APRIL 2020

Data Blending:

Haven’t We Been Doing This for Years?

MDI Report

AUTHORS

Lisa Singh, Georgetown University
Michael Traugott, University of Michigan
Leticia Bode, Georgetown University
Ceren Budak, University of Michigan
Pamela E. Davis-Kean, University of Michigan
Ramanathan Guha, Google
Jonathan Ladd, Georgetown University
Zeina Mneimneh, University of Michigan
Quynh Nguyen, University of Maryland
Josh Pasek, University of Michigan
Trivellore Raghunathan, University of Michigan
Rebecca Ryan, Georgetown University
Stuart Soroka, University of Michigan
Laila Wahedi, Georgetown University (now at Facebook)

MDI
GEORGETOWN UNIVERSITY
MASSIVE DATA INSTITUTE
McCourt School of Public Policy

Data Blending – Motivation

Data blending is the process of associating a variety of organic and observational data with more traditional administrative and survey data. It combines data obtained from designed studies with more ad hoc, non-designed, and/or organic (behavioral and digital trace data shared at scale) data sources. While data blending is currently resonant as "Big Data" gains prevalence in social science research, it is important to note that data blending is not new. Many different disciplines have been *blending, merging, integrating, harmonizing, triangulating, linking,* and *joining* data for a long time. The practice describes a process by which data are combined from multiple types of primary or secondary source material such as news reports, records prepared by administrative institutions, economic data generated by industry, documentation provided by non-governmental or not-for-profit organizations, social science surveys, observational studies, experiments, clinical data, and the like to produce research datasets for analysis.

Given the long history of data blending, why is the topic resurfacing now? We believe there are two reasons: the accessibility of new forms of organic data, and the advantages of using these combined data to accelerate our understanding of phenomena that are difficult to understand. Contemporary data blending typically combines the methods and analytical skills of the social sciences with those of computer science and data science to conduct new forms of analysis and to produce new understandings of social phenomena. This bridging of disciplines is advantageous because data exist at varying granularities, and varying levels of quality, requiring new methods that are robust to these variations. It further requires an understanding of when new predictive methods or machine learning algorithms are applicable and reliable enough for specific research objectives.

New Forms of Organic Data

More and more unstructured, organic data related to human behavior, beliefs, and opinions are being shared online. Because of their availability, richness, and "real time" nature, these data constitute an important new source of information for social scientists and policymakers who want to characterize and predict human and societal dynamics. Blended data offer insights that traditional surveys and behavioral data can miss; however, they can be unwieldy to deal with in terms of volume, reliability, validity, and representativeness.

Historically, many of these organic forms of data required extensive human preparation, coding, and processing. They were also limited in scope, with respect to the space-time window they covered and the units of analysis at which the information was available and then aggregated. In recent years, new forms of computer-readable data are being evaluated to better understand the opportunities for blending them with more traditional types of data using automated methods. Among these new types of data sources are social media data, information on online usage and behavior (e.g. Google searches, profile clicks), mobile device data, data collected by health or home devices, satellite imagery, surveillance data, and reports by citizen journalists. What most of these data have in common is that they are often generated in an automated or computer-aided fashion and provide information on a massive scale (i.e. "Big Data").

New Insights into Old Problems

Blended data offer the possibility of increasingly *holistic* approaches to explain and predict a wide range of social, health-related, and political phenomena. Theories about our social world have traditionally been constrained to variables that are relatively proximate to each other, either logistically or theoretically speaking. We may think of voting behavior as being predicted by certain attitudes, as well as by political-contextual variables (e.g. electoral systems or administrative regulations) or even by the weather. We may think of successful physiotherapy for a knee as being dependent on other parts of a leg, the back, or a person's weight. We have focused our energy on those variables that are likely to be most significant. Now, with the availability of these new forms of data and automated ways of connecting them, we can easily test more distal relationships – and we can begin to accumulate more experience identifying what those distal relationships might be. For example, using lifestyle related metrics from Internet of Things (IoT) devices may provide additional insight into successful physiotherapy for a knee.

Blended data, in a research environment in which data availability is ever-expanding, offers an opportunity to broaden the focus of both our empirical testing and our theorizing. Health outcomes may be affected by environmental variables — from the weather to the air to whether one's commute involves a lot of hills. Political outcomes may be affected by the combined effects of exposure to many different media sources, by one's work environment, or by how much sleep one gets. Our new theories are not just driven by researchers, but also by data producers. And it is not just the new data we have access to, but also the new methods that are being developed to identify some of these new relationships at scale. This is a new frontier that we are embarking on. Exploratory methods using increasingly blended data may suggest connections that we have not yet thought of. This expansion of the scope of our empirical and theoretical investigations is perhaps the most consequential promise of blended data.

Data Blending Challenges

Data digitization and increased availability have the potential to democratize data science. These trends allow diverse sources of data to be blended and used by a heterogenous community of researchers. Yet, this broader access also introduces important challenges. An analogy can be drawn to new software solutions that make data science more accessible. The use of statistical software libraries (e.g. scikit-learn in Python, gamer in R) allow researchers to employ complex statistical methods using only a few lines of code. But the abstraction of the statistical methods internal to such code can obscure the assumptions embedded in their algorithms, making it easier to draw incorrect inferences.

The overarching challenge that arises from these new types of data is to assess their quality and determine how they can be effectively blended with more traditional forms of data to inform and advance scientific knowledge (whether by descriptive, inferential, or causal methods). This assessment of quality is non-trivial and manifests itself in different ways. For example, unique biases are associated with different data sets. What do these biases mean and how do they impact the quality of individual measures in the data? What specific constructs do these data measure? How do such biases propagate as growing numbers of data sets are blended together or as variables are combined into indexes and scales? How should such information be revealed without overburdening the researcher? Quality issues can arise when linking data from a non-random sample. The blending process may reduce the external validity of the conclusions that can be drawn from the blended data. Another quality issue arises as more of our input data values come from outputs of machine learning algorithms, e.g. estimated or inferred demographics of individuals. When those outputs become inputs for other predictions, researchers regularly use point estimates without factoring in prediction confidence measures in the calculation of additional error from the input sources. How should we be calculating those elements so that there is a consistent error calibration across different types of data sets? What does it mean to researchers that certain inputs are generated from machine learning algorithms when they make quality estimates of such factors as the reliability and validity of those measures?

