Project Portfolio Milestone

Brian P. Hogan Jr.

SUID: [redacted]

Master of Science in Applied Data Science

Syracuse University

December 6th, 2020

Author Note

This paper prepared for Panel of Faculty Experts Syracuse University iSchool

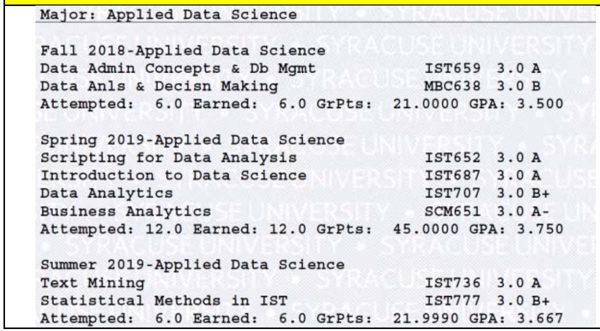
Contents

Introduction

Thank you committee for reviewing my Project Portfolio Milestone cataloging and synthesizing analysis work across courses taken in pursuit of my masters in Applied Data Science. Referring to Table 1, an overall strategy of gaining as much exposure to programming classes was pursued to help expand knowledge across numerous data science disciplines, such as text mining or Natural Language Processing (NLP), as available. This focus was also supported by program committee acceptance of a petition to substitute Dr. Stanton's statistics course for an Analytics Application core course.

All project materials, code, data, and reports are available at the following github portfolio. Upon review some sections will be password protected to ensure data and techniques are not available for Github data mining. The link is: bbe2/Portfolio: graduate portfolio (github.com)

Course Load with Focus on Expanding Programming Knowledge

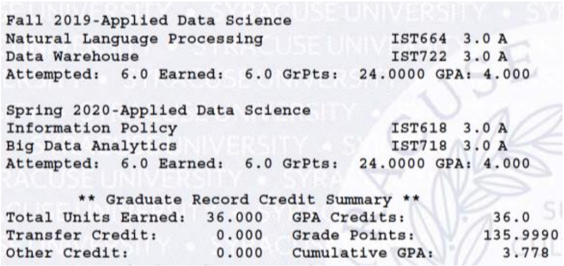


```
Major: Applied Data Science

Fall 2018-Applied Data Science
Data Admin Concepts & Db Mgmt           IST659  3.0 A
Data Anls & Decisn Making               MBC638  3.0 B
Attempted:  6.0 Earned:  6.0 GrPts:  21.0000 GPA: 3.500

Spring 2019-Applied Data Science
Scripting for Data Analysis             IST652  3.0 A
Introduction to Data Science            IST687  3.0 A
Data Analytics                          IST707  3.0 B+
Business Analytics                      SCM651  3.0 A-
Attempted: 12.0 Earned: 12.0 GrPts:  45.0000 GPA: 3.750

Summer 2019-Applied Data Science
Text Mining                             IST736  3.0 A
Statistical Methods in IST              IST777  3.0 B+
Attempted:  6.0 Earned:  6.0 GrPts:  21.9990 GPA: 3.667
```



```
Fall 2019-Applied Data Science
Natural Language Processing                 IST664  3.0 A
Data Warehouse                              IST722  3.0 A
Attempted:  6.0 Earned:  6.0 GrPts:  24.0000 GPA: 4.000

Spring 2020-Applied Data Science
Information Policy                          IST618  3.0 A
Big Data Analytics                          IST718  3.0 A
Attempted:  6.0 Earned:  6.0 GrPts:  24.0000 GPA: 4.000

      ** Graduate Record Credit Summary **
Total Units Earned:  36.000   GPA Credits:          36.0
Transfer Credit:      0.000   Grade Points:     135.9990
Other Credit:         0.000   Cumulative GPA:      3.778
```

Table 1

A. Background

After spending a 15-year career as a business process engineer using discrete-event modeling and simulation I had come to learn of numerous gaps in my science background as the field of data science matured. My goal in pursuing a Master's at Syracuse was to have access to faculty, seasoned approaches to building competence in data science, and building a deep understanding of what machine learning was. In preparation for the program I took a certificate course at Johns Hopkins and quickly realized the complexity I was undertaking and how my skills in discrete-event modeling were obscure in a decaying programming language. Courses in graduate statistics had carried me during this period relying upon descriptive analysis, generic data fitting, and regression for informing models and prediction outcomes.

After a couple of years and study and the Covid-19 pandemic interrupting employment as a data scientist, I can point to a new welcomed data science landscape opened from my Syracuse University experience. Many professors, to name a few, such as Dr. Ami Gates and Dr. John Santerre, would take the time to connect classroom learnings to practical use cases. Dr. Nancy McCracken built an amazing NLP landscape essentially enabling analysis across book corpus I didn't even know possible but had certainly pondered. Dr. Lin expanded R programming knowledge with direct machine learning cases in a clear and concise manner. Dr Ryan made statistics exciting and useful again in everyday life. And Dr. Stanton's work in Bayesian statistics opened an entire new world based on posterior probabilities. With this growing knowledge I was connected to active online data science research communities, such as Microsoft's Dr. John Langford (hunch.net) and the International Conference on Machine Learning (icml.cc) to help understand the difference between the work a of data science "research scientist" performs versus an autonomous vehicle signal data scientist versus use of NLP at hospital on electronic health records (EHR).

What has resulted is an ability to direct my data science field of interest, understand required skill gaps, and know "where to go" to build the scientific and programming knowledge necessary to be competent. For example, in order to prepare for an employment opportunity in a medical data science laboratory, such as Prof. Gil Alterovitz (harvard.edu), it is first necessary to synthesis how ML algorithms are being used and in what data universe. For example, it was necessary for me to apply my data science learning applying to Alterovitz's lab focusing on predicting disease prevalence in breast cancer cohorts. Without the tools learned in the Syracuse program I would not have been able to interview and be competent in topics such as "clinical phenomics is the measurement of diversity of disease states across human subjects. The massive accumulation of clinical data accrued automatically inside electronic medical records with each episode of patient care through clinical, laboratory, and billing systems has enabled a new type of phenomic research using clinical data. When such pheno-type data are extracted, these large data sets, called phenomes, can provide useful snapshots of disease prevalence, distribution, and correction" (Warner, Denny, Kreda, Alterovitz, 2014, pg 324).

While I was not selected for that Alterovitz position my determination only increased. Quickly I rebounded and prepared for rather intense interviews at Harvard University Business School in organization behavior data science. This work challenged my nascent data science underpinnings to understand AI-notebooks related to the field of people analytics (PA). According to Gal, Jensen, and Stein (2018), the increased use of algorithmic management should allow employer decision-makers to use evidence based, bias free, objective decisions facilitating growth of employee talent. However, great ethical challenges are associated with systemic person metrics resulting from fallible companion technology, such as internet tracking tools and data mining of emails. Organization behavior is actively using machine learning to "towards and ontology of people analytics" and development of more robust human resources practices deciphering employee commitment and satisfaction requires as much organization people data as possible to profile "real-time assessments" versus traditional post-hoc survey methods (Gelbard, Ramon-Gonen, Carmeli, Bittmann, and Talyanksy (2018). What is key is with my university training I was competent speaking to data science data of acquiring, cleaning, transforming, exploring, and reporting on human resource data depositories. I could answer why unsupervised algorithms were better for text mining email data with latent derelict algorithms (LDA) and association rules may lead to predicting what employees are more likely to be absent from work based on various data set features.

The following projects and their descriptions attempt to detail paths pursued to build competence in data science techniques. Each class spent programming in **R** or Python built a new appreciation for programming using notebook approaches or even integrated development environments (IDE) with translators for multiple languages in a single notebook, such as available at Amazon Web Services (AWS). More importantly though is Syracuse's approach to ensuring ML algorithm math is covered developing an appreciation for how calculations are being performed in an engine.

In the journey for becoming an employed what I can stress is the personal responsibility necessary to quickly talk about methods, go and test pulling and cleaning data, and essentially not having a portfolio based on ***collegiate tinkering***. It is also necessary to maintain weekly if not daily ties with active ML communities such as participating in Alexa "skill" learning sessions on "Twitch" and engaging with seminars such as AWS: ReInvent 2020.

While my personal efforts are spent weekly becoming a "production" data scientist, deep business analyst, or researcher I hope the following details the learnings outcomes sufficient to illustrate competence necessary to be awarded the title of "Applied Data Scientist."

I. Project 1 – LA Parking Tickets with Machine Learning Algorithms in R

A. Description

A largescale analysis across 500,000 Los Angeles tickets was performed assessing the impact of street cleaning and using random forest machine learning algorithms to predict city revenue. Other approaches included street ticket descriptive, k-means density, predicting total revenue with random forest algorithms, and using ggplot2 to graph geocode density and abundance.

