Big data biomedicine offers big higher education opportunities

John Darrell Van Horn[a,b,1]

I like to tell my students a story about a time back in the "olden days" of the early 1990s when I was a newly minted postdoctoral fellow at the National Institutes of Health (NIH). Our laboratory conducted brain imaging studies using positron emission tomographic imaging and would routinely obtain 6–12 functional image volumes per subject in our experiments. The director of our neuroimaging laboratory asked me to explore the purchase of a new computer hard drive capable of storing our growing collection of brain imaging data files. At that time, we had already accumulated quite a number of subjects and wanted to easily access their data for a variety of different statistical analyses—and we expected to obtain much more data.

Navigating big data, to maximize its utility, will require specially designed training programs directed at early-career scientists. Image courtesy of Dave Cutler.

So I searched, and I compared and contrasted, rating drive quality, price, and capacity. With our not-so-expansive $4,000 budget, I eventually decided on the drive that would surely satisfy our needs. It was among the largest self-contained external hard drives of its kind at that time. Its massive 4 gigabyte capacity seemed infinite. People would come to visit the laboratory to just gaze upon it, jealously oohing and ahhing. Their own puny little hard drives paled in comparison with the storage capacity brute that stood before them. It would never fill up, I thought. I could not have been more wrong.

Today, with access to my institution's modern brain imaging systems, I can easily eat up more than 4 gigabytes—on a single research subject in an afternoon. Indeed, the data storage capacity at the University of Southern California (USC) Mark and Mary Stevens Neuroimaging and Informatics Institute, where I work now [and where my responsibilities also include directing the NIH's Big Data to Knowledge (BD2K) Training Coordinating Center], is on the order of 6 petabytes–where 1 petabyte equals 1,024 terabytes, and a terabyte is 1,024 gigabytes. Clearly, we have come quite a long way in our collective ability to acquire, store, and process large amounts of digital scientific information.

Physicists earn their pay by making things like MRI scanners collect more data per unit time, capabilities we neuroscientists gladly take advantage of. Researchers, in turn, push the physicists to turn a few more screws, apply new mathematical methods, etc., to wring still more data out of those scanners. This applies also to genomics and proteomics, where technology and research questions drive each other forward. Not surprisingly, this cycle has greatly amplified the amount of data being acquired. One can be assured that scientists will not collect less data—they will only collect more. All this means a dramatically new landscape for young biomedical researchers interested in collecting data with ever more detailed levels of spatial and temporal resolution. Training, however, is paramount. Fortunately, plans are afoot to develop the computational skill sets of those who can work with the latest "big data" technologies.

[a]The Institute for Neuroimaging and Informatics, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA 90032; and [b]Laboratory of Neuro Imaging, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA 90032

Author contributions: J.D.V.H. wrote the paper.

Any opinions, findings, conclusions, or recommendations expressed in this work are those of the author and do not necessarily reflect the views of the National Academy of Sciences.

[1]Email: jvanhorn@usc.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1607582113

Dealing with the Data

We frequently hear about the emergence of big data, referring to the large amounts of information now being mined in relation to identifying trends in social media, to economics and consumer preferences, and to voter profiling, among other applications. In the case of biomedicine, large-scale data are being gathered from across a range of disciplines, stored, combined, and used in new and creative large-scale informatics applications. This includes massive archives of genomic, proteomic, and phenomic information, as well as the data being continuously obtained from mobile devices. It involves data from exercise monitors, tablet computers, and a dizzying variety of websites with millions of users. How such data are organized, managed, manipulated, and visualized is an active area of data science research.

The analytics being performed on such data promise to improve the precision and specificity of health monitoring, telemedicine, and medical decision-making. Recent progress in information technology applied to biomedical archives and individual data is now altering the landscape of patient privacy and personal information, with patients getting more control of their health information but with concerns over who else might have access, as well. But to maximize the research potential, a new breed of data scientists must be trained—early on in their careers—with a carefully tailored, specially designed course curriculum.

NIH has recognized this movement to gather and analyze ever-larger biomedical data, spearheading the BD2K program (https://datascience.nih.gov/bd2k)—a $656 million initiative to support development of the science of big data biomedicine. Currently funding several major centers of excellence around the country (laboratories within my own institution being among them), numerous individual data science research grants, and a flock of training grants, the BD2K effort seeks to develop and deploy new analytic tools, modern data management techniques, and new capabilities for just moving data from place to place. Although it is in its infancy, the BD2K effort will undoubtedly shift how the world thinks about biomedical data, how to work with it, and what it has to tell people about their personal health status.

