Analysis and Visualization Considerations for Quantitative Social Science Research Using Social Media Data

J. Bradford Jensen[1], Lisa Singh[1], Pam Davis-Kean[2], Katharine Abraham[3], Paul Beatty[4], Leticia Bode[1], Duen Horng Chau[5], Tina Eliassi-Rad[6], Rich Gonzalez[2], Rebecca Hamilton[1], In Song Kim[7], Theresa Kuchler[8], Jonathan Ladd[1], Kristina Lerman[9], Maggie Levenstein[2], Zeina Mneimneh[2], Quynh Camthi Nguyen[3], Josh Pasek[2], Trivellore Raghunathan[2], Rebecca Ryan[1], Stuart Soroka[2], Mahlet Tadesse[1], and Michael Traugott[2]

July 28, 2021

[1] Georgetown University
[2] University of Michigan
[3] University of Maryland, College Park
[4] U.S. Census Bureau
[5] Georgia Institute of Technology
[6] Northeastern University
[7] Massachusetts Institute of Technology
[8] NYU Stern
[9] University of Southern California

This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. To learn more about The Future of Quantitative Research in Social Science research project, visit www.smrconverge.org.

1. Project Overview

In 2019, a group of computer and social scientists began a project to 'converge' the two fields of research, with the aim of harnessing data from social media to improve our understanding of human behavior. People all over the world use social media, search engines, smart devices, and other technologies that record their moment-to-moment behaviors (often called "digital traces"). Some digital traces are more passively collected by devices that are being carried or worn by people, like location tracking on a smartphone or measuring steps on a FitBit device. Other types of digital trace data are more actively or deliberately created, like social media posts. All these data provide massive amounts of insight into the day to day lives of humans. Social media, in particular, provides a massive amount of information on the everyday activities, opinions, thoughts, emotions, and behaviors of individuals, groups, and organizations in near real-time. Today, most adults in the US use some form of social media (Perrin & Anderson, 2019) to share and discuss topics as wide-ranging as politics, employment, parenthood, leisure activities, travel, sports, and health. As such, these platforms provide new ways of gathering information on constructs relevant to all social science fields.

Expanding the availability and utility of this extremely rich but still underutilized data source in the social sciences requires attention to the unique features of these data. However, unlike many forms of typical social science data, social media data have a structure that is not the product of a designed process initiated by the researcher to answer specific hypotheses or questions. Instead, the data are provided "as is," which often means they are raw, complex, and highly sparse in nature. Moreover, these data introduce unique bias concerns not typically at issue in traditional social science methods, including a good deal of imprecision about who generated the data, or what population those data represent. The organic nature of the data, along with their magnitude and complexity, require methods for managing, structuring, and understanding these data to create useful measures for social scientific inquiry. Appropriate methods for doing so are found within the toolbox of computer scientists, making a convergence of computer science and social science methods potentially very fruitful.

While employing methods from computer science to wrangle social media content in order to answer social science questions has enormous potential, it also presents a number of challenges. First, neither computer scientists nor social scientists are especially well versed in the others' methods, conventions, or language. So before social scientists can begin using ideas and algorithms from computer science, they need to learn how to work with large-scale unstructured organic data and understand the general principles, tools, and methods used by computer scientists. Likewise, computer scientists can reach inaccurate conclusions if they fail to understand key considerations and objectives within social science research that may not traditionally apply in computer science. Second, it is often unclear who or what, exactly, are behind the accounts that produce the data appearing in social media data sets, and how the entities creating trace data might relate to the larger groups of people that social science researchers would like to understand. For many questions, social media data contain information

about the presence of a behavior, but do not have information about why or under what circumstances that behavior may have occurred (or may not have occurred) -- which is a key area of focus for many social scientists. Third, ethical questions around the use of digital trace data in research contexts require collaboration across these disciplines. Understanding what we need to know about the data that are gathered, what other data are needed to supplement them to answer central research questions, and how to do so responsibly is critical for digital trace data to live up to their full utility. These are just some of the challenges involved in the social+computer science study of social media data.

The advantages to social science in effectively harnessing social media data are clear. But for computer scientists too, this convergence holds great opportunity. Designing algorithms with a new set of constraints and optimizing existing algorithms for this large-scale, real-time domain, while addressing privacy, bias, misinformation, and algorithmic fairness concerns, will also advance computer science research.

To initiate this convergence, our group planned a set of topical meetings bringing together social scientists from multiple disciplines, including economics, psychology, political science, communications, sociology, and survey methodology with data scientists and computer scientists with the goal of creating a common set of understandings and methodologies for how to study complex human behaviors using social media data in a scientifically rigorous manner. The topics of these meetings addressed each stage of the research process as we have defined it: study design; data acquisition, sampling, and preparation; measurement and feature engineering; model construction; analysis and visual analytics. At each meeting, we also discussed criteria for the responsible conduct of research with social media data.

This is the fifth in a series of white papers providing a summary of the discussions and future directions that are derived from these topical meetings. This paper focuses on issues related to analysis and visual analytics. While these two topics are distinct, there are clear overlaps between the two. It is common to use different visualizations during analysis and given the sheer volume of social media data, visual analytic tools can be important during analysis, as well as during other parts of the research lifecycle. Choices about analysis may be informed by visualization plans and vice versa - both are key in *communicating* about a data set and what it means. We also recognized that each field of research has different analysis techniques and different levels of familiarity with visual analytics. Putting these two topics into the same meeting provided us with the opportunity to think about analysis and visual analytics/visualization in new, synergistic ways.

Finally, we pause to mention that we have defined a large number of concepts in the other white papers and will only define the subset that are necessary here. We refer you to other white papers for definitions that are more relevant to study design, sampling, measurement, and modeling (Bode et al., 2020; Mneimneh et al., 2021; Ladd et al., 2020; Budak et al., 2021).

2. Terminology and Background

Researchers in different disciplines think about analysis in different ways. Is the analysis descriptive or explanatory? Is it cross-sectional or longitudinal? Key questions about the processes that produce social media data and the research questions scholars wish to answer with these data have implications on the type of analyses and visualizations that may be appropriate and/or helpful. Because of the variety of ways that data may emerge and the numerous goals to which researchers may leverage those data, there are many different resolutions and dimensions along which analyses can be defined. We begin this section by reviewing some common analysis frames and their uses as a way to help structure the discussion that follows and, perhaps, broaden our collective perspectives on the range of analysis options (Section 2.1). We then switch gears to define and frame information visualization and visual analytics, highlighting differences and similarities across different disciplines (Section 2.2).

We want to reiterate that analysis and visual analytics should not be viewed as synonyms or the same concept. This white paper is not advocating for that. Instead, because there are synergies between analysis and using visualizations and visual analytic tools to support the analysis process, there is overlap when thinking about them in the context of social media data. Therefore, when the overlap is large, we discuss the two at the same time and when only an analysis concept or a visualization concept apply, we will focus on that instead.

2.1. Analysis Resolutions and Dimensions

Here we present three alternative ways of thinking about analysis: based on level/scale, based on analysis type, and based on data properties. We borrow from framing across different fields, blending concepts when appropriate.

Level/Scale: One dimension considers the level or scale of the study, whether it is about individuals or groups or populations (micro to macro). The micro-level perspective of social media is research that focuses on the actions and behaviors of individuals participating on the platform. For example, one might study the differences in messaging between Trump and Biden supporters as it relates to Covid-19. The meso-level perspective of social media is research that focuses on the actions and behaviors of groups participating on the platform. Here, groups can be constructed in different ways. A group may be explicitly defined as an entity on the platform. For example, on Facebook, there are explicit group pages, e.g. parenting groups, women's groups, university student groups, etc. These groups may exist only online or both offline and online, e.g. church groups. A group can be constructed based on shared interest, e.g. those conversing about the same topic (people using #meToo or #baltimoreRavens). A group may exist based on external classifications, e.g. U.S. Senators or S&P 500 Companies. An example research question related to groups may be the following - how are Republican senators vs Democratic senators discussing Covid-19 policy? Finally, the macro-level perspective of social media is research that focuses on aggregate level actions and behaviors of the population participating on the platform. For example, researchers may be interested in understanding public opinion about

Covid-19 stay-at-home orders on a specific social media platform, or how opinions vary across platforms.

Analysis Type: Another dimension along which we can organize analyses that is more prevalent in the social sciences is whether the analysis is descriptive or explanatory, where the goal of a descriptive analysis is to use data to explain or describe what is happening and the goal of explanatory analysis is to use data to explain what has occurred. We note that in the context of social media data, many causal questions cannot be answered - descriptive analyses are the norm. An example of a descriptive analysis would be looking at the prevalence (or dearth) of different Covid-19 related myths. An explanatory analysis example would explain why some misinformation spreads more rapidly than others.

