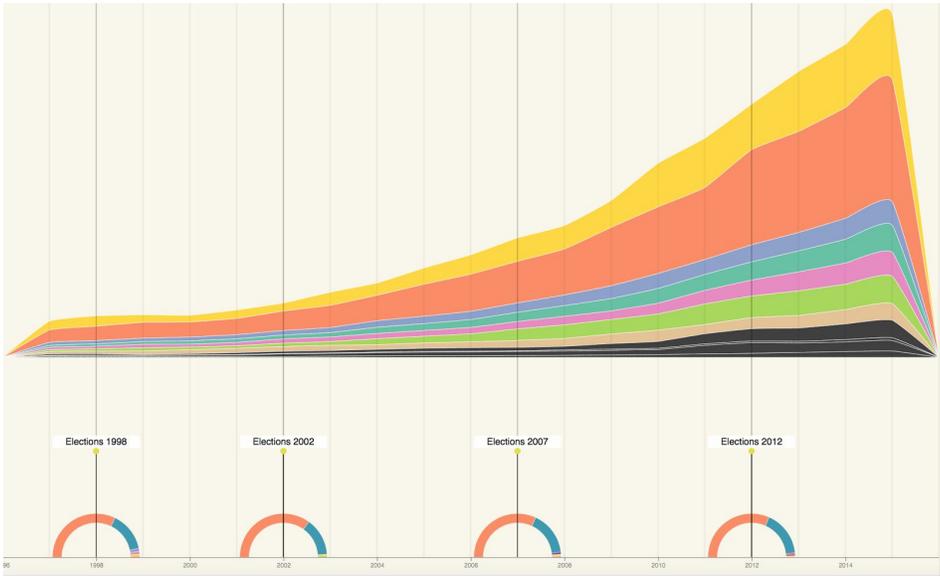**Figure 7 : Mathematical representation of Streamgraph**

Figure 8:: Streamgraph with baseline at x = 0

When the baseline of the streamgraph is $g_0 = 0$, it makes the size of the sum easy to read but at

the same time leads to legibility of design issues .The wiggle in any of the time series in the

graph is carried to the upper layers decreasing the legibility of the trends associated with them.

This is corrected by shifting the baseline and making the geometry symmetrical around the x

axis. This would improve the legibility of the visualization as the silhouettes are closer to the

central axis. Also the variation in slope due to the wiggle being carried forward, will be reduced
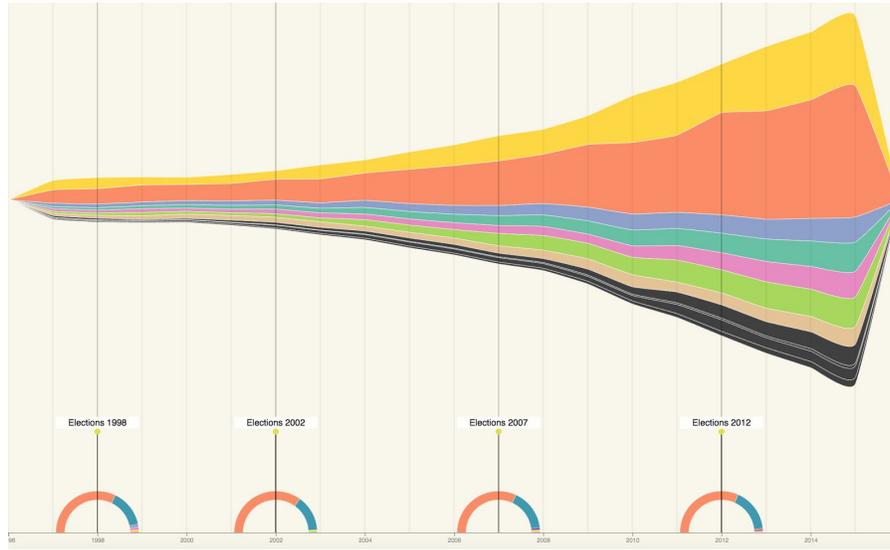
as it is now spread in both the directions.

Figure 9 : Streamgraph with centered baseline

The visualization for this segment also visualizes the variation of parameters in these sectors along with the changing governments by highlighting the election years.

*Historical Election Data Visualization Prototypes*

The first prototype of the visualization was an attempt to combine the exploratory and explanatory aspects of the visualization using a combination of waffle charts [7] and histomaps [8] . In order to make the visualization explanatory the visualization was segregated into slides, with every slide visualizing a data attribute (Continuous Candidate appointments, Constituency Type, Winning Party, Women Candidates and Voter Turnout).The visualization was made exploratory by adding user interactivity and enabling a user select a constituency to look into. Upon selection the user could view the temporal visual data points linked with that constituency.

24

Every election year (1998, 2002, 2007 and 2012) was a horizontal line and every constituency was depicted by the waffle charts' component, which is a rectangle.

The first slide is the Continuous Candidacy which visualized the candidates who won elections over the years from the same or different constituency. The continuous candidacy is represented by lines connecting the rectangles of that particular candidate over different constituencies and years. The user could highlight the trajectory of a particular candidate by selecting a candidate line.

The second slide, on the constituency type, visualized the different types of constituency (GEN, SC and ST). This visualisation was achieved by color mapping of the waffle components (rectangles) to the constituency type attribute.

The third slide, about the winning party for individual constituencies, visualized the parties of the candidates that won the elections in that particular year from that particular constituency. This visualisation was again achieved by color mapping of the waffle components to the constituency party attribute . Also, the constituencies were sorted according to the winning party to enable the user to understand how the political demographics changed over the years. The sorting along with polygon extension of the waffle chart to visualize evolution of the political landscape over the years was inspired by the histomap visualization [8].
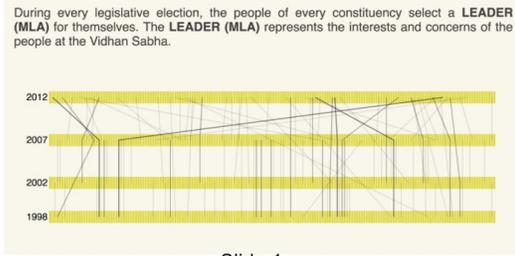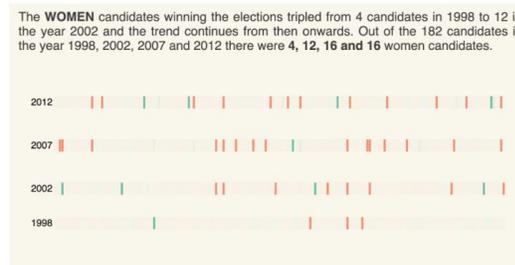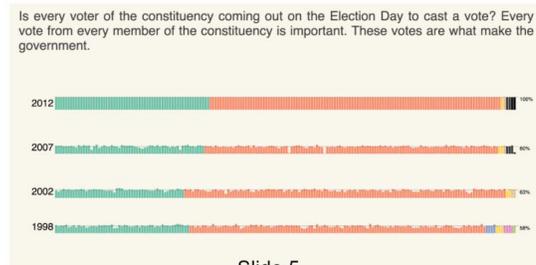
Slide 1
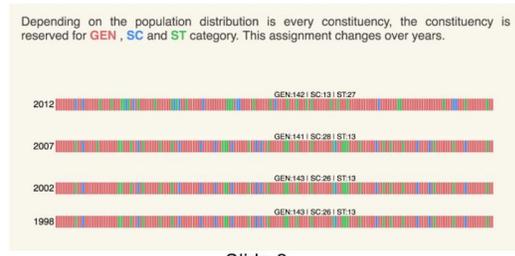
Slide 2
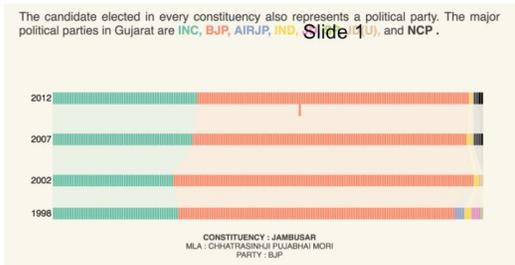
Slide 3

Slide 4

Slide 5

Figure 10 : Histomap and Waffle Chart Based Visualization

The fourth slide about reserved women candidacy, visualized the constituencies where the winning candidates were women. This was achieved by opacity mapping, with the women candidate data point having a higher opacity.

