M.S. in Applied Data Science  
**Program Reflection**

Lauren Lawless | Syracuse University | September 2020

Table of Contents

[Achievement of Learning Outcomes 3](#_Toc50043409)

[Describe a broad overview of the major practice areas in data science 3](#_Toc50043410)

[Collect and organize data 4](#_Toc50043411)

[Identify patterns in data via visualization, statistical analysis, and data mining 4](#_Toc50043412)

[Develop alternative strategies based on the data 4](#_Toc50043413)

[Develop a plan of action to implement the business decisions derived from the analyses 5](#_Toc50043414)

[Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization 5](#_Toc50043415)

[Synthesize the ethical dimensions of data science practice (e.g., privacy) 6](#_Toc50043416)

[Key Conceptual Works 6](#_Toc50043417)

[The Obama Campaign 6](#_Toc50043418)

[Social Desirability Bias 7](#_Toc50043419)

[Application of Learning Outcomes 7](#_Toc50043420)

[Preparation for Work in Chosen Field 7](#_Toc50043421)

[Areas of Strength and Challenge 8](#_Toc50043422)

[Life-long Learning 8](#_Toc50043423)

[Appendix A: GitHub Repository 9](#_Toc50043424)

[Appendix B: Course Key 10](#_Toc50043425)

[Works Cited 11](#_Toc50043426)

# Achievement of Learning Outcomes

## Describe a broad overview of the major practice areas in data science

Over the course of the Master of Applied Data Science program, I have encountered and explored multiple facets of the data science field: from dabbling in the foundations of database management in IST 659 to comparing the costs and benefits of different machine learning algorithms in IST 707 and IST 718. I have studied the particulars of text mining in IST 736 and natural language processing in IST 664 and the additional considerations required for applying these same algorithms to text data, and I have applied the product of my analyses to informed decision-making and campaign strategizing in MAR 653.

While I have centered most of my coursework around political ideology and voter behavior, I have also built a database for tracking breeding pairs of birds nesting at the Fisherville Brook Wildlife Refuge, examined access to protected land areas in California through the lens of environmental justice, provided policy alternatives for addressing inequity in access to gifted education programming for minority student populations, and predicted who would run in the presidential primaries based on tweets.

The projects described in the following sections are the collective artifacts of my work in the data science program. Below, I describe the goals of each project briefly in order to provide some context to the discussion that follows:

* In my first quarter, in IST 659: Data Administration Concepts and Database Management, I built a database to store the data I collected on nesting pairs of birds at the Fisherville Brook Wildlife Refuge in Exeter, Rhode Island in the summer of 2018.
* In the second quarter, in IST 687: Introduction to Data Science, I evaluated access to protected land areas in California through the lens of environmental justice, using census tract proximity to determine whether there was a pattern of inaccessibility to protected lands by historically disadvantaged population groups.
* In the third quarter, in IST 707: Data Analysis, I used several classification algorithms to estimate to which party a voter’s candidate choice was most likely to belong in a generic ballot election, based on 2018 midterm election U.S. House candidate choices.
* At the same time, I dove into PAI 897: Fundamentals of Policy Analysis, a course in the Executive Master of Public Administration curriculum, to conduct a policy analysis of equity in access to gifted and talented education resources by minority population groups in Virginia’s public schools.
* In the fourth quarter, in MAR 653: Marketing Analytics, I analyzed survey responses of Georgia voters to develop a campaign strategy for a fictitious run by Stacy Abrams for U.S. Senate.
* Also during the fourth quarter, in IST 718: Big Data Analytics, I analyzed a different segment of the same dataset to estimate the likelihood that an individual respondent was being dishonest in response to the question of candidate choice in the 2016 presidential election.
* In the fifth quarter, in IST 736: Text Mining, I analyzed tweets by U.S. politicians to predict who would run in the 2020 presidential primaries.
* In IST 652: Scripting for Data Analysis, I used a different voter dataset than previously to predict for which candidate a respondent would choose to vote in the 2020 presidential election.
* Finally, in my last quarter, in IST 664: Natural Language Processing, I am building a classification model to detect hate speech in tweets; and in IST 719: Information Visualization, I am telling the story of the 2016 Trump electoral victory and how shifting political ideologies might play out on the stage of the 2020 Electoral College.

