Big Data? Big Issues: Degradation in Longitudinal Data and Implications for Social Sciences

Matthew S. Weber
Rutgers University
4 Huntington St.
New Brunswick, NJ 08901
matthew.weber@rutgers.edu

Hai Nguyen
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854
hdn11@cs.rutgers.edu

ABSTRACT

This article analyzes the issue of degradation of data accuracy in large-scale longitudinal data sets. Recent research points to a number of issues with large-scale data, including problems of reliability, accuracy and quality over time. Simultaneously, large-scale data is increasingly being utilized in the social sciences. As scholars work to produce theoretically grounded research utilized "small-scale" methods, it is important for researchers to better understand the critical issues associated with the analysis of large-scale data. In order to illustrate the issues associated with this type of research, a case study analysis of archival Internet data is presented focusing on the issues of degradation of data accuracy over time. Suggestions for future studies are given.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]; H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Management, Measurement, Reliability, Verification.

Keywords

Keywords are your own designated keywords.

1. INTRODUCTION

Recent estimates project that there are more than 4 zettabytes of digital data produced per years [1], and that number has likely increased over the past two years. Digital data are produced by a variety of sources, from major research programs, to corporations, to individuals utilizing smart phones. Large sets of data are produced nearly every day, and increasingly large-scale data is being utilized for social science [2]. As large data sets are increasingly utilized in the social sciences, it is becoming more important to understand the implications and quality of large-scale data sets. Recent work has demonstrated that findings from large-scale data analysis may not always prove reliable [3]. Others have worked to understand the biases in the collection of large-scale data, pointing to many issues that exist with regards to data reliability [4]. This research outlines some of the key issues with regards to data reliability and large-scale data, and subsequently we focus on the issue of data degradation in dealing with large-scale longitudinal data sets. In the closing section, we provide suggestions for addressing these issues in future scholarship.

2. BIG DATA AND THE SOCIAL SCIENCES

Big Data, also referred to as large-scale data, is increasingly being utilized in the social sciences for research purposes. Large-scale data provide opportunities to explore previously untouched phenomena. For instance, large-scale data allow researchers to explore entire ecosystems of interaction, examining the way a large body of entities interacts with one another, while still being able to focus in on individual actors. Large-scale data is being utilized for a wide array of social science investigations, including research examining social movements [5], healthcare and health messaging [6, 7], collective action [8], news media [9, 10] and even as a tool for qualitative research [11]. Specific areas of focus include social media research, the utilization of big data in health, and the use of archival data to examine trends over time. Each is considered in brief in the following sections; these examples are intended to illustrate the intersection of Big Data and the social sciences.

2.1 Social Media Research

Social media provides fertile grounds for examining a wide array of issues in various domains of the social sciences. Social media data sources vary widely, but a significant portion of research in this space has focused on commonly used platforms such as Twitter.com and Facebook.com.Twitter data has been utilized to explore a number of mimetic processes focuses on the diffusion and replication of information [12, 13]. Much of that research has focused on political issues, including agenda setting and information spread [14]. Related studies have also been conducted analyzing blogging data [15]. Similarly, Twitter data has been utilized by a number of scholars to explore social movements and social coordination, including the organization by activists in the Occupy Wall Street movement. [8, 16, 17].

Facebook data has been utilized for similar types of research. The richer context available in Facebook data (via profiles, photos and other rich media) has often allowed for greater context in analysis, but issues with regards to data access have often created challenges for large-scale studies. Focusing on smaller subsets of Facebook data, researchers in communication have explored how identities are represented and shaped via Facebook profiles and posting [18]. Because Facebook is data is proprietary, it is often difficult to gain large-scale access. In one exception, a study conducted by Facebook explored an aspect of social influence by measuring peer-to-peer impact in driving voting behavior; the study had more than 60 million participants [19].

WebSci'15, Jun 28 – July 1, 2015, Oxford, UK.
Copyright 2015 ACM 978-1-4503-3672-7

2.2 Big Data and Health

Large-scale data has the potential to be used to better understand health issues and to improve a wide range of health outcomes; much recent work has focused on utilizing large-scale data to improve understanding of health outcomes, and to trace patient care practices [6]. In addition, large-scale data such as Twitter has been utilized to examine health campaign effectiveness, examining, for instance, the way that audiences accept or reject health messaging in the public sphere [7].