The final Voter Turnout Slide visualized the voter turnout percentage in constituencies. This was visualized by mapping the length of the waffle representing every constituency to the voter turnout.

Failures and Learnings

1.   Parameter Inclusion / Complexity : The quality of a visual is defined by the number of data dimension visualized. In this visualization, on every slide, this value reaches a maximum of four . Hence there was a need to increase this complexity to improve the quality of visualization.

2.   Varying Parameter Mapping : In an explanatory setup the visualization becomes inefficient if the data attribute defining a visualization attribute is changed at different stages. For example : The color mapping of every waffle must remain the same throughout the visualization.

3.   Unique Constituency Position : Ideally, every unique constituency must have a unique position along one of the axis. This is not achieved here. Along every year a constituency may or may not have the same position along x. This results in a confusion about the presence or absence of a constituency in a particular year.

The above learnings led me to develop the second prototype which in addition to combining the exploratory and explanatory approach of the previous prototype, also derives inspiration from the waffle charts. At the same time the second prototype counters the failures of the first prototype. The explanatory approach takes the user through the visualization step by step to explain them what visual component represents which data attribute. This enables the user to understand the visualization component in great detail and enable better exploration.
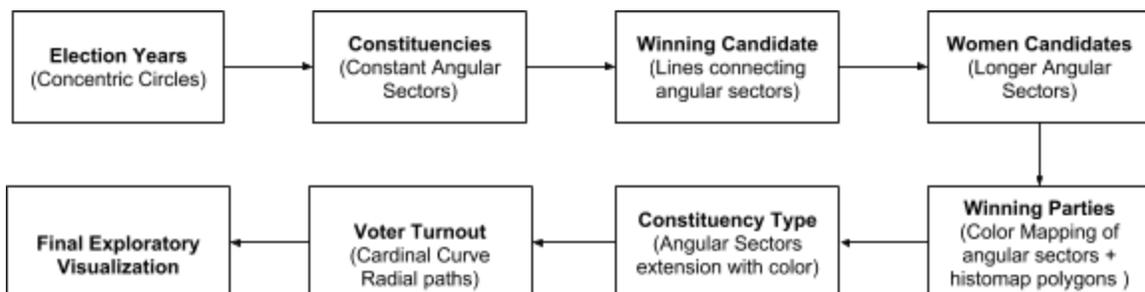

Figure 11 : Explanatory Flowchart

Once the explanatory slides are encountered, the user lands on the exploratory visualization. Here the user can choose what aspect of the visual do they want to explore - constituency type, wining party, women candidates or voter turnout. This way the user is easily able to walk him/herself through the political landscape on Gujarat and understand the historical significance of their representation.

In order to enable the correct placement every constituency must have unique position along the x axis. If executed in a linear space the visualization becomes space extensive leading to massive reduction in the interaction space for every constituency and leading to overall visual clutter. This led to prototyping a radial visualizations combining the three families of circular visualizations [11] to overcome the failures of the previous visualization.
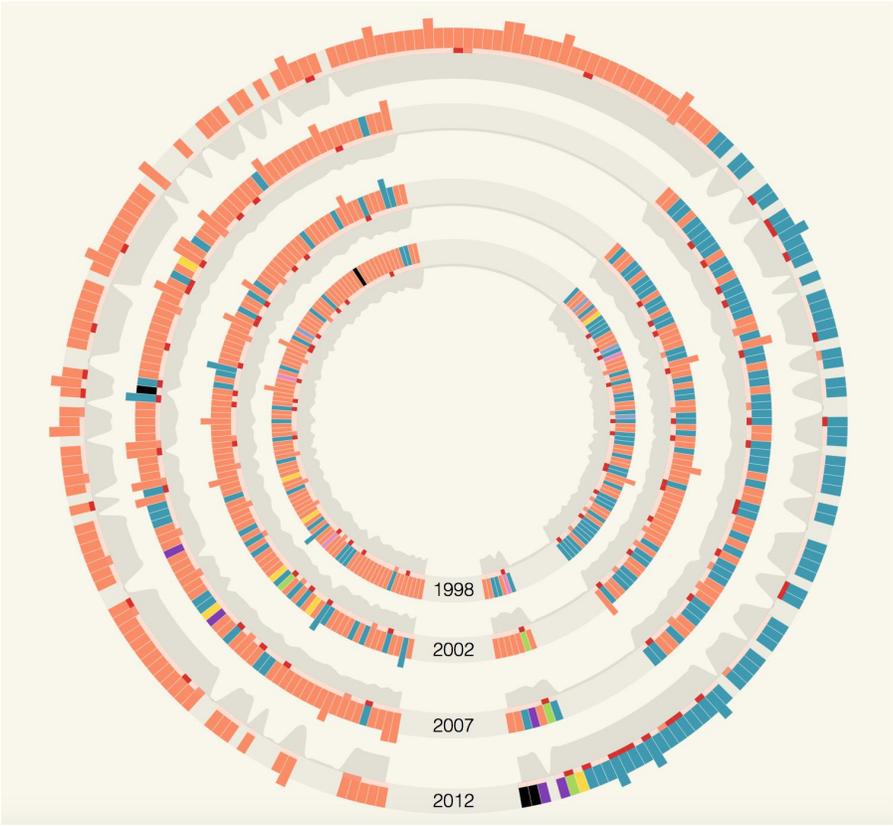

Figure 12 : Grid- Graticule Radial Visualization

The use of circular space to visualize the data point to conserve space, gives more space to every datapoint. For a given square area on a page (e.g. square of side x), the circular layout can support a larger data domain (i.e. $2\pi x \sim 6x$) than the square (i.e. x). Also a circular form encourages eye movement to proceed along curved lines, rather than in a zigzag fashion in a square or rectangular figure. [23]

3.4 Visualization Development and Deployment

Matdan is a web based application to ensure the outreach of the project. The front end development majorly comprises of data visualization development. The technology stack comprises of HTML, CSS and Javascript. The visualization is majorly developed using D3, the javascript based visualization library and JQuery. D3 combines powerful visualization and interaction techniques with a data-driven approach to DOM manipulation. To develop the visualizations, D3 is extensively used to bind data to SVG components and manipulate styling attributes of the same on the HTML page.

The back end of project, which deals with data extraction, transformation and loading utilizes Python. Python is an interpreted high level programming language for general purpose programming. The availability of data extraction and wrangling packages like Beautiful Soup and Pandas makes Python an obvious choice for the back end development of a Data Visualization Application.

The developed project is deployed on the Google App Engine cloud. Google App Engine is a web framework and cloud computing platform for developing and hosting web applications in Google-managed data center. To minimize the capital usage of the project, Google App Engine

is utilized as it comes with a higher bandwidth and processing power when compared with AWS (Amazon Web Services) and Azure.

3.5 User Study

The outreach and success of the project is evaluated by conducting a user study for the project. The user study is divided into three sections Pre User study, User Study and Post User Study. The pre and post user studies are conducted using google forms. The pre user study profiles the user and helps us understand them better. The main user study is conducted by video recording, screens of the users while they interact with the application. In the main user study, we particularly look into the amount of time spent by the user exploring every section and the features used by them. We also investigate the intuitiveness and usability of features within the application in the main user study. Finally the post user study helps us understand the impact and outreach of the application.

