FORS
explore.understand.share.

Brian Kleiner, Alexandra Stam and Nicolas Pekari

Big data for the social sciences

Lausanne, April 2015

FORS Working Papers 2015-2

FORS Working Paper series

Copyright and Reserved Rights

Editorial Board

How to cite this document:

Kleiner, B., Stam, A., & Pekari, N. (2015). Big data for the social sciences. *FORS Working Paper Series,* paper 2015-2. Lausanne: FORS.

ISSN 1663-523x (online)

FORS
c/o University of Lausanne, Géopolis
1015 Lausanne
Switzerland
E-mail: paperseries@fors.unil.ch

Big data for the social sciences

Brian Kleiner, Alexandra Stam & Nicolas Pekari[1]

1. Introduction

In an age where new technologies are enabling the production and control of massive amounts of information, “big data” has been heralded as a new frontier not just for the business world, but also for the social sciences. In addition to corporations, governments are increasingly paying heed to the economic and scientific potential of harnessing the power of digital data, pouring fortunes into innovative projects that aim to develop new methods and tools for capturing, managing, and exploiting enormous volumes of information (e.g., the ESRC’s Big Data Network in the UK, and the U.S. NSF’s Big Data Research Initiative). These major initiatives signal a strong belief that data and related technologies can accelerate growth in economies and ultimately lead to better quality of life for individuals. The European Commission cites significant big data benefits in various sectors, such as in healthcare, transport, the environment, agriculture, and manufacturing[2].

Big data means many things to many people. Key features are repeated in the literature, most notably that this involves very large and complex datasets, and that the data are produced and diffused rapidly, often in real time. Big data, also known as “organic data” (Groves, 2011), are often created automatically by systems, and thus while rich may be unstructured and variable. Couper (2013) describes three main types of big data: 1) administrative data, which are provided by individuals or organisations to governments; 2) transactional data, which are generated as by-products of transactional activities (e.g., credit card transactions, phone records, browsing behaviour); and 3) social media data, which are created online by people who want to share information about themselves (e.g., Tweets, Facebook data). We

[1] FORS – Lausanne
[2] https://ec.europa.eu/digital-agenda/en/what-big-data-can-do-you

would add a fourth that seems to stand apart – large text corpus data in digital form that are culled from various online sources (e.g., political speeches, legislative texts, blog posts). This category does overlap to some extent with social media data and transactional data, which may also involve text, but at the same time it is different in that large text corpus data are far less structured and are generated with a greater diversity of purposes (see section 2.4).

But what does the "big data revolution" mean for the social sciences? Will the rapid availability of massive amounts of data from diverse sources render traditional scientific practices obsolete? Will the social sciences at some point abandon current standard research methodologies in favour of newer more powerful ones? The purpose of this paper is to address in broad strokes the question of big data for the social sciences, its real potential and challenges, and then, based on this, to state our institutional orientation.

FORS is an infrastructure institution and serves the Swiss social science research community. We produce, preserve, and provide national data for secondary analyses, data that can be used to address important research questions within the social sciences. Further, we have the obligation to be aware of and to take into account important developments that can affect knowledge production in the social sciences. Therefore, we feel it is important to take an informed position with respect to big data. During 2014, FORS held periodic meetings of a working group on big data among interested staff, with the goal of understanding the scope, definition, benefits, impact, and problems concerning big data for the social sciences. This paper summarizes the conclusions of these meetings and reflections. Further, it outlines the ways that FORS will likely use big data in the near future to benefit the research community that it serves, and addresses the extent to which we may modify or expand our services to this effect.

2. Big data for the social sciences

According to its advocates, what big data provokes is not only the productive exploitation of larger volumes of data, but perhaps more importantly the conviction that naturally occurring data, advanced tools, and innovative techniques will open unexpected avenues of analytic potential, allowing naturally occurring data to reveal new patterns and insights into human society, politics, and economics. Some contend that these new sources of data will inevitably replace more costly and time-consuming traditional methods of gathering information, e.g., surveys or in-depth interviews (e.g., Savage and Burrows, 2007).

Use of organic data, such as administrative data, social media data, transactional data, or large text corpuses offers specific advantages for social science research. First, in theory the data can be used on a larger scale than with traditional sampling methods, reflecting the habits and characteristics of greater numbers of people (i.e., with less reliance on probabilities and margins of error). Also, organic data come right from the source and presumably exhibit real and naturally occurring behaviours (e.g., spending with credit cards, phone records, social networks in social media data), thus allowing a more direct insight into human activities, without depending on self-reports. Further, the rapid and automatic production of such data facilitates examination of real-time processes – the data are current, and this makes it easier to study change over time. The following sections illustrate the potential of the four types of big data with several examples from the literature.

2.1 Administrative data

Broadly speaking, administrative data is information that was primarily collected for administrative purposes. It is often collected routinely as part of the daily activities of both public and private instances, and it may either be longitudinal or cross-sectional in nature (Calderwood and Lessof, 2009). Administrative data usually consist of numbers, and while they are "big" with respect to quantity, and reach large populations, they cannot compare to the "volume, velocity, and variety" of the other organic data sources (Von Gunten et al., 2014). Examples include data on taxes, income, labour status, government benefits, education, social care, vital events, and health (Laurie and Stevens, 2014). Following the UK Administrative Data Liaison Service[3], two areas of research – education and health – have so far resorted considerably to administrative data. Although important access problems remain, it is highly likely that social scientists will engage more strongly with such data in the near future. This section focuses on public administrative data, and we will use the terms ‘administrative data’ and ‘register data’ interchangeably.

[3] http://www.adls.ac.uk/

Towards a new research environment

The research landscape is changing, and the use of administrative data to either replace or complement survey data is becoming increasingly common. Some countries are clearly ahead when it comes to producing and using register data. For example, 96 percent of Statistics Finland data come from registers and administrative records. In many countries decennial censuses have been replaced by register data, or by a combination of both register data and survey data, as is the case of Switzerland. These shifts have been made possible by technological developments that facilitate the creation of large databases. For public institutions like National Statistical Offices, register data present two main advantages. First, they seem to be cost-effective alternatives to the production of data, and second, they reduce the burden on citizens by not having to ask questions for which data already exist.

Discussions surrounding administrative data are increasingly prevalent at the level of the European Commission. Administrative data play a central role in the future vision for the European Statistical System (Holmberg, 2012), and recent years have seen the development of a growing number of initiatives aiming at promoting their use. For example, the ESSnet initiative of data integration and the 7th framework program Blue Enterprise and Trade Statistics both give important consideration to key methodological issues in the use of such data for research purposes (Bakker, 2012).

While many believe that administrative data have the potential to provide datasets of significant value for social science research, there are still important questions to be answered regarding their validity as well as their usage – on their own or combined with other data sources. There has been in recent years a mushrooming of publications and national reports on the quality and reliability of register data. Statistica Neerlandica, for instance, devoted a special issue in 2012 to the methodological challenges of register-based research (Bakker and Rooijen, eds.). The following sections will describe how administrative data have been used for research purposes, followed by two examples in the form of brief case studies. We then discuss the main advantages and drawbacks of this type of data.

