Open Science Index, Computer and Systems Engineering Vol:10, No:8, 2016 publications.waset.org/10007377.pdf

Road Accidents Bigdata Mining and Visualization Using Support Vector Machines

Usha Lokala, Srinivas Nowduri, Prabhakar K. Sharma

***Abstract*—**Useful information has been extracted from the road accident data in United Kingdom (UK), using data analytics method, for avoiding possible accidents in rural and urban areas. This analysis make use of several methodologies such as data integration, support vector machines (SVM), correlation machines and multinomial goodness. The entire datasets have been imported from the traffic department of UK with due permission. The information extracted from these huge datasets forms a basis for several predictions, which in turn avoid unnecessary memory lapses. Since data is expected to grow continuously over a period of time, this work primarily proposes a new framework model which can be trained and adapt itself to new data and make accurate predictions. This work also throws some light on use of SVM's methodology for text classifiers from the obtained traffic data. Finally, it emphasizes the uniqueness and adaptability of SVMs methodology appropriate for this kind of research work.

***Keywords*—**Road accident, machine learning, support vector machines.

I. Introduction

IN this ever growing population, coupled with technological trends and fast growing vehicles on roads, it is quite common to expose to several road accidents, all over the world. Currently, there are over 1 billion cars on roads and this number is expected to double by 2020 (IBM 2014). Vehicular traffic increased by 236 percentage from 1981 to 2001 (IBM 2014) while the world population grew only by 20 percentage [11]. Increased urbanization has impacted the mobility of people in cities [11]. Zero traffic fatalities and minimizing traffic delays are some of the grand challenges in Cyber-Physical Systems [10]. Fast growing technology has been a helping hand in support to several traffic control and road safety issues. In recent era, this has become a bottleneck issues in most of the developed nations. Driving is the task that ones life can be at danger if one is lacking utmost concentration as well as physical and cognitive co-ordination [7]. It is prevalent that drivers usually involve themselves in tasks like chatting,eating,enjoying music which interrupts them from their most prior task driving and often may lead to unexpected accidents [7]. Past research shows, machine learning (ML) techniques have been popular among the several available technological strategies. These are especially used in addressing several issues and challenges like optimizing the road safety. On the other hand, since several decades, data has been a cornerstone in addressing several road safety and accident issues. This is especially true for analyzing, diagnosing (road crashes) and finally for monitoring road safety. The data-led diagnosis plays a pivotal role in achieving a sustainable road safety improvement and addressing several road safety management problems. Initially we have segregated the road safety problems and its associated risk factors and priorities. Then we formulated a strategy in addressing targets and monitor performance. At the same time, the needed data is very hard to reach, sometimes not available in a friendly format, but is quite rich in contents and poor in intelligence. This situation, forces us to refine data, using several advanced techniques like visualization, clustering analysis, linear regression, and dimension regression. To this effect, there are several algorithms in force to free the noise, and make it more useful while achieving higher road safety. The primary research focused in this work is solely concentrated towards road safety expectation and management as described below.

II. Research Projection and Expectation

This research is mainly investigating the given data from two different schools of thought, viz., (a) what are the road safety expectations vital for today? and (b) keeping these expectations in mind, what are the strategic road safety management issues needs to be addressed? Using the correlation and regression analysis on the given data, we have surfaced certain pertinent results. For example, the first step in traffic control on roads is to look for vehicles and their behavior pattern. Initially this research analysis establishes a correlation between the engine make, vehicle type such as make, model coupled with the gender of the driver involved in each accident incident.

The root cause analysis of several road accidents may also include weather and road conditions. There by the weather and daylight conditions, has been playing a vital role in determining the road accidents. This gives us scope and a pertinent need for establishing a correlation between accidents caused, particularly due to bad weather and daylight conditions. A standard computing hardware and software is used for necessary computational work is

Usha Lokala is with the Department of Computer Sciences, Wright State University, Dayton OH 45435, USA (Corresponding author, e-mail: lokala.2@wright.edu.

Srinivas Nowduri is with the Department of System Security and Law Enforcement Technology, Farmingdale State College, State University of New York, Farmingdale NY, USA (e-mail: nowdurs@farmingdale.edu).