Pre-User Study Questionnaire

1. How old are you ?

2. Are you an Indian ?

3. Have you voted in an Indian Election ?

4. Which of the following affect your voting decision ?

    a. Prominent leading party

    b. Candidate Profile

    c. Historic Development of the State

    d. Socio Economic Parameters

    e. Family

Post- User Study Questionnaire

1. How likely are you to use this application before casting a vote ? (1-5)

2. How likely are you to use the Current Candidate Data Visualization before casting a vote ? (1-5)

3. How likely are you to use the Socio Economic Data Visualization before casting a vote ? (1-5)

4. How likely are you to use the Historical Election Data Visualization application before casting a vote ? (1-5)

5. How informative was the application ? (1-5)

6. How intuitive was the application ? (1-5)

7. Do you think an application like this is required to educate the voters ? (Yes/No)

8. How do you think the application can be further improved ? (Short Answer)

4. Data Visualization

4.1 Visualization 1 : Radial Dendrogram Visualization for Multivariate Data

The current election candidate data is a multidimensional dataset where in every data point creates a profile of the contesting candidate. Every data point is 7 dimensions or attributes - 4 categorical, 2 continuous and 1 distinct dimension. The categorical dimensions are visualized using a radial dendrogram visualization. The 2 continuous dimensions are visualized using the radial lines of distinct colors. The distinct dimension is the candidate's name which is represented by every link on the dendrogram.

The dataset is first formatted using the d3 stratify function to form top-down hierarchical agglomerative clusters on the basis of 3 categorical dimensions. The tree data structure which is created out of the stratification has a depth of 4. The first level clustering forming the depth 1 data points is performed using the candidate_party attribute. The second level clustering (Depth =2) is done using the candidate_education attribute and finally the third level clustering (Depth =3) is done using the candidate_criminal cases attribute. The candidates are the leaf nodes forming the last level of the tree. The parent nodes of the candidate

The tree data structure created out of the stratification of the data is then used to create the radial dendrogram visualization. The nodes of depth = 0, represent the political parties and are the innermost and largest nodes in the radial dendrogram. The color of the links moving down the tree from the party nodes, representing every candidate is derived from the node or the candidate party color itself.

The nodes of depth equal to 2 visualize the candidate's education. These are color coded with decreasing saturation of green with decreasing education level. The final splitting is on the nodes with depth level 3 visualizing the number of criminal cases against the candidate. These are color coded with increasing saturation of red with increase in the number of criminal cases against the candidate. The leaf nodes of the tree are the candidates and the nodes are low white opacity circles.

The continuous dimensions, which include the assets and liabilities of the candidates, are represented by the radial lines. The data points constructing these radial lines are positioned at a scaled radial extension from every candidates leaf node position.
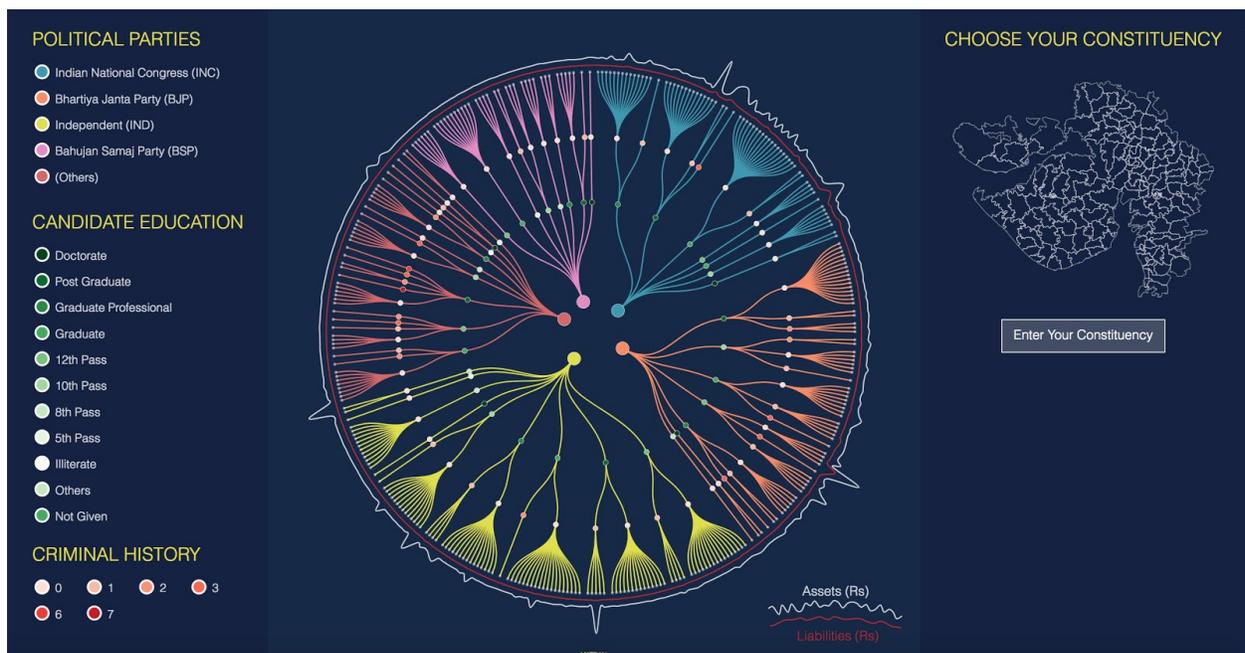

Figure 13 : Radial Dendrogram Visualization Screen

Once the user gets the overview of the data, the user interacts with the application to get further insights about the data. The user can interact with the visualization using three approaches :

hovering, filtering and selection. At any time the user can hover over the visualization nodes to get more details about the same.
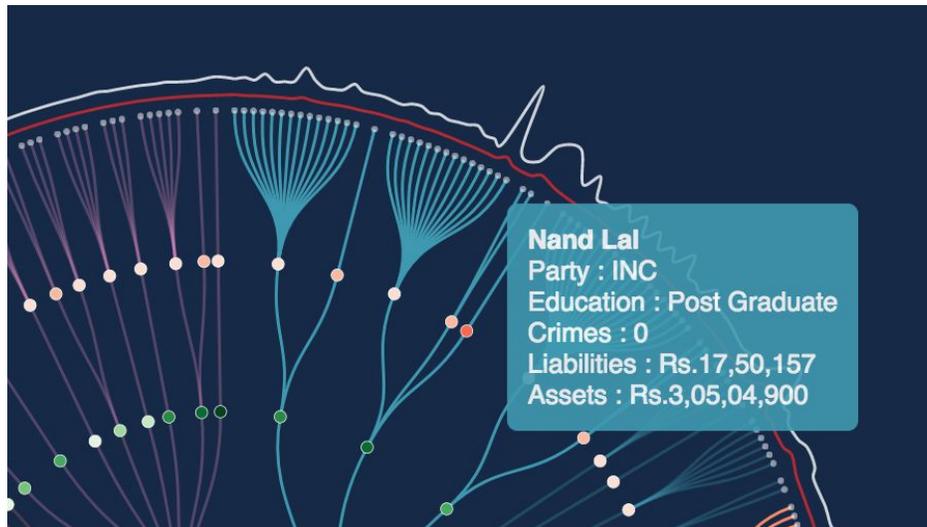

