Please cite this article in press as: Voytek, Social Media, Open Science, and Data Science Are Inextricably Linked, Neuron (2017), https://doi.org/10.1016/j.neuron.2017.11.015

Neuron
NeuroView

Social Media, Open Science, and Data Science Are Inextricably Linked

Bradley Voytek[1,*]
[1]Department of Cognitive Science, Neurosciences Graduate Program, Halicioglu Institute for Data Science, University of California, San Diego, CA 92093-0515, USA
*Correspondence: bradley.voytek@gmail.com
https://doi.org/10.1016/j.neuron.2017.11.015

Should scientists use social media? Why practice open science? What is data science? Ten years ago, these phrases hardly existed. Now they are ubiquitous. Here I argue that these phenomena are inextricably linked and reflect similar underlying social and technological transformations.

Something Is Changing in Science

On the morning of September 14, 2015, a 200-ms "chirp" was detected by the Laser Interferometer Gravitational-Wave Observatory (LIGO) laboratories in Hanford, Washington, and Livingston, Louisiana. Just over 2 years later, this "chirp"—the first experimental confirmation of the existence of gravitational waves—earned Rainer Weiss, Barry Barish, and Kip Thorne the Nobel Prize in Physics. During the intervening 2 years, the LIGO laboratories, which cost over $600 million to build, collected more than 4.5 petabytes of data.

In January 2009, just as LIGO was getting a major technological upgrade, Fields Medalist Tim Gowers wrote a post on his personal blog wherein he invited his readers to help him solve a problem in his field of mathematics. According to author Michael Nielsen in his book *Reinventing Discovery*,

> …over the next 37 days, 27 people wrote 800 mathematical comments, containing more than 170,000 words. Reading through the comments you see ideas proposed, refined, and discarded, all with incredible speed. You see top mathematicians making mistakes, going down wrong paths, getting their hands dirty following up the most mundane of details, relentlessly pursuing a solution. And through all the false starts and wrong turns, you see a gradual dawning of insight.

In the comments of Gowers' blog, and the blog of fellow Fields medalist Terence Tao, the problem was eventually solved. Now dubbed "The Polymath Project," Gowers reflected that:

> …something I found more striking than the opportunity for specialization of this kind was how often I found myself having thoughts that I would not have had without some chance remark of another contributor. I think it is mainly this that sped up the process so much.

Echoing the power of collaboration, when reacting to the recent announcement of winning the Nobel Prize, the *New York Times* noted that, "Dr. [Rainer] Weiss said that he considered the [Nobel Prize] as recognition for the work of about a thousand people over 'I hate to say it—40 years.'" Similarly, Dr. Thorne "said that as the resident theorist and evangelist on the project he felt a little embarrassed to get the prize. 'It should go to all the people who built the detector or to the members of the LIGO-Virgo Collaboration who pulled off the end game.'"

While both LIGO and the Polymath Project involved prize-winning researchers, in most other respects, they could not appear to be more different: one is a big data technological marvel that cost hundreds of millions of dollars to construct; the other was borne out in the comments sections of personal blogs. Despite the surface-level differences in scale, the "comments sections of two personal blogs" are *themselves a technological marvel*, albeit one that is easily trivialized.

In this piece, I argue that several major trends in modern science—social media; open science, reproducibility, and data sharing; and data science and big data—are not distinct, separable phenomena; rather they are inextricably linked and reflect the same underlying social and technological transformations. That is, none can exist without the others; it is no coincidence that so many major technological events, each of which influenced scientific practice, occurred within such a short, 5-year time frame (Table 1).

Though an incomplete accounting, each of the events listed in Table 1 marks a significant change in the scientific/technological landscape. I group these changes into three categories, discussed in detail below: Social Media, Open Science, and Data Science.

Should Scientists Use Social Media?

Despite existing for barely more than a decade (Table 1), major social media services such as Facebook and Twitter, which opened to the general public in 2006 and 2007, respectively, have significantly shaped the nature of social, political, and scientific discourse. While much has been said regarding whether or not scientists should "use" social media—and how they should do so—what is becoming more evident is that *social media can use scientists*, whether they wish to be involved or not.

