Review Article

2

Big Data Between Technology and Science: Challenges for Psychology and Social Sciences

Bojan Musil
Department of Psychology, Faculty of Arts, University of Maribor, Slovenia

Nenad Čuš Babič
Construction IT Centre, Faculty of Civil Engineering, University of Maribor, Slovenia

Abstract

Among phrases that originated in the context of IT and have become virally immersed in everyday life is "big data". Especially with the work of Kosinski and his colleagues, it has gained considerable attention in psychology and related disciplines, and following the Brexit and the US presidential election campaigns in 2016, has spread beyond the sphere of strict science. In the article we question whether big data research represents a distinctively new endeavour in the empirical sciences. From a methodological and analytical perspective, the three Vs of big data, volume, velocity and variety – represent challenges that foster new (interdisciplinary) strategies to handle them properly, but from the more traditional psychological perspective, a fourth V – veracity, addressing the importance of data truthfulness – raises additional questions. However, from the perspective of psychology and the social sciences, ethical considerations, addressing questions of privacy, anonymity and autonomy are even more important.

Keywords: big data, psychology, ethical considerations, privacy

Big Data in Social Sciences and Psychology: A Recent Endeavour?

In recent years many phrases from the world of information technology have gained special attention from the professional and general public, and with the immersion of social networks into our everyday life, one of these phrases is definitely "big data".

Following Harlow and Oswald (2016), "big data involves the storing, retrieval and analysis of large amounts of information" (p. 447). To put it more simply, it involves work with large data sets.

Some authors have ascribed the origin of the concept to a computer scientist from Silicon Valley, John Mashey in the late 1990s (Diebold, 2012; Lohr, 2013). At first, big data research received considerable interest in the fields of computer science, business and statistics, the latter especially in the domain of government, medicine and public health.

According to the results of search engines in the most distinguished scientific databases, most of the scientific work relating big data to the social sciences (and psychology) dates from recent years (or even months). Could we thus make an assumption that big data studies are relatively recent in the research field of social sciences, and psychology in particular? Or could we even call it a new endeavor in the social sciences and psychology?

If we interpret the definition of big data as merely a huge amount of data, the answer would be definitely no. Two classic examples from the domain of psychology, with a prominent influence on the broader (social) sciences, could be quite useful in providing an answer to the previous question.

In the late 1960s and early 1970s, Hofstede conducted a study on a huge sample of more than 110,000 workers in a global multinational company (the company was later revealed to be IBM). From the data gathered, he elaborated his model of cultural value dimensions, and his most popular value dimension, individualism versus collectivism, made an enormous impact on future research in diversity among social science fields (e.g. Hofstede, 1980, 2001).

In the late 1980s, a new model of personality structure was proposed, which claimed to be universal. It comprises the personality traits of Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. The era of the Big Five model (later also known as the OCEAN model) was on the rise (e.g. Goldberg, 1981). Studies of the model in different demographic and sociocultural groups (gender, age and culture) used samples as large as 11,000 or 23,000 participants (e.g. Costa, Terracciano, & McCrae, 2001; McCrae, Terracciano et al., 2005a, 2005b).

However, both examples represent research where data sets have a limited (manageable) number of items (i.e. constructs in the analysis); the large quantity of data is thus a consequence of the enormous number of cases (subjects or participants in the study). In the analytical sense, these data with relatively simple structures and consequent straightforward (statistical) analysis does not represent any serious challenges.

However, if we interpret big data as a mixture of data size and complexity at the same time, a better example is the World Values Survey project (WVS). WVS is a worldwide research project, conducted on representative samples from all parts of the world, in more than 100 countries and with almost 400.000 respondents (WORLD VALUES SURVEY, 2017). It explores a variety of concepts, from values and beliefs, to specific attitudes, and through relatively structured and stable questions and items, addresses these issues in different time sequences, with the first wave between 1981 and 1984. Wave 6 is currently available, covering data from the surveys between 2010 and 2014 (WORLD VALUES SURVEY Wave 6 2010-2014).

In our opinion, the above-mentioned interpretations and examples do not fully represent the modern definition of big data. With the advent of the web and online social network sites (OSNs), the meaning of big data changed to incorporate the changing nature of the number of (online) participants

and incorporated constructs, indicators or other information about participants. Modern-day big data are very dynamic systems, incorporating more and less structured information: to analyse everything that someone is online, together with everything that someone does online, and possibly executing this in real time.

