From big data to deep insight in developmental science

Rick O. Gilmore

The Pennsylvania State University, The Databrary Project

Author Note

Rick O. Gilmore is in the Department of Psychology, The Pennsylvania State University, University Park, PA 16802, rogilmore@psu.edu. This work was supported by the National Science Foundation (BCS-1238599), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (U01-HD-076595), and the Society for Research in Child Development. Any opinions, findings, and conclusions or recommendations expressed in the material contributed here are those of the author and do not necessarily reflect the views of the National Science Foundation, the Eunice Kennedy Shriver National Institute of Child Health and Human Development, or the Society for Research in Child Development. I thank Karen E. Adolph for her comments on the manuscript.

Abstract

The use of the term “big data” has grown substantially over the past several decades, and is now widespread. In this review, I ask what makes data “big” and the implications of size, density, or complexity of datasets for the science of human development. A survey of existing data sets illustrates how large, complex, multi-level, multi-measure data have been used to reveal the complexities of human development. “Big data” poses significant technical, policy, ethics, transparency, cultural, and conceptual issues that must be addressed. Most big developmental science data are currently hard to find and cumbersome to access; the field lacks a culture of data sharing; and there is no consensus about who owns or should control research data. But, these barriers are rapidly dissolving. Developmental researchers are finding new ways to collect, manage, store, share, and enable others to reuse data about the multiple facets of human development. This promises a future in which “big data” can lead to deep insight about some of the most vexing questions in behavioral science.

From big data to deep insight in developmental science

Introduction

A search on the term "big data" yields more than 49 million hits on Google, more than 100,000 results on Google Scholar, and 13 million hits on Bing. The results return in less than a second. A search of the term using Google's Ngram viewer (`https://books.google.com/ngrams`) that indexes terms in digitized books shows the first appearance of the term around 1900 with a steady rise in frequency from the 1950s to around 2000. Clearly, as measured by search engine matches of electronic documents on the Internet, scholarly documents, or digitized books, the use of the term "big data" has grown substantially over the past several decades, and is now widespread. Moreover, the fact that these basic statistics about a particular phrase can be determined in an instant speaks to the rapid progress in networked computing, search engines, and databases. Most of the tools that enable it have been created in the last 20 years. "Big data" has become a significant cultural phenomenon (Borgman, 2015; boyd & Crawford, 2012), with frequent feature articles in the popular (Lohr, 2012; Marcus, 2013) and specialist press (McAfee & Brynjolfsson, 2012; Press, 2013).

In this review, I show how increased availability of and interest in big data sets promises to alter the study of human development. I begin by asking what makes data "big" and the implications of size, density, or complexity of datasets for understanding human development. Then, I survey case studies to illustrate how large, complex, multi-level, multi-measure data sets have been used to reveal the complexities of human development. I conclude by discussing key questions that "big data" approaches pose for the future of developmental science pose.

I disagree with the claim that an era of "big data" signals the end of theory (Anderson, 2008). Nor are "big data" analyses especially new in developmental science. Although some claims about "big data" revolutionizing scientific understanding are exaggerated (boyd & Crawford, 2012), developmental science tackles questions that can

benefit substantially from larger, richer, more widely shared, and more readily inter-operable datasets. Significant novel insights into human development emerging from the era of "big data" depend on the size, density, and complexity of the datasets, on how widely and openly the data are shared, and on how readily researchers are able to combine or link datasets across levels of analysis. In turn, these specific innovations depend largely on small, probably manageable, but nonetheless thorny problems related to policy, scientific culture, individual researcher behavior, publisher priorities, and research funding levels. Technology may accelerate the "big data" era, but the challenges it poses turn out to be relatively unimportant for advancing research in developmental psychology.

What Does "Big Data" Mean in Developmental Science?

According to Laney (2001) the volume, velocity, and variety of data streams make data big. Of course, general statements about the total quantity of data generated per day (IBM, 2015) make little sense outside of a specific research context. High volume data for a developmental psychologist–an archive of 10 terabytes (TB) of video and flat-file data, for example–is tiny to a physicist working on the Large Hadron Collider (LHC) that generates 30 petabytes in a year (`http://home.web.cern.ch/about/computing`). Similarly, what is "big" depends on how volume is measured. The Inter-university Consortium for Political and Social Research (ICPSR) (`https://www.icpsr.umich.edu`), one of the largest data repositories for data from the social sciences, consists of more than 500,000 files in 16 specialized data collections. Yet, until the recent acquisition of video data from the Gates Foundation-funded Methods of Effective Teaching (MET) Project (`http://www.metproject.org`), the entire repository totaled about 10 TB of digital storage.

Even more important than the quantity of information stored is the kind of information: What are the targets of inquiry? The range of phenomena covered by developmental science spans multiple levels of analysis in space (from genes to geography)

and time (from microseconds to millennia) all aimed at answering two questions: "What develops?" and "Why does it develop this way?" In seeking answers, developmental scientists have long recognized the importance of multiple, nested influences on developmental processes arising at different temporal and spatial scales (Elman et al., 1998; Gottlieb, 1998; Oyama et al., 2000; Vygotsky, 1980). For social and behavioral scientists, the goal is to describe change in the psychological processes of individuals, groups of individuals, or families. Data about neighborhoods, schools, and broader social, cultural, political and environmental influences primarily inform thinking about the development of individual or family behavior. Similarly, the assessment of biological data–genes, hormones, physiological responses, brain activity, body dimensions, brain structure, disease or disorder status–inform thinking about how within-person factors influence changes behavior.

Consequently, the aspects of data *volume* pertinent to most developmental scientists include the number of participants or families and/or the number of measurement time points. Some datasets have hundreds or thousands of participants. Such high volume datasets enable more precise estimation of small effects, especially for rare qualities or conditions. Similarly, the aspects of data *velocity* most relevant to developmental researchers relate to the frequency or spacing of measurements. Velocity can span many orders of magnitude, from physiological measurements collected at millisecond time scales to longitudinal research spanning years or decades. High volume or velocity data may inform the estimation of trends within and between people across time (Rietveld et al., 2014). *Variety* encompasses the range of measurement types employed across developmental research–Biological, behavioral, contextual, and cultural/historical/evolutionary–and the use of different types of measurements to address the same underlying construct. In addition to the "three Vs", intra-individual *variability* and complexity, or the mutual interdependence of individual measures, might also be included.

The collection, management, and analysis of data high in volume, velocity, variety,

variability, and complexity poses significant practical challenges for data collection, capture, storage, sharing, analysis, transfer, visualization, and analysis. Big data may also magnify the challenges facing researchers in maintaining participant privacy, in part because the more data that are collected, the more likely it is that individual identities can be discovered (Sweeney, n.d.). Big data also pose theoretical challenges: For example, how do micro-scale factors influence macro-scale phenomena? Nevertheless, big data offer the possibility of tackling some of the most profound and vexing questions in the field if the relations among data components can be revealed in ways that do not undermine ethical research principles. Greater openness and more widespread sharing of research data than is current practice will be essential to realize this scientific promise. Developmental researchers can look to the recent past for examples of big datasets that have already been collected and in some cases, widely shared for inspiration.

