BEHAVIOURAL CONTEXTUAL PREDICTION IN THE WILD USING MACHINE LEARNING

Analyzing the users behavioural and predicting them

Research in Computing

National College Of Ireland

Aswin Sivam Ravikumar
x16134621
Masters in Data Analytics
6th August 2017

Abstract

The objective of this paper is to exemplify the behavioural contextual prediction in the real life. In the present the biggest question raised over in the big data industry is that the data can predict the human behaviours and we want to know is that whether we can use this data to measure what people are doing in the world right now, and possibly even to predict what they are going to do in future because the most the actions and work they done in past they are mostly going to do the same in the present and future. The main aim of this project as it can serve many domains for example think about the medical industry if it is used they can monitor physical activity, generally natural behavioural is complex but once it found out it would be helpful in many aspects.

METHODOLOGY To predict the behavioural prediction here I am going to use both qualitative and quantitative analysis so here I am using the machine learning techniques so it helps to find that where they are, with whom they are and their other behaviours the data was captured in a direct observation for 7 days with the help of 60 persons. An elaborate information about the data is given below.

CONCLUSION Human behaviour is an important one by predicting it can be used in the medical industries, and it can contribute to the health monitoring, aging care and on par as it will be helped in many industries to predict the customers behaviour so it can help the company to run in a profit direction, in bank sectors and to reduce the crime

KEY-WORDS- human behavioural, machine learning, quantitative and qualitative analysis, real life.

Contents

1 Introduction

The modelling and the prediction of the human behaviour it can be described as a well set of dynamic models with the help of the Markova chain they are sequenced and related together.(Pentland and Liu; 1999) Alex Pentland and Andrew liu students of MIT have stated that every human will have the different behaviours but if it can be predicted it can be used to improve the human machine systems on par as it would antedate the human behaviours. The main problem in predicting is the internal states of the human are not directly accessible or observable so for that to find out the patterns we need some models to find out them so they have developed the markov models. This is the first model that researchers were trying to predict using the markov models in the year of 1999.

After some years the big data have started to rule the world there is a strong belief that the big data have more potential to solve anything from the normal social problems to the longstanding business problems. (Li et al.; 2012) A brief study was published on the 2011 by the Jian li whats the behavioural context so they developed a novel framework for automatic discovering the behavioural context specially about the behaviour spatial context, correlation context and temporal context A workshop was conducted in New York during the year 2014 named the social, cultural and each dimensions of the big data so they have stated that the predicting human behaviour can gives a more benefit or it will be more problematic.(Rosenblat et al.; 2014)

Machine learning the which helps the computers to automatically learn and improve without the explicit programming the aim of the ML is focus on the computer programs in order to improve them. (Valletta et al.; 2017)When the world has Moven in to the big data the researchers used the machine learning to predict the animal behaviours so most of them have the questions whats the use of predicting the animal behavioural if some natural disasters is going to be happen the animals would react little bit weird by predicting their behaviours some precaution measures can be taken so that the damage cause will be reduced, they are predicted using the both supervised and the un supervised learning stated by john joseph Valletta, Colin tourney in the applications of machine learning in the animal behaviours.
Behaviour prediction can be predicted using the higher granularity temporal information stated by the students of Vanderbilt university they have collected the data for week of the students with the two different and they are predicting whether they are going to drop out or not by using the data collected from here they have used the machine learning algorithms like the random forest, decision trees, support vector machines and predicted the students behaviour(Oliver et al.; 2000). The machine learning problems will be more challenging when it comes to the human behaviours because every part of users context is not observable, for instance the human audio-visual system will have the limited accuracy and some cognitive elements so at a certain its difficult to quantify and measure but to overcome that a new method of data collection was introduced using the smart watches and smart phones.
(Vaizman et al.; 2016)The Yonatan vaizman and the two other members published a paper that how

well the human context can be recognized using the smartphones and the smartwatches which consists of sensors to find out the complete situations and the positions of the people, for every single minute the multiple sensor measurements will be measured with the relevant to their context labels. (Epstein et al.; 2016)Ziv Epstein and the two others from yale university have analysed the decision making of human behaviours using the good, the bad and the selfish. Many of them have developed different algorithms human behaviour one of the most important one if its predicted it will be useful to police departments to find out the crimes, in education and many more industry sectors but it should be used wisely(Rosenblat et al.; 2014).

