Survey Research Methods (2019)
Vol. 13, No. 1, pp. 123-135
doi:10.18148/srm/2019.v13i1.7467

ISSN 1864-3361
http://www.surveymethods.org

Exploring New Statistical Frontiers at the Intersection of Survey Science and Big Data: Convergence at "BigSurv18"

Craig A. Hill
RTI International

Paul Biemer
RTI International

Trent Buskirk
Bowling Green State University

Mario Callegaro
Google London

Ana Lucía Córdova Cazar
Universidad San Francisco de Quito

Adam Eck
Oberlin College

Lilli Japec
Statistics Sweden

Antje Kirchner
RTI International

Stas Kolenikov
Abt Associates

Lars Lyberg
Inizio

Patrick Sturgis
National Centre for Research Methods
University of Southampton

Held in October 2018, The Big Data Meets Survey Science conference, also known as "BigSurv18," provided a first-of-its-kind opportunity for survey researchers, statisticians, computer scientists, and data scientists to convene under the same roof. At this conference, scientists from multiple disciplines were able to exchange ideas about their work might influence and enhance the work of others. This was a landmark event, especially for survey researchers and statisticians, whose industry has been buffeted of late by falling response rates and rising costs at the same time as a proliferation of new tools and techniques, coupled with increasing availability of data, has resulted in "Big Data" approaches to describing and modelling human behavior.

Keywords: Big Data; artificial intelligence; machine learning; survey research; official statistics

1 Introduction

The ever-increasing availability of compute devices and the data they spawn present special opportunities for researchers—as well as challenges for researchers who are using tools that do not take full advantage of "big data." Creating a big tent under which survey scientists could share thoughts, ideas, complaints, and praise with computer and data scientists seemed, when we birthed this idea several years ago, like the opportunity of a lifetime. Our hope was to assemble experts from around the world to exchange ideas about how best to leverage massive data using ever-evolving data management and analytic techniques to examine (and solve) social science problems to which we have long applied survey research approaches; in the end, we believe we succeeded in, at least, getting those conversations started.

Under the auspices of the European Survey Research Association (ESRA), the first conference on Big Data Meets Survey Science (BigSurv18) took place in October 2018 at the Universitat Pompeu Fabra's Research and Expertise Centre for Survey Methodology (RECSM) in Barcelona, Spain, attended by over 400 researchers and practitioners from 45 countries. Interestingly, nearly 30 percent of registrants self-identified as computer or data scientists.

BigSurv18 began with 143 delegates taking one of four stimulating short-courses and another 24 delegates participating in the "Green City Hackathon[1]," co-organized with the city of Barcelona. The main conference kicked off with keynote addresses: "Automating Metadata Documentation"

Contact information: Craig A. Hill, RTI International (E-mail: chill@rti.org)

Editor's note: This article is not the kind of paper ususally published in SRM, and did not undergo SRM's standard peer review process. The aim of this article is to disseminate the contributions of the BigSurv18 conference, funded by the European Research Foundation (ESRA). By publishing this report, SRM documents this important event of its funding institution and encourages its readers to become part of the endeavour.

The papers mention in this article can be found at https://www.europeansurveyresearch.org/conferences/bigsurv.

[1]The Green City Hackathon asked six teams to investigate bicycle use in Barcelona. Teams produced descriptive analyses, usage maps, predictive analyses, digital tools, etc. aimed at increasing sustainability. The hackathon took place 24–25 October 2018 and utilized data from the Open Data BCN portal (http://opendata-ajuntament.barcelona.cat/en/), including population and demography of specific neighborhoods, information about bicycle use, traffic and accidents, and complaints about traffic in the city. See https://github.com/bigsurv18/bigsurv18.

by Julia Lane (New York University); "Data Science for Public Good" by Tom Smith (Office for National Statistics, UK); and, a plenary session organized by Frauke Kreuter (University of Mannheim, University of Maryland), entitled "Big Data, Surveys, and the Privacy/Ethics Challenge."

The main scientific program spoiled delegates for choice with 51 individual sessions (including 2 poster sessions) scheduled across 2 full days, and organized into six main tracks: 1) The New Survey Landscape; 2) Total Error and Data Quality; 3) Big Data in Official Statistics; 4) Combining Big Data with Survey Statistics: Methods and Applications; 5) Combining Big Data with Survey Statistics: Tools; and 6) The Fourth Paradigm: Regulations, Ethics, and Privacy. Below, we provide some highlights from each of these sections.

2 The New Survey Landscape

The earliest hard drives on personal computers had the capacity to store roughly 5 megabytes; now, however, typical personal computers can store thousands of times more than that (around 500 gigabytes). Furthermore, data and computing are not only limited to supercomputing centers, mainframe, or personal computers, but have become more mobile and virtual—nearly ubiquitous. Data are getting larger and computing is commonly distributed across actual laptops or desktop computers or stored and processed virtually in the cloud. The definition of "computer" is also evolving to encompass more and more aspects of our lives: from transportation (on-board computers in vehicles and self-driving cars) to everyday living (televisions, refrigerators, doorbells, thermostats, and countless other "smart" devices comprising the Internet of Things, or IoT). The burgeoning growth of IoT and compute power, and the increasing ubiquity of devices, virtual, and mobile environments have certainly created an unprecedented opportunity for survey researchers, social scientists, computer scientists, and government officials to track, measure, and better understand public opinion and the world around us, resulting in a new survey landscape.

The BigSurv18 conference provided a unique opportunity to discuss, brainstorm, and learn how to leverage both the power of Big Data and data science to better estimate public opinion and improve official statistics. As demonstrated by many presentations at this conference, the field of public opinion research is in flux; within this so-called "new landscape," social and survey researchers—along with government officials—are expanding their palate, now tasting sources and methods along a broad spectrum, including (but not limited to): 1) reimagining traditional survey research by leveraging new machine learning methods that improve efficiencies of traditional survey data collection, processing, and analysis; 2) augmenting traditional survey data with non-survey data (administrative, social media, or other Big Data sources) to improve estimates of public opinion and official statistics; 3) comparing estimates of public opinion and official statistics derived from survey data sources to those generated from Big Data or other non-survey data exclusively; and, 4) exploring new methods for enhancing survey data collection as well as automating the collection of non-survey web data. Below, we highlight some of the presentations made showcasing this new survey landscape.

