# Data Analytics in Public Policy and Evaluation

Kirsten Ronning

Arizona State University

CPP 593: Applied Project

Professor Robert Rowley

July 9, 2021

## Abstract

In the past decade, data analytics has become a booming industry. Technological advancements in modern-day computers have allowed for large data storage systems and complex predictive algorithms. While much of the focus is often directed toward major corporate efforts to effectively utilize data, there are numerous other industries that have relied on data analysis techniques for centuries and continue to do so. This paper walks through the history of data analysis by focusing on the development of data visualization, regression modeling, and classification modeling technologies. This information is then applied to the public sector specifically, taking a focus on how public policy can be determined and improved upon by data analytic approaches to evaluation and policy modeling. A closer look is taken at three specific organizations: the U.S. Department of Housing and Urban Development (HUD), the Illinois Housing Development, and the UChicago Urban Labs. By deciphering how these federal, state, and local institutions operate, we can envision the future of data analytics in public policy.


*Keywords:* data analytics, public policy, policy modeling, program evaluation

## Contents

## Data Analytics in Public Policy and Evaluation

Terms like “big data” and “artificial intelligence” have become quite the buzzwords since the mid-2000s. The development and expansion of the Internet, coinciding with the personalization of social media access, has created ample opportunity for the creation and storage of large amounts of information. Although most people are marginally aware of the fact that their actions and decisions while utilizing social media and e-commerce platforms are large contributions to major data analyses, many do not understand the actual scope of data tools and technologies. It is common knowledge that “big data” requires advanced analysis, typically conducted by large technology corporations, but there are countless other instances that information and knowledge can be effectively utilized to tell stories and make predictions. The instruments developed to inform data analyses are not only applicable to for-profit companies working to expand their earnings, but they are also extremely beneficial in the public sector as well. Rather than politicians simply guessing what would best serve their communities, they now have access to scientifically developed tools that can not only communicate what their policies will do but can also determine if they will work before they are even implemented. Data analytics fuels effective public policy and evaluation decisions.

It is recognized that data analytics tools and technologies are beneficial in practically any decision-making field, but the history of how these were developed and how they work is still not well-known by the average person. This paper will dive into the history of data analysis, specifically in relation to public programs and policy. Data techniques that exist today, ranging from data visualization to statistical modeling and machine learning will be covered. From there, it will more specifically cover how these methods have been applied to public sector and program policy analysis. A majority of the focus will be spent on real-world examples provided

by three case studies that focus on public housing programs centered in Chicago, Illinois, United States of America. A thorough breakdown of data analysis in the Department of Housing and Urban Development (HUD), the Illinois Housing Development Authority, and the UChicago Urban Labs will be included in order to demonstrate the role that data analysis plays at the federal, state, and local levels of the public sector. By understanding how data science was originally created, as well as how its currently being applied to program evaluation and public policy, we can more effectively interpret how it will most likely continue to operate in and innovate the industry.

## Literature Review

### History of Data Analysis

The field of data analysis and data science has an extremely broad reach, and its scope covers a multitude of different areas. Some might associate data analytics with for-profit business enterprises, but the methods and technologies developed over time can easily be applied to any sector. Based off of research, there appear to be three main techniques that many data analysts employ: data visualizations, statistical modeling, and machine learning.

#### *Data Visualizations*

For many data analysts, visualizations are the primary and most effective means of communicating information to stakeholders of all levels. According to Friendly (2008), the first sign of data visualizations appeared as early as the 10$^{th}$-century in the form of a multiple time-series graph (p. 18). Graphs and navigational visualizations became very prominent throughout the 1700's, and what we know as modern graphics made their way into existence a century later. Little formal development in data visualization occurred until around the 1950's, coinciding with the development of the modern-day computer. Friendly (2008) writes, "It may be argued that the

greatest potential for recent growth in data visualization came from the development of interactive and dynamic graphic methods, allowing instantaneous and direct manipulation of graphical objects and related statistical properties" (p. 41). Azzam et al. (2013) corroborates this same timeline of visualization development over time in their work as well. Data visualizations have the power to communicate both complex and very simplified pieces of information about data, depending on the audience they are made for. Theories regarding design such as placement and color choices exist to better serve the science of data visualization.

### *Statistical Modeling*

Another technique involved in data analytics is that of statistical modeling. There are two types of models that data analysts typically utilize in their statistical analyses: regression models and classification models (Stobierski, 2021). Regression models are used in conjunction with experiments to determine which independent variables contributed the most to the resulting dependent variables. According to Stanton (2001), the development of regression analysis can be attributed to the work of Sir Francis Galton and Karl Pearson, who discovered what is now known as the regression slope. Their work proved that researchers could look at past contributing factors to understand the current state of their subjects. In modern times, regression models are a lot more complex than just a plotted line on an x- and y-axis graph, and statistical software exists for data analysts to rearrange variables and run their models.