Administrative data in practice

Although the use of administrative data among social scientists is still relatively rare – in particular due to restrictive access conditions – the literature has shown there to be three main applications. First, administrative data can be used for research instead of survey data. National economic and social statistics, such as for example employment or crime rates, can be produced through the direct use of administrative data. To increase the analytic potential of the data, it is often possible to link different years of data, or even different sources of administrative data. Nordic countries, for instance, have established a system of personal identity codes, which facilitates the process of linkage.

Second, administrative data may supplement survey data. It is possible to combine administrative data with survey data (both cross-sectional and longitudinal) to provide additional items. Linkage is usually achieved by using a limited set of socio-demographic variables (Jutte et al., 2011).

Third, administrative data can serve as a way of enhancing survey quality. For example, administrative data may provide sampling frames and contact addresses for respondents of a survey, and as such improve survey response and representativeness. Administrative data may also help to improve survey quality by making it possible to check for errors or impute missing data. They may also be used to validate survey data: similar items may be found in both surveys and administrative data, allowing advancement of knowledge on item non-response and if necessary helping to develop weighting strategies. If administrative datasets are population references, they can be used to inform estimates of under-coverage or attrition in surveys.

To better understand the benefits and limitations of these usages, two case studies are presented below. The first one draws on a study entirely based on register data. We selected a Danish example to examine the potential of administrative data in an "optimal" setting, where exhaustive register data sources are available and can be matched together. The second case study draws on the Swiss context and illustrates the combined use of both survey and register data, an approach that is increasingly advocated by scholars, but which also faces important technical and ethical issues, which will be discussed later on.

Case study 1: Christoffersen et al. (2003), a study on suicide attempts among Danish children born in 1966.

This study, which was based purely on register data, aimed at assessing the risk factors that may explain suicide attempts of Danish children. Data on different aspects such as health, education, family dissolution, suicidal behaviour, substance abuse, criminality, and employment were investigated for some 84’765 children born in 1966 and their parents. Analyses revealed some interesting findings, such as relationships between first-time attempts and parental psychiatric disorders, suicidal behaviour, violence, child abuse, and neglect. Those who suffered from psychiatric disorders or physical handicaps, as well as those who had been legally imprisoned, who were addicted to drugs, or who were jobless were also at higher risk of suicide attempts. However, the study also stressed a number of important limitations. Analyses included those who had been hospitalized following a suicide-attempt, therefore only representing a portion of the overall cases of suicide attempts. Those who did commit suicide were not included in the data. Turning to the factors that may explain suicide attempts, only serious cases would appear in the registers, for example when it comes to situations of child neglect. ‘Soft data’, such as psychological factors or personal events that may account for suicide attempts (like the failure to pass an exam or a sentimental break-up) are not measured in registers. The authors also noted that life events that occurred the same year as the suicide attempt could not be accounted for, as it was not possible to know what came first. Furthermore, young adults who had travelled to another country were absent from the registries. The study design also meant that immigrants and refugees were excluded.

Case study 2: Wanner (2006), a study on the professional behaviour of those approaching retirement age.

As part of a mandate for the Federal Social Insurance Office (FSIO), Wanner used administrative data as a complement to survey data to explore the socioeconomic situation of 60-70 year olds in Switzerland, and their various strategies when it comes to anticipating retirement or maintaining some professional activity following retirement. Two main registers were considered: cantonal fiscal registries and social insurance registries. Results highlighted strong relationships between people’s economic situation (both in terms of fortune and access to financial provisions) and their attitudes when it came to retirement or to continuing to have some professional activity beyond retirement. This served as a basis for wider reflection on the methodological perspectives and limits of the use of register data, which is the focus of Wanner’s 2006 paper.

Although the paper advocates combined use of both register and survey data, it clearly demonstrates the importance of register data for the study of some sensitive questions that traditional surveys fail to fully capture. Wanner illustrates his point with the Swiss Labour Force Survey (SLFS), which he claims is not adapted to measuring sensitive variables on financial situations, wealth, social protection, and income. For instance, tax declarations make it possible to distinguish income from each retirement pillar with a high level of precision, which is not possible with a survey. Plus, at the time of Wanner's study, the size of the SLFS made it difficult to draw representative conclusions about 60-70 year olds. Furthermore, the least privileged groups are likely to be underrepresented, while register data provides full coverage. Another drawback of the SLFS concerns the collected information, which does not make it possible to clearly draw links between anticipated retirement and people's financial situation, but also that does not fully capture professional activities undertaken after retirement. Register data should not replace but rather complement surveys. Important behavioural or perceptual information, such as attitudes toward the job market, aspirations for retirement, or indications of health, can only be collected through traditional surveys.

Pros and cons of administrative data

The two case studies presented above highlighted some important advantages when it comes to using administrative data. Christoffersen et al. illustrate the richness of the linkages that can be made: not only can different registries be linked together, provided that they are available and can be matched via a personal identification number, but it is also possible to match individual records with those of their relatives over a long time period. As such, there is the potential to build up a very rich and large (if not big) dataset. Wanner's study highlighted other aspects and the fact that in some cases register data may not only be exhaustive and cover an entire population but may also be of better quality, in particular with respect to the collection of sensitive information that people may be reluctant to provide 'on the phone', or factual information that people may not accurately remember – for instance, the income from the different retirement pension plans.

Even if they are not as 'big' as the other types of big data, administrative data present some common advantages, in particular that of being already collected, of being relatively cheap (though processing costs may be very high depending on the quality of the registries), and of not being intrusive to the target population. Those using administrative data in supplement to survey data can therefore concentrate on information not available elsewhere.

In general, though important variations may exist across registers, they are regularly and sometimes even continuously updated. They are usually collected in a consistent way, which makes it possible to build up good quality time-series. Depending on the producer or the distributor, rigorous quality checks may be undertaken. Administrative data distributed by the Swiss Federal Statistical Office, for instance, have undergone high quality checks and transformations to make them appropriate for the production of statistics. In some cases they provide an almost 100 percent coverage of the target population – this is the case for data coming from tax declarations. When a full population is considered, this allows for analyses at the small area level as well as of rare person groups.

As the case studies showed, however, there are also a number of drawbacks. Just like for other types of big data, administrative data are criticized for not addressing specific research questions, raising important issues about the underlying statistical paradigms (Wallgren and Wallgren, 2011). In some cases – though this again depends on the producer – documentation may be absent or of bad quality. Just like for any secondary data source, researchers have no control over the production of the data, and as noted by the UK Administrative Data Liaison Service, there is a lack of well-established theory and methods to guide the use of administrative data in social science research. Another disadvantage is that while there may be important demographic variables available, administrative data do not usually include information about intentions, motivations, or attitudes. Further, administrative data may in some cases face important administrative delays, and are also more vulnerable to changes to administrative procedures. These could affect definitions and make comparisons over time more difficult.

Most discussion in the literature on the drawbacks, however, concerns quality issues. Data may be missing or erroneous. Errors may easily happen, not only because administrative data rely on information provided by people (who may purposely or not give false information), but also because they are often processed by people. The administrative practice of the register keeper may lead to biased entries (Bakker, 2012). For example, in the case of Residents Registration Offices many different people are involved in the processing of the data, some of whom may be more careful than others. Also, some information may not be considered important by the administrators, and therefore updates may be neglected (for instance address details). Bakker (2012) distinguishes measurement errors from representation errors. While measurement errors relate to statistical concepts that do not match administrative ones, representation errors occur when the measured population differs from the target population. As we saw with the Christoffersen et al. case study, only those young people who had experienced major suicide attempts leading to hospitalization were included, while many other attempts were not captured.