First, in this track, we saw many examples of applying machine learning methods within traditional survey and social science research. In his recent book, Machine Learning, Alpaydin (2016) notes "Machine learning will help us make sense of an increasingly complex world. Already we are exposed to more data than what our sensors can cope with or our brains can process." At BigSurv18, this theme was constant: many presentations described using machine learning methods to improve various aspects of the survey data collection and estimation process. For example, Buskirk, Bear, and Bareham (2018) showed how unsupervised machine learning algorithms could be applied to create sampling strata from landline telephone bank information. Eck et al. (2018) applied supervised machine learning methods to create sampling frames of non-traditional sampling units, such as windmills, using satellite images. Amer et al. (2018) also applied machine learning methods designed for object-detection tasks to identify sampling units of interest within cells defined within a gridded population sampling frame. The estimated number of units within these smaller areas serves as auxiliary information for the gridded population sampling frames, thereby increasing the choices for sampling designs that can be applied as well as initial computation of sampling weights for these designs.

The application of machine learning methods extends beyond frame development and enhancement to the processing of data collected from fielded surveys: for example, McCreanor, Wronski, and Chen (2018) discussed how open-ended responses in the form of sentences, rather than single-word responses, from large scale surveys can quickly create Big textual Data. They demonstrated a new method that not only automates the labeling process, but allows a topic model to be applied to survey data across multiple time periods to compare trends in open-ended responses. Ye, Medway, and Kelley (2018) used data from a large national survey to examine ways in which natural language processing could be applied to classify open-ended survey responses. Using both supervised and unsupervised natural language processing methods, these authors showcased both the challenges and the potential gains in efficiency of using automated (compared to manual) coding methods. Matthews, Kyriakopoulos, and Holcekova (2018) also compared multiple machine learning methods for coding verbatim survey responses from the national crime survey for England and Wales. Their work led to the development of coding rules in which some responses can be auto-coded while others would be coded using human coders. Their method points to the potential to leverage the

power machine learning methods—in combination with the experience of human coders—to balance efficiency and error.

Machine learning can also be applied to survey data to generate estimates of substantive outcomes of interest: Bobashev and Wu (2018) compared various supervised machine learning methods including classification trees, lasso, random forests, neural nets, and others for predicting the propensity for multiple visits to the emergency room using multiple years of data collected from a national, probability-based repeated cross-sectional survey. During this presentation, they also examined how various methods can be employed to identify a key set of predictors from among the many available predictors for the propensity models more parsimonious.

Second, we also saw evidence of a new landscape being created by linking and combining survey data with non-survey data. Leveraging alternative data sources to improve or augment survey estimation is a growing trend. Big data and other non-survey data are often linked to survey data using probabilistic linkage models or through a common identifier such as a telephone number, address, or some other higher level geocode. Some of this work has focused on using these additional data sources to reduce response burden or as a way of completing otherwise missing items from surveys. For example, Mulry, Bates, and Virgile (2018) tested whether commercially-available lifestyle segmentation data improved response propensity models; indeed, they found a great deal of variation in predicted response scores between segments and very high correlation between mean segment response propensity and self-participation rates. Other work has focused on using alternate data sources to evaluate the survey response process and use the survey data to evaluate possible error sources in the alternative data. Datta, Ugarte, and Resnick (2018) provide an interesting example for linking a nationally-representative ABS sample survey to a commercial data source (Zillow.com) using the address as the primary linkage key, making use of the commercial data to evaluate survey eligibility and potential nonresponse bias, thereby improving the survey recruitment process. In the other direction, the linked survey data allowed the researchers to evaluate the coverage error for the commercial data that would otherwise not have been possible.

Third, another change we saw in this new landscape is employing big data and other non-survey data to generate estimates. Can, Engel, and Keck (2018) discussed the new role of surveys and social media data within the growing field of computational social science; specifically, they note that advances in computational capabilities and methods and increased availability of big data are changing the way social science researchers consider the use of surveys and survey data; instead, many researchers now want to examine multiple data sources to test theories and understand the world. Of course, different data sources may lead to different conclusions. For example, Pasek, McClain, Newport, and Marken (2018) found similar patterns across aggregated Twitter sentiment and daily probability-based survey estimates of presidential approval—but caution that underlying differences between Twitter data and survey data may be related to the fact that these data streams could be measuring substantively different things.

While the demise of "Google flu trends" appears to be the modern-day version of the "Literary Digest Poll of 1936," there is considerable variation in the degree to which these new data sources have been adopted and used in production. In this area, much work is currently focused on evaluation of new, non-survey data sources by: (1) comparing estimates derived from them to those derived from contemporaneous survey data or (2) using them directly to generate estimates, independent from survey data. For example, Hutchinson (2018) evaluated alternative point of sale data collected at the product, store, and national levels as a possible replacement for high-burdened enterprises that are likely not to respond to traditional retail surveys. Her work revealed a high level of consistency between estimates derived at the national and store level and tracking surveys administered over similar time periods. Can et al. (2018), cited above, applied topic modelling to compare the distribution of latent topics in open-ended survey data collected from probability-based surveys, nonprobability-based surveys, and data gathered from social media posts. While the distributions of the underlying topics were generally similar across the three different data sources, the authors did note some key differences (as hypothesized). Buelens, De Broe, Meijers, ten Bosch, and Puts (2018) combined weather data, tax rebate data, electricity meter readings, and other data to produce national estimates of solar power. Burke-Garcia, Edwards, and Yan (2018) provided a rich discussion of the opportunities and challenges of working with social media data from various platforms for measuring public opinion. Their work also offers examples and guidance on how various aspects of the survey process—from sampling participants to identifying eligibility to encouraging participation to data collection and mining of responses—can be carried out within the context of social media.

Finally, one of the emerging aspects of the new landscape relies on leveraging technology for increasing efficiencies in survey data collection or enumeration. For example, Amsbary and Dulaney (2018) used GIS tools, street view maps from Google, and other internet and commercial sources to create a virtual listing system that allowed field staff to virtually validate eligibility and add auxiliary data to a sample of commercial buildings. Beyond survey data, webscraping methods continue to advance as a viable method for data acquisition: ten Bosch, Windmeijer, van Delden, and van den Heuvel (2018) describe various web data sources that have been used to create or augment official statistics estimation;

additionally, they provided a framework within which to develop, deploy, and evaluate webscraping methods.

This new survey landscape appears lush and green, filled with opportunities. This new landscape looks as though it will provide gains in efficiency and reductions in costs or burden. It has new methods, approaches, and data sources that can be used in tandem with surveys—or as a substitute for surveys. As we trek through this new landscape, we will want to bring forward the expertise gained from many years spent in the old landscape, ensuring that those lessons are not forgotten as we tirelessly pursue answering the "why" questions we encounter every day in every landscape.

