Data Mining, Analysis and Prediction

Nirdesh Bhandari

Senior Capstone Research

May 2017

1. **Abstract**

In this survey paper I plan on looking into how data has been mined and analyzed in the past. I will look into the importance of data mining and how researchers have scraped, organized, analyzed and visualized data to catch trends and make logical predictions from them. I look at how data is analyzed using the top data mining algorithms and machine learning. I have also looked at some ways of visualizing data and why appropriate visualization can make the difference when it comes to interpretation of data.

1. **Introduction**

Data is the currency of this century**.** Withmore than 2.5 quintillion bytes of data being created every day, 90% of which has been created since 2010 alone, there exists immense policy and financial implications to manage, analyze and learn from this data [6]. In 2012 when Presidential candidates Barrack Obama and Mitt Romney debated, it generated 10 million tweets in the first two hours. Analyzing this showed the broad areas of public interest. Numerous such examples exist where data has been analyzed to test and discover things of interest. In their paper, *Data Mining With Big Data,* the authors describe the challenges of analyzing big data as being similar to having a number blind men trying to size up an elephant; each can only look at one incomplete section at a time and guess what the bigger picture looks like. The real case here, they say, is that the ‘elephant’ of big data is also continuously growing and the blind men need to constantly communicate with each other as well [6]. This survey paper aims to look at some prominent research that has been done in this field.

1. **Data Collection:**
   1. **Sources of Data for Research**

Data collection is a vital part of most academic research. Before the age of computers, data would usually be collected from repositories of government or other organizations that kept records for future reference. Other times, researchers would have to painstakingly go out and conduct surveys on population samples. With virtual storage of datasets and the onset of the Internet, researchers now have the luxury of collecting large datasets straight off the Internet. Moreover, using various scraping tools and methods of data monitoring, they are now able to harvest data straight off popular social platforms to test their theories. Let us look at how datasets have previously been categorized in the past and how each comes with its own unique advantages as well as challenges to data collection.