### *Machine Learning*

Classification models are an additional statistical method, and they can be considered synonymous with machine learning techniques. Stobierski (2011) describes classification modeling as, "…a process in which an algorithm is used to analyze an existing set of known points." This approach most often works to group variables based off of previous information in

order to make accurate predictions about these variables. Fradkov (2020) explains that Frank Rosenblatt, a psychologist, is credited as the first person to implement machine learning technologies in 1957 (p. 1385). Just like data visualization, classification modeling had its main developments coinciding with the further development of computers. With that being said, there have been some striking developments in machine learning throughout the past two decades, led by Ian LeCun, Yehoshua Benjo, and Geoffrey Hinton (Fradkov, 2020, p. 1388). Buzzwords from these technologies such as "algorithm", "artificial intelligence", and "big data" have become commonplace for many Americans, as machine learning has been integrated into most large business practices.

**Data Analysis in Public Programs**

Information and research regarding the history of these three techniques is very plentiful, and more exists regarding the implementation of data analysis in public policy and program analysis. Data analytic techniques are beneficial in every aspect of a program, including the problem structuring, forecasting, prescription, monitoring, and evaluation phases (Dunn, 2016, p. 8). According to Estrada and Yap (2013), the development of policy modeling occurred in conjunction with the availability of computers as well, and today a majority of policy modeling papers are related to the public sector (p. 178). Government and other public agencies utilize data visualization, regression modeling, and machine learning in order to inform their experiments and programs.

There appears to be a consensus among much literature that data analytics in the public sector is becoming more advanced as data has become more accessible. Pirog (2014) states that, "The data.gov Web site alone lists over 80,000 datasets, searchable by key words…" (p. 537). As data becomes more attainable, stakeholders will expect both clarity and a level of detail from

their program administrators. Azzam et al. (2013) predicts that stakeholders will not only possess a fluid knowledge of data visualization techniques, but they will expect program administrators to fully utilize their visualization knowledge in order to project the most accurate and influential information.

There are some ethical considerations to consider regarding data analysis techniques in public programs. Rodolfa et al. (2021) points out that there have been some shortcomings when it comes to the general development of machine learning (p. 1). Just as American society is based off biased and racist systems, the classification models have learned these themselves over time. If programs are utilizing prejudiced models to conduct their work, they will not be overall effective for the public they are intending to serve. Another ethical issue occurs in regression modeling, through the act of p-value hacking. Program administrators have to be sure that the experiments they run are not simply set up to prove their preconceived notions, but instead to actually find the best predicting variables for the change they want to see.

Many program studies that utilize data analysis methods are readily available in electronically published journals, ranging from public safety, the environment, public housing, and more. As more research is conducted, specific studies will be narrowed down to use for the case study aspect of this applied project.

## Methodology

The findings and results included in this paper are mostly derived from qualitative accounts published in academic journals, textbooks, and by public organizations themselves. A majority of the resources were accessed from the ASU Library's database. Most of the information regarding the history of data visualizations, regression modeling, and classification modeling came from peer-reviewed documents and textbooks. Primary resources were also

utilized, and some accounts date back as far as the 19th century. In regard to the current and future practices such as software and policy operations, reputable websites were used as references. This paper can be considered a summary of multiple resources to produce a qualitative account of data analytics in public policy and evaluation.

## Findings and Results

Data analysis can be considered the science of cleaning, altering, manipulating, and reorganizing information in order to effectively communicate a story. Although this definition may appear oversimplified, it becomes a bit more complex when the myriad of methods to conduct data analyses are considered. Data can be shaped to be used for important decision making, as analysts have the opportunity to evaluate current applications, while also making predictions about future best practices. The main methods that data scientists currently take advantage of are data visualizations and statistical modeling.

Many people have seen data visualizations, and they most often appear in the form of choropleth and heatmaps, charts, and other plots. Effective data visualizations are powerful tools to communicate information to stakeholders at every level, as well as the general public. Data visualizations are often the final product used for presentations and figures in journals, but statistical modeling procedures determine the information that is ultimately visually displayed. There are many methods of statistical modeling, but the two most commonly seen are regression models and classification models. Regression models can tell us how we arrived at where we currently are, while classification models can predict where we will be going. Data analysis tools such as data visualizations and statistical modeling are all made accessible through the use of computer technology. These techniques have a long history from before computers as we know them even existed.

## The Formation of Data Analysis

When referring to data, most people consider it to be bits of information stored and managed on advanced computer systems. However, data has existed in multiple capacities throughout all of history. Apsray (2019) asserts that, "…in the old days, we used to have the Library of Alexandria, and we used to have great collections in huge buildings such as the Library of Congress" (p. 1). Data was recorded and stored in numerous functionalities, and it was also analyzed. Although our modern-day data handling comes from techniques developed and refined only as recently as 2005, there is a complex and long-standing history of the progress of data science. In order to fully understand this history, a closer look at the advancement of data visualizations and statistical modeling can be taken.