Data Properties: Yet another dimension to organize analyses along is based on the data properties themselves. This type of organization is more prevalent within computer science. A text or image content analysis focuses on using the text or images within a post for the analysis, e.g. understanding the emotional content of a post. A data property that is commonly used across disciplines is time. Temporal or longitudinal analysis focuses on looking at trends/anomalies through time. Social media gives us the opportunity to vary the time scales considerably, e.g. minute, daily, weekly, annually. An interesting minute by minute analysis could be looking at how conversation or opinion shifts about the candidates during a debate (Freelon & Karph, 2014; Budak et al., 2020). Spatial or geographic analyses investigate attitudes or patterns of behavior using location. Again, social media data allow for analyses at a range of scales from geolocation point or neighborhood to states or countries. Finally, networks are another data property that analyses can be organized around. A network analysis not only considers the individual participants, but also the relationships among the individuals. Network analysis can range from ego-networks about individual entities to platform communities. Social media is particularly well suited for network analysis because networks can be constructed easily by looking at shared content, friendship/followers, and mentions to name a few.

We do not advocate for any one of these analysis framings, but we believe that it is important to recognize that these different types of organizational schemes are useful for social media analysis. We also refer the reader to the white paper focused on study design and the different types of studies that are well suited for social media analysis, including qualitative observational designs, experimental designs, survey designs, and data first designs (Bode et al., 2020). That white paper walks through how to translate the more traditional social science designs to effective designs for social media analysis, and also considers a "data first" design that maps to opportunistic data collection.

Finally, when conducting a social media analysis, issues of consent and privacy cannot be ignored. Across all of our meetings, there was concern expressed about the lack of uniformity around these issues. Previous white papers have discussed these issues in more detail, particularly the white paper focused on sampling and data preparation (Mneimneh et al., 2021).

2.2. Information Visualization and Visual Analytics

While visualizations have been used to support scientific research for a long time, in the last decade, the field of visual analytics has changed the role visualizations play in the research process. Kiem and colleagues (2008) define visual analytics as a field that integrates automated analysis techniques with interactive visualizations to improve understanding and reasoning from large, complex data sets. Ultimately, the goal is to analyze large volumes of data in order to determine and visually highlight valuable and important content or patterns.

Because we are still understanding the properties of social media data, visual analytics can be especially important for providing researchers an avenue to communicate the process of analysis, instead of only sharing results. It also gives researchers new tools for exploring the data and viewing it in innovative ways, thereby helping researchers understand the properties of the data and highlight their connections to different research questions. As an example, one can imagine seeing a network visualization that is colored to highlight different structural clusters. However, a researcher can explain much more about the network if he/she can show connectivity using different types of links and explain why he/she decided to use a particular network construction. This ability to visually interact with the data, not only for exploration, but also for explanation, is a strength of visual analytics. Finally, we point out that there is a difference between information visualization and visual analytics. Information visualization focuses on displaying information in ways that improve understanding of the data and does not focus on the analysis task or the algorithms being used. Visual analytics focuses on supporting specific analysis tasks and considers the algorithms being used. Visual analytics supports a process by which humans and machines interact with each other to explore and analyze data, beginning with overviews and drilling down to improve understanding of relevant details. In this white paper, we will discuss both information visualization and visual analytics.

Similar to analysis, there are many different ways that visual analytics and visualization approaches and ideas can be organized. Because visual analytics is task oriented, we will discuss that more in the next section. Here we consider the following categories of visualization that are particularly useful for social media data: visualizations that highlight aspects of the raw data (data visualizations); visualizations that show the inner workings of models, including diagnostic checks, e.g. QQ plots for regression (model visualizations); visualizations that can be used to present statistical or other analysis results (statistical visualizations); and visualizations that are used for storytelling, including infographics (storytelling visualizations). Visualization can also span multiple categories, like a scatterplot (data visualization) that shows a regression line (statistical visualization). All of the categories of visualizations can be static, dynamic, or interactive.

In the social sciences, visualizations presented in research publications tend to be data visualizations or statistical visualizations. Visualizations are primarily used to describe the data to better understand variability, central tendencies, and outliers. Diagnostics are typically presented in tables or plots. Depending on what is being emphasized in the paper, these

diagnostic visualizations may not be presented in the final paper but are part of the research record detailing the research process to answering the central questions of the paper.

For social scientists, the most common data visualizations include box and whisker plots, scatterplots, histograms, and bar charts to show distributions and frequencies. For example, a researcher can display with a chart the 25 most popular food mentions on Twitter. Network visualizations (nodes (circles) with lines (edges) between them) are also common data visualizations for research involving network structures. Statistical visualizations, typically used to display the findings of different statistical analyses, focus on line charts, scatterplots, and some more advanced visualizations like dendrogram and heatmaps. Even though the visualizations for many social science disciplines remains simple, (perhaps due to demands from style guides such as the American Psychological Association's Style Guide), there have been changes occurring in basic software programs such as Microsoft Excel and R, where options for visualizations have increased substantially. Additionally, social scientists with familiarity with ArcGIS software can generate maps that display the spatial distribution of variables in their study. Thus, this is a potential area where the advances in the subfields of information visualization and visual analytics within computer science can provide needed innovation in how the social sciences test, display, analyze, and interact with the complexity of their data, potentially leading to clearer explanations and understandings new models being used (model visualizations) and insights into beliefs and behaviors through visual storytelling. This is not only promising for social media data, but also for behavioral data collected with all types of methods (e.g. surveys).

In computer science, the types of visualization and the level of visual analysis conducted throughout the research process vary considerably. Some computer scientists only use data and statistical visualizations. In these cases, histograms, box and whisker plots, and line graphs showing efficiency or cumulative distribution plots are the norm. However, most computer scientists who work with large-scale data think about how to use visualizations throughout their research life cycle to look at data in different ways at different resolutions. In some sense, the way social scientists and computer scientists look at the data and interpret them may differ because of differences in disciplinary training. For many computational problems, the raw data or aggregates that highlight specific patterns or deviations may be more valuable than the specific model that is generated from the data. For example, a computer scientist developing a new clustering algorithm needs to design it based on specific features of the data itself, making the raw data central to the design of the algorithm. For social scientists, the model of the data, particularly one that is generalizable, tends to be more valuable than the raw data itself. This difference could have significant implications for the perceived value and ultimate usage of visualizations by social scientists throughout the research project lifecycle. Ultimately, however, analyzing data through visual data representations is more useful if the “true insight” is in understanding the strengths and weaknesses of different estimates that reside between visualizations of the raw data and visualizations of the statistical model.

Finally, it is worth noting that not all visualizations are successful. While the goal of information visualization is to create images that help users see patterns or highlight features using an intuitive visual design, some visuals do not improve the understanding of the data or help convey information easily to users. This can happen when visualizations attempt to explain too many things or do not summarize data in ways that are easily interpretable by researchers. Therefore, we should not use visualizations or visual analytics because we can. It is important to use them when they convey information in a meaningful way. It is also important to use them with care to ensure that they do not distort what the data are actually saying. They can give incredible insight when thought through carefully.

Given the disciplinary differences in using visualizations and visual analytic tools and techniques, one goal of this white paper is to help researchers recognize the range of visualizations and visual tools available to them.

3. Analysis Pipeline

When designing a study that incorporates social media data, a special emphasis needs to be placed on understanding the relationship between the research question and the available data. Because of the novelty of social media data for many social science researchers and because of the potentially (relatively) large scale and unstructured nature of social media data, it became clear that even at the initial stages of a project, researchers need to take extra time to think through the entire analytical pipeline: how to obtain the data, how to prepare it, how to construct measures, how to model it, how to assess the model, and how to discuss the analysis and results. However, one of the challenges of bringing computer scientists and social scientists together is the absence of a common vocabulary for similar concepts and tools. The absence of a common vocabulary necessitated the construction of a baseline understanding/vocabulary with regard to the other components of the methodology. To facilitate communication between the two groups, we introduced the notion of this "analytic pipeline" that incorporates the methodological components discussed in the other meetings. Figure 1 shows the components we included in the analytic pipeline: data acquisition, data preparation, construction and assessment of features and measures, mathematical model construction and assessment, and storytelling.

In this section, we draw on insights from previous meetings supported by the NSF Convergence grant, as well as discussions in the most recent meeting to develop high-level overviews and checklists containing six to eight questions that should be considered when designing a study that uses social media data. We also highlight the types of visual analytic tools that would be useful for each component of the analytic pipeline, and when possible, we mention existing tools that we are aware of. We also refer you to Chen and colleagues for a survey of different visual analytic tools that have been designed for social media (Chen et al., 2017).

