Public Internet Data Mining Methods in Instructional Design, Educational Technology, and Online Learning Research

Royce Kimmons & George Veletsianos

Kimmons, R., & Veletsianos, G. (2018). Public Internet Data Mining Methods in Instructional Design, Educational Technology, and Online Learning Research. *Tech Trends, 62*(5), 492-500.

Abstract

We describe the benefits and challenges of engaging in public data mining methods and situate our discussion in the context of studies that we have conducted. Practical, methodological, and scholarly benefits include the ability to access large amounts of data, randomize data, conduct both quantitative and qualitative analyses, connect educational issues with broader issues of concern, identify subgroups/subpopulations of interest, and to avoid many biases. Technical, methodological, professional, and ethical issues that arise by engaging in public data mining methods include the need for multifaceted expertise and rigor, focused research questions and determining meaning, and performative and contextual considerations of public data. As the scientific complexity facing research in instructional design, educational technology, and online learning is expanding, it is necessary to prepare students and scholars in our field to engage with emerging research methodologies.

Public Internet Data Mining Methods in Instructional Design, Educational Technology, and Online Learning Research

Data mining of the public internet has been an emerging research method for the past two decades as it has been applied to a variety of fields to help solve persistent problems like developing webpage recommender systems (Niwa, Doi, & Honiden, 2006), combating infectious diseases (Brownstein, Freifield, Reis, & Mandl, 2008), identifying cybersecurity threats (Maloof, 2016), improving network traffic (Wang, Madhyastha, Chan, Papadimitriou, & Faloutsos, 2002), and predicting political orientations (Colleoni, Rozza, & Arvidsson, 2014), just to name a few. Previous work has pointed out some of the technical opportunities and challenges of such methods (Andersen & Feamster, 2006), but public internet data mining has not yet been widely applied to addressing issues facing the field of instructional design and technology (IDT), and we do not fully understand the benefits and challenges of its application to our field. Furthermore, though some data mining methods are eagerly being applied in the realms of learning analytics and data dashboard visualization (Baker & Inventado, 2014), we have not as a field begun exploring the potentials and ramifications of using massive amounts of disorganized, publicly-available data to address persistent IDT challenges or determining how we must train new professionals to make use of the wealth of data available to them via the public internet. Data mining of the public Internet affords IDT researchers the ability to answer important questions that they have henceforth been either unable to answer or unable to explore using non-invasive methods on a large scale. To illustrate, in table 1 we provide a list of potential questions that are of interest to IDT that researchers may be able to address using data mining methods.

Table 1. A selection of typical IDT research questions that may be answered via data mining methods

Research Question	Public Internet data source
What sorts of IDT skills do employers require?	Job ad postings
What challenges do teachers face in integrating technology in K-12 classrooms?	Public discussion forums
What kinds of peer-support do online learners	Discussion forums found in public online courses

provide to one another?	
In what ways are particular web-based technologies used in k-12 courses?	Blog networks, wiki networks, etc
How do instructional designers describe the field to others?	Personal portfolios, discussion forums
What motivates individuals to contribute to informal learning communities?	Discussion forums
What sentiments does the public express toward particular educational and technological innovations (e.g., MOOCs, artificial intelligence, online education, adaptive learning, etc)?	Discussion forums, newspaper comments
What is the relationship between demographic variables (e.g., gender) and achievement in STEM courses?	Secondary data made available in public repositories

Over the last two years, we (the two authors of this paper) have conducted more than 10 studies using public data mining methods in IDT. These studies included extracting and analyzing publicly-available data from Websites (e.g., K-12 websites), social media (e.g., Twitter), and discussion fora (e.g., YouTube comments). They generated massive datasets and allowed us to conduct research pertaining to technology use, social media prevalence, equity, and civility in online discussions. In this paper, we will describe the benefits and challenges we encountered while engaging in public data mining and situate our discussion in the context of studies that we have conducted in order to present authentic examples of the ways that public data mining can be used in our field.

As we have put processes in place to collect and analyze public social media data, we have reached out to colleagues and secured funding for graduate students at other universities to conduct collaborative work with us. To date, we have collaborated with 17 scholars on these projects representing 10 universities in the U.S. and Canada, and our collaborators have included undergraduate, master’s, and doctoral students as well as tenure-track faculty. These efforts have allowed us to take on the role of mentors in public data mining methods to our colleagues, to expand the horizons of our own research, and to train young researchers in these emerging methods. By doing so, we have identified a curricular need facing our field that we will also discuss here. While the practice of IDT traditionally involves multidisciplinary collaboration (e.g., instructional designers, subject matter experts, assessment experts, and faculty may collaborate to create an educational intervention), the scientific complexity facing IDT research and practice is increasingly expanding. For instance, the infusion of technology in all aspects of education has provided access to a deluge of digital data that was previously unfathomable (Selwyn, 2015), and instructional designers may nowadays collaborate with even more actors, such as data scientists and learning analytics researchers. Thus, it is necessary for researchers in our field to explore and understand emerging research methodologies. This paper will conclude by arguing that doctoral preparation programs in our field should include interdisciplinary methodological training for IDT researchers as a core component.

