Exploring LEGO dataset using Visualizations

Vibhor Mishra, Vipra Gupta, Ali Raza, Vilok Krishnan, Hemang Bhansal

University of Colorado Boulder

Abstract

Visualizations are an effective way of identifying the hidden patterns and new themes from a large set of data. In this paper, we explored the Lego dataset which contained millions of records and answered questions relevant to the set size and themes produced over the years. We also detailed about our design decisions while producing two interactive visualizations; chord diagram and a mix of a histogram, pie chart. From these visualizations, we were able to extract some interesting patterns in the Lego datasets explaining which set and theme were produced mostly in the last decade and vice versa. Whereas we also used a filter in the chord diagram based on the themes and years. Finally, we also explained the process of our data cleaning and joining from different tables of the Lego database.

**Keywords**: Visualization, chord diagram, Lego.

**Index Terms**: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

# Introduction

Living in the data-driven age is full of unseen challenges and it pushes us to pursue new advancements in the technology for efficiently handling the data. Big data challenge is being tackled by communicating the data through the visualization [10]. We have selected Lego dataset to explore and tackle the challenge of big data. Lego is one of the biggest brands for manufacturing toy building bricks. They are sold in the form of sets for the construction of an object. Each set consists of the different number of size, shapes, and colors. The dataset that we are exploring consists of the inventory-sets, inventory-parts, sets, themes, parts, colors. It details about all this data from the start of the company till 2017. Getting context out of this dataset is important to us because it will help us understand how the company has evolved overtime and how their preferences have changed also in terms of designs, sizes, and themes. Using visualization, we want to answer these questions below:

Q1) How top themes of LEGO sets were reused over the years?

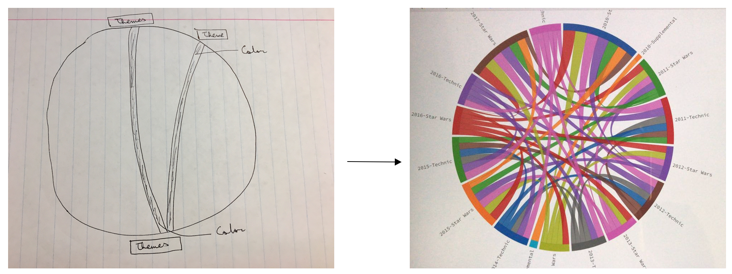
Q2) How has the size of sets changed over time?

To tackle this big data challenge which contains millions of records of the Lego sets. We tried to utilize different visualization options but then selected to implement chord diagram and histogram along with pie chart diagram to help answer our questions. It is being suggested that chord diagram is an optimal solution for representing the relationships between the large datasets [8, 9]. To identify how sets in the dataset changed overtime, we selected a combination of the histogram with the pie chart to show the difference between the sets over the years and also how we can project the top five sets that were retained over the years in the production of the Lego.

