WITTY TITLE HERE

BEN HARWOOD

Abstract. This paper is a... memoir of my journey through the Applied Data Science graduate program at Syracuse University. Submitted for fulfillment of the Portfolio Milestone requirement.

Introduction

My journey toward a master's degree in data science actually began in early 2018. To spare the reader a sob story, my family and I had just come off our worst year (medically, financially, emotionally) and I was ready for a change. Being the only person that could do anything in Excel besides enter values in a spreadsheet at the car dealership I work for, I had been doing some things with customer service survey data[1] and had become very interested in data analytics and data science as a result.

I applied to Syracuse and was offered a conditional acceptance (they wanted me to take IST 659 as I had no coding background of any kind, a course that I went on to miss 5 out of 300 points in), but I couldn't afford the tuition for a single class. As a result I went shopping for a different program and found one not too far away at what became known, affectionately, as the School-That-Shall-Not-Be-Named. Let's just say things there didn't work out[2], so I reapplied to Syracuse and was unconditionally accepted.

Admittedly, I was very nervous. I had no coding background (except LaTeX), had not done any real mathematics in 15 years, and had never taken an actual course in statistics[3]. Even still, I was confident that once I got back into the groove of school and adapted to both the synchronous learning model and the material in general that I would be OK.

Date: March 7, 2021.

[1]Including producing a paper for my managing partner with suggestions on how we could adjust the bonus/penalty structure for perfect and bad surveys (this is not included for privacy reasons), as well as studying the surveys themselves to see if I could predict who would complete their survey (also not included for privacy reasons, because **ethics** is importance in data science).

[2]I got into a philosophical argument with the sociologist teaching multivariate statistics (purely with SPSS) about the nature of arithmetic, and over the course of the program there was no SQL, only one class that used R, only one that used Python, and four that had nothing to do with data science whatsoever.

[3]I've often quipped that I took every course the mathematics department offered as an undergraduate (plus six independent studies) EXCEPT anything related to statistics.

If I've not yet bored you, the remainder of this paper is a reflection on my time in the program, a synthesis of (most of) what I've learned, and a consideration on the future. I have made a concerted effort not to make it a play-by-play report on each of the courses I took, but it is difficult to adequately reflect on my experience without discussing them in some way. The first appendix is a detailed listing of the projects mentioned throughout the paper, and the second appendix (which is totally optional, but included as an interesting aside and insight into a direction I may go in the future) includes a description of a problem I had hoped to undertake on the side in coordination with a member of the mathematics department but I simply didn't have enough time.

1. What is Data Science?

This is a question I asked myself over and over while I was investigating different programs. Imagine my surprise to learn that there really isn't a concrete definition of what data science is. From one perspective, it seemed like it was just statistics with computer programming laid on top, which quite honestly was not very appealing. However, I kept reading that while it was heavily focused on statistics, a strong grounding in multivariable calculus and linear algebra was essential for machine learning, so my interest continued.

Much like any other academic discipline, the road to really having a grasp on data science does not come in the beginning. Not knowing what to expect in any of the courses in the program left me somewhat disoriented as I began learning basic statistics. But the gravity of why statistics is so useful and how I could begin applying what I was learning, even at the beginning, hit me while deciding what to do for course projects in both IST 659 and MBC 638. Having to design and build a database for IST 659 (Harwood, 2019a) while also finding a problem to apply Six Sigma processes for MBC 638 (Harwood, 2019b), it seemed like I could use the same problem for both projects. This would allow me to approach the problem in different ways, while also applying the concepts of database management and decision making (and learning SQL). I took this opportunity to tackle a problem at the dealership, and while I was ultimately unsuccessful at solving it, I did gain valuable insight into several things related to data science in general.

The data I used for the 638 project did have potentially sensitive personal information included. In the interest of avoiding bias, I made sure not to include any of that in the analysis. This was not an issue with the project for 659, in fact the personal data was necessary, but privacy is still a concern, and if the

database and ui were to be actually implemented it would have to be setup so that the personal data was only available to specific individuals that would actually have need of the data. Again, **ethics** matters in data science. I thought I'd have all the data I needed. This turned out to be the Achilles heel of the project for 638, because while I was able to draw some conclusions, the data did not provide enough insight into the problem to make a solid recommendation to my managers. Additionally, I built this nice, clean, well organized inventory database especially for use in tandem with the 638 project, but it wasn't until the project was nearly complete that I realized I couldn't actually do anything with it besides add new records. So in the course of these two connected projects, I grasped the importance of both sufficient data and that the scope of a project should be considered (something I wish I had learned while working on my undergraduate thesis). Both of these lessons came back while working on the project for IST 687 (Bump et al., 2019). About halfway through we discovered that in order to have enough evidence to justify the adoption of the designated hitter in the National League we needed to **alter our approach** because we needed more information than just batting statistics: we needed to relate those batting statistics with wins (under the assumption that wins generates ticket sales). Thankfully there was another dataset about baseball team managers that had win records included, so we were able to integrate that into the study and run regression models to see which batting statistics were actually related to wins, and ultimately make our recommendation.

There it was, my first real data science-y study, one that reflected the "data science process" as outlined in O'Neill and Schutt (2014), as seen in Figure 1. The project followed every step of the process except for the building and deployment of a data product.

