J Comput Soc Sc
https://doi.org/10.1007/s42001-017-0013-6

SURVEY ARTICLE

Big data in social and psychological science: theoretical and methodological issues

Lin Qiu[1] · Sarah Hian May Chan[1] · David Chan[2]

Received: 16 November 2017 / Accepted: 27 November 2017

Abstract Big data presents unprecedented opportunities to understand human behavior on a large scale. It has been increasingly used in social and psychological research to reveal individual differences and group dynamics. There are a few theoretical and methodological challenges in big data research that require attention. In this paper, we highlight four issues, namely data-driven versus theory-driven approaches, measurement validity, multi-level longitudinal analysis, and data integration. They represent common problems that social scientists often face in using big data. We present examples of these problems and propose possible solutions.

Keywords Big data · Computational social science · Psychology · Social science · Social media · Methodology

Introduction

It has been almost a decade since the field of computational social science was advocated in a paper published in *Science* [18]. Remarkable advances have been made in social and psychological sciences through the analysis of big data since then. For example, Kern et al. [15] revealed age-related language patterns by studying 20 million Facebook status updates. Youyou et al. [36] studied millions of Facebook likes and showed that likes are more predictive of personality traits and life outcomes than self-report measurements. With a staggering growth of digital

✉ Lin Qiu
linqiu@ntu.edu.sg

[1] School of Social Sciences, Nanyang Technological University, Singapore, Singapore

[2] Behavioural Sciences Institute, Singapore Management University, Singapore, Singapore

Published online: 05 December 2017

data every year and an estimation of 35 trillion gigabytes of digital footprints in 2020 [13], big data provides unprecedented opportunities to study human behavior. However, a number of theoretical and methodological challenges need to be addressed before ground-breaking discoveries can be made in the coming decade. In the following, we discuss four issues and present possible solutions.

Data-driven versus theory-driven approaches

Traditionally, researchers in social and psychological research make sense of empirical data using theory-driven approaches to explain phenomena (i.e., how things happen) rather than merely describe them (i.e., what has happened). The theory-based focus increases our understanding of causal relationships in psychological processes and underlying mechanisms of social phenomena. However, with the emergence of big data research where computer scientists often use data-driven methods such as machine learning, social scientists have started to adopt bottom-up data-driven approaches that favor prediction over explanation [35]. For example, Schwartz et al. [31] proposed an open-vocabulary differential language analysis (DLA) approach to predict personality from Faceboook status updates. This approach was used by Liu et al. [22] to predict the personality of Twitter users and identify features in Twitter profile pictures that predict personality. Kosinski et al. [16] and Youyou et al. [36] developed predictive models of personality traits using Facebook likes. These studies relied on machine-learning algorithms to choose variables in their models to improve prediction accuracy.

While the above studies demonstrate that social media can be used to accurately predict psychological attributes such as personality, their use of data-driven approaches may result in overfitting of the prediction model to the existing dataset (e.g., Facebook status updates) and yield poor performance on new datasets (e.g., tweets). For instance, Ginsberg et al. [10] used machine learning to choose 45 Google search terms from 50 million queries, and developed a prediction model that can accurately predict flu pandemics faster than the official disease control and prevention agency. However, researchers later found that the model completely missed nonseasonal influenza, suggesting that it predicted seasonality rather than the actual flu trend [19]. The failure of big data use in this exemplary case stresses the importance of using theories to guide the research design. If predictors were chosen based on theoretical relevance, seasonality would have been included in the model because it is well-known that seasonality is strongly associated with flu pandemics.

The practice of including theoretically relevant variables as predictors has been well established in social science research and should not be replaced or compromised by bottom-up data-driven approaches. Instead, the heterogeneity in big data allows researchers to include more theoretically relevant variables such as time, location, or population density than in traditional laboratory studies. For example, when using social media data to predict individual differences, it is important to control for variables such as location because many psychological characteristics have been found to be geographically clustered [30]. Although building models with variables of theoretical relevance may result in lower prediction accuracy than those developed using machine learning, it can provide meaningful explanation of the phenomena of interest and avoid overfitting the model.