Big Datasets in Developmental Research

Existing big datasets in developmental science fall into one of three broad categories depending on who collected the data and what mandates or restrictions apply to who may access it. Many of the largest existing datasets available to developmental researchers come from data collected and disseminated by government entities such as the U.S. Census Bureau and the U.S. Department of Education. The collection and dissemination of these datasets is mandated by law and subject to provisions of a complex set of statutory and regulatory requirements designed to protect individual respondents' identities. Government-collected datasets tend to be the most open and widely available to researchers, but by their very design and legal mandate the datasets are not intended for answering questions about individuals, families, or small groups. Datasets collected by individual researchers or teams of researchers employed by institutions, usually colleges or universities, who are funded, usually by government or private funders, to collect large datasets form the next group. These datasets tend to be somewhat smaller in size, but

significantly more varied in terms of the types of information collected, the duration or intensity of data collection, the means of collection, the extent to which data are made available to individuals outside the research team, and the impact on scholarship. The extent to which datasets collected by government-grant-funded research teams are shared with other researchers depends on a set of factors including institutional research ethics and data privacy policies, funder and journal requirements, and individual researcher's support for data sharing. Finally, some large-scale datasets are collected by private entities for business purposes and may include sharing data about individual subscribers or users. The data collected by private entities about individuals largely concern consumer behavior because the data customers are other commercial entities seeking to sell products or services to individual users. But, commercial entities are beginning to collect and store health and fitness related activity (e.g., Fitbit, etc.) These datasets are sometimes made available to academic researchers, but the policies that govern data access are under the control of the commercial entities that provide the services.

Government-collected and Managed Datasets

Some of the largest and most widely used datasets in developmental science are population-based, often with a focus on health. The National Health and Nutrition Examination Survey (NHANES; `http://www.cdc.gov/nchs/nhanes.htm`) began collecting data in 1959 on the health and nutritional status of U.S. infants, children, and adults. The annual survey combines interviews with a geographically diverse representative sample of about 5,000 people. NHANES focuses on demographic, socioeconomic, dietary and health-related topics and physical examinations involving medical, dental, and physiological measurements, and laboratory tests of behavior. The products include standardized child growth charts and the closely associated National Youth Fitness Survey (NYFS; `http://www.cdc.gov/nchs/nnyfs.htm`), which reports physical activity and fitness levels in 3 to 15-year-olds Much of the NHANES and NYFS data are available to

the public through the National Center for Health Statistics at the U.S. Centers for Disease Control. However, access to some data is restricted, available only to individuals via one of the 19 Federal Statistical Research Data Centers locations across the U.S. By mid-2015 more than 110,000 publications had cited the NHANES dataset.

In contrast to the cross-sectional NHANES, The National Longitudinal Survey of Youth (NLSY; `http://www.bls.gov/nls/`) has examined the same cohort–almost 10,000 American youth born between 1957 and 1964–on a annual basis since the study began in 1979. NLSY data are now available for 25 survey rounds on questions relating to employment and other issues; extensions encompass wider age ranges. Most NLSY and NLS data are available free of charge by means of a web-based data portal to which users must apply for access, but access to geographic-related variables requires special permission. By mid-2015 more than 20,000 publications had cited the NLSY.

The Early Childhood Longitudinal Study (ECLS; `https://nces.ed.gov/ecls/`) is another longitudinal population-based study It focuses on child development, school readiness, and early school experiences in separate cohorts of children followed either from birth or from kindergarten that in some cases exceed 20,000 individuals. ECLS consists of both publicly available and restricted data archived at ICPSR. A Google Scholar search generates more than 8,000 hits, of which the ICPSR archive can verify 167 project-related publications.

Several large government-collected and shared datasets contain information about employment patterns and educational attainment in children and youth. The Bureau of Labor Statistics reports data about youth employment and unemployment patterns derived from the Census Bureau's Current Population Survey (`http://www.bls.gov/cps/demographics.htm`). The U.S. Department of Education's National Center for Educational Statistics (NCES) collects and analyzes statistics on U.S. primary, secondary, and post-secondary education, and many of the datasets are either public; data with personally identifiable data are available to researchers under a restricted

use agreement. The Trends in International Math and Science Study (TIMSS) has studied mathematics and science achievement in 4th and 8th grade for more than 20 years (http://timss.bc.edu/), with more than 20,000 U.S. students included in the 2001 sample. De-identified data from TIMSS and its sister study, the Progress in International Reading Literacy Study (PIRLS; http://timssandpirls.bc.edu) are available for public download.

Some large-scale longitudinal studies focused on children's growth, health, and education include participants from outside the U.S. The World Health Organization's (WHO) Multicentre Growth Reference Study (MGRS; http://www.who.int/childgrowth/en/) collected physical growth and related data from 8,500 children in Brazil, Ghana, India, Norway, Oman, and the U.S. The Japan Environment and Children's Study (JECS; http://www.env.go.jp/en/chemi/hs/jecs/) involves 100,000 parent-child pairs with the goal of evaluating the impact of environmental factors on children's health and development. Large-scale birth cohort studies in the U.K. include the National Survey of Health & Development and the National Child Development Study (NCDS; http://www.cls.ioe.ac.uk). NCDS is an ongoing longitudinal study that follows everyone in Great Britain who was born in one particular week in 1958, focusing on factors predicting wellbeing across the lifespan. NCDS data are available by registration from the U.K. Data Service (http://ukdataservice.ac.uk/get-data/key-data/cohort-and-longitudinal-studies). The British Cohort Studies have generated more than 3,500 publications to-date. The Organization for Economic Cooperation and Development (OECD) publishes the PISA comparative educational dataset (http://pisa2000.acer.edu.au/index.php), which generates more than 5,000 search hits.

Not all large-scale birth cohort studies succeed. A notable failure is the National Children's Study (http://www.nichd.nih.gov/research/NCS/Pages/default.aspx). Authorized by the Children's Health Act of 2000, the NCS would have followed 100,000

children prenatally until age 21. However, the NIH Director decided to close the NCS in 2014 following the recommendations of an advisory panel. Questionnaire, physical measures, biospecimens and environmental sample data from up to 5,726 participants in the NCS Vanguard study were collected in 2009-2014 prior to study closure. Those data are slated for release in a data archive sometime in 2015.

Several themes emerge from this sample of U.S. and international population-based studies. Datasets are large in terms of the volume of participants sampled, ranging in the thousands to tens of thousands. Some studies involve dozens of samples per individual over extended periods of time, and show significant variety in the measures collected. Most data are available to the public via web-based download or browsing/analysis portal, although access to data deemed sensitive may require a specific application and approval and possibly travel to a Research Data Center where special security provisions apply. Finally, the datasets have generated research publications that range from the hundreds to the tens of thousands, suggesting that the collection, curation, and preservation of these resources has had a significant impact scientific discovery. I turn next to datasets whose collection and management is under the control of individual investigators.

Public/Private Partnerships and Investigator-driven Project

Several big data studies in developmental science involve partnerships between government agencies and academic investigators. Others involve large-scale data collections initiated and managed by individual investigators and funded by government or foundation sources. These projects are diverse in focus and methods, including population-based studies with a health focus, behavior genetics, brain imaging, cognition and temperament, and education and employment.

Population-Based Studies. The Panel Study on Income Dynamics (PSID; `https://psidonline.isr.umich.edu`) is a population based survey study focusing on employment, income, wealth, expenditures, health, marriage, childbearing, child

development, philanthropy, and education. PSID began in 1968 with a U.S. nationally representative sample of 5,000 families and more than 18,000 individuals. As of mid-2015, PSID had generated more than 3,900 publications. The data is available publicly to registered users who agree to the terms of use conditions that ensure participants are not reidentified and that users properly cite and acknowledge use of the data. Some restricted data elements (e.g., geospatial data, school-level identifiers) are only available subject to a formal restricted use agreement between the investigator and the data owner, the University of Michigan.

The National Longitudinal Study of Adolescent Health (Ad Health; `http://www.nichd.nih.gov/news/releases/Pages/adolescent.aspx`) began as a survey study undertaken in response to a Congressional mandate designed to measure the multiplicity of influences on adolescents' health and unhealthy behaviors. Parts of the Ad Health cohort have been followed into young adulthood with four in-home interviews, and in a recent wave, biomarker data from a blood test has been made available. The Ad Health data are archived at ICSPR and are available for public use. As of mid-2015, more than 5,000 publications had arisen from the Add Health dataset.