Leskovec (2008), in his seminal work, claims that analysis performed on big data reveals otherwise invisible phenomena. In addition, the author points out an opportunity to analyse social phenomena like communication patterns, using anonymized datasets without compromising individual privacy.

Big Data and Psychology on OSNs: Case Kosinski

The digital age can best be illustrated with the phrase "datafication of everyday life", which is a product of the extent to which people use digital and online technologies, as well as the extent to which digital technologies have replaced older ways of life (Chen & Wojcik, 2016, p. 458).

As Kosciejew first quoted Wolf from *The New York Times*, "And yet, almost imperceptibly, numbers are infiltrating the last redoubts of the personal. Sleep, exercise, sex, food, mood, location, alertness, productivity, even spiritual well-being are being tracked and measured, shared and displayed" (Wolf, 2010), he later added:

The individual and (ideas of) individuality are being transformed into quantifiable, statistical, and objective data points that can (allegedly and supposedly) help reveal new possibilities, novel insights, and hidden facts about ourselves. A data-driven life, in other words, is datafying the individual (Kosciejew, 2013, p. 47).

To estimate the size of the Internet and the quantity of transactions on the web is an almost impossible mission. To illustrate it, Amin (2017) in the Big data overview 2013-2014, collected the following information: Google processes 100 petabytes per day on 3 million servers; Facebook has 300 petabytes, processes 500 terabytes per day and has 35% of the world's photographs; YouTube has 1000 petabytes of video storage and 4 billion views per day; Twitter processes 124 billion tweets per year. The importance of big data (for different purposes) is thus quite obvious – as is any elaboration of the definition of big data.

According to Laney (2001), big data can be described with three Vs or three dimensions: Volume, Velocity and Variety. Volume is related to the large size of the dataset, velocity to the high speed of data acquisition and processing and variety to the diversity of data types, being often an unstructured mixture of texts of speeches, photographs, videos and numbers (e.g. Chen & Wojcik, 2016; Cheung & Jak, 2016). Especially from the perspective of psychology, some authors add a fourth V – Veracity, addressing the importance of data truthfulness (e.g. Cheung & Jak, 2016; Saha & Srivastava, 2014). In more psychological or psychometric language, this V is related to the question of the data validity and the consequent conclusions based on data analysis. As a simple rule of thumb, big data refers to datasets that cannot be adequately processed by traditional IT and its tools (Chen, Mao, Zhang, & Leung, 2014).

This new dynamics and understanding of big data in the domain of psychology can be quite illustrative, as seen in the work of Michal Kosinski.

Kosinski (e.g. Kosinski, Stillwell, & Greapel, 2013) shows that using a very few, easily accessible indicators of human behavior, for example Facebook likes or Twitter followers and reactions (Quercia, Kosinski, Stillwell, & Crowcroft, 2011), can accurately predict many personal attributes of a user. For example, it is relatively possible to predict reliably some highly sensitive personal data like "sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender" (Kosinski, Stillwell, & Greapel, 2013, p. 5802), using just FB likes. The likes significantly correlate with the above-mentioned personal attributes. Ko-

sinski developed this algorithm correlating public FB data with the results of traditional personality assessment questionnaires, like the well-known personality questionnaire NEO-PI-R (Kosinski, Stillwell, & Graepel, 2013).

Among Twitter users, just by reading their three publicly available data points – "followers", "following" and "listed counts" – it is possible to predict personality traits according to the well accepted OCEAN (BIG5) model. This means personality can be effectively and easily predicted from public data (Quercia et al., 2011). In addition to simple online gestures, language pattern analysis could reveal transient personal states like emotions and also trait characteristics like personality (Schwartz et al., 2013).

This could be of great value for people when the findings support the needs of a given user. However, as Kosinski (2016) points out, one can also imagine that other people could use those algorithms to reveal attributes a person does not want to reveal, along with those attributes which could pose a threat to an individual. Another problem could be that other people could reach inappropriate conclusions about a person when the algorithm is wrong.

These examples show that even the most cautious users of social networks who protect access to their published content can easily be analysed and users have no insight into how this information will be used. Additionally, reliability of those assessments can be improved by integration of data from different sources. Finally, the data shared on social networks can be purchased from companies owning the systems and reused for unknown purposes.