By this I mean that post-publication review of research can occur in the public sphere with or without the participation of the primary researchers (Faulkes, 2014) and that this can and will be done by anonymous supporters and critics (Neuroskeptic, 2013). While frustrating

Please cite this article in press as: Voytek, Social Media, Open Science, and Data Science Are Inextricably Linked, Neuron (2017), https://doi.org/10.1016/j.neuron.2017.11.015

Table 1. Abbreviated Timeline Highlighting Recent Transformations in Scientific Practice

Year	Event
2003	Open Access publishing is boosted with *PLOS Biology* launch
2004	Google's Dean & Ghemawat publish MapReduce in *OSDI*
2005	Amazon launches Mechanical Turk
2005	Reproducibility goes critical with Ioannidis's "Why Most Published Research Findings Are False" in *PLOS Medicine*
2006	Facebook opens account creation to the general public
2006	*PLOS One* launches with a central goal of facilitating post-publication peer review
2006	Amazon launches Elastic Compute Cloud (EC2) as part of Amazon Web Services (AWS)
2006	Netflix announces the Netflix Prize competition
2007	Twitter becomes independent and "debuts" at South by Southwest (SXSW)
2007	iPhone launch sparks the "smartphone revolution"
2007	iPython changes scientific computing when introduced in Pérez & Granger, *Computing in Science and Engineering*
2008	GitHub launch makes scientific version control easier
2008	DJ Patil (LinkedIn) and Jeff Hammerbacher (Facebook) coin the phrase "Data Science" to describe their jobs

for many, the fact is that science does not "take place in a vacuum, and it is important to maintain sensitivity to the social implications, whether positive or negative" of scientific research, because that work "manifests in real-world social contexts" (O'Connor et al., 2012).

In addition to its use as a communication tool among scientists and between scientists and the public and media, social networks are research tools that scientists are leveraging for their research. While not strictly a social media service, Amazon's Mechanical Turk, launched in 2005, has become a major research platform used by social scientists and psychologists worldwide to aid in, and speed up, the collection of large amounts of behavioral data on demand. In addition, Facebook and Twitter have provided researchers with unparalleled access to data regarding human communication, interaction, relationships, and politics.

Despite its utility, many scientists still view social media as "frivolous" or a "waste of time" (Collins et al., 2016). A common counterargument to this frivolity perspective is that networking is good (for scientists specifically and science in general); however, because conferences are costly affairs only available to the privileged relative few, social media provide a platform to more cheaply run "the biggest research conference in the world" (Faulkes, 2014). As Stafford and Bell note, "academia aspires to openness, engagement, and respect for the principles of rational discussion. Social media facilitate these. The online community is free-flowing, somewhat chaotic, and information-rich—much the same as science has ever been" (Stafford and Bell, 2012). This chaos is daunting and, at times, overwhelming. While social media services provide unprecedented tools for scientific communication, data collection, and the study of human behavior, its amplifying power can make it feel like a Faustian bargain, opening scientists and their research projects to rapid, anonymous, and sometimes voluminous criticism.

Open Science Accelerates Innovation

For centuries, scientists have communicated research findings through the publication of peer-reviewed manuscripts in scientific journals. These results have remained locked in static documents—research papers—often closed to the general public unwilling to pay for journal subscriptions or individual article fees. Additionally, there are significant delays between original submission of the research paper and its final publication. While the slowness of scientific publication has been addressed through preprint publications, this practice has largely remained restricted to mathematics, computer science, and the physical sciences, primarily through deposition of results onto the arXiv server. The adoption of preprint publication by the broader biomedical science community has been slow, although it is rapidly accelerating in recent years with the creation of the bioRxiv preprint server by the Cold Spring Harbor Laboratory in 2013.