Some Benefits of Public Internet Data Mining

As interest in data mining takes hold in many industries, from healthcare to e-commerce, education researchers have started exploring the ways that both large and public datasets can contribute to making sense of issues facing educational practice and the science of learning. While substantial literature exists on the use of learning analytics in education (e.g., in Massive Open Online Learning [MOOC] contexts), much less is written about the use of public online data. The benefits or opportunities that mining of public Internet data engenders are numerous. These opportunities are practical and methodological, as well as scholarly. We organize these in the following themes:

- providing large amounts of data and allowing easy randomization;
- empowering both quantitative and qualitative analyses;
- connecting educational issues with larger public issues;
- enabling identification of subgroups/subpopulations for further research;
- and avoiding many biases.

Providing Large Amounts of Data

The data generated by contemporary Internet platforms, and made available to researchers through various means, are unprecedented. For instance, the data associated with posting one single tweet includes information about the person posting the tweet (e.g., username, name, biographic information, location, account creation date, and various statistics associated with the account holder such as total tweets posted and total followers), data associated with the actual tweet (e.g., the text of the tweet, the hashtags included in the text of the tweet, the time it was posted, the location associated with the device it was posted from, the application used to post the tweet, and various metrics associated with it such as number of times this particular tweet was retweeted), and similar data for any other accounts interacting with that particular tweet. In other words, a single tweet is associated with copious data points that IDT researchers have rarely seen. This data deluge present in Twitter is typical of online platforms. A similar situation exists with a variety of platforms that are used for teaching, training, and learning purposes (e.g., blogs, YouTube, Reddit, public websites, etc). To illustrate the magnitude of the data available, in a recent paper we sought to investigate time patterns in social media use (Veletsianos, Kimmons, Belikov, & Johnson, under review) and were able to identify a sample of academics on Twitter (n = 3,996) and retrieve more than 9 million tweets they posted along with associated metadata, yielding more than 100 million raw data points.

Good data enable one to answer the research questions he/she poses. While abundant data are not synonymous with *good* data, large amounts of data provide a number of opportunities for IDT researchers. Large-scale data allow researchers to examine whether the results generated by smaller-scale studies (e.g., case studies) hold up to scrutiny, investigate questions that can only be answered by larger datasets (e.g., investigations of populations vis-a-vis samples), and enable investigations of samples drawn at random from large populations.

Empowering Both Quantitative and Qualitative Analyses

Though data mining is often associated with analyses involving quantitative data, mining the public internet enables researchers to collect and analyze both quantitative and qualitative data. This method, therefore, accommodates a diverse range of research questions, data analysis methods, and approaches. In other words, as part of the IDT researcher’s methodological toolkit, data mining methods may enable the collection and analyses of different kinds of data in relation to the research questions being asked. Such versatility is important because it enables IDT researchers to use data mining methods across research paradigms, enabling the use of qualitative data to generate detailed and rich descriptions of phenomena, as well as the use of quantitative data to draw generalizable conclusions. For example, in investigating ways to scaffold student learning when interacting with a chatbot, data mining methods may enable IDT researchers to (a) code student prompts in order to develop a taxonomy of help-seeking questions, and (b) compute the frequency with which students ask different types of questions.

To illustrate, we were interested in examining the ways higher education institutions used social media for educational purposes with students and the broader public (Kimmons, Veletsianos, & Woodward, 2017; Veletsianos, Kimmons, Shaw, Pasquini, & Woodward, 2017). In order to explore this topic, we gathered quantitative data (e.g., number of tweets posted) and qualitative data (e.g., individual tweets and images) associated with the Twitter accounts of Canadian and US universities. We computed new variables using these data (e.g., number of replies, replies as a proportion of all tweets, number of tweets that included audiovisual elements) and also conducted descriptive, inferential, and qualitative analyses on them. Using this dataset, quantitative analyses enabled us to identify that higher education institutions in both countries mostly used Twitter to broadcast information rather than engage in dialogue. Qualitative analysis of a sample of tweets enabled us to discover that those broadcasted messages portrayed an overwhelmingly positive picture of institutional life. In other words, quantitative analyses enabled us to discover the frequency and type of Twitter use, while qualitative analyses allowed us to describe what such participation looked like. Data mining enabled us to develop a multi-layered understanding of institutional social media use, highlighting a finding that is core to IDT, namely that technologies are rarely neutral in their use (e.g., Twitter prompts users to broadcast messages) and that they can be appropriated to serve different needs (e.g., Twitter seemed to be used for promotion rather than educative purposes).

