

Prospectus

Wrangling Big Data with Machine Learning: A Survey
(Technical Topic)

The Role of Big Data Analytics in the Urban Policing Network
(STS Topic)

By

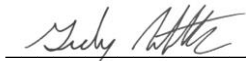
Grady Roberts

Fall 2020

STS 4500-004

Technical Project Team Members:
Daniel Collins

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed:  Date 11/02/2020
Grady Roberts

Approved: _____ Date _____
Rosalyn W. Berne, Associate Professor, Department of Engineering and Society

Approved: _____ Date _____
Aaron Bloomfield, Professor, Department of Computer Science

I. Introduction

Fifth-generation (5G) networking technology standards aim to improve every facet of networking. One aspect of 5G, Massive Machine Type Communication (mMTC), will expand the ability to network small sensing and monitoring devices, increasing the size of the Internet of Things (IoT). Chettri and Bera (2020) expect that over 80 billion devices will be connected by 2030 when 5G is more available. This will enable “smart” homes, cities, and industries powered by massive data collection and analysis. The total volume of enterprise data generated each year is expected to reach 175.8 zettabytes (ZB, equivalent to one trillion gigabytes) by 2025, up from 18.2 ZB in 2015 (Seagate, 2020). Seagate’s survey also found that only 32% of this data is available in a state ready to be analyzed. With data volume seemingly growing too fast for companies to keep up with, it is important to find more efficient ways to store, manage, and analyze data. One way to accomplish this is the application of machine learning (ML) in aspects of database design and query optimization. In the technical project, a literature review of the use of ML to address these problems of managing large-scale data will be completed.

The expansion of IoT, in addition to its many consumer benefits, also has the possibility to drastically increase the scope of police surveillance. Brayne (2017) explains the implications of big data policing, saying:

On the one hand, big data analytics may be a rationalizing force, with potential to reduce bias, increase efficiency, and improve prediction accuracy. On the other hand, use of predictive analytics has the potential to technologically reify bias and deepen existing patterns of inequality. (p. 978)

Brayne also describes a shift in policing practice from reactive to predictive—from responding to crimes to using big data analytics to predict the most likely place for crimes to occur and even who is most likely to be an offender. The STS research paper will describe the network of urban policing and analyze the changing role of ML and big data within it.

II. Technical Topic

The technical project will take focus from two upper-level CS courses: CS 4750 Database Systems and CS 4774 Machine Learning. The project will take the form of a literature review that summarizes and synthesizes a significant amount of current research at the interface between these two courses. The goal of the project is to understand how ML can inform good database design strategies, as well as optimize database operations. Understanding the current state of the art in these areas will provide insight into the best ways for companies to manage and extract useful information from the data they collect.

To perform queries efficiently, database designers create a data structure called an index to order data so that it can be quickly accessed. The type of index used depends both on the structure of the data and the types of queries that will be performed most often. The most common types of indices used are the BTree-Index for searching ranges, Hash-Index for single lookup, and BitMap-index to check key existence. All these types of indices are built using fixed data structures that do not change in response to changing data. Beutel et al. (2017) argued that these indices fail to exploit patterns in the data to further optimize database operation. Using indices created using neural networks (a common ML model), Beutel et al. demonstrated a 44% improvement in the speed of queries and a 75% reduction in the memory required to store the index compared to a BTree-Index on a read-only database. The choice of ML-based indices also

has drawbacks, such as planning to handle false negatives (i.e. the ML index reports that the key does not exist when it does) which do not occur with traditional indices.

One of the most expensive database operations is that of joining multiple tables together to construct the result of a query. Join operations are expensive partly because the database must determine the most efficient ordering of the tables in the join. This decision is typically done using heuristics to reduce the search space. Krishnan et al. (2019) showed that heuristics tend to fail to find the most efficient ordering when there is non-linearity in the database cost model. To correct this, Krishnan et al. created a reinforcement learning ML model capable of outperforming existing heuristics. The reason for the improved performance is that heuristics tend to be greedy, meaning they will take short-term benefit potentially at the cost of greater long-term benefit. In contrast, Krishnan et al.'s ML model is capable of learning to optimize for the long-term.

Another interesting area of research that can be explored in the literature review is the creation of new programming languages that create in-database ML models. These models can exploit the locality of the data and avoid lengthy pipelines to move data into another repository to be analyzed. One such language is MLog, developed by Li et al. (2017), a declarative programming language that interfaces directly with the database. MLog allows users to create complex deep learning models within the database in the same manner that users can open a SQL session and run simple queries against the data in-place. The software operates by viewing the data as tensors, and includes functionality to optimize the model-building process.

III. STS Topic

The STS research paper will analyze the role of big data and ML in urban policing. The goal of the paper will be to describe how the role of data has changed in this network over time,

as data has become more readily available. Focus is placed on urban policing in America because it is widely discussed in the literature and cities are more likely to be outfitted with extensive IoT networks than rural areas. The analysis will be guided by Actor-Network Theory (ANT). The primary stakeholders in this research are the urban residents, police officers and departments, judges, policymakers, and ML researchers. These stakeholders can be expanded to include the nontangible ANT actants of the ML algorithms themselves and the data they process. ANT is a good choice for this research because the goal is to describe the relationships between the actants in the network and, crucially, how they have changed in response to the growth of big data.

It is vital to understand the role of data and ML in this network because errors in policing and sentencing can have drastic and long-lasting impacts on people's lives. Digard and Swavola (2019) describe the effects of pretrial detention, the practice of holding defendants who have not been convicted of any crime in jail awaiting trial. They found that being exposed to pretrial detention made it more likely for defendants to be convicted and to receive harsher sentences. Even those that are not convicted are more likely to be involved in the criminal justice system again, and may lose their jobs while detained. These negative impacts fall mainly on low income defendants who are unable to post bail.