The paper has the following sections namely the section 2 has the literature review for the papers related to the behavioural contextual prediction, section 3 provides you the research plan which includes the description of data set and how I am going to implement and the last part is conclusion.

2 Literature Review

According to (Fehlmann et al.; 2017), the use of accelerometers gives out much more advantage in collecting the data with some restricted direct observations and also gives one more choice of habituation of wild primates. Moreover, the Optimal Foraging theory and the social behavior in wild primates can conclude something about the use of accelerometers. This paper basically includes a complete protocol from collar design construction in order to identify the behaviors from the accelerometers.

The journal Animal behavior by (Valletta et al.; 2017), states that when it comes to Machine Learning the techniques says the disadvantages of several classical statistical model in order to describe the data sets that are been captured at unprecedented rate in several fields of animal behavior related topics. Certainly, three cases were given out by the author in order to open up the use of machine learning in developing data analytical workflows to present regarding the biological questions. Translation of complex data sets into scientific knowledge will be very useful in ML.

The paper depicts the behavioural variability (Vaizman et al.; 2016) which is an inadequate representation in controlled studies. Which makes harder to understand the content in the previous cases and scenarios, so the precision has been challenged compared to those reported in the experiments which had some integrity over the behavioural conditions. The paper reflects a persons natural behaviour based on devices we use on a daily basis. There are examples describing the sensor fusion and also fusion multi-model sensors based on the finding from the legacy findings. The combined representation of the behaviour is very flexible and to fully broaden length of the context a user or a researcher can either use methods which are supervised and also turning their attention to newly targeted labels during the course of collecting additional data.

Our case study here only describes a very specific category of office workers excluding the complete scenario. Even though the results are promising,(Kanan et al.; 2014) we still couldnt conclude that all the other statistics are covered in this study. Further to add to this study we can say sensor-based Interruptibility is very much true and speech/voice recognition sensors are the most promising one in this scenario. Also, the other overall sensor models can be used for better result and clarity.

As we know this paper clearly demonstrates how we use ORMB (Ontology-based Restricted Boltzmann Machine)(Phan et al.; 2017) for the purpose of understanding the human behaviour prediction in the health social networks. There are several models/methods introduced in this paper such as self-motivation, get a control over social influence and environmental behavioural modelling in order to deal with the human behaviour prediction in the health social networks. Also this research can be extended in several directions such as obtaining descriptive explanations for predicted behaviour and next one is these research are rooted to RBM and also can be used for different strategies in any further work.

Its has been clearly discussed that we can precisely categorize human driving actions after the start of actions using the behaviour modelling methodology. As the basic nature of the driving task this methodology is considered to generalize to other dynamic human-computer systems. Which helps us to develop a system which depicts peoples indented action and possible to build a control over them using the system.(Rosenblat et al.; 2014)

This study gives us an idea about a how to use relational data to complete end-to-end system for data science. Also, a complete overview about deep feature synthesis which is nothing but a working algorithm which is in an automated form for synthesizing machine learning. (Kanter and Veeramachaneni; 2015; **?**)This paper can be concluded saying that the datasets for human solutions are tested and can say that data science machine has a major role to play in the process.

(Wright and Leyton-Brown; 2010)This paper investigates methods for predicting the human play so the author James R wright has proposed the four models namely the level k, quantal level k, closely related cognitive hierarchy and the quantum response equilibrium so the four are compared and then found out that which model is the best the one for the prediction so they used the different formula sets but it is difficult to understand the state at last the author have stated that the qlk is the best model, as the contents are not made clearly. On the other hand (Kanan et al.; 2014) Christopher Kannan and the four others reanalysed the data by using the some powerful machine learning algorithms named multi fixation pattern analysis it helps to track the eye data this algorithm can be used in the pattern recognition but the paper presented in the way is not an understandable one.

(Oliver et al.; 2000)Here the author uses the real-time computer vision and the machine learning techniques to recognize the human behaviour and also, he combines the artificial intelligence elements to predict the human behaviour data driven statistical approach is used, bottom up and top down approach is the main part tracking is done here, to find out the posterior state markov chain has been used and the sequence probility formula is given below the main disadvantage in this is only the limited no of examples can be trained, and also the Bayesian approach used for the best frame work but he failed in it because the selection of priors is difficult its one of the open issues in the Bayesian.