From Big Data and Psychology: Mere Application or Manipulation?

One recent example of the above-mentioned activities is revealed by claims from Cambridge Analytica (Persily, 2017) about successful social influence involving political propaganda based on psychographics, which is the exploitation of psychological profiles created from available digital data, like Facebook data in the first place and other (legally) bought consumer datasets (e.g. from magazine subscriptions, to airline tickets).

Are all these claims realistic?

Several authors (like TED speakers Goldbeck, 2013; Kohn, 2014; Nolan, 2012; Kaliouby, 2015; Lupi, 2017) in the field of information technology and big data have expressed concerns about the integration of information sources in an individualized form, because psychological profiles can be developed from these. Since widespread social networks, with their ubiquity and pervasiveness, could lead to important social implications, they suggest how to reestablish privacy. The commonly suggested solutions are naive, such as "we should pay attention to what we click online", or to resolving problems with technology by demanding even more technology. Despite these warnings, there are no scientific publications reporting how this knowledge about user personality can be (or has been) effectively used for manipulation.

As an active participant in US presidential campaign in 2016 helping first Ted Cruz and then Donald Trump, Cambridge Analytica is the most (publicly) salient case of using targeted communication based on psychometrics for social influence, even though there are still doubts about how factual these claims really are.

At first glance, such thinking could be supported by scientific research in psychology in the area of personality and attitudes. It is well known that cross-situational studies show that personality traits do not significantly predict behavior (Mischel, 1968). Similarly, behavior is often inconsistent with attitudes expressed by a person (especially more general attitudes) (Wicker, 1969; Ajzen & Fishbein, 1977). However, personality traits and attitudes, in combination with contextual influences, do affect behavior. With regard to the BIG 5 (OCEAN) personality model, research shows that low neuroticism, high openness and extraversion correlate with good response to messages promising comfort, and agreeable and con-

scientious subjects respond better to messages presenting utility aspects of objects and activities (Chen & Lee, 2008; in Gerber, 2013). Subjects displaying anxiety trait usually respond better to more attractive messages. On the other hand, less anxiety correlates more with better response to threatening messages (deBono, 1994; in Gerber, 2013). The most consistent claim across research studies is that subjects respond better to messages delivered by persons whom they evaluate as similar regarding their personality (Gerber, 2013). Hence, when manipulation activity is delivered in a context prepared for a specific target subject and according to their personality traits and attitudes, there are grounds for success. This kind of reasoning can also be identified in statements from Cambridge Analytica, who claim that they did not create the content of the messages for their clients (the presidential candidates), but that they analysed or influenced the context and advised when and where specific message would be most effective, and that they formulated targeted messages for specific recipients using their personality profiles.

How does this work on social networks? The principle was described in detail by Eli Pariser in his book "What the Internet is hiding from you" (Pariser, 2011). The author introduced the term "filter bubble", which is a principle where users on social networks get biased information adjusted to their past behavior. The more often you click on a certain type of content, the more of it you will be offered. Pariser focused particularly on the content and relationships available in the network. For example, he noticed that the newsfeed on his Facebook profile was biased toward messages from users consistent with his preferred political view.

The impact size of the filter bubble is questioned by other researchers like Liao and Fu (2013), who claim that, even when all information is present, users preferentially select information that reinforces their views, and also by Facebook itself, which claims that bias in newsfeeds is very low. Nevertheless, if personality information is used for customization of messages and customization of the environment according to the above-mentioned research, we can speculate that such a bubble could establish an effective context for manipulation.

Initial research performed during election campaigns shows that such manipulation is possible. For example, Bond, Faris and Jones (2012) proved that it is possible to increase participation in voting with a quasi-experiment in natural setting. They targeted 61 million Facebook users in the USA during elections and demonstrated that "go to vote" appeals in the form of social mobilization (showing photos of the user's friend who had already voted) did actually increase participation in elections. Even though the effect was not large (they report 0.14% actual increase), this is important in the context of big data and social networks, because the number of targeted users is enormous.

Gerber (2013) added personality dimension to this body of knowledge. The author explores the efficacy of different types of messages delivered to people of specific personality types at the level of attitude change and behavior change in "go to vote" campaigns. The results show that subjects displaying high openness are susceptible to different kinds of messages, and that other OCEAN traits contribute to effectiveness of more specific types of persuasive messages in this context.