2.3 Archival Data

Finally, archival Internet data provides a glimpse back into a wide variety of phenomena of interest to scholars. For instance, archival Internet data allows for the examination of discourse over time, changes in media coverage, and shifts in political positions. Archival news media has been used to understand the transmission of news articles across different media platforms [20]. Scholars have additional utilized archival Internet data to recreate hyperlinking patterns of news media organizations online, and to assess changes over time [10, 21]. Data extracted from Internet archives have also been utilized to explore social movements and collective action [22]. A number of ongoing research initiatives are aimed at opening up research access to archival data sources, and it is expected that this will become a larger source of research data in the near future.

3. CENTRAL ISSUES OF BIG DATA

Despite the growing arena of social science research related to Big Data, scholars and practitioners alike recognize that many issues exist that must be addressed. One area of concern is the issue of data privacy. A number of scholars have critiqued research utilizing Big Data due to concerns regarding deidentification and participants' right to privacy [23, 24]. The size of datasets, the aggregated information, and the inability of participants to opt-out, all may create ethical issues for researchers [24]. In addition, content archived into large data sets main contain information on sensitive topics, but individuals often do not know that these records are being used for research [25]. For instance, with regards to large-scale data and healthcare, there are concerns that inaccurate analyses will lead to inaccurate messaging based on false conclusions [26]. Another issues for researchers focuses on scale and access. Large-scale data sets often exceed the size utilized by researchers in the social sciences. As a result, there has been a prolonged learning curve, and scholars seek to adapt methods to work with large-scale data [27, 28]. Hand-in-hand with methodological issues, researchers face challenges in identifying the desired subsets within archived data sources; often only a portion of data is actually needed, but access and filtering is problematic [29, 30].

4. DATA DEGRADATION

The above sections have outlined much of the promise that exists for research at the intersection of Big Data and the social sciences; in doing so, we believe that we have only scratched the surface. Moreover, we have also reviewed some of the more salient critiques. Much criticism to date, however, has focused on pointing to issues in Big Data – either ethical or methodological – without pointing to clear solutions or giving a clear path forward.

Our intent herein is to dissect a particular issue in large-scale data analysis, and in our presentation of this case study, we point to a number of practical solutions that may help to guide future research. Thus, we focus on the issue of data degradation, with an emphasis on large-scale longitudinal datasets. Much of the information contained in large-scale datasets pertains to longitudinal phenomena. For instance, the General Social Survey (GSS) conducted by the National Data Program for the Sciences (norc.org) as part of the International Social Survey Program contains 5,597 variables dating back to 1972. Some data sets, such as the GSS, are collected with a clear methodology, and a rigorous set of documentation to outline the approach.

On the other hand, much of the present wave of Big Data, including data sources extract from Facebook and Twitter, as well as archival data from sources such as the Internet Archive, are unstructured in nature, and are collected without a clearly established approach to data collection. The result is that such archived data sources often face issues of degradation as one parses back through time. Degradation of data can take a number of forms. For instance, one notion of degradation refers to the concept that data is lost due to technical failures and poor storage as one looks further back in time. For instance, scholars have noted that more than half of web page queries looking for historical information end up connecting to "dead" web pages, where information is no longer accessible [31].

4.1 Degradation and Data Completeness

In the context of this discussion, however, degradation is defined as the decreasing accuracy of archived digital data in relation to the historical age of the archived content. In other words, the completeness of an archived data set varies over time. With regards to both social media and archived Web data (and likely other datasets), part of the challenge is that the global population is often an unknown entity. Thus, it is near impossible to know the degree of accuracy of a subset of data. It is often possible to estimate for present-day data, but as one traverses back through a data set it becomes increasingly difficult to know the overall population, and therefore to estimate the accuracy of the subset.[1]

Degradation is known to be an issue in archived Internet data sources. For instance, one study suggests that the average life of a website is 3 years [32]. Given that there are billions of web pages available online, this clearly suggests significant turnover. Archival Internet data is also incomplete; crawlers must observe robots.txt records, struggle to handle dynamic web content, and are simply unable to fully capture the scope of the entire World Wide Web [33, 34]. Issues of degradation plague social media data as well. A study looking at three years of Twitter data found that after two years only 41% of the original content had been archived future use, and 27% of the data had already been lost to future users [35].