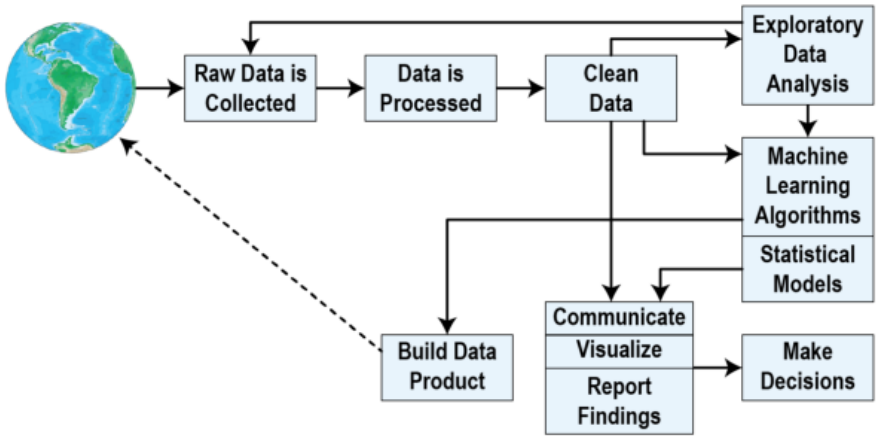

FIGURE 1. O'Neill and Schutt Data Science process

Having written (and published) numerous mathematics papers (see (Harwood, 2002) for example), this first statistically motivated paper was a challenge to be sure, as the format and style prevalent in most mathematics papers is much less... anonymous. Indeed that was essentially the only fault in the paper pointed out by our professor (costing us our single point reduction out of 25), we even earned a "Overall, this was one of the best and most comprehensive studies presented during my year teaching this course. Nice work!" for our efforts. Admittedly I was a bit taken aback by him pointing out the tone of paper since that's what I was used to using. That being said, I have learned to embrace it as this writing style and format dilemma was eventually solved for me (as a course requirement), as will be discussed in Section 3.

At this point I had a fairly good idea of what data science actually was. At least I thought I did. I remember being paranoid about ensuring we had sufficient visualizations of the data while writing the project for 687, but ended up producing some less than stellar plots. Such as in Figure 2. For some reason I (and my group) thought that plot was suitable for comparing the two batting stats, but looking back after taking IST 719 we were sadly mistaken. This was an important lesson to be learned, especially together with the writing style lesson. An important part of being a data scientist is the ability to effectively **communicate** your findings, so good write-ups and quality, effective **visualizations** are extremely important, as your results are useless if people can't make sense of them.

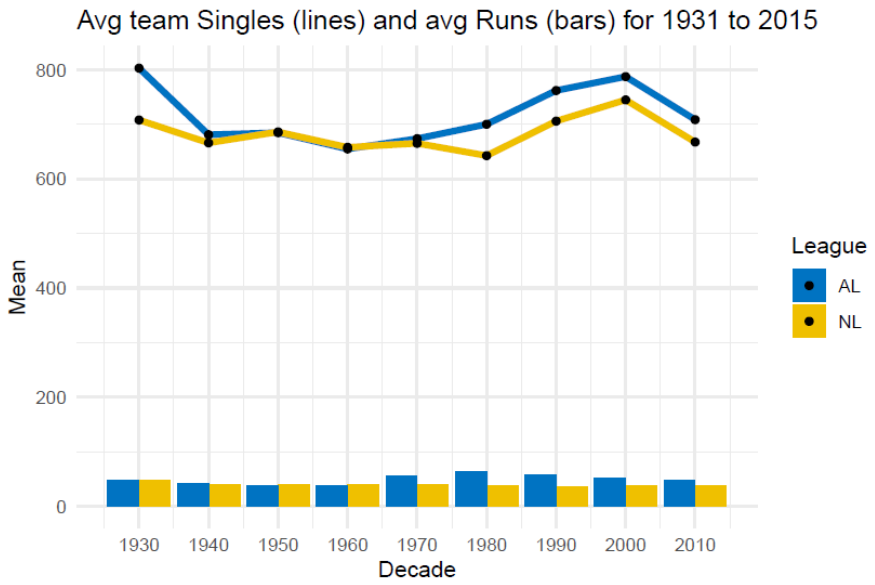

FIGURE 2. No good plot from IST 687 project

For example, consider the plot in Figure 3. This plot was part of my final poster for IST 719 (Harwood, 2020b). Without going into super detail, I was studying the competitive eSports side of the popular online game *World of Warcraft* and looking at how the different playable character classes were utilized. The plot in Figure 3 should be self-explanatory, but the goal was to show in a unique but still understandable way the most common DPS (damage-per-second) class combinations. This plot received rave reviews from my classmates during and after the presentation, and I like it because it conveys a good deal of information but in a way that is readable, aesthetically pleasing, and easy to interpret. While many other plots were made for that poster (it was, after all, a course on data visualization), this one feels like my visualization eureka moment.