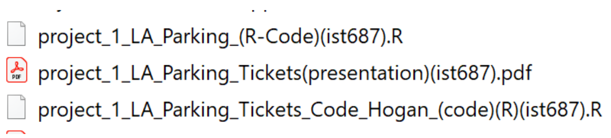
Files	Notes
 project_1_LA_Parking_(R-Code)(ist687).R project_1_LA_Parking_Tickets(presentation)(ist687).pdf project_1_LA_Parking_Tickets_Code_Hogan_(code)(R)(ist687).R	***Program***: **R** **Class:** IST687 Applied Data Science ***Data Science Methods:*** descriptive statistics, time series, ggplot geocode density graphs, association rules, randomForest, k-means clustering of highest offending ticket neighborhoods

Table 2

B. Methods, Experiments, and Results

Curiously, this effort helped to understand the wide availability of readily accessible data on the web, such as city parking tickets, and the ability to grab and munge data. The work Hadley Wickham in his book "Advanced R" helped with portioning of data *list* types and subsequent use of sqldf **R** library to query and sort data by car type and similar (Hadley, 2017). The original 2018 data set included 19 million records with 19 features but after data cleaning the resulting data set was 600,000 records and 17 features. Team peer Bray Coy was quite and expert at R developing cleaning routines beyond what was taught in class. This was an important journey learning as teamwork contributed invaluably to building skills through studying peer's code and coding approaches.

Referring to Figure 1, Data munging for this work was particularly challenging as conversions for latitude and longitude were necessary. A number of car types required further consolidation and accounting for bad data, such as no ticket violations, were addressed. The effort also highlighted the importance of identifying "relevant" data and ensuring a data set did not contain irrelevant information such as zero values.

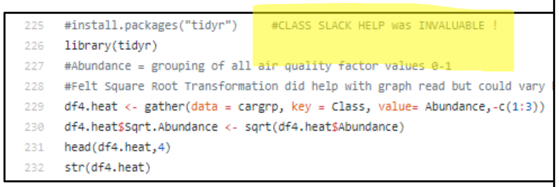
Munging…
1) Drop X column and check for duplicate ticket.number values
2) Break issue.date into months, weekday, keep issue date 'as is'
3) Convert issue.time into an actual timestamp, and bin into parts of day (morning, early afternoon, evening, etc.)
4) Clean up null, blank and NA values in all columns
-> insert 0 when necessary
5) Break up plate.expiry.date into month/year
-> flag for expired plates

6) Drop VIN column because all values are NA
7) Add a flag for import/domestic vehicle makes (checking for data quality issues along the way)
8) Simplify car color levels – 40 original colors
9) Clean up violation.code so formatted same ("80.69BS")
10) Run the lat and lon conversion

Figure 1

C. Learnings

Referring to Figure 2, Dr. Santerre encouraged use of a Slack channel to share approaches. Given I had no experience with R, and was only starting to learn dictionaries, teamwork shared the "tidyr" library contributing to learning the use of ***abundance*** gradient scale to represent a feature's density. This connection expanded on class exercises while facilitating the importance of data scale normalization with logarithmic or square root transformations. For my project contribution, such data analysis led to the finding in Los Angeles, California, there is in fact some benefit in owning a super luxury car, such as a Bentley, because when parked with a violation there is less likelihood of the city staff progressing to towing the automobile.



```
#install.packages("tidyr")     #CLASS SLACK HELP was INVALUABLE !
library(tidyr)
#Abundance = grouping of all air quality factor values 0-1
#Felt Square Root Transformation did help with graph read but could vary
df4.heat <- gather(data = cargrp, key = Class, value= Abundance,-c(1:3))
df4.heat$Sqrt.Abundance <- sqrt(df4.heat$Abundance)
head(df4.heat,4)
str(df4.heat)
```

Technique Summary ------>
i) sqldf to 62 car types
ii) Group cars into 6 bins
iii) Normalize "each" bin
iv) library(tidyr)\abundance

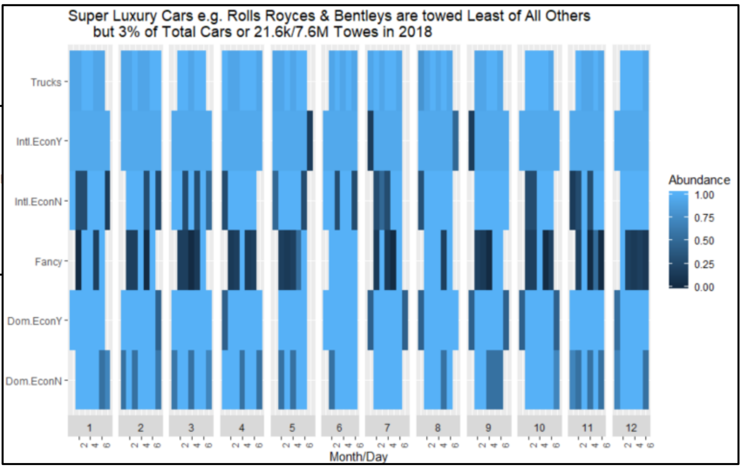


Figure 2

In terms of recommendations, random forest algorithm was helpful predicting citation revenue with only 4% on average being able actual data residual results. If we were able to alert drivers on when and where to park, parking between 9-12 was not advisable with most tickets occurring on West 5th Street, Hawthorn Ave in Hollywood, and Etiwanda Ave in Van Nuys.

II. Project 2 – Assessing NY Bridge Traffic with Twitter Text Mining in Python

A. Description

Referring to Table 3, assessment of New York State bridge data focusing on bridge condition, repair status, and traffic with respect to geographic locations. Python is used to assemble and preprocess data producing data summaries, list, and other structures. Text mining of Twitter feeds as performed to generate traffic word chatter to see how it may relate to heavy inbound/outbound traffic congestion days. Python QGIS was used for spatial analysis and text mining of Tweets to understand traveler sentiment.

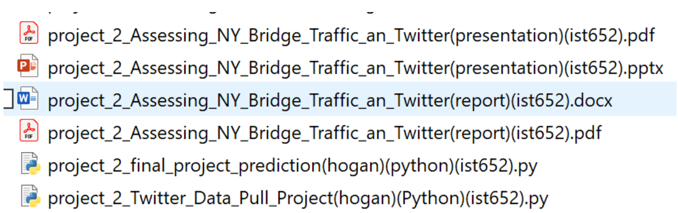
Files	Notes
 project_2_Assessing_NY_Bridge_Traffic_an_Twitter(presentation)(ist652).pdf project_2_Assessing_NY_Bridge_Traffic_an_Twitter(presentation)(ist652).pptx project_2_Assessing_NY_Bridge_Traffic_an_Twitter(report)(ist652).docx project_2_Assessing_NY_Bridge_Traffic_an_Twitter(report)(ist652).pdf project_2_final_project_prediction(hogan)(python)(ist652).py project_2_Twitter_Data_Pull_Project(hogan)(Python)(ist652).py	***Program***: Python **Class:** IST652 Scripting for Data Analysis ***Data Science Methods***: geocode density, text mining with stemming, lemmatization, and stop words of Tweets, linear regression predicting "good, bad, and ugly" condition states.

Table 3

B. Methods, Experiments, and Results

New York bridge statistics data is readily available from www.dot.ny.gov capturing engineering and repair statistics. Use of linear regression helped determine key variable associations including condition rating, inspection date, surrounding population, and curiously length of bridge in feet. My project partner, Katie Poole, used Python QGIS in visualizing good, bad, and ugly visualizations and I focused on Twitter sentiment in an attempt to assess commuter sentient. Overall, 264 poor bridge condition ratings were visualized by high, medium, and low traffic density in an effort to validate to Department of Transportation management bridge repair projects are properly aligned with traffic flow density.

My contribution centered on Tweet analysis. It was highly gratifying to create a Twitter developer account actively downloading tweets full of various characters, emojis, and non-value text detracting from sentiment by day by bridge area generation. Instead of identifying sentiment it led to analysis of good, bad, and ugly text information. Different data reduction processes, such as using Regex, were deployed with finally identifying key words, such as *blocked,* as a poor-person means to assess congestion sentiment.

C. Learnings

Interestingly, bridge area population was not a determiner of bridge conditions and specific Tweet handler, that is those who tweet "a lot", does not necessitate viable information, but rather over-Tweeting facilitates "disinformation" requiring data reduction in a production sentiment system. Incorporation of Regular Expressions into my data science toolkit will have long lasting value ultimately leading to valuable words, such as accident, slow, or service, necessary to assess Twitter traffic sentiment conditions.

III. Project 3 – Understanding Workplace Mental Health with Machine Learning in R

A. Description

Referring to Table 4, focused on applying data mining techniques to help understand a Kaggle data set on mental health focused questions by employees in information technology work environments. The team used several algorithms to assess respondent perceptions and predict the likelihood a respondent would or would not reveal a mental health issue with a potential employer during an interview. Modeling techniques included: association rule mining, clustering, naive bayes, k-means, support vector machine, and random forest.

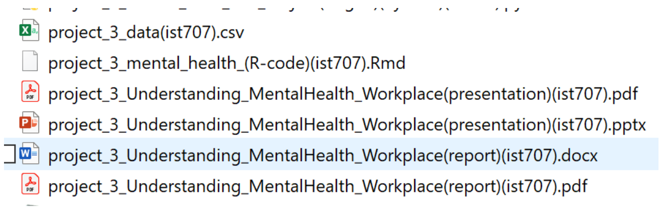
Files	Notes
 project_3_data(ist707).csv project_3_mental_health_(R-code)(ist707).Rmd project_3_Understanding_MentalHealth_Workplace(presentation)(ist707).pdf project_3_Understanding_MentalHealth_Workplace(presentation)(ist707).pptx project_3_Understanding_MentalHealth_Workplace(report)(ist707).docx project_3_Understanding_MentalHealth_Workplace(report)(ist707).pdf	***Program***: R **Class: IST707** ***Data Science Methods***: descriptive statistics, random forest inference, association rules, and decision trees find only 15% of respondents willing to reveal any mental health issues to their boss.