3 Total Error and Data Quality

This track focused on the errors associated with administrative data and other Big Data sources, particularly social media data. While a total error framework exists for complex survey data (c.f Biemer et al., 2017), no such framework has yet been developed or is in practice for data acquired from "Big Data" provenance.

Biemer and Amaya (2018) reviewed a range of total error frameworks that are available for assessing the quality of integrated and single source data sets as well as for evaluating the quality of hybrid-estimates—i.e., estimates derived from datasets that have been unified in some fashion. The authors described two of the frameworks in greater detail: one framework, the row-column-cell framework, considers the errors in a generic (rectangular) dataset that may affect the rows (primarily missing units), the columns (specification error), and the cells (content error and missing data). This row-column-cell framework also provides a starting point for a framework for estimation error; as such, it begins with a simple decomposition that parses the total error of the dataset mean into components for sample recruitment (i.e., the process of selecting population units for the data set) and data encoding (i.e., the process of recording information on characteristics of the recruited units). The sample recruitment component can be further divided into subcomponents for coverage, selection, and nonresponse, while the latter contains subcomponents for specification error, measurement error, and data processing error. Based upon this decomposition, the authors derived a formula to assess the total mean squared error of a sample mean.

The authors provided two illustrations demonstrating the utility of this formulation: one based upon data from an online real estate database (Zillow), which shows the total error in estimates of average housing unit square footage and a second from the 2016 U.S. Presidential election. The first example clearly emphasizes the importance of assessing both data encoding error and sample recruitment error. Many assessments of the quality of Big Data tend to focus on the latter error component—which may be miniscule in most applications. However, encoding error can quickly increase the total error and is often the driving error source in Big Data applications, as evidenced in the square footage data. Although the Zillow database has over 200 million records, its accuracy may be substantially inferior to a survey of 6000 housing units with a relatively low response rate, primarily as a result of data encoding error (i.e., the error in the recorded data). The second example uses data collected from national polls just prior to the election of Donald Trump as U.S. president in 2016, taken from Meng (2018). Although this dataset is comprised of over 2 million observations, the actual mean squared error of the estimated vote share for Donald Trump was equivalent to that of a random sample of 400 respondents.

Liao, Berzofsky, Thomas, Couzens, and Cooper (2018) discussed some strategies for assessing and addressing data quality issues for the largest administrative source of crime data in the United States: the FBI's National Incident Based Reporting System (NIBRS). NIBRS is an incident-based reporting system used by law enforcement agencies (LEAs) in the United States for collecting and reporting a variety of information on crime incidents. Still, only about 36% of the 18,000 LEAs in the United States submit their crime and arrest data to NIBRS; as a result, missing data is a big problem if it is to be used for generating national estimates. The authors described the National Crime Statistics Exchange (NCS-X) Initiative, the objective of which is to transition a sample of 400 non-reporting LEAs to NIBRS; combined with the 6,600 LEAs currently reporting, these additional LEAs could provide nationally-representative estimates of crime victimizations. The NCS-X involved two phases: first, the study team developed and tested methods for data quality assessment, unit and item nonresponse adjustments, and the generation of national estimates; in the second phase, they will construct a prototype automated system designed to produce national estimates much more quickly.

Liao et al. shared their experiences dealing with data quality challenges and producing timely statistics for NCS-X; in fact, many of these can be extended to other administrative data sources. For example, the hierarchical structure of the database results in missing data at each level of the hierarchy: a single crime incident can have multiple offense types, victims, and offenders, so the NIBRS data are broken down and stored at incident, victim, and offense levels. This paper examined data quality at each level as well as for combined or aggregated data; in addition, the authors developed methods for compensating for noncoverage and nonresponse in NIBRS while accounting for the complexity of the data structure. During the presentation, the authors also discussed the feasibility of producing the automated prototype for the second phase of the study. Finally, they reviewed data quality, accessibility, and timeliness issues for auxiliary data sources necessary to address estimation issues in NIBRS. This paper was an informative case study demonstrating several impor-

tant data quality and management issues associated with Big Data, as well as offering possible solutions.

Also in this track, two papers investigated the use of social media for informing decision making (Amaya, Bach, Kreuter, & Keusch, 2018; Pasek et al., 2018). These papers are particularly relevant given the near-ubiquity of social media use and its associated data stream. These data are constant, instantaneous, often free to access, and do not suffer from many of the types of errors that plague survey data; however, these data suffer from unique validity risks. Coding error may occur when incorrect models are used to convert streams of text to analytic variables. Or, missingness may occur if non-users are different than social media users or when users who post on the topic of interest are different than users who do not post on that topic. And, measurement error may occur when individuals post only part of their attitude on social media, obviating the ability of the social scientist to see the full picture.

Pasek et al. (2018) examine the theory that tweets (and other digital traces) reveal what political issues people are thinking about and what kinds of events they find salient. In the context of political expressions, this attention-based explanation has largely replaced an expectation that tweets about elections would reveal the likely winner. If true, an attention-based understanding of tweets would render social traces a powerful tool for testing a series of expectations about the conditions under which information in the news filters down to what people are thinking about.

Leveraging rich sources of social media and survey data, the authors examined the validity of social media as an expression of attention. They compared the presence and prominence of terms related to 242 keywords associated with 39 distinct events in both tweets and open-ended survey responses about the candidates running in the 2016 U.S. Presidential Election. Their analyses reveal that patterns of attention to events across data types are sometimes distinct and sometimes quite similar. Patterns of attention in the survey data, for instance, were much more variable over time than patterns of attention in the Twitter data; that is, event-related terms were concentrated on a smaller number of days in the survey data than in the Twitter data. Despite these systematic differences, the authors found that mentions of terms related to most events peaked at similar times across modes, that the pertinence of events to one candidate or the other was reasonably consistent across modes, and that the variations in attention to those terms over time (as measured with correlation coefficients) was similar, implying that there was a strong signal underlying both types of attention measures.

But although patterns of attention were similar for most events, several exceptions for each metric indicate that what Twitter users are talking about and what the public is attending to is not always the same. Thus, tweets may indeed help researchers track public attention, but what Twitter users are attending to is sometimes different from what survey respondents are thinking about. It is not reasonable to assume that the events that matter to the general public are the same as those that prove salient on Twitter.