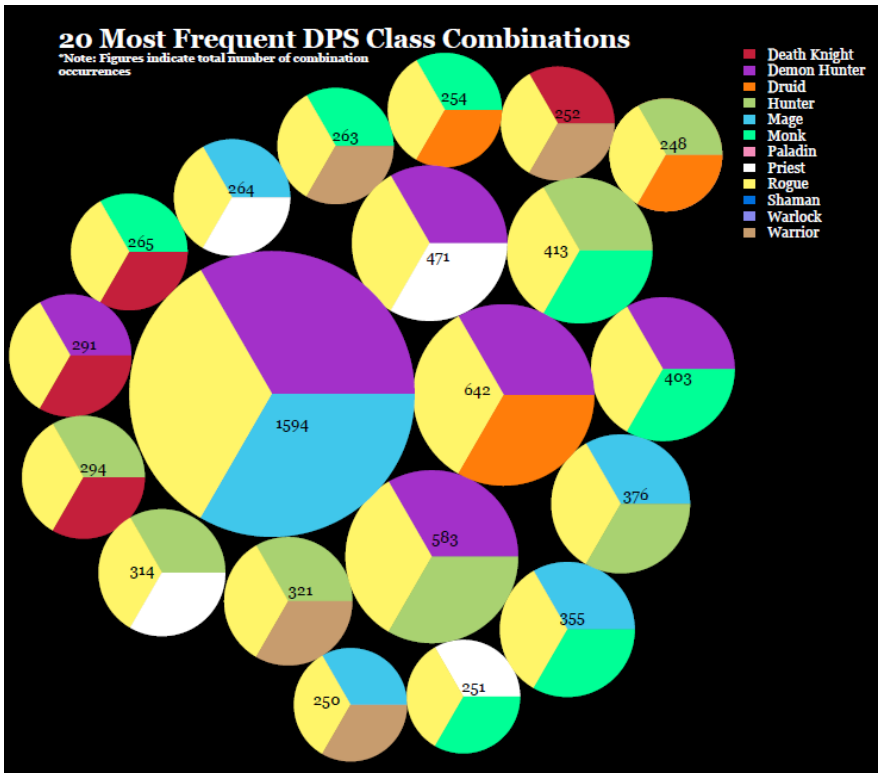

FIGURE 3. Circle packing plot from IST 719 poster

The project for 687 didn't use anything besides some basic statistics. There was no data mining, no machine learning, just some plots and linear regression. So maybe I didn't really know what data science was. At least not yet.

2. Progression

Note: this section really doesn't reflect much about the courses and learning, more the arduous process I went through determining my path of progression through the program electives.

Before the program even started, I settled on the following for my electives: IST 652; IST 719; IST 722; IST 769; something else (ultimately IST 772 at the suggestion of my success coach). Scripting seemed like a no-brainer since I had no computer science background and it would be additional Python exposure. Visualization seemed almost mandatory after 687 and 707 because of the emphasis on visualizing data. And it seemed like 722 and 769 would be good to have for additional SQL[4] and knowledge of additional database systems, because it felt at the time that having those skills in addition to Python and R would make me more... marketable to companies. So I enrolled in 719 and 769 for the Spring 2020 term. Then COVID hit and everything changed.

My multiple sclerosis treatment is immune compromising, putting me in the high-risk category. Thankfully my general manager is a decent human being and told me, "I appreciate you coming in and running around for us, but it isn't worth the risk to your health," and subsequently furloughed me. I was home from March 20 until June 14 (almost my six-year work anniversary). I mention all this because it obviously gave me a LOT of time at home to devote to school and helping my children with their schoolwork, and I remember walking around my basement having a conversation with myself[5] questioning whether I was on the right track with my classes. Then when classes started there was something about 769 that didn't jive with me, and I swapped into something completely random that I told myself I had no interest in (but it would provide additional Python): IST 664.

During this term (dubbed the COVID term), I began a job search. A car dealership didn't seem like an ideal place to flex my data science skills, and one of my overall goals of this journey was to get out of the dealership anyway.[6] During this job search I kept seeing things like Hadoop, Spark, Hive, and other advanced database software, which kind of made me kick myself for dropping 769. But at the same time I

[4]When I first started at the School-That-Shall-Not-Be-Named I saw a graphic that listed SQL, Python, and R as the most in demand languages for data science, and one course in SQL didn't feel sufficient to me

[5]Think Russell Crowe as John Nash in *A Beautiful Mind*

[6]This seemed like the right time to start this search because my success coach had told me early on that after 707 students reported "knowing enough to be dangerous".

truly enjoyed NLP (a course whose final project (Harwood and Vogel, 2020) laid the foundation for my final project in 652 (Harwood, 2020a) possibly 736), and I decided at the point I wanted to take text mining. But that left me with too many classes to take and not enough terms. I briefly thought about taking three classes at once, but quickly thought better of it after friends' reports of the workload in 722. Plus there was 772 (I figured it would be helpful to have taken a course in statistics that actually covered the concepts and wasn't a "how to use SPSS" class), and I still had the required 718 to take. Quite the pickle I found myself in, to be sure.

