WORKING AT A START-UP: EXPERIENCES AND LEARNINGS AS A DATA SCIENCE INTERN

A GLOOMY CLOUD: A SOCIOECONOMIC ANALYSIS ON THE USE OF DATA CENTERS IN NORTHERN VIRGINIA

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Ansel Sanchez

October 27, 2022

Technical Team Members: N/A

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

ADVISORS

MC Forelle, Department of Engineering and Society

Briana Morrison, Department of Computer Science

Introduction

In today's internet age, data is of utmost importance to both consumers and businesses, with concerns ranging from cloud storage to data processing needs. There has been a steady rise over the last six years that shows a 60% increase of internet users with a 440% increase in internet traffic (International Energy Agency, 2022). Moreover, the COVID-19 pandemic has added stress to the internet's infrastructure by emphasizing the importance of better internet connectivity and bandwidth for the millions of Americans that were working and continue to work at home (Vogels, Perrin, Lee, Anderson, 2020). Because of these needs and increasing internet demands, there has been rapid construction and development of data centers within the last two decades to help aid in supporting the internet's growing infrastructure (Schweitzer, 2021). With this recent boom of data center construction, one area has become a popular location for such development: Northern Virginia. However, with increased reliance on this infrastructure, one may wonder how sustainable these buildings really are in terms of their environmental concerns and the effects to the living community around them.

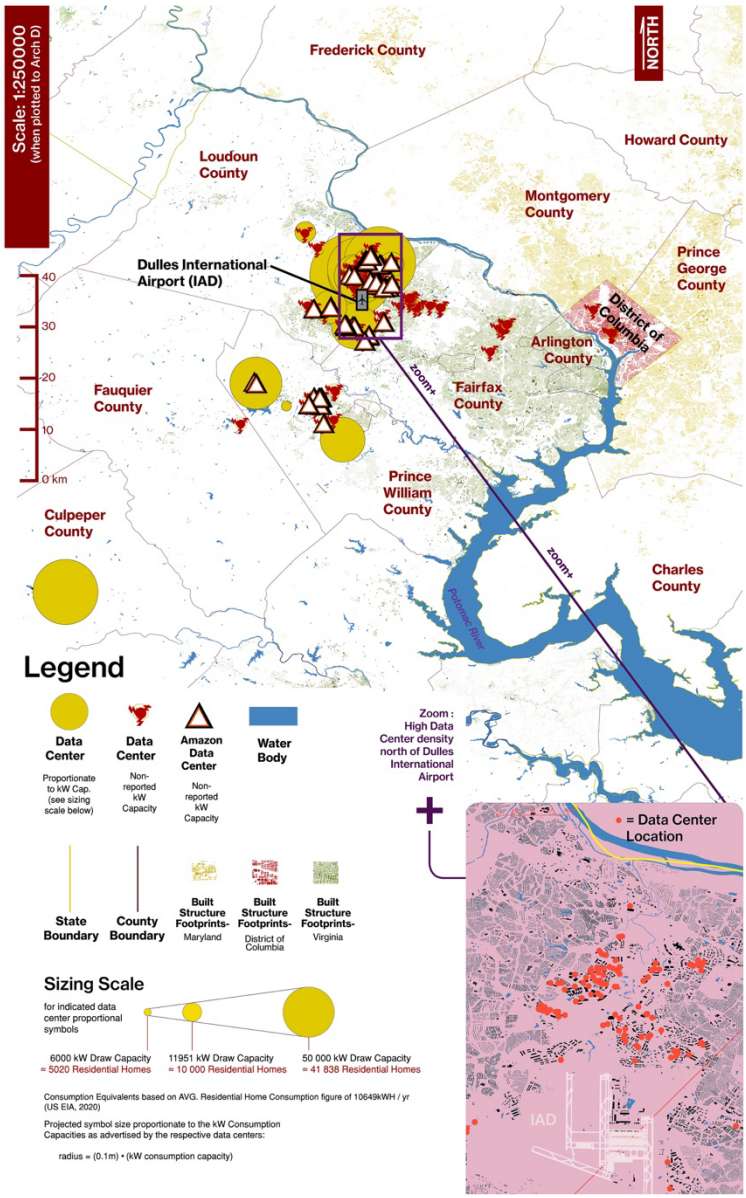
Data centers are defined as facilities that utilize computer systems to store and process data from the internet (Balodis and Opmane, 2012). Not only do these buildings hold the necessary computer systems to interact with user data but also house the necessary infrastructure to keep these buildings running, which includes power supplies, backup data servers, and cooling systems. Particularly with the power and cooling systems, data center facilities consume huge amounts of electricity to run their highly-energy demanding equipment as well as utilize a large volume of water for their cooling systems that keep running their equipment efficiently (Balodis and Opmane, 2012). Because of the need to store and process much of the world's internet data, data center facilities have a considerable land footprint that requires them to be constructed in

areas that are large enough to house at least one data center building, since it is not unusual to have a grouping of data center buildings in one campus.

