Journal of Information Technology Management

ISSN #1042-1319

A Publication of the Association of Management

A CONCEPTUAL FRAMEWORK FOR BIG DATA MODELS USING FILTRATION THEORY

ANIL AGGARWAL
UNIVERSITY OF BALTIMORE
aaggarwal@ubalt.edu

ABSTRACT

Big data and data analytics are becoming norm of the society. In the digital age, abundance of untapped structured and unstructured data is generated that need to be mined. This data may contain useful information that is still unexplored. Data are generated as soon as we step out of the house. This data is generated via social networking, sensors, mobiles, apps and many smart devices. The speed, variety and volume with which this data is generated is making current thinking obsolete. Both the private and the public sectors are thinking new ways of analyzing and using data for competitive advantage, e-services, e-information etc. -- big data analytics is an outgrowth of this thinking. Though private sector has achieved great success public sector is not far behind. Researchers have used statistical and optimization analytical technique to study various issues, however, visual analytical applications are still in their infancy. This may be due to separate streams of visualization and analytics research. Lately, both streams are being combined to create data-driven analysis using visual analytics. This paper uses filtration theory in developing a model and uses visual analytics, to study crime statistics in a metropolitan area.

Keywords: visualization, data analytics, big data

INTRODUCTION

Data is being generated in many varieties at lightning speed in large volume. Gartner group estimates, big data will grow at a 45% rate to 35 zettabytes annually by 2020. According to estimates from emarsysglobal.com, 21 billion Short Message Services (SMS) are sent and 1 billion users visit YouTube every day. 80% of online content is user generated. There are one billion Facebook members and almost 700 million Twitter users generating almost 600,000 tweets every second. This is generating interests from both public and private sectors. Figure 1 illustrates interest in four concepts over time on google.com.

As can be seen in Figure 1, interest in analytics is most profound and interest in visual analytics is just emerging. Analytics has different meaning for different users. According to Gartner group [10] data analytics is "…statistical and mathematical data analysis that clusters, segments, scores and predicts what scenarios are most likely to happen. Whatever the use cases, "analytics" has moved deeper into the business vernacular. Analytics has garnered a burgeoning interest from business and IT professionals looking to exploit huge mounds of internally generated and externally available data." Gartner group include visualization in advanced analytics and define it as, " … autonomous or semi- autonomous examination of data or content using sophisticated techniques and tools, typically beyond those of traditional business intelligence (BI), to discover deeper insights, make predictions, or generate recommendations. Advanced analytic techniques include those such as data/text mining, machine learning, pattern matching, forecasting, visualization, semantic analysis, sentiment analysis, network and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, neural networks". In addition, they project analytics application will become more mission critical in the future. Typically visualization and analytics research is done separately, however these two fields are merging giving interactive power to users. Following Gartner, we define visual analytics as, "…Data–driven analysis that facilitates visual interactive exploration and explanation using data analytic techniques".

The next section describes big data, followed by the conceptual model. Last section describes an application of crime statistics in a metropolitan area.