Table 4

B. Methods, Experiments, and Results

The Random Forest algorithm has an aptitude of rapidly generating decision trees finding and rating patterns with output score ranking in the R program such as the GINI index. Random Forest set the technical direction as it revealed *mental health consequence* feature as most significant across 56 categories. Technical challenges existed with the data set requiring careful curation of NAs as the data used a combination of blank values, -1 values, and even 99 encoding. This speaks to the importance of engaging with descriptive and correlation statistics carefully to address such values if not identified in a data sets associated encoding book.

Curiously, a combination of both supervised and unsupervised learning helped present the data set issue of mental health (MH) stigma. Decision trees, a supervised approach, found MH "stigma" effect is revealed as respondents feel "maybe" there would be negative consequence if disclosed. A general atmosphere created by coworkers, i.e. "feeling comfortable discussing," seems to play a big decision for individuals not self-employed.

The use of unsupervised association rules (AR) discovered respondents are going to keep mental health issues to themselves if they do not believe their anonymity is protected. AR mining indicates not having wellness programs, believing MH disclosure would hurt their career, and keeping MH disclosure from coworkers validates a lack of MH communication will be present with a direct supervisor.

C. Learnings

Similar to my deep dive in people analytics, human resource functions are keen to track as much "opinion" information associated with a work position. Any simple periodic inventory of "day in the life" office questions could help build a company "pulse" profile providing a data depository to perhaps find those conditions leading to increased employee retention and satisfaction.

Some analysis in this project was not linear and required more "elastic thinking." During the class, Professor Ying Lin recommended Richard Feynman's "Cargo Cult Science" to help uncover clues that are pertinent versus thinking you already know what is pertinent (1974). While the stigma associated with any mental health conditions may seem obvious, 41% of respondents did not think it would affect their career one way of the other. However, upon closer look only 15% expressed a definite willingness to bring up this issue in an interview and such knowledge kernels were made possible with the assistance of decision trees.

IV. Project 4 – Understanding Shakespeare with Text Mining in Python

The entirety of class IST736 Text Mining experience altered focused my future data science interests. For many years I knew sentiment analysis was expanding but was not privy to its methods nor could decipher through internet research. Participation and digestion of the material in IST736 really speaks to the high quality and standards Syracuse University requires of its professors, educators, planners, video and IT staff, and any other categories not mentioned. IST736 helped build a new foundation in both data science text mining methods and appreciation for the power of machine learning to perhaps one day result in the auto generation of new books programmatically.

In particular, the foundational work of Socher, Perelygin, Wu, Chuang, Manning, Ng, and Potts in "Recursive Deep Models for Semantic Compositionality" for predicting sentiment classification labels based on a Sentiment Treebank paving the way to even more elaborate treebanks available today (2013). The work introduced me to sentence parsing, recursive neural networks, and building a gateway to building knowledge of NLP negation, amongst other, techniques. I was able to expand on the concept when assessing restaurant sentiment by building qualitative categories of dining out including moods, taste as social distinction, and deliciousness by able to categorize written review expressions (Ariyasriwatana, Quiroga, 2016).

Other research facets contributing to significant knowledge growth of Text Mining including probabilistic topic modeling (Chang, Boyd-Graber, Gerrish, Wang, Blei, 2009), Amazon Mechanical Turk workforce for label agreement and potential misuse of online crowdsourcing (Kan, Drummey, 2018), seminal movie review sentiment classification with naïve bayes and support vector machines (Pang, Lee, Vaithyanathan, 2002), the significance of stop words and their potential misuse of genre classification (Yu, 2008), and finally the ease of construction a corpus of 4% of the books ever published facilitating grammatical trend analysis and study of human culture (Michel, et. Al, 2010).

Text mining, under the instruction of Dr. Ami Gates, also dramatically improved machine learning methods established in IST707 with Dr. Lin. This included building numerous train, test data splits for validating accuracy, repeatably using confusion matrix and accuracy scores for evaluating outcomes. Becoming adept with outcome measures including precision, recall, F1 scores, and Cohen's kappa. Ensuring k-means clusters are reasonably configured with the right number of clusters. Carefully selecting distance measures to normalize document term or term document matrices with cosine, Euclidean, and Manhattan distance measure approaches. And perhaps most importantly during the data cleansing step addressing the removal of white space, commas, annotations, and other anaphora potentially skewing counts.

A. Description

Referring to Table 5, text mining analysis provides an array of capabilities to assess language, sentiment, and use of natural language processing techniques to distinguish bodies of work. The entire Shakespearean corpus is assessed in terms of vocabulary to discern differences amongst characters and tragedy, comedy, and sonnet types. Analytical techniques including multinomial naive bayes, support vector machines, k-means and ensemble methods, such as random forest, are used with labeled data to predict play and character types in a 70, 30 train-test model split environment.

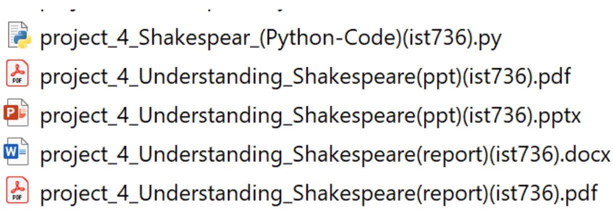
Files	Notes
 project_4_Shakespear_(Python-Code)(ist736).py project_4_Understanding_Shakespeare(ppt)(ist736).pdf project_4_Understanding_Shakespeare(ppt)(ist736).pptx project_4_Understanding_Shakespeare(report)(ist736).docx project_4_Understanding_Shakespeare(report)(ist736).pdf	***Program***: Python **Class: IST736 Text Mining** ***Data Science Methods***: corpus statistics, k-means clustering, stop word discrimination, support vector machines, cosine distance normalization, Naïve Bayes play character distinction, decision trees, and accuracy scores correctly labeling Shakespeare works by play type comedy, history, tragedy.

Table 5

B. Methods, Experiments, and Results

A four-person team contributed to analyzing 39 plays and over 835,997 words. Unlike other project efforts, each person worked independently to build their own corpus from the MIT core Shakespeare (The Complete Works of William Shakespeare (mit.edu)). It was far from an insignificant challenge to download and create the corpus supporting differentiation of characters ranging from 17 to 70, 3 play types, and scenes ranging from 9 to 40.

While team members used different machine learning approaches and angles to classify and predict characters, plays, or scenes, my deep dive into the value of stopwords was driven by the work of Dr. Bei Yu (2008) who found sentiment classification with literature texts, such as early American novel writer Emily Dickenson or mid-century playwright Willian Shakespeare, should pay special attention to English language grammar structures as “a common word in one collection might not be common in another one” (p. 330). Overall, the Natural Language Toolkit’s (NLTK) default stopword settings limited both k-means and support vector machines to classify comedy, tragedy, and history plays correctly. This effect held true across creation of different train-test corpus mixes.

C. Learnings

When using text mining, and natural language processing techniques as learning in IST664, it is imperative to thoughtfully assess their potential impact on data classification strategies. A simple lemmatization exercise for the word “breed” can have very different meanings and negatively impact feature weights with simple examples such as the word “breed” when also known as breeding, breeds, and breeder. Such caution speaks to methodically reading existing research on the corpus type being investigated, thoroughly becoming familiar with lexical and semantic constructs, and testing multiple machine learning algorithms before even beginning the process of forming a recommendation from observations. Referring to Figure 3, corpus rendering after significant cleansing still resulted in 5 classifications speaking to need for more complex hyperplanes and distance measures to discern comedy, tragedy, and history plays.

K-means Clustering of Shakespeare’s Works

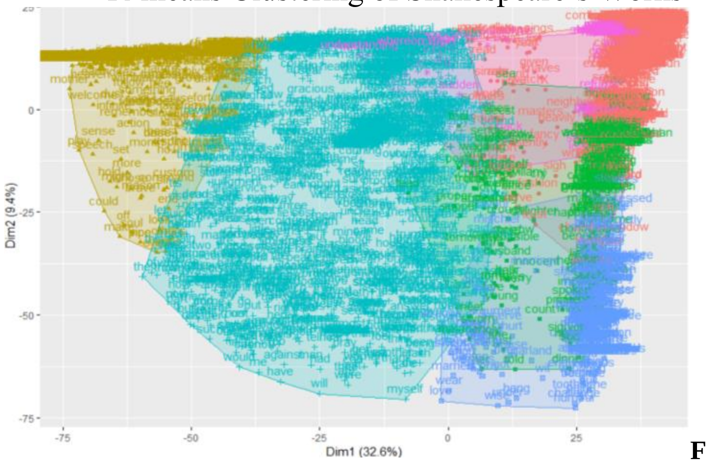


Figure 3

V. Project 5 – Deep Investigations into Convolutional Neural Networks, Facial Recognition, and other Applications in Natural Language Processing

A. Description

Referring to Table 6, facial and image recognition is the wave of the future. China has at least two companies who are image labeling factory building a warehouse of image features facilitating some of the most sophisticated image recognition capability the world has ever seen. Performed a deep investigation into the mechanics of convolutional neural networks (CNN) to understand how machine learning is accelerating artificial intelligence across pictorial landscapes.