In order to address issues of degradation, a number of strategies have been suggested. For instance, researchers have proposed focusing archival efforts on capturing data that changes the most frequently, in order to capture the majority of new content [36]. Elsewhere, researchers have suggested that crawling strategies should prioritize archival efforts based on the size and relative position of websites within their larger ecosystems [37]. Such approaches are useful, and undoubtedly it is important for practitioners to pay close attention to improving crawling strategies in order to insure future accuracy, but these approaches do not provide guidance for addressing existing issues with data.

[1] In instances where a dataset represents a complete archive, this is not an issue. For instance, many social tagging applications have made entire longitudinal datasets available to researchers (see http://svn.citeulike.org for one example).

4.1.1 Why Degradation Matters

In part, degradation matters because the size of the sample matters. When seeking to draw conclusions about social behavior – whether one is discussing information diffusion or message effectiveness – a researcher must be able to specify the boundaries and accuracy of a dataset. For instance, a comparison of Twitter data collected via a public API and data collected from a "fire hose" provided by GNIP PowerTrack, found significant differences between the two datasets. In most cases the PowerTrack data proved to be more powerful, but there were also significant differences in the nuance of the data, which made the API preferable in some cases [4]. As the authors note, the choice of one data source over the other will have implications for research design, and such choices cannot be made without understanding the sampling of the data.

4.1.1.1 Statistical Power

In addition to descriptive issues, the sample population has important implications from a statistical perspective. For instance, data degradation has the potential to impact the degree of variance in a data source, and more importantly, to impact the sources of variance error. If there is a clear pattern of changes in the completeness of a crawl, there is in turn the potential for a patter of variance error. From a statistical perspective, this likely introduces a violation of the core assumption of homocedasticity that is made when dealing with linear regression. Thus, it is important to understand the completeness of a crawl, and to be able to defend or explain the sampling method in a given dataset in order to understand the degree of homocedasticity, and to correct for such issues. Moreover, other relations may exist that need to be better understood. By examining the completeness, or lack thereof, of the data, it is possible to better assess these issues.

5. CASE STUDY: DEGRADATION IN ARCHIVED DATA

5.1 Data Source

This case study utilized data from the Internet Archive (archive.org). The Internet Archive is considered the single largest source of archived Internet data, containing considerably more data than comparable sources such as the Library of Congress or the British Library's digital collections. In 2013, the Internet Archive contained more than seven petabytes of data; the organization's archives contain records dating from 1995 to present, and the archive continues to expand.

5.1.1 HistoryTracker

The data analyzed in this case study is part of the HistoryTracker project [38]. The HistoryTracker project, funded by the National Science Foundation, aims to build tools for social scientists to better access archived Internet data. The initial focus of the HistoryTracker project is on social network data; the current tools available allow researchers to extract social network data from the subsets of data from the Internet Archive.

The topic areas included in the subsets of data include United States-based media organizations, Occupy Wall Street, .GOV (Senate and House data), and natural disasters (focused on coverage of Hurricane Katrina and Superstorm Sandy). The datasets range in size from 1TB (Hurricane Katrina) to 20TB (US Media). The datasets were created by taking a curated list of seed uniform resource locators (URLs) and extracting all related URLs by crawling out from the seed list, and cover a period from 2000 onwards (each data set has its own specification).

In order to improve analytical efficiency, HistoryTracker leverages the Web Archive Transformation (WAT) format of the Web ARChive Record (WARC) files contained in the Internet Archive. This is, in essence, a meta-data format of meta-data. The advantage of utilizing the WAT files is that they contain key data such as hyperlinks and anchor text, but remove much of the coding and text not needed for this type of study.