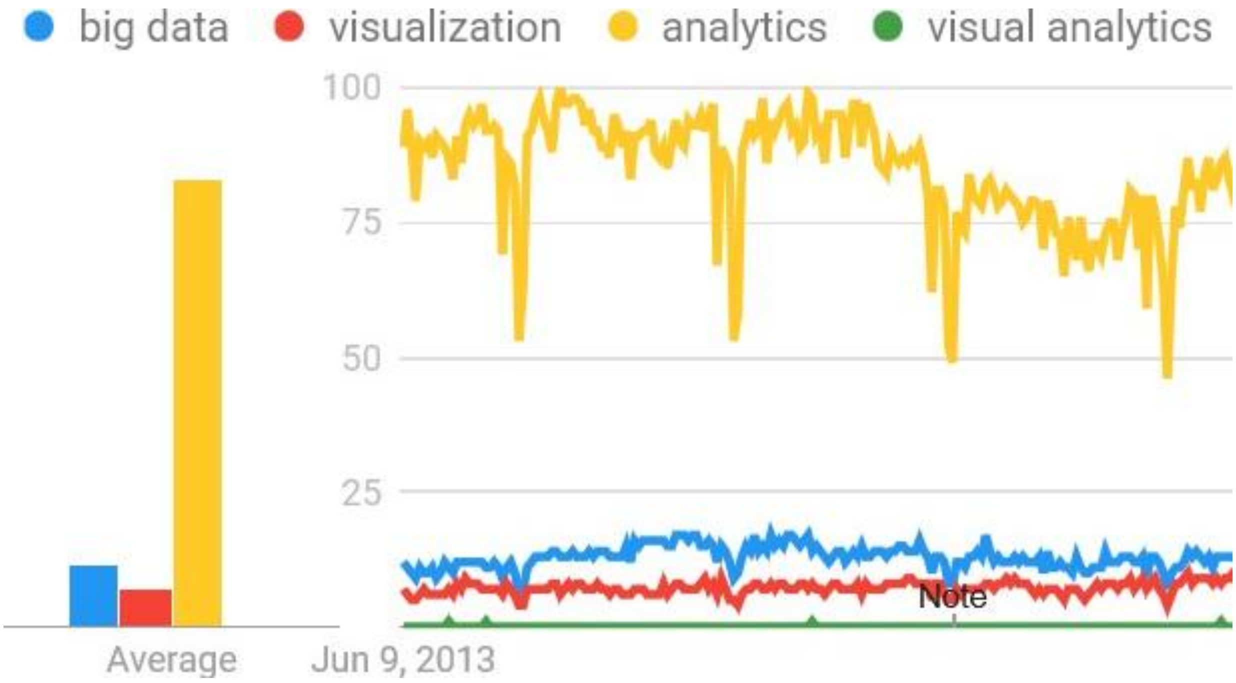


Figure 1: Interest in four concepts over time

BIG DATA

Big data is "big" but beyond that it is still a mystery. There is some agreement on the terms that characterize big data as, volume, variety, velocity and veracity. The first three Vs (velocity, volume and variety) refer to "raw" data as it is generated, whereas the fourth V, veracity, refers to the source or truthfulness of data, which is a debatable term. Veracity aspect can be different for different individuals since individual have their own frame of reference for veracity. According to the National Institute of Standards and Technology (NIST) [24], "Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets." In addition, "Big Data consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis". It appears that common are the three factors (volume, velocity and variety). Researchers have argued that new methodologies are needed. Not all data will have all three properties. We provide some examples of big data in Table 1.

Examples in Table 1 are not necessarily mutually exclusive in terms of the three V's (volume, variety and velocity), but based rather on factors that may be dominant in a given situation. Not all data contains all components of big data characteristics. Filtration model provides step by step convergence towards a goal of creating value. Though, Filtration model has been used in social context it can also be extended/modified to filter data. Steps identified in filtration can also be useful in big data analysis. The next section describes a conceptual model using filtration theory.

Table 1: Examples of Big Data

The Vs of Big data	Typical Examples	Vs intensity	Examples of data
Volume/ Variety/ Velocity	Celebrity following	Low volume, Low velocity, Low variety	celebrities tweeting, retweeting in text, pictures, (typically one way communication)
Velocity/ Volume	Tragedies in schools	High volume High velocity (via social network sites)	Interviews with school children, authorities
Variety	Tracking criminals	High Variety	Sensor data, biometric data, social network feeds

THE FILTRATION MODEL

Current data storage, manipulation and processing techniques are not suitable for big data as new Data is both document and user generated and consists of both structured and unstructured formats. User generated data is creating opportunities for organizations and governments to make their services more user-friendly. Initial concept of 3V's was extended by several others V's. For example, Veracity, Validity and Values were introduced by several authors and companies like IBM, SAS and Oracle. We will discuss several V's in developing our model.

We use filtration theory proposed by Klenke et al. [16]. Filtration theory has been applied in context of social culture and selection (filtration) of mate. They defined following filters:

1. Married Couples → Eligible Partner
2. Compatibility Filter → Eligible Individuals Attracted to Each Other
3. Physical Attractiveness Filter → Homogamous Potential Partners
4. Similar and Complementary Views Filter
5. Potential Field of Partners
6. People Who Live in Proximity → Total Field of Potential Partner

Parts 5 and 6 relate to outcome and actions that create "value" after the first four filter have been applied. Though they do not directly apply to big data we can, however, see similarities in the two. Next section discusses similarities in applying filtration model to big data. Aggarwal [1] proposed a model for big data analysis. The model, however did not extend to include data analytics. We have modified Aggarwal [1] hybrid model by adding a value step. This step combines filtered data with analytics. Figure 2 summarizes the conceptual model. Following sections discuss various filters.

Filter 1: Domain-related filter (Eligible partner and compatibility)

This filter checks for domain compatibility. Social media is generating lots of data some related and some unrelated to topics. It is not uncommon to find data are that are mixed with personal data which may include personal promotions, new product offerings, job applications etc. In general, these postings may have nothing to do with the problem at hand. We suggest this filter should be used to remove "noise" to create eligible and compatible data. Cheng et al [7] propose a Content-based Collaborative Filtering approach (CCF) to bring both Content- based Filtering and Collaborative Filtering approaches together and suggest using CCF to filter and to inspect rich contexts of the recommended items. It is recommended only eliminate noise from a web user profile but learn prior to elimination. Onyancha et al [25] propose a machine learning model that removes noise from data to improve quality of (in this case) user profile. HADOOP and Mapreduce could also be useful in removing content-based noise using text analytics, key word or sematic analysis.

Domain related filtering does make data more useful this, however, alone may not be enough to generate meaningful values/results. Data needs to be further filtered/cleaned to make it meaningful. Next section describes "truth" and "fake" data filtration.