Connecting Educational Issues with Larger Public Issues

One of the pressing challenges facing our field is in pursuing an understanding of sociocultural and public issues pertaining to education, teaching, learning, scholarship, and technology (Veletsianos &

Kimmons, 2012). Such issues may involve access, equity, civility, socioeconomic divides, and sociotechnical issues (e.g., the impact of social media algorithms on opportunities for informal learning). While some of the field's research examines issues of broader concern, by and large the focus is on pedagogical applications of technology, with little attention being paid to the social, cultural, and political aspects and implications of instructional design and educational technology use. We need to pay close attention to these issues because of their societal significance and implications for practice. What is the public concerned about with regards to teaching and learning? In what ways can IDT re-imagine teaching and learning on a massive scale? In what ways are racism and sexism evident in our designs and educational offerings, and what does the field need to do in order to alleviate these problems? We believe that these types of questions (amongst many others) should be central to the field for they aim toward developing a more just and fair society. Public Internet data mining methods may provide opportunities for researchers to examine societal issues of broad concern, and enable the field to take a more active role in societal conversations of interest. For instance, in the same way that Rowe (2015) examined (in)civility in online political discussions occurring on the Washington Post Facebook account, IDT researchers might use data mining methods to investigate (in)civility on public platforms hosting educational interactions such as *CrashCourse* and *Physics Girl* on YouTube and develop ways to address this problem.

To illustrate how IDT research can be connected to issues of broader concern via data mining, consider the research we reported in Authors (2018). In that study, we sought to connect the educational uses of YouTube to gender issues. While typical IDT research might examine the pedagogical implications, opportunities, promises, drawbacks, and affordances of video-sharing technologies, we were interested in the sentiment that individuals faced when they asked to go online to share their research or to post their course assignments. We were also interested in examining whether different people faced different sentiment. By examining the sentiment expressed in response to TEDx and TED-Ed talks posted on YouTube we found that videos of male presenters showed greater neutrality, while videos of female presenters saw significantly greater polarity in replies. Such findings have significant implications for our field, because they question the oft-repeated optimistic narratives of contemporary technologies as necessarily positive for all people.

Enabling Identification of Subpopulations for Further Research

Due to the massive amounts of data available online, public Internet data mining methods enable researchers to identify particular subpopulations for further inquiry. Granular approaches to identifying participants are important, because they enable researchers to focus on typical, unique, or otherwise significant subpopulations of interest. For instance, considering Twitter as a platform of interest, data mining methods enable researchers to identify and study IDT issues pertaining to professors who tweet frequently (e.g., Kimmons & Veletsianos, 2016), educators who engage with a particular topic or affinity space (e.g., Paskevicius, Veletsianos, & Kimmons, in press), community members who comment on educational content (e.g., Veletsianos, Kimmons, Larsen, Dousay, & Lowenthal, in press), doctoral students who have a large number of followers, teachers who reside in a particular geographic area, faculty members who mention their teaching evaluations, undergraduate engineering students who tweet about positive/negative learning experiences, or IDT faculty who attend both IDT and Learning Sciences conferences. Further, the identification of specific subpopulations enables comparisons between groups. For instance, one could examine whether there are differences between science students' perceptions of positive learning experiences and humanities students' perceptions of said experiences.

In one of our research studies, we sought to understand how the content MOOC participants post on social media varies with the role they espouse (Veletsianos, 2017a). After identifying a MOOC provider that included hashtags with every course offering, we examined what messages were posted to the course hashtags and how those varied by user role. Following traditional content analysis methods and categorization according to roles, we identified variations in the messages posted by different groups of users. For instance, we found that institutions and the MOOC provider posted more promotional messages than faculty and learners, while MOOC-dedicated accounts and instructors posted more instructional messages. Such results highlight the need for looking deeper into participant subpopulations to identify and examine the differential practices that subpopulations may employ, especially in the context of open-ended and flexible learning environments.

Avoiding Many Biases

It is widely recognized and acknowledged that conscious and unconscious biases have significant impacts in research outcomes. To mention a few, such biases might include Hawthorne effects (e.g., a teacher engages in behaviors perceived to be desired by a researcher observing their instruction), self-reporting biases (e.g., a student provides biased self-assessed measures of the time they spent studying for an exam), and self-selection biases (e.g., faculty in support of open access publishing in IDT self-select to participate in a study examining open access publishing in the field). Such biases adversely affect our understanding of issues related to IDT, and, even though researchers are trained to recognize and account for them, we are not always able to control for them.