Measurement validity

Traditional social science methods such as surveys and laboratory experiments allow researchers to carefully design their studies and determine how to measure variables of interest. However, in big data science, researchers often need to work with second-hand data collected by others such as social media services or mobile phone companies. There are three issues that can introduce measurement errors.

First, big data usually contains a large amount of noise. Researchers need to carefully examine the data and take multiple steps to remove such noise. For example, social media datasets often contain non-individual accounts such as spammers or news agencies. When studying individuals' language styles, researchers need to remove these non-individual accounts because they generate much more content than average users and can significantly bias the results. Researchers can use software programs such as spam detector to identify these accounts, or use traditional statistical methods to find outliners. After removing these accounts, any texts that are not written by the user (e.g., retweets, URLs, time stamps, and ads) needs to be removed because they do not reflect the language style of the user. While the above steps can reduce key sources of noise in the data, there could still be unforeseen noise remaining in the data.

Second, the software tools used to process the data may introduce measurement errors. For example, linguistic inquiry and word count (LIWC) is a widely used software tool to measure psychological processes from writing samples by counting word frequencies in pre-defined categories (e.g., positive affect; [32]). LIWC categories were developed based on psychological measurement scales and have been validated by independent judges [25]. However, LIWC may still generate inaccurate assessment that results in inaccurate interpretation of the data. For instance, because the word "great" belongs to the positive affect category, "a great amount of rain" will be categorized by LIWC as expressing positive emotion. Tov et al. [33] showed that LIWC coding of positive emotion did not reliably predict self-reported positive emotion in two diary studies. The inconsistency between machine-generated coding and self-reported measurement could be due to the error-prone results produced by the software.

Third, researchers in big data studies need to use proxies for their variables of interest. However, due to the lack of ground truth, it is unclear how accurately these proxies represent their corresponding variables. For example, emotional expressions on social media are often considered as proxies for users' actual emotional states in daily life (e.g., [11]). However, research has shown that users' online emotional expressions could be influenced by their impression management concerns and social network structure [20]. They selectively express more positive relative to negative emotions and present better emotional well-being on Facebook than in real life [26]. Therefore, users' emotional expressions online may not be a reliable measure of the frequency and valence of their actual emotional states. Empirical studies are needed to establish the validity of using online emotional expressions as a measurement of offline emotional states. For example, a diary study can be carried out to have participants report their emotional experiences every day, and compare their self reports with their Facebook status updates to estimate to what extent Facebook status updates reflect actual emotional experiences.

The three aforementioned issues inevitably produce measurement errors in big data research. They pose significant methodological and theoretical challenges. When findings from big data are inconsistent with existing theories, researchers cannot be sure if it is due to measurement errors or inherent problems in the theories. For example, when Liu et al. [21] found that positive emotional expressions online were not related to self-report life satisfaction, it is difficult to argue against past theories on the connection between life satisfaction and positive emotion because the inconsistency could very well be due to measurement errors.

One way to address the problem of measurement error is to conduct additional laboratory studies to validate the results from big data. For example, Doré et al. [8] analyzed tweets after the Sandy Hook Elementary School shooting, and found that spatial and temporal distance were positively associated with anger but negatively associated with sadness. They explained the associations using construal level theory, and conducted a follow-up laboratory study where abstract (vs. concrete) thinking was manipulated and showed to change emotional responses in the corresponding directions. Nai et al. [23] found that people in more racially diverse areas used more prosocial languages in their tweets. They validated the findings by showing supporting evidence from follow-up survey studies where people in more racially diverse neighborhoods were found to be more likely to offer help after a disaster and report having helped a stranger in the past month. These studies provide exemplar cases of how to complement big data research with traditional research methods.

