Data Science: A Global Presence

The field of data science evolves every day, demanding an ever curious, determined mind of those who choose data science as a profession. Technology, mathematics, and business acumen are deeply rooted in the field and are constantly changing fields that require skilled and rapidly fluid individuals. Trying to pin point a true definition of the field of data science is a challenge as it means different things to different sectors of the professional arena in both the public and private sectors. Data science has humble beginnings, first arriving on the lips of academic researchers in the early 2000s, during the heat of the dot-com bubble. During this time, Professor William S. Cleveland published the academic paper "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.". In his paper, Cleveland accurately and systematically outlines a perfect marriage between statistics, modeling data, computing systems, tooling for analysts, and theory derived from evaluating the resulting process he calls "data science".[1] Cleveland alludes to a process where data is analyzed with the objective of deriving direct value and implementing that value into the respective organization. To add to Cleveland's robust and accurate assessment, outside of pure technical skills (which are undoubtedly a necessity), a structured analytical framework to which problems are processed with the goal of determining an approach that meets business or research questions is essential within the process of data science. To summarize, data science is an interdisciplinary field that uses scientific approaches, methods, statistical evaluation, algorithms and systems to conduct thorough analysis to answer questions and derive concise strategical insights for decision makers. This definition will indisputably change over time but seems to be robust enough for the time being.

Looking at the state of the field nearly twenty years after Cleveland's groundbreaking work, many of the necessities and challenges outlined by Cleveland continue to drive the field forward. Big data processing, data privacy, increasing need for technically trained individuals, and auditing the entire data science process from aggregation to black box machine learning algorithms remain critical research and business obstacles for the field of data science for the current and next generations of data scientists to wrangle and answer. Data science no longer remains an obscure corner of computer science research in academia. On January 29th, 2019, the heads of the US Intelligence Community (IC) testified before the Senate Intelligence Committee on global threats to the United States. In their joint assessment, the dangers presented by leveraging and weaponizing big data from global threat actors such as China and Russia were highlighted followed by a call to service to data scientists and data analysts echoed by each respective IC leader. The now global impact data science has in everyday governmental, business, and academic operations emphasizes the continued growth of the field and need for talented, motivated individuals.

The cutting edge, constantly changing and challenging elements of the field is what brought me to this program here at Syracuse, and the everchanging obstacles will continue to fuel my professional growth. Throughout the course of the Masters' in Applied Data Science at Syracuse University, certain learning objectives are laid out to ensure the quality of students coming out of the program to meet and exceed expectations of their employers or future academic endeavors. These learning outcomes include but are not limited to:

- Describing a broad overview of the major practice areas in data science
- Collect and organize data
- Identify patterns in data via visualizations, statistical analysis, and data mining

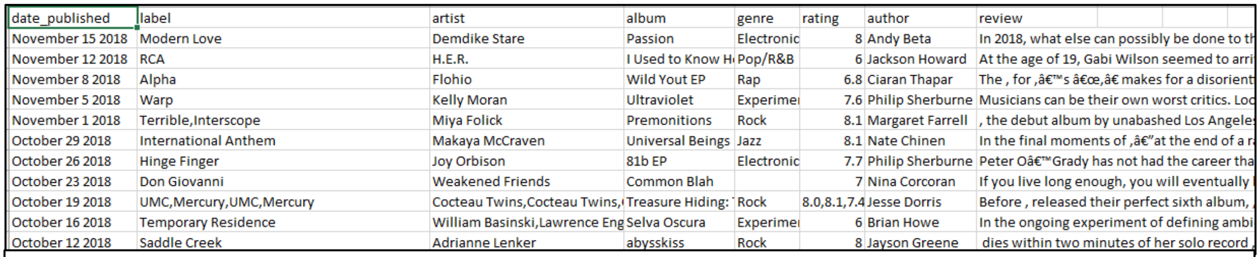
- Develop alternative strategies based on data
- Develop a plan of action to implement the business decisions derived from the analyses
- Demonstrate communication skills regarding data and its analysis for mangers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
- Synthesize the ethical dimensions of data science practice

Through 36 credits, and several different projects, these learning outcomes were obtained and will serve as a fantastic base for a career in data science and analytics to be built upon. While the course load unquestionably suffices for a strong starting point, side projects will remain a key aspect for any individual to remain sharp and knowledgeable on the advancements in the field. The following sections of the paper will highlight a cohesive portfolio of work that aims to display effectively meeting the previously listed learning objectives to attain approval towards my continued progress in the field of data science.