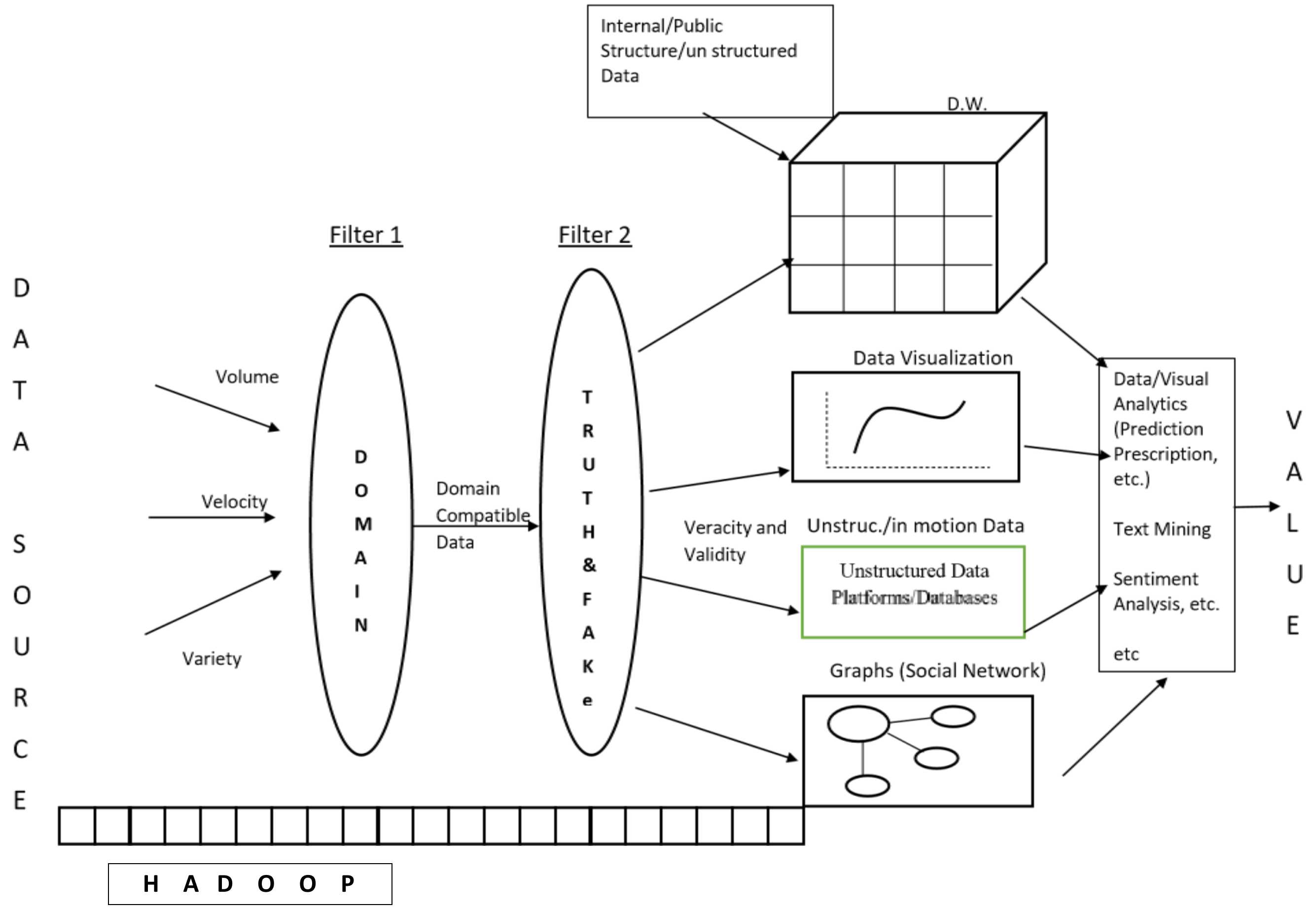


Figure 2: A Filtration Model for Big Data

Filter 2: Truth & Fake Filter (Homogamous & Similarity)

Once data becomes domain compatible there is still problem of "truth" and "reliability" of data. Given the "open" nature of social media it is possible to post fake and untruth data. Data can be fake in the sense it does not exist and/or it can be untruth because truth is known or visible. Individuals post data/text which is fake or untrue. This may be due to the need for inclusion, deception or financial gain. Agichtein et al. [3] noted, "The quality of user-generated content varies drastically from excellent to abuse and spam. As the availability of such content increases, the task of identifying high-quality content sites based on user contributions -- social media sites -- becomes increasingly important".