Six courses and two terms to take them in with only two courses per term. I'm good at mathematics but that arithmetic doesn't work in any number system I'm familiar with. There were really only two possibilities: 1) Skip a class that is either of high interest or is (at least perceived) high importance; 2) Take an additional term, stretching out the program (and adding more student loans). Neither option was especially appealing, and I actually bounced back and forth several times, finally settling on taking the additional term. Ultimately, I was, and am quite happy with this decision, as I will end up with the education I was anticipating and feel will best prepare me for a new career.[7]

3. Getting My Legs

I mentioned at the end of Section 1 that after two terms I still didn't really have a strong feeling for what data science is. More specifically, I knew what it was but hadn't really DONE any data science yet. That changed while working on the project for IST 707 (Harwood et al., 2020). This was the first time where the fact that there is more to data science than just statistics really became evident. We spent a long time even figuring out what to do for the project, ultimately settling on seeing if we could determine or even predict how and where the United Way places its locations. This was a good **data collection** lesson, because we primarily used Census data, however we did have to manually find some of our own data about the United Way. Additionally, we had to create our own variables and form different data frames to use for association rule mining, support vector machines, and other algorithms. Then there came writing the report. We had a very specific format that we were required to follow, one that felt excessive and unnecessary at first, but

[7]Indeed Professor Fox told me, when I swapped out of 718, "... I would say (as a part time corporate recruiter) that focusing on SQL/data models/databases is a very good choice - too many candidates these days can use python/R to run a model but the really talented data science folks can also get the data and then code the algorithm in a way that is efficient..." which served as a level of confirmation of my decision.

after using it half a dozen times it became clear how beneficial it was. The very distinct and clear format allowed for very effective **communication** of the results (as long as it was well-written, of course). This provided a strong framework most assignments and every subsequent project submission, except of course for IST 719 which was a poster.

Speaking of 719, since this course is called "Information Visualization" it should come as no surprise that this course was reinforcement that adequate **visualization and representation of data** is paramount to the understanding of data, not just to the data scientist during exploration but also in the presentation and communication of findings. Additionally, the idea of seeing at the same data in different ways was instrumental in completing the poster. It was at this point where the first dilemma described in section 2 popped up. But we'll get to that later, we need to talk about something that kept coming up in every class first: statistics.

I don't remember seeing IST 772 on the electives list,[8] but it seemed a very good course to add to the mix. Despite having no interest in statistics for much of my adult life, this ended up being one of the highlights of the program for me, and while neither midterm nor the final for this course were actual projects, they were set up in such a way that I was able to complete the tasks and formulate the responses as if they were actual studies. These were both opportunities to not only complete statistical analyses in (simulated) real-life scenarios, but they were great practice in **communication**, and the midterm had inherit ethics and **data privacy** concerns to consider.[9] My submission for the final (Harwood, 2020c), whose only feedback was, quote, "Outstanding work, Ben! I like the way you brought everything together as a cohesive research project and presentation. The scored rubric is attached, but it's perfunctory," I consider to be one of my best pieces of work in the entire program.

4. Politics

(Don't worry, we're not actually talking politics)

As a teenager, I despised studying grammar. I remember distinctly doing very badly on grammar quizzes my sophomore year in high school, but that was primarily because much of what was being quizzed on built

[8]Though it was almost three years ago when I was initially started this entire process, so my recollection of this could be... fuzzy.
[9]Really, the data was not provided, only certain aspects of the data and results of certain analyses that we just had to interpret, so that allowed for the simulation of not sharing data due to it being proprietary or sensitive.

from concepts I had not been taught in elementary and middle school (different school districts). Ironically, bad grammar has become a personal pet-peeve. Because of this apprehension for studying grammar, NLP was not something I was remotely interested in when the program started. But as discussed in section 2, it seemed a good opportunity to get some Python under my belt.[10] So imagine my surprise when I found myself sucked in after the first assignment. While not included in this portfolio, my submission for the first assignment looked at the similarities between the inaugural addresses of the current and previous governors for the Commonwealth of Kentucky, which conveniently represented each party. This assignment served as a starting off point for not only that course's final, but also the final projects for both IST 652 (Harwood, 2020a) and IST 736 (Devito et al., 2020). Ultimately, the NLP course (and its final project) served the purpose I was hoping for, as it helped me bridge into Python (even though both the asynchronous videos and professor in the live sessions specifically said it WAS NOT a Python course). Indeed the process of developing the final project code helped me develop some methods and techniques that I re-used in 652, 736, and 718. Said project was supposed to be doing one of three classification tasks, however I had become so enamoured with the possibility of being able to call BS[11] on a politician while they were speaking that I asked to expand on the first assignment and see if political affiliation of a governor could be determined by the text of their inaugural address. This project was my biggest source of difficulty in terms of **data collection and organization**, because we wanted to use consecutive governors where the part in power switched (e.g. Kentucky in 2016 and 2020) for each state. This proved problematic because there does not appear to be an archive of gubernatorial addresses, and in some states the party in power has not changed for many years, thus the vocabulary in those instances is likely very different because of time period. This was also a good exercise in **ethics** because we had to make a concerted effort to not introduce our own political beliefs or viewpoints into the investigation.

Challenges in **data collection** cropped up once again while completing the project for IST 652. Once again I went looking for governor speeches, this time for state of the state addresses from 2016. My goal was to see if there was a relationship between the content and/or sentiment of the governor's address and how their state voted in the 2016 election. Finding the election data was straightforward, but once again I

[10] It was also the only course that had open seats to swap into at the time.
[11] Please forgive the colloquialism

was impeded by a lack of a centralized archive of governor speeches. In addition, I was only able to locate speeches for 36 states (due to seven governors not giving addresses that year, and the others just not being available). Not to be deterred, I proceeded with the study, **adapting my strategy** along the way and including not just the sentiment of the speech but also considering demographic variables that may have contributed to a county and/or state switching parties from the 2012 election to the 2016 election. This one is included because it is one of the few solo final projects I've had to complete in the program, and used elements of multiple previous classes to get to the final result.