Big Data and Psychology: Pros and Cons

Harlow and Oswald (2016), in the Introduction to the special issue of *Psychological Methods*, highlighted the common themes that emerge in psychological research in the area of big data. First, there are the mutual benefits of collaboration among diverse of disciplines, such as those from social sciences, applied statistics and computer science. Second, the availability of large data sets from different OSNs provides a psychological window into the attitudes and behaviors of a broad spectrum of the population, and, in a methodological sense, there arises the opportunity for and also the necessity of testing the diversity of predictive models in big data. However, in the process of acquiring and processing large data sets

from public or private sources, there are important ethical considerations.

Since psychologists remain mostly suspicious about big data movement, primarily because of widespread subjective perceptions about their impairment in the area of computer programming and related IT skills and lack of access to big data (e.g. Cheung & Jak, 2016); there is much less restricted optimism in the field of computer science. Montag, Duke and Markowetz (2016) promote the emerging research discipline Psychoinformatics, the interdisciplinary cooperation between psychology and computer science in handling large data sets derived from a range of heavily used IT devices.

However, Cheung and Jak (2016) argued that psychologists equipped with the knowledge, skills and capacities of psychological and behavioral theories, psychometrics and statistics, are valuable in understanding and processing big data. Specifically, this is the case in the phase of data collection (which data are collected and how), whether the data have adequate psychometric properties, and in subsequent phases of construction and testing the hypotheses and models, to explain behavior(s) with advanced statistics, such as multilevel modelling, structural equation modelling and meta-analysis. Additionally, what psychology can contribute to big data science is a strong theoretical background, which broadens the meaning of the findings. As Kosinski & Behrend (2017) pointed out, the data can only predict the future if it is consistent with the past.

There are numerous advantages of big data in psychological research. Besides classical psychological research techniques such as questionnaires and experiments conducted online, there is an unrestricted opportunity for the third fundamental technique: observation of human-computer interaction on a very large scale (Montag et al., 2016). Data can be divided into a large spectrum of samples, encompassing distinct groups, socio-structural groups, subcultures and cultures. Moreover, a (cross-)cultural perspective can even be analysed in a historical manner (for how big data can trace cultural change over time with Google Books Ngram Viewer, see Pettit, 2016). Because data on human-computer interaction represents directly recorded behavior(s), some biases or drawbacks of standard psychological techniques and measures are omitted - such as the tendency to produce socially desirable answers on self-report measures or, despite tracking of real behavior, tracking only behavioral intention or perception of behavior from past experience.

As an example, research in clinical psychology shows that psychopathology or related risks could be identified and dealt with in very early stages, reducing the cost of interventions and preventing severe outcomes. Also, the patients could better monitor their condition and the therapists could adjust medication in response to the present situation (Markowetz, Błaszkiewicz, Montag, Switala, & Schlaepfer, 2014). Luhmann (2017) reports that using big data one could successfully analyse the current state of subjective well-being on individual and social levels and predict changes in subjective well-being over time, especially in cases of rare events like natural disasters or terrorism where traditional research methods are not very useful.

In general, big data from OSNs represents a shift of research focus from subjective internal states or psychological constructs to objective, observable behaviors or results of behavior(s). By omitting subjective internal states, such as personality traits, emotional states, mood, intentions, interests and attitudes, this new research endeavor (i.e. psychoinformatics) can be seen as new form of behaviorism – cyberbehaviorism – the interrelation of digital behaviors without the necessary awareness of the individual (human).

Again, we face the question, whether digital or virtual behaviors (clicking or tapping) truly or authentically represent the individual. Or, can we adequately infer internal processes from these behaviors? Or vice versa? From the case of Kosinski et al., the answer would be affirmative.

However, to these questions, we can add three important (contextualized) issues. First, social

interaction through computer-mediated communication (CMC) represents a significant part of our everyday lives, but is still a specific kind of interaction. Fullwood (2007), from the work of McKenna et al. (2002), summarized the factors that set apart online spaces from the offline world: a greater propensity for anonymous interaction; a reduction in the importance of physical cues or appearance; a higher degree of control over time and space of the interaction; relative ease of finding similar others, and the additional factor of control over the content that is generated online. These characteristics have a significant impact on our self-presentation to others in the online world and its social relations. The question of whether our online self is authentic, or if we are the same online as offline, is still debated. It is hard to predict the future, but the idea that our real-life selves will in the future move in the direction of our online selves would probably gain a considerable number of votes from the professional or scientific public.