### *Data Visualizations*

Data visualizations come in many shapes and forms. They can range from something as simple as a bar chart of three variables that was automatically generated by Microsoft Excel to an entire dashboard created by the Center for Disease Control (CDC) that automatically updates the world-wide spread of the COVID-19 pandemic. While some might not be intentionally aware of it, most people have interacted with a data visualization at some point in time. They exist to visually depict quantitative or qualitative information, and they serve us at both professional and personal capacities. By walking through the history of data visualizations, it becomes apparent that they have been truly relevant in more areas than generally expected.

#### The Origins of Visualizations and Social Implications

The first key historical data visualization appeared in around 150 B.C. Claudius Ptolemy, an Egyptian astronomer and mathematician, created a map projection of the world utilizing longitudinal and latitudinal coordinates for the first time (Friendly, 2008, p. 18). The techniques

that Ptolemy used in his work became the basis for most geographers and astronomers for centuries to come. Shortly after Ptolemy's creation, in approximately 950 A.D., an anonymous astronomer drew the first quantitative data visualization, a time-series graph that depicted the "…seven most prominent heavenly bodies over space and time…" (Azzam, 2013, p.11) using an x- and y-axis approach. The next major development in visualizations occurred many years later, when Scottish engineer William Playfair developed multiple modern-day charts, including line graphs, pie charts, bar charts, and circle graphs. By the start of the 19th century, scatter plots and histograms were also established. With that being said, most of the focus remained on mapping techniques. One of the most credited maps of the time was formed by Dr. John Snow, an English physician, who used dots on a map of London to represent the number of deaths from the cholera outbreak of the time. This map ended up assisting health officials in understanding what exactly was causing the decades-long cholera outbreak, as a majority of the dots were recorded around the Broad Street pump (Friendly, 2008, p. 27). To this day, data visualizations concerning information about the spread of disease and illness remain helpful sources of information for professionals and the public.

The latter half of the 19th century can be considered the most productive and innovative time for illustrated data visualizations. Technologies like 3-D mapping, area differentiation, and scales and legends, were formed as public officials began to make use of visualizations for documentation and policy decisions. Another major epidemiological advancement occurred in 1857, when Florence Nightingale, an English statistician working as a nurse at the time, created a polar area graph to convey the importance of sanitation throughout the Crimean War (Hedley, 2020, p. 27). This graph is exceptionally notable, because it is regarded as having effectively persuaded both public officials as well as the public. At this point, data visualizations were not

only utilized by high-ranking professionals, but they were designed with rhetorical and influential purposes as well. Throughout Western Europe, entire statistical atlases with economic, financial, and spatial data began to be published so that public officials could use them for city planning (Friendly, 2008, p. 34). The United States of America experienced the benefits of improved data visualizations, one of the most notable affecting the country's largest federally run data collection operation: the United States Census. In 1874, Francis A. Walker created an entire atlas of tables and visualizations from the data collected in the ninth census (Walker & Bien, 1874). Various charts and maps were generated using color and other design techniques to illustrate information about immigration, wealth, disability rates, education, and more. By the end of the 1800's, most of the Western world was actively utilizing data visualizations.

The next few decades did not produce nearly as much improvement or modification in data visualizations. Although, they were still actively utilized and refined over time. An extremely important development occurred in 1962, when John W. Tukey, a United States statistician, released his paper, *The Future of Data Analysis*. In this paper, he asserts that statistics and data analytics should become their own separate fields of study (Tukey, 1962). While this pronouncement was undoubtedly influential for the field of data analytics, Tukey did even more. He specifically focused on the exploratory data analysis subsector of the field, and he even created stem-and-leaf displays and box-and-whisker plots (Hoaglin, 2003, p. 313). His work coincided with the evolution of modern-day computer systems and software. As the creation of the personal computer and the Internet occurred in the late 1980's, data visualizations continued to advance. One of the latest developments was the Hans Rosling interactive visualization, presented in his 2007 TED Talk (Azzam, 2013, p. 12). Visualizations are no longer

static images on a piece of paper, but they now have the capacity to be dynamic and collaborative experiences for all to experience.

### Modern Guidelines for Effective Data Visualizations

The history of the development of data visualizations is vast, and this area of the field of data analysis has already contributed to societal advancements. With that being said, there are still aspects of visualization techniques that continue to be explored and codified. Entire sets of guidelines and principles have been developed to ensure that data visualizations work as they are supposed to. Now that the various types of charts and graphs have been developed, it is now up to the individual analyst to present them in a manner that effectively communicates their message to their stakeholders. Ajani et al. (2021) determine that the major focus of every data visualization should be to declutter and focus the visual (p. 2). In order to declutter visualizations, it is recommended to be cautious of aesthetics and artistry taking over from the actual message of the product. Luckily, techniques to focus the visualization can assist in ensuring that the story is told. Ajani et al. (2021) suggest annotating and pointing out specific highlights and trends to the readers, so that they can clearly decipher the information being presented (p. 3).