Future studies may also use agent-based modeling (ABM; [14]) to validate the phenomenon observed in big data. ABM allows researchers to specify a theoretical model in a computer simulation and test if the simulation can generate the phenomenon observed in big data. For example, Gao et al. [9] developed an agent-based model showing how people's interpersonal communication styles and their acceptance to social influence may result in different patterns of opinion diffusion. This model may be used to generate simulated results to match geographic distributions of political preferences or brand shares found in big data. The match between simulated and empirical patterns can provide a logical explanation of how micro-level psychological processes and interpersonal communication lead to macro-level social phenomena.

Multi-level longitudinal analysis

Big data share a similar structure with traditional data in social and psychological research, which is one where the data are often longitudinal and hierarchical because they reflect the temporal and multilevel nature of the substantive phenomenon under study [1, 2, 4–6]. This provides great opportunities to study the interaction between individuals, organizations, and environments. However, current big data research mainly focuses on cross-sectional studies at the individual level. For example, studies have used big data to examine how individuals' temporal orientations are associated with their personality and well-being [24] and how political orientation affects subjective well-being [34]. A limited number of studies performed longitudinal analysis to examine the change of psychological processes. For instance, Golder and Macy [11] revealed individual-level diurnal and seasonal mood rhythms using millions of tweets across 84 countries. Liu et al. [21] showed that negative emotional expressions on Facebook within the past 9–10 months (but not beyond) predicted life satisfaction.

The use of big data should be maximized to explicate and test cross-levels interactions and inter-individual differences in intra-individual changes over time [1, 3, 5]. For example, big data about employees are hierarchical because each employee belongs to a team within a company. To understand how the mood of employees affects their company's performance, multi-level longitudinal analysis could be performed. Furthermore, there could be changes over time in an inherently cross-level construct such as person–group fit which is a composite construct involving two levels [3]. More importantly, the different facets of changes over time explicated by Chan [1] should be conceptualized and assessed. For example, any observed changes over time need to be decomposed into random fluctuations versus systematic changes in the focal variable. When systematic change over time exists, the trajectory of a variable may have time-varying correlates and the trajectory may affect or be affected by the trajectories of other variables, such that we need multivariate models that specify and test relationships linking changes in different focal variables. Finally, there may be between-group differences in one or more of the various facets of changes over time, and these groups may be observed groupings such as gender and culture groups or unobserved (or latent) groupings distinguishable by distinct characteristics of changes over time. Understanding the above complexities and the various facets of change over time, in terms of both the conceptual and methodological considerations, is necessary to make adequate substantive inferences from the longitudinal assessment of changes. Big data researchers can use advanced statistics models such as structural equation modeling and latent growth modeling to address the complexities involved in a variety of these changes and uncover the dynamics of social and psychological processes.

Data integration

Although existing big data studies often examine very large volumes of data, few studies have obtained and analyzed the *full* data (e.g., all the data on Facebook or Twitter). Researchers should try to analyze as much data as possible, because conclusions based on a subset or a particular type of data may be different from those obtained from the full data [7]. There are three challenges involved in obtaining and processing the full data.

First, due to privacy and proprietary concerns, organizations or companies rarely share their raw data even if they can be anonymized. Researchers often need to rely on a subset of data from a single data source. This greatly limits the ecological validity of the results. One possible solution is to use a divide-and-conquer approach proposed by Cheung and Jak [7]. For example, researchers can specify the data analysis procedure (e.g., regression, reliability tests, factor analysis, and multilevel analysis) and ask thousands of banks to run the procedure on their own consumer records. Each bank can then share the results of their analysis (e.g., regression coefficients, sampling covariance matrices), and researchers can use meta-analytic models to combine the results and estimate the effect size at the individual level. Such a divide-and-conquer approach allows researchers to perform analysis on data from different sources without accessing the raw data. It protects information privacy, and therefore, reduces the obstacles in data sharing.