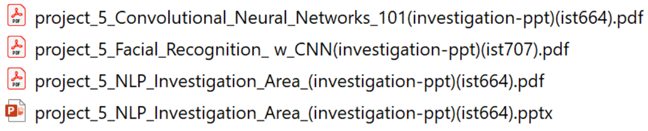
Files	Notes
 project_5_Convolutional_Neural_Networks_101(investigation-ppt)(ist664).pdf project_5_Facial_Recognition_ w_CNN(investigation-ppt)(ist707).pdf project_5_NLP_Investigation_Area_(investigation-ppt)(ist664).pdf project_5_NLP_Investigation_Area_(investigation-ppt)(ist664).pptx	***Program***: R & Python **Class: IST707 & IST664** ***Data Science Methods***: convolutional neural networks for image recognition and use of CNN in natural language understanding

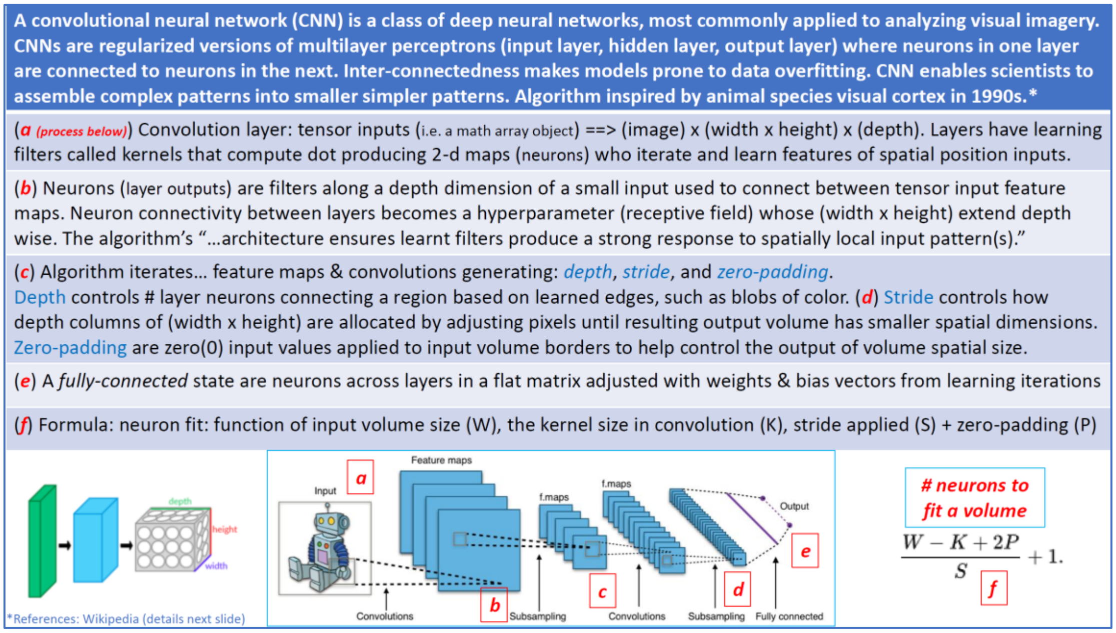
Table 6

B. Methods, Experiments, and Results

Both IST707 and IST664 offered opportunities for building a deeper appreciation of technology, algorithmic processing, and advances in algorithm techniques such as pattern recognition. Such classwork helped broaden my appreciation of algorithm complexity and the “below the hood” programming required to operationalize an algorithm. This experience yielded the appreciation of how to contribute to the scientific community with algorithmic feature enhancements if such an opportunity and data is able to present itself.

Referring to Figure 4, is classwork IST707 analysis into the breakdown of generic CNN algorithmic processing excluding the specific math vectors built to support tabular execution. While the algorithm speaks to depth, stride, and zero-padding the real work is in arrays created to tabulate computations based on dynamic neurons calculated across and image or video segment.

Convolutional Neural Network Basics



A convolutional neural network (CNN) is a class of deep neural networks, most commonly applied to analyzing visual imagery. CNNs are regularized versions of multilayer perceptrons (input layer, hidden layer, output layer) where neurons in one layer are connected to neurons in the next. Inter-connectedness makes models prone to data overfitting. CNN enables scientists to assemble complex patterns into smaller simpler patterns. Algorithm inspired by animal species visual cortex in 1990s.*

(*a* *(process below)*) Convolution layer: tensor inputs (i.e. a math array object) ==> (image) x (width x height) x (depth). Layers have learning filters called kernels that compute dot producing 2-d maps (neurons) who iterate and learn features of spatial position inputs.

(*b*) Neurons (layer outputs) are filters along a depth dimension of a small input used to connect between tensor input feature maps. Neuron connectivity between layers becomes a hyperparameter (receptive field) whose (width x height) extend depth wise. The algorithm’s “…architecture ensures learnt filters produce a strong response to spatially local input pattern(s).”

(*c*) Algorithm iterates… feature maps & convolutions generating: *depth*, *stride*, and *zero-padding*.
Depth controls # layer neurons connecting a region based on learned edges, such as blobs of color. (*d*) Stride controls how depth columns of (width x height) are allocated by adjusting pixels until resulting output volume has smaller spatial dimensions.
Zero-padding are zero(0) input values applied to input volume borders to help control the output of volume spatial size.

(*e*) A *fully-connected* state are neurons across layers in a flat matrix adjusted with weights & bias vectors from learning iterations

(*f*) Formula: neuron fit: function of input volume size (W), the kernel size in convolution (K), stride applied (S) + zero-padding (P)

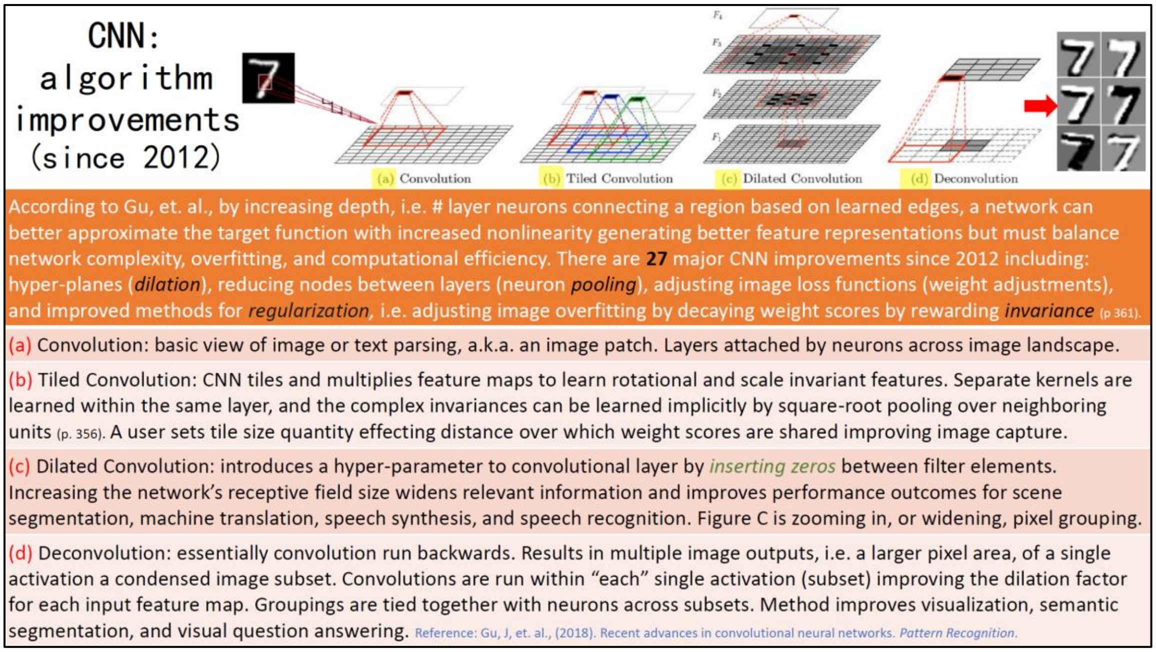
# neurons to fit a volume

$$\frac{W - K + 2P}{S} + 1.$$

*References: Wikipedia (details next slide)

Figure 4

Referring to Figure 5, speaks to work performed in IST664 expanding on CNN algorithmic enhancements where smaller scale tile areas are partitioned and then a forward and backward process ensues, called deconvolution, dramatically enhancing layer neurons and improving performance for speech or image recognition. Such expansions have dramatically altered the current state of artificial intelligence making Amazon Web Service tools such as Alexa and dynamic video capturing available in the game changing technology Reckognition.



CNN: algorithm improvements (since 2012)

According to Gu, et. al., by increasing depth, i.e. # layer neurons connecting a region based on learned edges, a network can better approximate the target function with increased nonlinearity generating better feature representations but must balance network complexity, overfitting, and computational efficiency. There are **27** major CNN improvements since 2012 including: hyper-planes (*dilation*), reducing nodes between layers (neuron *pooling*), adjusting image loss functions (weight adjustments), and improved methods for *regularization*, i.e. adjusting image overfitting by decaying weight scores by rewarding *invariance* (p 361).

(a) Convolution: basic view of image or text parsing, a.k.a. an image patch. Layers attached by neurons across image landscape.

(b) Tiled Convolution: CNN tiles and multiplies feature maps to learn rotational and scale invariant features. Separate kernels are learned within the same layer, and the complex invariances can be learned implicitly by square-root pooling over neighboring units (p. 356). A user sets tile size quantity effecting distance over which weight scores are shared improving image capture.

(c) Dilated Convolution: introduces a hyper-parameter to convolutional layer by *inserting zeros* between filter elements. Increasing the network's receptive field size widens relevant information and improves performance outcomes for scene segmentation, machine translation, speech synthesis, and speech recognition. Figure C is zooming in, or widening, pixel grouping.