Figure 14 : Hovering Functionality

The user can also use the node buttons on the left side of the screen to filter candidates of certain type. For example the user can choose to see all candidates of a political party A and B who have education C and D, and E criminal cases. On activating the buttons of these attributes on the UI, only these candidates on the visualization will be highlighted with a higher opacity compared to the others.
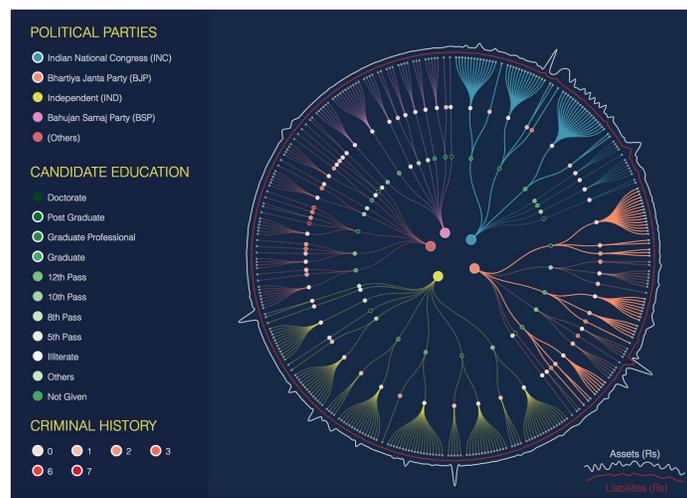

Figure 15 : Data Filtering Functionality

The user can use the UI on the right side of the visualization to make a selection of the candidates contesting from a certain constituency. The input textbox to select the constituency is a pure svg element developed from scratch. On selection of a constituency, the constituency is highlighted on the map of the state. Also, upon selection of the constituency the leaf nodes and link of the those candidates extend outside and are highlighted for better viewing of the user.
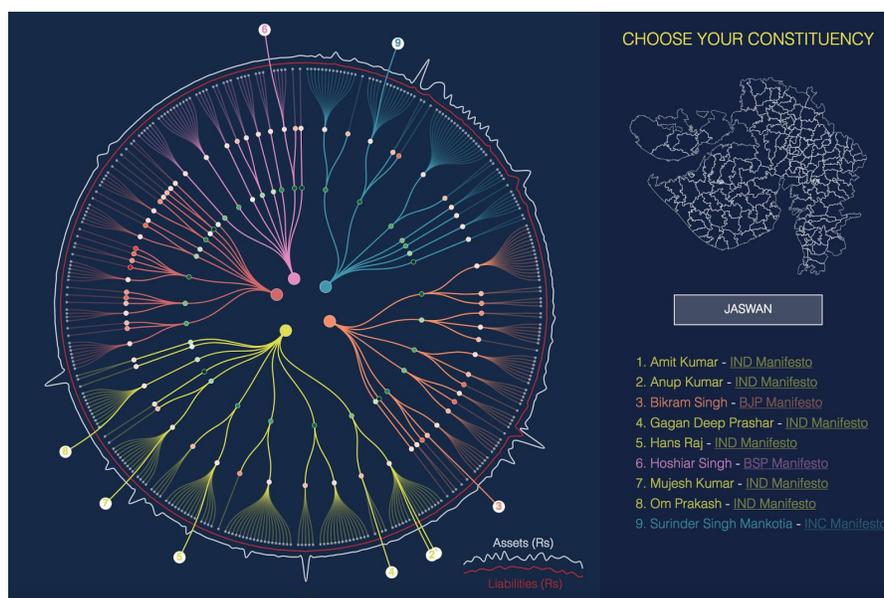

Figure 16 : Data Selection Functionality

4.2 Visualization 2 : Stream Graphs

The socio-economic parameters is a collection of time series data. The data is visualized by creating a stream graph [17] visualization. Stream graphs are a modification of the stacked bar chart, where in the values are displaced along a baseline and the value mapping along the time axis is a continuous area called stripes instead of discrete blocks of area.

The user interacts with the application to generate stream graphs for different sectors (Economy, Education, Health and Women Safety) according to their explorative needs. When the user selects a particular sector, explanatory text giving an overview of the attributes is updated as well. The visualization gives the user an overview of the trends in the different time series parameters associated with every sector. Once generated, the user can scroll across the streamgraph to perform an in-depth exploration. The scrolling appends annotation comprising of the attribute, value and timestamp of the datapoint.

The Streamgraph is also annotated with the election years, so as to recognise a change in the government and conduct a comparative analysis of the variation in parameter values during different governances. The composition of the government elected in that year is visualized using a donut chart.
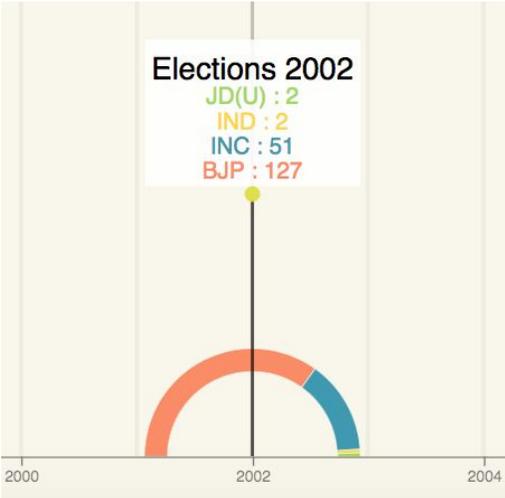

Figure 17 : Election Year Donut Visualization

The streamgraph construction is smoothened by utilising cardinal curves instead of straight lines to generate the stream areas. The generation of stack values utilizes a reverse stack order for easier color mapping readability and silhouette offset in order to shift the baseline so that geometry is symmetrical around the x axis.

36

Economy : The economy sector visualizes the GSDP (Gross State Domestic Product) of different commercial sectors. The commercial sectors are visualized here are Agriculture, Manufacturing, Transport, Real Estate, Banking, Construction, Electricity, Railways, Mining and Communication. The color of every stripe represent individual sectors. The width of the stream represents the value of GSDP of that particular sector.
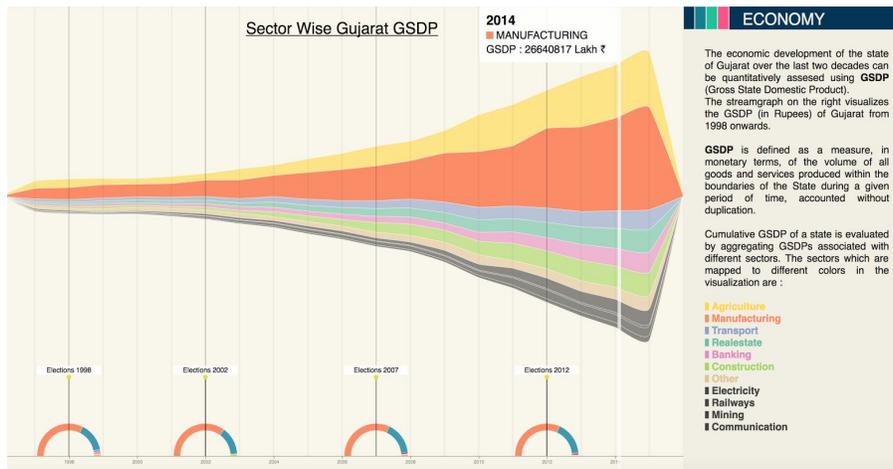