Purpose of this step is to generate "trust" which is important factor in any data study. User must be assured that data is from trusted sources. There are many generally accepted government (Department of Labor, Human and Health Services, CDC etc.) and private (IBM, SAS, STARBUCK, NYTimes, Washington Post, Wall Street Journals, etc.) trusted sources. However, big data also comes from many open public sources. This data may contain Fake accounts and fake postings, as recently has been reported. (Possible Russian troll interference in US elections, BritExit, Europe elections using fake accounts and postings, etc.) There may be paid/unpaid people posting reviews of movies they have never seen or products they have never used. This data would bias the results and must be filtered. Untruth data is reported by individuals about other individuals, which in many cases have led to suicides.[1] Rubin et al [29] describe types of

[1] https://www.nytimes.com/2016/09/10/nyregion/conviction-thrown-out-for-rutgers-student-in-tylerclementi-case.html

fake news that analyst should be aware of. Given these issues, it is important to remove fake and unreliable data from domain compatible data. Fake data is usually generated by weird host names or names very similar to known names. These hosts can easily be separated/flagged by using known key words. Researchers have used regression analysis and other techniques (Morris [20]), to identify important contents and source based features, which can predict the credibility of information in a tweet. Zubiaga et al. [35] developed a rumor classification system that consists of four components: rumor detection, rumor tracking, rumor stance classification and rumor veracity classification. Gupta et al. [12] adopted a supervised machine learning and relevance feedback approach using the above features, to rank tweets according to their credibility score. Castillo et. al [6] analyzed microblog and classified the postings as credible or not credible, based on features extracted from them. They used features from user posting and re-posting ("*retweeting*") behavior, from the text of the posts, and from citations to external sources. Martens et al. [22]. provide an overview of the relevant economic research literature on the digital transformation of news markets and the impact on the quality of news. Mao et al. [21] present a model to investigate information spreading over cyber-social network of agents communicating with each other. Del et. al. [9] introduce a general framework for promptly identifying polarizing content on social media and, thus, "predicting" future fake news topics. They use Italian Facebook postings and identify topics that are susceptible to misinformation with 77% accuracy. Pennycook et al. [26] used crowdsourcing to gauge user's ratings of news outlet trustworthiness. They found ”… rather than being initially agnostic about unfamiliar sources, people are initially skeptical – and thus a lack of familiarity is an important cue for untrustworthiness”. In addition, many techniques are emerging that suggest how to remove fake data (see: https://carloseo.com/removing-google-analyticsspam/; https://www.optimizesmart.com/geek-guideremoving-referrer-spam-google-analytics/ etc.)

Once data is verified and validated, the next step is to analyze it to create alternatives and eventually create “value” (comparable to filtration theory steps of people who live in proximity and eventual partner). Next section describes analytics that can be used in value creation.

Value Creation

This step is similar to finding potential match (alternate analysis) and finding a suitable match (developing a solution/value) of the filtration theory. Value creation implies using analytics to create meaningful insights like trends, likes, sentiments etc. We discuss several methodologies to achieve this.

Once data is cleaned of noise analysts can created value using analytics techniques. Data can be used for explanation, opportunities, challenges sentiments, exploration, hidden relations, tracking etc. There are infinite opportunities, however one should be careful of “fake’ relationships. With more data there are more chances of fake relationships. Meier, et al. [19] apply measurement theory to administrators' self-perceptions of organizational performance and show that spurious results are not only common but that they might include as many as 50% of all statistical tests. In addition, Calude et al. [5] argue correlations in big data are only due to size and in most cases “spurious”. Creating value can be quite challenging. Spurious relations can emerge in big data which can lead to erroneous conclusions. Researchers have argued against data deducted relationships without any theory or modeling background. Hosni et al. [14] suggest that the simple-minded ideas are not tenable. They suggested that data science performs at its best when coupled with the subtle art of modelling. L'heureux et al.[17] discuss emerging machine learning approaches and highlight the cause-effect relationship by organizing challenges according to Big Data Vs or dimensions that instigated the issue: volume, velocity, variety, or veracity. They also argue in favor of data modeling. Big data, however, provides opportunities that were not available before. Sentiment analysis can be used to study citizen's reaction to “rate hikes” to “new treaties” to “Brit exit”. Visual analytics can be used to explore complex relationships or to explain causes of disease spread. Social network analysis can help government agencies to track/trace suspects and their accomplices. Numerous analytical techniques are available and can be used to create values. We do not discuss these data analytics since most are well established in management science and statistical literature. Due to space constraint, we only briefly describe following platforms/options that can be used to create values.

- Data warehouse
- Social networking
- Unstructured/ in motion
- Data Visualization

Data Warehouse

In many cases unstructured data can be converted to structured data. For example, locations can be converted into longitude and latitude. Pictures can be converted as jpeg addresses. This data can be combined with structured data and can be used in data warehouse for further analysis and data mining. Cheng et al. [7] discuss a system which consists of a data collection layer with a unified standard, a data management layer for distributed storage and parallel computing, and a data-oriented service layer. Zhang et al. [33] argue that the technologies of cloud and big data can be used to enhance the performance of the healthcare system so that humans can then enjoy various smart healthcare applications and services. There are many examples where researchers have integrated structured/unstructured data to create value (Aslam,et al. [4]; Chin-Ho, et al. [8]). Data warehouse is a developed field (see Inmon, [15]; Mankad, [18] etc.) for additional information).