(d) Deconvolution: essentially convolution run backwards. Results in multiple image outputs, i.e. a larger pixel area, of a single activation a condensed image subset. Convolutions are run within "each" single activation (subset) improving the dilation factor for each input feature map. Groupings are tied together with neurons across subsets. Method improves visualization, semantic segmentation, and visual question answering. Reference: Gu, J, et. al., (2018). Recent advances in convolutional neural networks. *Pattern Recognition*.

Figure 5

C. Learnings

The significance of CNN advancements must be appreciated as the "technology with the greatest potential to change policing is also the least visible to the public" (Economist, 2019). There is both widespread and vast disinformation being distributed on the use of facial recognition software resonating in comments such as "there is a feeling that everything you say and do is being monitored. It is terrifying." (Economist, 2019, p. 71).

Consider the following, when walking into a large box store such as Home Depot, an individual has already given up their pictorial rights. Home Depot resides on private property and video monitoring is simply a function of security for the facility and individual coming on the premise to help assess for adverse events. What is not known by the public, is if Home Depot has an Amazon Web Services contract it can log and count individuals coming onto the property and even associate their driver licenses through the purchase of public data from a state of residence. Sure enough such profiling is perhaps only for the benefit of customer cohort segmentation given resident area statistics but certainly speaks to the broader implication of the ethics of privacy, access, and security in data science endeavors.

And finally, referring to Figure 6 CNNs make speech to picture rendering feasible.

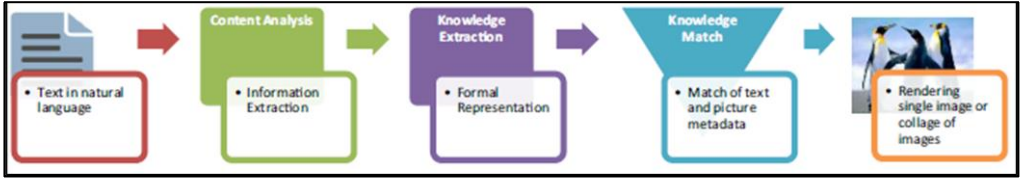


Figure 6

VI. Project 6 – Essays of Data Science Privacy and Access Ethical Dimensions

Having the opportunity and access to build a background in ethical theory was a strong component of my master's education in data science in course IST618 Information Policy with Professor Ian MacInnes. The works of Schultz helped understand the concepts of right, good, and just and how they are acted upon in end-based or duty-based activity pursuits (2006). Schultz further expanded on enabling values, theory of justice, and Rawls distinctive features of social contracts and building an appreciation for assessing fairness. According to Warschauer, a binary divide in the digital divide continues to widen in the digital age between have and have nots and how literacy will play a major role as the digital divide delineates those with reduced education achievement support (2002). The overall importance of content creation suggests an individual's information participation can lead to social inequity from being a content consumer versus being a content creator (Hargittai and Walkejo, 2008). Finally, building proficiency in digital human capital helps appreciate that information is complex and its necessary to perhaps mandate access to technology and computer literacy else contribute to a lack of policy assuring competency in technology that may play a significant component in an individual's life and livelihood (Bach, Shaffer, and Wolfson, 2013).

A. Description

Referring to Table 7, data science privacy and security surrounding Electronic Health Records (EHR) provide an array of challenges ensuring the safety of patient records while record updating and sharing is performed on essentially all aspects of these overly sensitive records. Complex security algorithms are continually challenged with the growth of machine learning algorithms put pressure on governments and the private sector to ensure client confidentiality. This type of reflection work broadens the data scientist perspective to have an eye towards policy, security, and an understanding of how to consider these factors and ethics in data science endeavors.

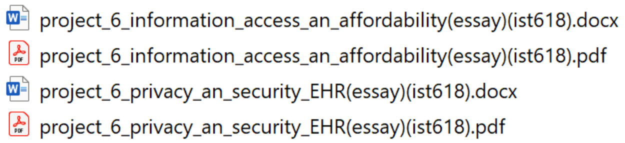
Files	Notes
 project_6_information_access_an_affordability(essay)(ist618).docx project_6_information_access_an_affordability(essay)(ist618).pdf project_6_privacy_an_security_EHR(essay)(ist618).docx project_6_privacy_an_security_EHR(essay)(ist618).pdf	***Program***: none – written analysis **Class: IST618 Information Policy** ***Data Science Methods***: understanding the ethics of data science privacy and implications of policy formation around its government encapsulation.

Table 7

B. Methods, Experiments, and Results

Course offerings contributed to performing several essay analyses. In particular focused on privacy and security of electronic health records and the image of technology access and affordability on marginalized youth populations. The following captures intent pursued in building competency of ethics in data science.

i. Marginalized Young People (MYP) have a range of challenges both domestically and internationally regarding access and navigation of health-care systems. Use of digital mental health (DMH) information combined with a communication technology system (ICT), is a solution to help underserved populations gain information channels and improve access to correct health service information (Schuller, Hunter, Figueroa, and Aguilera, 2019). MYP have differing cohorts including homeless, rural, race, and LGBTQ. Robards, Kang, Usherwood, and Sanci (2017) performed a systematic review of MYP across 1,796 articles and developed themes relating to an "…ability to recognize and understand health issues" including a professionals' knowledge, service environments, and ability to assess one's health.

ii. Privacy and security surrounding Electronic Health Records (EHR) provide an array of challenges ensuring the safety of patient records while record updating and sharing is performed on essentially all aspects of these highly sensitive records. Complex security algorithms are continually challenged with the growth of machine learning algorithms putting pressure on governments and the private sector to ensure client confidentiality. This requires providing a system that is increasingly cost effective, secure, and enabling of data mining efforts to help provide higher standards of care to all patients. There are clear advantages and disadvantages in gathering and managing health information but there are fundamental ethical dilemmas both the government and private sectors face. The following generates an understanding of this landscape and reviews policy adjustments to help broaden ERH's transparency between individuals, institutions and the government whose concomitant supports EHR accountability and best practices.

C. Learnings

This course work offered an opportunity to perform quality construction of both analytical and argumentative paper types. Platform building of ethical constructs will inform future situations where questionable outcomes from data science decisions will require thoughtful analysis.

VII. Project 7 – Natural Language Processing Custom Feature Algorithms for Tweet Classification

The coursework of IST664 was perhaps the most robust academically perhaps next to IST707 with Dr. Lin. Professor Nancy McCraken built and incredible course and I will be continuing to incorporate learning from coursework for the foreseeable future.

Numerous articles, reference materials, treebanks, parts of speech tools, and similar were provided to help build a thorough understanding the NLP universe. Some critical works building substantial knowledge substrate included "Annotation Guidelines for Twitter Part-of-Speech Tagging" thoroughly detailing nuisance manage in Tweet wrangling (Gimpel, Schneider, and O'Connor, 2013). Gimpel expanded knowledge in the Twitter arena with "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments contributing to algorithm advancements with negation (Gimpel, et. al., 2010). Noisy data is a fundamental component of most NLP undertakings given the generation of uncurated text across numerous applications and use of micro-blogs provide some means to facilitate proper tagging (Dercynski, Ritter, Clark, and

Bontcheva, 2013). None of work involved in NLP would even be possible without the work of Jurafsky and Martin in *Speech and Language Processing* (2019). Finally, learning how to figure out ways to assess sentiment on a national level helped really augment NLP thinking approaches with classwork reading Dodds and Danforth work on how to measure happiness based on Presidential songs and blogs (2017).

Dr. McCracken provided very relevant and updated material making the course highly transferable in conversations with prospective hiring managers.

A. Description

Referring to Table 8, this effort will build an understanding of natural language processing (NLP) features working with Twitter tweets. The data is from the Semantic Evaluation conference, a.k.a. SemEval, and has labeled data to assess message sentiment. Several NLP features, such as bigrams, parts of speech (POS), and negation, will help decipher and characterize tweet sentiment. ML train and test approaches with are explored.

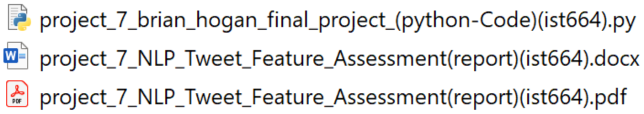
Files	Notes
 project_7_brian_hogan_final_project_(python-Code)(ist664).py project_7_NLP_Tweet_Feature_Assessment(report)(ist664).docx project_7_NLP_Tweet_Feature_Assessment(report)(ist664).pdf	***Program***: Python **Class:** IST664 Natural Language Processing ***Data Science Methods***: NLP feature construction with bigrams, parts of speech, and negation. Use of logistic regression was best suited for assessing classification outcomes.

Table 8

B. Methods, Experiments, and Results

Engineered an algorithm with parts of speech, bigrams, and negation words such as “no” or “not.” This resulted in the construction of feature sets which are used for running against Tweets to classify their positive, negative, and neutral sentiment. A gold-standard labeled data set was provided to facilitate feature engineering ultimately leading to the correct classification.

The effort was only able to achieve an accuracy of near 55% after extensive text cleaning with regular expressions was performed. This led to increasing the most common words from 2500 to 3000 to help improve classification. This was performed along with bigram chi-square value setting to 600. The chi-square value was set so high simply to ensure the most significant bigram scores were used in the data set given 18,018 words in the training and 8,298 words in the test data sets. This approach resulted in a reduction in accuracy instead of an improvement. Furthermore the “most informative features” was unchanged. This suggests other types of experiments incorporating techniques such as subjectivity, LIWC, or trigrams will be necessary to assist with the classification performance.