In a similar vein, the paper by Amaya et al. (2018) reported on research aimed at determining whether social media data can be used to measure the strength of attitudes, noting that limited research has been conducted to date that would isolate the reasons for error in social media data. They examined two specific research questions: 1) whether non-traditional data (text data) can be used to produce similar attitude distributions as survey data, and 2) what might account for observed differences. To answer these questions, the authors scraped and coded over 400,000 Reddit posts on 367 subreddits based in German-language countries to construct attitude scales on seven social topics: political ideology, interest in politics, immigration, the European Union (EU), trust in individuals, gay rights, and climate change. These social media-based distributions were compared to the attitude distributions acquired from the German version of the European Social Survey. The authors found significant and large differences between the two data sources. To isolate the error, the authors supplemented these data with two additional data sources: a survey conducted on Reddit and the survey respondents' Reddit posts. The authors conclude that error stems from several angles: the topic models did not accurately identify relevant posts; the Reddit population is not representative of the general population; and, redditors' posts do not reflect their true attitudes.

Also in this track, the paper by Vanhoof, Lee, and Smoreda (2018) presents an extensive empirical analysis of home detection methods using a national mobile phone dataset from France. The authors analyze the validity of nine simple, but different, Home Detection Algorithms (HDAs) and assess various sources of uncertainty in locating home origins. Based on 225 different set-ups for home detection of around 18 million users, the paper discussed different validation measures and investigated sensitivity to user choices such as HDA parameter choice and observation period restriction. They find that nationwide performance of home detection is moderate at best, with correlations to ground truth maximizing at 0.60. Additionally, the authors found that the time and duration of an observation have a clear effect on performance, and that the effect of HDA criteria and parameter choice are actually rather small compared to other uncertainties. These findings represent welcomed insights to other practitioners who want to apply home detection on similar datasets, or who need an assessment of the challenges and uncertainties related to leveraging mobile phone data for official statistics.

4 Big Data in Official Statistics

Producers of official statistics are currently facing challenges arising from increasing expectations associated with the near-ubiquity of data: users want—and expect—data that are both timelier and are made richer by tapping into extant administrative (or other) data. At the same time, official statistics based on traditional survey data are falling out of favor because nonresponse is increasing at alarming rates, threatening validity, and costs for survey data collection are increasing at equally alarming rates. Coupled with now-common demands for reduced respondent burden, statistical agencies have been forced to investigate new ways to produce official statistics. More and more frequently, statistical agencies are now considering using big data in its various forms more systematically and more routinely, sometimes in combination with other data sources (including surveys), to replace current data collection procedures. In this track, we saw many examples of this strategy.

Japec and Lyberg (2018) kicked off the track by describing several ongoing initiatives at national statistical agencies regarding the use of big data. Those initiatives concern not only the use of big data sources per se, but also changes regarding the way user needs regarding speed, contents, quality, and costs can be addressed. They pointed out that the world is transitioning from a probability-based sampling paradigm to a multiple data source paradigm, while, at the same time, stepping back from conceptual or theoretical purity to using best-available data sources (Citro, 2014). As a result, statisticians (and others) will have to embrace new theoretical developments regarding, among other things, nonprobability-based sampling and associated inferential issues, and how to combine or integrate different data sources; in other words, assert the authors, we are on the cusp of replacing "model-assisted surveys" with "survey-assisted modeling" in which survey data are but one component of the estimation process. Many national agencies now have their own Big Data centers where survey statisticians work together with data scientists; and, indeed, some agencies have now formed partnerships with academic and commercial big data units.

Japec and Lyberg (2018) described several potential use cases for big data in official statistics, including: replacing surveys entirely; combining big data with other, different kinds of data sources; exploring wholly new topics and concepts; and, performing data mining to identify new patterns and models. Their paper provided examples of applications of big data currently conducted by agencies, calling out the data sources. They also point out, however, that these new types of data and data acquisition would benefit from an adjusted quality framework (see above); in addition, they note, we have not yet completely solved issues related to privacy and confidentiality when national statistical agencies use administrative data or the implications that privacy concerns may have on participation, data capture, or resulting estimates.

The paper by Hutchinson (2018) described attempts at the U.S. Census Bureau to use point-of-sales or scanner data to reduce the extensive respondent burden in economic data collections, especially the multitude of monthly and annual retail surveys. The respondent burden in U.S. economic surveys can be measured in different ways, including actual costs, disruptive information demands (sometimes requiring several respondents within a store and whose accumulated responding work can add up to hundreds of hours annually), and number of forms received annually (for instance, 40% of retailers receive 6–10 forms each year). This heavy burden leads some retailers to simply abstain from participating in these voluntary surveys, reducing the response rate; as noted above, this falling response rate is colliding with the rising clamor for richer and timelier data. The Census Bureau's choice of scanner data as a first option regarding an alternative data source is logical, since it has been used in other countries to help produce consumer price indices. However, whenever an alternative data source is considered, quality requirements must be maintained. Scanner data, acquired from a third-party vendor that curates datasets for the Census Bureau, has been compared with survey data on the national, store, and product levels. The results are promising: correlation between the two sources is quite good and differences observed between them are often relatively small. The author was reporting on a pretest which, of course, has limitations: the pretest is small, and the Census Bureau has not yet settled on data quality metrics. Despite that, the results for good reporters were encouraging and the imputation for monthly nonresponders has worked well. They conclude that it is evident that just because there are data sources other than surveys available does not mean they can be used without proper testing.

Braaksma, Zeelenberg, and DeBroe (2018) provided examples from Statistics Netherlands such as the use of scanner data, social media messages, traffic-loop data, and mobile phone data to produce official statistics. The authors evaluated big data on three fundamental quality dimensions for a National Statistical Institute: accuracy, objectivity, and reliability. They make clear that there are, indeed, great opportunities for using big data in official statistics, which could result in increased efficiency, reduced respondent burden, and improved quality. But there are also challenges for such use, including coverage error and selectivity bias—and they note that some data sources may change or disappear altogether. Correcting for these biases often requires a modelling approach, but, traditionally, national statistical offices have been hesitant or reluctant to use model-based methods to produce official statistics, worried that model-based estimates would be viewed as less objective or trustworthy. Yet it is true that, in specific statistical areas, national statisti-

cal agencies have already employed model-based methods, for example, to correct for non-response, and to compute seasonally-adjusted time series, and to calculate preliminary macro-economic estimates. The authors argue that, since models are already being used in official statistics, the reluctance to use model-based methods for treating big data sources may be misplaced. To settle this issue, they emphasize the need for clear principles and guidelines (Buelens, de Wolf, & Zeelenberg, 2014) about the use of models in the production of official statistics. Furthermore, the authors stress the importance of transparency and that any use of models should be documented and made explicit to users.