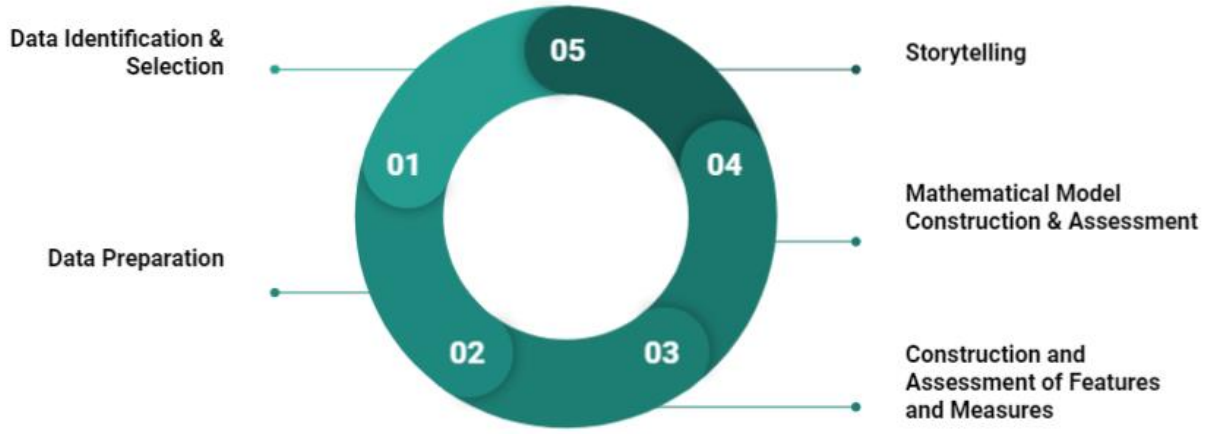

Figure 1. Analysis Pipeline

Finally, we pause to mention three broad observations we have made with regard to visualizations and the analysis pipeline. First, many visualizations have been designed to highlight univariate or bivariate relationships between the data/model/analysis. While useful, given the high-dimensionality of social media data, visualization tools beyond univariate or bivariate relationships are important. Second, for data exploration, measurement construction, and model exploration, interactive multi-dimensional visualizations can be very useful to further enhance our understanding of the data, model, or results. Third, visualizing millions of data points is not likely to be useful. So aggregating and slicing data in different ways becomes imperative given the scale of the data. Addressing these key needs is one area where collaboration between computer scientists and social scientists could have a potentially large payoff, for both better communication regarding existing tools and interaction to develop new tools.

3.1. Data Identification, Acquisition & Sampling

One of the challenges of using social media data is identifying and obtaining the specific sample of interest. A previous meeting highlighted the different ways data can be obtained, from working with social media companies to using platform-provided APIs (Application Programming Interfaces) to data scraping (Mneimneh et al., 2021). We also discussed the limited demographic data that accompany many social media samples of individual-level posts, for example, and considerations when designing a probability-based sample vs. a non-probability-based sample. It is not unusual for a researcher to need to adapt the question to the available data or design a study around the available data.

Data acquisition and sampling checklist

Which social media platform(s) is suitable for the research study?
What data are available to researchers on these platforms? Are these raw data or have they been curated or post-processed by the platform?
What are the properties of the data sample? Can the sample be used to generalize to the target population or are there subpopulations that are missing?
What are the data access restrictions? Does it change if the study is international?
How will researchers collect and store the data?
Can the original data be shared after the study is completed by the original team?
How will the data limitations impact the study design?

Visual Insight: A visualization that would be particularly useful would be one highlighting different data sources and the meta-data associated with the data sources. It would also be useful to have a visualization of the data schema or data model. There are many tools that exist for this, particularly for relational data models. For example, entity relationship (E/R) diagrams describe the relationships that exist between different entities or objects, e.g. user accounts, in a data set. These diagrams also keep track of the attributes or features associated with each entity and the data type of each attribute, e.g. decimal (float) for salary. There are a large number of tools for generating E/R diagrams, including LucidChart[10] and DBdiagram.io[11]. Platforms have tools and application programming interfaces (APIs) for data collection, but more visual tools that can be used by non-programmers would also facilitate research with these data. For example, Google public data lets users select subsets of data and shows visual representations of the selected subsets.[12] Figure 2 shows an example from the Google public data sets that shows fertility rates in different regions of the world using World Bank data.

[10] Lucidchart. (2021). *Lucidchart.* https://www.lucidchart.com/pages/
[11] Holistics Software. (2021). *Dbdiagram.io.* Holistics Software. https://dbdiagram.io/home
[12] World Bank (2020.) *World development indicators* [Dataset]. Available from Google Public Data https://www.google.com/publicdata/directory

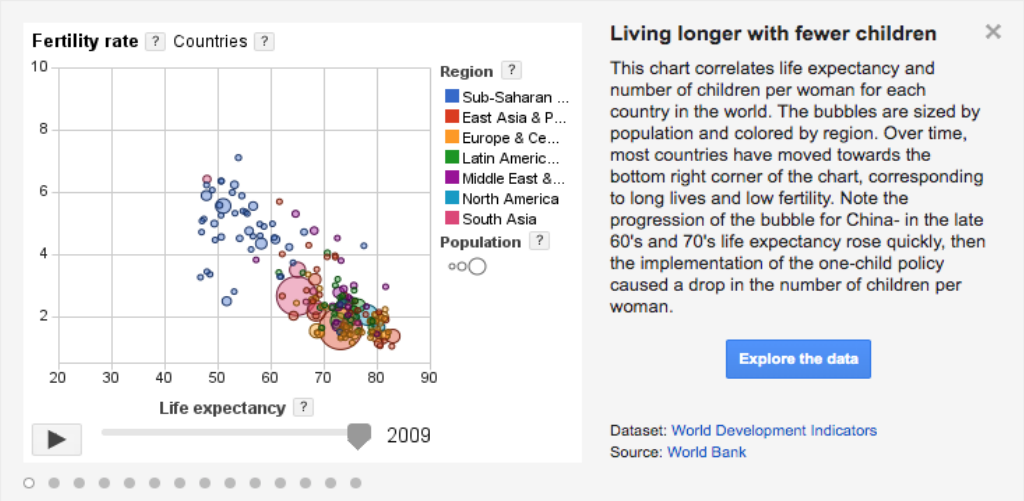

Figure 2. Example visualization using World Bank data through the Google public data webpages. The "Explore the data" link lets users see the detailed data and build new visualizations using subsets of data

3.2 Data Preparation

Once collected, the data need to be prepared and cleaned for analysis. There are many considerations that arise during data preparation, including what should be included/excluded from the study and how to standardize the text and/or image data so that reliable, meaningful variables can be constructed (Mneimneh et al., 2021). Previous meetings also discussed the biases that may be inherent in the data and options for correcting for or adjusting for these biases.

Data preparation checklist

- How much processing was already done during data collection?
- Have missing data been imputed? Does systemic missingness exist, and if so, should it be addressed in some way?
- Has the platform updated their interface or data in a way that may require updates to the data to make them more consistent over time? For example, does the longer post length on Twitter impact the researcher's analysis? Similarly, is there evidence that norms and patterns of use have changed over time such that steps need to be taken to ensure the compatibility of the data?
- How should the data be cleaned? Should some data be excluded (for example, removing bots)?
- Does documentation of any coding standards including where the data were collected from, cleaning steps, aggregation, imputation, etc. exist or does it need to be created?
- How do we account for the presumed biases in the data? Is adjustment or weighting necessary?
- How will the data preparation choices impact the final analysis?
- What statistics are most useful for helping researchers conduct high quality data preparation?

Visual Insight: There are a number of different types of visualization tools that are beneficial during data preparation. For example, histograms are useful for looking at aggregates of the data properties. If a researcher is processing a text data set, it can be useful to cluster words and concepts or view a preliminary word cloud to see what the main themes of the text are. Data cleaning is an important step. Interactive visual tools that let researchers view instances of anomalies, e.g. missing data or outliers, and remove them interactively through either individual selection or automated rule generation can improve the data cleaning process. Scatterplots are useful for correlations between continuous variables; box plots for seeing the distribution of a variable and possible outliers or to visualize the distribution of a continuous variable across levels of a categorical variable; violin plots also show the probability density of the data; tree maps can be useful for viewing the relative size of groups and subgroups within groups; and spaghetti plots can help with sensitivity analyses. There are a number of these types of visual supports in statistical packages like R[13] and data mining packages like Orange.[14] GIS mapping visuals are also an important way to see the spatial distribution of the data.

3.3 Measurement and Feature Design

Determining how to construct the measures and features that are important for an analysis is a central issue for many researchers. The previous meeting on measurement highlighted different ways social scientists and computer scientists construct and evaluate measures and features (Ladd et al., 2020). Social scientists evaluate measures based on how well they represent a theoretical construct (i.e., their validity). Computer scientists evaluate features (i.e., variables used within machine learning algorithms to model a predictive task) based on how well they capture a facet of the data that may be useful for the descriptive or predictive task at hand. These different viewpoints mean that measures and features are designed differently and that both communities need to work to bridge the gap. There are important concepts from both communities that should be used to assess the quality of measures and features, including, reliability, validity, bias, computational accuracy, and computational efficiency.

[13] The R Foundation. (2021). *The R Project for Statistical Computing.* https://www.r-project.org/

[14] Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research*, *14*(8): 2349−2353. https://orangedatamining.com/

Construction and assessment of features and measures checklist

- What are the measures and features most useful for the research?
- What is the reliability and validity of the measures and features?
- If the number of measures and/or features is large, does it make sense to remove some or use a dimensionality reduction technique?
- Does the measure or feature seem reasonable? (qualitative assessment)
- Are the feature distributions as expected or are the features capturing anomalous characteristics of the data? (quantitative assessment)
- Are there unobserved confounders that are creating a bias?
- How sensitive are the measures to data collection / algorithm / cleaning / imputation / construction choices?