To finish our political discussion, we have to talk about IST 736. Similar to NLP, I hadn't planned on taking text mining initially. I didn't see the general benefit of it. A combination of the professor, my enjoyment of NLP, and what I perceived to be an extension of NLP techniques AND additional Python exposure, I jumped on it. Admittedly, I wasn't as enamoured with the course as I thought would be, mostly because much of what was covered was a rehash of things covered in other courses, just done in Python and focused on text data, but the problem my team and I chose for our final project was truly interesting. The course took place during the election, and we wanted to see if we could predict the party affiliation of the person making a tweet. Once again, **data collection** was a challenge, because we had to determine the right tweets to use for training our models. We settled on collecting tweets from 9 left-leaning accounts (Barack Obama, MSNBC, etc) and 9 right-leaning accounts (Mitt Romney, Sean Hannity, etc). This allowed us to label the tweets and accurately build our models. But once we had the models, we still needed something to apply the model to. Sure, we could determine the political parity of a tweet, but we started wondering what benefit could come from this. Since this was done during an election where the pollsters got things very wrong, we thought it would be interesting our models could determine if, without loss of generality, a seemingly democratic tweet could actually be more republican. So we **adapted** our methodology so that we could **implement** our study and collected 250 tweets each for 2 slogans related to the Biden campaign and 2 for the Trump campaign. Amazingly, our models were able to identify tweets with slogans from one side whose verbiage classified them as the other side. For example:

> "Trump lied to MAGA and the American people Demand a Covid Commission Coronavirus updates Every hour Americans are dying AstraZeneca plans additional trial after error hospitalizations hit record again via usatoday"

Now there is a lot in this tweet, and half of it seems to be a redirect ad to USAToday, but while it has "MAGA" included (which is almost a conservative rallying cry at this point), the tweet is not exactly a positive use of the slogan. The model predicted, the authors believe correctly, this to be a democrat tweet.

Conclusion

It's been quite the interesting 21 months, to be sure. Friends and family have continually asked how school is going (though I often suspect it's purely out of courtesy), and as the program winds down I look back fondly. I've made good friends, my LinkedIn network has grown by over 25%, and I have submitted an application to continue my education in the iSchool's Doctorate of Professional Studies in Information Management. The experience has been the... reset that I needed, has opened new doors for me, and I will be a strong proponent in the future should my input ever be requested.

Appendix A. Individual Class Project Synopses

This appendix is an itemized list of the different projects referenced throughout the paper. Links to each project in my Github repository may be found in the references.

MBC 638 Data Analysis and Decision Making.

Objective: Identify and improve upon a process with Lean Six Sigma methodology.

Problem: Reduce the time new cars sit stagnant on a dealership lot. This problem was actually posed to me by my general sales manager. The issue is that as vehicles sit on the lot they count against our floor plan. The goal was to determine what aspects of a vehicle (paint, interior color, features, etc.) may contribute to how long it sits before selling.

Solution: Done mostly with Excel (and a bit of R), statistical methods such as control charts, correlation, and linear and logistic regression, it was determined that the only factor that had any definitive impact on the length of time a vehicle sat prior to sale was the month in which the car was wholesaled to the dealership. With that said, important data regarding the individual options

and packages for the vehicles was not available, and it is highly likely that this data would have shed more light.

IST 652 Scripting for Data Analysis.

Objective: Write a Python program to access and collect data from multiple sources and types (both structured and unstructured), and process and study the data.

Problem: Determine how much influence a governor can have on how their state votes in presidential elections

Solution: I collected data from the 2016 presidential election, as well as state of the state speeches for 37 states (some governors didn't give a speech that year, and some I just couldn't find). The election data did not require any real processing, except for dropping the state of Alaska because it had no data represented. A list of counties that switched parties from the 2012 to the 2016 election was created. Next, the speeches were processed using NLP and had their overall sentiment determined. Everything was collected together in anticipation of regression models to see which variables might impact "jumping" (switching parties as mentioned earlier), and a χ^2 test for independence between the party of the governor and who a state voted for in 2016.[12]

IST 659 Data Admin Concepts & Database Management.

Objective: Design and implement a database to solve a data management problem of your choice.

Problem: An inventory management system for an automotive dealership

Solution: I chose this to serve as a kind of custom solution to help with the project topic for MBC 638. Using Microsoft SQL Server I developed a database with custom procedures to add vehicles into inventory, create a list of clients, "sell" a vehicle to a client (removing it from inventory but retaining record of it), and performing some basic analytics tailored to the MBC 638 project. For usability, I opted to build a fairly straightforward Microsoft Access front-end.

IST 664 Natural Language Processing.

Objective: Classification of Text

[12]This χ^2 test was not significant, so independence was rejected

Problem: Inaugural Gubernatorial Speeches

Solution: We collected Democrat and Republican inaugural addresses for 10 states when the party changed (e.g. Kentucky in 2018 switched from Republican to Democrat). We looked at each parties speeches as a collection to see how much of what they said was substantive (for example, we discovered that while the Republicans used 907 more unique words than the Democrats, their speeches were 54.47% stop words, so they had more to say but more of it was "fluff"), and any common themes, surprisingly finding that the Democrats used the phrase "God bless" more frequently than the Republicans. The speeches were tokenized into sentences and numerous multinomial Naïve Bayes models were built for different choices of features (negation, subjectivity, etc) following the following methodology:

- Each of the top 2,000 words that appear in a sentence is assigned a binary label reflective of whether it matches the current feature of choice
- The collection is then divided into training and testing sets (initially varying the training/testing ratio)
- The training data was used to train a Naïve Bayes classifier algorithm within NLTK, and the resulting model was applied to the testing data
- Accuracy was measured, using n-fold cross-validation to ensure a good model had been developed
- Micro-averaged F1 score was calculated for a better measure of performance of the model.