(Ye et al.; 2015)Predicting the variety of temporal features across the students, predicting the students behaviour and finding out whether they are going to drop out or not ,in a college two different programs were taken based on the video lectures released and the a weekly basic quiz the attributes are formed in the two different types one is the students who have seen the lecture video and not taking the quiz and the students who have watched the video and who takes some quiz based on this the attributes are formed and by using the random forest, linear regression support vector machine they are tested and some preliminary tests are made and the results are explained clearly. With the help of these paper I can use the algorithms for my research purposes.

In a similar study carried out by (**?**), the factors were collated and analyzed using the decision tree algorithm. In this study, a combination of statistical methods and machine learning techniques was implemented. This led to an efficient collation of data from different sources. The data was split into testing and training models and the efficiency across each sphere was calculated. This can lead to a highly normalized view of each parameter and the variation across a particular value. This study seeks to adopt a similar approach but with the necessary testing and training data in place. This would ensure the accuracy of the results and the outcome that can be obtained.

(Li et al.; 2012)Jian li and the two others proposed a framework for the automatic discovering the behavioural context, they have specially mentioned the three types of behavioural context they are Behavioural spatial context Behavioural Correlation context Behavioural temporal context they have used the visual context and the existing works were well used the regional and the multiscale context learning is used and they have also succeeded and the accuracy is about 87.4
Here the author Jonathan gratch(Kanan et al.; 2014) have a different method for predicting the human behaviour using the negotiation outcomes, here by using the nonverbal factor and depending upon the proposer and the responder behaviours the data set is collected for the different males and females aging from 19 to 25 and they are finding out the mean accuracies of the negation outcome are determined but the paper presented here is not an easy one to understand and its a complex one. (Kosinski et al.;

2013)The attributes and the traits are one of the most important one to predicting the human behaviour without that we cannot get the accurate prediction Yonatan vaizwman and the other PhD students were collected the data using the sensors in the watches and by using the app in personal mobiles called extrasensory. The way the data collected and they integrated using the python on par as they have used the linear regression in predicting the human behaviour the paper was well designed and easily understandable this paper address most of the questions raised in the bigdata industry about the behavioural contextual prediction in the real life(Vaizman et al.; 2016).

2.1 Summary of literature review

By reviewing so much papers to predict the behavioural contextual prediction machine learning would be better part and what algorithm we gone use so by doing this literature review i also have an brief idea on the ML and to use which type of algorithm can be used predict better.the following section will consists of my research question and how its going to be implemented.

3 Question

Whether the output obtained from the machine learning with the help of a large number of attributes helps to predict the users personal contextual recognition in the real life and how well the behavioural contextual prediction can be used in the real world?

4 Project plan

4.1 1.Data extraction and pre-processing 2.Desired software installation 3.Machine learning algorithms implementation 4.Comparing the algorithms 5.Final Result and Documentation

4.2 Data extraction and pre-processing

Data extraction is process of analysing the data to obtain the relevant information from the data sources like data sets and databases in a specific format, after data extraction data pre-processing is a data mining technique it used to convert the raw data into understandable one its nothing but changing the formats most of the datas that present in the real word are incomplete or inconsistent so to overcome we are performing the data pre-processing. The basics steps in the data extraction and data pre-processing are Data capturing,Data cleaning

4.3 Data capturing

In the field of human behavioural, the datas present in the open sources are relatively small with having the small percentage of information its difficult perform the analysis. Then open sources data will have only the small amount of information mostly its a sample data so it becomes difficult to predict the human behaviour with the small quantity of information provided. As I strongly believe that the to predict the human behaviour the information can only be collected using the direct observations of samples. Its a mandatory process. More open data repositories are available like Kaggle, UCI data repository, GitHub we can use these open sources to gain the information with relating to its respected journals, documents, articles and related conference papers.

