PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

The Virtuous Cycle of a Data Ecosystem

Bradley Voytek*

Department of Cognitive Science, Neurosciences Graduate Program, Institute for Neural Computation, and Kavli Institute for Brain and Mind, University of California, San Diego, California, United States of America

* bradley.voytek@gmail.com

Overview

Modern science is creating data at an unprecedented rate, yet most of these data are being discarded. Raw scientific data, when they are published at all, are provided in a very limited form. Large, multidimensional datasets—rich with hidden information—are reduced to summary statistics filtered through limitations imposed by contemporary methods and technologies, and through the biased lens of the originating research group. The massive loss of raw data currently underway, and the lack of a system for discovering them, hinders scientific progress. In this Perspective, I argue that our contemporary limited view of the long-term scientific and medical benefits that could be made possible by data sharing masks the benefits for doing so. This, in turn, makes the costs of data sharing seem higher than they are.

Citation: Voytek B (2016) The Virtuous Cycle of a Data Ecosystem. PLoS Comput Biol 12(8): e1005037. doi:10.1371/journal.pcbi.1005037

Editor: Philip E. Bourne, National Institutes of Health, UNITED STATES

Published: August 4, 2016

Funding: This work was supported by the UCSD Qualcomm Institute Calit2 Strategic Research Opportunities program and a Sloan Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author declares that no competing interests exist.

Introduction

Digital data of all types are being created at an ever-increasing rate, doubling approximately every two years. Annual data creation rates are estimated to reach 44 trillion gigabytes by 2020 [1]. Similarly, the rate at which primary scientific data are being collected is accelerating [2]. This astounding growth in scientific data creation has led to the contemporary discussion of scientific data sharing policies. Many of the criticisms levied against data sharing have focused on practical issues such as the economics and logistics of data storage, technical challenges for doing so, or appropriate attribution of credit [2–9]. In contrast, the arguments in favor of data sharing have focused largely on scientific replication, reproducibility [10], facilitation of collaborative research, and increased citations for publications that share data [11]. This is largely an ethical argument wherein there is an obligation to share data collected using public funds [3–6,12,13].

Rather than focusing on the much-discussed arguments against data sharing—cost, infrastructure, curation, privacy, and attribution/credit concerns—in this Perspective, I outline the overlooked benefits of data sharing: novel remixing and combining as well as bias minimization and meta-analysis. I argue that we must consider the weight of the costs against the true value of the possible benefits. If the decision for any individual researcher, university, or funding agency to implement data sharing policies comes down to a cost—benefit analysis based solely on replication versus storage, the cost—benefit analysis may be artificially tipped in favor of not sharing data caused by overlooking more subtle—but critical—benefits. These hidden benefits of data remixing cannot be appreciated when considering each individual dataset as an independent entity, and thus a richer consideration of those benefits is warranted.

Although there is some evidence that, on the local scale, research groups may not make use of shared data [14], in this Perspective, I outline the ways in which research groups are beginning to take advantage of open data in novel, and sometimes surprising, ways. Rather than

arguing for a centralized, large-scale data repository, I am advocating for a more organic development wherein we, institutionally, encourage the growth of a data ecosystem. This can be done via multiple venues, such as the general scientific data sharing sites figshare (https://figshare.com/) or the Dryad Digital Repository (http://datadryad.org/), each of which, in addition to Nature Publishing Group's recently launched peer-reviewed data sharing journal, *Scientific Data* [15], provides citable Digital Object Identifiers for the data themselves. Such developments are addressing concerns regarding credit and help motivate data curation and contextualization. A data sharing ecosystem provides space for multiple diverse datasets to intermingle to encourage new, multidisciplinary discoveries for current and future scientists.

Data Sharing Benefits

Data remixing and combining

One of the potentially most powerful yet underrated benefits of releasing data is the opportunity to reanalyze older data using contemporary methods. There are countless examples of data (broadly construed) being used in novel ways to generate new insights in domains far removed from their original source. Below, I cite four general cases.

1. Reanalyzing old data using new methods. Exoplanets were discovered in decades-old data collected by the Hubble Space Telescope [16]; 19th century naval logbooks were used to extract weather data to model climate change [17]; epigenetic changes in DNA methylation were identified as a function of prenatal exposure to famine as documented by health records preserved from the 1944–45 Dutch Hunger Winter [18]; ink traces of electrophysiological data collected from the human cerebellum in the 1930s and 1940s were digitized and analyzed using modern methods to uncover novel functions of this brain region [19].