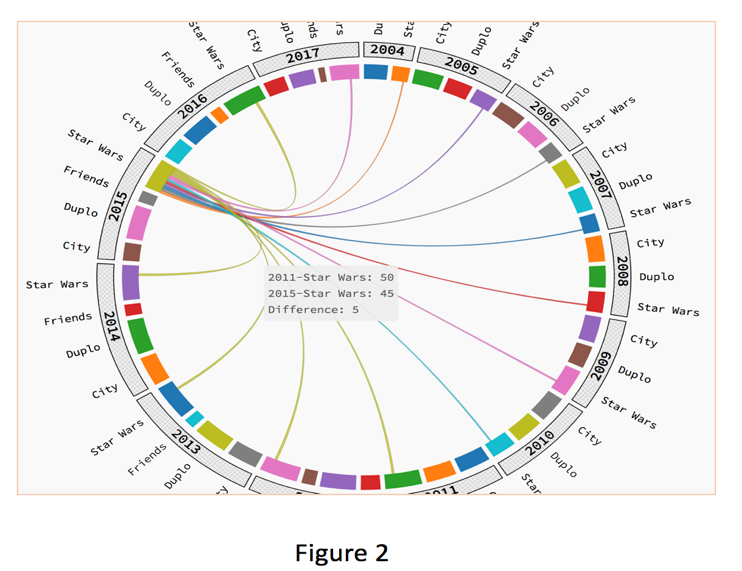
# Related Work

Graph layout has been used as a representation in biology to compare the information between the biological data. Fruchterman Rheingold graph layout was used as of the algorithm for the visualizations. The algorithm was able to create 2D and 3D visualizations that showed the difference between the protein sequencing and other visualizations. Both the graphs can be zoomed, rotated and translated. It also provided users with the option of linking the graph to the database so more information can be extracted from the links [1]. A user study was done to understand the cognitive difficulties people have to go through when they are interacting with some uncommon visualization. In this experiment, three visualizations were selected; parallel coordinate plot, chord diagram, treemap. The research participants in the research selected the chord diagram mostly for attempting to understand the data. But it made the least sense to the participants in terms of the understanding the colors of the chord whereas participants were easily able to identify the meaning about the size of the chord arcs [2]. The Chord diagram is categorized as an interactive way to show relationships within a dataset to highlight the hidden patterns. The width of the chord shows the strength of the relationships and colors show the different types of the categories. The chord diagram also has the ability to represent big data based on interactions for a year or more [3]. Chord diagrams are also used for the data analysis purpose of the protein-protein interaction in a tool called NetworkAnalyst. Along with the chord diagram, heatmaps are also used as a visualization technique for the analysis of the interactions. The visualizations allow researchers to explore the large datasets and recognize connections, patterns between these interactions so new biological hypotheses can be generated. In these chord visualizations, the arcs represent the shared genes and the chords represent the expressed genes [4]. Semantic trajectory mining was used to characterize and detect the audience of the Mobile world conference held in the Barcelona, 2012. Whereas chord diagram was used to represent the semantic origin- destination matrix of twitter trajectories. The visualization showed that before and during the conference week professional and other places trajectories are the most whereas after the conference food and shop services are increased in comparison to the other projections [5]. Chord diagrams are used in the business process management to enhance the abstraction level of the employees when they are visualizing the dense social networks. Chord diagram have shown the employees to recognize the contributions and associations of the teams that help them greatly in understanding the process management [6]. Cloudopsy is proposed as a system to help the user better understand how their personal data flows in the cloud services they are using. A Chord diagram is used to provide an overall information about the flow of the user information. Whereas chord diagrams are also used to provide information to the service provider. The system also enabled the user without any technical background to understand how their personal information is floating in the third-party cloud services [7]. The Chord diagram is also helpful in recognizing the differences, patterns of the genome types and it is providing A base for the necessary identification and analysis of the genomes. This is achieved utilizing a system called Circos, where users can also build scatterplots, heat maps, line graphs [8]. Network anomaly detection results have been displayed using the chord diagram to help the network analysts. Chord diagram was able to show a large number of detection results that include similarities and differences. By increasing the number of detections, chord diagram was able to show the relationship without affecting the readability of the visualization [9].

# Description of project and design elements

To address our first research question, we decided to create a chord diagram. A visual description of the complete process reaching on reaching out to this design decision is explained in figure 1(a) and 1(b) below. Firstly, we tried to categorize the themes that were mostly used based on the colors. The figure 1 (a) shows a mockup in which we brainstormed on how we can project this idea on the screen. Whereas Figure 1(b) shows the resulted image that we implemented.