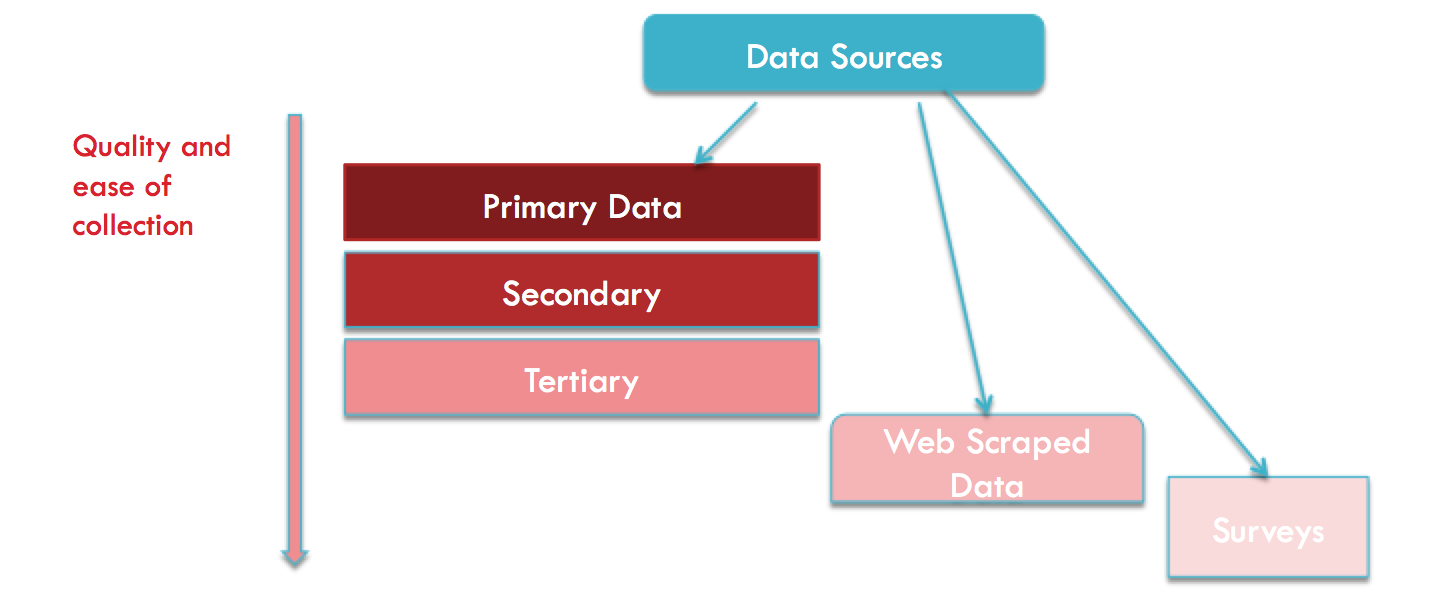


Figure: Sources of data

* + 1. **Primary data from public datasets:**

Primary data sets are structured datasets that are maintained by organizations when they can justify the cost of storing these records. These could be stored by branches within the government such as the FBI or by private companies such as Uber or Amazon. Retrieving data from these datasets is relatively easy and working on them requires minimal efforts since they will usually be stored in some logical manner [3]. For example, the supermarket chain Walmart makes over 7 billion transactions over a year. AT&T carries over 70 billion long distance calls annually. And, NASA hosts a wide range of space data from the Earth Observing System, which generates data in the order of 50 Gigabytes an hour [2]. Many such publicly available datasets are uploaded online for the general population to study. Finding and utilizing these datasets will be of top priority for me since, compared to other datasets, these data are often stored in bulk and are reliably organized.

* + 1. **Secondary data from aggregator and archive domains**:

Secondary sources include aggregator websites that often combine and store datasets from multiple primary sources. Secondary sources offer researchers the freedom to compare data side by side and are preferred over primary sources if someone has already compiled a more comprehensive dataset for which we are looking. For example, prosper.com keeps data on person to person lending, the search engine Pricewatch helps compare prices of products across different web platforms, Internet Archive stores all webpages since 1996 (around 150 billion of them), and ComScore tracks purchasing and browsing done by millions of users [3].

* + 1. **Tertiary data from network monitoring websites**

Another source of data collection can be network-monitoring websites/tools that analyze trends or changes in behavior to make logical predictions. Google trends can be one such example where trends in search patters are used to guess what topics are becoming more popular over a certain geographical area. These tertiary sources don’t provide direct or detailed data but can be useful in catching patterns if the researcher knows what correlating factors to monitor [3].

* + 1. **Experiments and Surveys**

Another method of data collection is to conduct online surveys. Online surveys have mechanized the process of creating, posting and hosting surveys for efficient data collection. The use of databases to store these data and run queries on it in order to compile research has now become a much easier task. Tools such as *Palm* Pilot or *Visor* and web platforms such as *SurveySolutions* have made things simpler for researchers [1]. Contrary to traditional survey methods where samples of people used to be asked questions to test a hypothesis, social media data now provides the opportunity to do similar research in a much more accurate and efficient manner. Users of social media leave digital traces and passive data about their behavior. Analyzing this data presents immense opportunity in finding social trends and understanding patterns across groups and nations [11].

* 1. **Data Scraping**

Often times when publicly available data sets are insufficient for the data we need, web pages and social media websites can be scraped to collect data. Automating the collection of online data not only offers the opportunity for analytically fruitful social research, but it also helps harvest from a vast mine of big data left in the form of web presence and digital traces. Using these scraped data, social scientists are now able to not only quantify the “sociological laws of human behavior” but are also able to find trends and predict social or financial instabilities. Moreover, data scraping offers researchers the ability to create structured datasets that chart the relationship between entities and attributes from extracted unstructured sources [13].