Social Networking

Social networks provide connectivity information. It can be used by companies to see how new products are being viewed, solve customer complaints, launch new products and many other citizen-centric applications. Social networks are open and everybody has equal access which is a boon but also a problem. Analyst should be careful about fake postings and troll that may pass through even after filtration. Zannettou, et al. [32] provide examples of trolls and their impact on fake news.

However social media also provides important information about communities and is helpful to government agents in tracking undesirables. Companies and governments can mine data to find sentiments, group workings, loner and influential people in the networks by analyzing graphs available in popular sites like Twitter, Facebook, Instagram, Pinerest etc. Much research has been done in this area. (see Glasman et al. [11] etc. for further details).

Unstructured/ Streaming Data

Many times data cannot be converted in structured form. In such cases different analytics are needed. There are several new concepts and databases are emerging for processing unstructured data. (Dynamo, MongoDB, HIVE, PIG, NOSQL etc.) For data in motion Hirzel et al. [13] describe IBM approach (developed a language) that can process data in motion. Zikopolous et al. [34] describe IBM InfoSphere BigInsights (Big Data at rest) and IBM InfoSphere Streams (Big Data in motion) technologies. This platform can process both streaming and data at rest. Zaharia et al. [31] describe Apache's SPARK system for processing streaming data. Many companies are developing integrated platforms that can process structured/unstructured and streaming data simultaneously. This field is still developing and new systems are emerging every day.

Data Visualization

It is well known that "A picture is worth a thousand words". Visualization involves two parts, exploration and explanation. Both parts are important to understanding system's behavior. As previously mentioned, in early stages visualization and analytics were two separate research areas, however new technologies are combining the two and creating a new environment called visual analytics. This not only allows analyst to create visuals but also allows them to create a "story" that can explain a behavior, cause or opportunity. Visualization tools like (treemap, heatmap, pie charts, bars, 3-D pictures, gantt chart, polygon etc.) can be combined with analytics (cluster analysis, decision tree, RFM, Neural networks etc.) tools to explore, analyze data and provide possible values. Deak (https://public.tableau.com/enus/s/gallery/mapping-1854-cholera-outbreak) created a visualization to explain great cholera outbreak that killed 616 people. By using visuals and creating a story he was able to explain the cause of outbreak to a faulty pump. Some other examples are available where visualization has helped solve or analyze a problem (Molesworth et. al. [23], Varshney et. al., [30]; Ranganathan et al[28]; Polwart, [27] etc.) Visual analytics is also being used for real time reporting. For example, yahoo.com; Cnbc.com etc. shows heat map of stock markets in real time.

Visual analytic applications are beginning to show results and it is not far when many more application of big data will emerge in both public and private sector.

The approaches described above are used to create value. However, these are not exhaustive or mutually exclusive. Isolated approaches may not produce desired results. Analysts must select approach(s) that will suit the problem at hand. Many times these approaches are used together to provide best results. The filtration model described above is the first step in providing a systematic process of filtering big data to make it useful. This is not an end in itself but only the beginning. It is a moving target. The challenge is to select trusted "source(s)", apply appropriate filters and correct "analytics" to create value for your audience at any given point in time.

The next section describes an application that implicitly uses the visual analytics of filtration model in its application.

FILTRATION MODEL APPLICATION USING VISUAL ANALYTICS: METROPOLITAN ISSUES

Many urban cities are facing crime issues. Citizens are looking at government for safety, accountability and privacy. Cities want to provide transparency and make their services more citizen-centric. We look at one such city to study real time crime statistics and develop a story for the audience (imaginary high level). Crime data is available through city's web site. Data could be a combination of variety of data sources, from police cameras, to street cameras, to GPS to simple data entry by police. Data is combined in one file and contains 313,204 records. Location was translated as latitude and longitude to allow geographic display of data. No filters were applied here since data already was from a trusted source (the city) and there was no fake data (since it was entered by police and automatically from crime sites). In this example data is already filtered for first two filters. The third step value creation was done using visualization. Data was explored using visual analytics. After analysis a story line was created for high level city officials. A dashboard was created to show crime summary by month, district and serious crimes. Figure 3 shows the dashboard.

As can be seen in the Figure 3 dashboard, crime has decreased from 2012 (May) to 2018 (May), a decrease of 16% (from 4648 to 3891). This is, however, still a staggering number. Data was further explored for time of crime and districts of crime. It appears serious crime (shooting) occurs between midnight to 2 pm in western and eastern districts. Table 2 shows serious crime by District.

