education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 25 Number 94 August 28, 2017 ISSN 1068-2341

Education Policy Research in the Big Data Era: Methodological Frontiers, Misconceptions, and Challenges

Yinying Wang
Georgia State University
United States

Citation: Wang, Y. (2017). Education policy research in the big data era: Methodological frontiers, misconceptions, and challenges. *Education Policy Analysis Archives, 25*(94). http://dx.doi.org/10.14507/epaa.25.3037

Abstract: Despite abundant data and increasing data availability brought by technological advances, there has been very limited education policy studies that have capitalized on big data—characterized by large volume, wide variety, and high velocity. Drawing on the recent progress of using big data in public policy and computational social science research, this commentary discusses how to approach big data and how big data can be used in education policy research. First, I introduce big data that is potentially relevant to education policy research. I then present methodological frontiers by examining the assumptions, key concepts, merits, and caveats of three commonly used analytical approaches to mining massive amounts of text data: topic models, network text analysis, and sentiment analysis. Next, to ensure the veracity of using big data in education policy research, I debunk three methodological misconceptions. This commentary concludes with a discussion on developing interdisciplinary research capacity and addressing the privacy concerns and ethical conundrums as we explore a research agenda of using big data in education policy.

Keywords: big data; education policy; network text analysis; sentiment analysis; text mining; topic models

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 5/5/2017
Revisions received: 20/6/2017
Accepted: 5/7/2017

Investigación de política educativa en la era de los grandes datos: Fronteras metodológicas, equívocos, y desafíos

Resumen: A pesar de la abundancia de datos y del aumento de la disponibilidad de datos traídos por los avances tecnológicos, hubo estudios de políticas educativas muy limitadas que usaron datos importantes, caracterizados por gran volumen, gran variedad y alta velocidad. En base al reciente progreso del uso de grandes datos en investigaciones de políticas públicas y de ciencia social computacional, este trabajo pretende demostrar el potencial de datos importantes y la gran cantidad de datos que pueden ser utilizados en la investigación de políticas educativas. En primer lugar, introduzca datos importantes que son potencialmente relevantes para la investigación de políticas educativas. Puedo, entonces, presentar fronteras metodológicas, examinando los supuestos conceptos clave, méritos y salvedades de tres enfoques analíticos comúnmente usados en la minería de grandes cantidades de datos de texto: modelos de tópicos, análisis de conexiones textuales y análisis de sentimientos. A continuación, para garantizar la veracidad del uso de grandes datos en la investigación sobre políticas educativas, desenmascaramos tres equívocos metodológicos. Este artículo concluye con una discusión sobre el desarrollo de la capacidad de investigación interdisciplinaria abordando las preocupaciones de privacidad y los enigmas éticos a medida que exploramos una agenda de investigación de uso de datos importantes en la política educativa.
Palabras-clave: datos grandes; política educativa; análisis de conexiones textuales; análisis de sentimientos; minería de textos; modelos de temas

Pesquisa de política educacional na era dos grandes dados: Fronteiras metodológicas, equívocos, e desafios

Resumo: Apesar da abundância de dados e do aumento da disponibilidade de dados trazidos pelos avanços tecnológicos, houveram estudos de políticas educacionais muito limitados que usaram dados importantes, caracterizados por grande volume, grande variedade e alta velocidade. Com base no recente progresso do uso de grandes dados em pesquisas de políticas públicas e de ciência social computacional, este trabalho pretende demonstrar o potencial de dados importantes e a grande quantidade de dados que podem ser utilizados na pesquisa de políticas educacionais. Primeiro, eu introduzo dados importantes que são potencialmente relevantes para a pesquisa de políticas educacionais. Posso, então, apresentar fronteiras metodológicas, examinando os pressupostos conceitos-chave, méritos e ressalvas de três abordagens analíticas comumente usadas na mineração de grandes quantidades de dados de texto: modelos de tópicos, análise de conexões textuais e análise de sentimentos. Em seguida, para garantir a veracidade do uso de grandes dados na pesquisa sobre políticas educacionais, desmascaramos três equívocos metodológicos. Este artigo conclui com uma discussão sobre o desenvolvimento da capacidade de pesquisa interdisciplinar abordando as preocupações de privacidade e os enigmas éticos à medida que exploramos uma agenda de pesquisa de uso de dados importantes na política educacional.
Palavras-chave: dados grandes; política educacional; análise de conexões textuais; análise de sentimento; mineração de textos; modelos de tópicos

Introduction

The rise of big data—characterized by large volume, wide variety, and high velocity (boyd & Crawford, 2012)—has breathed new life into the social sciences (King, 2011). The progress in acquiring, processing, and analyzing big data has enlightened many fields in the social sciences, such as political science (Bode, Hanna, Yang, & Shah, 2015; Schwartz & Ungar, 2015), public health

(Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Bates, Saria, Ohno-Machado, Shah, & Escobar, 2014), economics (Einav & Levin, 2014), and criminology (Chen, Cho, & Jang, 2015), to name a few. In the field of education, despite the fast growing body of literature on learning analytics—collecting and analyzing big data to optimize student learning, particularly in online learning environment (Baker & Yacef, 2009), there has been limited scholarship on capitalizing on big data in education policy.

To discuss the potential of big data and how big data can be used in education policy research, in this commentary I first introduce big data that is potentially relevant to education policy research. Given the conspicuous absence of education policy studies using big data, I primarily draw upon the literature on big data in public policy and computational social science—the emerging field at the convergence of computer science and the social sciences, using computational modeling to analyze massive amounts of digital data harvested mostly from digital media sources to study social phenomenon (Lazer et al., 2009; Shah, Cappella, & Neuman, 2015; Watts, 2013). Grounded in the broad literature relevant to education policy research, I then introduce three methodological frontiers of mining massive amounts of text data (i.e., a corpus of texts; Fleuren & Alkema, 2015; Hearst, 1999): topic models, network text analysis, and sentiment analysis. In particular, I examine the assumptions, key concepts, merits, and caveats of each of the three analytical approaches. Next, to ensure the veracity of using big data in education policy research, I debunk three methodological misconceptions. This paper concludes with the discussion on developing interdisciplinary research capacity and addressing the privacy concerns and ethical conundrums as we explore a research agenda of using big data in education policy.

Big Data in Education Policy Research

What is big data? In fact, big data is very loosely defined. There is no arbitrary cutoff point regarding how big the data must be in order to be considered as big data. Generally speaking, big data has three distinct features: large volume, wide variety, and high velocity (boyd & Crawford, 2012). In addition, some posit a fourth feature of veracity—the trustworthiness and accuracy of big data (Bello-Orgaz, Jung, & Camacho, 2015). In this commentary, I draw upon the literature of big data in public policy research and computational social science to delineate each of the first three features of big data, followed by the potential value of big data in education policy research. I then address the fourth feature of veracity as I present the methodological frontiers and debunk potential misconceptions. Here I proceed to introduce the first three features of big data: large volume, wide variety, and high velocity.

Large Volume

The first defining feature of big data is its massive volume. While the name of "big data" suggests large volume, there is no arbitrary cutoff point distinguishing "big" from "small" data. Oftentimes, the volume of big data is staggeringly large, such as over 118,000 speeches in the U.S. Senate (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010), over 16,000 articles in the journal *Science* from 1990 to 1999 (Blei & Lafferty, 2007), 13,246 political blog posts on the 2008 presidential election (Roberts, Stewart, & Tingley, 2016), and over 16,000 documents on climate change (Boussalis & Coan, 2016).