2. Text mining for scientific discovery. Text was extracted from millions of books published across hundreds of years to model language evolution and cultural phenomena [20,21]; freeform text from patients writing in online forums was analyzed to aid in clinical discovery [22]; online food recipes were used to uncover cultural taste preferences [23].

3. Data remixing and combination. Data from studies in archeology, criminology, economics, geography, history, political science, and psychology were used to analyze the effect of climate on human conflict [24]; neuroscientific textual information from millions of peer-reviewed papers was compared against human brain gene expression data to identify brain structure, function, and disease relationships [25]; spatial information about the functional relationships of the human brain, as mined from thousands of peer-reviewed papers, was combined with spatial information on human gene expression data to identify novel gene—cognition relationships [26].

4. Semi-automated, or algorithmic, hypothesis generation. Neuronal electrophysiological data were aggregated to study neural diversity [27,28]; research maps of experimental results were created to extract the weight of evidential support or results [29]; possible novel hypotheses were uncovered by analyzing missing connections between scientific topics [25,26,30].

This last point—semi-automated or algorithmic hypothesis generation—has enormous potential to speed scientific discovery. Hypothesis-generation algorithms thrive in an environment rich with independent data sources. The above examples all come from the neurosciences, a field that poses unique challenges for data mining [31]. These projects represent largely independent, parallel efforts operating at different conceptual scales ranging from subcellular to psychological. As more neuroscientific datasets become available, it will become increasingly possible to statistically link multiple domains, including gene expression [32], neural diversity [28], functional neuroimaging [33], neural activity [34], and cognition [35]. Once

these datasets can be aligned in a common format, hypothesis generation algorithms can be deployed to identify candidate links between genes, neural activity, cognition, and disease.

Bias minimization and meta-analysis

Another benefit of large-scale data availability is that it could uncover sampling bias by allowing researchers to combine data from multiple studies. For example, sampling bias is rampant in psychology, in which 96% of studies published from the top six psychology journals consisted of data collected from people living in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies [36]. Furthermore, many datasets, both human [36,37] and rodent [38], are biased in their gender sampling, calling into question the generalizability of many biomedical findings. By combining data from sources collected from animals of different ages and genders, or people from different cultures, the generalizability of the results can be assessed.

Similarly, unless raw data are shared, access to them is limited to those who collected it (and their collaborators). Given that the vast majority of scientific research is conducted by industrialized societies, this limits the interpretation of those data through a narrower cultural lens. There is ample evidence that culture at all levels affects data collection and interpretation, ranging from the "publish or perish" culture of modern academic science biasing what results are published to larger, more macroscale political and social influences in how findings are contextualized [39–41].

One way of minimizing bias is through meta-analysis. However, these analyses, wherein the results of many peer-reviewed studies are aggregated, are limited by the massive data reduction that results from reporting summary statistics. This data reduction—taking a rich, multivariate dataset and summarizing it for publication using measures of central tendency, confidence intervals, *p*-values, and effect sizes—removes the opportunity for future scientists to apply new algorithms, methods, and transdisciplinary ideas that could yield unforeseen insights and discoveries [42,43]. This is because future reanalyses of existing data are restricted to looking only at whatever summary statistics the authors decided to include in their original manuscripts. Given that the majority of raw scientific data are reported to be inaccessible or lost [44], future opportunities to put historical results in context are limited.

Thus, it is important to ensure that data are discoverable and that access to these data be open—similar to the current PubMed search engine and PubMed Central manuscript repository—to limit the currently large digital cultural divide [37]. Closing this divide allows access to those who may not have sufficient resources to run large-scale experiments on their own. It also opens up the opportunity for broader interpretation and contextualization of those data, as well as democratization of the scientific process through citizen science, which has proved to be a highly successful model such as Foldit [45], EyeWire [46], and Galaxy Zoo [47].