## Collect and organize data

The AP VoteCast 2018 survey and the Democracy Fund VOTER (Views of the Electorate Research) Survey datasets have both been invaluable in my effort to focus my work in the data science program on the field of voter behavior. (Associated Press, 2019; Democracy Fund Voter Study Group, 2019) Both surveys are extensive in the number and breadth of questions asked to their respective groups of voters, but there are key differences:

* The AP VoteCast has many more observations of survey respondents, but raw data are only partially encoded, i.e., responses follow the format (*response number) response text*, so much effort was devoted to preparing the data before even exploratory analysis was possible.
* The VOTER Survey polled fewer respondents but is already in a useable format and is longitudinal, with responses from the same set of voters in December 2016, July 2017, May 2018, and January 2019. These respondents came from a set that were surveyed in December 2011 and November 2012 as part of the 2012 Cooperative Campaign Analysis Project (CCAP).

As a result of the nature of the VoteCast survey, I have refined my skills in data cleansing by writing robust functions to automatically transform unencoded data into a usable format and by implementing iterative feature and observation selection to remove missing values while maintaining a sufficiently large and interesting dataset. Additionally, having used both datasets to answer multiple research questions, I have extensive practice in data segmentation and isolation of key variables and observations in an effort to narrow the focus of analysis for each project.

## Identify patterns in data via visualization, statistical analysis, and data mining

Practically every course has required some component of data visualization, whether for exploratory analysis and gathering a preliminary understanding of the data or for final presentation of analytic results to a non-technical audience. As I progressed through the program, the visualizations that I created became more engaging and concise in the messages that were being conveyed. In my final quarter, I am taking IST 719, which focuses entirely on the technical aspects of visualization that clearly define what makes for effective and interesting representations of data.

While good visualization tools provide intuitive evidence of patterns in data, most people outside of the data science space are primarily concerned with the bottom line, often in numeric terms. Thus, a complementary tool in identifying patterns in data is any machine learning algorithm for which the mechanisms affecting the prediction are viewable to the user, such as linear or logistic regression, where the coefficients for each variable are readily available, or decision trees, where the path from input to output is traceable. Both of these techniques were explored in my coursework: logistic regression in IST 718 and decision trees in IST 707, both applied to the VoteCast dataset to answer separate research questions.

## Develop alternative strategies based on the data

Although offered outside of the applied data science curriculum, the Fundamentals of Policy Analysis course (PAI 897) provided valuable instruction in creative problem solving under various constraints. The key deliverable of the course was a thorough analysis of an existing problem and the provisioning of three carefully considered and mutually exclusive solutions to that problem. While the course was taught in the context of economics and the social sciences, and the solutions were based on more humanities-style analysis than is typical in the iSchool, the investigative and interpretive skills I learned are not only translatable to data science, but are also more nuanced because of the diversity of approach to which I have now been exposed.

Applied in a more technical and data-centered context, this skill set has provided flexibility and creativity when working with group members on different course projects to develop diverse strategies of exploration and analysis given each data and project context, such as in developing several solutions to the text mining problem of using tweet content to predict who was planning to launch a presidential campaign, where ultimately, hierarchical clustering, support vector machines, and Multinomial Naïve Bayes were all used to explore and apply predictions to the data.

## Develop a plan of action to implement the business decisions derived from the analyses

One of the greatest values of data science is the direct connection between statistical evidence and actionable recommendations. While some business leaders and executives get by on experience and “gut feelings”, much guesswork can be reduced in business decision-making by providing demonstrable validation of solutions. One of the most robust examples of this aspect of data science is exhibited by the course deliverable from Marketing Analytics (MAR 653).

The Georgian voters from the VoteCast dataset were isolated and analyzed to find each individual’s propensity to vote for Stacey Abrams in the fictitious scenario where, following her near victory in the 2018 Georgia gubernatorial race, she decides to run for U.S. Senate in 2020. The product of the analysis was a targeted marketing strategy focused on three distinct demographic voting groups. Based on evidence that the greatest chance of victory would be the result of convincing supporters who typically stay home to get out and vote rather than a strategy of converting non-supporters, alternate plans were pointedly designed for capturing each voting bloc based on estimated levels of engagement and enthusiasm over the course of the campaign.

## Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

Separation of language styles is a fundamental communication strategy that is taught in most courses, but it was especially emphasized in IST 736 and PAI 897, where the ability to summarize findings and suggestions using broadly accessible language was as critical to success as the ability to conduct the research and detail the results. Both courses culminated in research papers of substantial length that were required to be digestible by both highly technical audiences that would read the reports cover-to-cover and for higher-level leaders or executives who may only have sufficient time to read an executive summary or introduction and conclusions. Each audience has unique levels of experience and engagement with the material, and they are typically interested in different things. Thus, it is important to clearly summarize key outcomes in the obvious, exterior locations and make the fine details available to interested parties in interior sections.