4.4 Data cleaning

Data cleaning is nothing but the process of removing duplicate, null and inconsistent data from the data source or the data-set to gain the more accurate prediction and the outcome the main aim of this process is altering the data in the resource storage to make it correct as well as accurate various softwares can be used to clean the first one Microsoft excel and the second one Google refine, its one of the most powerful tool used in the big data industry. The data cleaning can also be mentioned as the data scrubbing. The open source has more cleaning tools as it can be used.

4.5 1.Desired software installation
2.Machine learning tool
3.Statistical tool
4.Visualization tool
5.Machine learning tool

To perform the prediction on behavioural context I am going to use the scikit learn in python and the weka.

4.6 SCIKIT LEARN

Its a python module which helps to integrate the wide range of machine learning algorithms for both the supervised and the UN supervised learning. The main of this package is bringing the beginners to use the high-level language python is one of the most powerful languages in the scientific computing. When comparing the scikit-learn, mlpy, pybrain, pymvpa, mdp and shogun the scikit is the best, it selects the parameters using the cross validation. It provides the easy comparisons of methods to the given problems.(Pedregosa et al.; 2011)

4.7 WEKA

Its an open source software its a collection of machine learning algorithms these algorithms can be applied directly into the dataset or by using the java code. This software was founded and developed in the university of Waikato by the machine learning department.

4.8 STATISTICAL TOOL

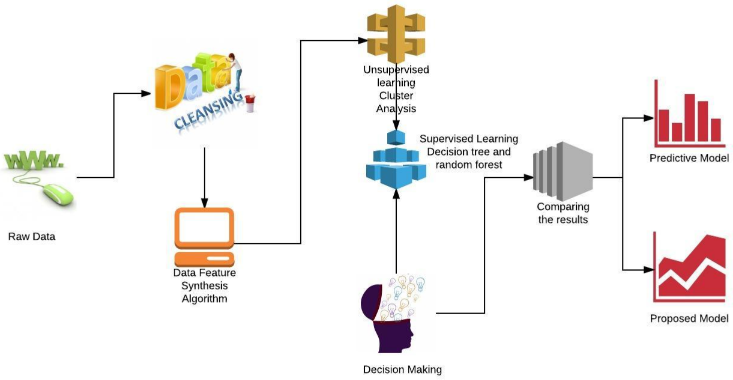
Here I am going to use the IBM SPSS statistical tool for my analysis to find out the correlation between the human behaviours that who have the strong and weak links, finding correlation between the persons is also helps to avoid the crime percentage by predicting them before.

4.9 VISUALIZATION TOOL

In good information visualization, there are no rules, no guidelines, no templates, no standard technologies, no style books. You must simply do whatever it takes its a famous quote about the data visualization by Edward tufte, visualization is one best part of understanding the information. Visualization usually consists of many charts such as bar, histogram, box plot etc so we can how they differ in one week its one of the uncomplicated way to understand the results. Here I am using the python using the package mat plot, and qlik sense its one of the visualization tool which uses the in-memory concept the data load can be done quickly using the. QVD format.

4.10 IMPLEMENTATION OF MACHINE LEARNING

For every project implementation is the heart where the process of planning, decision and the execution this process includes the following steps. 1.Importing data sets 2.Deep feature synthesis 3.Unsupervised learning- dimensional reduction, feature extraction, clustering, K means 4.Supervised learning- Decision tree, Random forest 5.Comparison of different algorithms 6.Key findings

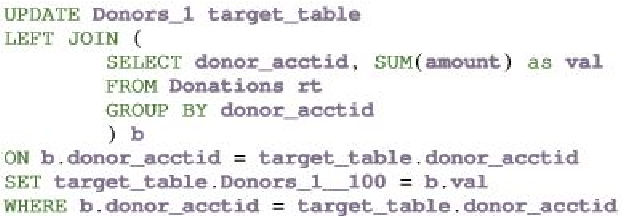

Raw Data
Data CLEANSING
Data Feature Synthesis Algorithm
Unsupervised learning Cluster Analysis
Supervised Learning Decision tree and random forest
Decision Making
Comparing the results
Predictive Model
Proposed Model

4.11 Importing Datasets

The first process is to load the data before starting our machine learning project, one of the most common format is the CSV file to load in the python or weka the dataset, the datasets can be obtained from the different sources first we need to integrate it together, so it should have a unique key so our first impression will be how will be the relationship and communication between them.