Conclusion

Modern science is massive in scale; the data we are generating are evidence of our advancing knowledge. The simultaneous growth of data collection techniques [48] along with data aggregation and mining algorithms [49,50] provides an unprecedented opportunity for rapid knowledge discovery [51]. We cannot know what other discoveries lay hidden in our data, similar to how even the most innocuous-seeming scientific results can lead to important breakthroughs. To give but a few examples of this: studying monkey social behaviors and eating habits led to insights into the origins of HIV [52]; research into how algae move toward light paved the way for optogenetics—using light to control neural activity [53]; and black hole research spurred the development of algorithms eventually used as part of the 802.11 specifications ubiquitously

used in modern Wi-Fi [54]. The ideas spawned from the above projects (and countless others) could never have been anticipated. They cut across broad research domains well outside their original fields. However, the possibility for a breakthrough can't exist if we base our decision-making on the immediately obvious and predictable outcomes.

Of course, there are concerns for sharing data, and privacy and consent issues surrounding the sharing of human data are complex [55]. Privacy issues are compounded by the fact that even data that have been de-identified can be re-identified [56], so care must be taken to ensure individual privacy until de-identification has been proved to be secure. Nevertheless, encouraging the growth of a data ecosystem should be a priority among scientists. By basing the decision of whether or not to share data solely on whether replication and reproducibility is worth the cost of curation and storage, we are limiting the opportunities for future scientists to make novel use of our data in ways that we could never predict. By sharing the raw data, we can create a virtuous cycle that allows researchers to remix and reanalyze data in new and interesting ways. It is our duty to preserve our data so that future generations will not be hindered by our prejudiced interpretations and analytical limitations.

Acknowledgments

I would like to thank Scott Cole, Tom Donoghue, Richard Gao, Roemer van der Meij, Torben Noto, Erik Peterson, Tam Tran, and Shreejoy Tripathy for their comments, edits, criticisms, advice, and mockery regarding previous versions of this manuscript.

References

1. Turner V, Gantz JF, Reinsel D, Minton S. The digital universe of opportunities: Rich data and the increasing value of the internet of things. International Data Corporation; 2014.
2. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. Neylon C, editor. PLoS ONE. 2011; 6: e21101–21. doi: 10.1371/journal.pone.0021101 PMID: 21738610
3. Berman F, Cerf V. Who Will Pay for Public Access to Research Data? Science. 2013; 341: 616–617. doi: 10.1126/science.1241625 PMID: 23929969
4. Sterling TD, Weinkam JJ. Sharing scientific data. Communications of the ACM. 1990.
5. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics—re-shaping scientific practice. Nat Rev Genet. 2009; 10: 331–335. doi: 10.1038/nrg2573 PMID: 19308065
6. Koslow SH. Sharing primary data: a threat or asset to discovery? Nat Rev Neurosci. Nature Publishing Group; 2002; 3: 311–313.
7. Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE, Martone ME. Big data from small data: data-sharing in the "long tail" of neuroscience. Nat Neurosci. Nature Publishing Group; 2014; 17: 1442–1447. doi: 10.1038/nn.3838
8. Carroll MW. Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol. 2015; 13: e1002235–11. doi: 10.1371/journal.pbio.1002235 PMID: 26313685
9. Ribbon B. The Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2008). Interim Report. 2008.
10. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. SCIENTIFIC STANDARDS. Promoting an open research culture. Science. 2015; 348: 1422–1425. doi: 10.1126/science.aab2374 PMID: 26113702
11. Pienta AM, Alter GC, Lyle JA. The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. 2010.
12. Soranno PA, Cheruvelil KS, Elliott KC, Montgomery GM. It"s Good to Share: Why Environmental Scientists" Ethics Are Out of Date. BioScience. 2014; 65: 69–73. doi: 10.1093/biosci/biu169 PMID: 26955073
13. Duke CS, Porter JH. The Ethics of Data Sharing and Reuse in Biology. BioScience. Oxford University Press; 2013; 63: 483–489. doi: 10.1525/bio.2013.63.6.10

14. Wallis JC, Rolando E, Borgman CL. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. Nunes Amaral LA, editor. PLoS ONE. 2013; 8: e67332–17. doi: 10.1371/journal.pone.0067332 PMID: 23935830

15. Launch of an online data journal. Nature. 2013; 502(7470): 142. http://www.nature.com/news/announcement-launch-of-an-online-data-journal-1.13906.