Finally, data linkages may also be a source of problems. Not only can the process of linking survey and administrative data be costly, time-consuming, and complex, but it is not always possible. Linkages often result from negotiations between the different producers. If data owners are willing to proceed with the linking, and if the legal basis allows it, further problems may still occur. Lynn, Calderwood, and Lessof (2009) noted three main reasons why linkage is not always possible: consent for linkage, success of linkage, and the completeness of the administrative data. In addition, confidentiality and ethical concerns are a major potential barrier.

2.2 Social media data

Hundreds of millions of people use social media platforms every day for diverse activities such as networking and communicating (Facebook, Twitter, Skype, emails), sharing photos (Flickr, Pinterest), videos (Youtube), personal stories and opinions (blogs), or participating in collaborative projects (Wikipedia) – to name only a few. With most of the population actively using the Internet, it is no surprise that social media is today one of the most important areas of the rapidly growing data market (Puschmann and Burgess, 2013). More information is collected nowadays within a short period than existed in the whole world only a few decades ago, raising new challenges as to how capture, handle, and make sense of such data for scientific purposes. While social media data were long considered to be the domain of information and computer scientists due to the new skills and tools such unlimited and unstructured data required, they are now becoming of increasing interest to social scientists. Thanks to collaborations with platform owners as well as facilitated access through Application Programming Interfaces (API's), social media data are now within reach for scientific purposes.

Sometimes presented as a goldmine of what people are thinking (Brid-Aine Parnell, 2014), such data are seen as offering unprecedented opportunities for the collection of experimental and observational data. In particular, they provide the ability to observe the online behaviour of hundreds of millions of people in almost real time, and therefore to measure small effect sizes that may otherwise have gone unnoticed (Golder and Macy, 2014; Khoury and Ioannidis, 2014). This led Golder and Macy (2014) to describe the scale as being both massive and microscopic at the same time.

Not only have researchers access to the most recent data available (Khoury and Ioannidis, 2014), but social media data also offer information on what people do and say “in the wild” (Tinati et al., 2014), rather than retrospectively. By capturing “everything” within a particular field of a particular platform (Tinati et al., 2014) they also open up new research venues, making it possible to study behaviours that were unsolicited and unprompted by the researcher in the first place (McCormick et al., 2013). Social media data may also be used to inform traditional research, for example by uncovering new associations that may be further tested, or may serve as a pre-test to see whether behaviours or attitudes not currently in a survey are evident among the larger population.

However, along with big opportunities are also big challenges, and sometimes big problems. To best account for both the opportunities and shortcomings of such data, the following sections will address the two social media data sources most used currently by social scientists, namely Facebook and Twitter data, with over one billion and 300 million users worldwide, respectively (Golder and Macy, 2014). These sources offer contrasting perspectives, not only in how data are being accessed, but also how they are being analysed. Two contrasting short case studies will be presented. The first presents one out of many uses of Twitter data in social science research. Highly descriptive, it illustrates some of the analytic and methodological limitations of such data. The second, drawing on Facebook data, describes a controversial recent study on “emotional contagion”, which involved collaboration between Facebook and academic researchers. Experimental research was conducted without the consent of Facebook users, which recently became something of a scandal and is currently the driving force for new considerations about ethics and ethical research in the rising era of big data. The final section will then discuss some of the shortcomings of social media data for scientific research.

Twitter data

Among all social media sources Twitter data are without doubt the most widely used for social research. Examples are numerous and attest to the relative ease of access to the data, which Twitter has made partly available through an Application Programming Interface (API). As explained by Boyd and Crawford (2012), however, only a fraction of Twitter material is made available through its APIs. The “firehose” supposedly contains all public tweets (with some exceptions), but very few researchers have access to this exhaustive dataset. Most researchers access a “gardenhose” (about 10 percent of public tweets), or a “spritzer” (about 1 percent).

Research based on Twitter data is extremely varied. Some recent topics include for example studies on political activism (Tinati et al., 2014; Theocharis, 2011) or health promotion (Khoury and Ioannidis, 2014; Jashinsky et al., 2013). While often observational, a number of recent studies, particularly in the field of health are used to predict outbreaks and particular health-related behaviours (see for example Young, 2014). The following case study, which investigates mood rhythms, is just one out of many possible illustrations as to how social scientists can analyse the digital archives of online activity.

> *Case study 1 : Golder, S.A. and Macy, M.W. 2011. Diurnal and seasonal mood vary with work, sleep, and day length across different cultures. Science, vol 333.*

Using Twitter data, the authors considered 509 million public messages from 2.4 million individuals across the world between 2008 and 2010 to investigate diurnal and seasonal mood. Analysis of the messages was conducted using a text analysis software program called Linguistic Inquiry and Word Count (LIWC) to determine positive and negative affect. Altogether 64 behavioural and psychological dimensions were created using emotion rating scales and thesauruses. Hourly, daily, and seasonal changes were analysed at the individual level in 84 countries. The authors noted the prevalence of robust rhythms across cultures.

Findings show that people appear to be in better moods in the morning, which then deteriorates throughout the day, plus they have better moods on the weekends and when days get longer. As noted by the authors, such research also has important limitations. For instance, there is "little data on conditions that may influence mood, including demographic and occupational backgrounds that may influence when and how much people sleep, the level and timing of environmental stress, susceptibility to affective contagion, and access to social support". Furthermore, lexical analysis measures the expression of affect rather than the experience. Expression of different moods, and the moment they occur during the day may also be affected by cultural norms.

Facebook data

As opposed to Twitter data, Facebook data contain rich demographic data, which can include full names, dates of birth, geo-location, affiliations with friends and organisations, political and social movements, or even cultural tastes (Golder and

Macy, 2014). Its access is however limited, and is often provided as part of a collaboration with Facebook research staff, or as part of special relationships with platform owners. Other alternatives include the development of specific applications that once adopted by users may allow researchers access to the data of those users (Golder and Macy, 2014). However, this latter option, relying on self-selection, may reinforce sampling biases. There have been several articles based on collaboration with the Facebook team, such as for example the controversial study by Bond et al. (2012) on social influence and political mobilization.

The case study presented below features an atypical piece of research in that it involved collaboration with Facebook staff and an experimental study that led to a current debate on research ethics. It raises important questions about the analytical potential as well as the possible dangers of big data research.

> *Case study 2: Kramer, A.D.I., Guillory, J., and Hancock, J.T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. PNAS, 24:111.*

Seeking to address gaps in the understanding of emotional contagion, the authors conducted an experiment among 689,003 Facebook users, dividing them into two randomly selected groups. The first group was exposed to positive emotional content in the News Feed, while the others were exposed to negative content. In particular, researchers wanted to test whether the exposure to a particular emotion led people to post messages of the same emotional nature.

The News Feed, which is determined by an algorithm, is the information people see about their friends when they first log into Facebook. During one week in January 2012, algorithms were manipulated so that users could be exposed mainly to one or the other emotion. Posts were determined to be positive or negative if they contained at least one positive or one negative word. In total, 3 million posts were analysed. The authors stated the hypothesis that if verbal expressions can influence affective states, people would be less positive if exposed to negative information and vice versa. This was consistent with the research findings that showed emotional contagion. Emotions expressed by friends influenced moods, even though the effect sizes from manipulations were small. While the methods can be discussed – particularly the fact that the research did not (or could not) take into account the wider influences that may affect moods, rendering the results rather simplistic, it nonetheless triggered important debate among scholars and others alike.