Public Internet data mining approaches avoid many such biases. For instance, researchers are able to unobtrusively observe behavior *in situ*, mitigating the potential for Hawthorne effects, and self-reporting and self-selection biases. As an example, our investigation of the types of messages posted by IDT departments on social media sites (Romero-Hall, Kimmons, & Veletsianos, in press), relied on identifying and categorizing the actual messages already posted by IDT departments online. Thus, IDT department behavior was not impacted by virtue of the study being conducted, and self-reporting and self-selection biases were avoided because all available actual messages were collected and analyzed rather than depending on analyzing IDT departments' perceptions about those messages. It is important to note, however, that it is impossible to account for all potential biases. For instance, in the aforementioned study results are based on the sample of IDT departments identified, and the methods used to identify the specific departments to include in the study may have led to some departments being included/excluded.

Some Challenges of Public Internet Data Mining

Despite these benefits, public internet data mining as a research method presents a variety of noteworthy challenges. These challenges revolve around technical, methodological, professional, and ethical issues that arise from using massive amounts of public observation data from people and organizations. We have organized these challenges into the four following themes:

- multifaceted expertise and rigor requirements;
- focused questions and determining meaning;
- performative and contextual considerations of public data;
- and emergent ethical dilemmas.

Multifaceted Expertise and Rigor Requirements

The first challenge and largest barrier to entry for most education researchers who might have an interest in public internet data mining is that collecting, cleaning, organizing, and analyzing these data at any scale relies upon various technical skills that are interdisciplinary (at best) or not taught at all in most education research programs. This is in part due to the relative newness and ever-evolving nature of the internet (e.g., the emergence of APIs) but is also due to the siloed and specializing nature of the academy, which requires education researchers to utilize increasingly specialized methods of inquiry in order for their work to be considered valid. For instance, researchers who have already devoted years to becoming expert at phenomenological inquiry or structural equation modeling might understandably be slow to venture into a new realm of inquiry that might require them to learn equally specialized technical methods such as website scripting, API querying, tokenization, and so forth. In the reverse situation, however, web developers, data scientists, and internet marketing professionals might have a variety of skills necessary to do public internet data mining, but they will equally lack the content area expertise necessary to ask meaningful questions of the data and will make various assumptions about educational phenomena, institutions, and stakeholders that are controversial, unwarranted, or just wrong. Thus, especially in the case of small-budget projects (such as theses and dissertations), it becomes very difficult for a single researcher or even a small group of researchers to have all of the expertise necessary to do this kind of work in a way that will be viewed rigorously by education, web development, and data science communities alike.

To illustrate some of the expertise required, we will briefly explain some of the data collection steps that we undertook in a recent study of U.S. university Twitter accounts (see Kimmons, et al., 2017 for a complete explanation of all steps undertaken). After identifying two pre-existing lists of university websites, we used keyword identifiers and manual coding to merge the lists into a relational database to match Carnegie classifications with university website addresses. We then wrote a series of scripts that systematically opened and parsed the contents of all the university website homepages, searching for embedded Twitter feeds, links, or keyword references to an institutional Twitter account (e.g., "Follow us @OurUniversity"). The script stored all referenced accounts in the relational database with a unique university identifier. Another script we wrote queried the Twitter REST API, retrieving the Twitter user objects for all university accounts and storing them in the relational database. Next, we read through all account information (e.g., screen name, location, description) and manually coded accounts as either the primary institutional account or other (e.g., athletics department, registrar). This resulted in a maximum of one primary institutional Twitter account for each university (n = 2,411), and we excluded other accounts from further analysis. We then wrote another set of scripts to again query the Twitter REST API for all available Twitter activity for each account and stored returned tweet objects in the relational database (n = 5.7 million tweets). Following these data collection steps, we developed scripts to clean the data, developed scripts to identify multimedia in tweets, used an open-source sentiment analyzer, operationaled items of theoretical interest, identified representative samples, and conducted descriptive, inferential, and content analyses.

As this highly abridged narrative of some of the steps taken suggests, this one study required many technical steps to complete that required web scripting, quantitative analysis, qualitative coding, SQL querying, API querying, JSON parsing, keyword searching, database management, image analysis, sentiment analysis, and so forth. Furthermore, each study that is undergone in this way may have many unique elements to it that prevents the development of a one-size-fits-all approach to data collection and analysis. These challenges may be alleviated most readily by building functional teams of researchers (e.g., a web programmer, a quantitative methodologist, and a qualitative methodologist), but they also introduce challenges of getting the work published, because just as it is highly infeasible for one researcher to have all of the expertise necessary to conduct a study like this, it is equally infeasible that a single reviewer or editor can meaningfully evaluate a completed study's significance and rigor.