Another challenge is the lack of understanding about the generation process of these data. Large scale data hosting platforms like government websites (e.g. data.gov), academic data portals (e.g. Harvard's Dataverse), and open access data exchanges (e.g. Humanitarian Data Exchange), can lead people to use datasets without providing insight into the process by which the data were created—if such data hosting systems are not designed properly or if the platform creators are not forthcoming and transparent. Also, it is usually unclear how representative the datasets are of a specific population. Even though big data are large, potentially containing millions of observations, we often do not understand the generation process sufficiently to generalize findings to a broader population of units of analysis. What types of information should be provided about the process of generating the resulting datasets? Who is choosing to participate and who is not? What sampling strategies make sense for different types of big data in order to reduce them to manageable analytic size?

A challenge that is very specific to newer organic datasets is the proliferation of misinformation, disinformation, and other poor-quality information. Misinformation is generally characterized as the unintentional generation of incorrect information, while disinformation is the deliberate posting of falsehoods. This is still different from bots that generate misleading quality or quantity information. Whether humans or bots, the goal of those posting poor quality information is to influence opinions and beliefs. How do we identify and account for both unintentional and deliberate misinformation when blending multiple organic sources? Are there reliable ways to measure the impact of misinformation on public opinion and beliefs when doing so?

Although there are many other challenges, the final one we mention is an ethical consideration. Researchers have access to more public data at the individual level than ever before. Some researchers have begun using this potentially identifying data without the consent of the subjects. Moreover, even if consent is given by individuals, under conditions where the original investigator promises confidentiality and anonymity, additional inferences about individuals' identities may be possible once multiple data sources are combined. Is this ethical? Were the individuals in the separate studies aware of these possibilities after datasets might be combined? Were the principal investigators who collected the original datasets even aware of these possibilities? When people publicly share their thoughts and opinions online, should we extend the use of these data without their explicit consent? In what situations could it be reasonable?

Even though new abilities to generate and blend data overcome several limitations inherent in merging multiple sources of data, they also introduce new challenges that were not part of the data blending frames of the past. These challenges will need to be addressed as we continue to move forward with combining data from different designed and organic data sources.

Objective

The remainder of this white paper provides a brief overview of both old and new versions of data blending as a way of illuminating the differences between the two. We do this by sharing examples of different domains in which new forms of data blending have been successfully used. We then outline methodological, computational, and ethical considerations when blending different forms of data. Finally, we discuss best practices. Our hope is that this white paper will serve as a reference guide highlighting research potential and benefits, while pointing out methodological and ethical challenges.

Data Blending – Old and New

To illustrate the challenges and opportunities of data blending, we begin by sharing examples of different domains with a history of data blending, how new forms of data blending in these domains have been successfully used, and how these forms differ from previous data blending in the field.

Well-Being

In order to better understand the well-being of people in different populations, researchers need to identify, measure, and understand the conditions and policies that lead to healthy populations. The number of health determinants are large but can be categorized as follows: genes and biology, health behaviors, social or societal characteristics, and health services (CDC, 2014). There are many sources of health information, and no single data source is ideal for studying all factors related to well-being. Federal and state government officials have spent the last decade putting these data online and making them available for both individual analysis and data blending across different organizations. The data vary, from information about individuals participating in different local, state and federal programs/initiatives, to surveys about different socioeconomic, behavioral, and interaction factors that relate to well-being, to data about cases and the spread of different viruses, diseases, and toxic substances, to health claims data that identify where and when different negative health outcomes are occurring. Most of these examples of data fall into four categories: administrative data, biological/genetic, clinical trial data, and survey data. Some classic examples of the relationship between policy and well-being that have been studied for years include tobacco warning labels and television ads featuring smokers with lung cancer as ways to promote quitting smoking, vaccinations and health screening to proactively prevent illness, seatbelt and smoking restrictions laws to make the default a healthy choice, and increasing the availability of preschool education to increase education equity across socioeconomic classes (Frieden, 2010).

Data blending gives us the opportunity to integrate exposures from the environment with diagnostics of the disease, including the interplay of psychological and biological issues, in order to get a more holistic perspective of the relationship among determinants of well-being. Data blending in the arena has been used to extend coverage of survey data, bridge transitions in reporting, and improve analyzes (Schenker and Raghunathan, 2007). A well-known example of data blending in public health is the National Longitudinal Study of Adolescent to Adult Health (Add Health), a longitudinal study of adolescents in grades 7 to 12 from 1994 to 2008 (Harris, 2013). This collection effort incorporated different forms of data, including school and in-home surveys of adolescents, parents, siblings, romantic partners, and school administrators, as well as biomarkers, spatial data from GPS devices, and community data on poverty, unemployment, availability and utilization of health services, crime, church membership, and social programs and policies. The initial study design of these data was a mixed methods approach to data collection for four waves. This classic example of data blending by design has led to 1000s of publications, some of which relate to specific areas of well-being, including mental health (Hargrove et al., 2020; Amin et al., 2019), peer effects (Bond, 2019), neighborhood inequity (Brazil, 2019), violent traits (Markowitz et al., 2015), and substance usage (Deutsch et al., 2015).

A more recent study involves blending two survey data sources (the Medicare Current Beneficiary Survey and the National Health and Nutrition Examination Survey) and administrative data from three sources to study the prevalence of specific health conditions and costs of treatment, thereby improving our understanding of national health spending patterns (Cutler et al., 2019; Raghunathan et al., 2020). The goal of the project was to explain the potential reasons for the slowdown in Medicare spending beginning in 2009. The major conclusion was that an increased use of pharmacological treatment for cardiovascular risk factors may explain a decrease in cardiovascular diseases, and hence the slowdown in medical spending. Both the Add Health project and the Cutler et al. study highlight the prevalence of data blending in well-being, and more broadly public health, research.

Newer studies are beginning to use machine learning to construct variables from available organic data. For example, a recent study used Google satellite imagery from the Street View Image API to create indicators of environmental characteristics, including greenness and building type (Nguyen et al., 2018). The authors built a computer vision model to automatically annotate the images and then used these variables to estimate the association between neighborhood environmental features and the prevalence of obesity and diabetes. In addition to these new data allowing public health officials to better understand neighborhood quality at the regional and national level, combining these data with existing environmental and socioeconomic indicators could further improve our understanding of how different neighborhood conditions influence different disease outcomes.

The complexity of data blending for studies of health and well-being means that many analyses require varying levels of data, as well as data generated in different ways about different subjects employed in the same analyses. While data blending is part of the tradition in public health research, the availability of new forms of spatial and social data may allow for insights that are not as easy to capture in existing surveys or administrative data. Also, public health researchers have access to large-scale data not currently being utilized to help make decisions. To take advantage of both traditional and newer forms of available data, large-scale data processing practices for public health need to be standardized and democratized, i.e. shared methods, resources and training for everyone to use. Otherwise, utilization of available data will not increase. This in turn would lead to less evidence or more incomplete evidence when evaluating different public policy alternatives.