Greenaway (2018) discussed three pilot studies carried out by the UK Office of National Statistics (ONS) designed to see if economic metrics could be improved or replaced with big data. The first pilot estimated the proportion of businesses conducting e-commerce on their websites by scraping them. Currently, ONS conducts an annual e-commerce survey of 5,000 businesses; in this pilot, they instead crawled the web to identify the websites of the businesses in the sample and then used a search API to look for the businesses' names. From this website content, they attempted to determine whether the website supported e-commerce activity. In the end, the estimate produced by using webscraped business data produced an estimate within the confidence interval of the survey estimate. Because this pilot covered only one time period, they now plan to attempt it across multiple periods to ensure that the method is robust.

ONS' second pilot investigated the possibility of estimating job vacancies by using data from online jobs portals and websites. ONS wants to provide short-term statistics like this to measure current (and fast-changing) economic conditions; to do so, ONS carries out a quarterly job vacancy survey, publishing the data about two months later. In this pilot, ONS compared estimates of job-vacancy counts from job portals to estimates from the job vacancy survey. They tested several different nowcasting models and an aggregate-level time-series approach showed quite a bit of promise, and this will be investigated further when more data are available.

Classification of businesses by type of economic activity is central to economic statistics. In the third pilot, free-text data describing the economic activity of UK businesses was used in order to gain new insights for industries not covered by Standard Industrial Classification (SIC). The results from the pilot show that it is possible to identify form of activity, which is not currently captured in SIC. Combining these data with survey or administrative data, it may be possible to produce estimates about the size of new economic sectors.

Tam's (2018) paper, "Mining the New Oil…" alludes to the statement by UK mathematician Clive Humby: "Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value." It is likely true that the value to users relies a great deal on how data are created and refined for statistics production. One common view is that data should be representative and as free as possible from measurement errors, but there is now an abundance of data sources available from the IoT (and other sources) that may not, at first blush, satisfy that view. Rather than discard these data, practitioners in official statistics need to find ways to refine them so that official statistics can produce richer data in a timelier fashion at a lower cost. This paper detailed efforts made at the Australian Bureau of Statistics of combining traditional sample data with some other Big Data sources. Tam showed that Big data sources—which usually suffer from coverage bias—can be integrated with probability sample data, yielding efficient finite population inference. These methods can also address situations where the variables in the Big Data source or the probability sample suffer from measurement errors, and when there is unit nonresponse in the probability sample. Tam demonstrated the efficacy of the methods using simulation and two case studies.

5 Combining Big Data with Survey Statistics: Methods and Applications

The exponential growth of computing power and the availability of cheap data storage—the age of ubiquity—have given rise to big data applications, including artificial intelligence and machine learning. BigSurv18 featured a variety of presentations investigating the potential of these applications and methods for survey research demonstrating, for example, how big data can be used to improve sampling efficiencies in traditional survey data collection, how big data can be used to complement or replace traditional survey data, or how machine learning can be used to improve the accuracy of estimates.

One example to use big data to improve traditional sampling approaches was presented by Ridenhour, McMichael, Krotki, and Speizer (2018). The authors demonstrate that the sampling efficiency of incomplete specific lists of target populations—such as lists of boat registrations—can be increased by appending target population flags from big data for all U.S. households; in this specific example, they used an address-based sample containing information from the U.S. Postal Service's Computerized Delivery Sequence file enhanced with auxiliary data from commercial vendors. The authors used this combined dataset to build a model identifying the target population of boat owners and then applied this model to all cases of the "big data." This approach generated a sample that not only increased coverage of the target population and screening rate, but also lowered overall data collection cost. Other applications using big data as auxiliary information to increase sampling efficiencies include stratification (Ridenhour & McMichael, 2017) and predicting eligi-

bility, and nonresponse adjustment (West, Wagner, Hubbard, & Gu, 2015).

Other research presented at BigSurv18 leveraged alternative data sources such as crowdsourcing data, image data, and digital media data instead of more traditional survey data to map individuals' attitudes and perceptions. These approaches may have the potential to reduce data collection cost and respondent burden. The study conducted by Buil-Gil, Solymosi, and Moretti. (2018), for example, used crowdsourcing instead of survey data, i.e., "outsourcing" of a problem usually solved within an organization to a "crowd" of volunteers (Salganik, 2018), to model crime and safety perceptions in Greater London. One of the big advantages of their approach was that data collection was more cost-efficient compared to traditional survey data collection; however, these data are not necessarily representative of the target population and may be of limited utility to researchers and policymakers. To resolve these issues, the authors employed a combination of a non-parametric bootstrap and small area estimation to mitigate bias in the crowdsourced data. Survey data in this example exclusively served as an external validation to assess bias and variability of these estimates. The simulation study and application showed that estimates improved (in terms of bias and variability) using the suggested approach. Diego-Rosell, Srinivasan, Dilday, and Nichols. (2018) explore the use of streetview and satellite imagery data to predict subjective well-being and health outcomes in Baltimore. The authors employed multispectral analysis of satellite imagery and crowdsourcing for image labeling. The image labeling task required significant labor (and thus cost) and had limited scalability and poor data quality (i.e., low reliability) for some indicators. Nonetheless, the authors demonstrated that the predictors derived from the imagery data explained 57–63% of the variability in the subjective well-being and health outcomes. Extrapolating model predictions from one city to the other, however, decreased explained variability, ranging from 13–18%. Similar to Buil-Gil et al., the authors relied on survey data to train and evaluate their models.

Another alternative to survey data is publicly-available digital media data that can be acquired, for example, via webcrawling. Hinz, Laufer, Walzenbach, and Weeber (2018) use publicly-available media data from digital archives and web chronicles in combination with geocoding to create an event dataset to investigate spatial and temporal diffusion of xenophobic attacks in Germany between 2015 and 2017. The authors demonstrate that the combined data can be a valuable alternative to data collected by state-level statistical offices in Germany.

Social media are, of course, not the only source of Big Data. Sensor data generated by smartphones are yet another example of data sources that can enrich, complement, or even, in some cases, replace traditional survey data. The presentation by Haas, Keusch, Kreuter, and Trappmann (2018) investigated the use of differential incentives to increase app installation rates on smartphones and the extent to which users activate different data sharing functions, including location information and activity data, among android users in a German panel study. While higher incentives were associated with a higher installation rate, the authors found no effect of incentives on the type of information users agreed to share. Furthermore, these effects were stable across different subgroups, including a general population sample and a sample of welfare recipients. The results of this study suggest that the findings from the survey research literature on incentives may hold even in this new landscape.