This last point is important for any researcher who is expected to publish their work in certain types of venues, because all journals have a niche audience and rely upon reviewers that have a unique set of beliefs, attitudes, and skills. When submitting studies like the one described above to the journals we are most interested in publishing in, we have found that reviewers and editors typically come at the study either from an education perspective (and thereby want to see rich, meaningful results in terms of students' and educators' lives) or from a computer science or methodological perspective (and thereby want to see conformity to expected norms of data collection and classification as well as methodological insights). This can require the researcher to essentially serve two masters wherein one wants more qualitative examples and less technical jargon while the other wants the opposite and is exacerbated by word limit requirements that essentially require the researcher to choose one over the other. We have found that this issue must be navigated on a study-by-study basis wherein the researchers must iteratively work with the editor and reviewers to determine which elements of the study should be emphasized and which elements can be effectively summarized, placed in an online supplement, or ignored.

Focused Questions and Determining Meaning

Second, when working with a pre-existing, massive dataset like the internet, as researchers it is sometimes difficult to navigate the relationship between our research questions and the data. The traditional social science research approach, for instance, is for the research question to come first and for it to guide the collection and analysis of our data. However, with a pre-existing dataset this approach often feels inappropriate, because the researchers are simultaneously constrained and empowered by the parameters of the data, which may not allow them to answer questions that they are interested in but may also empower them to answer new questions that they did not know were possible to answer. It has been our experience that often when embarking on these studies our initial questions become reshaped or somewhat refined as we immerse ourselves in the data and contemplate their possibilities, but at the same time this often leads to scope creep, wherein we quickly try to tackle too much because we feel that the data are so rich, and theoretical drift, wherein we move away from our theoretically-grounded emphasis to focus on disconnected, emergent issues that we thought were novel and interesting. Both scope creep and theoretical drift are problematic for a variety of reasons not least of which is that they lead to studies that overreach or that can delve into areas far outside the researcher's realm of expertise, and discerning audiences are quick to point this out.

This situation has led us to enter these types of studies with focused research questions at the outset and to be much more careful in safeguarding against drastic changes late into the research process. Though we feel that there should always be some flexibility to refocus research questions in light of emergent data issues, those embarking on studies like these should never approach a massive dataset with a "we'll see what the data can tell us" attitude, because the data are often so rich that they can become more of a distraction than a tool of inquiry.

A related issue is how we think about significance and meaning and how our qualitative or quantitative traditions might prepare us to approach massive pre-existing data in inappropriate ways. For instance, in a traditional education research study that employs a quasi-experimental design, a researcher might study as few as 10 or as many as 1,000 participants and look for statistically significant differences between participant groups based upon a set of *a priori* factors. In such a scenario, a statistically significant result is the typical goal, and though such significance is discernible even with a small sample (given large enough effect sizes), larger samples are generally preferable, because they allow researchers to discern differences at a finer granularity (thereby reducing Type II error likelihood). However, in the case of massive datasets that rely upon millions of data points, it becomes possible for virtually any difference between groups to be detectable, even those with effect sizes that have no reasonable meaning. For instance, in testing the theoretical notion of the *romance of the public domain* in the adoption of open source software among schools, we were able to determine that this phenomenon did in fact exist but that it only represented about 7% of the variance (Kimmons, 2015). Depending upon the factor being tested, 7% might be considered large or negligible, and as a result, significance testing in such studies should only be a precursor to discussion of what constitutes meaningful significance for the specific theoretical constructs in question.

A similar issue faces researchers intent on using these sources for more qualitative purposes. In a traditional qualitative study, researchers will typically try to target outliers or valuable informants for conducting interviews and focus groups, favoring rich information, nuance, and transferability to quantitative generalizability. The benefit of a massive dataset is that it gives qualitative researchers the ability to find outliers or instances representing anything, thereby treating a once-in-a-million tweet as something special rather than as something inconsequential. However, such power requires a certain level of restraint and circumspection on the part of the researcher, because by hyper-fixating on outliers, research can (perhaps inadvertently) lead to misconceptions about the phenomenon or serve to reify unjust biases. That is, even if qualitative researchers do not make claims of generalizability, their work can lead to a general perception that may be inaccurate, and massive datasets can provide infinite fodder for doing this. One example of this that we point to in a previous study regards our general perceptions of trolling, negativity, and toxicity in social media (Kimmons, McGuire, Stauffer, Jones, Gregson, & Austin, 2017). Though such behaviors certainly exist online, our common perceptions of the overall negativity of online spaces may actually be based upon an overemphasis on a small minority of interactions, thereby allowing those who exhibit such behaviors to dictate our perceived norms of these media. Researchers can mirror this phenomenon by focusing on minority behaviors, such as trolling, and can thereby lead readers to develop views of these media that misinterpret targeted instances as generalizable evidence rather than as outliers.