Since the 1990s, there has been an increased interest in attracting large tech companies to build data centers and host their internet traffic in the Northern Virginia area (Northern Virginia Technology Council, 2016). This interest has been driven by state and local officials wanting to stimulate the Virginia economy by ushering in a large tech presence in the Northern Virginia area. Initially attractive for its cheap land, Northern Virginia, and Loudoun County in particular, boast large areas of open space that make it prime for building sprawling data center campuses. Furthermore, the relative closeness of the Potomac River to this area allows for easy access and use of this water system for intensive cooling arrangements, which can be seen in Figure 1 along with the high electrical energy needs of these data centers (Bast, Carr, Madron, and Syrus, 2022). By granting large tech companies a presence in Northern Virginia, the creation of thousands of jobs that lead to the influx of high-tech talent to oversee and interact with such facilities has been a trend over the last few years (Joint Legislative Audit and Review Commission, 2019). Although there are great economic benefits that come with bringing in these data centers, there are also hidden costs. These costs include negative effects to the environment through large energy consumption and outputting of carbon as a byproduct as well as negative effects to the local living communities in terms of housing prices and overall cost of living.

Figure 1

Concentration of Data Centers in Northern Virginia and their Energy Needs



Note. From "Four reasons why data centers matter, five implications of their social spatial distribution, one graphic to visualize them", by D. Bast, C. Carr, K. Madron, M., and A. Syrus, 2022, *Environment and Planning A: Economy and Space*, *54*(3), 441–445 (https://doi.org/10.1177/0308518X211069139)

With data needs being apparent to the largest tech giants, even the smallest companies need to think about ways to store, process, and monitor their data in sustainable ways. As an intern at a local Charlottesville start-up named Babylon Micro-Farms, I, along with another intern, was tasked with developing a new dashboard of metrics that would monitor and visualize their data coming in from Internet of Things (IoT) sensors in their smart hydroponic farming cabinets. The development of this full-stack application was achieved with sustainability in mind, specifically regarding software robustness and in how the application can be supported for better maintainability and future use. With companies both large and small thinking about data as a primary concern for their business, it is important to consider how the process of working with this data is affecting those both those inside the operation, such as engineers, and out, such as the community at large. Regarding the STS project, the concern will be looking at the effect of Northern Virginia data center usage on the local living community and environment around it and in regard to the technical project, building a full-stack data monitoring and visualizing application that emphasizes sustainable software practices for endured development and use.

Technical Project

A local Charlottesville start-up, Babylon Micro-Farms, which specializes in developing smart cabinets for automated hydroponic produce farming, was rapidly growing and needed to innovate its produce monitoring system. Their main product, the Galleri, is a smart cabinet that can grow more than 45 different produces in a hydroponic way, meaning with no soil and using nutrient packets (The Editors of Encyclopedia Britannica, 2022). The growing conditions of these plant varieties are managed and controlled remotely by Babylon's AI software through IoT sensors (Babylon Micro-Farms, 2022). As a data science intern, I was tasked with building a new

dashboard of sensor metrics that would be easy to use and be highly presentable both to engineers and potential investors. Prior to my time of being hired as an intern in the summer of 2020, the engineers would use their app to individually look at each farm's data and did not have a single, easily accessible place to view all their data at once.

Working in an Agile environment to complete the project, I used languages and technologies I was familiar with such as Python, Linux, and version control on top of a myriad of new tools learned throughout the internship experience, such as Prometheus and Grafana for data monitoring and visualization. In the end, I developed, for both internal and client use, a full-stack data scraping, visualizing, and alerting application to fully monitor all of the company's farms.

As larger companies think about how to sustainably run their data center operations, I also thought about the sustainability of data needs, but more so through a technical standpoint. For example, I thought about how the software could be built so that it is easily maintainable for future development and use. The key part of achieving this was by using Docker, an open-platform software that allows engineers to separate different services of an app into isolated containers so that integration, testing, and deployment can be done quickly and easily (Docker, 2022). By using this platform, engineers are able to refactor code in the visualization part of the application without needing to change anything in the data scraping portion of the application (and vice versa) and makes it easier for engineers to build long-lasting, robust code base for continued development. Furthermore, an application like the one built, would allow for better visibility of the company's data that makes it easier for engineers to do their job and diagnose problems in their sensors to optimize their growing algorithms. Furthermore, this would help attract more capital to grow their expanding start-up by having investors see a visually appealing and more sophisticated data monitoring system.