The NICHD Study of Early Child Care and Youth Development (SECCYD; `https://www.nichd.nih.gov/research/supported/Pages/seccyd.aspx`) prospectively followed the experiences of 1,364 infants born at 10 locations throughout the United States in 1991. The goal was to learn more about the kind and quality of child care experiences and the effects of different child care settings on developmental outcomes. Measures included video, surveys and interviews, behavioral tests, biomarkers such as salivary cortisol, blood pressure, anthropometrics, and demographic information. Unlike the Ad Health study, data in SECCYD access are restricted, available only by application, with approval granted on a case-by-case basis. Staff at the National Institute of Child Health and Human Development manage the data access and approval process. The SECCYD has generated more than 214 publications as of mid-2015.

Investigator-initiated and managed, the Family Life Project (FLP; http://flp.fpg.unc.edu) used a population-based approach to study the early development of children living in rural, largely poor communities in North Carolina and Pennsylvania. A birth cohort of 800 children were studied in a series of home, child care visits, and phone calls when the infants were 2, 6, 15, 24, and 36 months of age. Data included interviews, questionnaires, videotaped interactions, and biological samples yielding information about economic and community factors, family, home, language and cognition, stress, and genotype. Data from Phase I of the FLP are archived at ICPSR, but are subject to restrictions on access due to the sensitive and identifiable nature of the data. The FLP site lists about 50 publications through 2013.

Behavior Genetics. Large data samples have long been required in behavior genetics to detect small effect sizes. Several large-scale genetics studies that are not specifically developmental deserve mention. The Genome of the Netherlands Project(http://www.nlgenome.nl) is a publicly available dataset from 250 families, consisting of two parents plus one of their adult children. The Psychiatric Genomics Consortium (PGC; http://www.med.unc.edu/pgc) contains data from more than 170,000 individuals with psychiatric diagnoses or who are at-risk. Many disorders in the PGC dataset have developmental dimensions(autism, attention-deficit hyperactivity disorder, and schizophrenia). Results may be viewed in a specialized web browser tool (http://www.broadinstitute.org/mpg/ricopili/).

The U.K.-based Twins Early Development Study (TEDS; http://www.teds.ac.uk/) contains survey data from 15,000 English and Welsh families who gave birth to twins between 1994 and 1996, with lab-based behavioral task, survey, and DNA samples from a subset of participants. TEDS has generated more than 350 publications to-date. Researchers wishing to gain access to the data may do so by submitting a formal data request specifying the aims of the planned research and the variables needed to satisfy the aims. Proposals that do not overlap with analyses already being planned or carried out by

the TEDS research team or other collaborators are usually approved.

The Twin and Offspring Study in Sweden (TOSS; `http://ki.se/en/meb/twin-offspring-study-in-sweden-toss`) is another twin-focused study aimed at uncovering genetic and environmental contributions to measures of family relationships and mental health. TOSS extends the Twin Moms Project to include a sample of twin fathers (320 twin pairs and their families). In addition, TOSS plans to collect another 250 pairs of twin mothers for a target sample of 900 twin parents (3,000 individuals in all) and their families. While still relatively early in data collection, more than 24 publications reference the TOSS dataset. Data sharing plans have not yet been announced.

Among behavior genetics studies, those focusing on adoption provide a unique window on the mix of genetic and non-genetic factors that influence child development. The Colorado Adoption Project (CAP; `http://ibg.colorado.edu/cap`) provides an illustrative example. Since its inception in 1976, CAP has enrolled more than 2,400 participants from more than 450 families, collecting data on cognitive abilities, temperament, and demographics. CAP has generated more than 200 published articles, and data collection continues. Data from 1976-1989 are archived at Dataverse and can be accessed by application.

Brain Imaging. Several large-scale studies have focused on developmental patterns in brain structure or activity. One of the earliest studies was a combined longitudinal and cross-sectional study of child and adolescent structural brain development led by Jay Giedd at NIMH (Giedd et al., 1999). The study collected structural brain information from 145 4 to 20-year-olds along with behavioral and clinical data. More than 100 of the participants were scanned on more than once, about two years apart. The original paper describing this study has been cited more than 3,300 times, as of mid-2015. The dataset has generated other highly cited papers (Gogtay et al., 2004), but the data are not available for analyses by researchers outside the original investigative team.

The NIH MRI Study of Normal Brain Development

(http://pediatricmri.nih.gov/nihpd/info) collected multi-modal structural MRI from 554 children from 4 to 18 years of age. A second cohort of children from birth to 4 years was scanned longitudinally with up to 10 scans per child. Demographic, hormonal, cognitive, affective, and psychiatric data were also collected. The data are available to qualified researchers by application. John Richards and colleagues have combined data from the NIH MRI Study, data collected in their own labs, and other public sources to create average structural brain templates that can provide more accurate bases for developmental functional neuroimaging studies using EEG and fMRI (http://jerlab.psych.sc.edu/NeurodevelopmentalMRIDatabase/). These data can be openly accessed with appropriate citation.

More recently, the Pediatric Imaging, Neurocognition, and Genetics (PING; http://pingstudy.ucsd.edu/) Project collected multimodal neuroimaging data, genotypes, neurodevelopmental histories, and information about cognitive and social and emotional function in more than 1,000 participants 3-20 years of age recruited from 8 U.S. cities. Data are available to the research community by application. As of mid-2015, the project had generated more than 20 publications.

Outside the U.S., The Developing Human Connectome Project (http://www.developingconnectome.org), led by investigators at King's College London, Imperial College London, and Oxford University goes against the trend of imaging studies that focus on older children. This project aims to study the connectome, a map of human brain connectivity in fetuses from 20 to 44 weeks post-conception. Funded by the European Research Council, the imaging data are accompanied by clinical, behavioral and genetic information. Investigators hope that the data set will provide a basis for studying genetic and environmental risks that could lead to neurodevelopmental disorders such as Autistic Spectrum Disorder or Cerebral Palsy. The project has generated 17 publications as of mid-2015.

Language, Cognition, and Temperament. The child language community pioneered the aggregation and sharing of datasets. The CHILDES/TalkBank (MacWhinney, 2001) archive is one of the largest and most well-established archives in the behavioral sciences. It consists of transcripts and audio and video recordings of children's utterances along with recent data from adults with aphasia. CHILDES/TalkBank has generated more than 10,000 citations, and the datasets are available to the research community, many of them publicly.

WordBank (`http://wordbank.stanford.edu/`) is an open access archive that consists of data from more than 40,000 samples of the MacArthur-Bates Communicative Development Inventory (CDI). Catherine Tamis-Lemonda has released video data (`https://nyu.databrary.org/volume/8`) from an NSF-funded longitudinal study of child language consisting of more than 1,000 sessions of infants and children (9 mos-7.6 years) and their mothers carrying out a series of semi-structured tasks. The Human Speechome Project (`http://www.media.mit.edu/cogmac/projects/hsp.html`) at MIT's Media Lab recorded 10 hours of video from one child's home on a daily basis from birth to age three. The project has generated more than 78 publications, but the data are not available outside of the original investigative team. Data were gathered at an average rate of 200 gigabytes per day, necessitating the development of sophisticated data-mining tools to reduce analysis efforts to a manageable level, and transcribing significant speech added a labor-intensive dimension. The LENA (Language ENvironment Analysis; `http://www.lenafoundation.org/`) Foundation created a technology framework that allows children's speech in natural settings to be recorded and analyzed. The system provides an automatic language collection and analysis tool for speech language professionals and parents. No large-scale archive for LENA data currently exists, but some researchers who use the tool have begun to explore the creation of a data repository specialized for these sorts of data.