5.2 Approach to Data Collection

The HistoryTracker tool crawls through a specified subset, and either extracts (a) all of the hyperlinks in the dataset or (b) data that matches against a pre-specified white list. The outputted linklist format contains data that specifies the connections that exist between webpages (specifying the date of the link, and the nature of the originating webpage, including the type of content, the size of the file, and any meta-text that may describe the page). The tool functions by ingesting a seed list (the either crawls all data, or extracts specified URLs), and matching against fields contained in the JSON format meta-data envelope associated with each record. To process data, the HistoryTracker tool utilizes Apache Pig, with queries in Pig Latin, compiled into low-level Map Reduce jobs.

5.2.1 The United States Senate

For the purposes of this case study, the HistoryTracker tool was utilized to analyze a subset of data tracking the web content of the United States Senate web domain (senate.gov). The senate.gov subset contains 26,965,770 captures of web pages representing 8,674,397 unique URLs. The crawl represents web pages for the 109th through the 112th US Congressional Sessions, with web pages covering dates from 2005 through 2012. The WAT file transformation compresses this dataset to 41GB, as opposed to approximately 1.2TB of WARC files.

5.2.2 Degradation in Senate.Gov Data

In order to analyze data degradation in the senate.gov data, it is necessary to compare the captured archive data against a known entity. Unfortunately, as with much archive data, it is not possible to know the exact scope of data in prior periods. On the other hand, the WAT file contains a record of all outgoing hyperlinks from a given webpage. Therefore, there is a known record in that each WAT file contains (a) a source URL, (b) a list of outbound URLs and (c) a record of captured URLs in the dataset.

Captured Date	Number of captured linked URL(s)	Number of actual linked URL(s)
1/1/09	2	19
11/18/2010	85	207
1/6/2011	77	207

Table 1. Sample of Data Degradation

Therefore, it is possible to compare the percent of captured URLs versus the percent of outbound (actual) URLs. The aggregate result is a measure of the completeness of a data subset – or at least a strong approximation. Table 1 presents a snapshot of three archived records for the URL http://www.senate.gov (as the root URL for this domain), and illustrates the difference between the captured URLs and the actual number of expected URLs. Thus, there is a measureable differences between the two datasets, where the difference is measured as:

$$\Delta_t = \text{count}(\text{actualURLs})_t - \text{count}(\text{capturedURLs})_t$$

Building out this approach, it is possible to replicate the analysis over the breadth of an entire dataset. Figure 1 illustrates the difference curve based on the comparison of 12 randomly selected subdomains of the senate.gov dataset.

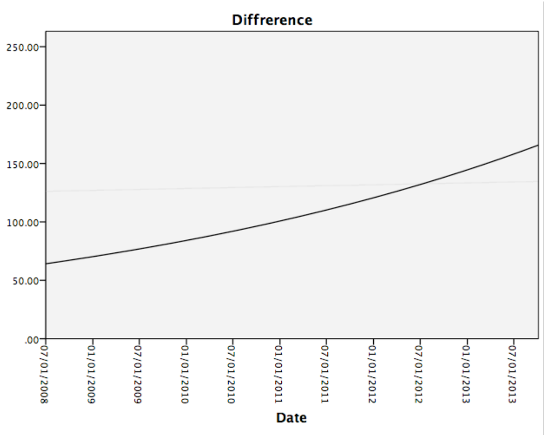

Figure 1. Difference Curve for Senate.Gov Sample

The curve estimation was generated using a standard curve estimation procedure. Curve fitting is not an ideal measure, but provides a suitable approximation. In the case of the senate.gov subsample, the curve was acceptable ($R^2 = 0.18$). The above curve is fit utilizing an exponential growth function ($y = c * e^{b}$). The curve fit of the senate.gov data yields a constant, c = 0.0029, and a growth function, $b = 0.0057$.

5.3 Utility for Research

The curve fit alone is not particularly interesting, as it is simply a representation of the data across time. The function, however, provides an extrapolated estimation of the degree of difference over time in a given sample. First, this provides an estimation of the amount of difference between the subset of data and the estimated actual data. In the case of the senate.gov data, this estimation shows that over time the amount of difference increases, suggesting that in this case as the domain increased the amount of archived data did not increased at the same pace. Ultimately, this would point to an increase in error over time when analyzing this dataset. There is, therefore, potential to correct for inaccuracies in data. Statistical corrections can be made to compensate for known errors in a dataset. The exact nature of a correction is dependent on the planned analysis.