Performative and Contextual Considerations of Public Data

Though one of the great benefits of these data is their observational nature and the invisibility of the researcher in the data collection process, thereby avoiding biases as noted above, this invisibility quickly becomes a double-edged sword if researchers fail to interpret observed behaviors through the performative and contextual lenses of the media they are studying. Much has been written about imagined audiences, and as users participate in commutative acts, they do so with certain assumptions about who they are participating with and how their behaviors will be interpreted (Kimmons & Veletsianos, 2015; Marwick & boyd, 2011). This negotiation of context and behavior makes interpretation of intention and authentic identity based upon observations alone difficult (Kimmons, 2014) and requires researchers to adopt more nuanced, contextually-adaptive constructs of identity than are generally used (e.g., Kimmons & Veletsianos, 2014). For instance, in one study we found that professors and graduate students showed differing levels of engagement with potentially controversial political topics on Twitter (Veletsianos & Kimmons, 2016). Based upon this result alone, one might be tempted to conclude that professors are simply more politically active or opinionated than their students, but such an interpretation would ignore the power structures that influence students' willingness to take a stand on controversy (e.g., vulnerability when seeking a job), the performative nature of the behaviors (i.e., they are performing for potentially different intended audiences), and how professors' intended purposes for using Twitter (e.g., outreach and awareness) may be different from students' purposes (e.g., sharing their work and job seeking).

This same consideration applies to anyone using any medium, because as children and adults or students and instructors use these tools, they will adapt their behaviors based upon a variety of factors including audience, purpose, and technological limitations (such as word length, available memes, or solicited participation norms). Thus, as researchers attempt to make sense of these behaviors, they cannot be evaluated in a context-free manner or in a manner that ignores how such factors might vary between groups. It also means that claims about monolithic identity constructs (i.e., how a person *is*) are difficult, because identity transcends singular contexts, and attempting to interpret a person's identity from a singular context yields results that are transitional, socially-responsive, and necessarily incomplete (Kimmons & Veletsianos, 2014).

Emergent Ethical Dilemmas

Finally, conducting research by mining public internet data provides a variety of emergent ethical dilemmas that researchers should be aware of in order to mitigate any negative consequences. In most institutions, institutional review boards (or IRBs) assist researchers by ensuring that their work meets legal and ethical guidelines, but ethics guidelines are not keeping pace with new research practices and possibilities (Taylor & Pagliari, 2017). This is especially true in matters of public and mental health and has ramifications for studies that involve the interpretation of public data in ways that might be problematic or harmful for individuals.

The U.S. National Institutes of Health (NIH), for instance, define "human subjects research" as research that involves living individuals "about whom an investigator … conducting research obtains data through intervention or interaction with the individual, or identifiable private information" (NIH, 2018). In the case of public internet data (such as tweets, forum comments, or blogs), data collection does not typically involve an intervention or interaction with the author, and the collected data is typically not private, because it is made publicly available on the internet by the original author. As with most data provided on the internet, there is no reasonable expectation of privacy to Twitter posts or YouTube comments, and the NIH would therefore likely not classify their analysis as human subjects research. Thus, when IRBs consider ethical ramifications of potential data mining studies, they typically either do not know how to deal with them or ignore them as non-human subjects research.

However, just because data is public does not mean that it has no potential ramifications for private individuals, and just because a person posts a message publicly to Twitter without any expectation of privacy, it does not follow that they were conscientious of the possibility that their messages would be aggregated, studied, and reported on. Thus, though they may *consent* to their data being public via a site-specific terms-of-use agreement, they may not *assent* to their data being used in various ways. Sharf (1998) pointed out this issue two decades ago with the example of a public breast cancer discussion forum, wherein participants knew that their words were publicly available but they sometimes felt like using their words for research was unethical without their expressed consent. Sharf's response was to suggest that internet researchers should essentially treat such participants like participants in any other research project through introductions, gaining consent/assent, and minimizing risks. Though such a set of guidelines seems reasonable, Sharf explains that "there must be room for the researcher to exercise judgment" (p. 255) insofar as internet spaces have different contextual considerations and the internet continues to evolve. Today the shear amount of available data and participants dwarfs what was available in the 90's, and the type of data allows for new threats (e.g., geolocation). These are realities of an ever-evolving internet, wherein public data are regularly used for everything from marketing to search engine optimization, but they bring with them ethical considerations that researchers have not previously needed to grapple with.