Second, from the methodological viewpoint, the data are collected and aggregated from multiple sources. This implies the data come from different contexts, using diverse data collection procedures, time spans, and error rates, which could all lead to incorrect statistical inferences. Consequently, Fan, Han & Liu (2014) call for new statistical methods in scientific research using big data.

Third, OSNs represent the complexity of social interaction. They do not merely serve as platforms for self-presentation by demarcated individuals, but are important place(s) of social influence, thus having a significant effect on someone's thoughts, feelings and behaviors and consequently her/his self-definition and identity. For example, the number of friends, number of likes, number and nature of comments can have a tremendous informative (i.e. what is right) or/and normative (i.e. what is consensual) impact on the individual, and consequently influence her/his future decisions and behaviors.

These issues address some of the theoretical or conceptual dilemmas, but probably the most important source of inconvenience that psychologists face in the context of big data can be attributed to ethical concerns. There is an established code of conduct in classical psychological research; however, in the context of big data, there are some potentially ambiguous situations. As, for example Montag et al. (2016) illustrate, a researcher or research team can deduce personality features of a user from his/her online behavior and hence have the potential to deny him/her a particular contribution. Or, in the analysis of big data, the characteristics of people that were primarily not in the scope of the particular big data sample can be processed (e.g. see article Webb, 2013).

This leads us to the major ethical concerns of big data processing: the issues of privacy, anonymity and autonomy. How should researchers deal with private information in the phase of collecting, processing and disseminating the data? Is the anonymity of research subjects preserved? Are participants informed about the research, and do they have a right to opt out of it? In sum, how is respect for persons (i.e. ICT users) addressed?

According to the previously mentioned cases of big data research in psychology and its application, these answers are generally not so clear-cut. In the context of big data health research, Rothstein (2015) argued that traditional research regulations should apply, and among essential things, individuals "…ought to be consulted and asked for permission before their specimens and data are collected, analysed, stored, and used for research" (p. 427), or, in other words, individuals "…ought to have the ultimate right to decide whether to participate in research" (p. 426). In this regard, researchers would be obliged to seek and obtain informed consent from the research subjects.

In summary, Kosinski et al. (2015) from various sources extracted that, since social scientists are relatively slow in embracing research on big data from OSNs, data-driven research has increasingly been left to computer scientists, who unfortunately, often lack the appropriate theoretical background and ethical standards relating to personal data protection. This is probably also a reason for psychologists (and other experts from the social sciences and humanities) to take a more active role in this interdisciplinary field.

Conclusion

In The Global Information Technology Report 2008-2009, Pentland (2009) proposed a "new deal on data", which will, according to the author, be the first step towards open information markets. In reference to Old English Common Law, he postulated the three basic principles of ownership:

1. Rights of possession: You have a right to possess your data – e.g. you can open an account and remove your data whenever you'd like.
2. Full control over use: You must have full control over the use of your data – e.g. everything must be opt-in, but with regular reminders that you can optout.
3. Right to dispose of or distribute your data: You have a right to dispose of or distribute your data. If you want to destroy it or remove it and redeploy it elsewhere, this is your call. (from Pentland, 2009, p. 79)

However, Pentland (2009) added one more principle, which addressed the combination of massive amounts of anonymous data to promote the common good. Aggregate and anonymous data can dramatically improve society: for example, people's movement could be used for early identification of infectious disease outbreaks, for protection of the environment and public safety. A step towards the regulation of the processing of personal data, which addresses these principles in practice, is recently implemented general data protection regulation (GDPR; Regulation, 2016) in the European Union.

On the other hand, there are new endeavors promoting the common good, which are based on massive amounts of data that is NOT anonymous. Recently, at a national level, China started several experiments on how to build "social credit system that covers the whole society" (Hvistendahl, 2017). The aim of the system is to rate the reputations of individuals, businesses and government officials, using data from public and private sources. Initially, people can join the system voluntarily, but it is foreseen that it will mandatorily include all citizens by 2020. In parallel to this state-led project, there is a private credit system initiative by China's largest marketing group Alibaba, named Zhima Credit. There are some indications that both systems will eventually merge or at least share data. Both systems use not only personal data but also a person's social networks to calculate a credit score.