Programmatically this effort required nearly 1000 lines of code built with lessons throughout the semester and gathering engineering from other classes. The robustness of the code would not have been possible with taking IST736 Text Mining prior to this class.

C. Learnings

NLP is a vast topic requiring significant engineering, experience, and training. The coursework provided laid a solid foundation so when working with available production systems, such as Amazon Web Services Alexa, Polly, or its NLP Compute platform, I am far more confident and able to progress in the technology and interpret engineered Python notebooks.

VIII. Project 8 – Machine Learning with Big Data Understanding Music Choice

A. Description

Big Data analysis in IST718 with over a million music data records using machine learning algorithms, such as XGBoost, to predict what song a user is likely to listen to next. While Association Rules was not entirely suitable to this work it was the approach I took to look closer at what genre is more likely to follow another finding "country music" as being unique amongst 17 different music types.

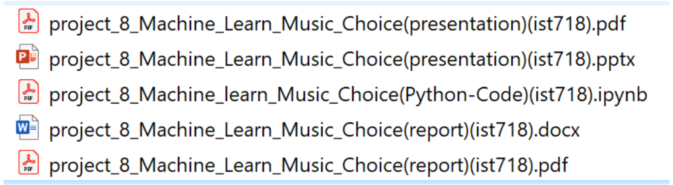
Files	Notes
 project_8_Machine_Learn_Music_Choice(presentation)(ist718).pdf project_8_Machine_Learn_Music_Choice(presentation)(ist718).pptx project_8_Machine_learn_Music_Choice(Python-Code)(ist718).ipynb project_8_Machine_Learn_Music_Choice(report)(ist718).docx project_8_Machine_Learn_Music_Choice(report)(ist718).pdf	***Program***: Python **Class: IST718 Big Data** ***Data Science Methods***:

Table 9

B. Methods, Experiments, and Results

The main challenge of this work was melding 1 million records across three data sets including music song detail, artist detail, member listening content. The data set extracted from Kaggle was not suited for machine learning exercise until data was associated across the three tables on a per record basis. The majority of this work was done in Pandas which required a significant learning effort. Once orchestrated, the running of both supervised and unsupervised models was straight-forward after cleaning NA values and some negative numbers.

C. Learnings

The key attribute that drives prediction accuracy is the length of a user's membership measured in days, which infers a larger amount of behavior and preference data for a particular user. Additional analysis showed that location, both country and city, and age of users impacted accuracy of users. Prediction error was primarily due to false positives where a song was recommended, but the user did not listen to more than once. The XGBoost algorithm was the most performant of explored models with a weighted precision and recall of 69%. This model demonstrated similar error predicting a user would like a song when only afforded a single listen. Additional analysis with association rules indicated some genres showed Country music had the highest predictability. Decision trees were least helpful.

VIII. Project 9 – Bayesian Statistics

The learning of Bayesian inference and its inclusion of posterior probabilities certainly put my frequentist training into a whole new perspective. Recent articles, such as *Moving to a World Beyond "p<0.05"* was very timely and certainly built an understanding of signal and noise data do not fit the mold of frequentist decision making (Wasserstein, Schirm, and Lazar, 2019). I had spent much of time on the side of *Objections to Bayesian statistics* but time spent performing the course work and exercises generating and interpreting Bayesian inference made all the difference to being comfortable analyzing outcomes per 100,000 observations and outcome expression regions for mean differences (Gelman, 2008).

A. Description

Referring to Table 10, coursework in Professor Stanton's *Reasoning with Data* made Bayesian analysis understandable in chapter 5 *Bayesian and Traditional Hypothesis Testing*.

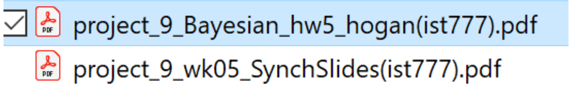
Files	Notes
 ☑ project_9_Bayesian_hw5_hogan(ist777).pdf project_9_wk05_SynchSlides(ist777).pdf	***Program***: R **Class: IST777 Reasoning with Data** ***Data Science Methods***: MCMC

Table 10

B. Methods, Experiments, and Results

Referring to Figure 7, the high-density interval (HDI) boundaries are -0.37 and 1.14 so there is a 95% probability the population mean difference between the control and treatment-1 groups is in this range. The greatest likelihood for a population mean difference is near 0.385 roughly between the region of 0 to 1. The expression: 14.5% < 0 < 85.5% means 14.5% of mean differences run in MCMC were negative while 85.5% were positive. This implies the chances the control group were equal to or better than the treatment group "are not" close to zero meaning 85.5% of the population means are different.

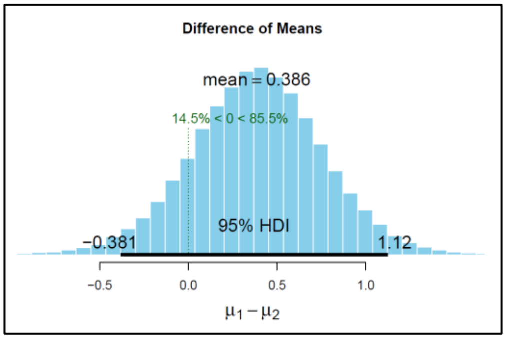


Figure 7

C. Learnings

In terms of synthesis, being able to express to managers the concept of a population mean, mu, and a model's mean difference is either not close to zero, i.e. not different, or away from zero meaning the means across 100,000 samples are statistically different.

Path Forward

It would not have been remotely possible to be actively reading work by the head of Microsoft's Machine Learning Research group, Dr. John Langford, without the tools and skills built in this program. Microsoft and other tech giants are on to tackling algorithms associated reinforcement learning and construction of high-quality decoders necessary to process rich sensory information generated in devices such as megapixel camera images (Misra, et. al., 2019).

Given the competitiveness of data science positions and reality that for some in the Boston there can be upwards of 30 Ph.D. applications alongside a suite of master's candidates it is necessary to build differentiation skills ***daily***. My current strategy in the Covid pandemic client is building competency and certification in Amazon Web Services machine learning platform SageMaker. While some ML enthusiasts may fashion this as being subservient to an over engineered system it provides a relevant and reachable means to employment. In my current position, I deploy my skills wherever possible in a lumber distribution warehouse to improve sales while at night studying my AWS skills platform.

In terms of the future I now plan to start applying for professor roles with the completion of my masters. This is a major life-long goal given my past two careers, experience, and investment in my Syracuse education. Appendix C details my Teaching Philosophy statement as I endeavor to gain employment helping fresh, young minds to develop programming and competency with data reasoning. Appendix B details my current resume and I am proud my education was also able to secure me part-time sporadic work as a scientific editor of data science manuscripts for peer-review publication. This is my very first step to contributing directly to a scientific body of knowledge and have edited three papers as of this portfolio submission.

Lastly in terms of life-long learning, influenced by my advancements in text mining and NLP I hope to build research into merging field of *disambiguation of disinformation*. I remain terribly dissatisfied as a citizen, trained scientist, and consumer of false internet noise of all forms and would like nothing more to find a means to better assess, rate, and distribute quality knowledge or at least a reliable information rating mechanism. I would very much like to have a publication in my name, and any associates, and feel equipped with the tools necessary to perform the proper science methods required.

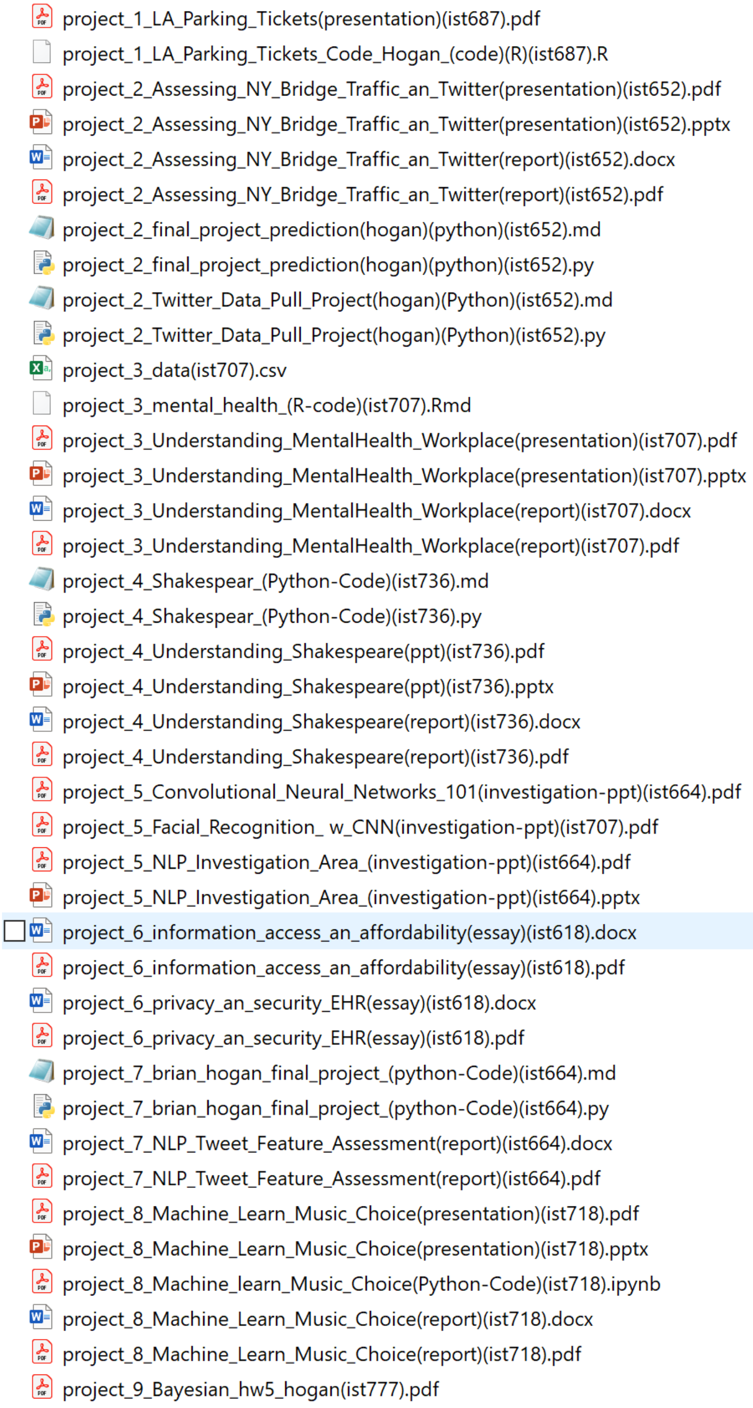
Thank you committee for this consideration of this portfolio milestone requirement and the opportunity to gain my science degree from your esteemed university.