From the perspective of police departments and officers, ML is an opportunity to more efficiently allocate policing resources. If the police can accurately predict where crimes will occur, they can allocate more officers to this area to prevent the crime or more quickly respond. The idea of assigning criminality to physical places has existed for many years (Brantingham & Brantingham, 1995), but is now expanded with the use of ML. A modern example is Wheeler and Steenbeek (2020), who developed a random forest ML model to predict robberies in Dallas,

Texas in 200 foot by 200 foot “micro-places”. Ensign et al. (2018), as ML researchers, are skeptical of this practice. They describe the issue of feedback: the ML algorithm’s decisions influence the data it is trained on. This means the algorithm only “sees” crime in places it has previously decided to send officers, potentially leading to runaway positive feedback loops and over-policing of certain areas.

Police officials also believe that ML can reduce human bias. It is common practice for judges to use actuarial methods to predict recidivism risk when considering sentencing and parole. Berk and Hyatt (2015) and Rigano (2019) argue that this process is prone to the personal biases of judges and that using ML algorithms to assign quantitative risk scores to defendants for judges to consider in their decision will reduce bias. Specifically, it is believed that ML algorithms trained on datasets that strip attributes such as race will produce unbiased models. Bostrom and Yudkowsky (2011) argue that this is not sufficient, that ML models can still exhibit bias such as by using the defendant’s address history to predict race. Additionally, the use of biased black box ML models may provide an inscrutable scapegoat that is more difficult to correct than a biased human judge.

Vaccaro and Waldo (2019) performed experiments that combined human judgment with the output of a ML recidivism risk predictor, finding that the addition of ML-predicted risk scores did not improve human prediction performance of actual recidivism. Vaccaro and Waldo’s experiments also showed that human participants were inclined to agree with the algorithm’s prediction even when it was intentionally faked. This disconnect between the belief in the correctness and unbiased nature of algorithms despite a lack of evidence will be further explored by examining the relationships between the ANT actants.

ML algorithms are not infallible, and can be prone to malicious attack. Sharif et al. (2016) demonstrated an attack on state-of-the-art facial recognition systems that use ML. Their attack worked by wearing specially crafted sunglasses that could obscure facial detection. This realistic attack scenario shows that this and other ML systems may not be ready for widespread adoption, even though it is being pushed by many. This may indicate, as suggested by Brayne, that the rush to adopt ML-based predictive policing may be because they are convenient and save money, rather than because they are more effective. Another reason for the push to use predictive policing may come from a secondary stakeholder, corporations that sell ML algorithms to police departments. Benbouzid (2019) describes one such corporation, PredPol, that may have exaggerated the abilities of its model to sell more software to police. These and other forces influencing the role of ML and data in the network will be further investigated in the research paper.

IV. Timeline and Expected Outcomes

The technical project aims to provide an overview of the current state of the art in the use of ML for database design, implementation, and management in the form of a literature review. Following the BSCS capstone procedure for the 2020-2021 academic year, work on the project will begin in the Spring 2021 semester. An abstract will be submitted by the 5th week of the semester and the final deliverable will be submitted by the 12th week. The topic of the technical project is also subject to change pending faculty reviewer feedback.

The STS research paper intends to describe the network of urban policing and contribute an explanation of the impact of ML and big data analytics on this network. The STS research will take place in the Spring 2021 semester and culminate with the submission of a thesis paper. The

combined results of the STS and technical projects will provide key insights on how to manage ever-increasing volumes of data as well as how this data may affect our relationship with the police.

References

- Benbouzid, B. (2019). Values and consequences in predictive machine evaluation. A sociology of predictive policing. *Science & Technology Studies*, 32(4), 119–136.
- Beutel, A., Kraska, T., & Polyzotis, N. (2017). *A machine learning approach to databases indexes*. Neural Information Processing Systems 2017.
- Berk, R., & Hyatt, J. (2014). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, 27(4), 222–228.
- Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge: Cambridge University Press.
- Brantingham, P. L., & Brantingham, P. J. (1995). Criminality of place: Crime generators and crime attractors. *European Journal on Criminal Policy and Research*, 13, 5–26.
- Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review*, 82(5), 977–1008.
- Chettri, L., & Bera, R. (2020). A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems. *IEEE Internet of Things Journal*, 7(1), 16–32.
- Digard, L., & Swavola, E. (2019). Justice denied: The harmful and lasting effects of pretrial detention. *Vera Evidence Brief*, 17.
- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. *Conference on Fairness, Accountability and Transparency*, 160–171.
- Krishnan, S., Yang, Z., Goldberg, K., Hellerstein, J., & Stoica, I. (2019). Learning to optimize join queries with deep reinforcement learning. *ArXiv:1808.03196*.
- Li, X., Cui, B., Chen, Y., Wu, W., & Zhang, C. (2017). MLog: Towards declarative in-database machine learning. *Proceedings of the VLDB Endowment*, 10(12), 1933–1936.
- Rigano, C. (2019). Using artificial intelligence to address criminal justice needs. *NIJ Journal*, 280, Article 280.
- Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1528–1540.

- Seagate. 2020. *Rethink data: Put more of your business data to work—From edge to cloud.*
- Vaccaro, M., & Waldo, J. (2019). The effects of mixing machine learning and human judgment. *Communications of the ACM*, 62(11), 104–110.
- Wheeler, A. P., & Steenbeek, W. (2020). Mapping the Risk Terrain for Crime Using Machine Learning. *Journal of Quantitative Criminology*.