But a critical element of the big data biomedical enterprise is its potential role in undergraduate, graduate, and postdoctoral education. How does one begin to learn about how big big data is and how big it will get? What forms do these data take? How do we mine, combine, and link these archives? How can we visualize results in intuitive ways? What does it mean to put these results into the practice of medicine and precision health?

An important and prescient provision made by the NIH program officials in the development of the BD2K program was to specifically include a role for education and outreach across the board for the efforts the program funds (https://datascience.nih.gov/sites/default/files/BD2K Training Summary_website.pdf). For instance, each major BD2K center of excellence has a specific budget dedicated to training-related efforts related to big data. But more than that, new, individual research programs are being designed, along with career development efforts, and Massive Open Online Courses are soon to get underway. Emerging programs also cover a range of workshops and seminars for young investigators, "big data experiences" for postdocs, and, importantly, new university courses specifically targeted to attracting undergraduate students into the realm of biomedical data science. More than 50 distinct NIH awards related to individual training and, notably, the development of new course curricula have been funded.

New Training Programs

New courses are essential for nurturing the coming generation of biomedical data scientists. They should seek to introduce, among other things, how to access large data archives, how to wrangle thousands or tens of thousands of records, how statistics can be performed with such large samples, data mining, time series analysis, and the domain-specific science behind it all. The University of Washington in Seattle and Northwestern University in Chicago, for example, each have announced brand new educational programs focused on introducing the ways of big data biomedicine to undergraduates. In addition, the California State University campuses at Northridge and at Monterey Bay, as well as at the University of Puerto Rico, are each developing programs that have a unique focus on the inclusion of underrepresented minorities and first-generation students with interests in large-scale biomedical data education.

A critical element of the big data biomedical enterprise is its potential role in undergraduate, graduate, and postdoctoral education.

Each of these, as well as other BD2K-seeded programs, necessitates the development of new course curricula, new standards for academic evaluation, and the development of new educational communities. These and all other BD2K training efforts work closely with the BD2K Training Coordinating Center, which I direct here at USC, to ensure that their programs are academically well designed but also to help evangelize about them to interested students. Once fully under way, these new educational programs can be expected to positively augment the data science skill sets of such students when they eventually enter the bioscience workforce.

Insisting on the training of a new crop of young scientists, schooled in the ways of massive data analytics, is the NIH's way of doubling down on their view that big data is here now and expected to get bigger, and that the current focus on health informatics depends greatly on a labor force having the skills to turn data into actionable information relevant to biomedical science. In particular, working to attract undergraduates to this emerging field of data science, even before they have begun to focus on traditional "bench" science, will not only help fill the existing need for people trained in big data science but also underscores that the NIH views the future of biomedical research as computationally vast, digitally precise, and ever more personal.

Undergraduate education is an ideal place to start. Now is the time for more universities to place an emphasis on educating this new generation of scientists, focused on the science of large-scale biomedical data. In so doing, universities will help hasten a time in the future when working petabytes and petabytes of big biomedical data might, in fact, be no big deal for many of us. The opportunities are enormous for budding computer scientists, mathematicians, and engineers to meld with trainee biologists, neuroscientists, and clinicians in novel university courses constructed to focus on the truly 21st century phenomenon that is the data avalanche from large-scale biomedicine.

With the ever-increasing role of digital information being gathered in domains such as neuroimaging, genetics, and "-omic" sciences, it is clear that the data we gather today, though seemingly big, might eventually be considered "cute." Indeed, an old adage notes that data expand to fill the space available to them for storage (a corollary of the so-called Parkinson's Law). Thus, even all those petabytes noted above will eventually be filled, and storage on the order of exabytes (1,024 petabytes) will be considered to replace them. Interestingly, this suggests that an apt way for defining big data is that it is the large-scale data we haven't even begun collecting yet. In other words, if you think data are big now, just wait. To make good use of it all, the next generation of biomedical "big data" scientists must be thoroughly trained to find the cleverest ways for analyzing and mining the data onslaught.

Acknowledgments

The author thanks Michelle Dunn, Carol Shreffler, and Erica Rosemond from the National Institutes of Health; Jeana Kamdar, Crystal Stewart, Sumiko Abe, Caroline O'Driscoll, Jose-Luis Ambite, Kristina Lerman, Lily Fierro, Michael Taylor, and John Berardo from the University of Southern California; and the staff of the USC Mark and Mary Stevens Neuroimaging and Informatics Institute. The author directs the NIH's Big Data to Knowledge (BD2K) Training Coordinating Center (www.bigdatau.org; NIEHS U24 ES026465-01).