Throughout the 2010's, many other recommendations just like the one described above have been developed by data analysts and graphic designers. For example, Evergeen and Metzner (2013) assert that the construction of data visualizations should concentrate on simplification and emphasis (p. 6). Whether its declutter and focus, or simplification and emphasis, a similar theme of getting to the point and being specific is apparent. However, more precise instructions and protocols have been catalogued. One specific set involves the Gestalt Principles of Design, which breaks down exactly how viewers respond to and interpret the

information that is in front of them. Turner and Schomberg (2016) explain that the Gestalt Principles include figure-ground segregation, closure, proximity, continuity, similarity, past experience, and symmetry. What makes the Gestalt theory, coined by Max Wertheimer, Kurt Koffka, and Wolfgang Kohler, especially beneficial is that it is based off of inclusivity and ethics. Using these principles, data analysts can ensure that their visualizations are effective for those with color blindness or other visual disabilities as well. These advancements have created opportunity for visualizations to communicate to even more individuals both accessibly and equitably.

### Data Visualization Software

While the original data visualizations created by Claudius Ptolemy and William Playfair were meticulously hand-drawn, modern day computer technology allows for a much more accessible and time-efficient approach to developing data visualizations. Data software not only saves analysts time, but the most advanced can also ensure that any graphics produced adhere to design best practices. Marr (2017) lists software such as Tableau, Qlikview, and Plotly as leading software of the modern time. Not only are they easy to operate, with drag and drop and other interactive functionalities, but they are also easily connectable to large databases stored in programs such as Python, R, Hadoop, Amazon AWS, and SQL. Visualization software can identify what type of graphs or charts would be the most useful based off of the data inputted. Many also offer the opportunity to create interactive dashboards for both presentations and every-day use. Data analysts have never been more equipped to tell stories at such a grand scale.

## *Regression Modeling*

Data visualization is an effective method of telling a story, but before an analyst can get there, they have to decipher exactly what narrative the data is communicating. Oftentimes, there

are thousands of rows and columns of data, more than what the average human eye could fully comprehend. Because of this, analysts are tasked with having to rearrange and shape the data so that clear conclusions can be drawn. Through statistical modeling, analysts can further understand how the data interacts with one another and can expect how it will continue to operate. According to Stobierski (2019), two of the most common types of statistical modeling approaches include regression models and classification models. Regression models are utilized to determine what effect, if any, the independent variable(s) of a study had on the dependent variable(s). When conducting technical and scientific experiments, regression models can be very straightforward. However, most experiments, especially in the social sciences, are based off of data that are not always collected under perfect conditions with precise controls. Because of this, many data analysts employ quasi-experiments and run variations to regression models such as interrupted time series, difference-in-difference, fixed effects, and more (Lecy & Fusi, n.d.). Regression models are effective at determining whether or not a change was detected, as well as how much of a change it was.

**History of Statistical Science and Social Implications**

Sir Francis Galton and Karl Pearson, English mathematicians, are credited as the first to discover regression modeling. In 1877, Galton set out to conduct an experiment on sweat peas. His findings, what he called reversion at the time, were not only beneficial to the biological and biomedical communities, but also to the statistical sciences (Ariew et al., 2017). The discovery of regression to the mean would be a direct influence to linear regression approaches. Around the same time, in 1899, English statistician Udny Yule utilized regression analysis in the realm of public policy. He created a regression equation to analyze if access to housing had an effect on the amounts of paupers (Freedman, 1999, p. 247). Based off of his model, he concluded that

those municipalities that provided homes for paupers had more paupers. Although Yule's study was profoundly advanced for the time, further reflection reveals that he did not control for many variables, and his findings are not actually statistically significant. All of that considered, this study was revolutionary for using regression modeling in data analysis.

Another key figure in regression modeling is someone previously introduced: Dr. John Snow. Not only did his 1855 dot map contribute to the development of data visualizations, but his approach to determining the cause of the cholera outbreak was highly influential for the development of quasi-experiments and regression models in social statistics. He resolved that the only difference between those who were suffering from cholera and those who were not was their water, and he was able to hold all other variables as controls (Freedman, 1999, p. 246). The science of regression models continued to develop throughout the 20th century, and by the arrival of the 21st century, computer technology was available for advanced models to be run at high speeds and in real time.

### Statistical Modeling Programs

Data analysts now have the opportunity to make quick adjustments to their models, providing for more accuracy and certainty. Most statistical software has the capacity to run regression models. Coincidentally enough, many of the programs that data visualization software connects to, like R and Python, are heavily utilized for running regression models. Other leading software includes SPSS, SAS, and even Microsoft Excel. Analysts can promptly understand the relationship between the variables of their data, which is incredibly beneficial when it comes to evaluation of current policy and interventions.