Prabhakar K. Sharma is with the Department of Computer Sciences and Information Technology, North West Missouri State University, Maryville MO, USA (e-mail: S521803@mail.nwmissouri.edu).

involved in this research, detailed in successive sections. In particular to data visualization activity, we have also adopted Tableau, due to its adaptability and capability, for huge data manipulation; coupled with R for the problem solving. Effective strategies to improve traffic operation and safety simultaneously require profound understanding about their features and relationship and in the age of information, these objectives could be efficiently realized through Big Data applications [6]. Thanks to the advantages brought by Big Data, researchers are able to analyze the traffic condition for each crash case and draw general conclusions using individual crash data [6]. Before we go further technically in this direction, let us throw some light on source of data and data collection, as described below.

III. Sources of Data and Dataset Collection

The existence of typical data mining methods are not sufficient to use, due to the process of knowledge building that should store data temporary in the memory. The fact that data is continuously becoming huge through time, so we need to find a way that could automatically adapt to process data and make predictions [5]. The sources of data and dataset collection have been significantly impact several experiments, within the purview of machine learning. The past research, in this direction has clearly demarcated and observed that the nature of data and sources of data, and then collecting dataset play a vital role. A particular emphasis is given to the sources of data and dataset collection in this research due to two primary reasons detailed below:

A. *For Better Quality and Quantification*

In order to visualize the datasets to be huge and accurate, we made a thorough statistical analysis for the data set provided by Department of Transport UK [1][16]. This analysis is made primarily based on several attributes such as circumstances of personal injury road accidents, including the types of vehicles involved and the consequent casualties.

B. *As the Support Vector Machines (SVM)*

which may not perform accurately on imbalanced and highly skewed data sets. Also We choose multi-class SVM to classify more than two labels. This may force the dataset need to be more classified with different class labels like age band of person, vehicle type, type of weather on the day the accident occurred. This data set supports both regressions as well as classification tasks and finally it can help multi class SVM train the models which can handle multiple continuous and categorical variables.

In this research the basic idea behind a data set is primarily based on road accidents, which are particularly reported to the police within the United Kingdom. Fortunately this has been provided by the Department of Transport UK, upon the authors request, which has been used in this research. Additionally, this data has also categorically separated and presented, according to the circumstances of personal injury road accidents. Furthermore these datasets also reflects the type of vehicles involved in an accident, coupled with the consequent casualties. The other sources are primarily related to certain attributes of the incident such as:

1) Road safety and regulations involved.
2) Associated hospital admission formalities
3) The number of death casualties registered
4) Coroners reports,
5) National travel survey,
6) Crime survey from England and Wales
7) Finally the existing statistics on breath tests and motoring offenses from Home Office and Ministry of Justice.

Among the several tables available within the provided datasets, it has been a challenging issue for the researcher to demonstrate the importance and dominance of one attribute over the other. This is primarily needed to establish the inter dependency among the tables. At the same time, these attributes were also helpful in projecting the scope of the research from different dimensions as described below.

IV. The Scope of the Research Project and the Building Models

Initially, this research project analyzes and address the correlation between the engine make, vehicle type make/model and gender of the driver involved in each accident incident. Next, it also highlights the correlation between accidents caused and its root cause analysis such as bad weather, daylight and road conditions. Finally it is going to use data visualization tool called tableau for data visualization [4], along with usage of certain libraries within R environment, in establishing the problem solving strategy. We now throw some light on those attributes of datasets that are accepted and vital for software in projecting better results

A. *Data Collection*

As mentioned earlier, a dataset is solely based on road accidents which are reported to the police in the Department of Transport UK. Further these datasets also includes, the needed statistics regarding the circumstances of personal injury road accidents, type of vehicles (make and model) involved in the accident and consequences of casualties.Department for Transport (DFT) traffic [1] statistics provide estimates of the vehicle miles traveled each year in Great Britain, by:

- Vehicle type
- Road category
- Region

B. *Data Cleaning*

Within the machine learning environment, the primary role of data cleaning is mostly concerned about the consistency and completeness of the data, for its accuracy. It is also realized that the obtained data is composed of several incomplete values, null fields and various repetitive values. This is precisely called the noisy data, which demands some special tools, for cleaning. Sometimes this noisy data also called inconsistencies of data, can be cleaned through intentional and binary values, using Trifacta Data Wrangler; a tool released by Stanford university [2].