Voting Behavior

Voting is the essential form of political participation in a democracy, as it produces either elected leaders or new policies through referenda. Social scientists have traditionally studied it as an individual behavior, through survey research and in aggregated form through election returns at many different levels of analysis. Analysts have explained variations in voting with individual demographic characteristics or various attitudes, elements of administrative arrangements related to registration or polling procedures, or cross nationally with variations in such arrangements as legislative district characteristics or levels of competition between parties or coalitions. The latter type of work employs multilevel statistical models.

One recurring focus of this research is the impact of weather on turnout, which illustrates the difference in the way data have traditionally been used and the possibilities of new data resources to increase our understanding. In the broadest theoretical sense, Downs (1957) wrote about the costs and convenience of voting; his work suggested that bad weather could hinder turnout because it made it more costly in terms of time and effort to get to the polls. Weather conditions were an appealing set of variables in turnout models because they were so clearly exogenous to that behavior, creating causal leverage.

In the earliest studies, researchers used county-level and state-level data as the units of analysis (Andrews, 1966). For the dependent variable, they had the number of voters and a turnout rate, as well as the partisan division of the vote. Because the range of the geographical area of units for counting votes was so large, the weather (inches of rain) was an average measure across the geographical unit or taken from a central weather station in the county. The results of this early research examining the relationship between the weather and turnout were mixed, with some indicating turnout went down when it rained and others suggesting no effect. Some research suggested turnout increased when temperatures went up (Van Assche et al., 2017). This research did highlight the possibility of alternative turnout effects, such as linking a party's success to its current control of offices. For example, increased turnout could reflect anti-incumbent sentiment generally (Hansford & Gomez, 2010). There was also an interest in whether increased or decreased turnout helped one party or the other – and whether decreased turnout under poor weather conditions had the same or a different effect.

Later the focus shifted to individual-level data from the American National Election Study (ANES) which conducted surveys in presidential election years before and after election day. Focusing on voting as an individual decision and using measures of partisanship, personal efficacy, and demographic characteristics, the ANES measured intention to vote before the election and whether a respondent reported voting in a post-election interview. These survey records could be merged with weather data in or near the place the respondents lived (Knack, 1994). Again, the results were inconclusive, partly as a function of the reliability of the self-reports of whether a survey respondent had actually voted and for whom, as well as a lack of information about whether or not voting was habitual for that person. At the same time, the measure of precipitation from weather reports could not be linked directly to the location of an individual's polling place or the time of day they voted.

Most recently, the administration of some pre-election polls has turned to samples of individuals drawn from registered voter rolls (Kennedy et al. 2018). While for whom a person voted remains confidential, whether the individual is registered and voted in a particular election is a matter of public record. This information, available over time, reduces the bias in self-reports that consistently inflates turnout rates and also makes it easy to identify whether a respondent is a habitual voter, an intermittent or casual voter, or a nonvoter. Because these records typically contain the name, address, and age of each registered voter, additional information about them can be merged from other administrative data or commercially available data with the same identifiers.

Some are now using surveys (Stewart & Ansolabehere, 2015), Twitter (Mebane et al., 2017), and direct observation (Stein et al., 2019) to gather data about incidents at polling places and the length of time it takes to vote in a precinct. This information is important for understanding whether the costs of voting, in terms of time and convenience, are equitably distributed in a jurisdiction, a state, or a nation. This research suggests they are not, with the experiences of Milwaukee voters in the 2020 primary election, when 180 polling places were reduced to five, being just the latest example (McCormack, 2020).

What will the blended dataset of the future look like and what kind of research will it support? It could begin with individual voting records purchased from a county clerk or a secretary of state that contain personal information, geographical indicators of residence, and polling place location. This could be further merged with other administrative files or commercially available data, e.g. occupation and income range. The geographical data could be linked with their precinct's voting returns over time and in the current election, as well as the readings from the neighborhood Weather Underground (2020) recording station nearest to their home or their polling place. These data could be enhanced with geo-tracking of the voter's travels on election day to indicate what time of day they voted, how far they traveled to the polling place, and how much time they spent there[1]. And the possibility of adding social media data from the individual and others in their social, political or work networks would further enhance the analysis. Indeed, existing research has already linked voter registration and Twitter data to examine voter behavior online (Grinberg et al., 2019). Integrating all these forms of independent, multi-resolution data require the development of standard practices and procedures that consider the different models of the data and the difference contexts associated with the data collection. This in turn will lead to a better understanding of fraud, inequity and inconsistencies associated with voting and elections, evidence that is beneficial for developing public policy for maintaining a healthy democracy.

Parenting and Child Development

In the last two decades, scholars across the social sciences have increasingly sought to understand how children develop within multiple contexts, with a particular focus on the role of the family background or parenting environment (Collins et al., 2000; Cunha, 2015). For example, researchers have found that poverty and food insecurity predict higher levels of internalizing and externalizing behavioral problems in children and lower academic test scores. Studies of family contexts have found that characteristics like engaged parenting and positive behavioral control predict better child health outcomes, such as lower rates of respiratory illness (Serbin et al., 2014; Gurney et al., 2006). Public policies like the 2002 No Child Left Behind Act initially increased student engagement (which predicts both achievement and socio-emotional well-being), but over time these effects ultimately reversed (Markowitz, 2018).

One classic example that blends multiple sources of data looks at the effects of neighborhood on children (Chetty et al., 2016). These economists took a randomized control trial (RCT) funded by the Department of Housing and Urban Development and designed by developmental psychologists and sociologists to examine the impact of moving families from housing projects/poor neighborhoods to mixed income neighborhoods through housing vouchers. The data were initially collected using a survey. Then Chetty et al. blended the survey/RCT data with tax data from the IRS to look at long-term impacts of moving on college attendance, earnings, and single parenthood rates.

While there are numerous examples of blending data in the child development literature, researchers have focused on combining data from multiple large-scale surveys or combining survey data with administrative data. Nearly all the information scholars have about parenting comes from large-scale surveys or small-scale direct observational studies where the data are

[1] It would also be necessary to account for the fact that 25% of the votes cast in the 2016 presidential election were cast by mail.

useful, but potentially biased (Sleap & Warburton, 1996). Specifically, self-reported data on parenting can suffer from social desirability bias, as parents aim to present ideal versions of their parenting to researchers. At the same time, observational data, though more objective, are labor-intensive to record and encode, and can be conducted only in small samples that often lack generalizability. Studies that use survey or observational data also focus on topics of interest to the scientific community, not necessarily topics that parents themselves find most relevant to their lives.