In other areas of cognition, a number of measures that have become standards and

their widespread adoption has resulted in data that are large in volume, velocity or variety. Davida Teller (Teller, McDonald, Preston, Sebris, & Dobson, 1986) pioneered empirical techniques for measuring visual acuity in preverbal children, and the use of Teller Acuity Cards has resulted in the publication of age-based norms for visual development in infants to four-year-olds (Mayer et al., 1995) based on a sample of more than 400. A companion study on children in Brazil (Salomão & Ventura, 1995) was conducted with an even larger sample.

The Bayley Scales of Infant and Toddler Development (Bayley, 2006) is widely used in clinical and epidemiological settings such as the EPICure (`http://www.epicure.ac.uk/`) because of the published norms even though there are questions about its validity (Hack et al., 2005). However, data about those norms is under the control of the Pearson Publishing company that publishes the Bayley and licenses its use. Some Bayley score data are available from the Carolina Abecedarian Project and the Carolina Approach to Responsive Education (CARE; `http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4091`) project on ICPSR. The data are freely available to registered users of ICSPR.

The standardized Wechsler intelligence tests were developed in the 1930's. The Wechsler Intelligence Scale for Children-III (WISC-III) is designed for children ages 6 - 16, and the Wechsler Preschool and Primary Scale of Intelligence-R (WPPSI-R) is designed for children age 4 - 6 1/2 years. The current version (WISC-V) is published by Pearson (`http://www.pearsonclinical.com/psychology/products/100000771/wechsler-intelligence-scale-for-childrensupsupfifth-edition--wisc-v.html`). Data about the norms are not available to researchers. However, data from the Project on Human Development in Chicago Neighborhoods (PHDCN) that collected WISC-R data on 6,000 children, adolescents, and young adults is housed at ICSPR and are accessible by specific authorization.

In the realm of emotion and temperament, two academic-investigator-initiated

measures have been widely adopted. The Infant Behavior Questionnaire (IBQ) is a parent survey measure, and the Laboratory Temperament Assessment Battery (Lab-TAB; http://www.uta.edu/faculty/jgagne/labtab) is an observational battery, using a set of 3-5 min episodes mimicking situations in everyday life. IBQ data on 1,388 participants are archived at ICSPR as part of the Maternal Lifestyle Study (MLS; https://neonatal.rti.org/about/mls_background.cfm), a longitudinal multi-site observational study of the long-term effects of in-utero exposure to cocaine on child development. IBQ and Lab-TAB data are also available from ICPSR from the Family Life Project.

Education and Learning Science. Several studies in the education and learning sciences studies deserve mention because of their volume, velocity, or variety or because of their availability for reuse by others. The 1999 TIMSS video study recorded math and science teaching practices in seven countries, and some of the materials are available for public use through a repository hosted at UCLA (http://www.timssvideo.com/). The Gates Foundation Measures of Effective Teaching (MET; http://www.metproject.org/) Project videotaped more than 3,000 classroom teachers beginning in 2009, with follow-on studies continuing currently. The MET video data are stored and shared at ICSPR. The Spatial Intelligence and Learning Center (SILC; http://spatiallearning.org/) is an NSF-funded initiative designed to develop a multi-disciplinary science of spatial learning, including the development of tests and instruments useful in research. LearnLab (http://learnlab.org), another NSF-funded center, combines cognitive theory and computational modeling to understand changes in student knowledge in the context of computer-based math, science and language courses. LearnLab hosts a repository, DataShop (http://learnlab.org/technologies/datashop/index.php), to store and openly share detailed data from classrooms employing cognitive tutors. DataShop enables data storage, management, visualization and analysis in a web-based tool.

Summary. The studies involving partnerships with government agencies or those initiated and managed by individual investigators exemplify the high volume, velocity, variety, and variability of big datasets. The studies employ a range observational measures, including video recordings, population-normed test instruments, biological measurements of physiology, and brain structure or function. However, the extent to which data are available for secondary reuse, and the process for acquiring access is more variable than for datasets initiated and managed by government entities. Only some of the datasets are stored in data repositories, for example. Many large-scale developmental dataset are usually housed locally, on project-specific sites, not on centralized servers that aggregate data across studies and sources. In many cases, the original investigative team retains control over the use of data by other researchers, including the kinds of questions that third parties may ask. In some cases, the original investigative team must be included as an author on publications. Perhaps as a consequence of these factors, these datasets have generated fewer scientific publications than government-sponsored efforts.

Non-Academically Initiated or Managed Sources of Big Data

Some large-volume sources of developmental data are collected and managed by private, non-academic or government entities. More than 1.6 million high school students (Lewin, 2013) take standardized tests or provide financial aid information via measures developed and managed by The College Board and ACT, Inc. The College Board shares Scholastic Aptitude Test (SAT) and college cost and scholarship data with the research community by application. So does the ACT (`http://www.act.org/research/`).

Internet-based for-profit service providers operate at an even larger scale. Google's Gmail has more than 900 million users worldwide (Lardinois, n.d.), and Facebook has more than a billion (Protalinski, 2014). According to the YouGov site (`https://yougov.co.uk`), 17% of Gmail users in the United Kingdom are 17-24 years of age. Facebook's policies require that users be at least 13 years of age

(https://www.facebook.com/help/157793540954833), but detailed information about user demographics for Facebook or other social media popular among children and adolescents is not openly available. Of course, detailed information about users, their characteristics, and preferences is the primary asset social media companies mine and market to the business community. Users receive free services in exchange for providing these data. Both Google and Facebook have arms that conduct research and cooperate with academic researchers albeit with significant public criticism about the ethics of certain research projects (Meyer, 2014). The primary criticism concerns whether Facebook users had given informed consent to participate in the manipulation of their newsfeeds as would be required by research ethics boards if a similar study were undertaken in a laboratory context.

The scale of data collected and managed by non-academic entities dwarfs that of government or academically-managed initiatives. Because the data are collected for proprietary business purposes, it is difficult to assess their current or potential impact on the scholarship of human development.

The Future of Big Data in Development

Clearly, the collection, analysis, and sharing of large, multilevel datasets has been part of the fabric of developmental science for a long time. In this section, I discuss a range of technical, conceptual, and theoretical issues that arise in thinking about the future of big data in developmental science.

Technical

Technical issues associated with big data in developmental science center on collection, storage and retrieval, data management, provenance, and analysis.

Collection from Multiple Sources and in Diverse Formats. Developmental scientists collect data from sources representing multiple levels of analysis. Increasingly,

measurement devices provide data and metadata in structured, organized, and machine-readable formats.

Although some researchers continue to use paper and pencil measures to collect survey information, many universities now have site-licenses for web-based tools such as SurveyMonkey and Qualtrics that reduce the manual labor involved in preparing a survey and processing completed data for analysis. Behavioral measures involving computer-based tasks are commonly used in developmental research, but most rely on custom, project-specific software. So, the output data files, while often in an electronic form, may require significant post-processing to be linked with other data. New tools like Amazon's Mechanical Turk (`http://www.mturk.com`) or Apple's HealthKit (`https://developer.apple.com/healthkit/`) are empowering behavioral and health researchers to conduct large-scale behavioral science experiments using tools specialized for psychological research (e.g., `https://psiturk.org`), and with data delivered in well-structured electronic formats. Amazon's terms of use prohibit minors, but developmental researchers have found ways to secure video-based informed consent from parents to enable their children to participate in looking time studies (`https://lookit.mit.edu`) over the web.

Video and audio recordings are a long-standing mainstay of developmental research. Video captures the complexity and richness of behavior unlike any other measure, and so video provides a uniquely valuable source of information for researchers who study behavior in laboratory, home, classroom, or museum contexts. Images and recordings are larger in file size, denser than text or flat-file data, and come in a diverse formats. Thus, a high priority is the development of tools to enable storage and sharing of images (including brain images), and audio and video data.