C. Data Integration

Within the purview of machine learning strategies, data integration establishes the logical connectivity and extraction of needed information, among the several tables. Therefore, the very idea behind data integration establishes a smooth transition across several attributes of many tables available within the data sets, through an integration process. For example, the dataset provided by the traffic department of UK, belongs to accidents occurred in the year 2014. This is basically including the tables related to road accidents such as road casualties, vehicle involved in accident, vehicle make and model etc. Within the shadow of data integration, it is realized that all these tables have a unique attribute called accident index that has come in handy for data integration. For this, the authors were used a special tool called Rapid Miner [3] which has several built-in functions that help in preparing data for building predictive models.

D. Importing Datasets and Designing Plots in R

The open question that triggered the researcher at this junction is to validate, the data consistency and its completeness with R. This has thrown some more light in following research direction. The integrated dataset is initially imported into R Studio, by installing all the required packages into R. Then the plot between location of accident and accident severity is carefully designed in R, to observe the issue such as, the accident severity is more with respect to location in terms of latitudes and longitudes. The obtained plot is finally captured as shown in the Fig. 1.

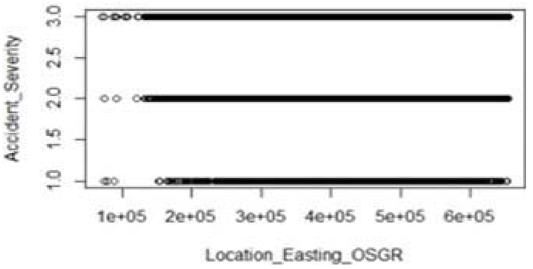


Fig. 1 Design Plot in R

E. Data Redundancy - CHI Square Test

A chi-square test is a technique to compares the observed distribution of data to an expected distribution of data. There are two types of chi-square tests: Chi-square test for goodness-of-fit and Chi-square tests of association and independence [8]. In this research both have been used to test the independence. Test of independence is used to determine whether the observed value (Recorded) of one variable depends on the observed value (Expected) of a different variable [8].

The data redundancy, in this research is primarily to establish the dependency between several factors such as accident severity and weather conditions. Therefore the correlation analysis in this research is done in R using a Chi-Square test. The problem analysis within the dataset, in this research about the accident severity has been divided into three values viz., 1- Fatal, 2-Serious, 3-Slight. At the same time, the weather condition has nine significant values such as high winds, rainy, foggy, snowy, etc. All the given data is basically in numeric form. Therefore it helps in predicting the accident severity against the weather conditions with the table function in R. This resulted in so called contingency table of two variables. We test the hypothesis if the accident severity is independent of weather conditions, at 0.05 significance level. Herewith we attach the data set in R studio, then load MASS library in R and finally find the resultant dataset. Further we have applied Chi-square test to calculate P value to check the above correlation. The formula used in Chi-Square test includes,

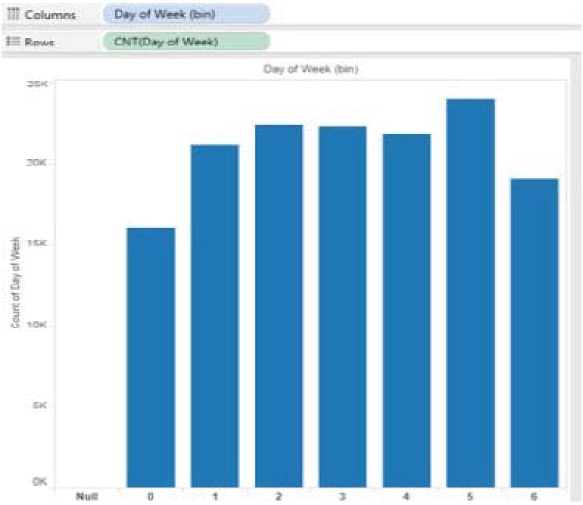

Columns: Day of Week (bin)
Rows: CNT(Day of Week)

Fig. 2 Weekdays to No of accidents in Tableau

$$X^2 = \sum (Observedvalue - Expectedvalue)^2 / Expectedvalue$$

F. Results of CHI Square Test

Once the testing is done, the hypothesis, whether the weather conditions are independent of accident severity, we have arrived at some population p values. In this work it is observed that, the p-value we derived is far less than the .05 significance level. Thereby this firmly rejects the hypothesis: the accident severity is independent of weather conditions. The results of the calculation part are shown in Table I. At this junction it is worth noticing that, the accident severity is highly dependent on type of weather on that particular day.