Data Quality is Essential

Data is everywhere one looks. As an aspiring data professional, the amble amount of data to hone and practice skills on allows for the creation of infinite problem sets. While it is easy to go to numerous websites and download pre-structured, tabular data, it is much more challenging to build a dataset from scratch using numerous techniques. In Text Mining, taught by Dr. Yu, the final project revolved around analyzing a corpus of text. To build knowledge of data collection practices, I built a web scraper using the open source Python tool Scrapy.[2] The goal of my project was to collect album reviews from the website Pitchfork.com, a globally renowned music review company.

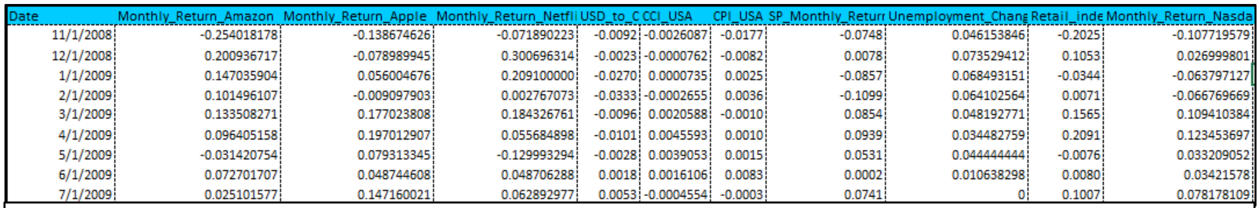
By building a corpus of album reviews coupled with author, music genre, overall rating and date, overall trends based on certain authors, and sentiment analysis, key takeaways could be extracted from the data. By taking the unstructured data from a website and giving it a tabular layout using a scraping tool, analysis is more efficient to conduct. By structuring and organizing the data, the entire data science and analysis pipeline becomes reproducible, giving an individual or an organization ease of access into the data for analysis. Also, a key part of this process is the cleaning of data, especially when it is being pulled directly from a website. Often, website data will have numerous special characters, spaces, and punctuation throughout text fields, making it essential that a tool like Pandas or Tidyverse is used to ensure and optimize data quality as well. Below, an example of the data has been displayed.[34]



date_published	label	artist	album	genre	rating	author	review
November 15 2018	Modern Love	Demdike Stare	Passion	Electronic	8	Andy Beta	In 2018, what else can possibly be done to th
November 12 2018	RCA	H.E.R.	I Used to Know H	Pop/R&B	6	Jackson Howard	At the age of 19, Gabi Wilson seemed to arri
November 8 2018	Alpha	Flohio	Wild Yout EP	Rap	6.8	Ciaran Thapar	The , for ‚â€™s â€œ‚â€ makes for a disorient
November 5 2018	Warp	Kelly Moran	Ultraviolet	Experime	7.6	Philip Sherburne	Musicians can be their own worst critics. Loo
November 1 2018	Terrible,Interscope	Miya Folick	Premonitions	Rock	8.1	Margaret Farrell	, the debut album by unabashed Los Angele
October 29 2018	International Anthem	Makaya McCraven	Universal Beings	Jazz	8.1	Nate Chinen	In the final moments of ‚â€”at the end of a r
October 26 2018	Hinge Finger	Joy Orbison	81b EP	Electronic	7.7	Philip Sherburne	Peter Oâ€™Grady has not had the career tha
October 23 2018	Don Giovanni	Weakened Friends	Common Blah		7	Nina Corcoran	If you live long enough, you will eventually
October 19 2018	UMC,Mercury,UMC,Mercury	Cocteau Twins,Cocteau Twins,	Treasure Hiding:	Rock	8.0,8.1,7.4	Jesse Dorris	Before , released their perfect sixth album,
October 16 2018	Temporary Residence	William Basinski,Lawrence Eng	Selva Oscura	Experime	6	Brian Howe	In the ongoing experiment of defining ambi
October 12 2018	Saddle Creek	Adrianne Lenker	abysskiss	Rock	8	Jayson Greene	dies within two minutes of her solo record