6. FUTURE DIRECTIONS

The above discussion is a case study. There are clear limitations to this type of analysis, and this work is intended primarily to point to future directions. First, and foremost, the difference measure of degradation utilized above will be expanded to test other datasets. As the degradation of data analysis is expanded, this research will yield a robust analysis of the change in this dataset over time. Moreover, by fitting the degradation across datasets, it is possible to create a more accurate measure of degradation.

Beyond the limitations of the case study outlined above, this research attempts to move critique of Big Data research in the social sciences beyond criticism to implementation and correction. Research such as the recent work illustrating the faults of Google Flu Trends predictions point to the many pitfalls associated with statistical analysis and trend predication based on historical longitudinal data [3] (subsequent work has shown that other data sources may prove to be more reliable [39]). The research presented in this article picks up that gauntlet and move it forward, albeit a short distance. Our hope is that scholars at the intersection of computational research and social sciences will continue to address these challenges, and further the potential for Big Data in the social sciences.

7. ACKNOWLEDGMENTS

The authors wish to acknowledge support from the National Science Foundation (Grant #1244727), the NetSCI Network Science Research Lab at Rutgers University, as well as the Internet Archive (archive.org) in making this data available.

8. REFERENCES

[1] Tien, J. M. Big data: Unleashing information. *Journal of Systems Science and Systems Engineering*, 22, 2 2013), 127-151.

[2] Armstrong, K. Big data: a revolution that will transform how we live, work, and think. *Information, Communication & Society*, 17, 10 2014), 1300-1302.

[3] Lazer, D., Kennedy, R., King, G. and Vespignani, A. Big data. The parable of Google Flu: traps in big data analysis. *Science*, 343, 6176 (Mar 14 2014), 1203-1205.

[4] Driscoll, K., & Walker, S. Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, 8, 20 2014), 1745-1764.

[5] Driscoll, K., Ananny, M., Guth, K., Kazemzadeh, A., Leavitt, A. and Thorson, K. Big bird, binders, and bayonets: Humor and live-tweeting during the 2012 US presidential debates. *Selected Papers of Internet Research*, 32013).

[6] Chawla, N. V. and Davis, D. A. Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine*, 28 Suppl 3(Sep 2013), S660-665.

[7] Emery, S. L., Szczypka, G., Abril, E. P., Kim, Y. and Vera, L. Are you Scared Yet?: Evaluating Fear Appeal Messages in Tweets about the Tips Campaign. *Journal of Communication*, 64(Apr 2014), 278-295.

[8] Agarwal, S. D., Bennett, W. L., Johnson, C. N., & Walker, S. A model of crowd enabled organization: Theory and methods for understanding the role of twitter in the occupy protests. *International Journal of Communication*, 8, 27 2014), 646-672.

[9] Leetaru, K. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16, 9 2011).

[10] Weber, M. S. Newspapers and the Long-Term Implications of Hyperlinking. *Journal of Computer-Mediated Communication*, 17, 2 2012), 187-201.

[11] Bisel, R. S., Barge, J. K., Dougherty, D. S., Lucas, K. and Tracy, S. J. A Round-Table Discussion of "Big" Data in Qualitative Organizational Communication Research. *Management Communication Quarterly*, 28, 4 2014), 625-649.

[12] Jungherr, A. The Logic of Political Coverage on Twitter: Temporal Dynamics and Content. *Journal of Communication*, 64, 2 2014), 239-259.

[13] Park, J., Baek, Y. M. and Cha, M. Cross-Cultural Comparison of Nonverbal Cues in Emoticons on Twitter: Evidence from Big Data Analysis. *Journal of Communication*, 64, 2 2014), 333-354.

[14] Vargo, C. J., Guo, L., McCombs, M. and Shaw, D. L. Network Issue Agendas on Twitter During the 2012 U.S. Presidential Election. *Journal of Communication*, 64, 2 2014), 296-316.