New forms of data, including data from social media, by contrast, offer a source of information about parents' interests and behaviors potentially without social desirability bias, but that also reflects the everyday lives of a large (though non-representative) number of parents and families. Three-quarters of parents in the U.S. use some form of social media, and the vast majority of them receive support and information on parenting issues through social media (Duggan et al., 2015). To identify the relevant content within the social media landscape in ways that are useful for analysis, different types of data can be collected. Examples include: tweets associated with particular accounts that are important to this community (hubs and/or authorities), e.g. parenting-related Twitter accounts like Parenting magazine and BabyCenter; randomly sampled tweets from followers of the hubs/authorities; tweets associated with a random set of parents; Reddit groups discussing parenting topics; Facebook groups discussing parenting topics; and search term trends from Google. These data provide potentially novel insights into parents' behavior, attitudes and beliefs that are plausibly more grounded in their everyday lives than responses to survey questions or behaviors identified during a structured laboratory task (Ryan et al., 2020). Understanding the relationship between parenting strategies/behaviors and child outcomes can help identify areas where government programs and public policy can benefit communities.

COVID-19

Given the current COVID-19 pandemic, there are some novel approaches to using blended data to improve our understanding of the impact of the disease. We focus our discussion on blending social media data with survey data. We cite examples using Twitter data as it is the most commonly available social media source, although there are other platforms that also might be useful. The general approach is to employ an interrupted time series design or analysis, analyzing tweets before, during, and after the pandemic to look for changes or shifts in trends.

There are several standard approaches to analyzing tweets, involving topic analysis, sentiment analysis, and network analysis. All these analyses presume or require an assembly of a sample of tweets over time, where the initial time is prior to the first COVID-19 case. Depending on a researcher's available computational infrastructure, sampling may be necessary because the actual volume of tweets per unit of observational time (day or hour, for example) is so great. There are many choices when it comes to sampling: identifying a random sample of the Twitter population, following a set of hashtags and/or keywords that are associated with the topic of interest, e.g. #COVID-19, or identifying individual accounts that serve as authorities on the topics of interest, e.g. epidemiologists, health care workers, policy makers, or individuals who have been diagnosed with the disease.

Analyses could be conducted to consider the volume of conversations and/or the variation in their content. For example, preliminary findings show that both geotagged volume and location conversation volume are highly correlated with new COVID-19 cases in different countries (Singh et al., 2020). This may indicate that a social indicator blended with more traditional health indicators may add insight, particularly in areas where traditional health indicators are less reliable. Another interesting possibility would be to analyze topics more generally. We would expect health-related concerns to become more frequent after the onset of the coronavirus. When sentiment is associated with these topics, the question is whether these tone indicators become indirect proxies of health factors over time. This inevitably raises the question of mental health status in the population.

Another design strategy would be to draw a sample of tweets, identify their source, and then construct the networks for those tweeters. Here again, researchers would employ a standard set of network analyses to look at network size and how it shifts over time, the number of edges in each, and the centrality of specific individuals in the networks. In these types of analyses, it is also possible to develop an understanding of how different content spreads – what topics are shared more frequently and what topics die out. We can then analyze misinformation to see who the influencers are and what segments of the population they are influencing.

Google searches linked to their geographical origin could also be used to assess physical health in the population and thereby map the rate of increase in COVID-19 cases by state over time (Stephens-Davidowitz, 2020). The author suggests two ways in which the searches might be used. In an ex post facto analysis, after reports of individuals saying that they lost their sense of smell before symptoms of the virus appeared, he looked for the number of searches on the phrase "I can't smell" over time and by state. The frequency of searches by day over time was an indicator of the relative rate of infections. The number of such searches was also strongly correlated with the number of positive cases of coronavirus. The results were replicated when the Italian phrase "non sento odori" was used as the key phrase in the searches. These results are similar to the findings of a previous study of early indicators of pancreatic cancer suggested by Google searches on such terms as "indigestion" or "abdominal pain" and could be linked to subsequent searches for the equivalent of "just diagnosed with pancreatic cancer" (Paparrizos et al., 2016).

If and when confidence in the social media measures can be established, then it should be possible to produce a continuous monitoring of the "validated" measures on a timely basis that is more frequent than the survey measures will appear. Of course, the validation efforts will have to be ongoing, as relationships can and do change over time (Lazer et al., 2014). But as confidence in the social media time series in the United States increases, it should be possible to extend the methodology to other countries where equivalent developmental work can be replicated.

Methodological Considerations for Data Blending

Traditional research bringing together data from multiple sources can take different forms. In general, these data combinations can be classified based on whether the data used to link sources comes from a single common level of observation (unit of analysis) or instead are combined across multiple levels of observation. Single-level data consist of cases where different kinds of information exist in different places about a particular unit and need to be merged to produce a new data record with the combined information. For example, medical records from one set of patients (either from doctors' visits or from hospitalizations) might need to be combined with survey responses from that same group of individuals. Multi-level data consist of cases where information from some set of units needs to be linked to information that either exists for subcomponents of those units or for larger aggregations of those units. For example, survey data might be aggregated to generate group-level estimates, and then group-level comparisons might be conducted. Or information about individual patients and their treatments for specific symptoms might be combined with diagnosis and treatment of all the patients who visited that same doctor or hospital. Alternatively, information about attributes of large geographical areas where people reside might be treated as valuable contextual information to predict individuals' attitudes and behaviors.

Challenges When Considering Single-Level Data

Linking data across a single level of observation is typically conducted using a matching process. This consists of assigning common case identifiers across datasets and supplementing information from some of those identifiers with data from another dataset. Often, when these types of merges are used to analyze social questions about individuals, they fall under ethical standards requiring that each individual subject consents to the linkage. Today, more secondary data are available and being used without explicit consent because the secondary analyst is remote from the individuals from whom the information was originally collected. Furthermore, the dataset which the researcher received for secondary analysis usually masks the identification of the individuals so the secondary analyst would have no way to contact individual subjects in order to obtain permission for the analyses she intends to perform. There have always been key questions about the conditions under which data can be linked across cases and when there is uncertainty about the linkage because of missing values or the quality of the secondary sources. Different methods for handling missing values have been proposed including probabilistic record linkage (Sayers et al., 2016; Hejblum et al., 2019) and multiple imputation (Sterne et al., 2009).