Visual Insight: Similar to data preparation, face validity of measures can be looked at using line graphs and histograms, while violin plots, and heatmaps can be useful for seeing the distribution of the data. Biplots, displaying information of both samples and variables in a two-dimensional plane, are also useful for data exploration when determining features and/or relationships among variables. For example, they can be used to look at the relationship between the dependent and independent variables or for looking at projections of data samples (PCA biplot). Area charts are useful for comparing temporal variation of quantitative data. Circular plots or network visualizations can be useful for seeing connections among individuals. There are a number of visualizations that have been proposed for looking at data after dimensionality reduction, including t-SNE (t-distributed stochastic neighbor embedding) and ISOMAPS. Again, libraries exist for most of these visualizations in statistical software, data mining software, and more general-purpose programming languages like Python and Java.

3.4 Model Construction

There are a number of characteristics that make modeling social media data challenging, including a lack of understanding about who and why people post, the language, spatial and temporal variation of the data, and the quality of the data. Our meeting on this topic highlighted how differently social scientists and computer scientists think about modeling (Budak et al., 2021). Social scientists think about it upfront and focus on how to model a theoretical construct. Computer scientists think about modeling more broadly, considering feature design,

measurement, and analysis within the modeling process. Part of this stems from the difference in their final research products. Computer scientists construct mathematical models for description and prediction, comparing them against each other to understand their strengths and weaknesses. Social scientists tend to use a specific set of mathematical models they were trained to use and then focus more on using the model to explain a behavior, attitude, or phenomenon. This checklist focuses more on the construction and assessment of mathematical models as they relate to social media since it maps more readily to our conceptual lifecycle.

Mathematical model construction and assessment checklist

What are the independent and dependent variables for the study? What is the relationship structure between those variables?
What is the best model structure given the variables of interest and the scale and dimensionality of the data?
How should the model be estimated? What is the correct loss function? Should estimation from the model be regularized or smoothed?
How well does the model perform? What are the best diagnostics or assessments of the model? - Are the errors normally distributed? Do they vary by category? - How does the model perform on in-sample and out-of-sample data? (model reliability)?
How sensitive is the model to variation in key parameters?

Visual Insight: When evaluating the best model for a problem, researchers tend to look at different statistics related to its errors and model fit. There are meaningful plots that exist for all of these, including combined line and scatterplots that show the fit of the model to the data, histograms that show the importance of each feature for a machine learning model, visual representations of different diagnostics for understanding error, e.g. residual plots that show the lack of fit of data to a regression model, or glyphs that use shape, color, size, and orientation to highlight different aspects of the data in the "space" of the model. It can also be useful to employ faceting to visualize different groups next to each other to see how the relationships between the dependent and independent variables differ. Statistical packages have many visualizations to support model analysis. Python and R have packages that incorporate machine learning model analysis visualizations, including the package caret[15] in R and scikit-learn[16] in Python. Network

[15] Kuhn, M. (2019). *The caret package.* https://topepo.github.io/caret/
[16] Scikit-learn developers. (2021). *Scikit-learn.* https://scikit-learn.org/

analysis libraries that incorporate data visualizations include igraph[17] in R and Python and networkX[18] in Python.

3.5 Analysis

This component is a core part of this white paper as it focuses on understanding the outputs of the previous components. How should we interpret the results and what do they tell us? It is the glue that connects all the previous components and attempts to explain the answer(s) to the research question. At the meeting, many of the researchers felt that the final analysis using social media data would be similar to when other data sets are used. The largest difference is explaining to readers that the design of the study, the sample, and the measures constructed were all appropriate for the analysis being performed. The skepticism of readers is likely due to features like the novelty, multi-dimensionality, scale, and unstructured nature of social media data; the relatively recent application of existing algorithms and tools; and the development and deployment of new algorithms and tools to analyze these types of data.

Analysis checklist

How would a researcher interpret the results of the analysis?
Is the analysis sufficient or does the researcher need to consider a different sample, measure, or model?
Are there specific biases in the analysis that result from using social media data?
What are the central errors that can be introduced when using this analytic approach and can we test how likely those are to interfere with conclusions?
What are the strengths and limitations of the analysis?

The visual tools in the modeling and the storytelling capture those that would be useful for the analysis component of the pipeline.

3.6 Storytelling

Because social media data are relatively new, when discussing an analysis, the decisions made throughout the research lifecycle must be explained in more detail. While we have highlighted the core checklist in the analysis discussion, we believe that storytelling and visualization can be particularly useful for this domain because of the scale and complexity of the data.

[17] The igraph core team. (2021). *I-graph – The network analysis package.* igraph. https://igraph.org/
[18] NetworkX Developers. (2021). *NetworkX – Network analysis in Python.* NetworkX. https://networkx.org/

Storytelling checklist

How do we discuss the different parts of the research project (origin of data, preprocessing, modeling, analysis) in a clear and consistent way?
How do we relate the use of social media data to data that a specific research community is more familiar with?
How do we describe the analytic results in a way that maintains objectivity and aligns with the understanding that the audience has?
What is the best way to answer the "so what" question and relate the research to constructs and ideas a specific research community is familiar with?

Visual Insight: Because the use of social media data is less familiar to many social science researchers, it is important to "tell the story" about the entire analytic pipeline: the sources of the data, the preprocessing, the modeling, the related research. All of these aspects are needed to contextualize the findings. Visually showing these connections can help those new to these data better understand the research design and the decisions made. Researchers must explain the decisions in a bit more detail because of the scale of the data and the organic and novel nature of the data (text, images, audio). The "so what" question also becomes more critical and having a clear and consistent way to explain it can help with acceptance of the research results. Also, the researcher needs to describe the analytic results in a way that appeals to the understanding of the audience and the research methods and analyses they are most familiar with. Seeing visuals of how the story changes when a researcher changes his/her assumptions at different points throughout the pipeline is important for improving the audience's overall understanding. For example, suppose the researcher changes parameters of a model. How does that impact the results? Having an animation with a slider bar that visually highlights the differences or similarities is a possible way to incorporate visual analytics within storytelling. The Our World in Data project has a number of examples of animated visualizations, including one that looks at the relationship between living standards and urbanization over time[19].

There are a number of visual tools that support storytelling, including interactive charts, dashboards, and presentation software like Powerpoint or Prezi. Infographics are also a nice concise way to relate data to analysis results. These are typically made using publishing software like Adobe.

4. Social Media Case Studies

To help researchers think through different social media analyses, we present three case studies that we believe contain examples of some typical study designs for social media. The first one is a study focusing on understanding unexpected events or system shocks, e.g. a pandemic, a

[19] UN World Urbanization Prospects, Bolt, J. & van Zanden, J. (2020). *Urban population vs. GDP per capita, 2016* [Data visualization]. Retrieved from https://ourworldindata.org/grapher/urbanization-vs-gdp

mass shooting, or a terrorist attack. Our next example considers a planned event. Examples of planned events include a presidential election, a sporting event, or a holiday. Our final example involves monitoring specific conditions or behaviors over time, e.g. an index of happiness, social connectedness, or financial distress. We note that this section focuses on the analysis component of the white paper. However, throughout the section, we also discuss how visualizations and visual analytic tools can support the specific cases.

4.1. System Shock Case

Overview: In this case study, we consider when a system shock or an unexpected event has occurred. There are many examples of this including those related to the coronavirus pandemic and racial injustice. It is not uncommon for social media to be cited as a good data source for understanding system shocks.

We begin by considering research questions that analyze pre and post attitudes given a system shock. For example, shifts in racial attitudes towards blacks before and after the death of George Floyd, or Asians before or after the beginning of the pandemic. Or the growth or decline of the antivax movement after the COVID-19 vaccine became available. Another thread of research questions revolves around behavioral changes. How has alcohol use and other substance use patterns changed from before the pandemic? As the pandemic restrictions are being lifted have people returned to their old habits (social gatherings, commutes, snow days rather than online school days, etc.) and at what pace? A third thread of research focuses on the resilience of systems and structures during and after a shock. This thread includes questions about the impact of Covid-19 on companies in different industries, the health care system, or different government programs. In these examples, the impact would be on people's interactions with these systems. Finally, computer scientists may be interested in understanding how to measure, adjust, and develop more robust models for prediction that can better compensate for when shocks occur. They may also be interested in understanding how new forms of data can be used to measure attitude or behavior shifts after a shock. For example, researchers may develop methods for measuring emotions like fear from memes to better understand the relationship between fear and the spread of misinformation related to Covid-19. Finally, the visual analytics community has developed a large number of event sequence visualizations for different domains (see Guo et al., 2020). Developing innovative ways to visually analyze social media conversation when sudden shocks occur is an important future direction.

While any one of these examples would be interesting to walk through, we will focus most of the remaining discussion on analysis related to racial attitudes.