STS Project

As a resident of Northern Virginia, specifically Ashburn where many of the Northern Virginia data centers are built, I witness firsthand the daily ongoing construction of these buildings and see completed ones littered across areas near my neighborhood. I understand the positive economic impact these centers bring to Virginia but have often wondered what their downsides are in terms of how they affect the environment and affect who is able to live in the area due to an increase of high-paying, technical jobs.

Many past studies have focused on the overall effects of data center usage in terms of both the environment and social impacts. For example, it has been shown that data centers can be responsible for at least 1% of the world's electricity consumption (Al Kez et al, 2022). Furthermore, it has been estimated that data centers contribute to 2% of the world's carbon footprint (Hewlett Packard, 2012). In a case study performed by Lam et al in Hong Kong, they found that not only does the price of land for data centers rise due to the scarcity of available area, but this also drives up residential living and housing prices. Companies have shown some promise, however, in their initiatives to try to curb these negative environmental externalities. But it has been shown through research analyzing major UK data center operators that although companies pout these initiatives to be more sustainable, they are doing so ingeniously, with the primary goal still being increased profits (Jones, Hillier, Comfort, and Clarke-Hill, 2013).

It is because of this disconnect between the profit-driven goals of tech companies and the living population of areas with heavy data center presence that begs an investigation on analyzing the cost versus the benefits of data centers on the surrounding community. As mentioned previously, one such data center centric community is Northern Virginia and living in

this community for half of my life, I have seen the influx of people move in while complaining about rising house prices and everyday living expenses. This makes it imperative to be able to understand the role that data center facilities play in affecting the community they lay their presence in.

In this analysis, I will be relying on the Social Construction of Technology (SCOT) framework presented by Pinch and Biker, which will help me look at the relationship between people, the problem, and technologies to really get a sense of how data centers are making, either a positive or negative, impact on things that surround it (Pinch and Bijker, 1984). Furthermore, by really analyzing what the technology, which is the data center, means to different relevant social groups, then I can be able to understand how it is viewed by the community and how integration of this technology can be improved. On top of this, by looking at the interpretive flexibility of the data center technology, I can then judge and weigh positive view points against the negatives to get a sense of how ethical and, ultimately, useful these facilities are to the surrounding community.

Research Methods

Due to the seemingly conflicting positives and negatives views of running data centers in Northern Virginia, my project aims at answering this question: how has the building of data centers for better cloud and data uses in Northern Virginia affect the surrounding community both environmentally and socioeconomically and are they ultimately ethical with all factors considered? This question is an important one to ask and answer as, stated previously, the world is becoming increasingly reliant on large scale infrastructure in the form of sprawling data center campuses to keep up with the data and internet demands. Soon, many other places may become like the Northern Virginia area and understanding the unseen affects in both a positive or

negative light is important to build better and more sustainable communities where technology and humans can live that complement each other. To accomplish this, a sociotechnical analysis will be completed by weighing the costs and benefits to determine the impact of data centers on the real, lived experiences of those in the Northern Virginia community. This will be performed with information gathered from previous literature review and would look at certain metrics of sustainability performance such as power usage effectiveness, a value that considers all heat sources and power consumption for a data center (Zhang and Liu, 2022). I'll also be reaching out and talking to at least 12 Northern Virginia community members. Amongst these 12 individuals will be at least six that live in the area and work in the technology field and at least six will be those that live in the area and are not employed by the technology field. Furthermore, I will ensure that I will get a mix of individuals within each social group that are of varying socioeconomic statuses. By interviewing these two relevant social groups and having a range of varying social statuses, I will be able to get a sense of how those that are directly associated with the technology feel about it while comparing it to the sentiments of regular Northern Virginia citizens as well as how it affects the lives of those that come from varying socioeconomic statuses. This will help me with the application of the SCOT framework by building up identities of relevant social groups, how they interpret the technology, the sentiments these groups have about them, and possible solutions to the problem to placate some of their concerns. On top of this, information from the literature review regarding environmental concerns of data centers will be added to the SCOT analysis to analyze and figure out solutions that combines the concerns of both environmental and social groups.