Best Regards,
Brian Hogan

Appendix A

Portfolio File Summary

All files available on Github: bbe2/Portfolio: graduate portfolio (github.com)
Note: pdf files substitute source files for immediate depository viewing



project_1_LA_Parking_Tickets(presentation)(ist687).pdf
project_1_LA_Parking_Tickets_Code_Hogan_(code)(R)(ist687).R
project_2_Assessing_NY_Bridge_Traffic_an_Twitter(presentation)(ist652).pdf
project_2_Assessing_NY_Bridge_Traffic_an_Twitter(presentation)(ist652).pptx
project_2_Assessing_NY_Bridge_Traffic_an_Twitter(report)(ist652).docx
project_2_Assessing_NY_Bridge_Traffic_an_Twitter(report)(ist652).pdf
project_2_final_project_prediction(hogan)(python)(ist652).md
project_2_final_project_prediction(hogan)(python)(ist652).py
project_2_Twitter_Data_Pull_Project(hogan)(Python)(ist652).md
project_2_Twitter_Data_Pull_Project(hogan)(Python)(ist652).py
project_3_data(ist707).csv
project_3_mental_health_(R-code)(ist707).Rmd
project_3_Understanding_MentalHealth_Workplace(presentation)(ist707).pdf
project_3_Understanding_MentalHealth_Workplace(presentation)(ist707).pptx
project_3_Understanding_MentalHealth_Workplace(report)(ist707).docx
project_3_Understanding_MentalHealth_Workplace(report)(ist707).pdf
project_4_Shakespear_(Python-Code)(ist736).md
project_4_Shakespear_(Python-Code)(ist736).py
project_4_Understanding_Shakespeare(ppt)(ist736).pdf
project_4_Understanding_Shakespeare(ppt)(ist736).pptx
project_4_Understanding_Shakespeare(report)(ist736).docx
project_4_Understanding_Shakespeare(report)(ist736).pdf
project_5_Convolutional_Neural_Networks_101(investigation-ppt)(ist664).pdf
project_5_Facial_Recognition_ w_CNN(investigation-ppt)(ist707).pdf
project_5_NLP_Investigation_Area_(investigation-ppt)(ist664).pdf
project_5_NLP_Investigation_Area_(investigation-ppt)(ist664).pptx
□ project_6_information_access_an_affordability(essay)(ist618).docx
project_6_information_access_an_affordability(essay)(ist618).pdf
project_6_privacy_an_security_EHR(essay)(ist618).docx
project_6_privacy_an_security_EHR(essay)(ist618).pdf
project_7_brian_hogan_final_project_(python-Code)(ist664).md
project_7_brian_hogan_final_project_(python-Code)(ist664).py
project_7_NLP_Tweet_Feature_Assessment(report)(ist664).docx
project_7_NLP_Tweet_Feature_Assessment(report)(ist664).pdf
project_8_Machine_Learn_Music_Choice(presentation)(ist718).pdf
project_8_Machine_Learn_Music_Choice(presentation)(ist718).pptx
project_8_Machine_learn_Music_Choice(Python-Code)(ist718).ipynb
project_8_Machine_Learn_Music_Choice(report)(ist718).docx
project_8_Machine_Learn_Music_Choice(report)(ist718).pdf
project_9_Bayesian_hw5_hogan(ist777).pdf

Appendix B

Resume

Brian Hogan, MS^, BS bphogan@syracuse.edu, https://www.linkedin.com/in/bbe/, https://github.com/bbe2/
284 Cross Street, Winchester, MA 01890, 757-477-8241. ^*4Q20 pending committee review*

PROFILE

- Solid working knowledge in data analysis methods, machine learning, deep learning, and statistical accuracy.
- Deep experience advancing infrastructure through quantitative and qualitative process engineering.
- Quickly provide relevant business insights for operation managers and executives to make informed decisions.
- Building proficiency in AWS SageMaker with machine learning models running parallel to data pipelines.

PROFESSIONAL EXPERIENCE

Supply Chain Associate, *Jackson Lumber & Millwork,* Woburn, MA 2020 -

- Material management and product forecasting.

Scientific Editor, *Accdon, LLC,* https://www.accdon.com/, Waltham, MA 2020 -

- Edit and prepare scientific manuscripts for peer-review journal publication with data science & nursing focus.

Graduate Student, *Syracuse University,* Syracuse, NY 2018-12/2020

- Courses: Information Policy, Data Scripting, Natural Language Processing, Big Data Analytics, Text Mining, Business Analytics, Bayesian Statistics, Machine Learning, Big Data, Data Warehouse. GPA 3.8/4.0.
- Heavy programming concentrating in machining learning algorithms, text mining, scripting, and NLP.
- Regular projects focusing on data sourcing, cleaning, analyzing, experimenting, and evaluating accuracy.

Academic Research Advisor, *Tutor Matching Service, https://tutormatchingservice.com* 2017 -

- Develop student’s logic ability in research and build confidence in their data science skills.

Project Manager & Consultant, *ProModel,* Allentown, PA 2001 – 2016

- Led customers in strategy of operational improvements designing and integrating discrete-event data solutions.
- Focused on productivity, demand/capacity, resource forecasting, >$75M capital budgets, and supply-chain.
- Discovered, with another, a novel patient enrollment feature by drug modality (SAS, VBA, factor analysis).
- Implemented 20-year predictive model for NASA program mgmt. budget (regression, discrete event, SQL).

TECHNICAL SKILLS

- Bayesian Analysis, LDA, Machine Learning proficiency (supervised, unsupervised, reinforcement), Adaline, structured probabilistic, (un)structured text mining, custom NLP, systematic literature review.
- ggplot, mongoDB, SQL, plotly, Pandas, Python, R, SMSS, Tableau, TensorFlow, PyTorch, VisualStudio.
- Extensive experience building content, training courses, and collaborating with subject experts on focus.
- Rapid systematic literature review across technical and qualitative material coalesced into consumables.

EDUCATION

Master’s of Science: Applied Data Science, Syracuse University, Syracuse, NY, 3rd Quarter 2020
Certificate: Data Science, Johns Hopkins University, Baltimore, MD, 2016
Graduate Certificate in Business Administration, Harvard Extension School, Cambridge, MA
Bachelor of Science: Business Administration & Psychology, Babson College, Wellesley, MA

EXTRACURRICULAR

Autopsy Respect and Dignity (research), GED Coach, Mechanical Turking, Golden Key Intl. Honour Society

Appendix C

Brian Hogan, Master's in Applied Data Science
Teaching Philosophy – 2021

Prior to training as a scientist I worked as a process engineering professional using discrete event modeling and simulation technology to design, augment, and optimize pharmaceutical development, packaging, and resource requirements from medical doctors to floor technicians. My work as a symbolic analyst focused on training leaders and managers in statistics and supervised methods. Consulting challenging me to quickly assess a business owner's experience, training, and build manager confidence in deliverables. I found great satisfaction in making connections and expanded this skill by running training courses teaching scientists, such as NASA engineers, simulation programming. I thrived on a constant idea flow and overcoming challenges to ensure successful outcomes. I knew I wanted to teach in academia and choose data science to augment my consulting technical skills with rigorous academic training for work in post-secondary education. Curiously, of my professors, Dr. Ami Gates @ Georgetown, commented "the world needs programmers" and I knew teaching would be the best avenue for me to generate societal equity by helping thriving minds with programming techniques, statistical reasoning, and the art of writing convincingly.

Interestingly, before I started my master's in 2018, I was advising individuals and small groups of graduate students in academic research, writing, statistics, and programming. During my master's in sociology at Boston College I became very committed to academic writing and began advising peer students. This experience taught me how to be a creative advisor who is able to connect to others, find the right message, and provide encouragement to advance work.

An important component of an instructor's role is to foster student motivation, strategize challenging assignments, and help students learn to trust their abilities. When I tutor students in programming I begin with an overview of systems theory and how to abstract interrelated components defining a structure or operating conditions. I find when students learn to abstract by drawing a box around an area of interest they can readily assess system relationships inputs, outputs, resources, and constraints. Abstraction also helps a student learn a languages' API, input and output storage containers, and how to classify objects, such as Python's lists, dictionaries, and strings. My goal is help students simplify large abstract universes into manageable chunks to practice coding exercises such as reading, cleaning, looping, and engaging library resources to learn other tools or methods available.