Analysis Plan: If we were using traditional data, we would use surveys to ask people about their attitudes on race. We may develop a set of statements and ask respondents if they agree or disagree with the statements. Ideally, we would do this immediately before and immediately after the shock. Unfortunately, because shocks are not anticipated, it is hard to get a survey into the field fast enough to capture the initial attitudes the public has about the shock event. Surveys also only allow for analysis of the specific questions in the surveys. Open-ended

responses can help increase insight about the topic. However, it is still not as insightful as hearing people's opinions in a more natural setting. Social media gives researchers the glimpse into conversations before and after the shock at different temporal resolutions.

If we conduct this analysis using social media, the most straightforward approach is to begin by looking at what users are posting on the topic of interest. The topic may be determined by looking at posts that are part of a topical group (BlackLivesMatter or meToo) or posts that contain topical words and/or hashtags (#BlackLivesMatter, #GeorgeFloyd). We may take a random sample of users who are posting on the topic or a stratified sample of users with particular demographics and platform usage properties (activity level - how often users post) to capture user-level attitudes for a subset of those posting. Part of this decision will be based on the availability of data, as well as the capability of the research team to collect and process the data.

We may also only be interested in looking at attitudes in different regions of the country or world. In this case, geography information about where the posts originated, timestamp of when posts were sent, and user activity level to capture frequency at which users post this content would all be important data to collect.

Once we identify our data set, we need to determine what measures can be constructed from the text and/or image data. Typically, the data are large and therefore, manual labeling of all the data may not be viable. Instead, any labeling effort would focus on creating a training data set that could be used by a machine learning algorithm (or dictionary-based method) to construct the measure of interest. In this example, we may decide that the stance on BlackLivesMatter (pro-BLM or anti-BLM) is a reasonable proxy for racial attitudes toward blacks. Given this, we would need to label a subset of relevant posts in order to create a training data set for a machine learning algorithm to predict pro-BLM or anti-BLM positions. Classic approaches to stance detection have been dictionary-based or machine learning based. We can test different machine learning methods (decision trees, Random forest, Naive Bayes, regularized logistic regression, etc.) to determine which model is the best given the sparsity of the data. Because of how noisy social media data sets are, people have begun using more sophisticated neural models to accurately determine stance from short, noisy posts (see Kawintiranon & Singh, 2021). We can also test those methods to determine whether or not they perform better, in terms of accuracy and reliability, on our specific stance task.

After labeling posts, researchers can aggregate the data to the user level to determine the fraction of users who are pro-BLM vs anti-BLM and monitor how that changes before and after other system shocks. One simple way to model this is as a dynamic sequence of binary variables, Y(I, p, t), where I is the individual, p is the post, and t is the time or sequence number. We can also consider how the shock relates to an individual's decision making. For example, what is the relationship between BLM stance and participation in protests?

Race, gender, age and socioeconomic status are key demographic characteristics for different user-level analysis. For many applications, geography either at the state level or preferably county or zip code level is also important. Typically, these are not all available to the

researcher and must be inferred using different algorithms. Demographic inference is a large area of research in computational linguistics and computer science. Methods have been developed for inferring gender, age, location, and other common demographics (Liu et al., 2021; Chamberlain et al., 2017; Chen et al., 2015; Ciot et al., 2013; Culotta et al., 2015; Hinds & Joinson, 2018; Ikawa et al., 2012; Mislove et al., 2011). One of the limitations of the existing methods is that they are built for a specific sample of data and do not transfer well to other sets of data. This is a result of a limited amount of labeled data. As labeled datasets get larger, these methods will also get more robust and be adaptable to new settings. Computer scientists can explore the use of transfer learning and active learning to reduce this limitation.

We note that these data can also be used to look at how conversation stance shifts over time when different events occur. Instead of focusing on the user level, the conversation stance analysis focuses on looking at all the posts about the topic to see how the volume and the content of the posts change over time. Changepoint analysis is a good option for this type of study. Topic modeling can also be used to group words into topics based on how frequently they are used together (see Churchill and Singh (in press) for survey of methods).

Another analysis may be at different geographic/spatial levels. For example, it may be interesting to identify levels of racism using stance at different locations to understand possible relationships between racism on different outcomes, e.g. adverse birth outcomes. Visuals can include maps of the geographic distribution of racism or plots to show variation across time.

There are many different mathematical models to choose from once all the variables of interest have been constructed, from traditional regression models to hierarchical regression models. Models can also be formulated within a Bayesian framework. Using hierarchical regression or hierarchical Bayesian models allow for incorporation of individual, group and state level structure or different temporal scales. Because of the potentially large amount of missing data, this flexibility can be important. When the number of variables is large, using heatmaps to see how highly correlated different variables are can help identify redundant variables within an analysis, as well as those that may be more unique. Diagnostic visualizations that highlight levels of convergence of different models can also be insightful.

Finally, network analysis can be used to investigate group formation (when do members join the group or begin sharing posts), community structure (how densely connected are communities with difference stance on BLM), or how information or misinformation travels through the network (do posts with certain content or from certain users spread faster?). Resnick and colleagues developed a suite of interactive tools for identifying rumors and assessing their impact visually (Resnick et al., 2014). There are also a large number of tools for visualizing networks. For example, Chau and colleagues have a tool that enables researchers to explore networks in their browser[20].

[20] Polo Club of Data Science. (2021). *Argo Lite.* https://poloclub.github.io/argo-graph-lite/

Analysis Challenge: One common type of error associated with social media, particularly after a shock, is selection bias. Not everyone uses social media, those who do use it with different frequency, and the majority of people sharing posts will be those who feel strongly about the issue. Selection bias can be investigated by potentially looking at the percentage of the population that uses the specific social media platform and whether different demographic groups are over- and under-represented. Additionally, internet access can be an issue in rural areas, further impacting population representation.

Measurement error is part and parcel of any model development, but may be more variable when a shock occurs. A standard regression model incorporates in the error terms both model deviation and measurement error. When the error is random, these are assumed to be random and zero centered. The key is to assess whether the errors can be assumed to be zero centered. For these data, sensitivity analysis may be more meaningful. That is, what is the extent of deviation from the zero-center that will make the findings a "null" finding? If it is a very large deviation then we can consider trusting the measure. If it is a small deviation, then it may not make sense to trust it.

4.2. Planned Event Case Study

Overview: In this case we are exploring how to use an announced event or a planned event like Thanksgiving or another holiday to examine social media data and the types of questions that could be asked of them. Examples of unique announced events include reopening dates, dates of NASA launches, and planned military withdrawals. Examples of recurrent planned events include holidays, elections, and the beginning of the school year. As an example of a visualization of the number of Covid-19 cases before and after the planned event of Thanksgiving, we refer you to a BBC visualization.[21]

For this case we explore changes related to the end of the pandemic in the United States as denoted by an announced event by the U.S. government. Just weeks after taking office and with three vaccines approved by the FDA (Pfizer, Moderna, Johnson and Johnson), President Biden noted that he hoped to see the end of all the Coronavirus pandemic restrictions by July 4th, 2021. Given this unofficial ending time when all restrictions would be lifted, researchers may be interested in what behaviors are occurring as a result of the expectation of the "end" of the pandemic. The questions that are of interest include will mask wearing be close to zero within the U.S by this date? Does this differ by demographic variables like completed education, income, political affiliation, race, and gender? Does it differ by case rates in the geographical area? Do we see sentiment changing (optimism vs. pessimism) as we get closer to July 4th? Computer scientists may be interested in building models for using the changes in conversation dynamics (topics, volume, etc.) to try to predict mask wearing in different regions of the country at different points in time.

[21] "Cases continued rising after Thanksgiving" from Horton, J. (2020, December 22). Covid: Thanksgiving the cause of a spike in US infections?. *BBC.* https://www.bbc.com/news/55363256

Analysis Plan: If we approach these questions using traditional methods, we might employ a time series design where we ask questions related to mask wearing and reasons for wearing or not wearing one when outside the individual's residence. The survey would be designed to answer some of the following questions:

- Are you currently wearing a mask when you interact with others outside your immediate family or residence?
- Can you tell us your reasons for wearing a mask? (case rates in your area, concern for the lack of vaccination rates in the area, general distrust of others, health reasons, other; this can also be a fully open-ended question)
- Can you tell us your reasons for no longer wearing a mask? (same options as above with slight differences in wording)
- When did you first remove your mask when going out in public?
- What was the reason for removing the mask?
- Do you think others should still be wearing a mask?
- Will you be attending a July 4th event and will you wear a mask?

The responses to these survey questions can be analyzed with simple descriptive statistics, a look at group differences, and time-series analyses. The questions are descriptive in nature and not predictive but one could imagine having specific prediction hypotheses related to different demographic groups having a stronger desire to stay protected against variants and so continuing to use masks in crowded situations but less so at smaller events.

Even though traditional survey methods could be used to assess this question, the addition of social media data would provide an added level of information that can be examined for differences prior to the Biden announcement of July 4th as the unofficial date for the ending of pandemic restrictions through time, until after the July 4th date. One could investigate how this is discussed on social media by looking at specific keyword mentions and hashtags (e.g. #nomoremasks, #maskless, #maskoff) related to mask-wearing or removal as well as the July 4th date (e.g. #maskindependenceday, #July4thmaskoff, #maskoff4th). Here is where data blending and triangulation of multiple sources of data (Singh et al., 2020) can provide a richer, more complete picture.