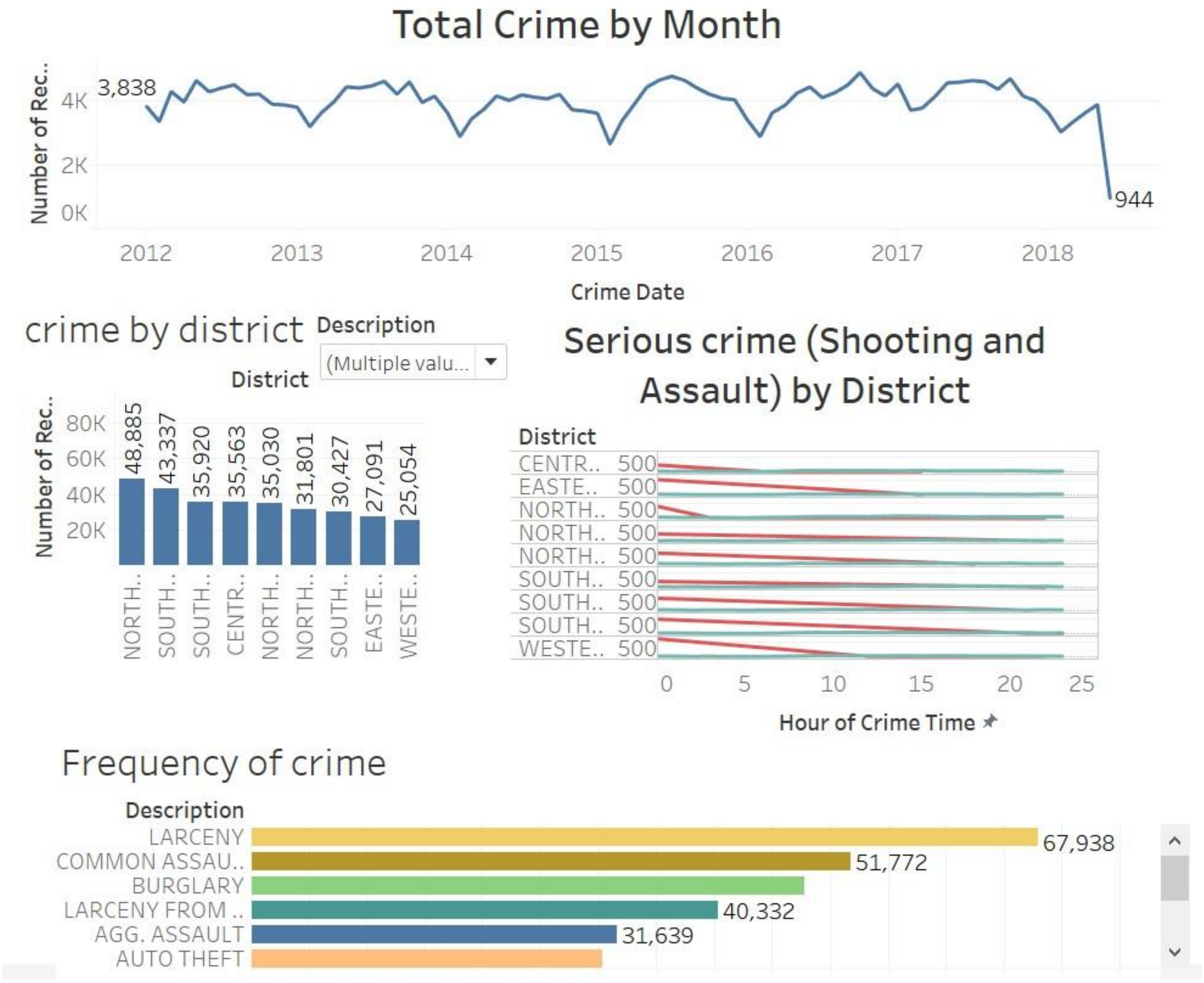
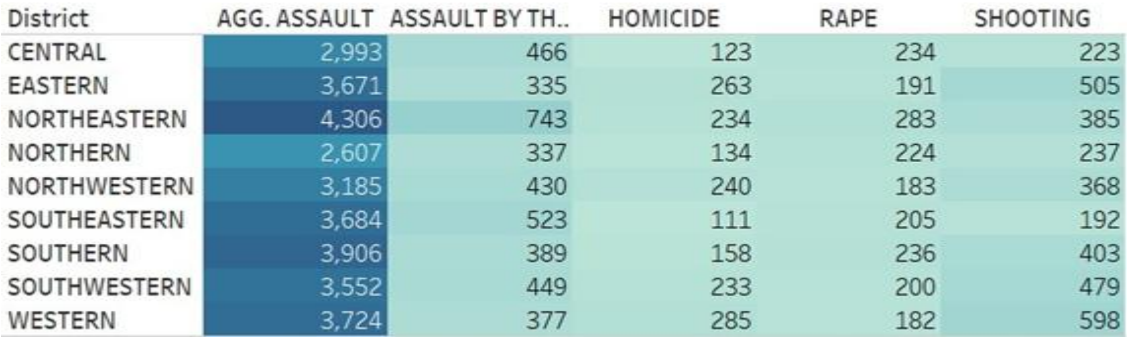