While these pre-peer review services are critical for quickly announcing new scientific results, much of the cumulative sum of peer-reviewed biomedical scientific progress is not on preprint servers. The dominant service for peer-reviewed biomedical publications is PubMed, which currently indexes more than 27 million articles. While PubMed allows for rapid, easy discovery of individual papers, it does little to aid in the recovery of specific information *from* those papers. The launch of the journal *PLOS Biology* in 2003 explicitly sought to change this. In announcing their rationale for the creation of a new journal, Brown, Eisen, and Varmus claimed that "freeing the information in the scientific literature from the fixed sequence of pages and the arbitrary boundaries drawn by journals or publishers—the electronic vestiges of paper publication—opens up myriad new possibilities for navigating, integrating, 'mining,' annotating, and mapping connections in the high-dimensional space of scientific knowledge." Thus, the goal of *PLOS* was not simply to open access of scientific results to everyone, but to create a publishing system *whose constituent publications could themselves be a source of data to be mined*. While still nascent, a number of neuroscientific publications have done just that, often with the goal of automating meta-analyses (Yarkoni et al., 2011) or hypothesis generation (Voytek, 2016; Voytek and Voytek, 2012).

Issues of open access and open data ignited the field in 2005 when John

Please cite this article in press as: Voytek, Social Media, Open Science, and Data Science Are Inextricably Linked, Neuron (2017), https://doi.org/10.1016/j.neuron.2017.11.015

Ioannidis published "Why Most Published Research Findings Are False." This article began from the fact that there are a number of standard methodological issues in psychological, biological, and medical sciences, such as using a standard $p < 0.05$ significance threshold, methodological flexibility, and small sample sizes. When combined with a general positive publication bias, the overall result is that most research findings in these domains will be "false" (that is, not reproducible). This assertion was explicitly tested in 2015 by the Open Science Collaboration—a consortium of 270 scientists (Open Science Collaboration, 2015). They found (with some contention) that when trying to replicate 100 experimental and correlational psychological studies, only 47% of the original effect sizes were within the 95% confidence interval of the (larger) replication effect size. Such replications are costly, in terms of both time and human hours of effort, and would be far less necessary in a culture of open data.

That said, open data as a scientific policy is not without its critics, who, in their most severe caricatures, contend that "a new class of research person will emerge—people who had nothing to do with the design and execution of the study but use another group's data for their own ends…" (Longo and Drazen, 2016). Such research persons, or, as the above authors refer to them, "research parasites," are imagined to engage in "stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited." While legitimate concerns regarding data sharing, such as credit, attribution, privacy, and so on, are warranted (Voytek, 2016), the above caricature captures the uncertainty felt by some in the greater scientific community.

So, why practice open science? Is it worthwhile? Open source software has facilitated the creation of numerous private companies by allowing them to repurpose and build upon the efforts of countless contributors. Many of the world's largest technology companies have benefited tremendously from open source software like the Linux operating system and Python programming language. In turn, those companies, at least partly built on open infrastructure, have contributed a great deal of their own open source software back to the community. This sets up a virtuous cycle of innovation. This sharing of code (and protocols, data, etc.) has been facilitated by version control software, such as GitHub, and software development environments, such as iPython. Just as open source software has created tremendous tools, technologies, and financial growth, so too can open data create "a virtuous cycle that allows researchers to remix and reanalyze data in new and interesting ways" (Voytek, 2016).

The Data Science Ascendency

As a byproduct of their sheer size, social networks and technological companies have generated huge amounts of data. As companies grew, the size of the datasets they needed to analyze quickly became larger than the memory available on any one given computer system. These "big data" sets posed new challenges for processing and analysis. This early hurdle was addressed by Google's creation of MapReduce, a system for splitting up a large task into smaller tasks that can each be processed on a single machine that are part of larger clusters. Although such large clusters were initially inaccessible to academic researchers, in 2006 Amazon launched Elastic Cloud Compute (EC2) as part of Amazon Web Services (AWS). This service provides on-demand cloud computing resources as needed (for a cost), leveraging system downtime within the greater Amazon computational infrastructure.

Given recent "big data" initiatives in neuroscience, ranging from the simultaneous collection of thousands of channels of electrophysiological data at tens of kHz resolution, to massive repositories of human functional and structural brain imaging, to the combination of thousands of small-scale studies' worth of data, cluster and cloud computing capabilities are quickly entering neuroscience. The ascendency of data in both the public and academic spheres has led to an entirely new set of technical and research problems to be addressed that do not cleanly fit into the traditional domains of computer science or statistics. For example, a researcher at Facebook may ask how they can statistically aggregate user data to understand user behavior when those data are in different formats that are difficult to combine: textual data from status updates, computer visual recognition from posted photos, in- and out-bound links within posted hyperlinks, the social network of the user, and the geographic locations and times from which those photos, links, and status updates were posted. Similarly, in neuroscience, we may begin to ask how we can aggregate data in different formats to address issues of behavior and mental health, including textual data from self-reports, brain imaging data, personal and family genetic information, electrophysiology, socioeconomic demographic information, and so on.