Lab-based tools for conducting physiological measurements such as EEG, heart-rate or skin conductance, produce electronic files of these time series. Genetic analyses from modern gene sequencing tools and reports from tissue, blood, or salivary samples typically

yield machine-readable outputs. Magnetic Resonance Imaging (MRI) systems produce electronic image data and machine-readable subject-level metadata; however, many research teams limit the amount and kind of subject-level metadata they enter into MRI databases because of the possibility of violating research participant confidentiality. But, unlike MRI, there are no standard file formats, and most data collection systems provide no standard subject-level metadata. Thus, the files require significant post-collection data processing prior to analysis.

New technologies, specifically the widespread use of smart mobile devices with embedded sensors, promise to make big data streams about individual participants' locations, physiological states (e.g., `https://www.empatica.com/`), activity patterns, and momentary cognitive, and emotional states broadly available to researchers. These tools will enable the collection of data from large numbers of participants in short periods of time, significantly enlarging the volume, velocity, and variety of data available for analysis.

Storage and Retrieval. Developmental researchers who wish to store and share big data face a bewildering array of options. These include individual or institutional websites, institutional repositories (e.g., `https://scholarsphere.psu.edu/`), cloud services (Dropbox, Box, or Amazon), domain or measure-specific repositories (ICSPR, Databrary.org, TalkBank.org, WordBank.org, OpenfMRI.org), domain general services (Researchgate.net, FigShare/SlideShare, Dataverse, and the Open Science Framework), and open source software web sites (GitHub). Some journals offer or require data storage, but these are typically limited to text-based flat-files used for statistical analyses and do not include raw images, videos, or physiological time series. The diversity of storage options can pose daunting challenges for researchers and institutions. Identifiable and sensitive data must be kept secure. Storage solutions must meet the needs of researchers during the active data collection phase of a study while not posing insurmountable hurdles to data sharing down the line. The effort to reconcile these competing demands led Databrary (`http://databrary.org`), for example, to build tools that allow researchers to

upload session-level video and flat-file data to a secure web-based server as it is collected, thus minimizing post-study data curation. The Open Science Framework (`https://osf.io`) offers similar data management functionality for non-identifiable data.

Where and how data are stored is only part of the problem. To foster increased reuse, data must be discoverable and accessible to researchers. At present, it is far easier to search and discover research publications relevant to a particular topic using web-based search tools than it is to find data. There are several reasons. Most research publications do not use data that are readily available to investigators outside of the research team. Datasets that are available may lack persistent, citable, identifiers. When data from a publication are available to other researchers, access is often restricted and requires a specific, time consuming application to a data repository or to the original data producer. In contrast, Databrary allows researchers access to a library of data under a single access agreement, an innovation aimed at accelerating reuse. Another barrier to reuse is the difficulty of finding data that meet specific task or demographic criteria. Some repositories such as ICPSR and the National Database for Autism Research (NDAR) maintain extensive standardized metadata about tasks and participant demographics. This can help investigators to search for specific data sources. But, even within domain-specific repositories variable-level searching is not available for all datasets. The problems of where to store and how to find and retrieve data will increase as datasets grow in size and complexity.

Coding, Analysis, and Provenance. Even easy-to-find datasets must be processed prior to analysis. Indeed, most data science involves "janitor work" (Lohr, 2014). The process of curation involves carefully documenting how raw information from a data stream was transformed into information used in formal analyses. Can the provenance of the data be recorded in ways that others can understand, reproduce, and rely upon? For example, physiological data are often filtered and smoothed, sometimes by the recording devices. Video data are usually coded by human observers and the codes transformed into quantitative measurements. What were the variables, units of measurements, calibration

properties of the instrument, and definitions of key terms and codes? Well-curated datasets usually report these components, but that curation takes time and specialized expertise that many individual investigators lack.

Several software tools have recently emerged that make it easier for researchers to produce and reproduce self-documenting data workflows. For example, the free RStudio (`https://www.rstudio.com`) and Jupyter (`https://jupyter.org`) environments allow researchers to create electronic notebooks that combine data, annotations and observations, statistical analyses, and visualizations in human-friendly web or document formats. The free, open-source Datavyu (`http://datavyu.org`) video coding tool allows automated data analysis and export schemes to be created with the Ruby scripting language. Many developmental researchers may be unfamiliar with these sorts of tools, but volunteer groups such as Software Carpentry (`https://software-carpentry.org`) provide researchers with on-site training in the use of tools for reproducible research workflows, including the use of version control and workflow scripting. Similarly, Databrary and the Center for Open Science (`http://centerforopenscience.org`) have initiated open office hours and conference-based and regional workshops to provide hands-on researcher training. But, the use of tools that produce well-curated, reproducible scientific workflows remains rare among mainstream developmental researchers.

Summary. Technical issues will continue to slow progress in many areas of developmental research that depend on big data. Critical challenges include getting data into open, standard, and easily manipulated electronic formats as soon as possible in the research cycle; the development and widespread adoption of data storage platforms or repositories that provide a degree of metadata standardization and enable search and discovery; the creation and adoption of data management practices that make curation part of the research workflow; and the creation of a cohort of developmental researchers who have the training and expertise to implement these techniques in their own labs. There is demonstrable progress on many of these fronts, and therefore cause to be

optimistic that the technical challenges can be overcome.

Research Ethics and Practice

Clearly the collection, analysis, and interpretation of large scale datasets present issues related to research ethics, participant privacy, and scientific transparency. Professional ethics require that special care be taken about what data are collected from research participants and who gains access to it. The focus in developmental science on studying vulnerable research populations magnifies these concerns.

Differing practices across cultures in terms of privacy pose challenges for collecting and aggregating datasets. In the U.S., researchers must navigate a regulatory environment in which different types of data are covered under different sections of Federal law. For example, the Federal Educational Rights and Privacy Act (FERPA; `http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html`) governs access to student educational records. The Health Insurance Portability and Accountability Act (HIPAA; `http://www.hhs.gov/ocr/privacy/hipaa/understanding/`) governs the disclosure of individually identifiable health information may be disclosed and to whom. The Code of Federal Regulations (CFR) Title 45 (`http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html`) governs research with human participants. If the data in question are audio or video recordings, different state provisions can come into play. For example, in two-party states (`http://www.dmlp.org/legal-guide/recording-phone-calls-and-conversations`) both the person making the recording and the person(s) being recorded must consent, making some forms of data collection using recordings problematic.

Research activities funded by the U.S. Federal Government are supervised by an Institutional Review Board or its equivalent. These entities are regulated by the U.S. Office of Health and Human Services (`http://www.hhs.gov/ohrp/`). Researchers supervised by IRBs must respect participants' privacy, secure informed consent, and

maintain confidentiality. These ethical principles have practical consequences for research. They limit the ways researchers may recruit participants. Minors must give informed assent to participate in a study with a parent or guardian giving formal consent. Data must be collected in ways that minimize the likelihood that information given by a participant will be disclosed or a participants' identity revealed outside the research team. This usually means that data items that could reveal a participant's identity are removed or altered to reduce the likelihood of disclosure. Whether deidentified data of this sort can be shared with researchers outside the original team that collected depends on several factors. One factor is the sensitivity of the data collected and the likelihood that specific identities or home locations or other items could be revealed. Another factor concerns whether participants were informed that deidentified data might be shared outside the research team. IRBs may view these matters differently, creating additional complexities for large-scale projects that span geographic areas. Some IRBs may require participants to be informed that deidentified data might be shared outside the IRB-approved research team, and others may deem that the analysis of deidentified data no longer meets the definition of human subjects research. Many big datasets in developmental science have restrictions on access either because the data were collected under Federal regulations that prohibit releasing identifiable data or because the participants were not asked for permission to share data with other researchers. From a big data perspective, if data cannot be shared outside the original IRB-approved research team, then the possible analyses are restricted to the interests, resources, and expertise of that team.