Another option is to collect survey data on social media platforms. A good example of this is work by the Delphi Group at Carnegie Mellon that has been tracking COVID-related data since the beginning of the pandemic by administering surveys through *Facebook* (2021). They provide a useful visualization[22] that shows the amount of mask wearing in the previous 7-days before the survey was administered. *Facebook* has a much broader user population than other social media platforms and that is good for getting a more diverse population, but the limitation

[22] Delphi Group (2021). Percentage of daily doctor visits that are due to COVID-like symptoms. https://delphi.cmu.edu/covidcast/classic/?date=20210603&region=42003&sensor=fb-survey-smoothed_wearing_mask_7d

of using surveys to gather real-time data the way you can get from *Facebook* or *Twitter* posts reduces some of the usefulness of these data. The visualizations that are used with these data are a good illustration of how mask wearing can be represented in relation to a time-series of survey data. This is also a useful way to examine social media data but with the added dynamic of seeing whether there is a gradual or dramatic shift in mask wearing on July 4th.

Another advantage of using social media to capture this planned event is the ability to examine the sentiment of these mentions and hashtags. In this case we may want to see if we can capture optimism or pessimism in the days leading up to and following this event, focusing on the changing dynamics. There are a number of factors that could affect relative optimism-pessimism at the individual and aggregated levels, including the pace of the pandemic (infections and hospitalizations should be declining), vaccination rates (should be increasing), and personal decisions about getting vaccinated. Sentiment is typically measured using dictionaries, but with the growth in neural models, more sophisticated techniques have emerged (Tian et al., 2020).

Social media methods can provide a very important enhancement to the survey methods regarding mask wearing and changes in adherence to mask wearing. Here, visualization can also give us important insight. For example, we can look at the relationship between changing sentiment on social media and changing attitudes in surveys and visually analyze whether or not a clear relationship can be extrapolated. We can also use visualizations to explore sentiment and discussion volume across different regions of the country. Who is still posting about wearing masks? What are the reasons given in the posts? Do we see sentiment changes toward mask wearing?

Finally, similar to the previous example, we can take advantage of network structures associated with social media data in order to understand how they may relate to disclosure of mask wearing or not. We cannot easily get this in traditional surveys other than asking if friends and family are wearing masks. It might also give us information on who their reference group is--more liberal, more conservative, highly concerned, no concerns at all. Also, social media would allow us to go "back in time" and pick up the initial adherence or rejection of masks. We might be able to observe this in existing data through surveys or public health data, but social media may provide a unique source of information about when it started and ended-which makes it a powerful data source.

Analysis Challenge: An important aspect of most social science questions is having a good understanding of the characteristics of the units from which the observations were made. Thus, like the previous case study, having demographics of the social media posters is important for understanding who is providing the information and how they match the general population. Along with demographics, case counts of COVID, and percent vaccinated in the geographical area will be useful in order to make any inference to behavior based on social media posts. One approach for handling this challenge is to consider blending or combining social media, survey, and administrative data. This would allow for better representation of the population and a richer data set for researchers. Of course, there are a number of challenges associated with blending

data including variability in data quality and reliability, lack of understanding of the data generation process, and ethical considerations to name a few (See Singh et al., 2020 for more discussion on this topic). One approach to mitigate some of these challenges is to model each data type and each data driver on social media (media coverage, presidential statements/activities, or disinformation attempts) separately. Then each can be compared to understand what variation is meaningful and what each data source adds to the analysis.

4.3. Economic Index Case

Overview: In this case study, we examine whether it would be possible to use social media data to provide timely and relatively inexpensive data to construct new indices of economic or political behavior and phenomena. Focusing on economic examples, the opportunities fall into three broad categories: (1) Can we replicate survey-based measures of economic activity or sentiment in a more timely or less expensive way using social media data? (2) Can we use social media data to construct new measures of previously unmeasured economic activity? (3) Can we use social media data to identify new forms of economic activity (or new products, services, or practices) that should be incorporated into existing survey-based measures?

Most previous work has focused on the first two questions. For example, in terms of survey replication, Antenucci and colleagues use data from Twitter to create indexes of job loss, job search, and job posting (Antenucci et al., 2014). Signals are derived by counting job-related phrases in Tweets such as "lost my job." The index had some strengths, like being able to see the impact of events like hurricanes and government shutdowns in real time. The authors also highlighted different challenges, including reliance on keywords that continually changed, requiring continual recalibration of the index. Political polling is another example that has received a lot of attention (Pasek & Dailey, 2018; Jungherr et al., 2011; Huberty, 2015; DiGrazia et al., 2013). Despite some apparent early successes, it has become clear that this is a relatively difficult problem and that much of what appeared to be success was actually some combination of luck and data massaging.

There is also work constructing new measures of previously unmeasured activity. A team of economists at NYU worked with Facebook to construct a social connectedness index (SCI). The measure is based on an anonymized Facebook social graph (geography and friendship likelihood) at various levels of geographic aggregation: US counties, other countries, subnational regions. The data show a strong impact of geography on social connections as can be seen in the New York Times interactive visualization of the connectedness index[23]. In a series of papers, Kuchler and co-authors examine the usefulness of the social connectedness index in explaining economic phenomena such as the impact of social connectedness on international trade, housing markets, and social behavior during COVID. (Bailey et al., 2018a; Bailey et al., 2018b; Bailey et al., 2020; Bailey et al., 2021; Kuchler et al., 2020; Kuchler & Ströbel, 2021; Ströbel et al., 2019).

[23] Badger, E., & Bui, Q. (2018). How connected is your community to everywhere else in America. *New York Times.* https://www.nytimes.com/interactive/2018/09/19/upshot/facebook-county-friendships.html

Finally, the third category is more speculative. For example, using social media to identify new forms of economic activity. We ask the following questions: Can we use social media data to inform traditional survey-based collection methods, for example, could social media data identify new practices, products, services, or issues that need to be incorporated in existing survey collection programs? Can we identify new and different measures of economic activity that help us understand changes in the structure of economic activity? Social media data seem well-suited to exploratory analysis to identify these kinds of new developments that traditional survey or administrative data-based systems would miss because of outdated classification systems or dependence on traditional, long-standing survey questions.

A specific example of this type of application is to consider whether and, if so, how employers are changing their remote work policies post-COVID-19? Which firms/industries/regions are changing them and how? How do employees feel about this, and what are the likely implications for labor markets? We will explore these questions for our final case study.

Analysis Plan: To consider how one might approach this issue, let's start with the challenge of measuring changes in firm behavior, rather than sentiment. Ideally, one could look at postings on, for example, LinkedIn, identify companies that mention remote work or flexibility, and look to see how many are posted over time and how many are shared over time. These counts could be used to capture the changing or stable dynamics of the firm's remote work policy at different points during the pandemic. One could imagine an index that monitors the changes in these policies over time. One issue with this approach is that it would only capture organizations that are hiring. Another option would be to look at discussion groups or other fora on Facebook or similar platforms of HR professionals who might post about policy changes. Additionally, one could try to capture posts by employees about going back to the office (or not) and if not, identify what the alternatives are. This may require linking accounts from multiple social media platforms, e.g. Facebook and LinkedIn, or at least considering ways to capture information across multiple platforms about specific companies. For example, researchers could identify words, phrases, and icons that employers use to indicate work from home or remote work flexibility in postings of new jobs (on LinkedIn, Facebook, Twitter), as well as identify postings by employees, potential employees, or past employees indicating job choice (quitting, taking, changing) and job behavior associated with remote work.

In order to assess how important changes in employer behavior might be in the labor market, it would be useful to assess employee (or potential employee) reactions. Are workers interested in the types of remote work or other alternative arrangements that employers are offering? One possible way to assess this would be to investigate worker sentiment and stance in social media postings. This would require identifying expressions of sentiment associated with remote work (missing colleagues, cursing employers who make you go back into the office) and changes in reported home location of social media posters who have not changed jobs, possibly through updates to a user's profile or changes in geotags of posts. For this case example,

machine learning methods may be more promising than dictionary-based ones since some of the meaningful phrases, words, and emojis may be more complex to determine manually. Specifically for stance, one could begin by learning words that are related to remote work policies. That could be done using topic modeling. Because the topic may not be that prevalent in the data set, new techniques may need to be developed to identify important low to mid frequency words. Using visualizations to see clusters of words and phrases that occur more frequently with relevant words and phrases of interest can help identify relevant concepts. Then tweets containing relevant words and phrases from the remote work policy topics could be used for stance analysis. A training data set that has examples of positive and negative stances can be created by experts and then used by the different learning algorithms to predict stance.