Today, this question is even more complex because different data that could be linked may not have been generated for the purposes they are being reused for when they are linked, e.g. using emotion or sentiment on social media to better understand well-being in different regions of the country. Computer scientists regularly construct variables and combine these noisy forms of data, but with a focus on algorithm design and development of inference methods. Little work is done in the area of calibration for different types of data sets or different types of social science research designs. While the generation process may be understood in some cases, it may not be as clear in other cases, making blending data from different sources that much more difficult. Finally, in the past, these types of samples were relatively small. Today they are regularly much larger because of the availability of administrative, corporate, and social media data, but also generally less representative.

Challenges When Considering Multi-Level Data

Linking data across different levels of observation typically requires a combination of multiple data sources that exist at a common level of analysis with other datasets that exist at alternative levels. This linking is typically performed by identifying the lowest common denominator in terms of data granularity - is it the individual, the family, the county, or something else? A researcher must then determine the relationship between the individual and aggregate levels of the data and specify the assumptions being made during this merging process. Typically, summary statistics accompany aggregated data, indicating how many of which units were combined to form what number of new units of analysis. However, new forms of metadata may also need to be part of this. For example, we may need to know about data provenance and parameters of algorithms used to construct variables. If these become part of the accepted standards when data are combined, we can more easily trace the lineage of original data variables and constructed ones.

The final scenario is linking across multiple independently generated datasets. This is typically done using multi-level modeling where data are hierarchically nested (Gelman et al., 2003). A strength of this approach is that there is clear parameterization and methods that are accepted as standards. Researchers know how to interpret them, e.g. Bayesian models, across disciplines. Unfortunately, there are several problems with some multilevel modeling approaches including strong assumptions of normality that can introduce large errors. Multilevel models are also not good at combining data that vary in both spatial and temporal resolutions. It is not unusual for researchers to merge without thinking about this mismatch, using data at different levels of analysis, repeating their observations and getting bad estimates as a result. A simple example of a spatial incompatibility involves zip codes. One data source might use ZIP Code Tabulation Areas (ZCTAs) to label individuals geographically, while another uses just ZIP Codes. These two schemes are similar, but they do not provide a one-to-one matching. A new variable needs to be constructed to align the data sets, or error measures need to be included when blending the data. Furthermore, decisions made by different researchers to handle this misalignment might vary—leading to inconsistencies across studies. There are still many questions surrounding the validity and reliability of multilevel modeling, and quantifying the measurement properties of different data sources. We need new methods to explore this more complex multidimensional space. Our arsenal is limited.

Data aggregation should be performed with caution. The inferences at the aggregate level can mask important distinctions between subgroups and sometimes even mislead researchers. One example way in which this happens is explained by Simpson's paradox (Simpson, 1951). For instance, while the examination of Twitter behavior might suggest that additional exposure by friends can suppress adoption (of a hashtag), researchers have shown that this finding goes away when the results are disaggregated by cognitive load (number of friends). Simpson's paradox is only one example of how individual (or subgroup) inferences can differ from group level inferences. This highlights caution that needs to be taken in multi-level data blending.

Other Challenges Specific to Social Media

There are a number of additional methodological challenges specific to using/blending social media data. The first concern stems from sample representativeness. Individuals on social media do not represent the entire U.S. population. For example, those who use Twitter differ demographically from those who use Facebook, no social media at all (Duggan et al., 2015), and differ from the general US population (Perrin & Anderson, 2019). Second, researchers do not know the demographics of social media users at the individual level. While these can be inferred by machine learning algorithms, measures of reliability are inconsistent (Cesare et al., 2019). However, cross-validating survey measures and social media indicators of the same underlying concept would be a promising way to deal with this (McClain et al., 2018), and might also provide the opportunity for more frequent measurement of indicators of interest, given that fielding surveys upon which we now rely is a major logistical effort, limiting the frequency with which they can be undertaken.

Another important methodological challenge is the differences in methods of analysis used by those working with different types of data (Stier et al., 2018). Researchers who use surveys typically use regression or structural modeling. Those who analyze social media data tend to incorporate machine learning, text mining, and/or network analysis. Understanding the strengths and limitations of these analysis methods, as well as important ways to extend them in the context of data blending is an important future direction.

Finally, analyzing organic data means social scientists need to draw them from existing sources with no control over the data generation process. Posing a research question that could be answered with social media data requires an understanding of how these data were generated and whether their design supports investigating the specific research question and its broad aim (i.e. is it descriptive, correlational, or experimental?). Trying to "fit" the data to traditional social science research questions, in contrast, is generally more problematic.

Computational and Algorithmic Considerations When Blending Data

Two major reasons for new forms of data blending are the plethora of new algorithms being developed for understanding human behavior, and the accessibility of large-scale compute infrastructure for data processing. While some disciplines, like public health, have been taking advantage of these trends for several years, some of the machine learning algorithms and statistical methods have not become as mainstream in other disciplines. Here we consider two areas of computational and algorithmic innovation that are helping change how data can now be analyzed.

Manual to Automated Coding

Traditionally, many of the variables that have been merged into other data sets have been human-coded - taking a certain type of content (text, audio, graphics, etc.) and applying a codebook to classify it into meaningful categories, facilitating quantitative analysis. These data in turn are often merged or blended with administrative variables or survey variables. Examples of coded data in political communications include media content from campaign coverage, open-ended survey responses to questions about candidates or issues in the campaign, the attributes of advertising including content and frequency of placements, and types of political systems in countries or at lower geopolitical levels (Bode et al., 2020; Soroka & Wlezien, 2010). Coding of this information has been done by trained research assistants as well as by experts on the topics. Researchers have traditionally tried to understand the biases in these coding methods, using measures of intercoder reliability or validating results from a small subsample of the data with other methods.

New software algorithms and inexpensive computing power provide automated ways of coding. Coding can be fully automated if a sample of labeled data exists. For example, if we want to label how a customer feels about different products from their textual reviews, having the numeric/star rating is a sufficient label for the customer's opinion of the product. In cases where labels are not automatically attainable, algorithms will need small samples (100s to 1000s of records) of human-coded labels to build a model. The number of samples needed depends on the complexity of the learning task and the desired quality/accuracy of the inference. The big advantage of an automated coding method is that it allows for coding large amounts of raw data into categories quickly. But greater reliability (i.e. less random error) and less bias (i.e. less systematic error) must remain as data generation goals for machine processing to be an advantage. Whether this approach is better than human coding or not needs to be evaluated with each new automated coding method, topic, and data source.

For example, topic modeling is an automated method that can help us extract keywords and concepts from text in order to identify topics summarizing the content from far more text than human research assistants could have read and coded in the past. Topic modeling can now be fully automated (Blei, 2012) or can involve human input to narrow down relevant words through an iterative, semi-supervised process. Alternatively, researchers may have specific concepts they wish to identify in a large body of text and hand code whether a sample of text contains a given topic to train a model to automatically code the rest. In all three of these cases, researchers can code far more text than possible with human coding alone.