In addition to written communication, visual aids, when composed well, are invaluable to effective communication during oral presentations. The conversion of work from written or spoken language, verbose in both content and form, to a visual summarization relies heavily on the ability to represent data in a meaningful graphic way and to succinctly outline when the points to be made are not statistical or numeric in nature. Again, using IST 736 and PAI 897 as examples, presentations were required to complement the submitted written work. In these cases, graphs, charts, and tables were used to convey key messages, where applicable, and brief reviews, otherwise. As discussed previously, effective visual communication was also taught in IST 719, where I learned to direct the viewer’s attention to central points and guide them through a visual story with minimal use of explanatory language.

## Synthesize the ethical dimensions of data science practice (e.g., privacy)

Since a large portion of my work in the Applied Data Science program has focused on American politics and voting decisions, it is natural for me to have deeply considered the issue of data privacy in the context of political campaigning and the spread of political information. Voter data contain large amounts of highly sensitive personal information from ZIP code and age to sexual orientation and religiosity, and all data, as a rule, should be anonymized to the greatest extent possible, in the interest of individual privacy. However, when considering the specific context of the political data realm, many voters wish to be more informed, so discretionary use of personal information can be justified in the interest of spreading honest, fact-based information.

Throughout this program, I have primarily used publicly available (or available upon request) datasets, with the exception of the text mining project in IST 736, where the text content of a collection of tweets was used to predict who would run for President in a 2020 primary election. This analysis focused on federal- and state-level politicians or those who had declared a presidential candidacy for 2016 and 2020. All profiles used for data collection were public profiles of public figures, so there is little to no risk in collecting and abusing sensitive information, especially since no metadata were considered in the analysis.

# Key Conceptual Works

## The Obama Campaign

The Obama 2012 presidential campaign profoundly influenced my work throughout the program. Sasha Issenberg’s piece on the campaign’s data team describes a highly organized operation that used multiple data sources to build complex and highly specific voter-level models, essentially predicting for each eligible voter in the United States his or her propensity to support Obama and his or her likelihood of voting in the general election, a modeling strategy I attempted to replicate in MAR 653 for the fictitious Stacey Abrams Senate campaign. (How Obama’s Team Used Big Data to Rally Voters, 2012)

Examining the Obama campaign also distinguished between proper and improper use of personal data in a political context. The Obama team used Facebook to launch a “targeted sharing” effort that searched for the Facebook profiles of individuals who had already engaged with the campaign (via donating, volunteering, signing up for newsletters, etc.) and matched those users’ friends list to the voter propensity models they had already built. Users were then prompted by the campaign to share voter registration and other election information with their personal contacts who the campaign deemed to be likely supporters. (Issenberg, 2012; Scherer, 2012)

Since the 2012 election, the use of data by political campaigns has become more controversial, in particular, as a result of the exposure of the mass use of raw Facebook data by British political consulting firm Cambridge Analytica in the 2016 presidential election. Cambridge Analytica used a Facebook quiz to collect data on users who took the quiz and then exploited a loophole in the Facebook API to collect data on those users’ friends as well. (Chang, 2018; Scherer, 2012) Compared to the more exploitative and disguised tactics, the Obama team’s method of optimizing the use of sensitive data while maintaining informed consent on behalf of the user and general transparency in purpose provides a template for upholding integrity in the field of data while still producing competitive results.

## Social Desirability Bias

One problem that is faced often in polling is social desirability bias, where it is believed that oftentimes, a poll respondent may purposefully misrepresent their thoughts, preferences, or behaviors if they can be construed as generally unpopular or undesirable. (Lavrakas, 2008) While there is evidence that the extreme divisiveness of the 2016 presidential election led to social desirability bias affecting some of Donald Trump’s supporters in 2016, it is unlikely to have significantly impacted the results of well-executed polls and their respective models, since many had margins of error that allowed for a theoretical, if unexpected, Trump victory. (Klar, Weber, & Krupnikov, 2017; Morris, 2020)

That said, my work in IST 718 focused on this aspect of the 2016 election in an attempt to estimate the likelihood that a poll respondent was lying about having voted for Trump, given their responses to other Trump-sensitive questions, such as favorability toward Trump and others. Although these questions seem to go hand-in-hand and that it would be unlikely that an individual would respond with positive favorability toward Trump but then claim not to have voted for him, there were many such individuals on whom the group’s analysis was focused. Overall, our findings were in agreement with existing research: that social desirability bias certainly appears to have impacted some individuals, but not in such a significant way that it would lead to wildly inaccurate polling models.

# Application of Learning Outcomes

## Preparation for Work in Chosen Field

In my application to the Applied Data Science program, I discussed my passion for political engagement and my ambitions to work in the politics using the skills I would learn to bring sophistication to the campaigns of progressive politicians and organizations. After realizing that the program was more private-sector oriented, I again cited this passion and professional drive in my petition to allow credit for the policy analysis course from the Executive Master of Public Administration program.