[15] Ifukor, P. “Elections” or “selections”? Blogging and twittering the Nigerian 2007 general elections. *Bulletin of Science, Technology & Society*, 30, 6 2010), 398-414.

[16] Bennett, W. L. and Segerberg, A. Digital media and the personalization of collective action: Social technology and the organization of protests against the global economic crisis. *Information, Communication & Society*, 14, 6 2011), 770-799.

[17] Bruns, A., Highfield, T., & Burgess, J. The Arab Spring and Social Media Audiences: English and Arabic Twitter Users and Their Networks. *American Behavioral Scientist*, 57, 7 2013), 871-898.

[18] Dubrofsky, R. E. Surveillance on Reality Television and Facebook: From Authenticity to Flowing Data. *Communication Theory*, 21, 2 2011), 111-129.

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

[20] Leskovec, J., Kleinberg, J. and Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1, 1 2007), 1-42.

[21] Weber, M. S. and Monge, P. Industries in turmoil: Driving transformation during periods of disruption. *Communication Research*2014), 1-30.

[22] Bennett, W. L. Social movements beyond borders: understanding two eras of transnational activism. *Transnational protest and global activism*2005), 203-226.

[23] Crawford, K., Gray, M. L., Miltner, K. Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication*, 82014), 1663-1672.

[24] Fairfield, J. and Shtein, H. Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism. *Journal of Mass Media Ethics*, 29, 1 2014), 38-51.

[25] Trevisan, F. and Reilly, P. Ethical dilemmas in researching sensitive issues online: lessons from the study of British disability dissent networks. *Information, Communication & Society*, 17, 9 2014), 1131-1146.

[26] Fallik, D. For big data, big questions remain. *Health affairs*, 33, 7 (Jul 2014), 1111-1114.

[27] Kaisler, S., Armour, F., Espinosa, J. A. and Money, W. *Big Data: Issues and Challenges Moving Forward*. City, 2013.

[28] Manovich, L. *Trending: the promises and the challenges of big social data*. University of Minnesota Press, City, 2011.

[29] Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. The meaningful use of big data: four perspectives--four challenges. *ACM SIGMOD Record*, 40, 4 2012), 56-60

[30] Busch, L. A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large-Scale Data Sets. *International Journal of Communication*, 82014), 1727-1744.

[31] AlNoamany, Y., AlSum, A., Weigle, M. C. and Nelson, M. L. Who and what links to the Internet Archive. *International Journal on Digital Libraries*, 14, 3-4 2014), 101-115.

[32] Agata, T., Miyata, Y., Ishita, E., Ikeuchi, A. and Ueda, S. *Life span of web pages: A survey of 10 million pages collected in 2001*. City, 2014.

[33] Ainsworth, S. G., Alsum, A., SalahEldeen, H., Weigle, M. C. and Nelson, M. L. How much of the web is archived? In *Proceedings of the Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (Ottawa, Ontario, Canada, 2011). ACM, [insert City of Publication],[insert 2011 of Publication].

[34] McKay, C. Ephemeral to enduring: the Internet Archive and its role in preserving digital media. *Information Technology and Libraries*, 23, 1 2004), 3.

[35] SalahEldeen, H. M. and Nelson, M. L. *Losing my revolution: how many resources shared on social media have been lost?* Springer, City, 2012.

[36] Spaniol, M., Denev, D., Mazeika, A., Weikum, G. and Senellart, P. Data quality in web archiving. In *Proceedings of the Proceedings of the 3rd workshop on Information credibility on the web* (Madrid, Spain, 2009). ACM, [insert City of Publication],[insert 2009 of Publication].

[37] Song, R., Liu, H., Wen, J.-R. and Ma, W.-Y. *Learning block importance models for web pages*. ACM, City, 2004.

[38] Weber, M. S. Observing the Web by Understanding the Past: Archival Internet Research. *WWW'14 Companion Proceedings*2014).

[39] Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y. and Priedhorsky, R. Global disease monitoring and forecasting with wikipedia. *PLoS computational biology*, 10, 11 2014), e1003892.