In conclusion, we see that the balancing dilemma between individual privacy, autonomy and personal benefits from available data, on the one hand, and collectively driven benefits, on the other, remains unsolved. The source of this problem lies in the process of mutual influence between a person and society. As much as a person and her/his life is influenced by society, society and its values are dependent on individuals that form the society. When looking from the perspective of an individual, it seems that active participation and emancipation of a vast number of IT users in the context of monitoring and responding to potential misuse of IT platforms and environments could resolve the problem of coercion and misuse of personal data. However, social forces go beyond the reach of any single person's strength or influence. The person is a part of the broader society and is always influenced by the culture of the society that strives constantly to remain in existence.

Within big data research where new phenomena and patterns are revealed only when analyzing very large data sets (Leskovec, 2008), we can perhaps find empirical proof for social phenomena that could not be reduced to or explained on the personal or group level. Hence, most probably, proper answers regarding the use of big data, privacy protection, autonomy of individuals and the common good should also be sought in the interaction between psychological factors, as well as social and cultural factors. Looking for technical solutions, however artificially intelligent, for these problems would be naïve.

References

Amin, K. (2017, May 10). *Big data overview 2013-2014*. Retrieved from https://www.slideshare.net/kms-technology/big-data-overview-2013-2014

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888-918. http://doi.org/10.1037/0033-2909.84.5.888

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*(7415), 295-298. http://doi.org/10.1038/nature11421

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. M. (2014). *Big data: Related technologies, challenges and future prospects*. Springer International Publishing. doi: 10.1007/978-3-319-06245-7

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458-474. https://doi.org/10.1037/met0000111

Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00738

Costa, P., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322-331. https://doi.org/10.1037//0022-3514.81.2.322

Diebold, F. X. (2012). *On the origin(s) and development of the term "big data" (SSRN Scholarly Paper No. ID 2152421).* Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2152421

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293-314. https://doi.org/10.1093/nsr/nwt032

Fullwood, C. (2015). The role of personality in online self-presentation. In A. Attrill (ed.), *Cyberpsychology* (pp. 9-28). Oxford: Oxford University Press.

Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M., & Panagopoulos, C. (2013). Big five personality traits and responses to persuasive appeals: Results from voter turnout experiments. *Political Behavior*, *35*(4), 687-728. http://doi.org/10.1007/s11109-012-9216-y

Golbeck, J. (2013). *Your social media "likes" expose more than you think.* Retrieved from https://www.ted.com/talks/jennifer_golbeck_the_curly_fry_conundrum_why_social_media_likes_say_more_than_you_might_think

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In Wheeler, L. (ed.), *Review of Personality and Social Psychology*, *2*, 141-165. Beverly Hills, CA: Sage.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447-457. https://doi.org/10.1037/met0000120

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage Publications.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks, CA: Sage Publications.

Hvistendahl, M. (2017, December 20). Inside China's vast new experiment in social ranking. *Wired*. Retrieved from https://www.wired.com/story/age-of-social-credit/

Kaliouby, R. el. (2015). *This app knows how you feel from the look on your face*. Retrieved from https://www.ted.com/talks/rana_el_kaliouby_this_app_knows_how_you_feel_from_the_look_on_your_face

Kohn, S. (2014). *Don't like clickbait? Don't click*. Retrieved from https://www.ted.com/talks/sally_kohn_don_t_like_clickbait_don_t_click

Kosciejew, M. (2013). The individual and big data. *Feliciter, 59*(6), 47-50.

Kosinski, M., & Behrend, T. (2017). Editorial overview: Big data in the behavioral sciences. *Current Opinion in Behavioral Sciences*, *18*, iv-vi. https://doi.org/10.1016/j.cobeha.2017.11.007

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802-5. http://doi.org/10.1073/pnas.1218772110

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543-556. https://doi.org/10.1037/a0039210

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493-506. http://doi.org/10.1037/met0000105

Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/ 01/ad949- 3D- Data- Management- Controlling- Data-Volume- Velocity- and- Variety.pdf

Leskovec, J. (2008). *Dynamics of large networks*. *Technical report CMU-ML-08-111*. Carnegie Mellon University. Retrieved from https://cs.stanford.edu/people/jure/pubs/thesis/jure-thesis.pdf