Finally, Liu and Wang (2018) compared three different machine learning techniques—random forests, support vector machines, and lasso—to a traditional logistic regression model for predicting the likelihood of follow-up survey participation in a panel survey. Since these methods require (almost) no model specification regardless of whether the true relationships are relevant or irrelevant, linear, nonlinear, or nonmonotonic, or a product of selection, they have the potential to increase prediction accuracy and decrease bias by more accurately detecting the true complexity of the underlying propensity model given the set of available covariates. The authors showed that, while the machine learning models did not outperform logistic regression in terms of accuracy, sensitivity, and specificity, these models provided valuable insights into the relative importance of individual variables for the prediction of follow-up survey participation.

Without question new data sources are emerging at a rapid pace. Many of these data sources convey information that may be of interest to public opinion, social science, and survey researchers alike. The key to a prosperous future for our fields moving forward will be to learn how to unlock the potential of these data sources either alone or in conjunction with more traditional data sources like surveys. Linking administrative, survey or other third-party data across sources requires careful data management to make sure that clear and consistent linking variables exist across the data sources or that there is enough information to create probabilistic linking models—not to mention explicit permissions. With big data comes big challenges, big potential, big responsibility and big opportunity. The new methods and applications presented at BigSurv18 are exciting examples of how we are beginning to tackle these big questions.

6 Combining Big Data with Survey Statistics: Tools

BigSurv18 also presented a platform for presenters to showcase a broad cross-section of the new computational tools that are becoming increasingly available to, and more often utilized by, survey methodologists and social scientists in general.

Emery (2018) presented a talk on the development and

early stages of deployment of the Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) data facility, a new platform to help bridge the vast troves of Dutch administrative data and social science research utilizing a high-performance computing environment. Prior to development of ODISSEI, Statistics Netherlands (CBS) was making microdata available to researchers via restricted access environments hosted on its own computing platform. While adequate for the analysis of person-level data with one cross-sectional record per person (about 18 million people), or several time points, that existing capacity is no longer sufficient. In its place, the ODISSEI data facility makes use of the national supercomputing infrastructure (SURFsara), originally intended for high performance computing in natural sciences, and has constructed an infrastructure that allows for secure data access to the potentially highly sensitive data in which social science researchers are interested (as opposed to, say, climate modeling or protein folding applications, where such concerns are generally irrelevant). Emery highlighted several test use cases to buttress these points.

For example, the very first project run on ODISSEI was the analysis of geospatial contextual determinants of schizophrenia. The study was conducted with 11,000 participants in the National Twin Registry (NTR), hosted by one of the members of the ODISSEI initiative, the Vrije Universiteit Amsterdam. The dataset includes genotype data generated from a genome-wide association study (GWAS) on the NTR participants, resulting in a dataset with, in essence, 11,000 rows and about 50 million columns. From this genotype data, scientists calculated a polygenic score designed to characterize the genetic risk for schizophrenia. This polygenic score was further linked to the geographic location of the study participant, allowing researchers to examine the association with urbanicity. While the cross-sectional association between schizophrenia and urbanicity has been established in the social psychology literature, further additional information on migration history may help establish the causality of the relation (i.e., that people with greater risk for schizophrenia are attracted to larger, more densely populated cities).

Another project that is currently under development on ODISSEI will measure population diversity at multiple geographic scales, aggregating individual characteristics starting at the lowest resolution (10m x 10m) and increasing scale from there. Both this project and the previously-described geospatial determinants of schizophrenia study highlight the challenges of disclosure control for social science analysis conducted at tight geographic resolution: researchers and official statistic authorities have yet to settle on how to characterize the risk of re-identification and personal characteristics disclosure for high-resolution choropleth maps.

Yet another project in the making on ODISSEI is the analysis of social networks: if one is interested in studying network formation or behavior for the entire population of the Netherlands (again, about 18 million residents), this cannot be done at the Statistics Netherlands computing facilities. The networked population would have, on average, about 4,000 links per adult (combined school, work, and neighborhood networks), producing a dataset of about 70 billion pairs or dyads. Emery concluded his remarks by stating that, in his opinion, the Netherlands is the best place for a computational social science researcher to be nowadays (a comment which created a lot of buzz on Twitter at #BigSurv18).

Amer et al. (2018) presented a different, but no less compelling, tool for making effective use of big data sources in social science research. They discussed the use of gridded population models as sampling frames where other population data may be unavailable or of poor quality. Gridded populations, as the name implies, are associated with a fine geographical grid; the tool is much more useful now because, whereas the grids from the early 2000s had a resolution of 1km x 1km, more recent grids sport a resolution of 100m x 100m. "Top-down" grids disaggregate the existing administrative geographic units, while "bottom-up" grids rely on high-resolution satellite imagery and can assign population counts using many different features, including building footprints, land use and settlement areas modeled from image texture, and energy consumption obtained from remote sensing in infrared spectra, among others.

Amer noted that there are still issues to be resolved when using gridded populations to sample actual human populations. Some of the challenges include potential mismatches between the artificial constructs of geographical grids vis-à-vis the human settlement processes: geographical grids are squares, but humans rarely actually settle in that way, instead flowing according to the topography. Also, researchers need to aggregate the grid nodes to meaningful administrative (or other sensible) units for the human population in question. And, available data may be misleading, if not incorporated carefully: for example, energy consumption data may reveal "ambient" population characteristics—that is, where people are during the day and, thus, consuming energy as opposed to where they reside and spend the night. Additional challenges include missing (e.g., clouds over the satellite image) or dated satellite data and modeling errors as a unique source of frame error.

BigSurv18 also saw several applications of machine learning methods in a variety of statistical tasks that arise throughout the survey life cycle. Cohen and Shorey (2018) presented a paper on one such application: using machine learning to enhance imputation of missing survey data. They considered the task of imputing of missing health care event expenditures in the Household Component (HC) of the Medical Expenditure Panel Survey (MEPS), the primary source of microdata in the United States on the cost of medical procedures and expenses incurred by individuals and insurance companies. These data are collected on approximately 35,000

individuals in approximately 15,000 households across the United States. Missing data are pervasive, as the U.S. health care system is opaque, and the sources and amounts of payments for medical procedures are often difficult to track. Approximately 50% of the expenditure data for physician (general practitioner) based visits are missing in MEPS data. This imputation project compared the "traditional" missing data imputation method, weighted sequential hotdeck (WSHD), which usually takes about 6 months to complete once the survey data is collected and weighted, with alternatives based on machine learning techniques. Cohen and Shorey compared a variety of machine learning methods, including classification and regression trees, neural network classifiers, *k*-means clustering, and random forests (RF), with the latter ultimately demonstrating better performance than other methods. Distributions of the imputed data were compared across WSHD and RF and were found largely similar in ten of the twelve variables imputed (exceptions: amount paid by the family out of pocket and amount paid by veteran insurance). Perhaps more importantly, given the (new) emphasis on timeliness noted above, these machine learning methods demonstrated the potential to process and impute the data in less than half the time that the traditional methods require.