As an example, Bode et al. used a semi-automated approach for identifying topics related to the 2016 US presidential election from open-ended responses to survey interviews, newspaper articles, and tweets (2020). The research team analyzed almost 60,000 interviews collected at a rate of 500 per day and a sample of 5,000 tweets per day, comparing what people recalled seeing or hearing about each candidate "in the previous couple of days" and the topics appearing in the tweets. While fully automated approaches were considered given the volume of data, the text from the survey responses entered by interviewers was short and noisy, e.g. incomplete sentences, misspellings, and the like. Therefore, it was more effective to begin with hand developed lists using one data source, employ automated approaches to identify other relevant words, and then have experts determine which words should be associated with which topics. As the volume of unstructured text data continues to increase, computer scientists will need to continue developing new fully and semi-automated algorithms for identifying topics in a body of text. In general, fully and semi-automated coding of data may be instrumental in helping increase the pace of social science research.

Different Learning Paradigms for Different Learning Tasks

All machine learning algorithms and approaches are not equal. One can group them based upon the learning style. Supervised learning algorithms predict an outcome variable (dependent variable) using a set of predictors (independent variables) by employing a function that maps the inputs to the outputs desired. For supervised learning tasks, labeled training data exist (i.e. a set of examples where the outcome variable is known for a set of predictors) and models are being built based on them. Some classic supervised learning models include decision trees, Naive Bayes, and support vector machines. Unsupervised learning does not have an outcome variable or labeled training data. Instead there is a large number of examples, but no outcome variable. These methods are used to split the population into different groups. Examples of unsupervised learning tasks include clustering and topic modeling. Finally, reinforcement learning is an option when supervised learning is not as effective or is not possible. Reinforcement learning deploys software agents in an environment to make decisions by maximizing a cumulative reward. The general idea is that instead of building a model directly, agents can apply different strategies in order to try to maximize their award. Each agent must balance between using its current knowledge to make a decision (exploitation) or looking at more unseen data examples (exploration). Two examples of reinforcement learning models are Markov Decision Processes and Q learning (Nilsson, 2015). Other learning styles exist, but supervised, unsupervised, and reinforcement learning are the most common. Therefore, understanding the type of data available and the appropriate learning style for the task is an important first step. Ultimately, computer and data scientists perform controlled experiments on the behavior of different algorithms on different types of data. If the algorithm has not been applied to the data type being used, e.g. short text or images, then experiments with sensitivity analyses of different hyperparameters need to be the next step so that researchers know whether or not an algorithm should be used for a specific learning task.

Many other algorithmic and computational considerations exist, including algorithmic bias, parameter selection, computational complexity and runtime of the algorithm, expectations of algorithms for specific underlying data distributions, and overfitting of models to training data to name a few. As new and promising models and computational methods emerge, researchers need to pause to understand the strengths, weaknesses, and assumptions associated with them. Only then can they be used effectively for good evidence-based decision making.

Ethical Considerations When Blending Data

Data ethics is emerging as a new discipline that investigates issues related to responsible data collection, data use, and data inference. The growth of organic data sets and open government data sets containing millions of data points has led to many ethically questionable practices that must be revisited to establish better standards. This section presents different issues and challenges that need to be thought through as we consider blending different data sets.

A large amount of personal, and even sensitive, information can be obtained from personal devices and online social media accounts. Much of these data can be obtained using APIs and apps without obtaining explicit consent of the people producing the information. Researchers also use big data to impute missing data values from surveys and other administrative data sets without consent. When should explicit consent be required? Is implied consent reasonable in some cases? How should traditional consent and data ethics guidelines be adjusted to address the availability and use of these kinds of information? Consideration of consent issues also leads to larger questions around data ownership, privacy, and the right to be forgotten. What should we expect from companies who handle our personal data? How do the obligations of federal statistical agencies who produce "official statistics" differ from commercial data providers who are selling products based upon the blending of big data? How do we ensure that people understand how their data are being used when we may not know how to reach them?

User consent is not the only ethical issue raised when blending data using large, online sources. Data mining and machine learning algorithms are a key to working with large data sets. Unfortunately, each algorithm generates a model with its own assumptions that are not readily apparent to researchers using the information they produce. These assumptions can lead to different biases and may even lead to models with fairness issues. A model that is built using biased or imbalanced training data, e.g. more examples from men than women, is likely to do a better job on predictions involving the majority class, e.g. men in this example. This exact scenario occurred when Amazon built a machine learning algorithm for selecting candidates to interview (Dastin, 2018). This reliance on learning algorithms for decision making without understanding the implicit biases built into the model is a cause for concern. This is even more relevant when using black box models instead of explainable ones. What levels of reliability and fairness are reasonable from an ethical perspective? While different metrics for assessing algorithmic fairness exist, which ones are best to use with different learning tasks and different types of data?

Further, as we blend and combine data, we may make them more identifiable than each of the original parts were individually. In this case, it is the act of blending the data that opens the door to privacy and confidentiality infringement.

In sum, social science research has always faced issues of privacy, consent, bias and fairness, but the impact of these ethical issues on research and available strategies for addressing them are larger and less well understood in the era of data blending with publicly accessible organic data sources.

Responsible data stewardship involves a careful consideration of subjects' rights at all stages of the research process. It begins with considerations of privacy and confidentiality in the design of the original data collection for which a researcher is responsible. It continues to the stage at which the investigator decides to share data with others and takes care to anonymize the data in terms of direct identifiers as well as the values of specific variables – as measured or created – that could be used to identify an individual because of relatively unique values. It extends to the treatment of the results from combining multiple data sources that could enhance others' ability to identify specific individual units to sharing individual anonymized posts that could be de-anonymized using online searches. As we begin to understand the different harms that can occur at different stages of the research process, new policies need to be developed to protect subjects.

Best Practices

While we have identified several challenges, here we highlight some best practices that are important when blending different forms of data.

Get Consent When Necessary: Some data have been anonymized and released, so getting consent is not possible. However, in cases where the data are not shared publicly by a user, or in cases where the user likely believes that the data are not being shared publicly, getting consent is important. Even if the secondary analyst does not know the identities of the subjects, the original investigators might. They should be contacted with a description of and a request for re-contacting the subjects to obtain their permission for the new intended use. It is a foundational pillar of good social science research.

Harmonization of Data: When blending multiple data sources that contain similar constructs that are measured differently, it is important to harmonize the measures to create a usable blended data set. As more data sets are combined, understanding the -similarities and differences in variables that are approximately the same and generating new variables that account for measurement error and bias for each source and level of data is an important step of the data blending process.