Figure 3: Interactive Dashboard

Table 2: Serious crime by District



District	AGG. ASSAULT	ASSAULT BY TH..	HOMICIDE	RAPE	SHOOTING
CENTRAL	2,993	466	123	234	223
EASTERN	3,671	335	263	191	505
NORTHEASTERN	4,306	743	234	283	385
NORTHERN	2,607	337	134	224	237
NORTHWESTERN	3,185	430	240	183	368
SOUTHEASTERN	3,684	523	111	205	192
SOUTHERN	3,906	389	158	236	403
SOUTHWESTERN	3,552	449	233	200	479
WESTERN	3,724	377	285	182	598

Additional analysis shows theses districts do not have highest crime rate but highest serious crime rate. Larceny and common assault occur at the highest rate in the city. Almost 26000 crimes were committed by firearms (see Figure 4).

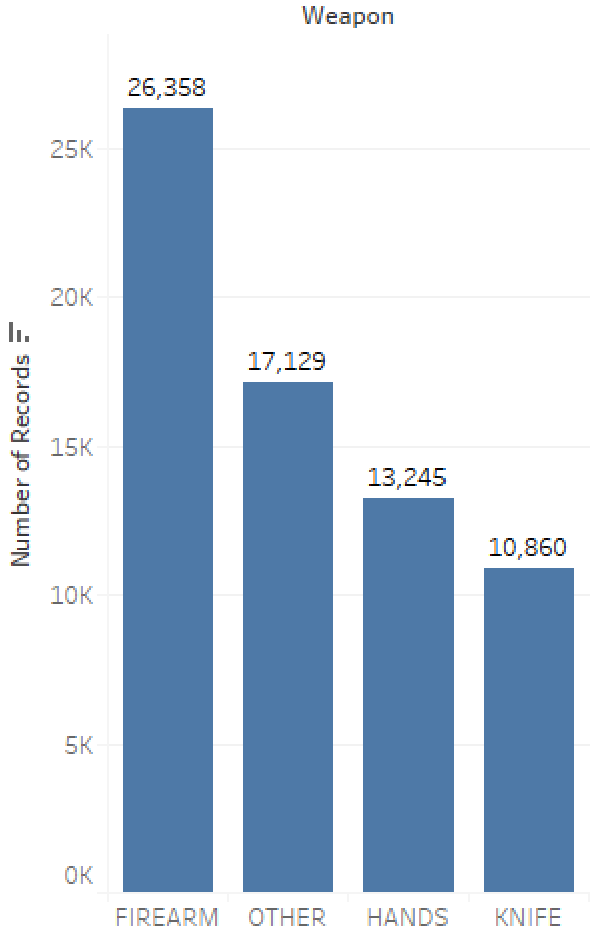


Figure 4: Weapons Used

There are disturbing signs of crime in the city. Gun control should be enforced vigorously and police petrol should be increased in western and eastern districts where serious crime is high. Additional data is needed to get the profiles of criminals, why are these crime happening? Is the unemployed youth committing most of these crimes? What can be done to address these issues? This additional data may reveal who is committing crime and what programs money may be used to deter crime.

CONCLUSION

Big data is coming increasingly under scrutiny because of ease of availability. This has both dark and bright side. Dark side deals with crime, deception, fake news and terrorism whereas bright uses it to provide customer and citizen-centric services by providing transparency to the whole process. When used appropriately big data can be beneficial to every customer and citizen. Big data is providing opportunities and every company should take advantage of it. Currently, however, there are no models that provide a systematic approach for managing such data. This paper provides a filtration model which combines structured with unstructured data to generate value for its users. In addition, an application was analyzed using visualization which provided important insights into crime statistics which can be used for policy making.

As internet of things start to generate more data (Ahmed [2]) filtration will become even more important. We do not discuss the technical details as they are beyond the scope of this paper. However, we have outlined the first step in big data modeling, and hopefully a standardized approach/model will eventually emerge from future discussions and research.

REFERENCES

[1] Aggarwal, A. K. (2016). A hybrid Approach to Big data Systems development, . *Managing Big Data Integration in the Public Sector*. IGI Pub.

[2] Ahmed, F. (2017). Implementation of Smart Cities under IoT & Big data Analytics. *IJCSNS*, *17*(10), 153.

[3] Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 183194). ACM.

[4] Aslam, J., Lim, S., Pan, X., & Rus, D. (2012). City-scale traffic estimation from a roving sensor network. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems* (pp. 141 -154). ACM.

[5] Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of science*, *22*(3), 595612.

[6] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 675-684). ACM.

[7] Cheng, W., Yin, G., Dong, Y., Dong, H., & Zhang, W. (2016). Collaborative filtering recommendation on users' interest sequences. *PloS one*, *11*(5), e0155739.