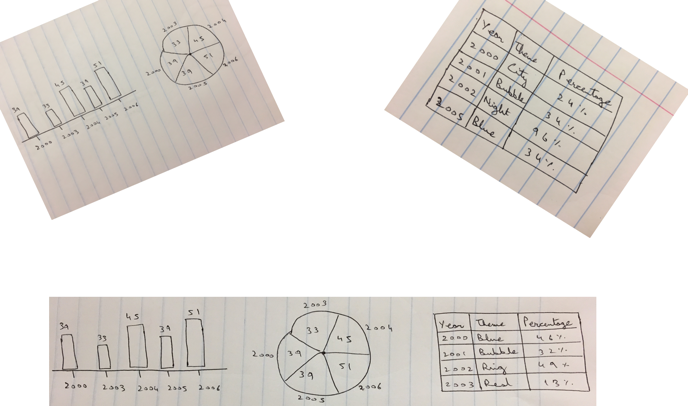
**Figure 1(a): Conceptualizing the visualization Figure 1(b): Implemented visualization**

After implementing Figure 1(b), we were unable to find any meaningful information out of this we thought about analyzing the themes based on the years. So, the resultant visualization was created, it is shown in figure 2 below.

**Figure 2: Top five themes based on years**

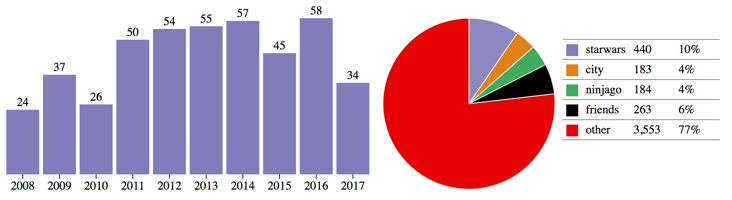
The figure 2 above is elaborates about the themes that were reused over the number of last twelve years in the Lego sets. It also explains information about the difference between the themes produced in the year selected as compared to the rest of the years. If the line is thicker on the opposite side of the theme selected it means, the year on the particular side produced more themes of it. Finally, we also applied filters on the chord diagram to provide us a detail view of the overall data based on the themes and years.

To answer our second research question, to explain how the size of sets has changed overtime. We implemented the histogram along with the pie chart that displays the top sets produced over the years and how they changed overtime. Initial mock-ups of the ideas are displayed in figure 3 below.



**. Figure 3: Conceptualizing second visualization detailing how size of sets changed**

For showing the trend of the sets that were used mostly in the Lego manufacturing, we thought it can be best projected using some mix of visualizations. But we encounter challenge on how to show merge these visualizations together and show meaningful data. The figure 4 below shows the implementation of the conceptual framework from the figure 3.



**Figure 4: Popular sets produced over the years**

The figure 4 details about the trend of the sets that were produced over the last decade in the Lego. The Histogram shows the overall sets that were produced in the last decades including the top four categories. Whereas clicking on any section of the pie chart details about the particular set that was produced over the year. In figure 4, it shows the percentage of Starwars sets produced over the last decade. Whereas in the table it explains about the top sets produced in the decade as compared to the overall sets.

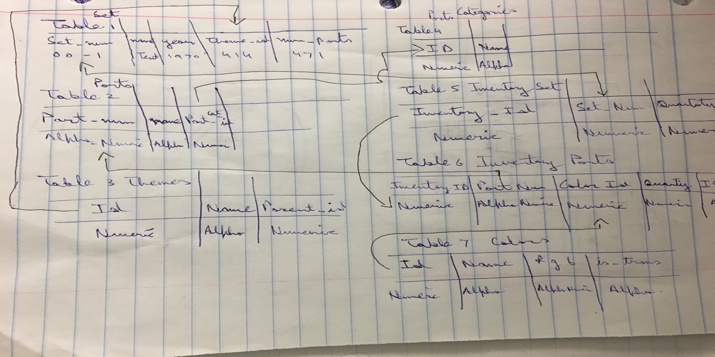
## Findings

For the chord diagram, we had to restrict our analysis to a certain data points because it was not possible to analyze overall data from the 1950s till 2017. So, we restricted ourselves to thirteen years for analyzing the chord visualization. From the chord diagram, we were able to find that Starwars was the most theme produced in the last thirteen years and it reached its maximum production during 2016. After Starwars theme, City theme was the second most common theme produced in the last thirteen years and it was produced most during the year 2014. Whereas Friends was the least theme produced during the last five years since Lego started manufacturing Friends theme from the last five years.