Figure 18 : Economy Data Streamgraph

Education : The education sector visualizes the number of public and private schools, enrollment in public and private schools, the number of teachers in public and private schools. There is a dearth of the number of data points collected by the government. The non availability of data in any particular year is represented a null value. The user can view one of three parametres enrollment, number of teachers or number of schools at a time. The selection of which parameter to view can be made from the UI on the right side of the screen. Every stripe and its color visualises whether the parameter value is attached with a private or government school. The width of every stripe visualizes the value of the parameter in the private or government school.
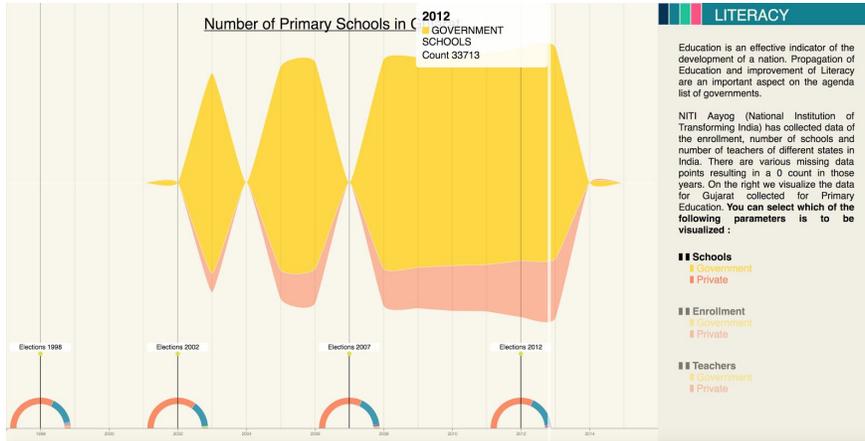
Figure 19 : Education Data Streamgraph

Women Safety : The women safety section visualizes the incidence count of categorized crime against women as disclosed by the National Crimes Record Bureau. The categories visualized here are Rape, Kidnapping, Dowry Deaths, Domestic Violence, Molestation, Eve teasing, Importing, SATI and Immoral Trafficking. Again the categories are visualized by every stripe and its color. The width of every stripe visualizes the number of incidences of crime of that particular category.
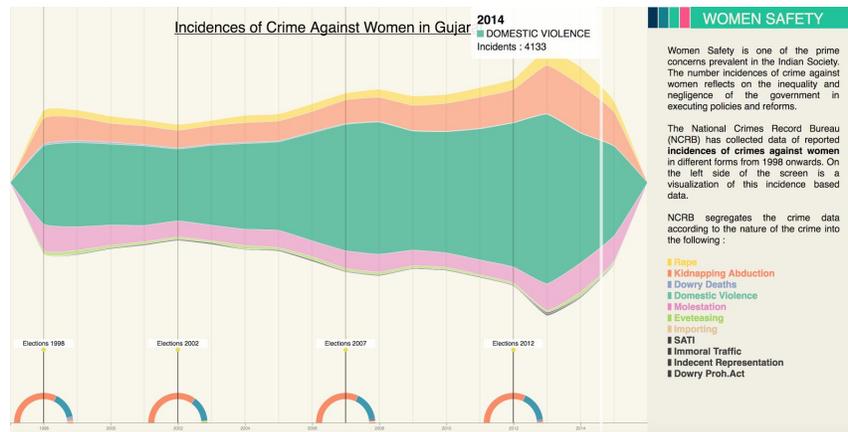

Figure 20 : Women Safety Data Streamgraph

Health : The health sector visualizes various health based parameters evaluated by the government. The parameters visualized are IMR (Infant Mortality Rate), Birth Rate and Death Rate in a rural, urban and combined demographics. The user can view each of the parameters separately in the application. Within every parameter visualization, every stripe and its color visualizes if the parameter is calculated in a rural, urban or a combined setting. The width of the stripe visualizes the value of the parameter in that setting.
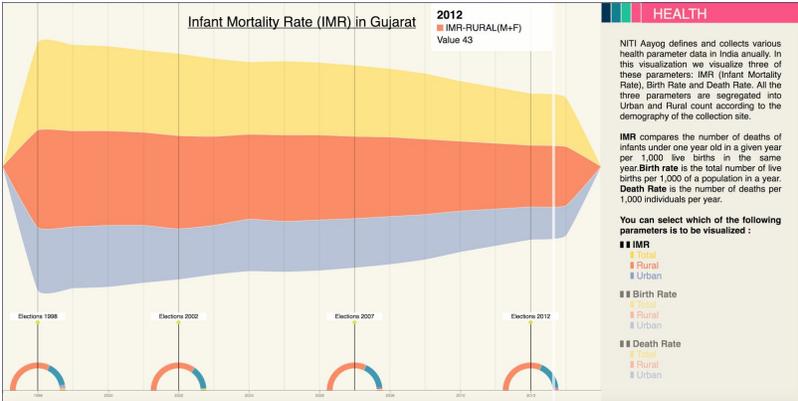

Figure 18 : Health Data Streamgraph

4.3 Visualization 3 : Grid-Graticule Radial Visualization

The historic election data visualization is a radial visualization combining three taxonomies [11] of circular visualizations. The visualization is spread across various screen called slides through which the user navigates to utilise the application. The visualization as suggested in the methodology is split into explanatory and exploratory segments. The explanatory section takes the user through the construction of the visualization and simultaneously briefs about the elections.

Election Years : Election years are visualized in Explanatory segment 1. Years are represented by concentric circles. The outermost circle is 2012 and the innermost in 1998.
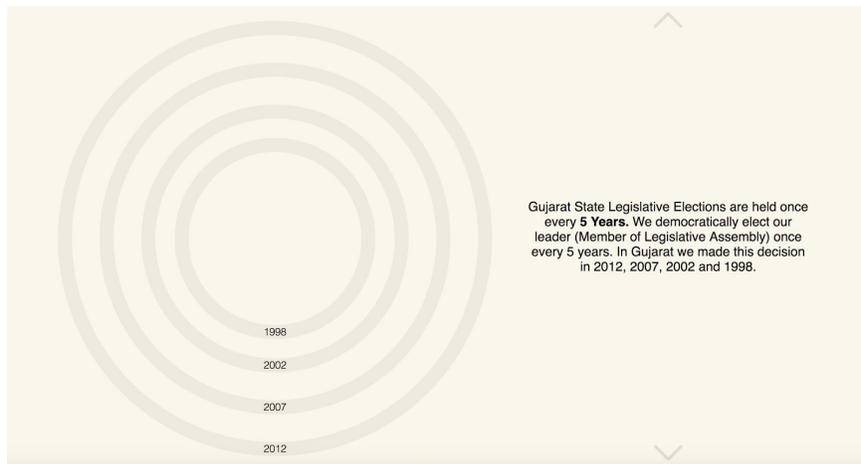
Figure 19 : Election Years

Constituency : Explanatory Segment 2 visualizes the constituencies declared in the state by the Election Commision of India every year. Every constituency is visualized using an angular sector. The starting angle and ending angle of a constituency over all the years remains constant. There are cases when certain constituencies are dissolved and new ones are created over the years. The absence of a constituency in a year is a crucial element and is represented by a blank angular sector.
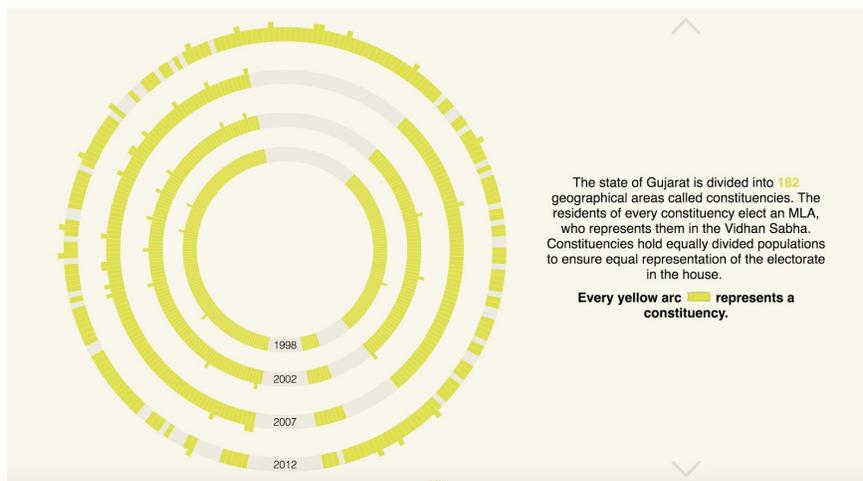