In more recent years, the Association of Internet Researchers (AoIR) has sought to help provide cross-disciplinary guidelines for researchers to follow that take into account the potential risks of certain types of internet research (e.g., those dealing with sensitive data) and the contextual factors that influence researcher decision-making, such as whether informed consent is necessary or advisable (Ess & Jones, 2002; Markham & Buchanan, 2012). The crux of the issue is that internet researchers need to find ways to do valuable research while at the same time minimizing risks for potential harm and valuing "the individual's integrity and right to self-determination" (Elm, 2008, p. 69). For instance, by determining a users' political attitudes, religious affiliations, sexual orientations, anxiety levels, or depression likelihood based upon public data, such information could be used by third-parties to target people for harassment or discrimination. As a concrete example, we are currently conducting studies wherein we analyze K-12 teacher Twitter activities and utilize IBM's Watson AI to correlate certain personality characteristics of teachers with school demographics, religious expressions, and political stances (Carpenter, Kimmons, Short, Clements, & Staples, under review; Krutka, Kimmons, Harding, & Harker, under review). In the case of such a study, if we provided an aggregate dataset of users that revealed political party, religious beliefs, personality traits, etc., or provided direct quotes by teachers in our reports about controversial policies (e.g., Common Core, the appointment of Betsy DeVos), then a hiring committee could conceivably use such a dataset when vetting prospective hires to deny employment to teachers based upon discriminatory factors. Though researchers cannot control the behavior of others, we should at least be mindful of how the datasets, tools, and conclusions we generate might impact our participants and strive to minimize such impacts whenever possible. This same issue also applies at least somewhat to groups of people and institutions, which may sometimes be identified as legally culpable through our work, such as schools and universities that are supposed to make their public websites accessible to users with disabilities (e.g., Kimmons, 2017; Veletsianos et al., 2017).

For this reason, we have reflectively developed our own research practices to err on the side of caution and respect for individuals' privacy in order to avoid our work from being used to support discrimination, prejudice, social shaming, or harassment, even though IRBs or publishers have not explicitly required us to do so. One simple and powerful example of this is in the generation of datasets and identification of users in reports. Since all public internet data begins as public artifacts, journal editors, peer-reviewers, and other researchers sometimes expect that restructured datasets should be openly shared and redistributed. Generally speaking, this is an important requirement for transparent scholarship that allows for verification and replicability. However, in the case of public data, doing so can often mean that our subjects are now identifiable in new ways (e.g., based upon their religious beliefs or trolling behaviors) and that anyone can find the exact person that was included in the study (e.g., based upon keyword searches of included content). To address such potentials for harm, in our studies we generally begin by following the same guidelines expected in human subjects research and de-identify users in our reports and datasets whenever there is a possibility of harm coming from the individual's inclusion in our study. However, even this is often not enough, because participants still may be identifiable by their artifacts (e.g., specific tweets), which means that sometimes we reword artifacts to prevent reverse lookup (e.g., rephrasing an illustrative tweet used in a report) or decline to provide our datasets altogether. Such a decision often runs contrary to journal and reviewer expectations, but we believe that it is sometimes an essential one given the relative infancy of such work and the potential for harm, and a more sophisticated treatment of these and other guidelines may be found in the AoIR report on the subject (Markham & Buchanan, 2012).

Furthermore, the ethical considerations surrounding responsible public internet data mining research are in a current state of flux and will continue to evolve in response to technical shifts and social expectations of privacy. It may be that future iterations of the definition of human subjects research used by NIH, IRBs, etc. will evolve to reflect our current realities, but in the meantime, it behoves responsible researchers to proactively seek to minimize harm that could be caused by their research efforts and to participate in the establishment of ethical guidelines for all.

Conclusion

As illustrated in our analysis and examples above, the practice of public internet data mining methods combines exciting opportunities with difficult challenges in wicked ways. Having engaged with these approaches in our own research, as well as in a mentoring and peer-reviewing capacity, we are convinced that doctoral preparation programs in our field need to prepare future IDT faculty and professionals with the skills and literacies to critically read, evaluate, and conduct emerging research methodologies. While this paper investigated public internet data mining methods as one such approach, doctoral preparation programs should prepare individuals to engage with a broader selection of methods.

Ethical approval: This article does not report on a study with human participants performed by any of the authors.

References

Andersen, D. G., & Feamster, N. (2006). Challenges and opportunities in Internet data mining. Parallel Data Laboratory, Carnegie Mellon University, Research Report CMU-PDL-06-102. Retrieved from https://pdfs.semanticscholar.org/8105/56e5f248e93f56e2ede855662fde9fad454d.pdf

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer New York.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS medicine, 5(7), e151.

Carpenter, J., Kimmons, R., Short, C. R., Clements, K., & Staples, M. E. (under review). Crossing the professional-personal divide: Teachers using twitter as a platform for expression and sharing.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. Journal of Communication, 64(2), 317-332.

Elm, M. S. (2008). How do various notions of privacy influence decisions in qualitative internet research? In A. N. Markham & N. K. Baym (Eds.), *Internet inquiry: Conversations about method*. Sage.

Ess, C., & Jones, S. (2002). Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee. Retrieved from http://aoir.org/reports/ethics.pdf

Kimmons, R. (2014). Social networking sites, literacy, and the authentic identity problem. *TechTrends, 58*(2), 93-98. doi:10.1007/s11528-014-0740-y

Kimmons, R. (2015). Open online system adoption in K-12 as a democratizing factor. *Open Learning: The Journal of Open, Distance and e-Learning, 30*(2), 138-151.