Figure 1 By making a tabular data structure, analysis and interaction with key data science packages becomes feasible.

Stock price analysis is one of the most common applications of data analysis and science, as the data is readily available for anyone who is interested via several different options. During my time in Data Analysis and Decision Making, taught by Professor Chernobai, my team and I chose to analyze Amazon's stock price. Instead of using Amazon's internal financial indicators, I took it upon myself to research and find external forces outside of the company that could potentially influence Amazon's stock price movement. In doing this, we attempted to break the conventional analysis structure by finding unique variables to include in our dataset. Using Amazon as our dependent variable, we sought out 15-20 independent variables that would then be used to conduct regression analysis to analyze and predict Amazon's stock price. Using a combination of competitors, economic measures, and ETFs, we were able to construct a data set looking at the monthly percent change between all our selected variables, approximately 120 observations and 20 variables. We were also strategic in our time frame we chose to use, dating back to 2007 to quantify both bearish and bullish market conditions within the dataset. By using percent change, our regression analysis outcomes were accurate and insightful, allowing us to further validate later portions of our analysis which will be discussed in the coming sections. Below, an example of the data has been displayed.



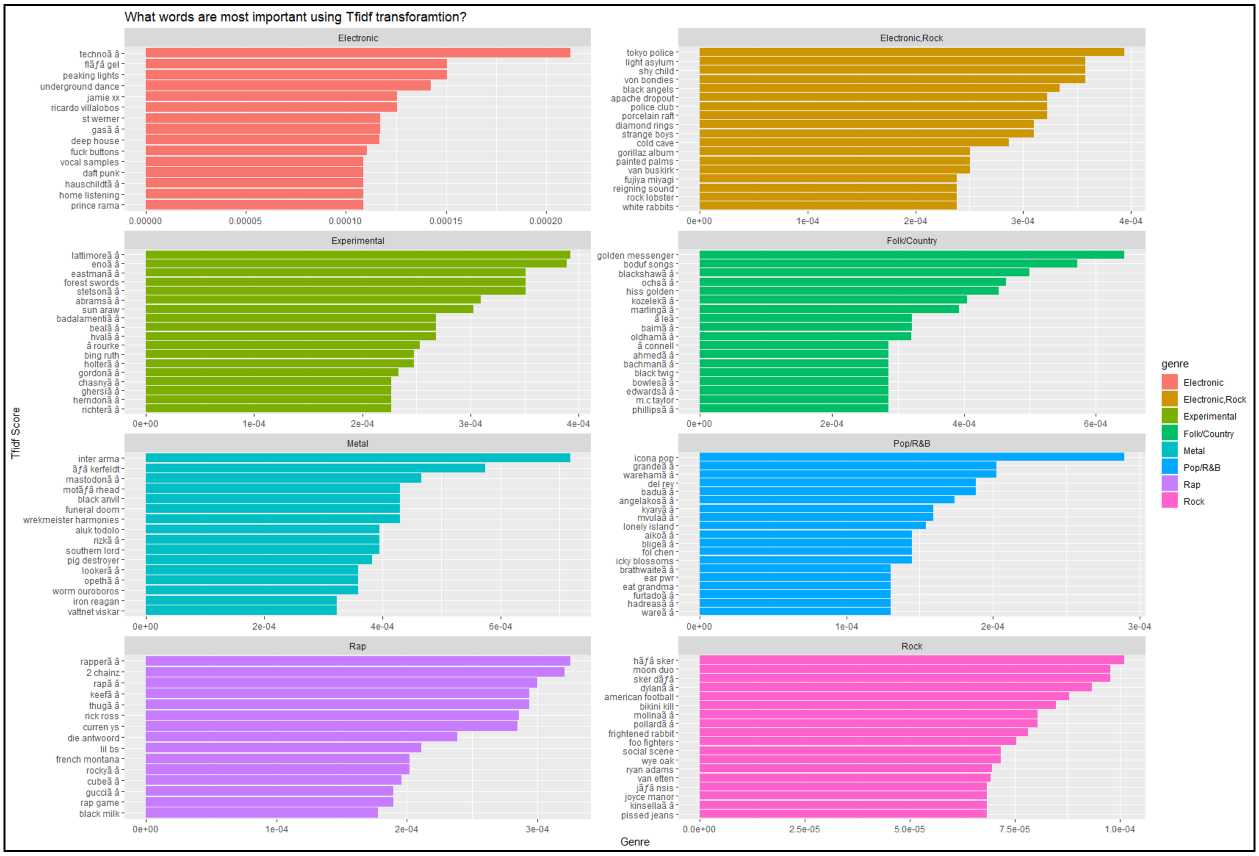
Date	Monthly_Return_Amazon	Monthly_Return_Apple	Monthly_Return_Netfli	USD_to_C	CCI_USA	CPI_USA	SP_Monthly_Returr	Unemployment_Chang	Retail_inde	Monthly_Return_Nasda
11/1/2008	-0.254018178	-0.138674626	-0.071890223	-0.0092	-0.0026087	-0.0177	-0.0748	0.046153846	-0.2025	-0.107719579
12/1/2008	0.200936717	-0.078989945	0.300696314	-0.0023	-0.0000762	-0.0082	0.0078	0.073529412	0.1053	0.026999801
1/1/2009	0.147035904	0.056004676	0.209100000	-0.0270	0.0000735	0.0025	-0.0857	0.068493151	-0.0344	-0.063797127
2/1/2009	0.101496107	-0.009097903	0.002767073	-0.0333	-0.0002655	0.0036	-0.1099	0.064102564	0.0071	-0.066769669
3/1/2009	0.133508271	0.177023808	0.184326761	-0.0096	0.0020588	-0.0010	0.0854	0.048192771	0.1565	0.109410384
4/1/2009	0.096405158	0.197012907	0.055684898	-0.0101	0.0045593	0.0010	0.0939	0.034482759	0.2091	0.123453697
5/1/2009	-0.031420754	0.079313345	-0.129993294	-0.0028	0.0039053	0.0015	0.0531	0.044444444	-0.0076	0.033209052
6/1/2009	0.072701707	0.048744608	0.048706288	0.0018	0.0016106	0.0083	0.0002	0.010638298	0.0080	0.03421578
7/1/2009	0.025101577	0.147160021	0.062892977	0.0053	-0.0004554	-0.0003	0.0741	0	0.1007	0.078178109

Figure 2 Enforcing data structure and organization is an essential part of any data workflow.