In education policy research, the large volume of data can come from a host of sources. First, data-driven education accountability systems provide the unprecedentedly large volume of data on students, teachers, administrators, schools, and communities. For a ready example, a recent study tapped into the record of about 200 million test scores in math and reading to examine educational inequalities (Reardon, Kalogrides, & Shores, 2016). Second, digital footprints—digital trace data we create as we use digital services and devices (Howison, Wiggins, & Crowston, 2011)—are another source of big data that are valuable to education policy research. The data generated from our online activities, either posting tweets on Twitter (e.g., Barberá, 2015) or reading the Facebook newsfeed (e.g., Kramer, Guillory, & Hancock, 2014), could serve as the proxy for online public opinion on education policy. For instance, the hashtags #CommonCore and #CCSS are among the most frequently used hashtags by the digital public when discussing the education policy on the Common Core State Standards on social media (Supovitz & Reinkordt, 2017; Wang & Fikis, 2016). One of the frequently co-occurring hashtags with #CommonCore and #CCSS is #OptOut, which refers to the movement of opting out of high-stakes standardized testing (Wang & Fikis, 2017). At the epicenter of the opt-out movement in the state of New York, the Long Island Opt-out group—a major advocate for opting out of state standardized testing—has tens of thousands of Facebook group members discussing testing and sharing resources on how to opt out (Wang, 2017). Third, technological advances in data acquisition increase data availability for education policy research. For instance, the optical character recognition (OCR) was used to convert the un-digitalized text data into the data that can be read in computers in a study examining the latent topics in 1,539 articles published in *Educational Administration Quarterly* from 1965 to 2014 (Wang, Bowers, & Fikis, 2017). Another common data acquisition approach is to use application programming interfaces (APIs) to retrieve data generated on the Internet. Many technology companies (e.g., Google, Facebook, and Twitter) use APIs to grant others limited access to their data so that more applications can be created using their data. Twitter API is one of the most popular APIs among researchers. For instance, Twitter API was used to collect 660,051 tweets containing the hashtags of #CommonCore and #CCSS to study online public opinion on the Common Core State Standards (Wang & Fikis, 2017). Fourth, collaborations with technology companies allow academics to access big data. An example is a collaborative research undertaking that carried out a randomized controlled experiment involving 61 million Facebook users to study social influence within online social networks (Bond et al., 2012). All these sources—the data-driven education accountability systems, the digital footprints generated by digital services and devices, data acquisition technologies, and collaborations with technology companies—provide the unprecedentedly voluminous amounts of data for education policy research.

Wide Variety

Big data does not simply mean large volume. It also refers to the wide variety of types of data, including, but not limited to, videos, images, audios, media coverage, blogs, social media posts, online comments, records of government agencies, mobile phone logs, data generated by wearable digital devices, and satellite images. These diverse data sources, along with the conventional data sources in education policy research (e.g., interviews, observations, artifacts, surveys, and archived documents), provide researchers with rich, multi-dimensional insights into the education policy issues of interest.

One potential for using big data in education policy research is to investigate public opinion on education policy. Since public opinion is one of the factors shaping policymaking (Burstein, 2003; Gormley Jr., 2016; Page & Shapiro, 1983), researchers have studied public opinion on a variety of education issues, including the policy in early childhood education (Gormley Jr., 2016), school reform (Henderson, Peterson, & West, 2016), school quality (Jacobsen, Snyder, & Saultz, 2014), and race-based and wealth-based student achievement gaps (Valant & Newark, 2016). Prior studies on public opinion on education and policy used the data collected from surveys, including the EdNext annual survey of American public opinion on education (Peterson, Henderson, West, & Barrows, 2017), the survey administered to the members of YouGov—a company conducting academic survey research and online political polling (Valant & Newark, 2016), and the survey fielded by Knowledge Networks to a sample representative of the U.S. population to study the effect of school report card formats on public opinion on school quality (Jacobsen, Snyder, & Saultz, 2014). In the big data era, in addition to survey data, a growing number of studies have examined public opinion on policy issues using social media data—the data generated by the digital public as they discuss policy issues on social media. A few examples could suffice. Over 120,000 tweets were collected to examine public opinion on health reform (King et al., 2013). Chung and Zeng (2015) developed a system called 'iMood' to identify opinion leaders, influential users, and community activists on the U.S. immigration policy by analyzing about one million tweets. Whitman (2015) detailed how the data from Twitter and the Google Trends were used to measure public opinion on the U.S. space policy. Reddicka, Chatfieldb and Jaramilloa (2015) examined public opinion on National Security Agency massive surveillance programs of Americans through a critical discourse analysis of tweets along with the survey data collected from a randomly sampled public opinion poll of Americans.

In addition to social media data, researchers have been exploring how to use other sources of big data, such as mobile phone metadata and satellite images. Toole and his team demonstrated the potential of using mobile phone metadata to improve the forecasts of critical economic indicators for governments' policymaking (Toole et al., 2015). Toole et al. used the mobile phone metadata (e.g., who called whom, the location of the cell towers through which the calls were made, the time of the calls, the total number of calls, and the number of incoming and outgoing calls) as a proxy to detect the changes in mobility and social interactions caused by layoffs, and then predicted the aggregated unemployment rates months before the official reports were released. Moreover, to leverage the value of image data, a novel approach is to use satellite-recorded nighttime lights to estimate the poverty and economic growth (Henderson, Storeygard, & Weil, 2012; Pinkovskiy & Sala-i-Martin, 2015; Xie, Jean, Burke, Lobell, & Ermon, 2016). Suffice it to say, these examples present the tantalizing potential of using big data in education policy research.

High Velocity

The third feature of big data is high velocity: the high speed of data generation. Unlike survey datawhich are generated periodically whenever the survey is administered, much of the big data—such as web browsing, Facebook status updates, YouTube videos, phone call logs—are generated in a constant stream. High-velocity streaming data pose both methodological challenges and opportunities of processing the constant incoming data, particularly processing data in a timely fashion, so that the data can be analyzed in real-time or near real-time. For instance, a group of researchers capitalized on the simplicity (i.e., no more than 140 characters) and immediacy features of tweets to detect the traffic events in urban areas (Gutierrez, Figuerias, Oliveira, Costa, & Jardim-Goncalves, 2015). In education policy research, a common limitation is the time lag of months, if not years, between data collection and result report. In the big data era, one solution to the time lag limitation is to capitalize on diverse data sources and automate or semi-automate data processing, analysis, and visualization, so that raw data can be processed, analyzed, reported, and visualized in a timely manner to inform education policymaking, implementation, and evaluation. In fact, the methodological advances in real-time analytics have been fast growing. The three methodological frontiers introduced in the following section can all be automated or semi-automated. The real-time analytics is definitely an area that those who are interested in using big data in education policy research should watch for.

Methodological Frontiers

In the big data era, the wealth of data available to social scientists has been considered as the microscope to microbiologists (King, 2011). It is certainly appealing that big data could enrich our understanding of social phenomena and relevant education policy issues. Yet big data, as King (2016) noted, "is not about the data" (p. iii). Rather, big data is about the *analytics*—the methodological approaches that extract insights from large-volume, wide-variety, and high-velocity data. To surmount the methodological challenges brought by big data, emerging analytical tools have been rapidly developed at the intersection of the social sciences and computer science to analyze massive amounts of digital data (Alvarez, 2016; Lazer et al., 2009; National Research Council, 2013; Shah, Cappella, & Neuman, 2015; Watts, 2013). Given the conspicuous absence of education policy studies using big data, I primarily draw upon the relevant literature on big data in public policy and computational social science to introduce three methodological frontiers—topic models, network text analysis, and sentiment analysis—along with their assumptions, key concepts, merits, and caveats. It is important to stress at the outset that the methodological approaches introduced here are only the commonly used ones (Alvarez, 2016; Feldman, 2013; Verd & Lozares, 2014). They do not represent a comprehensive survey of all the methodological tools to analyze big data. Neither do they supplant conventional methodological approaches in education policy research. By introducing the methodological tools, this article aims to invite education policy researchers to venture into big data by applying the tools to education policy research.