Kimmons, R. (2017). Open to all? Nationwide evaluation of high-priority web accessibility considerations among higher education websites. *Journal of Computing in Higher Education, 29*, 434-450.

Kimmons, R., McGuire, K., Stauffer, M., Jones, J. E., Gregson, M., & Austin, M. (2017). Religious identity, expression, and civility in social media: Results of data mining Latter-day Saint Twitter accounts. *Journal for the Scientific Study of Religion.*

Kimmons, R., & Veletsianos, G. (2014). The fragmented educator 2.0: Social networking sites, acceptable identity fragments, and the identity constellation. Computers & Education, 72, 292-301. doi:10.1016/j.compedu.2013.12.001

Kimmons, R., & Veletsianos, G. (2015). Teacher professionalization in the age of social networking sites. *Learning, Media and Technology, 40*(4), 480-501. doi:10.1080/17439884.2014.933846

Kimmons, R., & Veletsianos, G. (2016). Education scholars' evolving uses of Twitter as a conference backchannel and social commentary platform. British Journal of Educational Technology, 47(3), 445-464. doi:10.1111/bjet.12428

Kimmons, R., Veletsianos, G., & Woodward, S. (2017). Institutional uses of Twitter in U.S. higher education. *Innovative Higher Education, 42*(2), 97-111.

Krutka, D., Kimmons, R., Harding, T., & Harker, Z. (under review). Speaking out on twitter: Understanding teachers' expressed sociopolitical sentiments to improve policy making.

Maloof, M. A. (Ed.). (2006). Machine learning and data mining for computer security: methods and applications. Springer Science & Business Media.

Markham, A., & Buchanan, E. (2012). Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee (Version 2.0). Retrieved from http://aoir.org/reports/ethics2.pdf

Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, *13*(1), 114-133.

National Institutes of Health. (2018). *I am an investigator.* Retrieved from https://humansubjects.nih.gov/walkthrough-investigator

Niwa, S., Doi, T., & Honiden, S. (2006, April). Web page recommender system based on folksonomy mining for itng '06 submissions. In Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on (pp. 388-393). IEEE.

Paskevicius, M., Veletsianos, G., & Kimmons, R. (in press). Content is king: An analysis of how the Twitter discourse surrounding open education unfolded from 2009 to 2016. *The International Review of Research in Open and Distributed Learning.*

Romero-Hall, E., Kimmons, R., & Veletsianos, G. (in press). Social media use by instructional design departments. *Australasian Journal of Educational Technology.*

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society*, *18*(2), 121-138.

Selwyn, N. (2015). Data entry: towards the critical study of digital data and education. *Learning, Media and Technology*, *40*(1), 64-82.

Sharf, B. F. (1998). Beyond netiquette the ethics of doing naturalistic discourse research on the internet. In S. Jones (Ed.), *Doing internet research: Critical issues and methods for examining the Net*. Sage Publications.

Taylor, J., & Pagliari, C. (2017). Mining social media data: How are research sponsors and researchers addressing the ethical challenges?. *Research Ethics*, 1747016117738559.

Veletsianos, G. & Kimmons, R. (2012). Networked Participatory Scholarship: Emergent Techno-Cultural Pressures Toward Open and Digital Scholarship in Online Networks. *Computers & Education, 58*(2), 766-774.

Veletsianos, G., & Kimmons, R. (2016). Scholars in an increasingly digital and open world: How do education professors and students use Twitter? The Internet and Higher Education, 30, 1-10. doi:10.1016/j.iheduc.2016.02.002

Veletsianos, G. (2017a). Toward a Generalizable Understanding of Twitter and Social Media Use Across MOOCs: Who Participates on MOOC Hashtags and In What Ways? *Journal of Computing in Higher Education, 29*(1), 65-80.

Veletsianos, G. (2017b). Three Cases of Hashtags Used as Learning and Professional Development Environments. *Tech Trends, 61*(3), 284-292.

Veletsianos, G., Kimmons, R., Belikov, O., Johnson, N. (under review). Scholars' Temporal Participation on, Temporary Disengagement from, and Return to Twitter.

Veletsianos, G., Kimmons, R., Larsen, R., Dousay, T., & Lowenthal, P. (in press). Public Comment Sentiment on Educational Videos: Understanding the Effects of Presenter Gender, Video Format, Threading, and Moderation on YouTube TED Talks. *PLoS ONE*.

Veletsianos, G., Kimmons, R., Shaw, A. G., Pasquini, L., & Woodward, S. (2017). Selective openness, branding, broadcasting, and promotion: Twitter use in Canada's public universities. *Educational Media International, 54*(1), 1-19. doi:10.1080/09523987.2017.1324363

Wang, M., Madhyastha, T., Chan, N. H., Papadimitriou, S., & Faloutsos, C. (2002). Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic. In Data Engineering, 2002. Proceedings. 18th International Conference on (pp. 507-516). IEEE.