4.12 Deep feature synthesis

Deep feature synthesis is the algorithm developed in the MIT university by James max kantar and Kalyan veeramachaneni this algorithm automatically generates the features for the relational data sets, it follows a relationship between the data in a base field and then its starts applying the mathematical functions to create the final one. The main function is even it is an automatic algorithm it also needs some human intuition. This algorithm is built on top of the MySQL database by using the InnoDB. So how it works it automatically converts the raw data into an MySQL schema.



```
UPDATE Donors_1 target_table
LEFT JOIN (
        SELECT donor_acctid, SUM(amount) as val
        FROM Donations rt
        GROUP BY donor_acctid
        ) b
ON b.donor_acctid = target_table.donor_acctid
SET target_table.Donors_1__100 = b.val
WHERE b.donor_acctid = target_table.donor_acctid
```

The above example query is an auto generated MySQL query by the deep synthesis algorithm so there is also a generalized machine learning pathway to use the deep synthesis algorithm. The first step is what we are going to predict once we know that it will be called as the predictors after that data pre-processing and data modelling is done.

4.13 UNSUPERVISED LEARNING

(Gopalakrishna et al.; 2017)Depending on the goals of the application there are many ways and techniques for training the data the goal of the unsupervised learning is to find the patterns in the data and to build new and useful representations of it unsupervised learning is a powerful technique for the tasks like pattern recognition, data clustering, object recognition, feature extraction and data reduction so here I am going to use the K Means.

4.13.1 K- means

K means clustering its an algorithm used in the unsupervised learning when our dataset has the unlabelled data, or the undefined groups the K-means algorithm helps to find it, it assigns the data points to the k groups based on the dataset provided.so now the centroids will be formed centroid is nothing but the collection of the values. For this I am going to use the scikit in python.

4.14 SUPERVISED LEARNING

Supervised learning is nothing but teaching the model with an knowledge to predict the future instances so the model is trained on a labelled data set so I am going to use the decision tree and the random forest to predict my method.

4.14.1 Decision tree

Yes or no questions to find out the humans where they are at home, school, office etc by certain rules and conditions. Its easy to carry both the categorical and the continuous variables and they are robust The decision trees help to classify the main work of the decision trees are by splitting the train set into distinct nodes, where the one node contains most of data, one category of the data the categories can be called as the subsets with the data set divided, inputted out of sample data will be classified more easily when it falls into a node that is strictly one subset of the data if so there is a higher probability that the data point is the same classification as the node it fell under(Valletta et al.; 2017).

4.14.2 Random forest

When we use the different learning models we can increase the accuracy of the classification is the main idea of the technique called bagging the random forests works on the large correlated decision trees it creates a large of decision trees and then it makes a classification it can perform both the regression and the classification there will more robust to predict with the high accuracy each tree will give us a classification.

The main aim of using this both the classification and the regression task can be done, it can handle the missing values and it wont over fit the model and it can handle large data set.

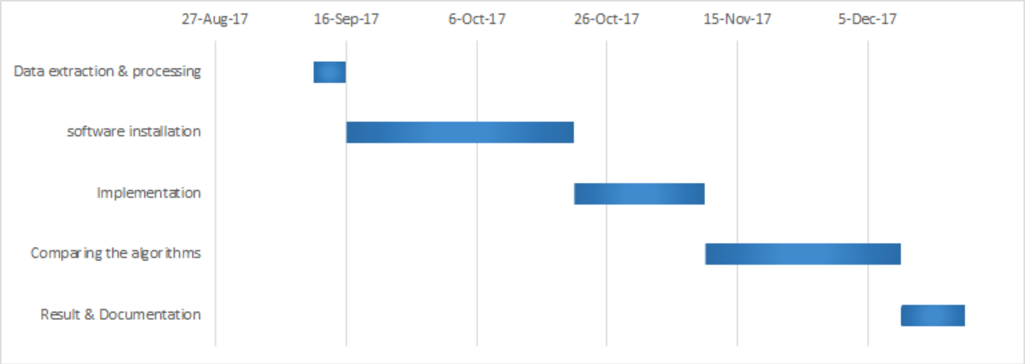
5 COMPARING THE ALGORITHMS

The previous work was done by the linear regression and the binary classifier no I am going to predict it using the deep feature synthesis, supervised and the unsupervised learning methods and going to compare the algorithms used here and trying to find out that which methods have more accuracy in the behavioural prediction.