16. Soummer R, Hagan JB, Pueyo L. Orbital motion of HR 8799 b, c, d using Hubble Space Telescope data from 1998: constraints on inclination, eccentricity, and stability. The Astrophysical . . .. 2011.

17. Allan R, Brohan P, Compo GP, Stone R, Luterbacher J, Broennimann S. The International Atmospheric Circulation Reconstructions over the Earth (ACRE) Initiative. Bulletin of the American Meteorological Society. 2011; 92: 1421–1425. doi: 10.1175/2011BAMS3218.1

18. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci USA. National Acad Sciences; 2008; 105: 17046–17049. doi: 10.1073/pnas.0806560105

19. Dalal SS, Osipova D, Bertrand O, Jerbi K. Oscillatory activity of the human cerebellum: The intracranial electrocerebellogram revisited. Neurosci Biobehav Rev. 2013; 37: 585–593. doi: 10.1016/j.neubiorev.2013.02.006 PMID: 23415812

20. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 2011; 331: 176–182. doi: 10.1126/science.1199644 PMID: 21163965

21. Lieberman E, Michel J-B, Jackson J, Tang T, Nowak MA. Quantifying the evolutionary dynamics of language. Nature. 2007; 449: 713–716. doi: 10.1038/nature06137 PMID: 17928859

22. Wicks P, Vaughan TE, Massagli MP, Heywood J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. Nat Biotechnol. 2011; 29: 411–414. doi: 10.1038/nbt.1837 PMID: 21516084

23. Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L. Flavor network and the principles of food pairing. Sci Rep. 2011; 1. doi: 10.1038/srep00196

24. Hsiang SM, Burke M, Miguel E. Quantifying the Influence of Climate on Human Conflict. Science. 2013; 341: 1235367–1235367. doi: 10.1126/science.1235367 PMID: 24031020

25. Voytek JB, Voytek B. Automated cognome construction and semi-automated hypothesis generation. Journal of Neuroscience Methods. 2012; 208: 92–100. doi: 10.1016/j.jneumeth.2012.04.019 PMID: 22584238

26. Fox AS, Chang LJ, Gorgolewski KJ, Yarkoni T. Bridging psychology and genetics using large-scale spatial analysis of neuroimaging and neurogenetic data. bioRxiv. 2014. doi: 10.1101/012310

27. Tripathy SJ, Savitskaya J, Burton SD. NeuroElectro: a window to the world's neuron electrophysiology data. Frontiers in . . .. 2014.

28. Tripathy SJ, Burton SD, Geramita M, Gerkin RC, Urban NN. Brain-wide analysis of electrophysiological diversity yields novel categorization of mammalian neuron types. Journal of Neurophysiology. 2015;: jn.00237.2015–40. doi: 10.1152/jn.00237.2015

29. Landreth A, Silva AJ. The Need for Research Maps to Navigate Published Work and Inform Experiment Planning. Neuron. 2013; 79: 411–415. doi: 10.1016/j.neuron.2013.07.024 PMID: 23931992

30. Poldrack RA, Mumford JA, Schonberg T, Kalar D, Barman B, Yarkoni T. Discovering Relations Between Mind, Brain, and Mental Disorders Using Topic Mapping. PLoS Comput Biol. 2012; 8: e1002707. doi: 10.1371/journal.pcbi.1002707.t002 PMID: 23071428

31. Akil H, Martone ME, Van Essen DC. Challenges and opportunities in mining neuroscience data. Science. American Association for the Advancement of Science; 2011; 331: 708–712. doi: 10.1126/science.1199305

32. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. Nature Publishing Group; 2012; 489: 391–399. doi: 10.1038/nature11405

33. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. Large-scale automated synthesis of human functional neuroimaging data. Nature Methods. 2011; 8: 665–670. doi: 10.1038/nmeth.1635 PMID: 21706013

34. Niso G, Rogers C, Moreau JT, Chen L-Y, Madjar C, Das S, et al. OMEGA: The Open MEG Archive. NeuroImage. Elsevier Inc; 2015;: 1–6. doi: 10.1016/j.neuroimage.2015.04.028

35. Poldrack RA, Kittur A, Kalar D, Miller E, Seppa C, Gil Y, et al. The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. Front Neuroinform. Frontiers; 2011; 5. doi: 10.3389/fninf.2011.00017

36. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world. Behav Brain Sci. Cambridge Univ Press; 2010; 33: 61–83.

37. Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information. 2012. doi: 10.1080/1369118X.2012.678878

38. Beery AK, Zucker I. Sex bias in neuroscience and biomedical research. Neurosci Biobehav Rev. 2011; 35: 565–572. doi: 10.1016/j.neubiorev.2010.07.002 PMID: 20620164

39. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. Public Library of Science; 2005; 2: e124. doi: 10.1371/journal.pmed.0020124

40. MacCoun RJ. Biases in the interpretation and use of research results. Annu Rev Psychol. 1998; 49: 259–287. doi: 10.1146/annurev.psych.49.1.259 PMID: 15012470

41. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials. 1998; 19: 159–166. doi: 10.1016/S0197-2456(97)00150-5 PMID: 9551280

42. Carney SL. Leroy Hood expounds the principles, practice and future of systems biology. Drug Discovery Today. 2003; 8: 436–438. doi: 10.1016/S1359-6446(03)02710-7 PMID: 12801791

43. Horn JDV, Grafton ST, Rockmore D, Gazzaniga MS. Sharing neuroimaging studies of human cognition. Nat Neurosci. 2004; 7: 473–481. doi: 10.1038/nn1231 PMID: 15114361

44. Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, et al. The Availability of Research Data Declines Rapidly with Article Age. Current Biology. Elsevier Ltd; 2014; 24: 94–97. doi: 10.1016/j.cub.2013.11.014

45. Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, Shen BW, et al. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat Biotechnol. Nature Publishing Group; 2012; 30: 190–192. doi: 10.1038/nbt.2109

46. Kim JS, Greene MJ, Zlateski A, Lee K, Richardson M, Turaga SC, et al. Space-time wiring specificity supports direction selectivity in the retina. Nature. Nature Publishing Group; 2014; 509: 331–336. doi: 10.1038/nature13240

47. Land K, Slosar A, Lintott C, Andreescu D, Bamford S, Murray P, et al. Galaxy Zoo: the large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey ★. Monthly Notices of the Royal Astronomical Society. 2008; 388: 1686–1692. doi: 10.1111/j.1365-2966.2008.13490.x

48. Brown HR, Zeidman P, Smittenaar P, Adams RA, McNab F, Rutledge RB, et al. Crowdsourcing for cognitive science—the utility of smartphones. PLoS ONE. 2014; 9: e100662. doi: 10.1371/journal.pone.0100662 PMID: 25025865

49. Yarkoni T. Psychoinformatics New Horizons at the Interface of the Psychological and Computing Sciences. Current Directions in Psychological Science. SAGE Publications; 2012; 21: 391–397. doi: 10.1177/0963721412457362

50. Yarkoni T, Poldrack RA, Van Essen DC, Wager TD. Cognitive neuroscience 2.0: building a cumulative science of human brain function. Trends in Cognitive Sciences. Elsevier; 2010; 14: 489–496. doi: 10.1016/j.tics.2010.08.004

51. Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. Intelligent Systems, IEEE. IEEE; 2009; 24: 8–12. doi: 10.1109/MIS.2009.36

52. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, et al. Hybrid origin of SIV in chimpanzees. Science. 2003; 300: 1713–1713. doi: 10.1126/science.1080657 PMID: 12805540

53. Zhang F, Vierock J, Yizhar O, Fenno LE, Tsunoda S, Kianianmomeni A, et al. The Microbial Opsin Family of Optogenetic Tools. Cell. Elsevier Inc; 2011; 147: 1446–1457. doi: 10.1016/j.cell.2011.12.004

54. Hamaker JP, O'Sullivan JD, Noordam JE. Image sharpness, Fourier optics, and redundant-spacing interferometry. J Opt Soc Am, JOSA. Optical Society of America; 1977; 67: 1122–1123. doi: 10.1364/JOSA.67.001122

55. Currie J. "Big data" versus "big brother": on the appropriate use of large-scale data collections in pediatrics. Pediatrics. 2013; 131 Suppl 2: S127–32. doi: 10.1542/peds.2013-0252c PMID: 23547056

56. McGuire AL, Gibbs RA. Genetics. No longer de-identified. Science. 2006; 312: 370–371. doi: 10.1126/science.1125339 PMID: 16627725