Of course, some data types like photographs and audio or video recordings contain identifiable information that cannot be removed or altered without reducing the value to others. Thus, data from photographs or recordings requires additional consideration and special care. Databrary, a digital data library specialized for storing, managing, and sharing video data from developmental research has an access model that empowers researchers who wish to share identifiable research data to do so with explicit permission of

the participants. Databrary has created template language to help researchers gain participants' permission. Furthermore, Databrary restricts access to identifiable data to researchers who have formally agreed to uphold ethical research principles and whose institutions approve of their access. The notion that research participants can consent to share identifiable or potentially identifiable data is relatively new. The Personal Genomes Project (`http://www.personalgenomes.org`), Open Humans Project (`https://www.openhumans.org`), and Human Connectome Project (`http://www.humanconnectomeproject.org`) embody similar principles. The experience of Databrary investigators is that a significant proportion of research participants and their parents or guardians will consent to sharing identifiable data, mostly video, with other members of the research community.

It is too early to predict whether it will become commonplace for academic developmental researchers to seek explicit permission to share identifiable research data with other researchers. But, there are reasons to be optimistic. In just over a year of operation, Databrary has secured formal agreements with more than 80 institutions in North and South America, Europe, Australia, and Asia allowing more than 140 researchers to access identifiable data. However, some leading developmental researchers have argued that the families of research participants forge a relationship of trust with a particular research team, formalized through the informed consent document (Eisenberg, 2015). The relationship might be harmed or the research project negatively affected if participants were asked to share data with other researchers. Sensitive to the latter argument, Databrary has recommended that permission to share be sought separately from consent to participant in research and after a given data collection has ended. Moreover, the fact that most families agree to share when asked suggests that the relationship of trust involved in research participation might be extended to a community of researchers, given suitable provisions and constraints. Undoubtedly, seeking explicit permission to share on a consistent and widespread basis would resolve any ambiguity about whether a given

dataset can be shared with whom and for what sort of purposes.

Greater transparency and more explicit clarification about what data is being collected and for what purposes could be sought from commercial entities as well. Social media companies like Google, Facebook, Twitter, SnapChat, and Instagram have business models that involve the collection, mining, and packaging of data, usually to advertisers, in exchange for services that are free to users. Although some services attempt to restrict the ages at which users can create accounts, the limits often lack rigor, and there is no parallel to the requirement of adult consent required in formal research contexts. The data collection and analyses carried out by private entities are subject to no supervision or formal regulation comparable to academic research. Instead, data use, analysis, and sharing provisions are governed by terms of use agreements that users acknowledge by clicking a button prior to using a given service. Unlike academic settings, where violations of research ethics principles may involve significant consequences for the researchers and institutions, violations of commercial terms of use require aggrieved parties to seek redress through litigation. The White House has recommended data privacy principles (The White House, 2012) that some software companies have adopted voluntarily.

Unresolved issues that could impact the availability of big data in the future include whether linkage across streams increases risk of reidentification, whether it is essential to reconsent minors when they become adults, a notion most researchers find totally impractical and a significant barrier to data sharing, and a general concern about the ethics of granting consent to share data for an indefinite period. Because data security cannot ever be guaranteed, risks can be minimized and managed, but not entirely eliminated. Finally, there are unresolved questions about privacy protections in the consumer domain that have the potential to influence academic research

Transparency and Reproducibility. An important dimension of scientific ethics concerns transparency and reproducibility. The social and behavioral sciences have incurred an unfortunate string of high profile cases of scientific misconduct in recent years,

including cases of fraudulent data (Singal, 2015; Bhattacharjee, 2013). The credibility problem is magnified by several factors. Lack of power and unrestricted exploratory analyses may mean that most research findings are false (Ioannidis, 2005). The actual effect sizes of published findings are unknown due to a bias toward publishing positive results. Most journals reject papers that report failures to replicate published findings, and as a result, few scientists attempt replications or are recognized and rewarded for doing so (Nosek, Spies, & Motyl, 2012). The problem is so serious that some have claimed that science as a whole faces a crisis of reproducibility.

To address this problem, the Center for Open Science has organized several large-scale replication efforts, including some in psychological science under the "Many Labs" project (`https://osf.io/ct89g/`; `https://osf.io/8cd4r/`). The results of these pre-registered, open, large sample replications have been mixed (Collaboration, 2015). Some published effects were replicated, but others were not.

Whether there are replicability problems exist in developmental science and whether they constitute a crisis is unknown. Developmental research reflects the same positive effects biases seen in other fields, and the same problem that null results often sit unpublished in file drawers–the so-called file drawer effect (Rosenthal, 1979). No failures to replicate developmental studies have been reported to Psychfiledrawer.org (`http://psychfiledrawer.org`), a resource designed to bring replication failures to light. As some developmental researchers have written (Bishop, 2012), replicating effects with developmental populations can be especially difficult and so even partial replications are noteworthy. No large-scale replication efforts in developmental science have been mounted, but there have been calls for changes in journal practices to give replications a more privileged place in scientific publications (Bishop, 2012). One barrier to more open data practices appears to be researcher's fears of having their reputation or abilities publicly undermined (Ascoli, 2006). So, changing views about replication may require shifts in the scientific culture. Researchers should work to reduce the extent of blame levied at

researchers whose initial positive findings fail to be replicated by others (Bishop, 2015). Technological tools that foster increased openness and transparency and more systematic research data management (OSF and Databrary) will also contribute to changing the scientific practices. So will the widespread adoption of more consistent journal practices related to transparent and open scientific practices (Nosek et al., 2015).

Still, the increasing availability of large-scale datasets about developmental questions promises to magnify problems at the intersection between exploratory and confirmatory research. Large volume, high velocity, and high variety datasets make it possible to explore and discover novel unpredicted patterns in data. But, novel findings might be spurious, and exploratory findings must be properly confirmed. Whereas pre-registration and pre-review have been suggested as one way to address the problem of spurious exploratory findings, these tools are not practical in all cases and could have a chilling effect on discovery. In contrast, increased transparency about the process that led to an exploratory finding and the steps taken to confirm it can bolster a finding's credibility. Thus, developmental researchers may find it essential to adopt more transparent and reproducible workflows using some of the new tools developed for this purpose (COS; Databrary).

Community Engagement and the Impetus for Change. Developmental researchers have clearly shown enthusiasm for sharing the results of their findings via publications, and in some subfields, the sharing of data, materials and methods is firmly established. Open sharing practices tend to be more common when there is a high cost, centralized source of scientific data that could not conveniently be owned or managed by individual researchers (e.g., space telescopes or the U.S. Census).

In addition to bottom-up/grassroots initiatives, journals and funding agencies continue to play a vital role in creating an impetus for change. Funders can require data management plans and mandate that data and research products be deposited into particular types of open repositories and provide funding to build and support big data infrastructure. Journals can also require that data be deposited in open archives as a

condition of publication in addition to adopting other transparent and open science practices for manuscripts they accept (e.g. PLoS). However, the problem with data sharing mandates from funders is that there is no specific mechanism to provide sufficient ongoing financial support for data archives. Few researchers budget funds to support data management and archiving. Some journals are willing to shoulder the burden of storing and sharing data associated with publications, but others refuse to accept supplemental materials of any kind (Maunsell, 2010). Thus, in the interest of promoting greater openness and transparency, funders and journals may create unfunded mandates that make it harder for researchers to make discoveries. For example, new regulations specifying when data must be deposited may be unwieldy and impractical for developmental scientists to carry out their work (Eisenberg, 2015; Group, 2015).