From the mix of the histogram, pie chart for exploring the manufacturing of sets over the years. We found that Starwars was most created individual set over the last decade. Based on overall production, it’s percentage consisted of almost ten percent. Whereas Friends was the least set produced from the start of its production from 2012-2017. Also, City set was least manufactured during the years 2012 when compared to its overall production.

## Cleaning dataset

We selected Lego database that contained millions of Lego records from more than fifty years [12]. The original dataset was sparse and we had to create a schema about the primary keys and how to merge the data together that we need to answer our questions. The figure 5 below shows the conceptual schema that was generated for each table.



**Figure 5: Schema of the dataset**

In figure 5 above, we extracted all the fields from the tables and highlighted the primary keys in them. Further, we mapped how the primary keys are working as the foreign keys in the other tables, which are highlighted using the black arrows. The schema helped us to identify which columns to merge together while creating the new table for answering our dataset.

After joining of the dataset, it contained manufacturing data of the Lego from the 1950s, but when we first mapped this data onto the visualization it showed immense values. So, we applied filters based on the themes and no of years on the chord diagram for facilitating the process of analysis and exploration.

# Discussion

We analyzed two visualizations; chord diagram and mixed histogram, pie chart based on this large dataset that we cleaned and joined together. We implemented chord diagram along with the filters; themes and years. This provides an overview of the complete dataset from the 1950s till 2017 in the chord diagram by changing parameters in one or both the filters. But analyzing visualization for the complete data was not possible because of the merging of the arcs in the chord visualization when selecting all themes and years. So, we decided to select data for the last thirteen years and top four themes that were produced. We found that Starwars was the most theme produced in the last thirteen years and reaching topmost production in 2016. The least theme produced during the last five years was the Friends. From the mixed histogram, pie chart diagram we found that Starwars is the most created individual set over the last decade. Whereas Friends was the least theme produced over the last five years. Our findings from both the visualization support the results of each other.

References

1. Enright, A. J., & Ouzounis, C. A. (2001). BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, *17*(9), 853-854.
2. Lee, S., Kim, S. H., Hung, Y. H., Lam, H., Kang, Y. A., & Yi, J. S. (2016). How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE transactions on visualization and computer graphics*, *22*(1), 499-508.
3. Keahey, T. A. (2013). Using visualization to understand big data. *IBM Business Analytics Advanced Visualisation*.
4. Xia, J., Gill, E. E., & Hancock, R. E. (2015). NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature protocols*, *10*(6), 823.
5. Gabrielli, L., Rinzivillo, S., Ronzano, F., & Villatoro, D. (2014). From tweets to semantic trajectories: mining anomalous urban mobility patterns. In *Citizen in sensor networks* (pp. 26-35). Springer, Cham.
6. Jalali, A. (2016, June). Reflections on the use of chord diagrams in social network visualization in process mining. In *Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on* (pp. 1-6). IEEE.
7. Zavou, A., Pappas, V., Kemerlis, V. P., Polychronakis, M., Portokalidis, G., & Keromytis, A. D. (2013, July). Cloudopsy: An autopsy of data flows in the cloud. In *International Conference on Human Aspects of Information Security, Privacy, and Trust* (pp. 366-375). Springer, Berlin, Heidelberg.
8. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, *19*(9), 1639-1645.
9. Mazel, J., Fontugne, R., & Fukuda, K. (2014, March). Visual comparison of network anomaly detectors with chord diagrams. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing* (pp. 473-480). ACM.
10. Keim, D., Qu, H., & Ma, K. L. (2013). Big-data visualization. *IEEE Computer Graphics and Applications*, *33*(4), 20-21.
11. Lego products, available at: <https://www.lego.com/en-us>
12. Lego database, available at: https://www.kaggle.com/rtatman/lego-database