Much of the subsequent discussion concentrated around ethics, in particular the fact that Facebook conducted an experiment without informing users, somehow playing with their emotions, and the more general observation that Facebook can manipulate its users. As the authors themselves noted, even small effects can have large consequences when it comes to big data. The finding itself means that Facebook or other service providers could learn from that experience to influence their users' behaviours. Even though the authors noted in their article that the researchers did not access the full texts, to be consistent with Facebook's Data Use Policy, this research took on unexpected dimensions, forcing the first author, who works for Facebook, as well as the Facebook chief executive to apologize for "the anxiety it may have caused", and for the way the research was communicated (Schroeder, 2014).

Limitations of social media data

The previous two case studies point to some interesting and innovative ways in which social media data have been used for social scientific research. There is no doubt that others will continue to develop and refine different techniques to exploit and analyse these types of data. However, several persistent analytic and methodological limitations of social media data should be noted. The first challenge that researchers face is that of accessing the data, and being able to handle the data, which ultimately will influence research results. As noted before, Twitter data are by far the most accessible, and resulted in a large number of papers, often very descriptive. The almost endless observational possibilities also raise questions about their academic pertinence and their validity. Some, like Tinati et al. (2014), express the concern that such a wealth of data may take researchers away from more important topics. As illustrated with the Twitter case study, analysis depends on the information that is available, and many factors needed to explain a phenomenon are missing from the overall picture. As Golder and Macy (2014) note referring to studies of network contagion, it is difficult to distinguish between homophily and contagion. This is made more problematic as socio-demographic information is mostly missing from these data sources (Ruths and Pfeffer, 2014; Golder and Macy, 2014). While some have tried to recreate demographic information by coding profile images (see for example McCormick et al., 2013), this remains an important barrier fully exploiting analytical potential.

There is also no guarantee that the information provided by users, whether demographic or substantial, is correct. For example, all kinds of new businesses are flourishing that enable people to hire a false girlfriend or boyfriend to post Facebook messages. Important too are measurement issues. We know when people write things, but we do not know when they experienced specific events. Users may also revise what they have written (Golder and Macy, 2014). Also, despite being obvious, one needs to keep in mind that people only write about what they would like to share.

Others have expressed concern about the false relationships such data may generate, since even small effects can become very significant, as the Facebook study showed. Khoury and Ioannidis (2014) warned against the accrued risks of the ecological fallacy, which may have drastic consequences, especially when the results are used to “inform and justify decisions and investments among the public and the industry and government” (Ruths and Pfeffer, 2014). As Khoury and Ioannidis note, the big data strength is in finding associations, not in showing whether these associations have meanings. Just as the Facebook study revealed, intentions of academic scholars are often different from those of businesses, raising important ethical considerations about how findings may be used and potentially misused (see for instance Savage and Burrows, 2007).

Some researchers tend to praise “big numbers”, as if these alone legitimize studies and provide analytic relevance, while neglecting discussion of the wider limitations (Boyd and Crawford, 2012). Beyond the value of size, three aspects of high importance for the quality of research with social media data need to be further developed: replicability, sampling, and platform design.

A first serious limitation is the fact that most social media data cannot be replicated. In an era where the possibility of replication is becoming more important, this may severely hinder the motivation of researchers to work with such data sources. Results from proprietary data (such as Google Flu Trends or Facebook) usually cannot be replicated. As Schroeder (2014) notes, internal researchers can overcome some problems to do with the data, but they will not make this public, which will prevent replication from people outside the company. Even when data are open (as in the case of Twitter) the proprietary company may apply changes to the platform or the underlying algorithms, which in many cases will not be documented. In the absence of an archive, replication may be strongly compromised, contrary to the emerging policies of many academic journals that increasingly require accessibility of the original data. The 'Retail Business Datasafe'[4] project from the Economic and Social Research Council, which received £14 million in funding to develop a repository, however, brings some hope that in the future analyses resulting from big data may be replicable.

Another important limitation is that of sampling. There are many ways in which population bias occurs. First, Facebook or Twitter users, like for any other social media platform, do not represent all people (Boyd and Crawford, 2012). There are many users worldwide who are not Internet literate, and even if they are, this does not mean they will be active on social media networks. As noted by Golder and Macy (2014), we have no idea how online and offline behaviour differs. Social media requiring technology and a number of skills means that a certain share of the population, like people with special needs, young children, or the elderly will inevitably be underrepresented in the data. Also, platforms themselves attract a specific category of users. Studies showed that users of specific platforms tend to share some specific socio-demographic characteristics (Hargittai, 2007). For instance, it has been reported that users of Pinterest are mostly females aged 25 to 34 (Ruths and Pfeffer, 2014). The attraction of a particular group to a specific platform may also change over time. A recent study showed that Facebook has become less attractive for teens but more so for older people (Washington post, 2014)[5].

Further, as researchers often work with a subset of the overall data, some information may not be made public, for example so-called private posts on Twitter. Social platforms do not usually provide information about who has been included or excluded (Ruths and Pfeffer, 2014). For example, is the sample a random subset of the entire dataset, or does it include data collected at a specific moment of time? As Boyd and Crawford (2012) note, without taking into account the sample of a dataset, and what it includes, the size of the dataset is meaningless. This may not only apply to users, but also to topics. Twitter for instance may decide to remove the tweets that contain problematic words, making the topical frequency inaccurate. Boyd and Crawford also draw attention to the differences between users and accounts – some users having multiple accounts and some accounts are used by multiple people. Some accounts may not even relate to users. Ruths and Pfeffer (2014) warn against the large numbers of spammers and bots present in social platforms.

[4] http://www.esrc.ac.uk/my-esrc/grants/ES.L011840.1/read
[5] http://www.washingtonpost.com/news/the-intersect/wp/2014/10/08/teens-are-officially-over-facebook/

Finally it is important to be critical of the platform itself, and how its design and way of functioning may generate specific behaviours. As noted by Ruths and Pfeiffer (2014), *“Social platforms also implicitly target and capture human behavior according to behavioral norms that develop around and as a result of the specific platform”.* Twitter may, for instance, be considered by users as a space for political discourse, therefore affecting the content of the communications. Platforms have not been constructed in the first place to collect good research data, and the way they are designed may affect both the content of what is being collected and potential analysis. For instance, as the authors further explain, Facebook integrated a “like” button, but the absence of a “dislike” button makes it more difficult to capture negative responses. Also, and more generally, it is worth remembering that technology is rapidly changing. As Kaplan and Haenlein (2010) note, what may be up-to-date today could disappear from the virtual landscape tomorrow.

2.3 Transactional data

Transactional data, as the name suggests, are derived from different types of transactions between humans and IT systems, where particular features of interactions are captured and recorded. The systems are generally designed to organise, process, and then document services or user activities. The collection of the data is normally automatic, and the individuals participating in the transactions may not be aware of how the data resulting from their activities will be used afterwards. The advent and wide application of such IT systems now makes possible the production of an enormous amount of data on people’s very specific behaviours in various domains.