These issues are complicated by lack of consensus about who *owns* research data (*Data Ownership*, n.d.). Federal funding agencies might argue that the public should own research data paid for by tax dollars, much like other data collected by government agencies such as the U.S. Census, National Weather Service, and U.S. Bureau of Labor Statistics. The institutions that employ, receive, and manage federal grants might stake a claim to ownership. Most investigators naturally feel a strong sense of ownership over their intellectual products, although formal copyright is often surrendered in the process of publishing, and that sense extends to data. Some have even argued that research participants themselves own their own data, and there are new business models emerging that may soon provide individuals an opportunity to sell data for personal gain (`http://www.datawallet.io`).

The lack of consensus about who owns data means that access is often limited in ways that impede reuse by others. Some investigative teams control who has access to datasets, for what purposes and for how long. That control may persist indefinitely. Others grant access to data only if co-authorship on any published product is guaranteed. Although legitimate arguments might be made in favor of embargo periods that enable

teams of researchers to mine and report findings from their research efforts, the ideal of fostering greater data reuse argues for the shortest possible periods. Establishing consensus about data ownership and the kind of control investigators can exercise over it will require conversations among researchers, institutions, and funding agencies. That consensus may well prove vital to achieving some of the benefits of big data analyses in development.

Conceptual and Theoretical Issues

The increasing availability of big datasets for analysis in developmental research poses significant theoretical and conceptual questions alongside the many pragmatic ones already discussed. Big(ger) data may help to overcome limitations with our existing knowledge base. One challenge that big data may help to address is a particular bias in existing samples. Developmental research typically purports to study what is normative about changes across time in human behavior. But, much of what we have learned about developmental processes comes from samples that represent only a small fraction of the world's population (Karasik, Adolph, Tamis-LeMonda, & Bornstein, 2010; Fernald, 2010) Developmental psychology, like other branches of the field, presents findings from Western, education, industrialized, rich, and democratic (WEIRD) societies (Henrich, Heine, & Norenzayan, 2010). So, to the extent that new tools enable research on development in non-WEIRD cultures and those data can be aggregated and combined will strengthen the ability to make claims about universal or near-universal components of developmental processes. However, developmental researchers are well aware of cohort effects–the notion that developmental processes can be influenced by changing social and cultural norms. Thus, even the most culturally diverse dataset possible may still yield conclusions that are locked in time.

A second challenge larger datasets may help to address is the fact that most social, behavioral (Maxwell, 2004) and neuroscience studies (Button et al., 2013) are underpowered. Most worryingly many published research findings are false in fields that

rely on small sample sizes, test multiple relationships between variables, engage in exploratory research, use diverse research designs, definitions, outcomes, and analytical modes across studies, and when more labs seek out significant effects (Ioannidis, 2005). The collection, analysis, and sharing of larger datasets will help to ameliorate these factors.

Developmental research faces a specific point of tension related to the measurement. Many of the measures for which high volume data are available come from proprietary, expensive instruments such as the Bayley and the WIPPSI. Free, academic instruments such as the Infant Behavior Questionnaire have no centralized data archive. Plus, the measures themselves have been revised several times, making it more challenging to compare data collected using different versions, especially across time. Similar problems arise when non-proprietary tasks are used. Most investigators customize even a well-known task to make it suitable for use with children, and the sharing of research materials is just as limited as the sharing of data. Efforts to encourage researchers to capture and record the conceptual structure of psychological tasks have been undertaken (e.g., The Cognitive Atlas; `http://www.cognitiveatlas.org`), but have not yet been widely adopted.

Finally, some critics have raised concerns that the rise big data means the "end of theory" (Anderson, 2008). In a provocative essay Anderson (2008) argued that large quantities of data mean the traditional model of scientific inquiry involving hypothesis testing will soon give way to model-free descriptions of data. Others note that bigger data don't necessarily lead to deeper insights (Graham, 2012). Boyd and Crawford (boyd & Crawford, 2012) note that beyond its scientific dimensions, big data is a sociocultural phenomenon that raises as many questions as it promises answers. Developmental science has a rich and rigorous intellectual history in which theory and experiment play central roles. That tradition is likely to continue.

Conclusion

As boyd and Crawford (2012) “The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions.” (p. 662). The clamor extends to the developmental and learning sciences where improving health and maximizing the potential for human achievement have significant consequences for daily living.

This review has shown that the collection, dissemination and analysis of data sets that are big in volume, velocity, or variety has a long and established history in developmental science. Many ‘big data’ studies have had substantial impact on scholarship, and in some cases, on public understanding and policy. For the most part, studies with the largest impact (as measured by the quantity of published papers) have been ones funded by and managed by government entities, perhaps due to widespread access to data for reuse. Those investigator-initiated and managed projects that have had the largest intellectual impacts have created significant intellectual communities around the datasets, communities that extend the beyond the boundaries of the original investigative teams. Although technical issues about data formats, storage, cleaning, visualization and provenance remain, there is significant progress. Developmental researchers have available a growing array of well-established data repositories (CHILDES, Databrary, Dataverse, ICPSR) and new data storage/management tools (Databrary, OSF). Research and data management practices have begun to converge on norms that will reduce the costs of preparing data for sharing in the future. New ethical procedures for seeking informed consent to share identifiable data have been developed and are being implemented in diverse research contexts that promise to accelerate future reuse.

On the other hand, many barriers remain. Most developmental science data are hard to find, and cumbersome to access, even for researchers. Most data linked to publications are not stored in open data repositories. This means that virtually all of the data from

unpublished studies remains unavailable, making the size of the file drawer effect unknown. Most investigators do not currently employ workflows that make it easy to share data or to document analysis pathways. With rare exceptions clustered around specific datasets there is no widespread culture of data sharing, and indeed some level of bias against the use of secondary data. Finally, there is no unified understanding or consensus within developmental science about who owns research data, whether it is essential or merely wise to share data, and when in the research cycle data should be shared. These factors limit the potential for discovery that the era of big data so seductively promises, especially in the face of what can seem a glacial pace of change in scientific culture.

Nevertheless, we should remember that Facebook was launched in 2004, Twitter in 2006, and the iPhone in 2007. It would be unwise to underestimate the speed with which new technologies, tools, and cultural practices can change. If developmental researchers can find ways to collect, manage, store, share, and enable others to reuse data about the multiple facets of human development, as many are beginning to do, we can look forward to a future rich in theory and understanding.

References

Anderson, C. (2008, June). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.* Retrieved 2015-07-27, from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/

Ascoli, G. A. (2006, September). The ups and downs of neuroscience shares. *Neuroinformatics*, *4*(3), 213–215. Retrieved 2015-05-08, from http://link.springer.com/article/10.1385/NI%3A4%3A3%3A213 doi: 10.1385/NI:4:3:213

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development.*

Bhattacharjee, Y. (2013, April). Diederik Stapel's Audacious Academic Fraud. *The New York Times.* Retrieved 2015-08-25, from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html

Bishop, D. (2012, Jan 19). *Novelty, interest and replicability.* Retrieved 2015-08-25, from http://deevybee.blogspot.co.uk/2012/01/novelty-interest-and-replicability.html

Bishop, D. (2015, Jul 11). *Publishing replication failures: some lessons from history.* Retrieved 2015-08-25, from http://deevybee.blogspot.com/2015/07/publishing-replication-failures-some.html

Borgman, C. (2015). *Big Data, Little Data, No Data.* MIT Press. Retrieved from https://mitpress.mit.edu/big-data

boyd, d., & Crawford, K. (2012, June). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. Retrieved 2015-07-27, from http://dx.doi.org/10.1080/1369118X.2012.678878

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013, May). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. Retrieved 2015-07-27, from

http://www.nature.com/nrn/journal/v14/n5/abs/nrn3475.html doi: 10.1038/nrn3475

Collaboration, O. S. (2015, August). Estimating the reproducibility of psychological. *Science*, *349*(6251), aac4716. Retrieved 2015-08-28, from http://www.sciencemag.org/content/349/6251/aac4716 doi: 10.1126/science.aac4716

Data Ownership. (n.d.). Retrieved 2015-08-25, from https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html

Eisenberg, N. (2015, June). *Thoughts on the Future of Data Sharing - Association for Psychological Science.* Retrieved 2015-08-25, from https://www.psychologicalscience.org/index.php/publications/observer/2015/may-june-15/thoughts-on-the-future-of-data-sharing.html

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1998). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press.