x-squared = 129.64, df = 16, p-value = 2.2e-16

G. Data Discretization: Multinomial Goodness of Fit

Another significant view of testing of any hypothesis is to establish the discretization, which establishes the collection of frequent item in a data set. In this research, a new data is called multinomial if its data belongs to a collection of discrete non-overlapping classes. Therefore the null hypothesis for goodness of fit test for multinomial distribution is that the observed frequency f is equal to an expected count e in each category. The best bet strategy at

TABLE I
CHI-SQUARE TEST RESULTS AND CONTINGENCY TABLE

	1	2	3	4	5	6	7	8	9
1	1409	154	3	18	25	1	16	17	15
2	17180	2227	28	262	312	8	124	272	263
3	100290	15284	196	1295	1996	69	613	2043	2202

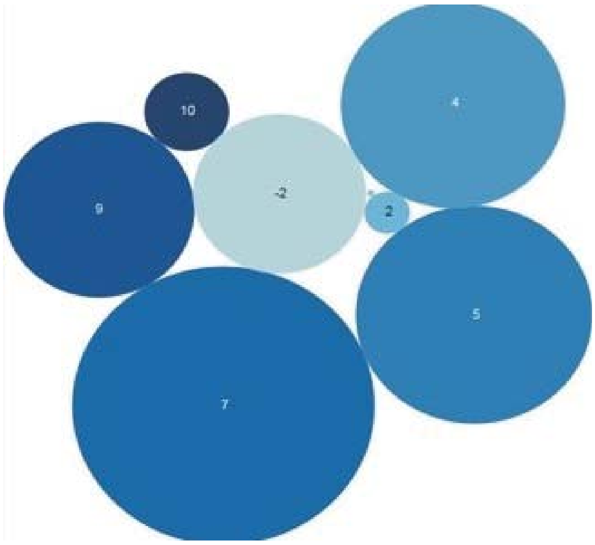


Fig. 3 AgeBand of Drivers visualization in Tableau

TABLE II
FREQUENCIES IN MUTINOMIAL GOODNESS OF FIT

-1	1	2	3	4	5	6
21002	91234	61234	133134	193213	584345	1910062

this junction is: It needs to be rejected if the p-value of the chi-square test is less than a given significance level i.e. 0.05. It establishes a fact that, if a new sample of data follows the similar trend observed in the past. In this connection the dataset age band of causalities is also considered. We can find the frequency distribution of each age band with the table function in R. We checked whether the new data sample also adheres to the usual frequency calculated in the data set. We saved the new data sample frequency in a variable name :Age band of causality prob. The formula used in this connection is

$$P[X1, X2, X3] = ((x1 + x2 + x3)!/x1!x2!x3!) * P1^{x1} P2^{x2} P3^{x3} \quad (1)$$

H. Results of Multinomial Goodness of Fit

We performed the Chi-Square test of age band of casualty and evaluated the probability against the whole data set to label the new dataset. When the test is done on whole dataset, it is labeled as an age band of 6 are doing more accidents. When tested with new dataset, it is assumed that p value is greater than 0.05, signifying that the new sample also follows the same trend. Therefore, it is finally concluded that the new sample strictly adheres to the multinomial goodness of fit. The results are well established and shown in Table II.

TABLE III
SVM RESULTS TABLE

Weather Condition	E.A.S	SVM Label	SVM Accuracy
Rainy High wind Fog Icy	1	1	0.89
Rainy Wet	2	2	0.77
Fine Dry	3	3	0.73
Fog Wet	2	3	0.89
Snow High Winds Wet	2	2	0.83