This process was performed with and without stop word removal. Ultimately, we were able to predict with 0.68 accuracy whether a sentence from one of the speeches was spoken by a Democrat or a Republican.

IST 687 Applied Data Science.

Objective: Pick and analyze a dataset.

Problem: My group and I studied data from Major League Baseball to determine the effectiveness of the Designated Hitter and whether it should be adopted by the National League.

Solution: For this project, we used multiple data sets found on Kaggle comprising batting statistics for the history of Major League Baseball. Using R, we looked at different batting statistics by decade for each of the leagues to see how they compared, and saw that after 1973 when the DH was introduced, the batting average lead switched from the National League to the American League, and the average team run production went from even to the American League having a nearly 50 point advantage. Given the assumption that more hits and runs leads to more exciting baseball which would lead to more ticket sales, we ultimately recommended that the National League adopt the DH.

IST 707 Data Analytics.

Objective: Solve a data mining problem

Problem: Can we predict where the United Way places in locations?

Solution: Using Census data and custom collected data on the United Way, we performed various data mining tasks to look for relationships. We used Kentucky and Tennessee as our baseline (mainly for demographic reasons). On a county-level, we assigned a label indicating whether there was a United Way location. In the top 20 most frequent itemsets from association rule mining, 12 indicated that a county in the fourth quartiles for child poverty, per capita income, and Native American population would have a United Way location. Various additional techniques were employed, specifically k-means clustering, Naïve Bayes, support vector machines, k-nearest neighbors, decision trees, and random forest. Each was done with 3-fold cross-validation, and k-nn was the most consistently accurate in its prediction on the test set. We then applied the models to the state of Ohio to see how well Kentucky and Tennessee data could predict Ohio, alas nothing could get above 0.59 accuracy on Ohio. This suggests that there may be other factors behind where United Way chooses its locations other than demographics.

IST 719 Information Visualization.

Objective: Create and present a poster

Problem: Class distribution in World of Warcraft Mythic+ Dungeons

Solution: As some one who played the game for (far too long) a time, and who enjoys the competitive E-sports side of the game, this seemed like a good topic. Using R I connected to the API (my first time using an API, in fact) from raider.io, a website and game add-on that tracks and ranks player performance in the game. Focusing on the top 100 runs for each dungeon in each "season" (translated: major game patch) for the most recent version of the game (as of May 2020), I looked at the different character class distributions for the entire expansion, the distribution of each class across the 12 dungeons available[13], how popular each class was for the three roles of tank, healer, and dps for each season, as well as the frequency of tank and healer combinations and the frequency of dps combinations. I shared the poster with one of the E-sports commentators online community and received quite positive feedback, and will probably do the same thing for the newest version of the game.

IST 736 Text Mining.

Objective: Solve a real text mining problem.

Problem: Can Twitter predict political affiliation?

Solution: This project was completed right around the 2020 election, so we thought there would be lots of good data. We collected 250 tweets from each of nine Republicans and nine Democrats. Stop words were removed, using a custom list consisting of common lists from the Natural Language Toolkit (NLTK), sklearn, and other words specific to tweets, such as 'rt', 'retweet', and 'amp'. A regular expression was also used to remove non-alphabetic words and word less than 4 characters in length. Finally, re-tweets and URLs were also removed. This was done with a custom-made function and was crucial for removing links. The tweets were then vectorized with different methods and had support vector machines and multinomial Naïve Bayes models applied to see how well the political affiliation of the user could be predicted. We then collected 250 tweets for each of four popular hashtags related to the campaign, to which we applied the models. We were able to show that even when the topic hashtag had a certain political leaning the word usage could reveal it was more a tweet from the opposite party.

[13]I fully acknowledge that the reader may not have the necessary context for this conversation

IST 772 Quantitative Reasoning in Data Science.

Objective: Final exam

Problem: Analyze multiple datasets pertaining to vaccinations to make recommendations to a scientifically inclined member of a state legislator's office

Solution: Multiple statistical methods (t-test, time-series analysis, linear and logistic regression, including Bayesian variants) were employed to address the following questions:

(1) How have vaccination rates at the national level varied over time?
(2) Which vaccinations have the highest and lowest rate of vaccination, and which vaccination rate was the most volatile?
(3) Is there a credible difference in reporting proportions between private and public schools?
(4) How do vaccination rates in California compare to national vaccination rates?
(5) How are individual vaccine rates related among districts, i.e. if a student is missing a single vaccine are they missing others?
(6) Can whether each school in a district reported vaccination rates be predicted, and if so how?
(7) Can the percentage of all enrolled students that are up-to-date on all of their vaccinations be predicted, and if so how?
(8) Can the percentage of all enrolled students claiming religious exemptions to vaccination be predicted, and if so how?