## *Machine Learning*

Another type of modeling that has emerged in recent history is classification modeling. While regression models can help explain what *X* variables lead to the resulting *Y* variables, classification models can take previous *X* and *Y* variables to predict future *Y* variables. Classification models use these training datasets in order to sort variables into different categories, or classifications. There are a couple of different approaches that a data analyst can take in order to utilize classification modeling. One of the main techniques is through utilizing a Naïve Bayes formula, derived from Bayes Theorem. This formula calculates the probability of an occurrence given that something else has already happened. Another approach to classification modeling is utilizing decision trees. A decision tree begins at the top, with a root node, that asks an initial question. From there, it breaks into binary branches, continuing to ask a new, more specific question until a classification is eventually derived. This method might seem arduous, however, when programmed into a computer, it can take only milliseconds.

### History and Development of Machine Learning

While many of the major advancements in data visualizations and regression modeling occurred throughout the 1800's, classification modeling did not break through until the end of the 1950's. Considering this mechanism heavily relies on computers, this should come as to no surprise. An American psychologist, Frank Rosenblatt, is often cited as the original developer of machine learning. Rosenblatt (1960) himself writes, "This program uses the IBM 704 computer to simulate perceptual learning, recognition, and spontaneous classification of visual stimuli in the perceptron, a theoretical brain model…" (p. 301). His intentions were to model all parts of the brain in order to train it to recognize letters of the alphabet. It did not take long for other researchers to continue utilizing machine learning techniques to advance their own fields. Yakov

Tyspkin, a Russian cyberneticist, made major developments not much longer after that would ultimately contribute to the way that today's algorithms function (Fradkov, 2020, p. 1386). Individuals and small teams of researchers continued to make contributions to the development of machine learning, until major technology companies had a stake in the success of machine learning. Now that machine learning was established, the goal was to use classification modeling as quickly and cheaply as possible.

### Machine Learning Software

Google was one of the first large companies to make major investments and advancements in machine learning. In 2004, they came out with MapReduce technology, and shortly after produced Hadoop (Fradkov, 2020, p. 1387). These technologies were radical at the time, and data analysts were able to access enormous amounts of data in an open-source format. In the fifteen years since this major release, many other companies have published their own variations of machine learning software. IBM Machine Learning, Google Cloud AI Platform, Amazon Machine Learning, and Anaconda are all leading and accessible programs (Boog, 2021). These platforms are used by data analysts at every scale who utilize classification modeling and machine learning techniques to inform their analyses and programs.

## Applying Data Analysis to Public Policy

Data visualizations, regression modeling, and classification modeling are just a few of the techniques that data analysts can put to use in order to draw conclusions and make further decisions. Through looking at the history, it is apparent that these methods were not formed for one specific industry. Collaborations and innovations advanced from astronomers, biologists, geographers, nurses, psychologists, and more. When the average person considers data analytics, their first assumption might be that it simply serves large technology companies and for-profit

corporations working with copious amounts of data. However, by thoroughly understanding the history of data analysis, it is clear that these systems benefit and contribute to a wide variety of professional fields. The public sector's use of data analysis, policy modeling, and program evaluation techniques has its own unique history that is worth recognizing.

### *Integration of Data Analysis and Public Policy*

Policy modeling as a tool in public planning has existed since the 1950's. However, public officials employed a much different approach rather than utilizing the data analytical techniques that were being systematized around the same time. In fact, policy modeling was mostly based off of economic models. At the time, the focus was simply to increase capital as much as possible, and policy was not people oriented. According to Estrada & Yap (2013), Antonio Maria Costa, Dominick Salvatore, and Douglas O. Walker, participants of the United Nations, came together in the 1970s to refocus policy modeling to make it less math-centered and more focused for social scientists (p. 170). This was perfect timing, as policy modeling was able to develop in alliance with the large advancements in computer technology. Policy analysts were now able to store, access, and clean manifold amounts of data in record time.

### *The Future of Policy Modeling*

Organizations and government-run programs continue to utilize data analytic techniques to construct, evaluate, and present their policies at every level. Decision making and problem solving have never been more efficient or effective. With that being said, there are still countless improvements that are expected to result from even more advancements in technology. Pirog (2014) predicts that access to public datasets will be one of the major progressions in policy modeling (p. 538). The push to democratize data and keep it open to anyone who would like to access it has led to more accurate and precise analyses. Large international organizations such as

the World Health Organization (WHO) and the World Bank have searchable and downloadable datasets open for anyone to use. The United States also has many similar resources. The U.S. Census Bureau, the country's largest data collection effort, keeps its data publicly available. Data.gov is also a popular resource, which includes datasets from federal, state, and local government initiatives (Patel, 2019). Policy analysts can take this open data and match characteristics based off of demographic, geospatial, and social aspects. This not only saves smaller organizations time and money, but it also creates for more precise policy interventions.