I. Data Classification Support Vector Machines

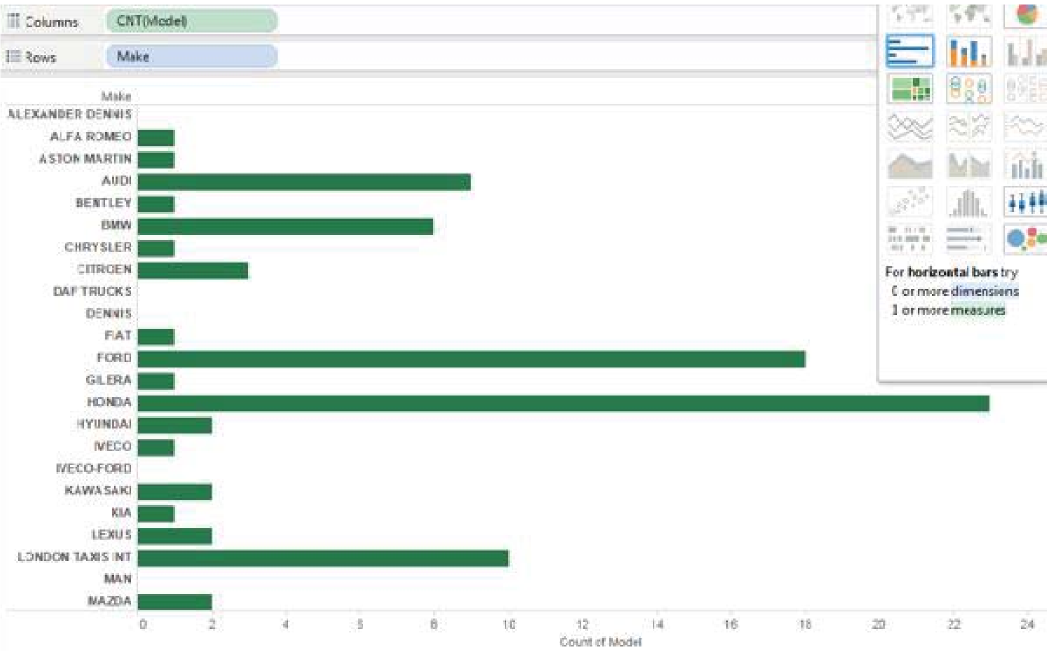
Building models to classify data into different categories can help more accurately to analyze and visualize big data and we can use common algorithms for performing classification which include support vector machine (SVM), boosted and bagged decision trees, k-nearest neighbor, Naive Bayes, discriminant analysis, logistic regression, and neural networks [9]. This research work basically encompasses the multi-class SVM for training the data model and make new predictions. The reason we are using support vector machines [13], [14] is that they are founded on the intuitive geometric concepts of large margin separation and regularized risk minimization, and they are well embedded into statistical learning theory [12]. The following are the main steps used in classifying text in R (with SVM):

1) Install the required dependencies and rattle.
2) Create Document term matrix .
3) Configure the training data into container.
4) Create Prediction term matrix and container.
5) Create and train the SVM model.
6) Predict with new data set.

The R text tools package used in this work provides a powerful way to generate document term matrix with the create matrix function. In order to train a SVM model with R text tools, one needs to put the document term matrix inside a container. In the container's configuration, we clearly indicate that the whole data set will be the training set. Ultimately one will test the model with new data set that was not present in the current training data. In particular, as my model is trained, it can be used to make new predictions.

J. Results of Support Vector Machine

As and when the training model is finally tested with this new dataset (provided by the author) such as: ”rainyhighwinds wet fog icy”, ”rainyhighwinds icy”, ”fine dry, rainy wet, fog wet, snow high winds wet”, we have obtained some significant predictions. These predictions are finally tabulated in the Table III. Furthermore,the test data are categorized as moderate, slight and fatal respectively. Also, their accuracies of happening are represented as probabilities



Columns CNT(Model)

Rows Make

For horizontal bars try
0 or more dimensions
1 or more measures

Fig. 4 Vehicle Make and Model Visualization in Tableau

TABLE IV
Age Band Labels

Label	-2	1	2	3	4	5	6	7
Age band	Missing	16-20	21-25	26-35	36-45	46-55	56-65	co-drivers

(likeness of event happening) of that test data. The following Table III establishes the obtained results.

K. Data Visualization - Using Tableau

As the time progressed, the ability of visualizing data has become more trending along with the importance of the ability to read and write [15].This visualization explains and exemplifies the power of data visualizations not only to help locate us in physical space but also to help us understand the extent and structure of our collective knowledge, to identify bursts of activity, pathways of ideas, and borders that beg to be crossed [15]. As part of the data post processing, several graphs are plotted and results are drawn from our road accidents dataset using Tableau [4]. As a result, we surface the following three seminal facts:

First, to predict on which day of the week, more accidents are likely to happen. This is vital, due to the fact that when plotted in Tableau, it is observed that more accidents are happening on Fridays. The results in this connection are shown in Fig. 2.

Second, to predict what age band of UK Population are more prone to several accidents. Our results establish that the age group of 36 to 45 years are doing more accidents. The results are visualized in Fig. 3 and the related bands are shown in Table IV.

Third, it is observed the correlations between model and make of the vehicle also affect the accidents. For example, it is established that the Honda vehicle are more prone to road accidents in UK. The results are shown in Fig. 4.