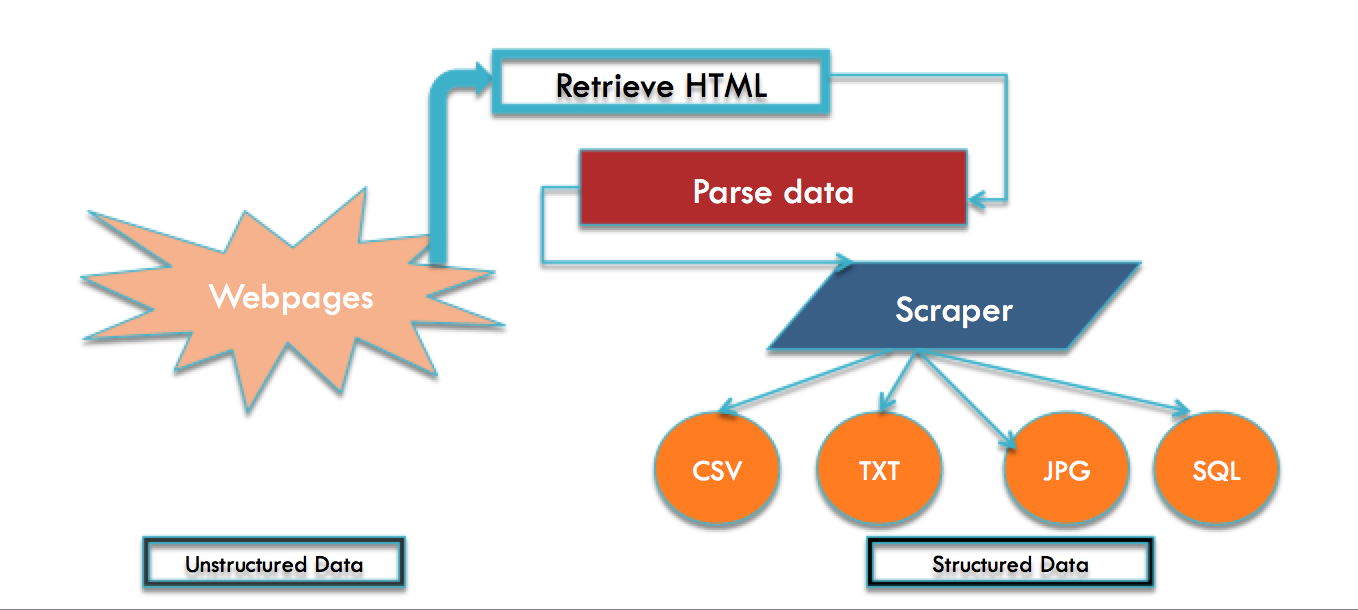
****

Figure: Data scraping process

The three steps to getting data from online sources include retrieving web pages from a web server, parsing the required extracted elements from the pages and then finally importing the data into a workable format [3].

Most commercial data scraping interfaces are rigid in their operation since they are designed to extract only a specific type of data and do not easily adapt to changes in settings in web pages. ScraperWiki is a widely used platform for developing and sharing data scrapers. It offers a great interface to design and build your own custom scraper and also includes data feed from media websites and government organizations. Data researchers have used similar platforms in the past to develop, archive and manage scrapers for efficient data collection and cross dataset analysis [13].

1. **Data Analysis and Prediction:**

In order to extract meaningful and actionable knowledge from datasets, the data needs to be analyzed using tools of analysis and statistical methods. Machine learning is the best method for data analysis; by teaching a machine to make and improve predictions after learning from a dataset, we can best improve our analysis [10]. Data analysis is a profitable industry and recently many public and private companies have started hiring data scientists within their ranks to maximize their profits. Analyzing and working with the findings of Big Data sets, McKinsey & Co. was able to save more than $300 billion in value annually. Recognizing the large scale benefits of working with Big Data, the Obama Administration dedicated $200 million in Research and Development budget to various government agencies to facilitate storing, visualizing and analyzing, large datasets. The Pentagon has pledged to spend $250 million more [5]. Analysis involves prediction, summarization, estimation and hypothesis testing.Tools such as DataWranger and OpenRefine can be used to clean the data that we have to make it easier to analyze [8].