Dutwin (2018) discussed an application of machine learning methods to enhance the U.S. general population sampling frames. While the U.S. lacks an official, government-maintained population register, there a quite a few commercial vendors that aggregate data sources, linking them with sampling identifiers such as addresses and phone numbers. These various data sources include voter registration records (publicly-available in many U.S. states), credit history records, magazine subscriptions, purchased product registrations, and many others. Additional information is often added from detailed Census tables at the tract level (about 4,000 people) or block group-level (about 1,000 people). Dutwin integrated these data with survey data, totaling about 380,000 survey responses from an omnibus simple random sample dual-frame design. Dutwin demonstrated how different methods fare when attempting to predict religion (specifically, Jewishness). He reported that random forest models far outperformed single CART models and CHAID models; further, the author noted that the tradeoffs between model sensitivity and specificity could be used to optimize survey costs when translated into incidence and coverage that can be used for stratification and oversampling survey design decisions.

7 The Fourth Paradigm: Regulations, Ethics, and Privacy

Arriving at what some would call an idealized future state in which data are shareable, discoverable, and accessible by virtually-connected communities of researchers (the Fourth Paradigm: see, Tansley and Tolle (2009)) will almost necessarily entail a re-working of regulations and the culture surrounding data privacy and ethics. At BigSurv18, we were fortunate to see evidence that many researchers are already considering these issues, not the least of which was the Plenary Session entitled, "Big Data, Surveys, and the Privacy-Ethics Challenge."

Increasingly, science—even social science—is a team sport, requiring close collaboration across multiple institutions and researchers (and their data). Hill (2018) pointed out that all science has become more and more computational (Third Paradigm), able now to harness the abilities of ever-more powerful computers and abundant, ubiquitous data, leveraging techniques such as simulation and modelling. Arriving at the Fourth Paradigm (also known as e-science or data-enabled science or data-intensive scientific discovery) will only occur if scientists and researchers are willing to consider, a proiri, adopting a data management life cycle approach as they plan and execute their studies. More specifically, because of the integral role that data play now in scientific breakthroughs, researchers need to plan—upfront—for the ultimate dissemination and stewardship of the study's data (in addition to the usual analytic plan). **Timpone18** echoed these sentiments, noting that platforms are now being built to aid in the exploratory aspect of data-enabled science.

For the Fourth Paradigm to truly take hold, collaborative research will need to be reproducible. McCoach, Dineen, Chafouleas, and Briesch (2018) are aligned with the importance of thoroughly thinking through one's data management approach from the outset of a study, and outline several lessons learned from a case study in field of education research. They noted that concerns over reproducibility have changed expectations about the research process and the products of research to the point where there is frequently the requirement that researchers make their data and code available; as a result, and as noted above, researchers will want to carefully document data provenance, the data management workflow, and any automation steps or approaches.

For ubiquitous data to be available and accessible for research, the scientific community will have to develop and adopt new ways and approaches of handling privacy and ethics concerns. As of this point, we are far from unanimity on how, exactly, this should be done. Keusch, Kreuter, Struminskaya, and Weichbold (2018) asked German smartphone users about their willingness to share their data with researchers. They found that users are far from willing to share all types of data and that already-existing attitudes about privacy and confidentiality predispose respondents' willingness to share, even with "trusted" organizations, such as universities.

Much of the power of Fourth Paradigm science emanates from scientists' ability to link data from disparate sources, creating the opportunity to find new patterns in the data. Fobia, Childs, and Eggleston (2018) wondered if there will be

challenges stemming from public perceptions of data linkage among US Federal agencies; i.e., that linking data across agencies will increase the risk of harm via identity theft, financial loss, or regulatory enforcement. They use both qualitative and quantitative data to understand these concerns in depth and, then, develop communication strategies to mitigate them. Similarly, Tolich (2018) employs a New Zealand-based case study to examine the ethics of using administrative data to measure the prevalence of illicit drug use in both large and small population centers. He notes that researchers must be attuned to the harm and stigmatization that can be caused by research outputs as well as the front-end ethical considerations of informed consent.

8 Conclusion

As the Scientific Committee for BigSurv18, the authors assert that, from the perspective of both survey scientists and computer scientists, BigSurv18 was a first-of-its-kind success, at the very least from the standpoint of drawing together various disciplines and perspectives in the same venue to consider how each other's work can benefit the collective understanding of the human condition. We saw not only novel applications of technology to enhance survey research, but also frontier-extending research advancing multiple scientific disciplines.

In our view, two primary intersections between survey research and computer science emerged at the conference. First, several presentations described how organic data sets created through technology (e.g., customer and electronic transaction data, physical sensor measurements, smartphone app ownership, web browsing histories) could be validated through data collected through survey instruments, as well as combined with survey data to provide greater context and insights than either source of data could achieve in isolation. In this way, we are seeing come to fruition the possibilities foreshadowed and anticipated by, among others, Groves (2011), Hill, Dean, and Murphy (2013), and the AAPOR Task Force on Big Data (**Japec15**); furthermore, this ongoing line of research provides new insights into how big data sources should be evaluated and consumed, which has wide ranging implications and benefits beyond the intersection of survey research and big data.

Second, much of the research presented at BigSurv18 described uses of machine learning and data mining as applied throughout the survey lifecycle. In evidence were approaches to assist or automate sample design and construction; or, provide real-time insights to monitor data collection efforts; or, aid in the analysis of collected data, including identification of error sources and imputing missing data. The machine learning and data mining methods employed were as varied as their uses, ranging from traditional decision trees and clustering methods for numeric data to natural language processing for enhancing the use of textual data in qualitative and quantitative analyses to cutting-edge deep learning approaches, such as convolutional and recurrent neural networks, for understanding spatial and sequential data. To us, this willingness to improve and enhance survey research will serve us all well as we move together into the new landscape.

References

Alpaydin, E. (2016). *Machine learning: The new AI.* Boston.