Second, it is important to combine data from multiple platforms, because individuals often use multiple platforms and display different behaviors on each platform. For example, an individual may indicate “in a relationship” on Facebook but keep silent of her relationship status on Instagram. Combining data from Facebook and Instagram will allow researchers to fill in missing information and cross-validate user input. However, existing studies mainly rely on one single data source such as Facebook or Twitter due to the difficulty of matching users from different platforms. This makes findings less generalizable because each platform has its own unique characteristics. For example, auto-completing mechanisms on Facebook and Google may work differently and lead to different frequencies of user inputs [17]. Hodas and Lerman [12] showed that the differences in the position of messages on Twitter and Digg resulted in different user behaviors. Qiu et al. [28] recruited participants who used both Facebook and Renren (a platform similar to Facebook in China), and found that their sharing behaviors differed due to platform-related cultural norms. These studies stress the importance of using data from multiple sources to cross-validate findings and avoid over-generalization.

Third, big data includes a variety of information, including user-generated content (e.g., tweets, photos, and videos) and ‘digital shadow’ (e.g., purchase records, web-surfing histories, and location information collected by cell phones). Each data type contains unique behavioral cues. For example, texts signal linguistic styles, photos contain facial expressions, and videos reveal gestures and bodily movements. These behavioral cues reflect different aspects of psychological characteristics. For instance, comparing personality expressions in tweets [27] and selfies [29], extraversion was not reflected by cues in selfies but by frequency of positive emotion and social-related words in tweets. In contrast, conscientiousness was indicated by the absence of private location information in selfies, but not related to any cues in tweets. Therefore, to have a holistic view of human behavior, studies need to combine multiple data types. Wojcik et al. [34] analyzed texts from tweets and public speaking records, and also photos from LinkedIn profiles and public pictorial directories, to reveal a stable connection between political orientation and subjective well-being. The use of multiple data types allows researchers to examine behavioral patterns from different angles and increase the reliability of their findings. However, it also poses significant technical challenges, because researchers need to use a broad range of software tools and techniques in data collection and processing.

Conclusion

Big data presents unprecedented opportunities to understand human behavior on a large scale. They can reveal patterns of human behavior that are difficult to observe in laboratory studies, and provide ecological validity that traditional research oftentimes lacks. In this article, we highlighted four issues in the current practice of big data research, namely data-driven versus theory-driven approaches, measurement validity, multi-level longitudinal analysis, and data integration. They represent common problems that social scientists often face in using big data. Exemplar studies have been shown to provide possible solutions to these problems. They help researchers to avoid biases, improve validity, and maximize the use of big data.

References

1. Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal means and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1*(4), 421–483.
2. Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*(2), 234–246.
3. Chan, D. (2005). Current directions in personnel selection. *Current Directions in Psychological Science, 14*(4), 220–223.
4. Chan, D. (2010). Advances in analytical strategies. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology* (Vol. 1). Washington, DC: American Psychological Association.
5. Chan, D. (2013). Advances in modeling dimensionality and dynamics of job performance. In K. J. Ford, J. Hollenbeck, & A. M. Ryan (Eds.), *The psychology of work*. Washington, DC: American Psychological Association.
6. Chan, D. (2014). Time and methodological choices. In A. J. Shipp & Y. Fried (Eds.), *Time and work: How time impacts groups, organizations, and methodological choices* (Vol. 2, pp. 146–176). New York: Psychology Press.
7. Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7,* 738. https://doi.org/10.3389/fpsyg.2016.00738.
8. Doré, B., Ort, L., Braverman, O., & Ochsner, K. N. (2015). Sadness shifts to anxiety over time and distance from the national tragedy in Newtown. *Connecticut. Psychological Science, 26*(4), 363–373. https://doi.org/10.1177/0956797614562218.
9. Gao, W., Qiu, L., Chiu, C-y, & Yang, Y. (2015). Diffusion of opinions in a complex culture system: Implications for emergence of descriptive norms. *Journal of Cross-Cultural Psychology, 46*(10), 1252–1259.
10. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014.
11. 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.
12. Hodas, N. O., & Lerman, K. (2014). The simple rules of social contagion. *Scientific Reports, 4,* 4343. https://doi.org/10.1038/srep04343.
13. IDC. (2010, May). *The digital universe decade—Are you ready?* Retrieved Nov 23, 2017, from https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf.