* 1. **Statistical methods- Accuracy, errors and bias**

When working with large volumes of data, the correct statistical tools must be employed in order to accurately quantify and interpret the results. Tools such as STATA and R have now made the task of statistical analysis easier and faster. The most common statistical analysis methods include calculating the mean and standard deviation, running a regression, sample size determination and hypothesis testing.

Each of these tools comes with unique advantages and pitfalls. While mean provides us an average of the dataset and is easy to calculate, it often fails in accuracy when it comes to datasets with a large number of outliers. Standard deviations are best for understanding the distribution of data but aren’t as helpful when distribution is skewed. Likewise, regressions are excellent at quantifying the relationship between dependent and explanatory variables, but outlying data points often throw off the accuracy of the results because they ‘pull’ the regression towards them. On the other hand, sample size determination helps you analyze the large datasets using a smaller sample from the data but is highly susceptible to errors. Hypothesis testing or t-stat testing is another prominent tool that checks the statistical significance of your results. It is the most widely used statistical method and provides accurate assessments about the population dataset based on a sample as long as one is careful about testing bias or common errors [14].

Furthermore, while these are some of the most common statistical tools used, there exist other more advanced methods such as First-Difference method, tests of auto-correlation and checking for Heteroscedasticity using the Breusch-Pagan test. These are analytical tools that can be used on the data to understand and fix errors in prediction in order to obtain the most accurate evaluation [14].An example of software that uses such statistical methods would be Lex Machina. Currently being used by Palo Alto, a California based firm, Lex Machina employs big data analysis where litigators weigh input variables to make case predictions for patent cases. However, their analysis cannot be seen as being completely statistical because the machine runs experiments for individual cases against their master dataset. The final accuracy of their results depends on the litigator’s ability to make the correct assessment [7].

* 1. **Algorithms**

Machine-learning algorithms are picked based on whether the data is continuous or categorical. Supervised algorithms work on categories of data that are already known while unsupervised algorithms discover the appropriate categories during their learning process [10]. There are many popular algorithms that have been used in the past for this task of analyzing big datasets.

C4.5 algorithm is one such supervised algorithm that builds decision trees to organize, learn and predict from a given dataset. K means clustering algorithm works with large dataset by placing points into a number of clusters and then uses the distance from the center of each cluster to any point to predict the outcome. Another prominent algorithm is the support vector machine algorithm or SVM, which is a binary classifier that designs hyper planes to classify instances and predict outcomes. Likewise the Apriori algorithm is best for finding item sets and association rules between sets; it works by utilizing set laws to run multiple passes through a dataset and finds frequent pairs [10].