Keep a Record: Once a data set is blended, it is easy to lose track of the origination of the different data elements. Therefore, it is important to keep information about the data's provenance. What was the original data value? Where was the data obtained from? What variables were used to merge or link the data? What algorithms and parameters were used to construct new variables? Maintaining a detailed account of how the blended data set was generated will help maintain the integrity of the data. Giving proper attribution for the data constructed throughout the blending process is just as important (Silvello, 2018).

Understand Algorithm Basics: Given the volume of data that we are now working with, manual processing is not an option. Using computational resources is a necessity, if not for accuracy, then because of the volume. Every algorithm used makes assumptions about the underlying data, e.g. the data are normally distributed or have a certain type of skew, and how to group the data. This means that we can compare the functions being optimized and the impact of any initial parameters, i.e. the sensitivity of the parameters. Just understanding these two properties allows us to compare different algorithms and understand differences in the results they produce.

Test for Fairness: Because models optimize for a function, they are not designed to recognize biases that may exist in their design as a result of insufficient, noisy, or imbalanced training data. Providing such information is the responsibility of the researcher, particularly social science researchers who are looking for generalizability from predictive models. New fairness measures are being developed and can be used to assess such properties, or even levels of accuracy, across population elements.

Transparency: Every measurement technique is subject to error. By combining multiple sources of data, however, the errors in sampling, measurement, and coverage may compound. Here we introduce the concept of Total Data Error (TDE), a way of understanding holistically how a blended dataset contains errors from each of its component parts, how those errors might interact with, aggravate, or possibly alleviate each other. TDE can be modeled on Total Survey Error (Groves & Lyberg, 2010), which considers errors at each stage of survey design and execution. TDE creates greater transparency related to large, blended datasets (Bode, 2018). It demands that researchers 1) know, 2) consider, and 3) report the choices they make, the assumptions of their models, and the sources of error from each component dataset, and how those errors may relate to one another, either theoretically or empirically.

Conclusions and Implications

As more data become available about individuals across multiple contexts and through time, it is important to think of ways to combine these sources of data to help in understanding the unique ways in which humans experience their lives. The social/behavioral sciences broadly considered, in combination with computer and computational scientists, are now positioned to connect and blend data across multiple designed and organic sources that allow for understanding individuals’ paths to physical, mental, social, educational and other outcomes of interest to practitioners, policymakers, and researchers.

The amount of money being spent by the government, institutes of higher learning, industry, and philanthropy to understand and try to help the human condition is vast. This money has been spent creating multiple levels of data from individual human cells to large societal programs (e.g. Social Security) to help us understand how society influences human changes across time. Being able to take advantage of the wealth of these data by combining them is an important research goal. However, this will not be done by any single discipline within behavioral science, the academy, industry, or government but instead by collaborations across these groups.

We are now at a point with technology and creative thought that allows a convergence of thinking and the production of blended data from multiple sources to understand the human condition in new and more complex ways. In order to maximize the potential research use of this blending, novel ways of extracting data from digital resources that were not primarily constructed for data use (text, video, audio) are needed. Statistical advancements to deal with the error and “noise” in the blended data will need to be constructed as well. There will also need to be decisions made on issues of privacy and ethics as datasets are created to maximize our understanding of individuals instead of groups. These are challenges that we can now meet with the right interdisciplinary collaborations across academia, government, and industry partners.

Acknowledgments

We would like to thank the Massive Data Institute for supporting the Data Blending Workshop (Data Blending: Tackling the Obstacles) in April 2019. This white paper is a result of that workshop and is co-authored by faculty who participated in the workshop and a half day meeting following the workshop. It also was foundational in developing some of the ideas that are part of the NSF Award #1934925. A special thanks to Rebecca Vanarsdall from MDI for her support during the writing process.

References

Andrews, W. G. (1966). American voting participation. *The Western Political Quarterly, 19* (December), 639–52.

Amin, V., Behrman, J., Fletcher, J. M., Flores, C. A., Flores-Lagunes, A., & Kohler, H. P. (2019). Mental health, schooling attainment and polygenic scores: Are there significant gene-environment associations?. *IZA Discussion Paper.*

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Bode, L. (2018). Everything Old Is New Again: Big Data and Methodological Transparency. In N. J. Stroud & S. McGregor (Eds.), *Digital Discussions* (pp. 36-49). Routledge.

Bode, L., Budak, C., Ladd, J., Newport, F., Pasek, J., Singh, L., Soroka, S., & Traugott, M. (2020). *Words that matter.* Brookings Institution Press.

Bond, R. (2019). Social network determinants of screen time among adolescents. *The Social Science Journal*.

Brazil, N., & Clark, W. A. (2019). Residential mobility and neighborhood inequality during the transition to adulthood. *Urban Geography*, *40*(7), 938-963.

Centers for Disease Control and Prevention (CDC). (2014). Introduction to Public Health. *Public Health 101 Series.*

Cesare, N., Grant, C., & Nsoesie, E. O. (2019). Understanding Demographic Bias and Representation in Social Media Health Data. *Companion Publication of the 10th ACM Conference on Web Science*, 7-9.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *American Economic Review*, *106*(4), 855-902.

Collins, W. A., Maccoby, E. E., Steinberg, L., Hetherington, E. M., & Bornstein, M. H. (2000). Contemporary research on parenting: The case for nature and nurture. *American Psychologist, 55*(2), 218–232.

Cunha, A. J., Leite, Á. J., & de Almeida, I. S. (2015). The pediatrician's role in the first thousand days of the child: the pursuit of healthy nutrition and development. *Jornal de Pediatria (Versão em Português)*, *91*(6), S44-S51.

Cutler, D. M., Ghosh, K., Messer, K. L., Raghunathan, T. E., Stewart, S. T., & Rosen, A. B. (2019). Explaining the slowdown in medical spending growth among the elderly, 1999–2012. *Health Affairs, 38*, 222–229.

Dastin, J. (2018, October 9). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://reut.rs/2Od9fPr

Deutsch, A. R., Chernyavskiy, P., Steinley, D., & Slutske, W. S. (2015). Measuring peer socialization for adolescent substance use: A comparison of perceived and actual friends' substance use effects. *Journal of Studies on Alcohol and Drugs*, *76*(2), 267-277.

Downs, A. (1957). *An economic theory of democracy*. Harper.