Every variable in one of the datasets was highly correlated, and all but two required transformations to reduce skew. Ultimately, the following was discovered:

(1) Vaccination rates have, overall, increased over the last 25 years, however some vaccines are more prevalent than others.
(2) Even though it has the highest vaccination rate, Hepatitis B seems to be a sticking point insofar as it is indicative of other vaccines being missed.
(3) Some focus needs to be placed on private schools, as they are less likely to even report their vaccination rates than public schools.
(4) With the exception of DTP, California is ahead of the US in terms of vaccination rates.

(5) The percentage of students with free meals, family poverty, and enrolled students in a district can predict the percentage of students without the HepB vaccine, which is a strong indicator that a student does not have the other vaccines.

(6) Except for child poverty, each demographic variable (individually) can predict whether a district was complete in its vaccination reporting, though the strength of the prediction is not large.

(7) The demographic variables together (again, except for child poverty) can predict the percentage of the student body that is up-to-date in their vaccinations.

(8) The number of enrolled students and (surprisingly) the percentage of students with free meals can predict the percentage of students claiming religious exemption to vaccinations.

Appendix B. Topological Association Rule Mining

Association rule mining fascinated me when I first learned about it. Being in retail for my entire adult life I easily related to the idea of market basket data. But I saw something deeper under the surface. This appendix will elaborate. Note to the reader: This appendix is rather technical, and could very easily be omitted entirely. However I have chosen to include the discussion and the advanced mathematics required because it does actually represent a level of synthesis of some of the learning objections of the program.

As a reminder and establishment of notation, let $I = \{i_1, i_2, \ldots, i_d\}$ be the set of all items in a market basket data set, and $T = \{t_n : t_n \subseteq I\}$ the set of all transactions. Any subset $X \subset I$ is an **itemset** and we say any transaction t_m such that $X \subset t_m \in T$ contains X. We consider the **support count** of an itemset X:

$$\sigma(X) = \big|\{t_j : X \subset t_j, t_j \in T\}\big|,$$

i.e. the number of transactions that contain X. The percentage of transactions that contain an itemset is called the **support**:

$$s(X) = \frac{\sigma(X)}{|T|},$$

and of course we call an itemset **frequent** if its support is above a certain threshold. We use these to develop **association rules**, implication expressions of the form $X \to Y$ where $X \cap Y = \emptyset$. We can measure the

support, confidence, and lift of these to measure the strength or quality of the rule:

$$\text{Support,} \quad s(X,Y) = \frac{\sigma(X \cap Y)}{|T|};$$

$$\text{Confidence,} \quad c(X,Y) = \frac{\sigma(X \cap Y)}{\sigma(X)};$$

$$\text{Lift,} \quad l(X,Y) = \frac{\sigma(X \cap Y)}{\sigma(X)\sigma(Y)}.$$

All of this should be familiar, modulo notation. Recall also the *APRIORI* principle, a theorem stating that any subsets of a frequent itemset are themselves frequent. Now, for any set, we define a **topology** as follows:

Definition 1. Let X be a set. Then a **topology** on X is a collection of subsets $\mathcal{U} = \{U_i : U_i \subseteq X\}$ such that the following are true:

(1) Both the entire set and the empty set are members of $\mathcal{U}$

(2) The intersection of a finite number of the U_i is in $\mathcal{U}$

(3) The union of an arbitrary number of the U_i is in $\mathcal{U}$

We call the U_i the **open sets** of the topology, and the pair $(X,\mathcal{U})$ is called a **topological space**. So what?

We know that subsets of frequent itemsets are frequent by the *APRIORI* principle. I started wondering if frequent items would form a topology on the set of transactions, and if so what would it mean? It turns out they do form a topology, but only on the maximal frequent itemset. Not especially interesting.

However, the use of algebraic topology to study data is an active area of research and practice. Carlsson (2009) lays the foundation for the use of what is called persistent homology to study the underlying structure of data. While a thorough treatment is far beyond the scope of this discussion, if you're still with me let me walk through the basics. Instead of a general discussion, this will be directly related to association rules, at least, as much as I've been able to put together.[14]

We need a notion of distance between transactions. Because they are sets of items, distance in the traditional sense doesn't make any sense. However, there is a measure of similarity that we could use. With

[14]I've never had a class in algebraic topology, or even standard topology, so everything is effectively self-taught and based on one conversation with Stephen Wehrli from the Syracuse Mathematics department.

notation as before, let $t_j \in T$. We define the **Jaccard Index** of two transactions as follows:

$$J(t_j, t_k) = \frac{|t_j \cap t_k|}{|t_j \cup t_k|}.$$

This measures the similarity between two transactions without considering those items that do not appear in either transaction. This is preferable since most transactions will have far more items NOT included than those included. While this is a perfectly fine measure of similarity, it is not a distance measure, because it doesn't satisfy the properties of a metric. However, $1 - J(t_j, t_k)$ does, as we now show.

Proposition 2. *The quantity $J_\delta(t_j, t_k) = 1 - J(t_j, t_k)$ is a metric on the set of transactions.*

Proof. The Jaccard index is necessarily positive, but also less than 1, because the size of an intersection is less than the size of the union. So J_δ is clearly positive.

Now, note that $J_\delta(t_j, t_k) = 0$ is equivalent to $|t_j \cap t_k| = |t_j \cup t_k| \neq 0$. The result is clear if $|t_j| = |t_k|$. Because $|t_j \cap t_k| = |t_j \cup t_k|$, then for every $x \in t_j \cup t_k$ we have $x \in t_j \cap t_k$, which implies $x \in t_j$ and $x \in t_k$ for every x. So every element of the union (which by definition is the collection of elements of both sets) is a member of each set, which implies $t_j = t_k$. This shows that, $J_\delta(t_j, t_k) = 0$ if and only if $t_j = t_k$.