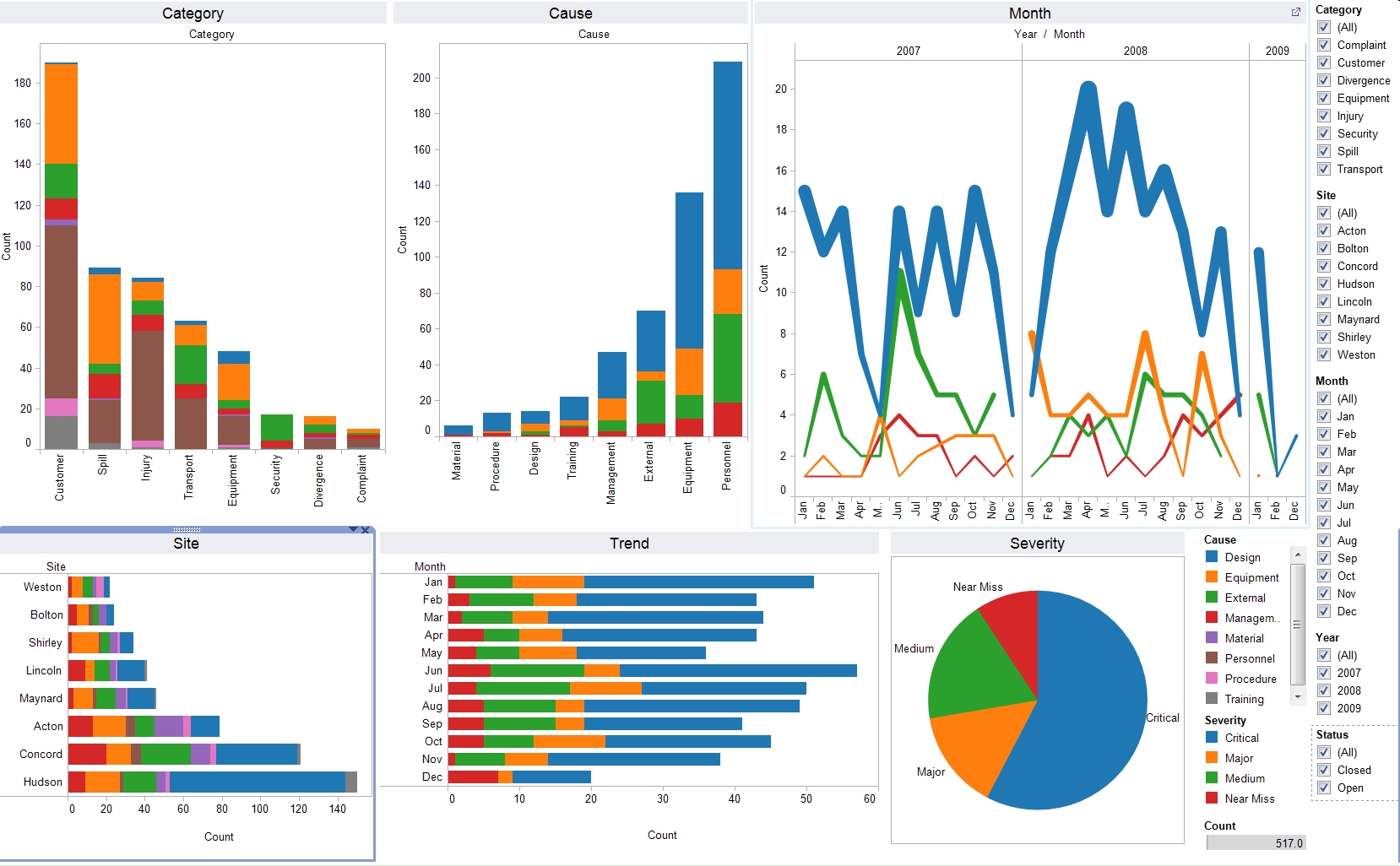
* 1. **Tools and Interface**

When it comes to working with large volumes of data, the interface and tools you use makes a difference on how accurate your tests are and how efficiently you can predict the results. While software such as Excel is widely used for general purposes because of its relatively user-friendly interface, Excel does not give users the freedom to many of the statistical tests mentioned above. This is where the R programming language comes in. R is made specifically for statistics and is a popular choice among data scientists, economists and researches. Its interface is designed for simplicity and logical operations on data. It offers python-like command options to work with data. It also helps you create graphs and charts to analyze data.

However, R can be a rather rigid interface and often lacks the required functionalities. R Studio is free open source Integrated development environment (IDE) for R. It allows you to embed R code and R output directly into pdf, document, HTML, supports LATEX and allows function extraction from a series of R command. R-Studio helps programmers automate their R code and work with uncluttered sequences on a much more customized environment. Because of these features, R-Studio is becoming increasingly popular among econometricians and data analysts [15].

1. **Visualization**

More so than often, researchers fail to properly present their findings from the data. The human eye is capable of processing visual information much faster than any other sense in our body. Our brains have a “pre-attentive visual processing” which enables us to process multiple visual elements in parallel to read information much faster. For example it is much easier and faster to find all the 3s in a series of numbers if only the 3s are in a contrasting color. Likewise, by reducing clutter in bar graphs, highlighting certain points in scatter plots or by organizing a pie chart, visualization of data can be optimized to enhance the viewer’s comprehension [12].