Conclusion

As a computer science student and user of the internet, I understand the value that big tech companies place on user data and the importance of processing internet traffic to the best of their technological abilities to keep the internet running. However, I'm also a resident of Northern Virginia, and can understand how infrastructure to support their digital endeavors in the form of data centers can be seen as invasive to the community. By investigating the effects and weighing the costs against the benefits of these facilities in the Northern Virginia area to determine if these are ethical, one could apply these findings to other areas where data centers are beginning to be constructed. At the end of the analysis, it would be expected that the views on the issue may be layered and complex. However, the information gathered and analyzed will ultimately prove useful in finding ways to better integrate data centers in a community. With internet and data usage being such a prevalent part of today's society, stepping back and looking at how our actions are affecting our communities is important to ensure both humanity and modern technologies can grow in healthy and sustainable ways.

References

Al Kez, D., Foley, A. M., Laverty, D., Del Rio, D. F., & Sovacool, B. (2022). Exploring the sustainability challenges facing digitalization and internet data centers. *Journal of Cleaner Production*, 133633. https://doi.org/10.1016/j.jclepro.2022.133633

Alkhatib, O.J., Abdou, A. An Ethical (Descriptive) Framework for Judgment of Actions and Decisions in the Construction Industry and Engineering–Part I. *Science and Engineeringß Ethics 24*, 585–606 (2018). https://doi.org/10.1007/s11948-017-9895-1

Balodis, R., Opmane, I. (2012). History of Data Centre Development. In: Tatnall, A. (eds) *Reflections on the History of Computing. IFIP Advances in Information and Communication Technology*, *387*, 180-203. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33899-1_13

Babylon Micro-Farms (2022). Introducing Galleri. Retrieved https://babylonmicrofarms.com/

Bast, D., Carr, C., Madron, K., & Syrus, A. M. (2022). Four reasons why data centers matter, five implications of their social spatial distribution, one graphic to visualize them. *Environment and Planning A: Economy and Space, 54*(3), 441–445. https://doi.org/10.1177/0308518X211069139

Britannica, T. Editors of Encyclopaedia (2022, October 8). *hydroponics*. *Encyclopedia Britannica*. https://www.britannica.com/topic/hydroponics

Docker. (2022). Docker Overview. Retrieved from https://docs.docker.com/get-started/overview/

Joint Legislative Audit and Review Commission. (2019). *Data Center and Manufacturing Incentives Economic Development Incentives Evaluation Series.* Retrieved from http://jlarc.virginia.gov/pdfs/reports/Rpt518-1.pdf

Jones, P., Hillier, D., Comfort, D. and Clarke-Hill, C. (2013), Towards sustainable data centre operations in the UK. *Journal of Property Investment & Finance*, *31*(1), 89-100. https://doi.org/10.1108/14635781311292999

Hewlett Packard (2012), "*Sustainability*", available at: www.hpl.hp.com/research/sustainability.html

International Energy Agency (2022). Data Centres and Data Transmission Networks. Retrieved from https://www.iea.org/reports/data-centres-and-data-transmission-networks

Lam, P. T. I., Lai, D., Leung, C., & Yang, W. (2021). Data centers as the backbone of smart cities: Principal considerations for the study of facility costs and benefits. *Facilities, 39*(1), 80-95. https://doi.org/10.1108/F-09-2019-0103

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science, 14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Schweitzer, A., (March 26, 2021). The Pandemic is Driving A Data Center Boom in Northern Virginia. *NPR.* Retrieved https://www.npr.org/local/305/2021/03/26/981557613/the-pandemic-is-driving-a-data-center-boom-in-northern-virginia

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist, 43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Northern Virginia Technology Council. (2016). The Economic and Fiscal Contribution that Data Centers Make to Virginia. Retrieved from http://biz.loudoun.gov/wp-content/uploads/2020/01/2016-Loudoun_DataCenters-NVTC-Report.pdf

Vogels, E., Perrin, A., Lee, R., Anderson, M. (April 30, 2020). 53% of Americans Say Internet Has Been Esssential During the COVID-19 Outbreak. *Pew Research Center*. Retrieved from https://www.pewresearch.org/internet/2020/04/30/53-of-americans-say-the-internet-has-been-essential-during-the-covid-19-outbreak/

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136. http://www.jstor.org/stable/20024652

Woolgar, S. (1990). Configuring the User: The Case of Usability Trials. *The Sociological Review, 38*(1_suppl), 58–99. https://doi.org/10.1111/j.1467-954X.1990.tb03349.x

Zhang Y, Liu J. Prediction of Overall Energy Consumption of Data Centers in Different Locations. *Sensors (Basel)*, 2022 May12; *22*(10):3704. doi: 10.3390/s22103704. PMID: 35632113; PMCID: PMC9145178.