Topic Models

Topic modeling (Blei, 2012; Blei, Ng, & Jordan, 2003) is one of the prominent, rapidly developing methods to infer latent topics in massive amounts of text data. The topic models assume that each document has multiple topics, and each topic can be inferred by the probability distribution over a set of words. For example, the topic of social justice is described by the words such as "inequity," "justice," "race," "disability," and "bilingual" (Wang et al., 2017). In addition to identifying latent topics in text data, topic models have been developed to uncover how the topics are related to one another—such as the correlated topic models (Blei & Lafferty, 2007), and how the topics evolve over time—such as the dynamic topic models (Blei & Lafferty, 2006).

The popularity of topic models is partly explained by the fact that they are effective and scalable to explore the latent topics in massive amounts of text data. Topic models can be either fully automated (i.e., unsupervised) or semi-automated (i.e., supervised) (Roberts et al., 2016). When unsupervised topic models are applied, no previous annotations or labeling of the documents is needed, as the topics are identified by the high probability words that describe a particular topic. Therefore, topic models are highly scalable, without the constraint of human cognition entailed in manual coding. Thanks to the scalability, topic models have been applied to analyze the large volume of text datasets from a variety of data sources: 24,236 press releases from the U.S. Senate (Grimmer, 2010), the opinion texts from the U.S. Supreme Court's 4,321 non-unanimous Court decision from 1949 to 2006 (Lauderdale & Clark, 2014), over 21,000 scholarly articles on literary studies over the last 120 years (Goldstone & Underwood, 2014), and 233,452 online healthcare chat logs (Wang, Huang, & Gan, 2016). Recently, topic models have been used for automated annotation of images (Feng & Lapata, 2010) and videos (Katsurai, Ogawa, & Haseyama, 2012).

In education policy research, topic models can be applied to infer latent topics in large corpora compiled from multiple data sources. They include records of government agencies, news coverage, and social media discourse throughout the policy processes from problem identification and framing to ongoing evaluations of existing policies. Since no study was found that has applied topic models in education policy research, here I present a topic modeling study in education research. Wang and her team used topic modeling and identified 120 latent topics in 16,524 documents published in the 116-year history of the longest running journal in education *Teachers College Record* (*TCR*) from 1900 to 2015 (Wang, Bowers, Chae, Fikis, & Natriello, 2017). The 120 topics were identified by generating two matrices: (1) a matrix of word probabilities by latent topics, and (2) a matrix of topic proportions by *TCR* articles (see Blei, 2012, for a thorough explication of probabilistic topic modeling). Among the 120 topics in education literature over the last 116 years, the topic of education in wartime disappeared after the end of cold war in the 1980s; the topic of social justice has been on the rise since the 1950s; the topic of measurement of student achievement has garnered persistent attention in the education research community since the 1900s.

Researchers who apply topic models to education policy research need to be aware of a couple of caveats. First, topic models, albeit valuable, are only a blunt tool to explore large corpora. Unlike the fine-grained results generated by manual coding, the results of topic models are a coarse-grained description of the text datasets. Further, topic models, even unsupervised topic models, necessitate researchers' subjective decisions on modeling and choosing the best fitting model. Most topic models do not run on every single word in the datasets. Rather, before running topic models, the text data need to be processed and cleaned. In this data cleaning process, a common practice is to remove the words that do not convey topical meaning or the words that are used most and least frequently. To do so, scholars may take different approaches, as "there is no globally best method" (Grimmer & Stewart, 2013, p. 3). A few examples would suffice. Blei and Lafferty (2007) removed the words by two criteria: the words occurred fewer than 70 times, and the 296 stopwords such as "a," "an," "the," and "around". Roberts et al. (2016) removed the words that occurred fewer than 1% of the 13,246 blog posts. Grimmer (2010) removed the words that occurred fewer than 0.5% and over 90% of the documents, as well as the stopwords. Grün and Hornik (2011) calculated the mean term frequency-inverse document frequency (tf-idf), and then only included the words that have a tf-idf value slightly above the median to remove the very frequently used words. All these examples demonstrate how running topic models entail the researchers' modeling decisions. Moreover, researchers interpret the topic model output and oftentimes label the topics through drawing upon the researchers' knowledge on the context of the text data. For instance, to label each topic, the researchers examined the topics that have been identified periodically in the existing literature in the field, and took account into the results generated from the topic models: the high-probability terms and the high-probability articles (Wang et al., 2017). This process of data interpretation relies on the researchers' contextual knowledge of the data. For this reason, topic models, while do not replace humans, "amplify human abilities" (Grimmer & Stewart, 2013, p. 4). To that end, the most promising way of automated text mining is not to negate the researchers' need to read the texts, but rather "to identify the best way to use both humans and automated methods for analyzing texts" (Grimmer & Stewart, 2013, p. 4).

Network Text Analysis

Network text analysis is another emerging methodological frontier to analyze text data. A network is formed by nodes and ties (Borgatti, Everett, & Johnson, 2013). To conceptualize texts as networks, the units of texts (i.e., words or concepts) are connected by the ties (i.e., the co-occurrence of words or the relationships between concepts; Verd & Lozares, 2014). Network text analysis assumes that the semantic meaning of texts is revealed by the patterns of network structure—how the units of texts are connected in the network (Diesner & Carley, 2005). For instance, in the network text analysis of stem-cell research literature, the words—such as "bone," "marrow," "transplantation," "tissue," and "peripheral"—are tightly connected by their co-occurrence ties

(Leydesdorff & Hellsten, 2005), denoting that the co-occurrence relationships between the tightly connected words are stronger than those words with others. In the network analysis of texts compiled by 1.9 billion anonymous queries on epilepsy, the nodes are the words that are connected by the co-occurrence ties; the tightly connected clusters of words suggest the topics, including the seizures and their effects on the body, anti-epileptic drugs and their side effects, and the life changes (e.g., driving and employment) related to epilepsy (Miller, Groves, Knopf, Otte, & Silverman, 2017). In addition to conceptualizing words as the units of texts in the networks, researchers also consider hashtags as the units of texts in analyzing the text data acquired from social media. For instance, in the network analysis of 660,051 tweets containing the hashtags #CommonCore and #CCSS (the Common Core State Standards), the nodes are the hashtags that are connected by the co-occurrence ties; the tightly connected clusters detected by the faction algorithm suggest the online discourse on the Common Core State Standards on Twitter centered around the topics, including the Republican Party's 2016 presidential candidates (e.g., #Trump2016 and #TedCruz2016), anti-Common Core (e.g., #StopCommonCor and #StopCC), education policy and reform (e.g., #NCLB and #ESSA), as well as teaching and testing (e.g., #teaching and #testing) (Wang & Fikis, 2017). Another novel approach to conceptualizing texts as the networks is to consider nodes as the nouns that are connected by the verbs as ties. For instance, in the network analysis of 130,213 news articles on the 2012 U.S. presidential elections, the nodes are the nouns such as Obama, economy, and efforts, and the ties are the verbs such as celebrate, acclaim, and blame (Sudhahar, Veltri, & Cristianini, 2015).