The traditional alternative to identify this type of information would be a survey of firms. While it would be more cost-effective if administrative data from firms were available to assess this question, administrative data do not usually include such detailed, rich information. It is possible that one might be able to infer changes in remote work if there were some shift in taxes paid to municipalities, i.e. a change in income tax when people moved to work from home. It is also possible that firm administrative systems are not robust enough to capture this type of information.

A benefit of this type of data collection is that we would be able to measure activity in real time, which is what we need as this is something that is happening right now. There have been previous surveys of measures of remote work and attitudes toward remote work, but practices and attitudes are changing rapidly as a result of work from home. Social media seems like a place that might allow one to capture those shifts.

Using social media data to construct new indices seems promising in that the new index could be better matched to the type of social media data available. The social connectedness index is a nice example of using the strength of social media data to construct something that would be difficult to do in a traditional survey context. In this particular application, the availability of a fairly large sample increases the likelihood of "representativeness" and the availability of geographic identifiers makes a number of interesting applications possible. Constructing new indices allows researchers to choose a use of data that capitalizes on social media data strength and avoids some of its weaknesses.

Analysis Challenges: The challenges mentioned for the other cases are present for this case. In each of these three potential applications (as well as the other two cases), the lack of information about the underlying sample of social media data could, to some degree, prove an analytical challenge. This issue poses the biggest challenge in the first potential application, replacing an existing survey with social media data. In this setting, not understanding the underlying population would pose significant challenges to successfully using social media data to replace survey data.

One challenge with this type of data collection is that it would be difficult to determine what the baseline was. Because we want to construct indices, we need to establish a norm (or a

norm band) from which we understand fluctuations and deviations. Another challenge, at least for some platforms like Twitter, is the difficulty of associating an account with a geographical place. This is a problem for developing measures associated with a particular geography.

In general, building indices using social media data depends on a series of assumptions about what is being measured and how that measurement relates to the population as a whole. For our case example, researchers need to believe that social media data capture something essential about firms and workers at the firms, that the process yields metrics which are either reflective of the population of firms and people at large or could be given some set of adjustments, and that that process is consistent in its reflection of the population over time, at least once any adjustments are accounted for. These are all relatively high bars.

Taking Twitter as an example, we know that one in five Americans are on Twitter and that representation on Twitter is lower in most other countries. Presumptions that the data reflect the population either depend on a joint assumption that the processes involved in joining and in posting relevant content on Twitter are more or less random -- which seems highly unlikely -- or that the data generating process yields an appropriately broad and sensitive coverage of work from home attitudes such that these can be treated as a proxy for the population as a whole. The latter might exist, for instance, if tweets are reflective of the same media coverage or messaging that shapes population-level attitudes.

The ideal measure for such a project is somewhat unclear. Some studies have attempted to use the mere mentions of one position versus another as the metric, others have examined the positivity or negativity of sentiment analysis, and one could even imagine approaches that are more data driven and allow any of a series of related constructs to predict. Notably, data driven metrics are difficult to reliably construct given the challenges associated with benchmarking this sort of data. They are also impacted by unusual events and may need to be continually recalibrated because of that. The recalibration of a social media index is not well understood.

5. The Ethics of Social Media Analysis

There are many different ethical and fairness issues that arise when considering the use of social media. We refer you to the previous white papers for a more detailed discussion of ethical and fairness issues specific to other components of the research life cycle. Here we focus on themes that emerged during this meeting and issues that arise during analysis and visualization.

There is an inherent tension when discussing confidentiality for data that are in the public domain. Because data do not “come from God,” is consent needed for public data? Because there are no consistent guidelines, it is important to make sure there is clear documentation about different forms of error (random/systemic), transparency about accuracy and fairness measures, clear lists of assumptions and a minimal assessment of reliability and validity.

In all of the cases presented here, demographic information was important. There are obvious ethical concerns related to accuracy, bias, and fairness when algorithms are inferring

demographics. When polled at our meeting, social scientists were comfortable with machine learning algorithms that were inferring demographics if the accuracy of the algorithms was over 75% and the algorithms were "fair" for each subgroup of the prediction task. There was some concern about inference of complex demographics like race, extremist behavior and suicidal intent. Part of the concern stemmed from certain uses of the data or entities who may use the inferences, e.g. government agencies like ICE inferring immigration status. The other part had to do with the complexity of the measure itself. Measures of suicide/depression are complex to develop with conventional data sources. Developing ones that use social media and maintain high levels of validity and reliability are not straightforward. In general, researchers do not have enough experience operationalizing more complex measures. How will errors in measurement impact individuals? Will errors be distributed randomly, i.e. fairly? Reporting differences in groups when there are errors that are not well understood could have major implications. Still, if we could quantify the measurement error, using it for social science research is potentially promising.

Another important ethical concern is reproducibility. This is typically not mapped to ethical concerns, but given that platforms continually change the way they present data and refine the algorithms generating recommendations, the stimuli for posts may also be changing over time. This means that without knowing all the changes made on a platform, it may be difficult to replicate conditions of an earlier study even if the researcher has access to the relevant data. We do not clearly understand the impact of this type of reproducibility issue for scientific progress.

Finally, bad visualizations can lead to misrepresentations of data and misunderstandings of research results. The public has limited experience understanding how to spot a misleading visualization. There are many examples of bad visualizations (see Wiseman, 2015 about poor quality map visualizations). When we produce visualizations and use them to convey research to the public, we must ensure that they cannot be used to misrepresent the work.

6. Accelerating Research in this Arena

The use of social media data in computer science and social science is relatively new, and it involves a number of challenges. Continuing experimentation with social media data will accelerate acceptance, use, and development of tools for analyzing and visualizing social media data.

It is important to think through the types of artifacts, tools, or directions that would help accelerate research using social media data. Here we group the ideas shared into five broad areas: shared research on social media data characteristics/quality, repositories and resources, analysis templates, modernization of publications, improved visual tools, and broadened access to computing infrastructure.

Repositories and resources: Because many social scientists are not trained to use visualizations throughout their research processes, one important step is to create a repository of

visual analytic tools for different parts of the analytic pipeline. Similarly, it is important to share R and Python packages and scripts that walk through a complete analysis of different social media data sets. A challenge to using voluntary repositories is that not all researchers archive their tools and, even if researchers do archive their tools, sometimes researchers do not update the tools with new libraries and packages as they become available. Developing norms within the variety of communities that use these types of tools would be a beneficial development and help accelerate research.

Analysis templates: It would be useful to have a set of templates for analyses conducted in different fields using social media data. This would also help students who have advisers less familiar with the issues and the techniques.

Modernization of publications: Many journals use static visualizations that do not change as more information is gained or as more research is completed. A new model may allow publications, specifically the visual results they include, to be updated in real time as new data are collected or processed. Embedded animations and dynamic visuals are now possible with pdfs. Having results that are "more alive" in journals and providing applications to create those visuals in a simple way, is important for moving the needle forward.

Improved interactive visual tools: While visual analytics have benefitted from great strides in recent years, there are still important ways to improve them. Examples include: multi-dimensional (3-D/4-D over time) scatter plots that allow (easy/rapid) changes to axes, tools to produce multi-dimensional scatter plots of regression estimates that allow researchers to iteratively make changes to regression model, and tools that easily show how models are working and the impact of different assumptions made.

Broadened access to computing infrastructure: Social media data are relatively large unstructured data sets that require significant storage and processing before use in analysis and visualization. Not all researchers have access to the computing infrastructure required to support these types of data collection and analysis. The research community needs to find ways to expand access to computing infrastructure. While NSF and other agencies have funded the creation of different infrastructure hubs, most researchers are not aware of what types of computing infrastructures are available or how to access them long term.

7. Conclusions

Empirical social science research has primarily focused on looking at differences between groups or trying to predict outcomes of interest. Social scientists have moved away from simple statistical description as their primary and only way to examine a phenomenon, e.g. see recent issue of Nature[24] and new journal on quantitative descriptive social science.[25] Social media provides an enormous amount of unstructured data that could potentially provide rich

[24] Skipper, M. (Ed.) (2021). Computational social science [Special Issue]. *Nature.* https://www.nature.com/collections/cadaddgige/
[25] Journal of Quantitative Description. (2021). https://journalqd.org/

descriptive information on beliefs and behaviors. However, the unstructured nature of the data poses challenges to the traditional analytic focus on hypothesis testing and inferential statistics that dominate social science fields. It is not even clear that significance testing is appropriate for this type of data due the amount of data collected (often in the 100's of thousands of posts) versus providing a rich description of the data. The convergence of descriptive analyses with visual analytics is a promising way to think about how to describe and show changes in beliefs and behaviors across time, but the challenge ahead is to the need to evaluate whether perceived changes are true signals or simply noise.

This white paper details analysis challenges associated with using social media data. It also provides a list of questions that should be considered by researchers about each stage of the analysis pipeline. This checklist can be used by researchers embarking on using social media data to ensure that they have thought through the main challenges of using these data. We then discuss three different types of cases (shock, planned event, index) that walk through possible research questions, data needs, measurement construction, model usage, and analysis. While a number of the same challenges emerge for each case, e.g. lack of knowledge about each platform's user population and missing geography information, there are also challenges specific to the different types of analyses. Fairness and ethical considerations can also not be ignored, particularly when algorithms are being used to infer important demographics, behaviors and attitudes of individuals.