Despite the differences between the progressive political sector and the traditional private sector, I have found myself amply prepared for work in either field due to the near-universal applicability of concepts and the flexible environment in which these concepts were taught, which enabled me to practice data science on politically relevant research questions and datasets while still working within the framework of business improvement and decision optimization.

By centering my projects around the political dataspace, I have gained exposure and valuable insight into survey design and data collection, especially where the topic is so sensitive in nature—as some of the questions asked in political polling are—that the question must be asked in a precise and neutral way in order to get both accurate and honest responses.

The techniques for data processing, exploration, synthesis, and analysis studied in the program, again, are relevant in almost any sector using data for insight. While topics and critical themes vary, the techniques taught for business operations improvement and marketing strategy optimization have parallel applications in campaign strategizing and voter targeting.

And finally, as in business, the leaders of large political campaigns and organizations have limited time and energy to devote to thorough investigations into highly technical spaces, so the ability to effectively communicate key results and their implications is as invaluable in this field as in any other.

## Areas of Strength and Challenge

The benefit of working in the political space is that elections occur almost every year, and each year, more research goes into understanding how voters behave. As a result, there is an ever-increasing supply of data on endless topics within the field. This, combined with my foray into policy analysis via PAI 897, provides the flexibility to work in many different subject areas, from policy reform to voter engagement to campaign strategy, while sustaining a demonstrated ability in performing both complex analysis of raw data when available and intricate synthesis of existing bodies of research when not.

Despite the abundance of data being generated in this space, many datasets present significant obstacles, including selection bias, the aforementioned social desirability bias, and privacy concerns. Not to mention the volatile nature of politics and politicians and that many voter data sets are not published. One technique that is often used in polling to mitigate the negative impact of some of these obstacles, especially in election polling, is observation weighting to make a sample more representative of the population. Unfortunately, this concept was not covered in any of the courses in which I was enrolled, so I will have to engage in extracurricular research in order to obtain the skills considered essential to poll analysis.

## Life-long Learning

As I begin to pursue a career in political data science, pivotal in my success will be the ability to keep up with new developments in the ever-evolving field of data science as well as to acquire the political field-specific techniques that I did not learn in the Applied Data Science program. Fortunately, this program has laid the technical and conceptual groundwork necessary to render each of these tasks not just achievable, but also personally gratifying. I look forward to continuing to develop unique solutions to complex problems that coincide with my personal interests.

# Appendix A: GitHub Repository

<https://github.com/lplawless/Applied-Data-Science-Portfolio>

# Appendix B: Course Key

|  |  |  |
| --- | --- | --- |
| **Term Enrolled** | **Course Code** | **Course Name** |
| Spring 2019 | MBC 638 | Data Analysis and Decision Making |
| Spring 2019 | IST 659 | Data Administration Concepts and Database Management |
| Summer 2019 | SCM 651 | Business Analytics |
| Summer 2019 | IST 687 | Introduction to Data Science |
| Fall 2019 | PAI 897 | Fundamentals of Policy Analysis |
| Fall 2019 | IST 707 | Data Analytics |
| Winter 2020 | MAR 653 | Marketing Analytics |
| Winter 2020 | IST 718 | Big Data Analytics |
| Spring 2020 | IST 736 | Text Mining |
| Spring 2020 | IST 652 | Scripting for Data Analysis |
| Summer 2020 | IST 719 | Information Visualization |
| Summer 2020 | IST 664 | Natural Language Processing |

# Works Cited

Associated Press. (2019, May 18). *AP VoteCast 2018*. Retrieved from The Associated Press and NORC: https://apnorc.org/projects/ap-votecast-2018/

Chang, A. (2018, May 2). *The Facebook and Cambridge Analytica scandal, explained with a simple diagram*. Retrieved from Vox: https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram

Democracy Fund Voter Study Group. (2019, January). *Views of the Electorate Research Survey*. Retrieved from Democracy Fund Voter Study Group: https://www.voterstudygroup.org/data

Issenberg, S. (2012, December 19). *How Obama’s Team Used Big Data to Rally Voters*. Retrieved from Technology Review: https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/

Klar, S., Weber, C. R., & Krupnikov, Y. (2017, February 22). Social Desirability Bias in the 2016 Presidential Election. *The Forum, 14*(4), 433-443.

Lavrakas, P. J. (2008). *Social Desirability*. Retrieved from SAGE Research Methods: https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n537.xml

Morris, G. E. (2020, June 14). *Are polls overestimating the Democrats again?* Retrieved from The Crosstab: https://thecrosstab.substack.com/p/are-polls-overestimating-the-democrats

Scherer, M. (2012, November 8). *How Obama's data crunchers helped him win*. Retrieved from CNN: https://www.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team/index.html