Both network text analysis and the aforementioned topic models are emerging methodological approaches to analyze text data. One might wonder: If we use the two methods to analyze the same text dataset, do their results differ? The answer, according to the literature (Leydesdorff & Nerghes, 2015; Wang & Fikis, 2016), is affirmative. In fact, the two methods can yield the results that are significantly uncorrelated. This by no means suggests topics models and network text analysis are invalid. Rather, it suggests the methods work well if applied for different purposes: the topic modeling is more appropriate to reveal similarities, whereas network text analysis is more appropriate to reveal semantic meanings. The different results yielded by using topic models and network text analysis are analogous to the different results generated from using different conceptual frameworks and coding schemes when manually coding the same dataset in qualitative research.

It is worth noting that when analyzing the hashtag co-occurrence networks, it is of critical importance to select the appropriate hashtags. The hashtags might keep changing and evolving over the process of education policymaking and implementation. Correspondingly, the appropriate hashtags should be broad enough to include all permutations of the words and phrases relevant to the topic of interest. However, if it is too broad, there is a risk of including irrelevant content, adding noise to data. Therefore, an iterative process is recommended to examine the data retrieved using the selected hashtags.

Sentiment Analysis

The third methodological frontier that holds great promise in education policy research is sentiment analysis, also known as opinion mining. Sentiment analysis identifies the sentiment and emotions in text data through detecting emotion-bearing words (Liu, 2010). For instance, the negative sentiment is considered to be expressed in the tweet "*Fear*, *retaliation* ruled @UserID HR department, employees say", because of the negative emotion-bearing words (i.e., fear and retaliation) were used; the positive sentiment is considered to be expressed in the tweet "*Thank* you for surprising one of our *amazing* teachers today!", because of the positive emotion-bearing words (i.e., thank and amazing) were used. Sentiment analysis can be applied at various levels (e.g., document, sentence, and aspect) using sentiment lexicons such as SentiWordNet (Feldman, 2013). With the wealth of digital data, sentiment analysis has emerged as a supplement to the existing methods (e.g., surveys and interviews) to gauge public opinion—an aggregate of individual views, attitudes, and beliefs about a particular topic expressed by the digital public.

The automated sentiment analysis has been applied to research in many fields to examine public opinion. In business, sentiment analysis has been used to evaluate the sentiment in financial news articles to predict stock prices (Schumaker, Zhang, Huang, & Chen, 2012). In political science, sentiment analysis was conducted to measure the public opinion on the 2008 Obama-McCain debate (Fernández-Gavilanes, Álvarez-López, Juncal-Martínez, Costa-Montenegro, & González-Castaño, 2016). In the field of public policy, sentiment analysis was conducted to gauge the public opinion on the U.S. immigration policy (Chung & Zeng, 2015) and crime problems (Jeremy, Paul, Krone, Spiranovic, & Cockburn, 2015). In the field of education policy, sentiment analysis was conducted to investigate the public opinion on the Common Core, and it was found that the negative sentiment surpassed the positive sentiment in all 50 states and the District of Columbia (Wang & Fikis, 2017).

If sentiment analysis is used to analyze social media data, researchers can often pinpoint the geographical locations of social media users by geographic identifiers, including geotagged locations and self-reported locations. Prior literature has consistently suggested that approximately 1% of tweets are geotagged, thus providing the data on latitude and longitude of the locations where the tweets are posted (e.g., Jahanbakhsh & Moon, 2014; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2012; Ram et al., 2015; Young, Rivers, & Lewis, 2014). In addition, around 70% of Twitter users self-report their geographic locations on their Twitter profiles (e.g., Mislove et al., 2012; Wang & Fikis, 2017). These data on locations, along with the data on the time stamps of social media posts, offer researchers opportunities to examine when and how public opinion emerge, fluctuate, and evolve at different geographical scales such as states, congressional and legislative districts, and large cities and towns. Moreover, when social media users post selfies like more than half of millennials have already done in the United States (Taylor, 2014; Wilson, 2014), the public opinion can be examined at an even more granular level by demographic groups such as African Americans, Hispanics, and Asian Americans. More intriguingly, public opinion is not a static entity. Rather, public opinion is contagious (Kramer et al., 2014), and they evolve throughout the policymaking and implementation process. As a corollary, sentiment analysis can be an instrumental tool in education policy research to glean insights into how education policy affect the public, and the interplays among education policy, public opinion, and policy outcomes in diverse political, economic, and cultural contexts.

Like other methodological approaches introduced earlier, sentiment analysis is a crude tool to detect sentiment in large-volume text datasets. Other than that, the automated sentiment analysis has a few unique limitations. First, the fully automated sentiment analysis does not detect sarcasm well. Sarcasm is a sophisticated expression, in which "praises" carries negative sentiment (Altrabsheh, Cocea, & Fallahkhair, 2015). Take the tweet "Brilliant! The real agenda behind #CommonCore!" as an example. The algorithms in sentiment analysis might deem "brilliant" as a positive-emotion-bearing word and categorize erroneously the tweet with positive sentiment. The second limitation of sentiment analysis is the lack of contextual understanding. Without it, the sentiment analysis algorithms might categorize the tweets "Testing is so green." and "Opt out is so White." as the tweets with the neutral sentiment, which might be in disagreement with manual coding results ("Testing is so green" might be manually coded as "testing industry wring profits from public education"; "Opt out is so white" might be manually coded as "The students and their parents who advocate for opting out of standardized testing are predominantly White"). The third limitation is the exclusion of the content on the webpage directed by the hyperlinks in tweets.

Consider the tweet "Must read. http://t.co/eUJqNEGZSm Is this where we are headed? Lots to think about." The text data in this tweet itself suggest neutral sentiment, but the content on the webpage directed by the hyperlinks in the tweet carries negative sentiment towards the proposed changes in the Elementary and Secondary Education Act (ESEA) Reauthorization. Given the limitations, sentiment analysis often needs some portion of manual coding to ensure the veracity of data analysis.

Debunking the Misconceptions

For all the richness of big data and emerging methodological tools at our disposal, if used properly, big data could be a gold mine of education policy research. However, big data is not a panacea. The label of "big" renders big data prone to misunderstanding. The fundamental premise to establish the veracity of big data—the fourth feature of big data—lies with not only aptly applying the aforementioned methodological tools and many others, but also exercising caution against the misconceptions of big data. The three misconceptions discussed below have been noted in the literature in public policy and computational social science. These misconceptions have the potential to misguide education policy researchers as they use big data. Here to ensure the veracity of big data in education policy research, I debunk the misconceptions as a cautionary note for future inquiry.

Bigger is Not Necessarily Better

The first misconception is that bigger is better. Not true! Big data, indeed, has much value to offer; yet it may delude us into thinking bigger is better. In fact, bigger is not necessarily better. Big data—the large-volume, wide-variety, and high-velocity data—inherently implies more noise in data, rendering more efforts to find the true patterns in data because "the signal-to-noise ratio may be waning" (Silver, 2015, p. 447). More importantly, big data and "small data" are not mutually exclusive. The techniques in traditional statistics are indispensable to examine the validity and reliability of big data analytics.