V. Conclusions and Future Possibilities

This research primarily projects the goal to optimize road safety and to contribute to achieve accurate prediction results by using several Data clustering, classification algorithms. In this vein, the data is cleaned and integrated in order to prepare it for future applications. Data being so huge and every data point being numeric, it automatically becomes a huge knowledge set to apply any machine learning techniques. As part of the road optimization, data discretization is done through CHI Square test and multinomial goodness of fit to surface new data that suits the already existing pattern within the given data. Also, several patterns in the data are visualized using Tableau Software. In addition to this, this work uses support vector algorithm for classifying textual data. Finally, the new data is more significantly predicted by training and testing the model. It is thus concluded that, the entire dataset is mined using data cleaning, integration, discretization, redundancy, classification and the results derived are found to be highly accurate. This research also establishes a new approach in optimizing the road safety, by successfully adopting several machine learning techniques. As a future prediction, this work can be extended as following: Since the dataset is so huge and consists of several lakhs of records, it might be a better idea to explore generating a frequent item sets in all tables, and then apply algorithms. In that case, it could provide more accurate results than direct mining the huge data set. Finally, the use of Artificial Neural Networks (ANN) might project more accurate results and produce more optimum results to achieve higher road safety.

ACKNOWLEDGEMENTS

We acknowledge partial support from the National Science Foundation (NSF) award: EAR 1520870: Hazards SEES: Social and Physical Sensing Enabled Decision Support for Disaster Management and Response. Any opinions, findings, and conclusions/recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Department for Transport. (n.d.). Retrieved January 28, 2017, from https://www.gov.uk/government/organisations/department-for-transport.

[2] Katie Feline, Ph.D. Proyecto Titi, Inc., Adrian McDermott, SVP Product Development, Zendesk, KlearSky, I. M. (2016, December 27). Trifacta Wrangler — Products. Retrieved January 28, 2017, from https://www.trifacta.com/products/wrangler/.

[3] Data Science Platform — Machine Learning. (2017, January 25). Retrieved January 28, 2017, from https://rapidminer.com/

[4] Tableau Software. (n.d.). Retrieved January 28, 2017, from https://www.tableau.com/.

[5] Wibisono, A., Jatmiko, W., Wisesa, H. A., Hardjono, B., Mursanto, P. (2016). Traffic big data prediction and visualization using Fast Incremental Model Trees-Drift Detection (FIMT-DD). Knowledge-Based Systems, 93, 33-46. doi:10.1016/j.knosys.2015.10.028.

[6] Shi, Q., Abdel-Aty, M. (2015). Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transportation Research Part C: Emerging Technologies, 58, 380-394. doi:10.1016/j.trc.2015.02.022.

[7] Taylor, P., Griffiths, N., Bhalerao, A., Xu, Z., Gelencser, A., Popham, T. (2015). Warwick-JLR driver monitoring dataset (DMD). Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - Automotive '15. doi:10.1145/2799250.2799286.

[8] Shanti Verma. 2016. Deciding Admission Criteria For Master of Computer Applications Program in India using Chi-Square Test. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '16). ACM, New York, NY, USA, , Article 103 , 4 pages.

[9] Solutions. (n.d.). Retrieved January 28, 2017, from http://www.mathworks.com/solutions.

[10] Rajkumar, R. R.; Lee, I.; Sha, L.; and Stankovic, J. 2010. Cyber-physical systems: the next computing revolution. In Proceedings of the 47th Design Automation Conference, 731 736. ACM.

[11] Anantharam, P., Thirunarayan, K., Marupudi, S., Sheth, A. P., Banerjee, T. (2016, February). Understanding City Traffic Dynamics Utilizing Sensor and Textual Observations. In AAAI pp. 3793-3799.

[12] Urun Dogan, Tobias Glasmachers, and Christian Igel. 2016. A unified view on multi-class support vector classification. J. Mach. Learn. Res. 17, 1 (January 2016), 1550-1831.

[13] Vapnik. Statistical Learning Theory. John Wiley and Sons, 1998.

[14] E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT 1992), pages 144152. ACM, 1992.

[15] Katy Borner. Data Visualization Literacy. In Proceedings of the 27th ACM Conference on Hypertext and Social Media (HT 2016), pages 1-1. ACM, 2016.

[16] Road safety dataset. Retrieved January 20, 2017, from https://data.gov.uk/dataset/road-accidents-safety-data