The systems that capture transactional data may be to track credit card or loyalty card use, subscription data, phone records, or web browsing, to name a few. These are designed usually for businesses or government sectors that aim to measure the activities of product or service users. This may be employed by companies to monitor use or to improve products and services, or for governments to understand behaviours for better security, services, regulation, or policy. Transactional systems reveal what people do, including their habits and preferences, but not why they do what they do. Also, the systems are set up to capture types of information that are expected to be useful to the organisations that provide the services.

Proponents argue that such data could replace in some ways more traditional practices, with various significant advantages, such as: less need for intrusive data collection; getting accurate and detailed information on real behaviours – less need for self-reports; and large datasets that cover broad populations and that allow sophisticated statistics. Ruppert and Savage (2009) note that transactional data could create new paradigms and models for thinking about social processes, with a shift in emphasis from the individual to relations: *"Whereas surveys focus on (sampled) individuals and hence assume these to be the centre of analytic attention, transactional data focuses more on specific transactions, which are more amenable to be understood in network, associational, and relational terms."*

Despite such claims, however, there appears in fact to be still quite little work using transactional data for social scientific purposes, with the exception of some published studies that focused on web browsing behaviours (see, for example, the work and literature review of Choi and Varian, 2012). The lack of scholarship to date using transactional data may be due to several reasons. First, there are practical and ethical constraints. Gaining access to transactional data can be difficult, due to data protection issues. Companies and governments have little incentive to share these data with researchers and risk violation of the confidentiality of their clients. Also, there are serious ethical questions concerning the use of individual data where the persons providing the data have not consented to their use for secondary purposes.

Second, transactional data are not generated with social science research questions in mind – fitting such existing data to questions may be challenging. In addition, the data themselves may be messy, not well documented, and difficult to use. Third, as noted before, such data are limited in that they reveal nothing about the motivations, beliefs, and intentions of people, which is usually of key interest for social scientists. Last, as noted by Couper (2013), such data may suffer from coverage problems – generalizing to whole populations may be impossible, since not everyone uses credit cards, loyalty cards, or mobile phones, and the use itself of such systems may be selective within individuals, for example, people who own credit cards but who buy alcohol or fast food with cash. This in combination with the usual lack of demographic information about users could discourage researchers from investing their time in such data, since they might be very hard to interpret. In sum, as stated by Ruppert and Savage, besides the ethical challenges, "There are also many methodological and analytical questions related to how such vast amounts of data can be effectively analysed."

Case study: Responsible Gambling Trust's machines research programme. NatCen

One study that nicely illustrates the potential of transactional data for social scientific research is the work from NatCen within the "Responsible Gambling Trust's machines research programme", described in Wardle et al. (2013) and Wardle et al. (2014). The purpose of the research was to identify harmful patterns of gambling in bookmakers (in this context, small casinos with up to four machines) in the UK.[6]

The study involved two principal forms of data. First, it included a random probability sample of 27,565 individuals from a frame of industry registers of "loyalty card holders". A survey was conducted then by telephone or by web with 4,727 loyalty card holders who agreed to participate, 4,001 of whom agreed additionally that their survey responses could be linked with their personal loyalty card data. Second, transactional data from machine usage was analysed for these same individuals. Like all loyalty cards (e.g., for grocery stores, airline mileage programs), loyalty cards for bookmakers allow for every transaction where the card is used to be recorded, thus making it possible to track usage patterns, such as: number of machine gambling sessions per day; average length of individual machine gambling session; average number of different games played per session; and average amount waged per bet. Generally, such systems exist to provide financial records of machine proceeds, needed for accounting and tax purposes.

The study then linked responses to the survey with the transactional data of individuals to compare problem and non-problem gambling behaviours, including analyses of their self-reported responses in relation to particular logged loyalty card transactions, selected by the researchers for their relevance to the key research questions. One of the findings was that lower income individuals with loyalty cards were more likely to be problem gamblers than those with higher income (i.e., to spend more overall, placed higher bets, have more sessions per week, etc.).

This study illustrates that under certain conditions transactional data can be used to address significant research questions for social scientists, and to open valuable analytic possibilities. There are several strengths to the approach employed that could

[6] The line between applied and basic research is difficult to determine here. This study was commissioned by an actor with an interest in learning about problem gambling behaviors for policy purposes, not with a particular interest in advancing sociological theory.

be noted. First, the detailed behavioural data gleaned from the loyalty card use is certainly far more accurate than any self-reported responses in surveys. In addition, having such transactional data available reduced the burden on the individual survey respondents – there was no need to ask them – as well as the survey costs.

There are also some limitations noted by the NatCen researchers. First, the transactional data from the loyalty cards varied across operators in different ways, and there were also inconsistencies regarding naming conventions, thus making it difficult to compare and work with the data. A certain amount of harmonisation was needed. Second, the industry-held data were very narrow in scope, with concentration purely on financial transactions, and this limited their potential research interest. Further, the transactional data contained no demographic information about the players, and no contextual information exists that might help explain the recorded behaviours (e.g., hours of operation, machine layout, geographic location of bookmaker).

It is most likely for these reasons that the researchers chose to supplement the transactional data with linked survey data. This made possible a clearer few of why people were behaving in certain ways, as well as of how different categories of people were behaving differently, both standard social science interests. In sum, the transactional data alone would have been extremely limited in addressing the research questions.

2.4 Text corpus data

We believe that if we were to limit ourselves to the three types identified by Couper (2013), we would exclude an important category of big data available to social scientists. While other types of big data, such as social media and administrative data, might involve text and its interpretation, the focus is usually not the text as a large corpus. Rather, tweets might be categorized according to their tone or content, or administrative data might be processed so as to convert textual information into categories, similar to recoding open questions in surveys. In research involving what we call text corpus data, the focus is on the text itself, and the units of text are usually relatively long, such as speeches, articles, or even books.

The amount of textual data available in digital form is massive and growing extremely rapidly. For instance, Google counts over 60 trillion individual web pages, and the size of its index is over 100 million gigabytes[7]. One large source of the constant increase of text available online is the amount of news and magazine articles published every day. There is also all the content Internet users generate through their different activities.

Of particular interest to political scientists, political processes by their very nature generate a great number of texts, which are increasingly easy to access, for instance through open government initiatives. There are also large projects lead by libraries and publishers that digitize old books, newspapers, and journals. The U.S. Library of Congress has, for example, scanned millions of pages of American newspapers from 1836 to 1922[8]. The Digital Public Library of America in turn contains over eight million items from libraries, archives, and museums[9].

This increase in the amount of textual data, the growth of computing power, as well as the development and refinement of tools for computer-assisted analysis of text has led to new possibilities for research in the social sciences to study social processes and dynamics in innovative ways. The sheer quantity of data and the fact that researchers have at their disposal variables that were not previously considered makes it possible to arrive at new discoveries that advance theory, as well as inferences that were until now impossible. This might also increase the importance of bottom-up theories that could suggest new avenues for research that can subsequently be tested with more established methods (Iliev et al., 2015). On the other hand, this data-driven approach has also been one of the main criticisms against big data.