Take sampling as an example. Some argue that sampling may outlive its usefulness as we can now access samples so large that $N \approx$ all (i.e., samples are close to population; Mayer-Schönberger & Cukier, 2013). Surveying the literature in the big data era, it is not uncommon to see the studies that analyzed multi-million pieces of data. For instance, 46 billion words posted by 63 million unique Twitter users were considered as the social sensors of human happiness (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011); a corpus that contained 509 million tweets posted by 2.4 million Twitter users from 84 countries was examined to detect people's seasonal mood changes (Golder & Macy, 2011); 9 million tweets posted by the Twitter followers of candidates for the U.S. House and Senate and governorship in 2010 midterm elections were used to study the digital public's political expression (Bode et al., 2015). However, the "$N \approx$ all" view has been challenged by many skeptics who have called to caution the biased sampling, particularly in those studies relying on specific social networking sites (boyd & Crawford, 2012; Hargittai, 2015; Sandvig, 2015). In other words, the statement "we examined how 10.1 million U.S. Facebook users interact" does not indicate a representative sample of the U.S. population, but rather the 4% of Facebook users who self-reported their political affiliation on their Facebook profile page (Sandvig, 2015). In fact, people do not randomly opt into the use of Facebook and social networking sites in general (Hargittai, 2015). Whether people use a specific social networking site is associated with a host of factors, including age, gender, ethnicity, education, and income. Specifically, younger people are more likely than older people to use three popular social media sites—Facebook, Twitter, and LinkedIn; those with more education and higher income are more likely to use these three sites than those who have lower socioeconomic status; women are more likely to use Facebook but less likely to use LinkedIn;

African Americans are more likely to be on LinkedIn and Twitter, but less likely to use Facebook than Whites (Hargittai, 2015). As a corollary, the online population demographics might shift constantly, and do not represent the ones in the offline world (Diaz, Gamon, Hofman, Kıcıman, & Rothschild, 2016).

More importantly, in an article titled "How Big Data is Unfair", Hardt (2014) noted that the algorithms are apt to "favor those who belong to the statistically dominant groups" (para. 9), and "a variable that is positively correlated with the target in the general population might be negatively correlated with the target in a minority group" (para. 12). Therefore, despite a large *N*, the sheer number of sample size alone is not the factor substantiating the claim that $N \approx$ all. Researchers must be cautious when making generalized claims based on a biased sample, as it is problematic to generalize the patterns in one demographic group at one place at one time to all demographic groups at all places at all times.

Theoretical Framing Is Important As Always

The second potential misconception of using big data in education policy research is that theoretical framing may become obsolete. Some predicted that the scientific approach of hypothesizing, modeling, and testing would become obsolete, because patterns can be revealed by crunching data through algorithms without being guided by theories (Anderson, 2008). This view is unfounded, because the revealed "patterns" might not exist and the "enormous quantities of data can offer connections that radiate in all directions" (boyd & Crawford, 2012, p. 668). The so-called "patterns" are deemed as "chance associations"—the discernable associations that would inevitably show up if we look at enough data (Leinweber, 2007). Without appropriate theories to frame and guide data mining, it is easy to fall into the trap of "data mining sins" (Leinweber, 2007, p. 15). An example is that the Google Flu Trends (GFT) was touted that it could predict flu by matching 50 million Google search terms to over 1,000 terms suggesting the propensity of flu, and the GFT could generate the flu prediction two weeks before the U.S. Centers for Disease Control and Prevention did. However, the GFT failed to predict the non-seasonal flu in 2009, partly because the GFT "was part flu detector and part winter detector" (Lazer, Kennedy, King, & Vespignani, 2014, p. 1203). This is why boyd and Crawford (2012) argued that data, big or small, lose value when they are taken out of context. In education policy research, the socially-constructed data should never override theory. Theory matters. Context matters. These principles apply to big data analytics as well. Technology truly helps satiate our voracious appetite for data; computer science provides the capacity to collect and analyze voluminous amounts of data. But still, in the big data era, theories and data are inextricably linked. Data alone tell us nothing. Data are merely a vehicle we can lean on to enrich our understanding of social phenomena. When we simplify human behaviors to numbers, we run the risk of losing the richness of human behaviors. It is the researchers' job to interpret data: unpacking the meaning behind the data in a given context (boyd & Crawford, 2012). Therefore, the data interpretation, data contextualization, and sense-making process must be guided by theories, rather than algorithms. In education policy research, on the one hand, theories are analogous to beacons guiding the data collection, analysis, and interpretation, so that researchers do not cherry-pick variables to blindly hunt for patterns; on the other hand, the results from a theoretical-guided analytical approach can help facilitate theory development and enrich our understanding of intricate education policy issues.

Even Automation Needs Human Involvement

Given the large-volume, wide-variety, and high-velocity features of big data, much of the big data analytics is automated by algorithms. The third potential misconception of using big data in education policy research is that automation does not need human involvement. This misconception runs the risk of overemphasizing algorithm-enabled automation at the expense of the veracity of big data. The sophisticated algorithms, along with the brute force of high-performance computing (also called supercomputer), do not mean that valid results are automatically produced after feeding data into the magic wand of algorithms. First, human judgment is needed in every step of big data analytics in education policy research to establish the veracity of research. The data quality needs to be evaluated by humans, as the data feeding into the algorithms can be unrepresentative of targeted population and error-prone (Tufekci, 2014). The model parameters used in the automated data analysis need to be set by humans, as the ones noted earlier in the section of topic models. The data interpretation entails researchers to draw upon their prior knowledge in the field to make sense of the data. As a corollary, to tap the potential of big data in the field of education policy, the key is not to blindly pursue clever algorithms or troves of data, but to augment automation and human judgment, so that they are complementary to each other.

Challenges

Amid the abundant potential of big data, pressing challenges abound in using big data in education policy research. Among them are developing interdisciplinary research capacity, and addressing the privacy concerns and ethical conundrums. Here I discuss these two challenges that may deter using big data in education policy research. More importantly, I encourage future researchers to offer viable strategies to unleash the potential big data in education policy research.

Developing Interdisciplinary Research Capacity

The first challenge we face is the deficiency of education policy researchers who are well versed in the fields of both education policy and computer science. In its essence, using big data in education policy research entails interdisciplinary efforts, in which the knowledge in education policy provides vital theoretical underpinnings that frame the studies, and the expertise in computer science enables the algorithmic approaches to data acquisition, mining, and analysis to scale up analytical capacity. Unfortunately, education policy researchers are often inadequately trained in computer science.

To surmount this challenge, I propose two viable solutions to build and maximize the research capacity in using big data in education policy research. The first solution is to develop interdisciplinary research teams. Stepping out the silos in individual researcher's disciplinary boundary, the researchers in different fields—such as education policy, computer science, and data science—collaborate and marshal the intellectual capital on data acquisition and analysis, as well as interpreting data in rich social contexts. Another solution to the deficiency of interdisciplinary research capacity is to build a pipeline of ambidextrous researchers who are deft in both education policy research and computer science. In fact, government agencies have already taken the initiatives to train aspiring socially-minded computer scientists and computationally-minded social scientists. For instance, National Science Foundation has been funding the initiatives to train interdisciplinary big data researchers (NSF, 2012). Moreover, many universities have begun building interdisciplinary programs that bring together social science, computer science, statistics, and data visualization (Wallach, 2015). It is thus recommended that future education policy researchers participate in such interdisciplinary training programs so that they will be well equipped to apply computational social science to education policy research.