Figure 20 : Constituency Visualization

Constituency Type : Depending on the electorate population composition of a constituency a constituency is classified into three types : General, Scheduled Caste and Scheduled Tribes. (

GEN, SC and ST). This aspect of the visual is presented in the Explanatory Segment 6 and is visualized as a colored extension of the angular sector.
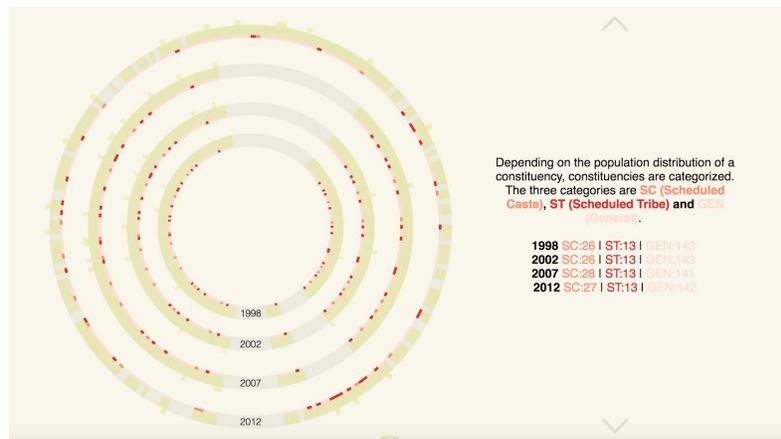
Figure 21 : Constituency Type Visualization

Winning Candidate Party : Every candidate who contests an election represents a political party. The political parties have manifestos and principles they represent, which helps a voter decide whom they want to elect. The candidate party visualization is explained in the Explanatory Segment 5.   The party of a candidate winning the election in a constituency is visualized by mapping discrete colors to the angular sectors derived by the constituency data.

Figure 22: Winning Candidate Party Visualization

Women Candidacy : Women Candidacy, is a measure of the number of women candidates who won elections in Gujarat and is represented in the Explanatory Segme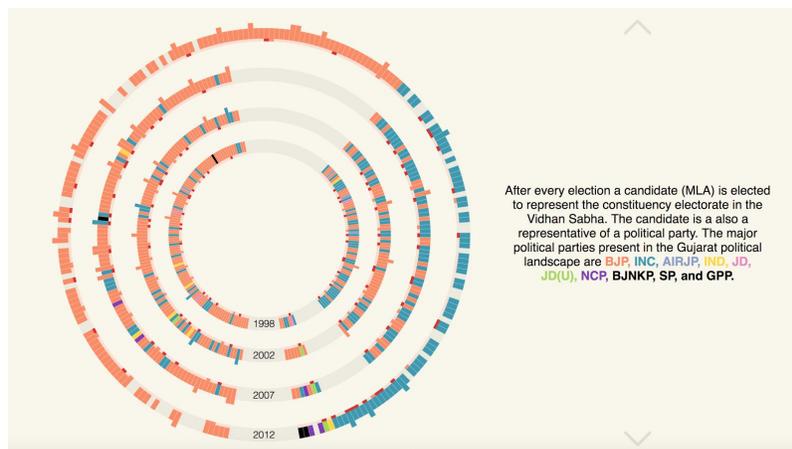nt 3. This is visualized by the length of the angular sector used to visualize the constituency. A longer arc represents women candidates and the normal arc length represents male candidates. This measure is important to be highlighted in order to educate the electorate of equal gender based representation in the state.

The representation of women in the political landscape in India is a prime concern. Over the years Gujarat has seen an increase in the number of women MLAs. Out of the 182 candidates in the year 1998, 2002, 2007 and 2012 there were 4, 12, 16 and 16 women candidates.
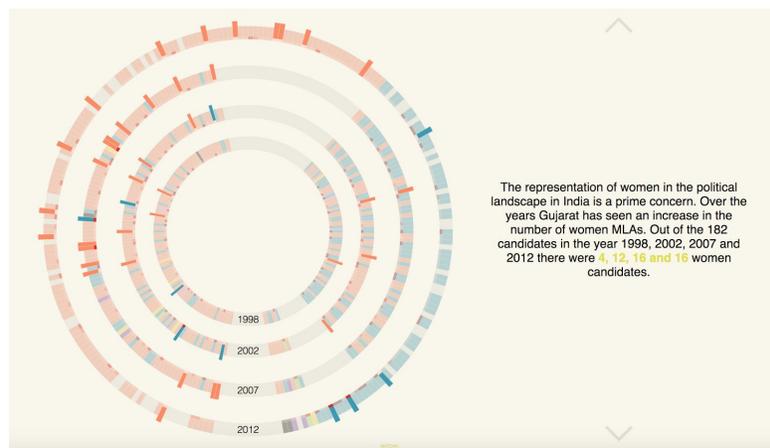
1998
2002
2007
2012

Figure 23 : Women Candidacy Visualization

Voter Turnout : Voter turnout is a measure of the what part of the electorate came out to vote on the election day. It is a measure which enables the user to understand their constituency electorate's awareness and participation in elections. A comparative low voter turnout must be highlighted to create more awareness in the constituency. This parameter is visualised using Sparklines [12] generated along the constituency data point position on the visual.
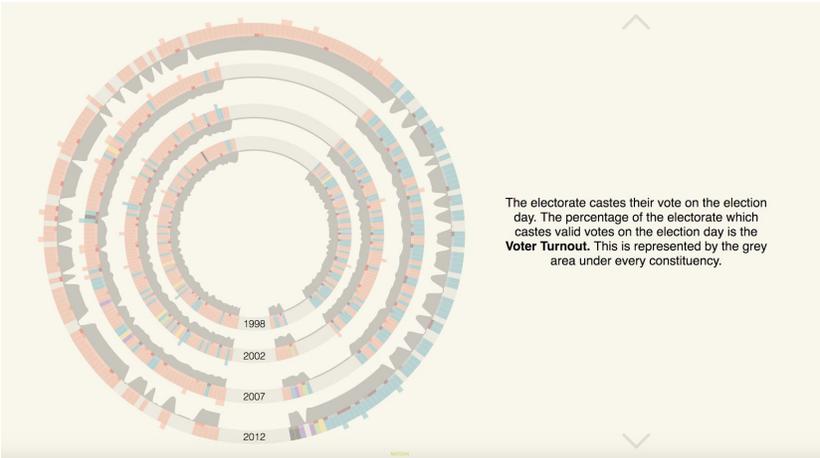
Figure 24 : Voter Turnout Visualization

The exploratory visualization allows the user to interact with the visualization and look into the measure that they would want to further explore. The user can also explore individual constituencies and the historical parameters attached with that constituency. This tool helps the user acknowledge who were the chosen leaders historically from the constituencies in Gujarat.
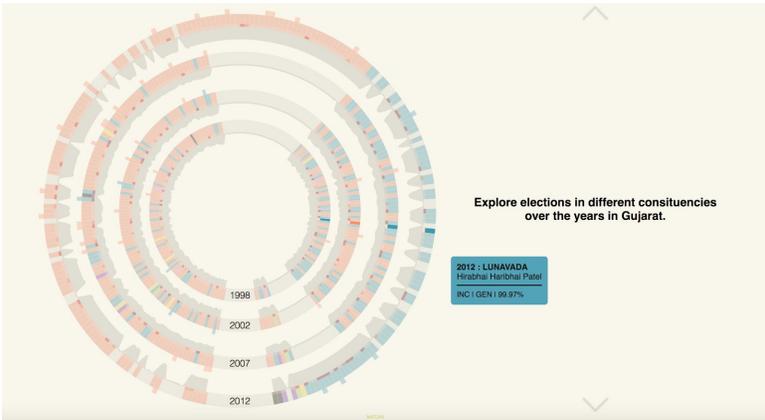

Figure 25 : Exploratory Visualization

5. Evaluation and Observations

5.1 User Study Based Insight

Project Matdan is an education tool for the electorate of Indian Democracy. The success and outreach of the project has been evaluated by conducting a user study, described in the methodology section of this document. The results and observations of this evaluation are summarised in this section.