Another major advancement in public policy modeling has been fueled by the development of machine learning and classification modeling techniques. The traditional process of creating a program involved identifying members of a community who were in need of assistance and then formulating the treatment based off of them. However, with classification models, social scientists can predict who will need the service or solution before they even show signs of it. Amarasinghe et al. (2020) state that, "In an early warning system, the ML [Machine Learning] model is used to identify entities… for some intervention, based on a predicted risk of some (often adverse) outcome…" (p. 2). This technology can ensure that our communities are being effectively served before the issue creates a long-term and perpetual outcome. Machine learning not only has the power to pinpoint members of the public who might need the program, but it can also predict when the actual program itself might become a problem. For example, the University of Chicago Center for Data Science and Public Policy (DSaPP) has created an Early Intervention System (EIS) to identify police officers who are predicted to have an "adverse incident" in the near future (Ackermann et al., 2018, p. 15). By utilizing a form of decision tree classification modeling called random forests, the team has been able to successfully create an algorithm that will hold those in power accountable.

### *Ethical Considerations*

With access to copious amounts of data and the ability to directly influence major policy decisions, it is imperative that program and policy analysts tread with care and responsibility. The most common perception of the dangers of artificial intelligence derives from the media depiction of artificial intelligence robots becoming more powerful than humans. While this is fairly dramatic and certainly not an immediate concern, there is the real possibility that the human-created artificial intelligence technologies continue to represent racist and patriarchal biases that programmers and analysts themselves hold. Unconscious biases and lack of representation run the risk of magnifying the issues that most public policy works to deter. Rodolfa et al. (2021) explain that when utilizing machine learning techniques, creators should be aware of, "…potential sources and mitigation strategies for biases, including the underlying data, labels, model training, and post-modeling adjustments to scores" (p. 2). This might appear daunting, but many studies have shown that when designers consciously consider equitability and fairness goals, the likelihood of an algorithm working in favor of every member of the community rises substantially.

## Case Studies

By examining the manner that data analytics is currently implemented in public policy and program evaluation, it is easy to see that there are a plentitude of exciting and innovative projects taking place. The fluid use of data technologies, sharing of open data, and the development of predictive algorithms have allowed for organizations of all sizes and locations to effectively utilize data to further benefit their constituents. In order to fully understand the scope of public policy analytics, we can take a deep dive into three case studies. The following section will focus upon housing policies at the federal, state, and local levels. The U.S. Department of

Housing and Urban Development (HUD), the Illinois Housing Development, and the UChicago Urban Labs all create policy that affects housing in the city of Chicago.

***U.S. Department of Housing and Urban Development -- Federal Public Policy***

The United States Department of Housing and Urban Development (HUD) is a federally run program designed to provide housing assistance and solutions for American citizens. It became a cabinet department in 1965 and has been producing policy ever since (HUD.gov, n.d.). The HUD provides housing and rental assistance and solutions, while also continuously utilizing data for research and innovations. They undertake multiple surveys annually, and they maintain datasets concerning operations and planning as well as program participation. A majority of their data is openly listed on their website, and qualified researchers and the public have access to download CSVs along with their corresponding Data Dictionaries.

One of their main and most popular initiatives is the Low-Income Housing Tax Credit (LIHTC). The HUD describes this program by stating that they give, "…the equivalent of approximately $8 billion in annual budget authority to issue tax credits for the acquisition, rehabilitation, or new construction of rental housing targeted to lower-income households" (HUD.gov, n.d.). Housing developers are able to use the Tax Credits awarded by the program to fund their housing projects. The program combines both public and private interests, as major investors are encouraged to back the improvement of low-income neighborhoods. The LIHTC has certain requirements for where housing can be built as well as who can live in the housing. They maintain their datasets on property, tenant, and difficult development areas. All of these datasets are aligned with the U.S. Census Bureau, so naming conventions as well as the division of neighborhoods by census tract are implemented. The LIHTC website even includes a Query

Tool so that policy analysts and the public can analyze the data in real time via their open-source format.

***Illinois Housing Development -- State Public Policy***

Only a couple of years after the HUD was initiated into the executive cabinet, the state of Illinois created their Illinois Housing Development Authority (IHDA) in 1967. Similar to the HUD, the goal of the organization is to provide affordable housing, specifically in Illinois. The IHDA cites that, "Since its creation, IHDA has provided more than $20 billion to finance more than a quarter million affordable homes" (IHDA, n.d.). It is clear that the scope of this state-run operation is smaller than the federally backed HUD. Even with this considered, the IHDA has continued to effectively provide solutions for Illinois residents to own homes through mortgage loans as well as rental assistance. Their efforts are funded by their own Illinois Affordable Housing Tax Credits, and they even are able to allocate the HUD's Low-Income Housing Tax Credits to their own programs.