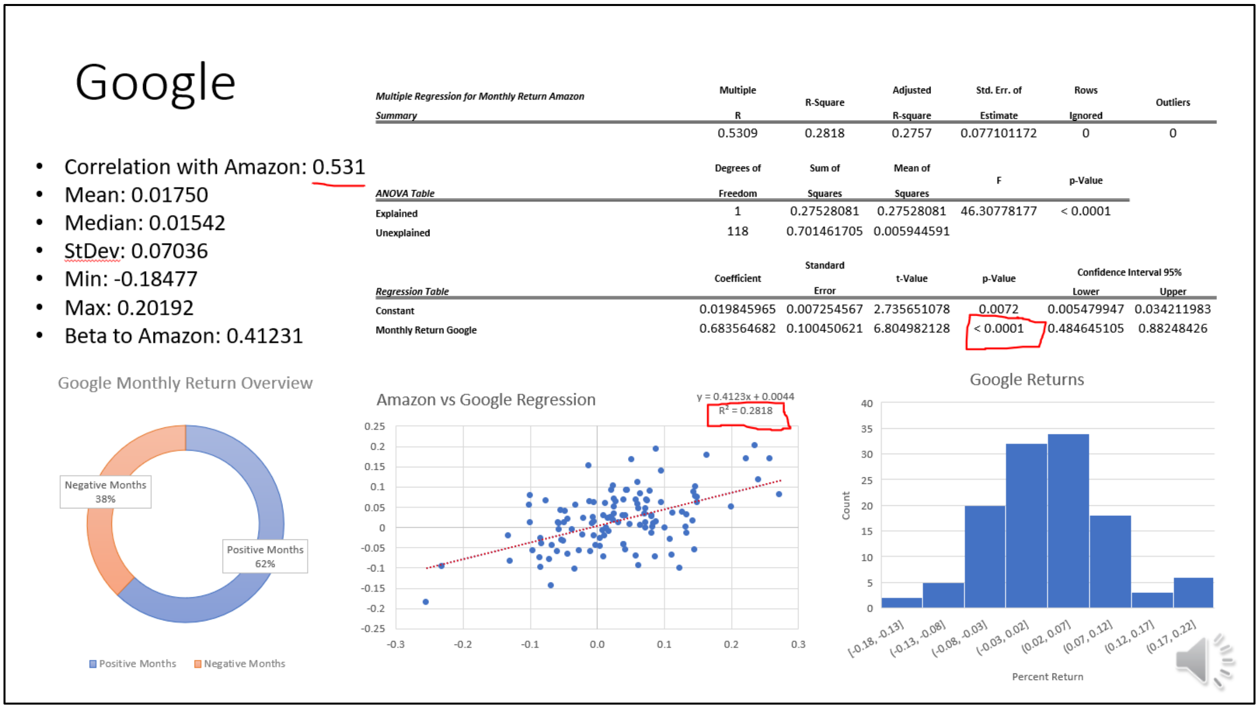
In both Text Mining and Data Analysis and Decision Making, the importance of data collection, organization, and cleaning were essential in beginning both individual and team-oriented workflows, improving the quality of the result. This idea of clean or "tidy" data was explored thoroughly by the researcher Hadley Wickham in his article titled "Tidy Data". Wickham states that roughly 80 percent of data analysis is consumed during the cleaning and organization process, with no real research on optimizing the clean process itself. Wickham goes onto define the very data structure that the R programming language is centered around, consisting of observations and variables, universalizing what "tidy" data should look like moving forward.[5] This paper has had ripple effects throughout the entire data analytics field, even inspiring the likes of Wes McKinney to develop Pandas, bringing much of the same speed and assurance of data cleanliness to the Python programming language as well. With Wickham's creation of the tidyverse and McKinney's Panda's, data analysis has been catapulted into the future as much of the workflow is now centered around visualization and analysis, allowing for grander and meaningful analysis to occur at an advanced rate.

Exploring the Data

With a clean, tidy dataset, taking a deep dive to identify key patterns, trends, and to display analysis through visualizations is a skill that will remain at the top of data scientist resume's for years to come. Communicating clearly what can be sometimes complex theoretical work is an important piece of displaying findings to the audience because without thorough understanding of the work being done, the work is often not trusted fully. This is the main reason exploratory data analysis and data visualization should always be a sharp tool in any data scientist's skillset. When working through the Pitchfork project for Text Mining, analysis was conducted using the tidytext packing in R, allowing for easy grouping of different categorical variables, such as music genre, to see the how sentiment scores and word importance differed across the data.[6] Looking at the visualization below, genres are represented in their graph with different colors allowing for the visual to easily read by the viewer. This visual aims to display the most important words using TFIDF scoring.