Pre User Study

The pre user study creates a profile of users taking the user study. Every datapoint of the data collected as a part of the pre user study comprises 4 parameters. These parameters are age of the user, the nationality of the user, whether the user has voted in the past and what factors affects their voting decision the most according to them. The observations from the pre user study are listed here :

1. The age of the user's range from 19 to 51 as shown in Figure 26. The mean, median and mode of the users ages are 28, 25 and 25.

2. All the users of the study have an Indian Nationality.

3. 60% of the users have exercised their right to vote in the past.

4. 2 out of 3 users believe that the most important information for them before casting a vote is the contesting candidates profiles.
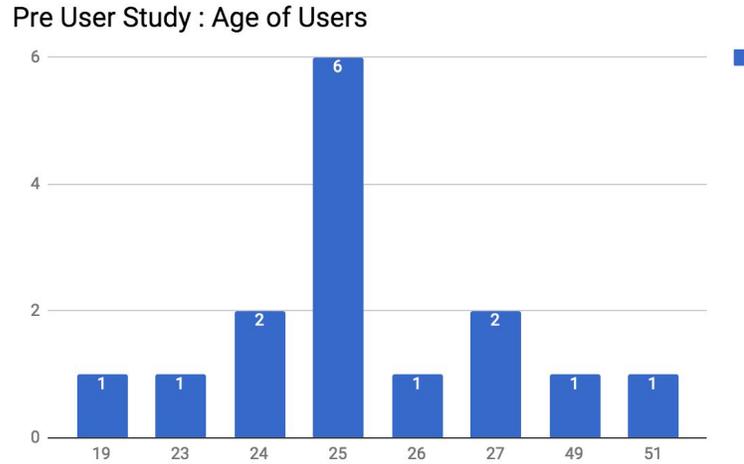
Figure 26 : Pre User Study Observation

User Study

The user study was conducted in two steps. First the user was given a small brief about the application. Then the users explored the application while their screens were recorded. The recordings were then monitored and observed to collect the user study data. The data collected from this study comprises of the time spent by the user exploring each section of the application and the features used by them in every section. Also, we observed the user's general behaviour to evaluate what could be improved in the application. The observations from the user study are listed here :

1. Users have spent the least amount of time exploring Section 3 : Historical Election Data Visualization. At the same times, users have shown a great amount of interest in exploring Data in Section 1: Current Candidate Data Visualization and Section 2: Socio Economic Data Visualization.
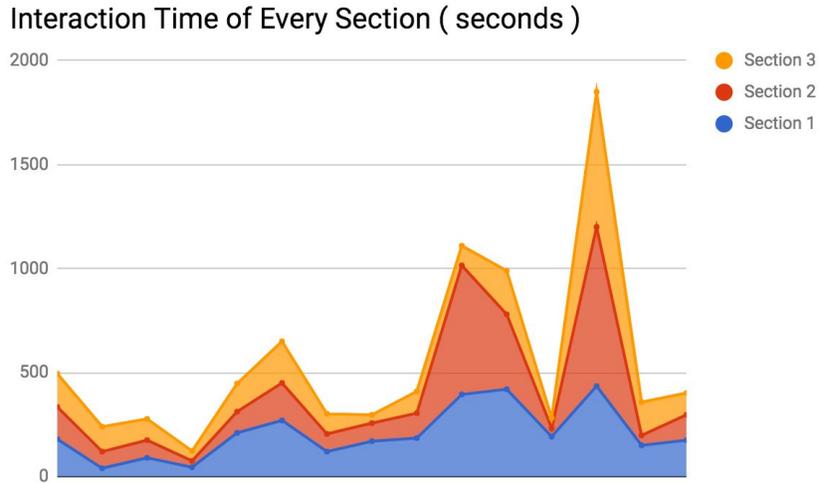
Figure 27 : User Study Observation -2

2. The user study comprises data about whether a user uses particular features in every

section or not. This enables us to understand which features can be removed or need more

work in terms of making it more user friendly and intuitive. The following visualization represents

the user interaction with the application.

S1	S1	S1	S1	S1	S2	S2	S2	S2	S3	S3
Hover Over Visualization	Filter Buttons on Visualization	Constituency Based Selection of Candidates	Hover Over Selected Candidates	Redirecting to Manifesto	Hover Over Streamgraph	Election Year Highlight	Interaction with Four Domains	Multi Parameter Viewing in Every Domain	Explanatory	Exploratory
Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No
Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Yes	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes
Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No
Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Yes	Yes	No	Yes	No	Yes	No	No	No	Yes	No
Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
1	1	2	3	6	1	4	4	5	0	3

Figure 28 : User Study Observation -1

a. Most of the features are well used by the users. The explanatory feature of the third section is used by all the users.

b. 4 features, redirecting to manifesto in section 1, election year highlight,interaction with the four domains and multi parameter viewing in every domain in section 2 is not very well used by the users and needs improvement in terms of being more user centered.

3. From the recordings it can be made out that users are spending most of the time exploring Section 1 : Current Candidate Data Visualization. There are certain UI features which need improvement to make the application more user friendly.

a. In section 1, users spend a lot of time working with selection and de-selection options for filtering the information. A Select All and Deselect All Button would help the user interact better with the application.

b. In section 3, the users at times start navigating from the top arrow instead of the bottom arrow. The direction of the narrative has to be made more clear to the user.

Post User Study

The post user study was conducted using google forms as well. The data in the post user study tells us about the user's experience of using the application. It gives us an insight about the probable outreach of this application.

1. Future Use of Application : The users were asked if they would want to use the application before casting vote the next time. This question was asked particularly for each section of the project. The following charts visualize the user output.
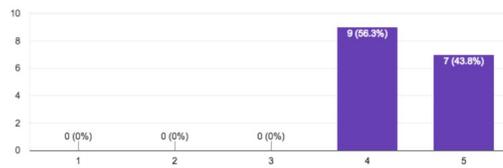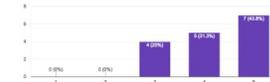
Figure 29 : Pre User Study Observation -1

From the post user study data, most of the users would want to use the application before casting or participating in an election. When we look into every section, users want to use section 1 the most , then section 2 and then section 3. Rather 25% users say that they are not very likely to use the 3rd section of the visualization. This behaviour correlates with the phenomenon identified from the user study, where a user spent most time exploring the first section compared to the other two.

2. All the user find the application to be informative. Not all users find the application to be very intuitive.
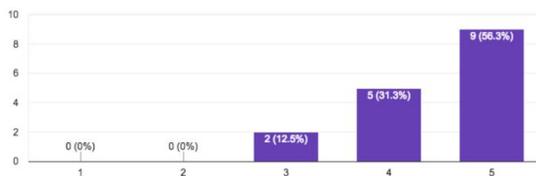
Figure 30 : Pre User Study Observation-2

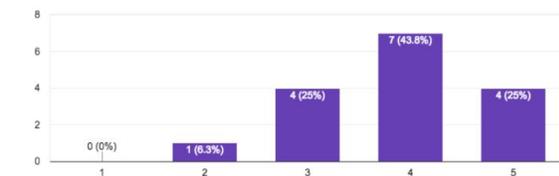3. 100% of the users think that an application like this is required for the electorate to cast an informative vote. Also all the users believe that the visualizations help them understand the data being presented better.