Recognizing people work, study, and grok at different rates has led me to strategize exercises and anecdotes to help students see the benefits of their endeavors and fill their *idea hopper*. To ensure students feel supported, outside the classroom I am available to discuss challenges critical to learning. Sometimes a quick response or short discussion maintains motivation to solve challenges and discover new approaches. In difficult assignments, I stress the importance of documenting thinking approaches for questions not completed to show class commitment when *everything else required to be human* disrupts focus. Most times, I find students are motivated to catch up and benefit from course material as often work performed builds gateways to new occupations and pillars of self-worth.

I am a firm believer in having high quality, researched, and tested materials to advance course objectives. While considered a standard for any accredited program, I highlight to stress to academic search committees my commitment to a student's education experience with an eye towards ensuring retention rate reduction. If materials need improvement, additional reference materials, or helpful examples I ensure they are generated. I was fortunate to have many amazing professors in my education experience and I endeavor to match their lecture quality and delivery styles. Teaching builds various mental pillars encouraging students to grow and consider new means to approach challenges.

Course materials typically include: a syllabus, reading materials, and asynchronous video lecture with supporting Microsoft PowerPoint training slides. When materials are carefully staged, students build reasonable expectations of weekly hours required to advance studies. Another crucial material set is lecture handouts. Providing detailed materials in class helps students build confidence in lecture theory, algorithms, and examples to support program learning. For example, when I was tutoring on Federalist Paper classification for Hamilton, Jay, or Madison authorship I discovered several students struggling to build the corpus in the statistics program **R**. By providing a process flow-chart and code component examples one student was able to build matrices and very quickly compare outcomes from supervised to unsupervised. I am always surprised by how ingenious students can be with vectors, matrix', arrays, and data frames.

As a research editor, I am committed to supporting science's body of knowledge and regularly provide students with relevant articles to further interest and have them build an appreciation of the research community. It is invaluable for students to learn the availability of journals and value of librarian science.

Finally, I work hard to balance homework, labs, and group projects to ensure institution course standards, active student participation, and student satisfaction.

Every student of any group, race, and orientation has an opportunity to learn and be successful. From the noted Roman Marcus Aurelius, "nothing has such power to broaden the mind as the ability to investigate systematically and truly all that comes under thy observation in life." Computer science provides a platform to promote equity, build effective thinking, and the data economy facilitates students being able to actively incorporate scientific methods into their daily work and activities.

Thank you for your time and consideration.

Best,
~Brian Hogan

Qualified to instruct:

- Introduction to Statistics / Data Analysis & Decision Making
- Introduction to Data Science
    - Machine learning algorithm theory
    - Example and code proficiency in both R and Python
    - Supervised, unsupervised, and avenues to deep learning
- Artificial Intelligence

  - NLP, neural networks, machine learning, cognition, vision (CNN), pattern recognition
- Text Mining
- Introduction to Python
  - Data Scripting (pull, parse, clean); Twitter, Facebook API data pulls
- Building AWS proficiency in Sagemaker
- Effective writing & communication strategies

References

Ariyasriwatana, W., & Quiroga,L. A thousand ways to say 'Delicious!'—Categorizing expressions of deliciousness from restaurant reviews on the social network site Yelp, Appetite, Volume 104, (2016), p.18-32. Retrieved from: https://doi.org/10.1016/j.appet.2016.01.002.

Bach, A., Shaffer, G., & Wolfson, T. (2013). Digital human capital: Developing a framework for understanding the economic impact of digital exclusion in low-income communities. *Journal of Information Policy (University Park, Pa.), 3*, 247-266. doi:10.5325/jinfopoli.3.2013.0247

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 288-296.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013, September). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 198-206).

Diamantini, C., Mircoli, A., Potena, D., & Storti, E., (2019). Social information discovery enhanced by sentiment analysis techniques, Future Generation Computer Systems, Retrived from: https://doi.org/10.1016/j.future.2018.01.051.

Dipendra, M. Henaff, M., Krishnamurthy, A., & Langford, J., (2019). Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning. Retrieved from: https://arxiv.org/abs/1911.05815

Dodds, P. S., & Danforth, C. M. (2009;2010;2017;). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of Happiness Studies, 11(4), 441-456. doi:10.1007/s10902-009-9150-9

Economist, Editor (2019, November 9). The first face-off. The Economist, 70.

Freedman, J., & Jurafsky, D. (2011). Authenticity in America: Class Distinctions in Potato Chip Advertising. *Gastronomica, 11*(4), 46-54. doi:10.1525/gfc.2012.11.4.46

Gal,U., Blegind -Jensen., Stein, M., Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics, Information and Organization, Volume 30, Issue 2,2020. Retrieved from: https://doi.org/10.1016/j.infoandorg.2020.100301.

Gelbard, R, Ramon-Gonen, R, Carmeli, A, Bittmann, RM, Talyansky, R. Sentiment analysis in organizational work: Towards an ontology of people analytics. Expert Systems. 2018; 3 5:e12289. Retrieved from: https://doi-org.libezproxy2.syr.edu/10.1111/exsy.12289

Gelman, A. Objections to Bayesian statistics. Bayesian Anal. 3 (2008), no. 3, 445--449. doi:10.1214/08-BA318. Retrieved from: https://projecteuclid.org/euclid.ba/1340370429

Gimpel, K. Schneider, N., O'Connor,B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. (2011). Part-of-speech tagging for twit- ter: Annotation, features, and experiments. In ACL.

Hargittai, E., & Walejko, G. (2008). The participation divide: Content creation and sharing in the digital age. *Information, Communication & Society*, *11*(2), 239-256. doi:10.1080/13691180801946150

Jiang, J., Wu, D., Chen, Y. *et al.*, Fast artificial bee colony algorithm with complex network and naive bayes classifier for supply chain network management. *Soft Computing* **23,** 13321–13337 (2019). https://doi-org.libezproxy2.syr.edu/10.1007/s00500-019-03874-y

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J: Prentice Hall.

Kan, I. P., & Drummey, A. B. (2018). Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce. *Computers in Human Behavior*, *83*, 243-253.

Kirch, B., Krupa, T., Luong, D. (2018). How do supervisors perceive and manage employee mental health issues in their workplaces? IOS. DOI 10.3233/WOR-182698

Liu, X., Li, J., Wang, J., & Liu, Z. (2020). MMFashion: An Open-Source Toolbox for Visual Fashion Analysis. *arXiv preprint arXiv:2005.08847*.

Momennejad, I., Learning Structures: Predictive Representations, Replay, and Generalization, Current Opinion in Behavioral Sciences. *Science Direct*, 32, 156-166, (2020). Retrieved from: https://doi.org/10.1016/j.cobeha.2020.02.017.

Morey, R.D., Hoekstra, R., Rouder, J.N. *et al.* The fallacy of placing confidence in confidence intervals. *Psychon Bull Rev* **23,** 103–123 (2016). Retrieved from: https://doi-org.libezproxy2.syr.edu/10.3758/s13423-015-0947-8

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Schultz, R. A. (2006). *Contemporary Issues in Ethics and Information Technology*. IGI Global.

Shen, M. J., Aiden, Y., Veres, A., & Gray, M., The Google Books Team, . . . Aiden, E. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 331*(6014), 176-182. Retrieved from http://www.jstor.org/stable/40986490

Silver, N., (2010,Jan). Obama's SOTU: Clintonian, In a Good Way. FiveThirtyEight. Retrieved from: https://fivethirtyeight.com/features/obamas-sotu-clintonian-in-good-way/

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

Spitzmüller, C. and Stanton, J.M. (2006), Examining employee compliance with organizational surveillance and monitoring. Journal of Occupational and Organizational Psychology, 79: 245-272. Retrieved from: https://doi-org.libezproxy2.syr.edu/10.1348/096317905X52607

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., . . . Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics - Association for Computational Linguistics, 26(3), 339-373. doi:10.1162/089120100561737

Warner, J., et al. “Seeing the Forest through the Trees: Uncovering Phenomic Complexity through Interactive Network Visualization.” *Journal of the American Medical Informatics Association : JAMIA*, vol. 22, no. 2, 2015, pp. 324–329., doi:10.1136/amiajnl-2014-002965. Accessed 6 Dec. 2020.

Warschauer, M. (2002). Reconceptualizing the digital divide. *First Monday*. Retrieved from: https://doi.org/10.5210/fm.v7i7.967

Wasserstein, R., Schirm A., & Lazar, N., (2019) Moving to a World Beyond “p < 0.05”, The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

Wetzels, R., Grasman, R., & Wagenmakers, E. (2012). A Default Bayesian Hypothesis Test for ANOVA Designs. The American Statistician, 66(2), 104-111. Retrieved December 6, 2020, from http://www.jstor.org/stable/23339468

Wetzels, R., & Wagenmakers, E. (2012). A default bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057-64. Retrieved from https://libezproxy-syr-edu.libezproxy2.syr.edu/login?url=https://www-proquest-com.libezproxy2.syr.edu/scholarly-journals/default-bayesian-hypothesis-test-correlations/docview/1242472691/se-2?accountid=14214

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, *23*(3), 327-343. doi:10.1093/llc/fqn015