14. Jackson, J. C., Rand, D., Lewis, K., Norton, M. I., & Gray, K. (2017). Agent-based modeling: A guide for social psychologists. *Social Psychological and Personality Science, 8*(4), 387–395. https://doi.org/10.1177/1948550617691100.
15. Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., et al. (2014). From “sooo excited!!!” to “so proud”: Using language to study development. *Developmental Psychology, 50,* 178–188.
16. 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, 110*(15), 5802–5805.
17. 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.
18. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721.
19. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science, 343*(6176), 1203.
20. Lin, H., Tov, W., & Qiu, L. (2014). Emotional disclosure on social networking sites: The role of network structure and psychological needs. *Computers in Human Behavior, 41,* 342–350.
21. Liu, P., Tov, W., Kosinski, M., Stillwell, D. J., & Qiu, L. (2015). Do Facebook status updates reflect subjective well-being? *Cyberpsychology, Behavior, and Social Networking, 18*(7), 373–379.
22. Liu, L., Preotiuc-Pietro, D., Riahi Samani, Z., Moghaddam, M. E., & Ungar, L. (2016). Analyzing personality through social media profile picture choice. In *Tenth international AAAI conference on web and social media.*
23. Nai, J., Narayanan, J., Hernandez, I., & Savani, K. (in press). People in more racially diverse neighborhoods are more prosocial. *Journal of Personality and Social Psychology.*
24. Park, G., Schwartz, H. A., Sap, M., Kern, M. L., Weingarten, E., Eichstaedt, J. C., et al. (2017). Living in the past, present, and future: Measuring temporal orientation with language. *Journal of Personality, 85*(2), 270–280. https://doi.org/10.1111/jopy.12239.
25. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. (LIWC.net, Austin, TX). Retrieved Nov 23, 2017, from www.liwc.net/LIWC2007LanguageManual.pdf.
26. Qiu, L., Lin, H., Leung, A. K.-Y., & Tov, W. (2012). Putting their best foot forward: Emotional disclosure on Facebook. *Cyberpsychology, Behavior, and Social Networking, 15*(10), 569–572. https://doi.org/10.1089/cyber.2012.0200.
27. Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you Tweet: Personality expression and perception on Twitter. *Journal of Research in Personality, 46*(6), 710–718.
28. Qiu, L., Lin, H., & Leung, A. K.-Y. (2013). Cultural differences and switching of in-group sharing behavior between an American (Facebook) and a Chinese (Renren) social networking site. *Journal of Cross-Cultural Psychology, 44*(1), 106–121.
29. Qiu, L., Lu, J., Yang, S., Qu, W., & Zhu, T. (2015). What does your selfie say about you? *Computers in Human Behavior, 52,* 443–449.
30. Rentfrow, P. J., & Jokela, M. (2016). Geographical psychology: The spatial organization of psychological phenomena. *Current Directions in Psychological Science, 25*(6), 393–398.
31. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791.
32. Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.
33. Tov, W., Ng, K. L., Lin, H., & Qiu, L. (2013). Detecting well-being via computerized content analysis of brief diary entries. *Journal of Personality Assessment, 25*(4), 1069–1078. https://doi.org/10.1037/a0033007.
34. Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M., & Ditto, P. (2015). Conservatives report, but liberals display, greater happiness. *Science, 347*(6227), 1243–1246.
35. Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393.
36. Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036–1040.