[8] Chin-Ho Lin; Liang-Cheng Huang; Chou, S.-C.T.; Chih-Ho Liu; Han-Fang Cheng; I-Jen Chiang, Temporal Event Tracing on Big Healthcare Data Analytics," (2014). *Big Data (BigData Congress), 2014 IEEE International Congress on* , vol., no., pp.281,287.

[9] Del Vicario, M., Quattrociocchi, W., Scala, A., & Zollo, F.(2018). Polarization and fake news: Early warning of Potential misinformation targets. *arXiv preprint arXiv:1802.01400*.

[10] Gartner.com, https://www.gartner.com/itglossary/analytics/

[11] Glasman, L. R., & Albarracin, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. *Psychological bulletin*, *132*(5), 778.

[12] Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. In *Proceedings of the 1st workshop on privacy and security in online social media* (p. 2). ACM.

[13] Hirzel, Martin, Henrique Andrade, Bugra Gedik, Gabriela Jacques-Silva, Rohit Khandekar, Vibhore Kumar, Mark Mendell et al.(2013). "IBM streams processing language: Analyzing big data in motion." *IBM Journal of Research and Development* 57, no. 3/4: 7-1.

[14] Hosni, H., & Vulpiani, A. (2018). Data science and the art of modelling. *Lettera Matematica*, 1-9.

[15] Inmon, B. (2012) *Building the Data Warehouse*. 1st Edition. Wiley and Sons.

[16] Klenke, Karin. (1981). Exploring human sexuality, New York, Van Nostrand Pub.

[17] L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access*, *5*, 7776-7797.

[18] Mankad, M. D., & Dholakia, M. P. (2013). The Study on Data Warehouse Design and Usage. *International Journal of Scientific and Research Publications*, *3*(3).

[19] Meier, K. J., & O'Toole, L. J. (2012). Subjective organizational performance and measurement error: Common source bias and spurious relationships. *Journal of Public Administration Research and Theory*, *23*(2), 429456.

[20] Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012, February). Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 441-450). ACM.

[21] Mao, Y., Bolouki, S., & Akyol, E. (2018). Spread of Information with Confirmation Bias in Cyber-Social Networks. *arXiv preprint arXiv:1803.06377*.

[22] Martens, B., Aguiar, L., Gómez, E., & Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/digital-transformation-news-media-and-rise-disinformation-and-fake-news

[23] Molesworth AM, Cuevas LE, Connor SJ, Morse AP, Thomson MC. (2003). Environmental risk and meningitis epidemics in Africa. *Emerg Infect Dis*. 9(10):1287-93.

[24] The National Institute of Standards and Technology (NIST) (2015) https://101.datascience.community/2015/04/23/nist-defines-big-data-and-data-science/

[25] Onyancha, J., & Plekhanova, V. (2017). A user-centric approach towards learning noise in web data. In Intelligent Systems and Knowledge Engineering (ISKE), 2017 12th International conference on (pp. 1-6). IEEE.

[26] Pennycook, G., & Rand, D. G. (2018). Crowdsourcing Judgments of News Source Quality.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=31184 71

[27] Polwart, N. Mobile Health Apps Have Role in Ebola Crisis, Information week (2014), available at: http://www.informationweek.com/healthcare/mobile-andwireless/mobile-health-apps-have-role-in-ebola-crisis/a/did/1306617

[28] Ranganathan, S., Nicolis, S. C., Spaiser, V., & Sumpter, D. J. (2015). Understanding Democracy and Development Traps Using a Data-Driven Approach. *Big Data*, *3*(1), 2233.

[29] Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

[30] Varshney, K. R., G. H. Chen, B. Abelson, K. Nowocin, V. Sakhrani, L. Xu, and B. L. Spatocco.(2015). Targeting Villages for Rural Development Using Satellite Image Analysis. *Big Data*

[31] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Ghodsi, A. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56-65.

[32] Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. *arXiv preprint arXiv:1801.09288*.

[33] Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M., & Alamri, A. (2017). Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Systems Journal*, *11*(1), 88-95.

[34] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

[35] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, *51*(2), 32.

AUTHOR BIOGRAPHY

Anil Aggarwal is a Professor in the Merrick School of Business at the University of Baltimore. Dr. Aggarwal was a Fulbright scholar and held Lockheed Martin Research and BGE Chair at the University of Baltimore. He has published in many journals, including Computers and Operations Research, Decision Sciences, Information and Management, Production and Operation Management, e-Service, Decision Sciences - Journal of Innovative Education, Journal of Information Technology Education: Innovations in Practice, Total Quality Management & Business Excellence, eService, International Journal of Web-Based Learning and Teaching Technologies and Journal of EUC and many national and international professional proceedings. He has published four edited books -- web-based education (2), cloud computing (1) and Big Data (1). His current research interests include Web-based education, business ethics, big data, virtual team collaboration and cloud computing.