Exploring the data through visualizations is extremely important as well, using statistical methods such as histograms, regression analysis, and scatter plots to explore the relationships between two variables. This approach was used extensively during my Amazon Stock Price Analysis, as we created visualizations that aimed to demonstrate a linear relationship between our selected variables. By creating an almost dashboard like layout to our slides for our presentation, our audience could make key takeaways in a small amount of time about the message we were portraying, with minimal explanation. When looking at a problem or dataset with an attempt to explore the nature of the data, the use of histograms allows for detection out outliers, if the data is skewed to the left or right, obtaining mean, standard deviation, and median, covariance, and correlation will give the analyst and the audience a sound idea of linear or non-linear relationships between two or more variable. This ensures quality of model fitness coupled with statistically reinforcing ideas and hypotheses that are being worked with. On the next page, a snapshot is taken from the presentation, displaying our quick visualizations for the audience.



Google

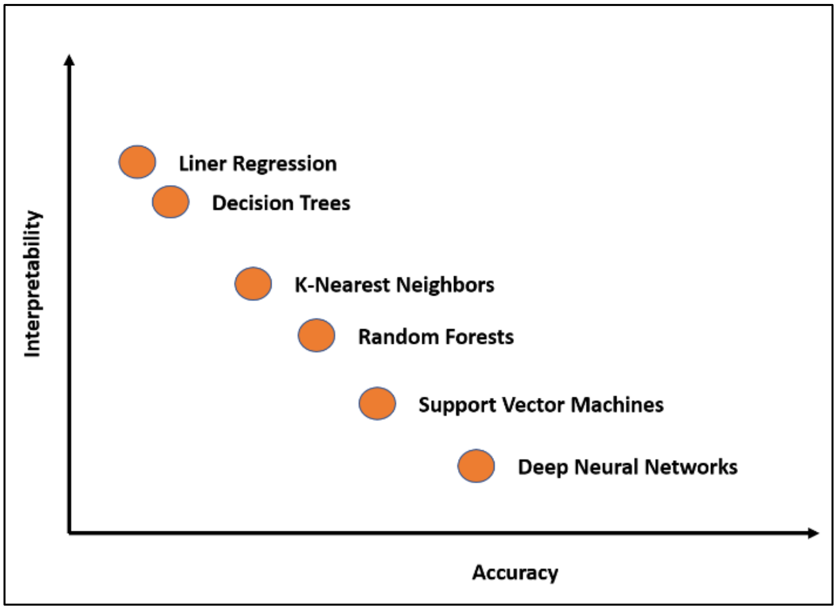
- Correlation with Amazon: 0.531
- Mean: 0.01750
- Median: 0.01542
- StDev: 0.07036
- Min: -0.18477
- Max: 0.20192
- Beta to Amazon: 0.41231

Multiple Regression for Monthly Return Amazon *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.5309	0.2818	0.2757	0.077101172	0	0

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value
Explained	1	0.27528081	0.27528081	46.30778177	< 0.0001
Unexplained	118	0.701461705	0.005944591		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	 Upper
Constant	0.019845965	0.007254567	2.735651078	0.0072	0.005479947	0.034211983
Monthly Return Google	0.683564682	0.100450621	6.804982128	< 0.0001	0.484645105	0.88248426

Actionable Insight and Implementation from Data



One of the most important skills any analyst or data scientist maintains is taking complex outputs from in-depth analysis and transforming findings into actionable insight and implementable strategy for a company or decision maker to derive value from. Months or even years of work can all be for naught if the implementation fails to be successful, making it a critical step in the data science and analysis pipeline.