Various services make it relatively easy to gather large amounts of textual data. Online databases of articles (e.g., ProQuest, Lexis Nexis, and J STOR) are one possibility. J STOR offers already structured data on its “Data For Research” page, which includes article- and title-level metadata. Various government bodies also provide different data. The Swiss Parliament, for instance, offers access to important data on parliamentary activities in this way[10]. Social networks can be accessed through APIs and various paid services that offer analysis and data, such as Crimson Hexagon[11],

[7] http://www.google.com/insidesearch/howsearchworks/thestory/
[8] http://chroniclingamerica.loc.gov/
[9] http://www.dp.la
[10] http://www.parlament.ch/e/dokumentation/webservices-opendata/Pages/default.aspx
[11] http://www.crimsonhexagon.com/

which collects data from the large social media sites, blogs, forums, YouTube, news sites, and consumer reviews. Access to data from social media sites such as Facebook can be more limited for individual researchers, as most of the content is not made public by the users. When specific tools or services are not available, scraping the web for text allows unlimited access to content, even though the cost of gathering and cleaning the data is much higher in this case.

A large number of studies in different fields have made use of text corpus data. Psychologists, for example, have been able to compare results from surveys about personality and mental diseases with the writings of the same individuals. This has, for example, allowed exploration of theories about the psychological processes behind certain mental states. In one such study, Rude et al. (2004) found that "depressed individuals are preoccupied not only by negative thoughts but by heightened self-awareness" and that "inhibition of thoughts and emotions […] plays a role in continuing vulnerability to depression". Easily available large amounts of time-stamped text have also made it easy to conduct detailed research on historical trends and cultural change. For instance, Kesebir and Kesebir (2012) found that the frequency of words that express concern about others as well as of those characteristic of moral virtue have decreased during the last hundred years. Greenfield (2013) observed that during the last two hundred years the use of words that are associated with individualism and independence have increased in frequency.

Another way text data has been used is for the identification of different features of the author behind the text. Diermeier et al. (2011) for instance were able to predict with 92 percent accuracy the party affiliation of U.S. senators using their speeches. Likewise, the political orientation of bloggers was predicted with 92 percent accuracy by Deghani et al. (2014). It has also been shown possible to identify with similar accuracy the gender (Mukherjee and Liu, 2010), age and native language (Argamon et al., 2009), personality dimensions (Oberlander and Nowson, 2006), and sentiments (Dave et al., 2003) of authors.

In political science, text corpus data has also been used extensively. Speeches are one interesting source of text and can give insights into the mechanisms behind the functioning of governments. Examples of this are the analyses of the influence of military and political elites on the decision of Russia to intervene in neighbouring countries using public statements made by Russian leaders (Stewart and Zhukov, 2009), or the analysis of problem solving in the United States congress (Adler and

Wilkerson, 2013). Floor speeches in legislatures have been analysed to determine political attention given to particular topics (Quinn, 2010), and policymaking processes in various policy categories have been analysed using congressional hearing, public laws, and other sources (Jones et al., 2009). Catalinac (2014) used Japanese election manifestos released between 1986 and 2009 in order to study how electoral strategies had changed, especially after the electoral reform passed in 1994.

> *Case study: Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia.*

Studies can also focus on the nature of the web itself. Wagner et al. (2015) for instance studied gender bias in Wikipedia in six language editions. They focused on articles on notable people and studied four dimensions of possible bias. First, the authors compared the list of people on Wikipedia with various independent databases to see whether there was a bias in the number of individuals represented on the site. They also studied the links between articles about women and men to see whether there were any structural biases. Then, they measured whether the way both genders were described and the words that were used in the articles differed significantly. Finally, they measured the bias in visibility by comparing how often women and men appeared on the English language Wikipedia main page. The authors found no bias regarding the representation of women and men on the site or the visibility on the main page. However, they did find that articles by women tended to link to those by men more often than the opposite, possibly putting women at a disadvantage in terms of accessibility of their pages. In the Russian, English, and German versions men also tended to be more central in network analyses. Overall, the strongest structural bias was found in Russian and English editions.

Regarding the lexical bias, various findings emerged. First, in articles about women, the fact that the person was a woman was more often emphasized. And, in particular, information on relationships and family were more often discussed when the article described a woman. This was true for all languages, but the bias was strongest in English and Russian. The authors concluded that even though there did not seem to be any overt discrimination of women on Wikipedia, more subtle biases do exist, which could be due to the overrepresentation of men amongst contributors to the site. Even though only one of the four dimensions studied in this research uses text as the main data source, this study is a good example of how the web can be searched for different kinds of information without it necessarily falling into one of the three well-defined types of big data. In addition, it shows how textual data can be combined with metadata associated with the text (e.g., links between the different pages) to provide a more complete picture of the reality.

Although numerous interesting studies in the social sciences have been conducted using text corpus data and important insights have been gained, the research does often show the limits of current methods, especially regarding the methods of automated analysis of text used, which cannot compete with the way a human reader can extract meaning from text. A large part of the research focuses on validating methods instead of testing news theories or bringing significant scientific advances. Results might also at times seem somewhat self-evident. Researchers in the field are confronted with many obstacles. For instance, there are many methods to choose from and a great number of steps in preparing and analysing large amounts of text that are not neutral and can impact and bias results. There is also still a lack of best practices and clear guidelines that could guide researchers with respect to these complicated tasks, and to help them avoid pitfalls. On the other hand, the possibilities brought by the deluge of textual data and the continuous increase in computing power seem to be almost endless, and the state of the art is progressing very quickly.

3. Opportunities and limitations of big data for social science research

The previous sections demonstrate the potential value of different sorts of organic data for the social sciences. The large quantity of readily available data from a multitude of sources makes possible the application of new methods and insights for addressing both old and new research questions. It is no wonder then that so many social science researchers are now turning to big data sources in their work.

One might argue that, along the lines of Savage and Burrows (2007), that various forms of "big data" hold to potential not only ultimately to render obsolete more traditional methods in the social sciences, but also to revolutionize theory and the sorts of questions asked by researchers. Indeed, we are optimistic that the new and available forms of organic data will increasingly enrich research prospects and change in important ways how research is carried out in the social sciences. We fully expect that different kinds of administrative, social media, transactional, and large text corpus data will more and more find their way into innovative and original research methodologies that will be refined and standardized over time.

On the other hand, we believe that it is important to carefully cull the real potential from the hype around big data. Indeed, as illustrated in this paper, a review of the literature reveals a wide range of critiques and caveats that suggest quite significant limitations in what can be done with organic data for scientific purposes, and that at least present a convincing case that social scientists should not quickly abandon their traditional methods.

An important principle within the social sciences is that a study design should be optimal in producing valid and reliable data that fit and address the research questions of interest. Big data are usually not generated following a design intended to address specific research questions, and so generally do not easily lend themselves to use by social scientists. As stated by Lazer et al. (2014): “The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis”. Rather, existing data must be fit to particular research questions. This point is reflected in the following more specific critiques and caveats.

3.1 Analytic utility

The usefulness of some types of organic data may be limited from an analytic point of view. First, the organic data used for social research, whether for transactional or social media data, are generated by individuals who are not necessarily representative of whole populations, and so may suffer from selection bias (Couper, 2013). For example, transactional data may be gathered from people who have credit or loyalty cards, but these people may be different from those who do not possess credit or loyalty cards. Also, Twitter analyses are based on people who have Internet access, and who actively use Twitter. Unfortunately, in many cases it is difficult to identify the social characteristics of the people included in the data, and so it is impossible to say with any certainty which sections of the population are represented. Again, size of a dataset is secondary: “A dataset may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a dataset, we need to know where data is coming from…” (Boyd and Crawford, 2011).