Finally, given the rich amount of data coming from social media at the descriptive level, the use of visual analytics is especially useful in showing how the data is connected across multiple levels from individuals to countries. Computer science has been especially innovative in creating useful ways to visualize the large amount of data disseminated through social media. However, the social sciences are lagging in the use of data visualizations for their data. They are often still using basic visualizations such as line graphs, histograms, and pie charts to describe their data at both the descriptive and inferential levels. This may be due to adhering to publishing rules that reduce the innovation of visuals or to the lack of knowledge of various ways to visualize data in new and interesting ways.

The good news is that some exciting examples are beginning to emerge in the social science literature and shown in this paper. Finding the most productive uses of social media data will require experimentation across a range of computer science and social science fields. Cross-fertilization between computer science and social science communities shows promise as differences in approach and emphasis appear likely to lead to productive collaboration.

Broadening access to existing analysis and visualization tools and developing new analysis and visualization tools for large, unstructured data will accelerate experimentation with social media data. The analysis lifecycle is more complex when considering social media, therefore, it is more important to explain how different parts of the analysis were conducted and how the analysis relates to other more established disciplinary work. We hope this paper provides a framework for exciting collaborations between computer and social scientists on

addressing rigorous analytics and interesting and novel visualizations for social media data in the future.

Acknowledgements

We would like to thank the National Science Foundation and the McCourt School's Massive Data Institute (MDI) at Georgetown University for supporting this collaborative meeting. This white paper is an output of that meeting and is co-authored by those who participated in the meeting. This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. A special thanks to Rebecca Vanarsdall from MDI for her support in planning the meeting and helping prepare this white paper. We also want to thank all the students and staff who helped facilitate and take notes during the meeting. Those notes were invaluable when putting this document together.

References

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). *Using social media to measure labor market flows* (No. w20010). National Bureau of Economic Research.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018a). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, *32*(3), 259-80.

Bailey, M., Cao, R., Kuchler, T., & Stroebel, J. (2018b). The economic effects of social networks: Evidence from the housing market. *Journal of Political Economy*, *126*(6), 2224-2276.

Bailey, M., Johnston, D., Kuchler, T., Russel, D., & Stroebel, J. (2020). The determinants of social connectedness in Europe. In S. Aref et al. (Eds.) *Social Informatics. SocInfo 2020. Lecture Notes in Computer Science*, vol 12467. Springer, Cham.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R., & Stroebel, J. (2021). International trade and social connectedness. *Journal of International Economics*, *129*, 103418.

Bode, L., Davis-Kean, P., Singh, L., Berger-Wolf, T., Budak, C., Chi, G., Guess, A., Hill, J., Hughes, A., Jensen, J. B., Kreuter, F., Ladd, J. M., Little, M., Mneimneh, Z., Munger, K., Pasek, J., Raghunathan, T., Ryan, R., Soroka, S., & Traugott, M. (2020). Study designs for quantitative social science research using social media. *PsyArXiv.* https://doi.org/10.31234/osf.io/zp8q2

Budak, C., Kawintiranon, K., Singh, L., & Soroka, S. (2020). Real-time analysis shows that the first debate shifted attitudes among Twitter users towards Biden and the second solidified them. *USApp–American Politics and Policy Blog*.

Budak, C., Soroka, S., Singh, L., Bailey, M., Bode, L., Chawla, N., Davis-Kean, P., De Choudhury, M., De Veaux, R., Hahn, U., Jensen, J. B., Ladd, J., Mneimneh, Z., Pasek, J., Raghunathan, T., Ryan, R., Smith, N. A., Stohr, K., Traugott, M. (2021). Modeling considerations for quantitative social science research using social media data. *PsyArXiv.* https://doi.org/10.31234/osf.io/3e2ux

Chamberlain, B., Humby, C., and Deisenroth, M. (2017). Probabilistic inference of Twitter users' age based on what they follow. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 191 - 203). Skopje, Macedonia.

Chen, X., Wang, Y., Agichtein, E., and Wang, F. (2015). A comparative study of demographic attribute inference in Twitter. In *Proceedings of the AAAI Conference on Weblogs and Social Media* (pp. 590 - 593). Oxford, United Kingdom.

Chen, S., Lin, L., & Yuan, X. (2017). Social media visual analytics. In *Computer Graphics Forum, 36*(3), pp. 563-587.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of Twitter users in non-English contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1136 - 1145). Seattle, WA, USA.

Culotta, A., Kumar, N., and Cutler, J. (2015). Predicting the demographics of Twitter users from website traffic data. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 72 - 78). Austin, TX, USA.

Delphi Group (2021). Percentage of daily doctor visits that are due to COVID-like symptoms. https://delphi.cmu.edu/covidcast/classic/?date=20210603&region=42003&sensor=fb-survey-smoothed_wearing_mask_7d

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one, 8*(11), e79449.

Freelon, D., & Karpf, D. (2015). Of big birds and bayonets: Hybrid Twitter interactivity in the 2012 presidential debates. *Information, Communication & Society, 18*(4), 390-406.

Guo, Y., Guo, S., Jin, Z., Kaul, S., Gotz, D., & Cao, N. (2020). Survey on visual analysis of event sequence data. *arXiv preprint arXiv:2006.14291*.

Huberty, M. (2015). Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. *International Journal of Forecasting, 31*(3), 992-1007.

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the German election of 2009 or the trouble with predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review, 30*(2), 229-234.

Kawintiranon, K., & Singh, L. (2021). Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4725-4735). Virtual.

Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In *Information visualization* (pp. 154-175). Springer, Berlin, Heidelberg.

Kuchler, T., Russel, D., & Stroebel, J. (2020). *The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook* (No. w26990). National Bureau of Economic Research.

Kuchler, T., & Stroebel, J. (2020). *Social finance* (No. w27973). National Bureau of Economic Research.

Hinds, J. and Joinson, A. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PloS one.*

Horton, J. (2020, December 22). Covid: Thanksgiving the cause of a spike in US infections?. *BBC.* https://www.bbc.com/news/55363256

Ikawa, Y., Enoki, M., and Tatsubori, M. (2012). Location inference using microblog messages. In *Proceedings of the International Conference on World Wide Web* (pp. 687 – 690). New York, NY, USA.

Ladd, J., Ryan, R., Singh, L., Bode, L., Budak, C., Conrad, F., Cooksey, E., Davis-Kean, P., Dworak-Fisher, K., Freelon, D., Hopkins, D., Jensen, J. B., Kelley, K., Miller, R., Mneimneh, Z., Pasek, J., Raghunathan, T., Gresenz, C. R., Roy, S., Soroka, S., & Traugott, M. (2020). Measurement considerations for quantitative social science research using social media data. *PsyArXiv.* https://doi.org/10.31234/osf.io/ga6nc

Liu, Y., Singh, L., & Mneimneh, Z. (2021). A comparative analysis of classic and deep learning models for inferring gender and age of Twitter users. In *Proceedings of the International Conference on Deep Learning Theory and Applications.* Virtual.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., and Rosenquist, J. (2011). Understanding the demographics of Twitter users. In *AAAI Conference on Weblogs and Social Media* (pp. 554 - 557). Barcelona, Spain.

Mneimneh, Z., Pasek, J., Singh, L., Best, R., Bode, L., Bruch, E., Budak, C., Davis-Kean, P., Donato, K., Ellison, N., Gelman, A., Groshen, E., Hemphill, L., Hobbs, W., Jensen, J. B., Karypis, G., Ladd, J., O'Hara, A., Raghunathan, T., Resnik, P., Ryan, R., ... & Wojcik, S. (2021). Data acquisition, sampling, and data preparation considerations for quantitative social science research using social media data. *PsyArXiv.* https://doi.org/10.31234/osf.io/k6vyj

Pasek, J., & Dailey, J. (2018). Why don't tweets consistently track elections?: Lessons from linking Twitter and survey data streams. In *Digital Discussions* (pp. 68-95). Routledge.

Resnick, P., Carton, S., Park, S., Shen, Y., & Zeffer, N. (2014). Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proceedings of Computational Journalism Conference* (pp. 1 - 5). New York, NY, USA.

Singh, L., Traugott, M., Bode, L., Budak, C., Davis-Kean, P. E., Guha, R., Ladd, J., Mneimneh, Z., Nguyen, Q., Pasek, J., Raghunathan, T., Ryan, R., Soroka, S., Wahedi, L. (2020). *Data blending: Haven't we been doing this for years?* [White paper]. Georgetown Massive Data Institute Report.

Ströbel, J., Bailey, M., Kuchler, T., & Farrell, P. (2019). Social connectedness in urban areas. *Centre for Economic Policy Research.*

Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., ... & Wu, F. (2020). SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint* arXiv:2005.05635.

Wiseman, A. (2015, June 25). When maps lie. *Bloomberg.* https://www.bloomberg.com/news/articles/2015-06-25/how-to-avoid-being-fooled-by-bad-maps