Fernald, A. (2010). Getting beyond the "convenience sample" in research on early cognitive development. *Behavioral and Brain Sciences*, *33*(2-3), 91–92.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., ... Rapoport, J. L. (1999, October). Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience*, *2*(10), 861–863. Retrieved 2015-07-16, from http://www.nature.com/neuro/journal/v2/n10/abs/nn1099_861.html doi: 10.1038/13158

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., ... Thompson, P. M. (2004, May). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(21), 8174–8179. Retrieved 2015-07-16,

from http://www.pnas.org/content/101/21/8174 doi: 10.1073/pnas.0402680101

Gottlieb, G. (1998). Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. *Psychological Review*, *105*(4), 792–802. Retrieved 2015-07-13, from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.105.4.792-802 doi: 10.1037/0033-295X.105.4.792-802

Graham, M. (2012, March). Big data and the end of theory? *The Guardian*. Retrieved 2015-08-25, from http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory

Group, A. D. S. W. (2015, Jun). *Data Sharing: Principles and Considerations for Policy Development*. Retrieved 2015-08-25, from http://www.apa.org/science/leadership/bsa/data-sharing-report.aspx

Hack, M., Taylor, H. G., Drotar, D., Schluchter, M., Cartar, L., Wilson-Costello, D., ... Morrow, M. (2005, August). Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. *Pediatrics*, *116*(2), 333–341. doi: 10.1542/peds.2005-0173

Henrich, J., Heine, S. J., & Norenzayan, A. (2010, June). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2-3), 61–83; discussion 83–135. doi: 10.1017/S0140525X0999152X

IBM. (2015, March). [CT000]. Retrieved 2015-07-08, from http://www-01.ibm.com/software/au/data/bigdata/

Ioannidis, J. P. A. (2005, September). Why Most Published Research Findings Are False. *CHANCE*, *18*(4), 40–47. Retrieved 2015-08-25, from http://amstat.tandfonline.com/doi/abs/10.1080/09332480.2005.10722754 doi: 10.1080/09332480.2005.10722754

Karasik, L. B., Adolph, K. E., Tamis-LeMonda, C. S., & Bornstein, M. H. (2010). Weird walking: Cross-cultural research on motor development. *Behavioral and brain*

sciences, *33*(2-3), 95–96.

Laney, D. (2001, February). *3D Data Management:Controlling Data Volume, Velocity, and Variety* (Tech. Rep.). META Group. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lardinois, F. (n.d.). *Gmail Now Has 900m Active Users, 75% On Mobile.* Retrieved 2015-08-03, from http://social.techcrunch.com/2015/05/28/gmail-now-has-900m-active-users-75-on-mobile/

Lewin, T. (2013, August). More Students Are Taking Both the ACT and SAT. *The New York Times.* Retrieved 2015-08-03, from http://www.nytimes.com/2013/08/04/education/edlife/more-students-are-taking-both-the-act-and-sat.html

Lohr, S. (2012, February). Big Data's Impact in the World. *The New York Times.* Retrieved 2015-07-08, from http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

Lohr, S. (2014, August). For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. *The New York Times.* Retrieved 2015-08-03, from http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

MacWhinney, B. (2001). From CHILDES to TalkBank. In B. MacWhinney, M. Almgren, A. Barreña, M. Ezeizaberrena, & I. Idiazabal (Eds.), *Research in Child Language Acquisition.* Somerville, MA: Cascadilla.

Marcus, G. (2013, March 13). Steamrolled by Big Data. *The New Yorker*. Retrieved 2015-07-08, from http://www.newyorker.com/tech/elements/steamrolled-by-big-data

Maunsell, J. (2010, August). Announcement Regarding Supplemental Material. *The Journal of Neuroscience*, *30*(32), 10599–10600. Retrieved 2015-08-25, from http://www.jneurosci.org/content/30/32/10599

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. doi: 10.1037/1082-989X.9.2.147

Mayer, D. L., Beiser, A. S., Warner, A. F., Pratt, E. M., Raye, K. N., & Lang, J. M. (1995, March). Monocular acuity norms for the Teller Acuity Cards between ages one month and four years. *Investigative Ophthalmology & Visual Science*, *36*(3), 671–685.

McAfee, A., & Brynjolfsson, E. (2012, October). *Big Data: The Management Revolution - HBR.* Retrieved 2015-07-08, from https://hbr.org/2012/10/big-data-the-management-revolution/ar

Meyer, R. (2014, June). Everything We Know About Facebook's Secret Mood Manipulation Experiment. *The Atlantic*. Retrieved 2015-07-08, from http://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015, June). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. Retrieved 2015-07-08, from http://www.sciencemag.org/content/348/6242/1422 doi: 10.1126/science.aab2374

Nosek, B. A., Spies, J. R., & Motyl, M. (2012, November). Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. Retrieved 2015-07-28, from http://pps.sagepub.com/content/7/6/615 doi: 10.1177/1745691612459058

Oyama, S., Taylor, P., Fogel, A., Lickliter, R., Sterelny, P. K., Smith, K. C., & Weele, C. v. d. (2000). *The Ontogeny of Information: Developmental Systems and Evolution*. Duke University Press.

Press, G. (2013, May 9). A very short history of big data. *Forbes*. Retrieved 2015-07-07, from http://www.forbes.com/sites/gilpress/2013/05/09/

a-very-short-history-of-big-data/

Protalinski, E. (2014, jan 29). *Facebook Passes 1.23 Billion Monthly Active Users.* Retrieved 2015-08-03, from http://thenextweb.com/facebook/2014/01/29/facebook-passes-1-23-billion-monthly-active-users-945-million-mobile-users-757-million-daily-users/

Rietveld, C. A., Conley, D., Eriksson, N., Esko, T., Medland, S. E., Vinkhuyzen, A. A. E., ... Koellinger, P. D. (2014, November). Replicability and Robustness of Genome-Wide-Association Studies for Behavioral Traits. *Psychological Science*, *25*(11), 1975–1986. Retrieved 2015-07-13, from http://pss.sagepub.com/lookup/doi/10.1177/0956797614545132 doi: 10.1177/0956797614545132

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi: 10.1037/0033-2909.86.3.638

Salomão, S. R., & Ventura, D. F. (1995, March). Large sample population age norms for visual acuities obtained with Vistech-Teller Acuity Cards. *Investigative Ophthalmology & Visual Science*, *36*(3), 657–670.

Singal, J. (2015, May 29). *The Case of the Amazing Gay-Marriage Data: How a Graduate Student Reluctantly Uncovered a Huge Scientific Fraud.* Retrieved 2015-08-25, from http://nymag.com/scienceofus/2015/05/how-a-grad-student-uncovered-a-huge-fraud.html

Sweeney, L. (n.d.). *Identifiability.* Retrieved 2015-07-08, from http://dataprivacylab.org/projects/identifiability/index.html

Teller, D. Y., McDonald, M. A., Preston, K., Sebris, S. L., & Dobson, V. (1986). Assessment of visual acuity in infants and children; the acuity card procedure. *Developmental Medicine & Child Neurology*, *28*(6), 779–789.

The White House. (2012). "consumer data privacy in a networked world: A framework for

protecting privacy and promoting innovation in the global digital economy. *Journal of Privacy and Confidentiality*, *4*(2). Retrieved from http://repository.cmu.edu/jpc/vol4/iss2/5

Vygotsky, L. S. (1980). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.