Furthermore, organic data may be rich in nature, but may still exclude important types of information needed for analysis. For example, they generally do not include enough socio-demographic information about individuals to allow for the kinds of systematic cross-group comparisons that are typical of social scientific investigations, including multivariate analyses. Also, they can provide exhaustive information about specific behaviours, but are usually silent about the motives behind these behaviours or about people's intentions for the future.

3.2 Quality and access issues

There are also criticisms regarding the quality of organic data, which may compromise their value for research purposes (again, following Couper 2013). First, the data that end up in the hands of researchers may be incomplete, that is, without the same information on all individuals in the file. The data may also be messy, unstructured, and poorly documented. The sources of error and total error involved in organic data are currently still not well understood. With respect to social media data especially, there are questions about the truthfulness of information gathered, since much of what is made visible online involves self-image management, no doubt with some distortions.

Another problem is that the sources of organic data may be relatively fleeting – online systems like Facebook come and go. This means that while such data may be useful for short-term time scales, it may not be possible to make comparisons over time, either because the system goes down, or because it changes its nature and/or its clientele. Issues of confidentiality always risk at any moment to become blockers of access for the research community.

Access to organic data is also a significant obstacle for researchers. Administrative data are notoriously difficult to obtain, as well as transactional data, which are usually generated by private companies that promise data protection to their clients. Although social media data can be accessed, it is usually only under certain conditions, and rarely is everything that is of interest available to researchers.

4. Big data at FORS?

FORS recognizes that organic data, in the form of administrative data, social media data, transactional data, and text corpus data have significant potential for advancing social scientific research. We look forward to following and learning from new techniques and innovations that will allow such data to be accessed and used in fruitful and original ways.

However, we recognize as well that there are still severe limitations to using such data appropriately in a scientific context, notably concerning their real utility, their quality, and their accessibility. If they are used at all, we believe that a critical view should prevail, informed by current social science best practice and expertise. Analytical strategies should be suitable for particular data types, should be improved, and error sources should be better identified and tackled. Also, ethical concerns are primordial – personal data must always be treated with utmost care, even if these are openly available.

Most importantly, we believe that at this time organic data should supplement, but not replace traditional methods and data sources in the social sciences. Within this perspective, we see several main ways in which our institution can employ big data/organic data in the near future at the service of the research community.

First, FORS can do more to facilitate the use of *administrative data* for research purposes. For example, we plan to map the existing provision of administrative data from various sources in Switzerland, so that researchers can have an overview of possibilities. This should include information on the procedures required to gain access to such data. In addition, FORS will examine possibilities for enriching FORS datasets by linking them with administrative data from the Swiss Federal Statistical Office and other federal offices. We will continue to conduct methodological research that depends on the register frame used for FORS surveys. And we will study questions concerning particular disclosure risks associated with administrative data use.

Second, some social media and text corpus data can be used as *contextual information* for surveys. For data analysis, there are interesting examples of combining survey data and text analysis from social media or other sources where individuals can be linked (Schwartz et al., 2013; Rude et al., 2004). It remains to be seen how feasible this is in the context of surveys at FORS, especially in terms of getting the consent of participants. This would be much easier for elite surveys such as the Comparative Candidates Survey (CCS), where it would be possible to link social media posts, blogs, articles, and other sources written by or about the candidates. For general population surveys, it would also be feasible to collect contextual data, for instance in-depth analysis of different media (including television using the transcripts already available in the form of subtitles), or studying public opinion on social media, blogs, forums, comment sections of newspapers, etc. It could also be thought of as a way to explore relevant categories and questions when designing questionnaires.

Even though open-ended questions in surveys do not count as big data, expertise in automated text analysis can help make better use of open-ended questions in surveys that are often underutilized. This might overcome problems of resources to code the questions, but also improve the analysis of the questions or use them in different ways. For instance, predefined coding schemes might not capture some differences that might prove important.

Third, our data service will do more to solicit, *curate, preserve, and disseminate* rich and diverse forms of data, including organic data, which can be used on their own for secondary analyses, or in connection with traditional data sources. This might include databases of different kinds of objects, e.g., job announcements, Twitter feeds, etc. We already have the infrastructure and know-how to handle diverse types of qualitative data. It should not be a large step to be able to integrate and disseminate organic data.

To take an example, it is important to be ready to store and disseminate data that have been prepared in studies using automated text analysis. The amount of work needed to prepare these data is intensive and many researchers might be interested in using the same data. This is especially the case for data collected from publicly available documents, such as information press releases and other communications of parties or institutions for instance. Many research questions could benefit from the same material, and as the methods are rapidly improving other researchers might want to try to replicate findings with refined methods. Also, journals increasingly require data to be made public for published articles, which will augment the likelihood of having requests to archive this kind of data. It is an open question whether FORS would be allowed to offer anonymized data from social media, traditional media, blogs, and other sources due to copyright or privacy concerns. In any event, we will continue to examine these questions and to prepare for the eventuality of archiving larger and more diverse types of data.

Finally, another role that FORS can play as a centre of expertise in the social sciences is to provide *general guidance to researchers* in Switzerland for working appropriately with various sorts of organic data. Through our own experiences, and by keeping up on current developments, we will be able to advise researchers on: identifying potential non-traditional data sources; accessing them; assessing their quality and real utility in addressing specific research questions; and avoiding misinterpretation. Toward these ends we will continue to train our staff and to study the ways in which big data can benefit the social sciences.

5. References

Adler, E.S., Wilkerson, J. 2011. *The Congressional bills project*. Available at: http://www.congressionalbills.org.

Argamon, S., Koppel, M., Pennebaker, J.W., and Schler, J. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM* 52(2): 119-123.

Bakker, F.M. 2012. Estimating the validity of administrative variables. Statistica Neerlandica 66(1): 8-17

Bakker, B., and Rooijen, J.V. 2012. Methodological challenges of register-based research. *Statistica Neerlandica* 66 (1).

Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. Latent dirichlet allocation. *The Journal of machine Learning research* 3: 993-1022.

Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D.I., Marlow, C., Settle, J.E., and Fowler J.H. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415): 295-298.

Boyd D, and Crawford K. 2012. Critical questions for big data. *Information, communication & Society* 15(5): 662-679.

Boyd, D., and Crawford, K. 2011. Six Provocations for Big Data. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September. Available at SSRN:http://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431

Budge, I., Klingemann, H-D., Volkens, A., Bara, J., and Tanenbaum, E. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press: Oxford.

Calderwood, L., and Lessof, C. 2009. Enhancing longitudinal surveys by linking to administrative data. In *Methodology of longitudinal surveys,* Lynn P (ed.). John Wiley & Sons: Chichester; pp. 55-72.

Catalinac, A. 2014. *Pork to policy: The Rise of National Security in Elections in Japan*. Unpublished manuscript.

Choi, H., and Varian, H. 2012. Predicting the Present with Google Trends. *Economic Record* 88: 2-9.

Christoffersen, M.N., Poulsen, H.D., and Nielsen, A. 2003. Attempted suicide among young people: risk factors in a prospective register based study of Danish children born in 1966. *Acta Psychiatrica Scandinavia* 108: 350-358.

Couper, M. 2013, Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, 7(3), 145-156.

Dave, K., Lawrence, S., and Pennock, D.M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews . In *Proceedings of the 12th International Conference on World Wide Web;* 519-528.