Addressing Privacy Concerns and Ethical Conundrums

In a data-rich environment, researchers have been wrestling with privacy concerns that derive from the inherent tension between data and privacy. The word "data" is a plural of the Latin word "datum" which means "something given", whereas privacy refers to "the ability to control the dissemination and use of one's information" (Wright et al., 2003, p. 148). As a result, big data and privacy have inherently contradictory goals (O'Leary, 2015). With Internet-connected digital devices at disposal, people might trade privacy for convenience. We disclose our whereabouts when we hail a ride with Uber or use Google Map to get to our destination. We share a part of our life when we use Facebook to stay in touch with families and friends. Further complicating the privacy issue is that once we share something online, either a picture or a comment, that "something" then has a life of its own on the Internet. That is, we no longer have much control over how the data about us are disseminated. In this digitally hyper-connected world, everything about our online activities, as boyd (2010) aptly put it, is "public by default, private when necessary" (boyd, 2010, para. 1). Still, for our researchers, as computerized algorithms allow us to scale up data collection and analysis through an automated process, we are now presented with the mounting challenges on how to strike a balance between strengthening the oversight of data access and advancing scientific discoveries.

Further, ethical conundrums loom large as we capitalize on big data in education policy research. First, readily access to data does not necessarily mean that the data are meant to be consumed by anyone (boyd, 2012). Rather, the easy access to big data highlights the need for greater oversight to deter prying eyes on personal data and prevent the potential abuse of massive datasets. A recent controversial study was about the massive-scale experiment (n = 689,003) on emotional contagion on Facebook (Kramer et al., 2014). To test whether emotions were contagious through Facebook's social networks without face-to-face communication or non-verbal cues, Kramer et al. manipulated the extent of the emotional content of Facebook users' News Feed. Kramer et al. stated that the data collection and analysis process was "consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this [Kramer et al.'s] research" (Kramer et al., 2014, p. 8789). This leads to a critical question: Are social media sites' Terms of Service the de facto informed consent for social experiments in the digital age? To date, this question has not received an agreed-upon answer yet. Most Institutional Review Board Protocols have not provided adequate guidelines on conducting large-scale social experiments on social media sites. This is particularly problematic when human participants are not fully informed and are not explicitly asked for informed consent for online social experiment participations.

Another ethical challenge centers on the use of social bots in the experiments. Social bots are the social media accounts that are operated by algorithms, instead of humans, to automate the content generation and interaction with other human users on social media (Ferrara, Varol, Davis, Menczer, & Flammini, 2016; Kollanyi, Howard, & Wooley, 2016). Recently, social bots have been increasingly used to contribute to and even steer the direction of online discourse on political elections and public policy issues (Bessi & Ferrara, 2016; Ferrara et al., 2016). A prime example is that the automated pro-Trump bots were used more aggressively than pro-Clinton bots during the U.S. presidential debates in 2016, and particularly in the final presidential debate the pro-Trump bots out-produced seven times more traffic on Twitter than the pro-Clinton bots (Kollanyi, Howard, & Wooley, 2016). Thus, social bots can potentially have a bearing on shaping and even intentionally manipulating the contours of online discourse, wielding subtle and even significant influence on public opinion and policy issues. In the research realm, the social bots have been used as the "virtual confederates" in social experiments conducted online to study human social behaviors, like the confederates—the people trained by the researchers to follow the pre-assigned scripts in social experiments—used in Stanley Milgram's experiment as the pedestrians walking on the street and looking at the balcony to examine how many other pedestrians followed the confederates' behavior (Milgram, Bickman, & Berkowitz, 1969), and in Solomon Asch's (1956) conformity experiment. As a result, the ethical challenges on the informed consent, deceptions, and direct harm must be addressed as the bots are used as the "virtual confederates" in the experiments to control the variables and create an artificial context (Krafft, Macy, & Pentland, 2016).

Concluding Remarks

Big data research has been growing by leaps and bounds. In education policy, researchers can collect big data from an array of sources, including, but not limited to, the data-driven education accountability systems, the digital footprints left by the digital services and devices used by education stakeholders, and data acquisition technologies such as the OCR converting the un-digitalized text data into digitalized data and API accessing data shared by technology companies. With the rich data, new models are rapidly being developed to analyze texts, videos, images, and audios, mobile phone call log data, and satellite image data. The wealth of data and analytical tools allow researchers to examine education policy research questions that might not be easily explored in the past. The three methodological approaches (topic models, network text analysis, and sentiment analysis) introduced in this commentary hold great potential to the real-time or near-real-time analysis of big data in education policy. In doing so, the timely results can be offered to inform education policymaking, implementation, and evaluation. For education policy researchers to venture into big data, it is important to develop interdisciplinary research capacity, as well as address the privacy concerns and ethical conundrums. This paper did not take a systematic approach to comprehensively surveying all methodological tools used in analyzing big data. Rather, to ensure the empirical examples presented in this paper are applicable in education policy research, this commentary introduces three analytical approaches used in the closely related fields of public policy and computational social science in an effort to invite education policy researchers to venture into big data. As we embrace big data in education policy research, many hurdles remain and new obstacles will emerge. It is the hope that this commentary could invite forward-looking discussions and the explorations of a research agenda of using big data in education policy.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S., & Liu, B. (2011). Predicting flu trends using Twitter data. *The IEEE Conference on Computer Communications Workshops* (pp. 702-702). Piscataway, NJ: IEEE. https://doi.org/10.1109/infcomw.2011.5928903

Altrabsheh, N., Cocea, M., & Fallahkhair. S. (2015). Detecting Sarcasm from Students' Feedback in Twitter. In G. Conole, T. Klobučar, C. Rensing, J. Konert, & É. Lavoué (Eds.), *Design for teaching and learning in a networked world* (pp. 551-555). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-24258-3_57

Alvarez, R. M. (Ed.). (2016). *Computational social science: Discovery and prediction.* New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781316257340

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine, 16*(7), 108–109.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1-70. https://doi.org/10.1037/h0093718

Baker, R., & Kalina Y. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-16.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis, 23*(1), 76-91. https://doi.org/10.1093/pan/mpu011

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123-1131. https://doi.org/10.1377/hlthaff.2014.0041

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2015). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45-59. https://doi.org/10.1016/j.inffus.2015.08.005

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday, 21*(11). https://doi.org/10.5210/fm.v21i11.7090

Blei, D. M. (2012). Probabilistic topic models. *Communication of the ACM, 55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty. J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning.* Pittsburgh, Pennsylvania, USA, June 25-29. https://doi.org/10.1145/1143844.1143859

Blei, D. M., & Lafferty. J. D. (2007). A correlated topic model of *Science*. *The Annals of Applied Statistics, 1*(1), 17-35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Bode, L., Hanna, A., Yang, J., & Shah, D. V. (2015). Candidate networks, citizen clusters, and political expression: strategic hashtag use in the 2010 midterms. *The Annals of the American Academy of Political and Social Science, 659*(1), 149-165. https://doi.org/10.1177/0002716214563923

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*, 295-298. https://doi.org/10.1038/nature11421

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks.* Thousand Oaks, CA: Sage Publications.

Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change, 36*, 89-100. https://doi.org/10.1016/j.gloenvcha.2015.12.001

boyd, d. (2010, January 25). Public by default, private when necessary [Web log post]. Retrieved from http://dmlcentral.net/public-by-default-private-when-necessary/

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Burstein, P. (2003). The impact of public opinion on public policy: A review and an agenda. *Political Research Quarterly, 56*(1), 29–40. https://doi.org/10.1177/106591290305600103

Chen, X., Cho, Y., & Jang, S. (2015, April 24). Crime prediction using twitter sentiment and weather. *Paper presented at Systems and Information Engineering Design Symposium*. Charlottesville, VA. https://doi.org/10.1109/sieds.2015.7117012

Chung, W., & Zeng. D. (2015). Social-media-based public policy informatics: Sentiment and network analyses of U.S. immigration and border security. *Journal of the Association for Information Science and Technology, 67*(7), 1588-1606. https://doi.org/10.1002/asi.23449

Cobb, W. N. W. (2015). Trending now: Using big data to examine public opinion of space policy. *Space Policy, 32*, 11-16. https://doi.org/10.1016/j.spacepol.2015.02.008

Diaz, F., Gamon, M., Hofman, J. M., Kıcıman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLoS ONE, 11*(1), e0145406. https://doi.org/10.1371/journal.pone.0145406

Diesner, J., & Carley, K. (2005). Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. In V. K. Narayanan, & D. J. Armstrong (Eds.), *Causal mapping for information systems and technology research: Approaches, advances, and illustrations* (pp. 81-108). Hershey, PA: Idea Group Publishing. https://doi.org/10.4018/978-1-59140-396-8.ch004

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE, 6*(2): e26752. https://doi.org/10.1371/journal.pone.0026752

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science, 346* (6210), 715-721. https://doi.org/10.1126/science.1243089

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82-89. https://doi.org/10.1145/2436256.2436274

Feng, Y., & Lapata, M. (2010). Topic models for image annotation and text illustration. *Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, CA, June 1-6. Retrieved from http://www.aclweb.org/anthology/N10-1125

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016). Unsupervised Method for sentiment analysis in online texts. *Expert Systems with Applications, 58*, 57-75. https://doi.org/10.1016/j.eswa.2016.03.031

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96-104. https://doi.org/10.1145/2818717

Fleuren, W. W. M., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods, 74*, 97-106. https://doi.org/10.1016/j.ymeth.2015.01.015

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333*(6051), 1878-1881. https://doi.org/10.1126/science.1202775

Goldstone, A, & Underwood. T. (2014). The quiet transformation of literary studies: What thirteen thousand scholars could tell us. *New Literacy History, 45*(3), 359-384. https://doi.org/10.1353/nlh.2014.0025

Gormley Jr., W. T. (2016). From science to policy in early childhood education. *Science, 333*, 978-981. https://doi.org/10.1126/science.1206150

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis, 18*(1), 1-35. https://doi.org/10.1093/pan/mpp034

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297. https://doi.org/10.1093/pan/mps028

Grün, B., & Hornik, K. (2011). Topicmodels: An *R* package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30. https://doi.org/10.18637/jss.v040.i13

Gutierrez, C., Figuerias, P., Oliveira, P., Costa, R., & Jardim-Goncalves, R. (2015). Twitter mining for traffic events detection. *Paper presented at the Science and Information Conference*, London, July 28-30. https://doi.org/10.1109/sai.2015.7237170

Hargittai, E. (2015). Is bigger always better? potential biases of big data derived from social network sites. *The Annals of the American Academy of Political and Social Science, 659*(1), 63-76. https://doi.org/10.1177/0002716215570866

Hardt, M. (2014, September 26). How big data is unfair. Medium. Retrieved from https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Hearst, M.A. (1999). Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3-10). Association for Computational Linguistics. https://doi.org/10.3115/1034678.1034679

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review, 102*(2), 994-1028. https://doi.org/10.1257/aer.102.2.994

Henderson, M. B., Peterson, P. E., & West, M. R. (2016). The 2015 Ednext Poll on school reform. *Education Next, 16*(1), 8–20.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems, 12*(12), 767–797.

Lazer, D., Kennedy, R., King, K., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*, 1203-1205. https://doi.org/10.1126/science.1248506

Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education, 121*(1), 1–27. https://doi.org/10.1086/678136

Jahanbakhsh, K., & Moon, Y. (2014). The predictive power of social media: On the predictability of U.S. presidential election using Twitter. arXiv:1407.0622. Retrieved from http://arxiv.org/pdf/1407.0622v1.pdf

Jeremy, P., Paul, W., Krone, T., Spiranovic, C., & Cockburn, H. (2015). Social media sentiment analysis: A new empirical tool for assessing public opinion on crime? *Current Issues in Criminal Justice, 27*(2), 217-236.

Katsurai, M., Ogawa, T., & Haseyama, M. (2012). A cross-modal approach for extracting semantic relationships of concepts from an image database. *Paper presented at the IEEE International Conference on Acoustics Speech and Signal Processing*, Kyoto, Japan, March 25-30. https://doi.org/10.1109/icassp.2012.6288392

King, D., Ramirez-Cano, D., Greaves, F., Vlaev, I., Beales, S., & Darzi, A. (2013). Twitter and the health reforms in the English National Health Service. *Health Policy, 110*(2-3), 291-297. https://doi.org/10.1016/j.healthpol.2013.02.005

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science, 331*, 719-721. https://doi.org/10.1126/science.1197872

King, G. (2016). Preface: Big data is not about the data! In R. M. Alvarez (Ed), *Computational Social Science: Discovery and Prediction* (pp. vii-x). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.001

Kollanyi, B., Howard, P. N., & Wooley, S. C. (2016). Bots and automation over Twitter during the third U.S. presidential debate. *Project on Algorithms, Computational Propaganda, and Digital Politics*. Retrieved from http://politicalbots.org/wp-content/uploads/2016/10/Data-Memo-Third-Presidential-Debate.pdf

Krafft, P. M., Macy, M., & Pentland, A. (2016). *Bots as virtual confederates: Design and ethics*. Retrieved from https://arxiv.org/pdf/1611.00447.pdf

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788-8790. https://doi.org/10.1073/pnas.1320040111

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science, 58*(3), 754-771. https://doi.org/10.1111/ajps.12085

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., … Alstyne, M. V. (2009). Life in the network: The coming age of computational social science. *Science, 323*(5915), 721-723. https://doi.org/10.1126/science.1167742

Leinweber, D. J. (2007). Stupid data miner tricks: Overfitting the S&P 500. *Journal of Investing 16*(1), 15–22. https://doi.org/10.3905/joi.2007.681820

Leydesdorff, L., & Hellsten, I. (2005). Metaphors and Diaphors in science communication: Mapping the case of stem-cell research. *Science Communication, 27*(1), 64-99. https://doi.org/10.1177/1075547005278346

Leydesdorff, L., & Nerghes, A. (in press). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (n < 1000). *Journal of the Association for Information Science and Technology*. Retrieved from http://www.mathpubs.com/detail/1511.03020v2/Co-word-Maps-and-Topic-Modeling-A-Comparison-Using-Small-and-Medium-Sized-Corpora-n-1000

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (pp. 627-666). UK. Chapman and Hall/CRC.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. London, UK: Eamon Dolan/Mariner Books.

Milgram, S., Bickman, L., & Berkowitz, L. (1969). Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology, 13*(2), 79-82. https://doi.org/10.1037/h0028070

Miller, W. R., Groves, D., Knopf, A., Otte, J. L., & Silverman, R. D. (2017). Word adjacency graph modeling: Separating signal from noise in big data. *Western Journal of Nursing Research, 39*(1), 166-185. https://doi.org/10.1177/0193945916670363

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., & Rosenquist. J. N. (2012). Understanding the demographics of Twitter users. *Association for the Advancement of Artificial Intelligence*. Retrieved from http://www.ccs.neu.edu/home/amislove/publications/Twitter-ICWSM.pdf

National Science Foundation. (2012). *Core techniques and technologies for advancing big data science & engineering*. Retrieved from http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf12499

National Research Council. (2013). *Frontiers in massive data analysis*. Washington, D. C: National Academies Press.