Symmetry is clear, leaving only the triangle inequality to prove. Unfortunately this is not a trivial matter and we refer the reader to Gilbert (1972) or Kosub (2016). □

It is important to note that J_δ is applicable to sets in general, not just market basket transactions. Applying this idea to general itemsets will help later.

With a metric on T, we have an inherit topological structure on T, by defining the open sets $U \subset T$ to be those such that for any $t_j \in U$ there exists $r > 0$ such that $B(t_j; r) = \{t_k \in T : J_\delta(t_j, t_k) < r\}$ is a subset of U. We now have a topological space created from our transaction data, with which all sorts of things can be done. This is one point where my investigation stalled due to time and lack of knowledge.

There is another direction that could be pursued which is in the realm of the contents of Carlsson (2009).

Definition 3. An abstract simplicial complex is a pair (V, Σ) where V is a finite set and Σ is a family of non-empty subsets of V such that $\sigma \in \Sigma$ and $\tau \subseteq \sigma$ implies $\tau \in \Sigma$.

Looking to our case, if V is the total set of items I and Σ the set of all *frequent* itemsets, we would have an abstract simplicial complex because any subset of a frequent itemset is frequent. Algebraic topology is centered on determining algebraic invariants of a topological space using homology (difficult to define, easy to compute) and homotopy (easy to define if you have the background, nigh impossible to compute) groups[15] which take a long time to define so we refer the reader to any standard text on algebraic topology, such as Hatcher (2002). Persistent homology is a process that re-calculates homology groups as some factor that impacts the nature of the space changes over time. In our example, the minimum support used to determine whether itemsets are frequent could be varied from 0 to 1, and the homology groups re-calculated as it varied. This is another point where my investigation into this topic stalled, both because of the inherit difficulty of the mathematics, but also what the homology groups might tell us about the items and transactions and what actionable conclusions could be drawn.

What's really interesting is there is a topological space associated to an abstract simplicial complex that effectively puts a geometry on the space, lending physical "shape" to the space. Admittedly, the idea of there being shape to a collection of transactions is a bit odd, but I was (and still am) enamoured by the idea and what conclusions could stem from it. To get this space, we define a bijection $\varphi : V \to \{1, 2, ..., N\}$ as the subspace of $\mathbb{R}^N$ given by the union $\bigcup_{\sigma\in\Sigma} c(\sigma)$ where $c(\sigma)$ is the convex hull of the set $\{e_{\varphi(s)}\}_{s\in\sigma}$, where e_i denote the $i$ht standard basis vector. In English, what this is describing is basically an N-dimensional analog of triangulation. Typically when calculating homology groups we use a convex hull of the set with an orientation (direction between the points). We could use J_δ to determine this orientation, say if $J_\delta(x_1, x_2) < 0.5$ then the orientation is "negative" and "positive" otherwise. I did start down this path (on a flight to North Carolina to retrieve a stolen vehicle[16]), but did not get very far as it is surprisingly difficult to represent a 35-dimensional space on paper.

References

Bump, R., Harwood, B., Noel, X., and Schaeffer, K. (2019). Ist 687. Github Repository.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):225–308.

[15]A group G is a set together with an operation (call it *) that satisfies four properties: if $g_1, g_2 \in G$, then $g_1 * g_2 \in G$; there exists an identity element e such that $g * e = e * g = g$ for all $g \in G$; each element has an inverse, i.e. there exists g^{-1} such that $g * g^{-1} = g^{-1}g = e$; the operation * is associative.

[16]long story

Devito, S., Harwood, B., and Webster, C. (2020). Ist 736. Github Repository.

Entertainment, B. (2004). World of warcraft [computer video game]. Blizzard Entertainment.

Gilbert, G. (1972). Distance between sets. *Nature*, 239(174).

Harwood, B. (2002). Quasi p or Not Quasi p? That is the Question. *Rose-Hulman Undergraduate Mathematics Journal*, Vol. 3(Iss. 2).

Harwood, B. (2019a). Ist 659. Github Repository.

Harwood, B. (2019b). Mbc 638. Github Repository.

Harwood, B. (2020a). Ist 652. Github Repository.

Harwood, B. (2020b). Ist 719. Github Repository.

Harwood, B. (2020c). Ist 772. Github Repository.

Harwood, B. and Vogel, K. (2020). Ist 664. Github Repository.

Harwood, B., Vogel, K., Wass, K., and Williams, N. (2020). Ist 707. Github Repository.

Hatcher, A. (2002). *Algebraic Topology*. Cambridge University Press, New York.

Howard, R. (Director). (2001). *A beautiful mind* [film]. Universal Pictures.

Kosub, S. (2016). A note on the triangle inequality for the jaccard distance. *Pattern Recognition Letters*, 120:36–38.

Munkres, J. (2000). *Toplogy*. Prentice Hall, New Jersey.

O'Neill, C. and Schutt, R. (2014). *Doing Data Science*. O'Reilly Media, Inc, Sebastapol, CA.

Stanton, J. M. (2017). *Reasoning with Data*. Guilford Press, New York.