Liao, Q., & Fu, W. (2013). Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2359-2368. http://doi.org/10.1145/2470654.2481326

Lohr, S. (2013). The origins of 'big data': An etymological detective story. *New York Times*. Retrieved from https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

Luhmann, M. (2017). Using Big Data to study subjective well-being. *Current Opinion in Behavioral Sciences*, *18*, 28-33. https://doi.org/10.1016/j.cobeha.2017.07.006

Lupi, G. (2017). *How we can find ourselves in data*. Retrieved from https://www.ted.com/talks/giorgia_lupi_how_we_can_find_ourselves_in_data

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypotheses*, *82,* 405-411. https://doi.org/10.1016/j.mehy.2013.11.030

McCrae, R. R., Terracciano, A., & Personality Profiles of Cultures Project. (2005a). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*, *89*(3), 407-425. https://doi.org/10.1037/0022-3514.89.3.407

McCrae, R. R., Terracciano, A., & Members of the Personality Profiles of Cultures Project. (2005b). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, *88*(3), 547-561. https://doi.org/10.1037/0022-3514.88.3.547

McKenna, K. Y. A., Green, A. S., & Gleason, M. E. J. (2002). Relationship formation on the internet: What's the big attraction? *Journal of Social Issues, 58*(1), 9-31. http://dx.doi.org/10.1111/1540-4560.00246

Mischel, W. (1968). *Personality and assessment*. NJ, USA: Lawrence Erlbaum Associates.

Montag, C., Duke, É., & Markowetz, A. (2016). Toward psychoinformatics: Computer science meets psychology. *Computational and Mathematical Methods in Medicine*, *2016*, 1-10. https://doi.org/10.1155/2016/2983685

Nolan, M. (2012). *How to separate fact and fiction online*. Retrieved from https://www.ted.com/talks/markham_nolan_how_to_separate_fact_and_fiction_online

Pariser, E. (2011). *Did Facebook's big study kill my filter bubble thesis?* Retrieved from https://backchannel.com/facebook-published-a-big-new-study-on-the-filter-bubble-here-s-what-it-says-ef31a-292da95

Pentland, A. (2009). Reality mining of mobile communications: Toward a new deal on data. In S. Dutta & I. Mia (Eds.), *The Global Information Technology Report 2008–2009*: *Mobility in a Networked World* (pp. 75-80). Geneva: World Economic Forum: INSEAD. Retrieved from http://www.weforum.org/pdf/gitr/2009/gitr09fullreport.pdf

Persily, N. (2017). Can democracy survive the Internet? *Journal of Democracy*, *28*(2), 63-76. http://doi.org/10.1353/jod.2017.0025

Pettit, M. (2016). Historical time in the age of big data: Cultural psychology, historical change, and the Google Books Ngram Viewer. *History of Psychology*, *19*(2), 141-153. https://doi.org/10.1037/hop0000023

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter profiles, our selves: Predicting personality with Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 180-185). Boston, MA, USA. http://doi.org/10.1109/PASSAT/SocialCom.2011.26

Regulation (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union,* European Parliament and European Council, 4.5.2016.

Rothstein, M. A. (2015). Ethical issues in big data health research: currents in contemporary bioethics. *The Journal of Law, Medicine & Ethics*, *43*(2), 425-429. https://doi.org/10.1111/jlme.12258

Saha, B., & Srivastava, D. (2014). Data quality: The other face of big data. In *2014 IEEE 30th International Conference on Data Engineering (ICDE).* Chicago, IL: IEEE, 1294-1297.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), e73791. http://doi.org/10.1371/journal.pone.0073791

Webb, A. (2013). We post nothing about our daughter online. *Slate.* Retrieved from http://www.slate.com/articles/technology/data_mine_1/2013/09/facebook_privacy_and_kids_don_t_post_photos_of_your_kids_online.html

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*(4), 41-78. http://doi.org/10.1111/j.1540-4560.1969.tb00619.x

Wolf, G. (2010). The data-driven life. *The New York Times*. Retrieved from http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html

WORLD VALUES SURVEY (2017, May 10). Retrieved from http://www.worldvaluessurvey.org/wvs.jsp

WORLD VALUES SURVEY Wave 6 2010-2014 OFFICIAL AGGREGATE v.20150418. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: Asep/JDS, Madrid SPAIN.