These questions, and similar new issues related to both the size and diversity of the kinds of data being collected, have led to a boom in "data science." The past 5 years have seen the creation of a number of independent data science programs and institutes. These include the Alan Turing Institute in the UK, the Department of Statistics and Data Science (formerly the Department of Statistics) at Carnegie Mellon University, the University of Washington eScience Institute, and UC Berkeley's Division of Data Science. Additionally, UC San Diego offers an undergraduate Data Science major, which is a joint effort between the Departments of Cognitive Science, Computer Science and Engineering, and Mathematics, to be administered by the new Halicioglu Data Science Institute.

But what is data science? At UC San Diego, rather than trying to define what data science *is*, we instead posed foundational questions unique to data science, questions to isolate what differentiates data science from existing fields, such as the Facebook data aggregation through experiment above. Although there is much to be argued about whether or not data science constitutes its own separate academic discipline, there is little doubt that the questions these new institutes and departments are tackling are significant in both their reach and impact. This impact has already been felt across society: Google indexed the Internet and made discoverable massive amounts of the world's information; the iPhone and subsequent smart phones provided mobile access to the Internet to billions of people; and Facebook and Twitter built

Please cite this article in press as: Voytek, Social Media, Open Science, and Data Science Are Inextricably Linked, Neuron (2017), https://doi.org/10.1016/j.neuron.2017.11.015

atop the Internet a dynamic system for sharing information and connecting with one another.

Social media allow for accelerating access to a diversity of people and ideas. Open science allows rapid, easy access to tools, technologies, and data that are the products of countless hours of human effort. Data science allows for new ways of thinking about combining and remixing those ideas, technologies, and data in ways that are shaping the future of science. None of these advances would be possible without the development of the original ARPANET, which was created as a means to speed the distribution of scientific results and software among scientists. Scientific necessity gave rise to the Internet, which in turn paved the way for social media, open science, and data science. All of these tools and technologies have incredible power to shape science and society in mutually beneficial ways and to enhance the quality of life for everyone. However, they are not without their own dangers and pitfalls and, as with other powerful tools, need to be wielded with care and thoughtfulness.

The fact that the rapid expansion of social media occurred at the same time as the open access and reproducibility movements in science, as well as during the emergence of big data and data science (Table 1), is not an accident, but a consequence of their inextricable interrelationships. That is, the big data and openness of LIGO and the social openness of the Polymath Project reflect the same general shift in the nature of the scientific process: social, reproducible, transparent, open, and data driven. And just as biodiversity is critical for vibrant ecosystems and neuronal diversity is critical for mammalian brain functioning, the mixing of diverse datasets, methods, tools, and ideas will allow science to flourish.

ACKNOWLEDGMENTS

B.V. acknowledges Curtis Chambers, Thomas Donoghue, and Jessica Voytek for their critical comments. B.V.'s laboratory is supported by a Sloan Research Fellowship and the National Science Foundation (1736028).

REFERENCES

Collins, K., Shiffman, D., and Rock, J. (2016). PLoS ONE *11*, e0162680–e10.

Faulkes, Z. (2014). Neuron *82*, 258–260.

Longo, D.L., and Drazen, J.M. (2016). N. Engl. J. Med. *374*, 276–277.

Neuroskeptic. (2013). Trends Cogn. Sci. *17*, 195–196.

O'Connor, C., Rees, G., and Joffe, H. (2012). Neuron *74*, 220–226.

Open Science Collaboration (2015). Science *349*, aac4716.

Stafford, T., and Bell, V. (2012). Trends Cogn. Sci. *16*, 489–490.

Voytek, B. (2016). PLoS Comput. Biol. *12*, e1005037.

Voytek, J.B., and Voytek, B. (2012). J. Neurosci. Methods *208*, 92–100.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Nat. Methods *8*, 665–670.