Duggan, M., Lenhart, A., Lampe, C., & Ellison, N. B. (2015). *Parents and social media.* Pew Research Center. https://www.pewresearch.org/internet/2015/07/16/parents-and-social-media/

Frieden, T. R. (2010). A framework for public health action: the health impact pyramid. *American Journal of Public Health*, *100*(4), 590-595.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2003). *Bayesian data analysis*. CRC Press.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374-378.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly,* 74(5), 849-879.

Gurney, J. G., McPheeters, M. L., & Davis, M. M. (2006). Parental report of health conditions and health care use among children with and without autism: National Survey of Children's Health. *Archives of Pediatrics & Adolescent Medicine*, *160*(8), 825-830.

Hansford, T. G., & Gomez, B. T. (2010). Estimating the electoral effects of voter turnout. *American Political Science Review*, *104*(2), 268-288.

Hargrove, T. W., Halpern, C. T., Gaydosh, L., Hussey, J. M., Whitsel, E. A., Dole, N., Hummer, R.A. & Harris, K. M. (2020). Race/ethnicity, gender, and trajectories of depressive symptoms across early-and mid-life among the Add Health cohort. *Journal of Racial and Ethnic Health Disparities*, 1-11.

Harris, K. M. (2013). The add health study: Design and accomplishments. *Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill*.

Hejblum, B. P., Weber, G. M., Liao, K. P., Palmer, N. P., Churchill, S., Shadick, N. A., Szolovits, P., Murphy, S.N., Kohane, I.S., & Cai, T. (2019). Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Scientific Data*, *6*, 180298.

Kennedy, C., Hatley, N., Keeter, S., Mercer, A. Igeilnik, R., & Frederic, T. (2018) *Comparing survey sampling strategies: Random-digit dial vs. voter files.* Pew Research Center. https://www.pewresearch.org/methods/2018/10/09/comparing-survey-sampling-strategies-random-digit-dial-vs-voter-files/

Knack, S. (1994). Does rain help the Republicans? Theory and evidence on turnout and the vote. *Public Choice*, *79*(1-2), 187-209.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203-1205.

Markowitz, A. J. (2018). Changes in school engagement as a function of No Child Left Behind: A comparative interrupted time series analysis. *American Educational Research Journal*, *55*(4), 721-760.

Markowitz, A. J., Ryan, R. M., & Marsh, A. A. (2015). Neighborhood income and the expression of callous–unemotional traits. *European Child & Adolescent Psychiatry*, *24*(9), 1103-1118.

McClain, C., Mneimneh, Z., Singh, L., & Raghunathan, T. (2018.) *Seeking the "Ground Truth": Assessing Methods Used for Demographic Inference from Twitter* [Presentation]. The BigSurv18 Conference, Barcelona.

McCormack, J. (2020, April 10) Why were only five polling places open in Milwaukee this week? *National Review.* https://www.nationalreview.com/2020/04/wisconsin-election-milwaukee-did-not-have-enough-polling-places/

Mebane, W. R., Pineda, A., Woods, L., Klaver, J., Wu, P., & Miller, B. (2017, April). *Using Twitter to observe election incidents in the United States* [Presentation]. Annual Meeting of the Midwest Political Science Association, Chicago.

Nguyen, Q. C., Sajjadi, M., McCullough, M., Pham, M., Nguyen, T. T., Yu, W., Meng, H. W., Wen, M., Li, F., Smith, K. R., Brunisholz, K., Tasdizen, T. (2018). Neighbourhood looking glass: 360º automated characterisation of the built environment for neighbourhood effects research. *J Epidemiol Community Health*, *72*(3), 260-266.

Nilsson, N. J. (2005). Introduction to machine learning: an early draft of a proposed textbook. *Stanford University*.

Paparrizos, J., White, R., & Horvitz, E. (2016). Screening for Pancreatic Adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice, 12*(8), 737-744.

Perrin, A., & Anderson, M. (2019). *Share of US adults using social media, including Facebook, is mostly unchanged since 2018.* Pew Research Center. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Raghunathan, T., Ghosh, K., Rosen, A., Imbriano, P., Stewart, S., Bondarenko, I., Messer, K., Berglund, P., Shaffer, J., & Cutler, D. (2020). Combining information from multiple data sources to assess population health. *Journal of Survey Statistics and Methodology*.

Ryan, R, Davis-Kean, P. E., Bode, L., Kruger, J, Mneimneh, Z., Singh, L. (2020). The new Dr. Spock: Analyzing information provided by parenting-focused Twitter accounts. [Unpublished paper under review].

Sayers, A., Ben-Shlomo, Y., Blom, A. W. & Steele, F. (2016). Probabilistic record linkage. *Int. J. Epidemiol. 45,* 954–964.

Schenker, N., & Raghunathan, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, *26*(8), 1802-1811.

Serbin, L. A., Hubert, M., Hastings, P. D., Stack, D. M., & Schwartzman, A. E. (2014). The influence of parenting on early childhood health and health care utilization. *Journal of Pediatric Psychology*, *39*(10), 1161-1174.

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6-20.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238-241.

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E. & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv preprint arXiv:2003.13907*.

Sleap, M., & Warburton, P. (1996). Physical activity levels of 5-11-year-old children in England: Cumulative evidence from three direct observation studies. *International Journal of Sports Medicine*, *17*(04), 248-253.

Soroka, S. N., & Wlezien, C. (2010). *Degrees of democracy: Politics, public opinion, and policy*. Cambridge University Press.

Stein, R. M., Mann, C., Stewart, C., Birenbaum, Z., Fung, A., Greenberg, J., Kawsar, F., Alberda, G., Alvarez, R. M., Atkeson, L., Beaulieu, E., Birkhead, N. A., Boehmke, F. J., Boston, J., Burden, B. C., Cantu, F., Cobb, R., Darmofal, D., Ellington, T. C., … Wronski, J. (2019). Waiting to vote in the 2016 presidential election: Evidence from a multi-county study. *Political Research Quarterly*.

Stephens-Davidowitz, S. (2020, April 5). Google searches can help us find emerging covid-19 outbreaks. *The New York Times*. https://nyti.ms/2UKc0sZ

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A.M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, *338*, b2393.

Stewart, C., & Ansolabehere, S. (2015). Waiting to vote. *Election Law Journal*, *14*(1), 47-53.

Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, M., Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., & Staab, S. (2018). Systematically monitoring social media: The case of the German federal election 2017. *arXiv preprint arXiv:1804.02888*.

Van Assche, J., Van Hiel, A., Stadeus, J., Bushman, B.J., De Cremer, D. & Roets, A. (2017). When the heat is on: The effect of temperature on voter behavior in presidential elections. *Frontiers in Psychology.*

Weather Underground. (2020). Retrieved from https://www.wunderground.com/pws/overview