6 KEY FINDINGS

Some of the key findings found in the model are given below Behavioural classification Human activities Activity tracking

7 Gantt Chart



8 Conclusion

This paper addresses the human behavioural in real life and trying to predict WHERE THEY ARE, WITH WHOM THEY ARE, WHAT ARE THEY DOING, BODY POSTURE STATE The proposed model will attempt to predict the humans behavioural prediction in the real life by the future works this model can be extended to analyze the human behaviours more accurately, this makes the contextual recognition is not an easy challenge when compared to the previous researchers. how the human behavioural will helps and whats the use of it, useful to the bank sectors with whom they speak and the tone of their voice, it can be used to reduce the crime to be in general it can be used in the human resources, educational and financial aid.

The nal outcome of this product also be used in the medical industry to know about their patient state In the day to day life it would help the parents to know where their children are they safe and its help improves the corporate relationships on their behavioural performance also its helps in the self-improvement.

The future works lead to the if some modern technologies are available for collecting the data it may help to reduce the annotations as well as it can be used in time series modelling that which will improve the context recognition.

References

Epstein, Z., Peysakhovich, A. and Rand, D. G. (2016). The good, the bad, and the unflinchingly selfish: Cooperative decision-making can be predicted with high accuracy when using only three behavioral types, *Proceedings of the 2016 ACM Conference on Economics and Computation*, ACM, pp. 547–559.

Fehlmann, G., ORiain, M. J., Hopkins, P. W., OSullivan, J., Holton, M. D., Shepard, E. L. and King, A. J. (2017). Identification of behaviours from accelerometer data in a wild social primate, *Animal Biotelemetry* **5**(1): 6.

Gopalakrishna, A. K., Ozcelebi, T., Lukkien, J. J. and Liotta, A. (2017). Relevance in cyber-physical systems with humans in the loop, *Concurrency and Computation: Practice and Experience* **29**(3).

Kanan, C., Ray, N. A., Bseiso, D. N., Hsiao, J. H. and Cottrell, G. W. (2014). Predicting an observer's task using multi-fixation pattern analysis, *Proceedings of the symposium on eye tracking research and applications*, ACM, pp. 287–290.

Kanter, J. M. and Veeramachaneni, K. (2015). Deep feature synthesis: Towards automating data science endeavors, *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, IEEE, pp. 1–10.

Kosinski, M., Stillwell, D. and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the National Academy of Sciences* **110**(15): 5802–5805.

Li, J., Gong, S. and Xiang, T. (2012). Learning behavioural context, *International journal of computer vision* **97**(3): 276–304.

Oliver, N. M., Rosario, B. and Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions, *IEEE transactions on pattern analysis and machine intelligence* **22**(8): 831–843.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python, *Journal of Machine Learning Research* **12**(Oct): 2825–2830.

Pentland, A. and Liu, A. (1999). Modeling and prediction of human behavior, *Neural computation* **11**(1): 229–242.

Phan, N., Dou, D., Wang, H., Kil, D. and Piniewski, B. (2017). Ontology-based deep learning for human behavior prediction with explanations in health social networks, *Information Sciences* **384**: 298–313.

Rosenblat, A., Kneese, T. et al. (2014). Predicting human behavior.

Vaizman, Y., Ellie, K. and Lanckriet, G. (2016). Recognizing detailed human context in-the-wild from smartphones and smartwatches, *arXiv preprint arXiv:1609.06354* .

Valletta, J. J., Torney, C., Kings, M., Thornton, A. and Madden, J. (2017). Applications of machine learning in animal behaviour studies, *Animal Behaviour* **124**: 203–220.

Wright, J. R. and Leyton-Brown, K. (2010). Beyond equilibrium: Predicting human behavior in normal-form games., *AAAI*.

Ye, C., Kinnebrew, J. S., Biswas, G., Evans, B. J., Fisher, D. H., Narasimham, G. and Brady, K. A. (2015). Behavior prediction in moocs using higher granularity temporal information, *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, ACM, pp. 335–338.