Dehghani, M., Sagae, K., Sachdeva, S., and Gratch, J. 2014. Linguistic analysis of the debate over the construction of the 'Ground Zero Mosque'. *Journal of Information Technology & Politics* 11: 1-14.

Diermeier, D., Godbout, J.F., Yu, B., and Kaufmann, S. 2011. Language and ideology in Congress. *British Journal of Political Science* 42(1): 31-55.

Eastham, L.A. 2011. Research using blogs for data: public documents or private musings? *Research in Nursing & Health* 34 (4): 353-361.

Eshbaugh-Soha, M. 2010. The tone of local presidential news coverage. *Political Communication* 27(2): 121-140.

Golder, S.A., and Macy, M.W. 2014. Digital footprints: opportunities and challenges for online social research. *Annual Review of Sociology* 40: 129-52.

Golder S.A., and Macy, M.W. 2011. Diurnal and Seasonal Mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051): 1878-1881.

Greenfield, P.M. 2013. The changing psychology of culture from 1800 through 2000.*Psychological Science* 24(9): 1722-1731.

Grimmer, J., Stewart, B.M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*: 267-297.

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

Hargittai, E. 2007. Whose Space? Differences Among Users and Non-Users of Social Network Sites. *Journal of Computer-Mediated Communication* 13: 276–297.

Holmberg, A. 2012. Discussion on assessing quality of administrative data. Statistical Neerlandica 66(1): 34-40.

Hookway, N. 2008. Entering the blogosphere: some strategies for using blogs in social research. *Qualitative Research* 8(1): 91-113.

Hopkins, D., and King, G. 2010. Extracting systematic social science meaning from text. *American Journal of Political Science* 54(1): 229-247.

Iliev, R., Dehghani, M., and Sagi, E. 2014. Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition*: 1-26.

Jashinsky, J., Burton, S.H., Hanson, C.L., West, J., Giraud-Carrier, C., Barnes, M.D., and Argyle, T. 2013. Tracking Suicide Risk factors through twitter in the US. *Crisis*.

Jutte, D.P., Roos, L.L., and Brownell, M.D. 2011. Administrative record linkage as a tool for public health research. *The annual Review of Public Health* 32: 91-108.

Jones, B., Wilkerson, J., and Baumgartner, F. 2009. *The policy agendas project*. Available at: http://www.policyagendas.org.

Kaplan. A.M., and Haenlein, M. 2010. Users of the world, unite! The challenges and opportunities of social media. *Business Horizon*s 53: 59-68.

Kesebir, P., and Kesebir, S. 2012. The cultural salience of moral character and virtue declined in twentieth century America. *Journal of Positive Psychology* 7(6): 471-480.

Khoury, M.J., and Ioannidis, J.P.A. 2014. Big data meets public health. *AAAS* 346(6213): 1054-1055.

Kramer, A.D.I., Guillory, J., and Hancock, J.T. 2014.Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111(24): 8788-8790.

Landauer, T.K., Foltz, P.W., and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3): 259-284.

Laurie, G., and Stevens, L. 2014. *The administrative data research centre Scotland: A scoping report on the legal & ethical issues arising from access & linkage of administrative data*. Research paper series 2014/35. University of Edinburgh, School of law: Edinburgh.

Lazer. D., Kennedy, R., King, G., and Vespignani, A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176): 1203-1205.

McCormick, T., Lee, H., Cesare, N., and Shojaie, A. 2013. *Using Twitter for Demographic and Social Science Research: Tools for Data Collection***.** Working paper, Center for Statistics and the social sciences. Available at: http://paa2013.princeton.edu/papers/130624

Mukherjee, A., and Liu, B. 2010. Improving gender classification of blog authors. In *Proceedings of Conference on Empirical Methods in Natural Language Processing,* MIT: Massachusetts; 207-217.

Oberlander, J., and Nowson, S. 2006. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions;* 627-634.

Parnell, B-I. 2014. *Scientists warn about bias in the facebook and twitter data used in millions of studies*. Blog available at: http://www.forbes.com/sites/bridaineparnell/2014/11/27/scientists-warn-about-bias-in-the-facebook-and-twitter-data-used-in-millions-of-studies/

Pfeffer, J., and Ruths, D. 2014. A few caveats for building social media research mavens. Available at: http://www.socialsciencespace.com/2014/12/a-few-caveats-for-budding-social-media-research-mavens/

Puschmann, C., and Burgess, J. 2013. The politics of Twitter data. Discussion paper no 2013-01. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2206225

Quinn, K. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54(1): 209-28.

Rude, S., Gortner, E.M., and Pennebaker, J. 2004. Language use of depressed and depression vulnerable college students. *Cognition & Emotion* 18 (8): 1121-1133.

Ruppert, E., and Savage, M. 2009, New Populations: Scoping Paper on Digital Transactional Data, CRESC Working Paper 74. Available at: http://www.cresc.ac.uk/medialibrary/workingpapers/wp74.pdf

Ruths, D., Pfeffer, J. 2014. Social media for large studies of behaviour. *Science magazine* 346(6213): 1063-1064.

Savage, M., and Burrows, R. 2007. The coming crisis of empirical sociology. *Sociology*, 41(5), 885-899.

Schroeder, R. 2014. Big Data: towards a more scientific social science and humanities? In *Society and the Internet*, Graham M, Dutton WH (eds.) Oxford University Press: Oxford; 164–176.

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., and Ramones, S.M. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8(9).

Stewart, B.M., and Yuri, M.Z. 2009. Use of force and civil–military relations in Russia: An automated content analysis. *Small Wars & Insurgencies* 20: 319-43.

Theocharis, Y. 2011. The influence of postmaterialist orientations on young British people's offline and online political participation. Representation 47:4: 435-455.

Tinati, R., Halford, S., Carr, L., and Pope, C. 2014. Big data: methodological challenges and approaches for sociological analysis. *Sociology* 48(4):663-681.

Von Gunten, L., Hümbelin, O., and Fritschi, T. 2014. *Administrative data: benefits and challenges for social security research*. Working paper. Berne University of Applied Science, Social work division: Berne.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *arXiv preprint arXiv:1501.06307*.

Wallgren, A., and Wallgren, B. 2011. *To understand the possibilities of administrative data you must change your statistical paradigm*. Section on Survey research methods. JSM. Available at: https://www.amstat.org/sections/srms/Proceedings/y2011/Files/300347_64422.pdf

Wanner, P. 2006. *L'utilisation de données administratives pour l'analyse des comportements professionnels en fin de vie. Exemple pratique, perspectives méthodologiques et limites.* Actes de colloques de l'AIDELF : 133-144.

Wardle, H., Ireland, E., Sharman, S., Excell, D., and Gonzalez-Ordonez, D. 2014. Patterns of play: analysis of data from machines in bookmakers. NatCen. Available at: http://www.responsiblegamblingtrust.org.uk/user_uploads/pdfs/patterns%20of%20play%20-%20analysis%20of%20data%20from%20machines%20in%20bookmakers.pdf

Wardle, H., Parke, J., Excell, D. 2013. *Machines Research Programme: Report 1 - Theoretical markers of harm for machine play in a bookmaker's*. NatCen. Available at: http://www.responsiblegamblingtrust.org.uk/user_uploads/pdfs/report%201%20theoretical%20markers%20of%20harm%20for%20machine%20play%20in%20a%20bookmakers%20-%20a%20rapid%20scoping%20review.pdf