Amaya, A., Bach, R., Kreuter, F., & Keusch, F. (2018). *Measuring the strength of attitudes in social media data.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Amer, S., Cajka, J., Chew, R., Jones, K., Unangst, J., & Allpress, J. (2018). *Household detection within gridded population area units: Producing small area population estimates in geo-sampling.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Amsbary, M. & Dulaney, R. (2018). *The view from above—virtual listing using GIS*. Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Biemer, P. & Amaya, A. (2018). *Total error frameworks for hybrid estimation and their applications.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L., ... West, B. (Eds.). (2017). *Total survey error in practice.* Hoboken, NJ: John Wiley and Sons.

Bobashev, G. & Wu, L.-T. (2018). *Comparison of simple and complex predictive models applied to the national surveys on drug use and health. example of multiple visits to emergency departments.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Braaksma, B., Zeelenberg, K., & DeBroe, S. (2018). *A framework for big data in official statistics.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Buelens, B., De Broe, S., Meijers, R., ten Bosch, O., & Puts, M. (2018). *From experimental to official statistics: The case of solar energy.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Buelens, B., de Wolf, P.-P., & Zeelenberg, K. (2014). *Model-based estimation at statistics netherlands.* The Hague, Heerlen: Statistics Netherlands.

Buil-Gil, D., Solymosi, R., & Moretti., A. (2018). *Non-parametric bootstrap and small area estimation to mitigate bias in crowdsourced data. simulation study and application to perceived safety.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Burke-Garcia, A., Edwards, B., & Yan, T. (2018). *The future is now: How surveys can harness social media to address 21st-century challenges.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Buskirk, T. D., Bear, T., & Bareham, J. (2018). *Machine made sampling designs: Applying machine learning methods for generating stratified sampling designs.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Can, S., Engel, U., & Keck, J. (2018). *Topic modeling and status classification using data from surveys and social networks.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Citro, C. (2014). *From multiple modes for surveys to multiple data sources for estimates: The role of administrative records in federal statistics.* Washington Statistical Society President's Invited Lecture, WSS Seminar on Administrative Records for Best Possible Estimates. September 18, 2014, Washington DC.

Cohen, S. & Shorey, J. (2018). *Ai and machine learning-derived efficiencies for large-scale survey estimation efforts.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Datta, R., Ugarte, G., & Resnick, D. (2018). *Using linked survey and administrative data to assess the quality of each contributing data source.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Diego-Rosell, P., Srinivasan, R., Dilday, B., & Nichols., S. (2018). *Assessing community wellbeing using google street view and satellite imagery.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Dutwin, D. (2018). *Feedback loop: Using surveys to build and assess RBS religious flags.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Eck, A., Buskirk, T., Fletcher, K., Stefek, P., Shao, H., Park, K., & Losch, M. (2018). *Machine made sampling frames: Creating sampling frames of windmills and other non-traditional sampling units using machine learning with neural networks.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Emery, T. (2018). *The ODISSEI data platform*. Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Fobia, A., Childs, J., & Eggleston, C. (2018). *Attitudes toward data linkage, privacy, ethics, and the potential for harm.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Greenaway, M. (2018). *Synthesizing big data and business survey data.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861–871.

Haas, G.-C., Keusch, F., Kreuter, F., & Trappmann, M. (2018). *Money for data. effects of incentives in smartphone data collection.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Hill, C. (2018). *Moving social science into the fourth paradigm: Opportunity abounds.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Hill, C., Dean, E., & Murphy, J. (2013). *Social media, sociality, and survey research.* Hoboken, NJ: John Wiley and Sons.

Hinz, T., Laufer, J., Walzenbach, S., & Weeber, F. (2018). *Social diffusion of xenophobic attacks in germany. an application of web crawling.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Hutchinson, R. (2018). *Using alternative data sources to reduce respondent burden in United States Census Bureau economic data products.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Japec, L. & Lyberg, L. (2018). *Big data initiatives in official statistics.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Keusch, F., Kreuter, F., Struminskaya, B., & Weichbold, M. (2018). *Combining active and passive mobile data collection: A survey of concerns.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Liao, D., Berzofsky, M., Thomas, I., Couzens, L., & Cooper, A. (2018). *Experiences in FBI's NCS-X NIBRS estimation project.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Liu, M. & Wang, Y. (2018). *Using machine learning models to predict follow-up survey participation in a panel study.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Matthews, P., Kyriakopoulos, G., & Holcekova, M. (2018). *Machine learning and verbatim survey responses: Classification of criminal offences in the crime survey for England and Wales.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

McCoach, D., Dineen, J., Chafouleas, S., & Briesch, A. (2018). *Reproducibility in the era of big data: Lessons*

for developing robust data management and data analysis procedures. Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

McCreanor, R., Wronski, L., & Chen, J. (2018). *Automated topic modeling for trend analysis of open-ended response data.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Meng, X. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 U.S. presidential election. *Annals of Applied Statistics*, *12*(2), 685–726.

Mulry, M., Bates, N., & Virgile, M. (2018). *Leveraging nontraditional data to improve response propensity models and design tailored and targeted geographical nonresponse interventions.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Pasek, J., McClain, C., Newport, F., & Marken, S. (2018). *Who's tweeting about the president? what big survey data can tell us about digital traces.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Ridenhour, J. & McMichael, J. (2017). *Propensity stratification with auxiliary data for address-based sampling frames.* Paper presented at the Annual Conference of the American Association for Public Opinion Research, New Orleans, LA, May 2017.

Ridenhour, J., McMichael, J., Krotki, K., & Speizer, H. (2018). *Using big data to improve sampling efficiency.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Salganik, M. (2018). *Bit by bit. social research in the digital age.* Princeton, NJ: Princeton University Press.

Tam, S.-M. (2018). *Mining the new oil for official statistics.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Tansley, S. & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery.* Redmond: Microsoft Research.

ten Bosch, O., Windmeijer, D., van Delden, A., & van den Heuvel, G. (2018). *Web scraping meets survey design: Combining forces.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Tolich, M. (2018). *Big data's front-ended ethical considerations ignore how results can stigmatize identifiable groups: Examining big wastewater data in New Zealand*. Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

Vanhoof, M., Lee, C., & Smoreda, Z. (2018). *Performance and sensitivities of home detection on mobile phone data.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.

West, B., Wagner, J., Hubbard, F., & Gu, H. (2015). The utility of alternative commercial data sources for survey operations and estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, *3*(2), 240–264.

Ye, C., Medway, R., & Kelley, C. (2018). *Natural language processing for open-ended survey questions.* Paper presented at the Big Data Meets Survey Science Conference, Barcelona, Spain, Oct. 2018.