Source: https://digitalhumanities.duke.edu/tools/tableau

Visualization of large volumes of data with multiple variables can be rather tricky and the graphical design process behind such presentations can be a challenge for many. But, by creating visuals through automatic presentations using algebraic specification language, the task of visualization can be simplified to a great extent. Tableau is excellent software that helps automate this task of data visualization. It comes with a set of user interface commands and defaults that help created automatic presentations. Using VizQl specification language, it comes with a *“Show me”* feature that automatically generates visualizations by setting the mark and view type. Tableau offers histograms, Gantt charts, scatter plots, heat maps, and highlight tables and other visualizations for datasets [16].

**6. Conclusion and Final Remarks**

In the course of this survey paper I have looked at how data is retrieved, analyzed and visualized. I have explored the various popular sources, tools and software that are used by researchers and data analysts to find, analyze and visualize data. In order to extract accurate knowledge from data, the right statistical tools need to be employed in the right datasets to through the correct algorithms. Only then will be able to correctly learn from the data to make predictions and visualize the results. Data is the currency of the future. As the volume of data collected each day grows, so does the potential and necessity to learn from the data.

**References:**

1. Topp, N., & Pawloski, B. (2002). Online Data Collection. Journal of Science Education and Technology, 11(2), 173-178.
2. Hand, D., Blunt, G., Kelly, M., & Adams, N. (2000). Data Mining for Fun and Profit. Statistical Science, 15(2), 111-126
3. Edelman, B. (2012). Using Internet Data for Economic Research. The Journal of Economic Perspectives, 26(2), 189-206.
4. Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science, 6(1), 3-5.
5. Grose, T. (2012). Delivering Big Data. ASEE Prism, 22(4), 26-31.
6. X. Wu, X. Zhu, G.-Q. Wu, W. Ding, "Data mining with big data", *Knowledge and Data Engineering IEEE Transactions on*, vol. 26, no. 1, pp. 97-107, Jan 2014.
7. Dysart, J. (2013). The Dawn of BIG DATA: Mining the information mother lode for pricing, practice tips and predictions. ABA Journal, 99(5), 42-47.
8. Varian, H. (2014). Big Data: New Tricks for Econometrics. The Journal of Economic Perspectives,28(2), 3-27.
9. Asadi, S., Hadavandi, E., Mehmanpazir, F., & Nakhostin, M. M. (2012). Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction. Knowledge-Based Systems, 35, 245–258.
10. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37. <https://doi.org/10.1007/s10115-007-0114-2>
11. COUPER, M. P. (2011). THE FUTURE OF MODES OF DATA COLLECTION. The Public Opinion Quarterly, 75(5), 889–908.
12. Schwabish, J. A. (2014). An Economist’s Guide to Visualizing Data. The Journal of Economic Perspectives, 28(1), 209–233.
13. Marres, N., & Weltevrede, E. (2013). Scraping the social? Journal of Cultural Economy, 6(3), 313–335. https://doi.org/10.1080/17530350.2013.772070
14. Dillard, J. (n.d.). 5 Most Important Methods For Statistical Data Analysis. Retrieved May 2, 2017, from http://www.bigskyassociates.com/blog/bid/356764/5-Most-Important-Methods-For-Statistical-Data-Analysis
15. Racine, J. S. (2012). RStudio: A Platform-Independent IDE for R and Sweave. Journal of Applied Econometrics, 27(1), 167–172. https://doi.org/10.1002/jae.1278
16. Mackinlay, J., Hanrahan, P., & Stolte, C. (2007). Show Me: Automatic Presentation for Visual Analysis. IEEE Transactions on Visualization and Computer Graphics, 13(6), 1137–1144. https://doi.org/10.1109/TVCG.2007.70594