The IHDA's main goals are to provide community and neighborhood growth and development. Similar to the data.gov website that launched in 2009, the state of Illinois has a data.illinois.gov website that stores and lists datasets for their 81 different state organizations. Researchers and the public can easily download large CSV files containing information about their programs, rental/ownership activity, closings, tax credits, and more. One of their largest data collection techniques is their issuing of the Affordable Rental Unit Survey (ARUS) that utilizes U.S. Census Bureau tracts and terminology to identify where affordable rentals are located. The ARUS survey is programmed in an open-source mapping tool, where anyone can search a choropleth map of the state of Illinois.

### *UChicago Urban Labs -- Local Public Policy*

Federal and state policy programs have access to advanced technology and large budgets, which allow for expansive data collection and data analyses efforts. Local organizations, such as the UChicago Urban Labs initiative, do not necessarily operate in the same manner with the same resources. The lab offers five different focuses: the crime lab, education lab, energy & environment lab, health lab, and inclusive economy lab (UChicago Urban Labs, n.d.). The inclusive economy lab particularly focuses on utilizing data science and public policy to ensure that all citizens of the city of Chicago have access to equitable opportunity and resources. The lab will often work with organizations of similar size to issue surveys and conduct analyses. Their analysts will create dynamic data visualizations and run models based off of the models to present their findings in reports. These reports are uploaded onto their website for the public to see, and they are also utilized to present their findings to major policy decision makers and leaders in the Chicago area. The UChicago Urban Labs have broader areas of study to focus on, and they have a bit of a more unique approach to conducting and displaying analyses.

## Discussion

The objective of this research was to comprehensively understand how the vast field of public data analytics was formed and how it currently operates. By considering the history of the technologies so often used today, policy analysts can better interpret these tools to make their programs as effective as possible. Understanding the history of data visualizations, regression modeling, and classification modeling, we can see that these methods influenced policy decisions even before they were officially integrated into the policy modeling field in the 1970's. A key example of this can be seen with Florence Nightingale and Dr. John Snow's utilizations of data

visualizations and regression modeling to communicate the spread and effects of disease. Their persuasive application of data made significant changes and saved countless lives. We can see these same techniques at a more expansive and refined level with the manner that public health officials and epidemiologists have utilized data analysis to track and communicate the spread of the COVID-19 pandemic as well.

The past functions of data analysis also effectively convey how there are even more possibilities for improvements and innovations in the future. While there is substantive work dating back to the 1800's, the most technological advances have occurred in only the past half century or so. After all, computers have experienced a majority of their major improvements in the past 30 years. The future of data analysis involves upgraded open-source software and interactive options for policy analysts and other stakeholders to easily utilize. Federal, state, and local organizations with similar goals will be able to effectively collaborate and create consistent analyses. Machine learning developments will also create new opportunity as policy can target individuals with a solution before they are even aware that they need it.

While the future of public policy data analysis is certainly exciting, it is imperative that developers and researchers remain conscious about relying too much on technology. People should continue to be kept at the forefront of the field's mind as programs should be as equitable and unbiased as possible. It is important that all algorithms that directly contribute to policy choices truly serve their purpose of working for the public. The ethics of data analysis will continue to be a consideration that analysts will have to make through their work. With that considered, when visualizations, regression models, and classification models are utilized with accessibility and representation in mind, there is an opportunity for exceptional improvements in our societies.

## Conclusion

Data analytics and policy modeling are currently at a pivotal moment in their application to public policy. Organizations and their constituents have never had more access to large amounts of clean, organized, and detailed data. As computers continue to advance, the field will be able to create powerful policy decisions that will lead to beneficial change. Our look at the history of analytics shows that leaders in many industries have used data technologies to make improvements in their lines of study. Data visualizations, regression models, and classification models are only a few tools that have developed and blossomed in conjunction with the development of personal computers and the Internet. The software that exists today will become even more functional and convenient for nearly any person with a computer to use. We can see that many organizations, including the U.S. Department of Housing and Urban Development (HUD), the Illinois Housing Development, and the UChicago Urban Labs actively utilize and create public datasets and models in open formats. As the ethics of sharing and accessing data continue to be considered, public data will become even more codified and consistently organized. The future is bright for program evaluation and policy analytics.

## References

Ackermann, K., Walsh, J., De Unanue, A., Naveed, H., Navarette Rivera, A., Lee, S.-J., Bennett, J., Defoe, M., Cody, C., Haynes, L., & Ghani, R. (2018). Deploying Machine Learning Models for Public Policy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* https://doi.org/10.1145/3219819.3219911

Ajani, K., Lee, E., Xiong, C., Nussbaumer Knaflic, C., Kemper, W., & Franconeri, S. (2021). Declutter and Focus: Empirically Evaluating Design Guidelines for Effective Data Communication. *IEEE Transactions on Visualization and Computer Graphics*, 1-1. https://doi.org/10.1109/tvcg,2021.3068337

Amarasinghe, K., Rodolfa, K. T., Lamba, H., & Ghani, R. (2021). Explainable Machine Learning for Public Policy: Use Cases, Gaps, and Research Directions. *Cornell University*. https://doi.org/https://arxiv.org/pdf/2010.14374.