O'Leary, D. E. (2015). Big data and privacy: Emerging issues. *IEEE Intelligent System, 30*(6), 92-96. https://doi.org/10.1109/MIS.2015.110

Page, B. I., & Shapiro, R. Y. (1983). Effects of public opinion on policy. *American Political Science Review, 77*(1), 175–190. https://doi.org/10.2307/1956018

Peterson, P. E., Henderson, M. B., West, M. R., & Barrows, S. (2017). Ten-year trends in public opinion from the EdNext Poll. *Education Next, 17*(1), 8-28.

Pinkovskiy, M., & Sala-i-Martin, X. (2015). Lights, camera, … income! Estimating poverty using national accounts, survey means, and lights. *Federal Reserve Bank of New York Staff Reports*, no. 669. Retrieved from https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr669.pdf

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science, 54*(1), 209-228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Ram, S., Zhang, W., Williams, M., & Pengetnze, Y. (2015). Predicting asthma-related emergency department visits using big data. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1216-1223. https://doi.org/10.1109/JBHI.2015.2404829

Reardon, S., Kalogrides, D., & Shores, K. (2016). The geography of racial/ethnic test score gaps. Unpublished manuscript.

Reddicka, C. G., Chatfieldb, A. T., & Jaramilloa, P. A. (2015). Public opinion on national security agency surveillance programs: A multi-method approach. *Government Information Quarterly, 32*, 129-141. https://doi.org/10.1016/j.giq.2015.01.003

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. In R. M. Alvarez (Ed.), *Computational social science: discovery and prediction* (pp. 51-97). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.004

Sandvig, C. (2015). The Facebook "It's not our fault" study. [web log post]. Retrieved from http://socialmediacollective.org/2015/05/07/the-facebook-its-not-our-fault-study/?blogsub=subscribed#blog_subscription-2

Schumaker, R. P., Zhang, Y., Huang, C., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464. https://doi.org/10.1016/j.dss.2012.03.001

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media. *The Annals of the American Academy of Political and Social Science, 659*(1), 78-94. https://doi.org/10.1177/0002716215569197

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The Annals of the American Academy of Political and Social Science, 659*(1), 6-13. https://doi.org/10.1177/0002716215572084

Silver, N. (2015). *The signal and the noise: Why so many predictions fail—but some don't.* New York, NY: Penguin Books.

Sudhahar, S., Veltri, G. A., & Cristianini, N. (2015). Automated analysis of the US presidential elections using big data and network analysis. *Big Data and Society, 2*(1), 1-28. https://doi.org/10.1177/2053951715572916

Supovitz, J. A., & Reinkordt, E. (2017). Keep your eye on the metaphor: The framing of the Common Core on Twitter. *Education Policy Analysis Archives, 25*(31), 1-29. http://dx.doi.org/10.14507/epaa.25.2285

Taylor, P. (2014). More than half of millennials have shared a 'selfie'." *Pew Research Center.* Retrieved from http://www.pewresearch.org/fact-tank/2014/03/04/more-than-half-of-millennials-have-shared-a-selfie/

Toole, J. L., Lin, Y., Muehlegger, E., Shoag, D., González, M. C., & Lazer, D. (2015). Tracking employment shocks using mobile phone data. *Journal of the Royal Society: Interface, 12*(107), 20150185. https://doi.org/10.1098/rsif.2015.0185

Tufekci, Z. (2014, May). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Paper presented at the 8th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*. Washington D. C.

Valant, J., & Newark, D. A. (2016). The politics of achievement gaps: U.S. public opinion on race-based and wealth-based differences in test scores. *Educational Researcher, 45*(6), 311-346. https://doi.org/10.3102/0013189X16658447

Verd, J. M., & Lozares, C. (2014). Reconstructing social networks through text analysis: From text networks to narrative actor networks. In S. Domínguez & B. Hollstein (Eds.), *Mixed methods social networks research: Design and applications* (pp. 269-304). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781139227193.014

Wallach, H. (2015). Computational social science: Toward a collaborative future. In R. M. Alvarez (Ed.), *Data science for politics, policy, and government* (pp. 1-11). Cambridge University Press. Retrieved from http://dirichlet.net/pdf/wallach15computational.pdf

Wang, T., Huang, Z., & Gan, C. (2016). On mining latent topics from healthcare chat logs. *Journal of Biomedical Informatics, 61*, 247-259. https://doi.org/10.1016/j.jbi.2016.04.008

Wang, Y. (2016). U.S. state education agencies' use of Twitter: Mission accomplished? *Sage Open, 6*(1), 1-12. https://doi.org/10.1177/2158244015626492

Wang, Y. (2017). The social networks and paradoxes of the opt-out movement amid the Common Core State Standards implementation. *Education Policy Analysis Archives, 25*(34), 1-27. https://doi.org/10.14507/epaa.25.2757

Wang, Y., Bowers, A., & Fikis, D. (2017). Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of *EAQ* articles from 1965 to 2014. *Educational Administration Quarterly*, *53*(2), 289-323. https://doi.org/10.1177/0013161X16660585

Wang, Y., Bowers, A., Chae, H. S., Fikis, D., & Natriello, G. (2017). Illustrating the epistemological identity of education research: Probabilistic topic modeling of 116 years of literature of *Teachers College Record* from 1900 to 2015. *Paper accepted for presentation at the 2017 annual meeting of American Educational Research Association*. San Antonio, TX.

Wang, Y., & Fikis, D. (2017). Common Core Standards on Twitter: Public sentiment and opinion leaders. *Educational Policy*, Online first. https://doi.org/10.1177/0895904817723739

Wang, Y., & Fikis, D. (2016, November). *Text mining social media data on the Common Core State Standards: Topic modeling and hashtag co-concurrence network analysis.* Paper presented at the 2016 Annual Convention of University Council for Educational Administration. Detroit, MI.

Watts, D. J. (2013). Computational social science: Exciting progress and future directions. *The Bridge Linking Engineering and Society, 43*(4), 5-11.

Wilson, C. (2014). The selfiest cities in the world: TIME's definitive ranking. *TIME*. Retrieved from http://time.com/selfies-cities-world-rankings/

Wright, R. N., Camp, L. J., Goldberg, I., Rivest, R., & Wood, G. (2003). Privacy tradeoffs: Myth or reality. *Financial Cryptography: Lecture Notes in Computer Science, 2357*, 147-151. https://doi.org/10.1007/3-540-36504-4_11

Xie, M., Jean, N., Burke, M., Lobell, D., & Ermon, S. (2016). Transfer learning from deep features for remote sensing and poverty mapping. *Paper presented at the 30th AAAI Conference on Artificial Intelligence*. Phoenix, AZ.

Young, S. D., Rivers, C., & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine, 63*, 112-115. https://doi.org/10.1016/j.ypmed.2014.01.024

About the Author

Yinying Wang
Georgia State University
ywang103@gsu.edu
http://orcid.org/0000-0002-9005-9641

Yinying Wang is an assistant professor of educational leadership in the Department of Educational Policy Studies at College of Education and Human Development, Georgia State University. Her research interests include social network analysis in educational leadership and policy, social media in education policy making and organizational communication, and educational technology leadership.

education policy analysis archives

Volume 25 Number 94 August 21, 2017 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.