For IST 707 Data Mining, currency exchange rates were analyzed using several different statistical and machine learning models. While Random Forest models offered better performance metrics, linear regression offers highly interpretable results. This is a common problem faced in the data science field, where scientists will sacrifice interpretably for accuracy of more complex models. Professionals in the field must determine if interpretability or accuracy will best solve or answer their business problem. For IST 707, I focused on interpretability, looking for different indicators of currencies that ultimately effect the UK Sterling Pound. These findings were then applied to foreign exchange trading strategies, thus making interpretation my focus. It seems this will continue to be a huge problem for data scientists, as it is easy to get lost in fine tuning the models rather than focusing on how these finding will be presented to a room of executives and then implemented into the business. Technology will continue to deliver groundbreaking results for businesses across the globe, but it remains in the hands of data scientists to bridge the gap between impressive business acumen and impressive technological acumen, delivering actionable insight and driving implementation in their respective settings.

Data Ethics – A Developing Conversation

Today, society continues to increase both reliance on and production of data. As with any commodity ever introduced to humans, data has been and continues to be exploited by large corporations and criminal actors alike. This is not to draw a parallel between the two groups but ethically, their practices are essentially the same. For example, not to pick the lowest hanging fruit, but Facebook serves as an example of how ethics of data can become a slippery slope in the matter of months and change the entire landscape of how data is viewed, shared, collected, and ultimately used by people across the globe. Facebook not only sold users data to secondary companies to analyze, compromising security on multiple levels, but also hid several massive data leaks from the public to protect public sentiment and stock price. This serves as the seminal case for social media companies, displaying just how much data they have on each one of their users and how intimate that data can prove to be. This has led to sweeping legislation across the globe, sparking the EU's GDPR, laying a groundwork that will tailor how companies use and share data for decades.

Unfortunately, the United States lags behind its allies in this regard. Fortunately, it is highly likely over the next 5 years, politicians will band against these technology companies (this has already begun) sparking legislation that will hopefully place the user at the center of the argument for protection. Maintaining ethics in data practices will continue to be a massive talking point, likely controlling political conversation in the years to come. Do not be surprised to see data ethics practices of major corporations becoming a vital political campaigning point as data literacy continues to rise outside of traditional data communities.

Conclusion – Where to Now?

Looking at where I began my data journey nearly 5 years ago, it was an admittedly intimidating challenge. Learning curves, extremely smart peers, and demanding subject matter have created a field in data science that still remains a buzzword for many in sectors across the economy and one that I am thrilled to be a part of. This, in a way, personifies the current data community as a whole, filled with buzzwords and hype. Ceilings continue to be shattered by technological innovation, creating an exciting and deservingly hyped field. Government funding, national coding initiatives, and other avenues have introduced new ways of becoming involved in the data community. The pendulum has swung nearly as far as it can in the positive direction for the data science field. What happens when the pendulum switches its trajectory towards the negative? This is a question that professionals entering the field now will likely encounter, creating an environment that will place data scientists, analysts, and engineers at the center of the conversation, highlighting the massive responsibility on their shoulders. Maintaining the integrity of the data science pipeline, from cleaning to presenting truthful and unbiased results to decision makers will become more and more important as data's value continues to be recognized by the globe. Ensuring and maintaining skill levels that appropriate for the challenge task ahead is essential as well. These are challenges we will face in the coming future and by following the principles outlined by this program and within this paper, we can make massive strides towards a world where data and humans live in synchronous harmony, placing users, customers, and stakeholders at the center of products, services, and analysis simultaneously. I have confidence in the community I have entered will face these challenges head on, with the perseverance and curiosity to solve problems that has propelled the field to where it is today, making the challenging future a worthwhile endeavor.

[1] https://www.jstor.org/stable/1403527?read-now=1&seq=3#page_scan_tab_contents
[2] https://docs.scrapy.org/en/latest/
[3] https://pandas.pydata.org/pandas-docs/stable/
[4] https://www.rdocumentation.org/packages/tidyverse/versions/1.2.1
[5] https://vita.had.co.nz/papers/tidy-data.pdf
[6] https://cran.r-project.org/web/packages/tidytext/index.html