Ariew, A., Rohwer, Y., & Rice, C. (2017). Galton, reversion, and the quincunx: The rise of statistical explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 66*, 63-72. https://doi.org/10.1016/j.shpsc.2017.08.001.

Aspray, W. (2019). *Historical studies in computing, information, and society: insights from the Flatiron lectures*. Springer.

Azzam, T., Evergreen, S., Germuth, A. A., & Kistler, S. J. (2013). Data Visualization and Evaluation. *New Directions for Evaluation, 2013*(139), 7-32. https://doi.org/10.1002/ev.20065

Boog, J. (2021, January 4). *10 Best Machine Machine Learning Software [2021 List]*. The QA Lead. https://thequalead.com/tools/machine-learning-software/.

Dunn, W. N. (2016). *Public policy analysis: an introduction.* Routledge.

Estrada, M. A., & Yap, S. F. (2013). The origins and evolution of policy modeling. *Journal of Policy Modeling, 35*(1), 170-182. https://doi.org/10.1016/j.jpolmod.2011.12.003

Evergreen, S., & Metzner, C. (2013). Design Principles for Data Visualization in Evaluation. *New Directions for Evaluation, 2013*(140), 5-20. https://doi.org/10.1002/ev.20071

Fienberg, S. E. (2011). Bayesian Models and Methods in Public Policy and Government Settings. *Statistical Science, 26*(2). https://doi.org/10.1214/10-sts331

Fradkov, A. L. (2020). Early History of Machine Learning. *IFAC-PapersOnLine, 53*(2), 1385-1390. https://doi.org/10.1016/j.ifacol.2020.12.1888

Freedman, D. (1999). From association to causation: some remarks on the history of statistics. *Statistical Science, 14*(3). https://doi.org/10.1214/ss/1009212409

Freedman, D. A., Collier, D., Sekhon, J. S., & Stark, P.B. (2010). *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press.

Friendly, M. (2008). A Brief History of Data Visualization. In *Handbook of Data Visualization* 9pp. 16-48). essay, Springer International Publishing.

Hedley, A. (2020). Florence Nightingale and Victorian data visualisation. *Significance, 17*(2), 26-30. https://doi.org/10.1111/1740-9713.01376

Hoaglin, D. C. (2003). John W. Tukey and Data Analysis. *Statistical Science, 18*(3). https://doi.org/10.1214/ss/1076102418

*HUD.gov / U.S. Department of Housing and Urban Development (HUD)*. HUD. (n.d.). https://www.hud.gov/.

*Illinois Housing Development Authority*. IHDA. (n.d.). https://www.idha.org/.

Lecy, J., & Fusi, F. (n.d.). *Foundations of Program Evaluation: Regression Tools for Impact Analysis*. Foundations of Program Evaluation: Regression Tools for Impact Analysis. https://ds4ps.org/pe4ps-textbook/docs/index/html.

Marr, B. (2017, July 20). *The 7 Best Data Visualization Tools Available Today*. Forbes. https://www.forbes.com/sites/bernardmarr/2017/07/20/the-7-best-data-visualization-tools-in-2017/$sh=2c2f1f2b6c30.

Patel, H. (2019, January 10). *These Are The Best Free Open Data Sources Anyone Can Use*. freeCodeCamp.org. https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/.

Pirog, M. A. (2014). Data Will Drive Innovation in Public Policy and Management Research in the Next Decade. *Journal of Policy Analysis and Management, 33*(2), 537-543.

Rodolfa, K. T., Lamba, H., & Ghani, R. (2021, May 7). *Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy.* https://arxiv.org/pdf/2021.02972.pdf.

Rosenblatt, F. (1960). Perceptron Simulation Experiments. *Proceedings of the IRE, 48*(3), 301-309. https://doi.org/10.1109/jrproc.1960.287598

Stanton, T. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education, 9*(3). https://doi.org/10.1080/10691898.2001.11910537

Stobierski, T. (2021, May 14). *What is Statistical Modeling for Data Analysis?* Northeastern University Graduate Programs. https://www.northeastern.edu/graduate/blog/statistical-modeling-for-data-analysis/.

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, *33*(1), 1-67. https://doi.org/10.1214/aoms/1177704711

Turner, J., & Schomberg, J. (2016, June 29). *Inclusivity, Gestalt Principles, and Plain Language in Document Design*. In the Library with the Lead Pipe. https://www.inthelibrarywiththeleadpipe.org/2016/accessibility.

UChicago Urban Labs. (n.d.). https://urbanlabs.uchicago.edu/.

Walker, F. A., & Bien, J. (1874). *Statistical atlas of the United States based on the results of the ninth census 1870: with contributions from many eminent men of science and several departments of the government*. map, New York, NY; United States Census Office.