5.1 Data Based Insights

As mentioned earlier, the application has three main sections. In the previous section I documented evaluation an observations based on the data collected from the user study. In Data based insights I document the insights generated from each of the three sections.

Current Candidate Data Visualization

Candidates Party Representation : Each candidate contesting in the upcoming elections represents a particular political party. From the visualization we derive that BJP (Bharatiya Janata Party) and Congress have an equal representation , with 66 candidates representing each party. 40 candidates represent the BSP party. At the same time, over 108 candidates are Independent Candidates.

Candidate Financial Ownership : An evident trend is visible from the visualization, wherein candidates of the congress party own more assets compared to any other party. Upon further data investigation it is found that , 15 out of 66 candidates of Indian National Congress had assets above average. Moreover the average assets of Indian National Congress was Rs. 87,71,763 which was more compared to BJP's candidates assets average which was Rs.7,02,722.

Candidates Education : The number of candidates with a college level education representing Congress in more than BJP. Where INC has 55 candidates with a college level education, BJP has 46 and BSP has 21 candidates.

Candidates Criminal History : Criminal Records are much more prevalent among candidates representing BJP than Congress. Where INC has 6 candidates with a criminal record , BJP has 23 members with a criminal record.

Socio Economic Data Visualization

Economy : The sector wise GSDP of Gujarat has seen a definite increment over the past 30 years. The highest rate of increment is seen from 2007 onwards. Also, among all sectors, maximum GSDP is contributed from the Agriculture and Manufacturing industries.

Crime Against Women : The number of incidences of crime against women have risen over the years instead of deprecating. It is worth mentioning that the number of incidences are visualized here and not the density as it morally inappropriate to do so. An interesting observation is a peak in the number of incidence the year after every election year, which is very evident here.

Health : The infant mortality rate as well as the birth rate both have had a steady decrease over the past 30 years. The death rate has been mostly constant, except for a slight evident dip in the years 1999 and 2003. This is clear reflection of the increase in deaths in Gujarat due to the massive 2001 Gujarat earthquake as well as the 2002 Gujarat Riots.

Education : There are various 0 intersections in the visualization of all three education parameters, enrollment, number of schools and teachers. This reflects on the dearth of data

points. Also between private and government organization based education, the later is more widespread.

Historical Election Data Visualization

1. Elections are held every 5 years in Gujarat, but there was a gap of only 4 years between the 1998 and 2002 elections.

2. There are 182 constituencies in the state of Gujarat. Due to the delimitation commision of 2001, during the 2012 elections the new constituencies were created and old were demolished. This is evident in the visualization.

3. The number of constituencies belonging to non General category has increased over the years from 39 to 42 , indicating the popularity of the caste system in the Gujarat.

4. Over the years BJP has won the elections is most of the constituencies and formed the government in the state of Gujarat.

5. The representation of women candidates in the political system is an important concern. The representation has increased over the years from 4 women candidates in 1996 to 16 candidates in 2012.

6. Conclusions

Matdan helps us to address the question raised as a part of the problem statement, that there is a dearth of neutral resources which provide the relevant data which would impact an informed vote in an holistic way. The visualizations act as a catalyst and a single point of information for a varied group of data sets. The data which was present in an obscured manner in a non user friendly manner earlier, has been made interactive and has been consolidated, which facilitates a user to understand and be aware about the decision that he is making regarding the leader he wishes to elect.

In accordance with what a user/voter finds most relevant, the user study clearly points out that the visualizations incorporate the current candidate data in a wholesome and informative way. The users unanimously agree that this data representation, is a better way for information compared to the existing mediums. With the current advent of technology this medium becomes all the more relevant, not just in the sense of being informative, but also in terms of outreach because of the infiltration of internet and social media.

This project also clearly demonstrates, the advantages and disadvantages of certain visualization methods. Through multiple iterations and prototypes it was realised that due to the complex nature of overlapping data, why it is not just important to make the visualization technically correct, but also to figure out whether a certain visualization is intuitive and friendly for the targeted user or not. This opens up the field for the developer to build upon existing methods and aim to cater the user in the best way simultaneously. Building on the insights gained from technical iterations and user study, the aim is to improve the application further to cater to a bigger electorate in an even better way in the future.

7. Future Work

The Project Matdan is an unbiased step towards catalysing India's transformation from a flawed to true democracy, by educating the electorate. It is is one of the only few web based data visualization platforms to educate the Indian Electorate. The project is currently built centered around only one Election, the Gujarat 2017 Legislative Elections. It will be able to further catalyse the transformation by expanding to other elections as well in the future. Currently the app navigates to a party's manifesto when they select a particular candidate. A string analysis and visualization of a party's manifesto upon selection of a particular candidate will be a very high value addition to the project from a user's perspective. A major part of the electorate in India is not very aware of the procedures involved in successfully casting a vote. Inclusion of material educating users about this procedure in the application, will make it one stop for election related queries. Also, the outreach of the project needs to be expanded and can be very well established by integration with social media platforms. Finally the insights a user needs from socio economic parameters would be more effective is the data utilised in the socio economic parameters data was collected for every constituency instead of the whole state. In order to make this change, the data collection strategy will need modification.

The project also demands improvisations from the development and evaluation perspective. The addition of sorting and searching based user interactions will empower users more and help them comprehend and utilize the visualizations better. Also, the user study now employs direct users who would use the application. Incorporating data visualization practitioners in the user study will help highlight and improvise on the presence of data visualization principles.

8. Bibliography

[1] The Economist Intelligence Unit's Democracy Index (2016), Revenge of "deplorable" : A report by The Economist Intelligence Unit.

[2] UN Sustainable Development Goals. (2016)

http://www.un.org/sustainabledevelopment/sustainable-development-goals/

[3] Conference Reader (n.d.), International Conference On Voter Education For Inclusive, Informed and Ethical Participation.

[4] Steele, J. & Iliinsky N. (2011) Designing Data Visualizations

[5] Tufte E. (1983) A Visual Display of Quantitative Information.

[6]Jensenius, F.,Verniers, G. Indian State Assembly Election and Candidates Data (1962-Present)

[7] Kosara and Ziemkiewicz (2010), Do Mechanical Turks dream of Square Pie Charts ?

[8] Sparks (1952) Rand McNally publication : Histomap

[9] Open Government Data Platform India , A Digital India Initiative (n.d.) https://data.gov.in/

[10] Datanet India Pvt. Ltd. (2000). *Indiastat.com*.

[11] Lima, Manuel (2017), The Book of Circles : Visualizing Spheres of Knowledge

[12] Tufte, Edward (2006-2013), Beautiful Evidence

[17] Streamgraph (n.d.), https://en.wikipedia.org/wiki/Streamgraph

[18] Streamgraph (n.d.), http://visualizingrights.org/kit/charts/stream-graph.html

[19] Circle Packing (n.d.), https://en.wikipedia.org/wiki/Circle_packing_in_a_circle

[20] https://en.wikipedia.org/wiki/Data_wrangling

[21] https://developers.google.com/chart/interactive/docs/gallery/sankey

[22] Circos http://circos.ca/intro/circular_approach/

[23] Lee Byron and Martin Wattenberg (2008), Stacked Graphs - Geometry and Aesthetics

[24] https://www.timeplots.com/collections/catalog/products/a-history-of-the-us-presidency

[25] https://en.wikipedia.org/wiki/FiveThirtyEight

[26] *Zinovyev, A (n.d), Data visualization in political and social sciences*