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 what implications size, density, or complexity of datasets have for the science of human development. A survey of existing data sets illustrates how existing large, complex, multi-level, and multi-measure data can reveal the complexities of developmental processes. At the same time, significant technical, policy, ethics, transparency, cultural, and conceptual issues associated with the use of big data 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 dissolving. Developmental researchers are finding new ways to collect, manage, store, share, and enable others to reuse data. This promises a future in which big data can lead to deeper insights about some of the most profound 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 (`http://www.google.com`), more than 147,000 results on Google Scholar (`http://scholar.google.com`), and 14 million hits on Bing (`http://bing.com`). 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 the use 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. In turn, 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 the 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 what implications the size, density, or complexity of datasets have for understanding human development. Then, I review and evaluate some of the existing big datasets in developmental science. I conclude by discussing key questions that big data approaches pose for the future of the field.

We will see that big data analyses in developmental science are not especially new. The field tackles questions that have benefited and will continue to benefit from large, rich, widely shared, and readily inter-operable datasets. So, big data approaches to development do not signal the end of theory (Anderson, 2008), nor will they necessarily revolutionize

scientific understanding (boyd & Crawford, 2012). Rather, significant novel insights emerging from the era of big data will depend not just on the size, density, and complexity of the datasets, but on how widely and openly data are shared, and on how readily researchers are able to combine or link datasets across levels of analysis. 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. Thus, technology may accelerate the big data era, but the challenges it poses may turn out to be less important for advancing research in developmental psychology than changes in scientific culture.

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 specific research contexts. High volume data for a developmental psychologist—an archive of 10 terabytes (TB) of video and flat-file data, for example—represents a tiny fraction of the 30 petabytes per year (`http://home.web.cern.ch/about/computing`) available to a physicist working on the Large Hadron Collider (LHC). Similarly, what constitutes big depends on how one measures volume. The Inter-university Consortium for Political and Social Research (ICPSR) (`https://www.icpsr.umich.edu`), one of the largest and oldest 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. Developmental science spans phenomena across 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 "How does development occur"? In seeking answers, scientists have long recognized the importance of multiple, nested influences on developmental processes arising across scales (Elman et al., 1998; Gottlieb, 1998; Oyama et al., 2000; Vygotsky, 1980). Most seek to describe change in the psychological processes of individuals, groups of individuals, or families. Data about neighborhoods, schools, and the broader social, cultural, political and environment primarily inform thinking about the development of individual or family behavior. Similarly, biological data—genes, hormones, physiological responses, brain activity, body dimensions, brain structure, disease or disorder status— are brought to bear to reveal the influence of within-person factors on changes in behavior.

Consequently, the aspects of data *volume* salient 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 the 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 informs 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, transfer, sharing, visualization, and analysis. Big data 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 pose theoretical challenges, too: For example, how do micro-scale factors influence macro-scale phenomena? Nevertheless, big data offer developmental researchers the opportunity to tackle some of the most profound and vexing questions in the behavioral science—if the relations among data components can be revealed in ways that do not undermine ethical research principles. Realizing this promise will require greater openness and more widespread sharing of research data than current practice. But, researchers may draw inspiration from examples of big datasets addressing developmental questions that have already been collected, and in some cases, widely shared. The next section highlights several.

Big Datasets in Developmental Research

Existing big datasets in developmental science fall into one of several broad categories depending on who collected the data, what mandates or restrictions apply to who may access it, and the diversity of measures represented.

Government-Collected and Managed Datasets

Many of the largest existing sources come from data collected and disseminated by government entities such as the U.S. Census Bureau. Table 1 summarizes information about some representative large, developmentally-focused datasets whose collection and hosting is managed by governmental entities. Several themes emerge from this sample. 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. 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. 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 in the U.S., possibly travel to a Federal Research Data Center where special security provisions apply.

Not all large-scale government-initiated studies succeed. A notable failure is the U.S. 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 data from up to 5,726 participants were collected in 2009-2014 prior to study closure. Those data are slated for release in a data archive sometime in 2015. A comparable study in the U.K. targeting 80,000 was cancelled in October 2015, just 8 months after launch, for failures to recruit sufficient numbers of participant families (Pearson, 2015). These failures highlight the significant challenges associated with designing and successfully implementing large-scale birth-cohort studies.

Researcher-Initiated and Managed Datasets

Datasets initiated and collected by academic or medical researchers form a second group. Table 2 summarizes information about some representative large, developmentally-focused datasets whose collection was initiated by individual researchers, and the data themselves are managed by non-governmental entities. These tend to be smaller than those collected by government agencies, but the data collected are more varied in type, means of collection, and duration or intensity. For example, investigator-initiated studies commonly collect observational measures, including video recordings, population-normed test instruments, biological measurements of physiology, genetics, and

brain structure or function. Unfortunately, the extent to which these data are available for secondary reuse, and the process for acquiring access is also more variable than for datasets initiated and managed by government entities. Institutional research ethics and data privacy policies, funder and journal requirements, and individual researchers' support for data sharing influence whether, when, and with whom data are shared. 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. Some require the original investigative team to be included as an author on publications. Most large-scale investigator-initiated developmental datasets are housed locally, on project-specific web sites, not on centralized servers that aggregate data across studies and sources. Only some are stored in open public data repositories, for example. Catherine Tamis-Lemonda's MetroBaby dataset (Tamis-LeMonda, 2013) hosted on Databrary is a notable exception.

Measure-Specific Data

Datasets representing a single test or form of measurement constitute another group. Table 3 summarizes information about some measures commonly used in developmental science research and datasets created around them. Many measures in this category derive from the use of standardized instruments with group norms. It is considered best-practice in many research communities to employ widely adopted, standardized behavioral tests with well-characterized psychometric properties and developmental, usually age-based, norms. This allows researchers to compare patterns of performance between groups. Perhaps surprisingly, most of the raw data underlying the norms remain private. So, with few exceptions, researchers seeking access to measure-specific data collected by others will find it almost impossible. A number of standardized measures are published by commercial entities, and so economic interests may conflict with the ideal of greater data availability. However, widespread data sharing remains relatively rare even where measures developed by academic researchers and made freely available are concerned. Data sharing initiatives

among child language researchers (CHILDES; WordBank; HomeBank) are notable exceptions.

Commercial Datasets

Large-scale datasets collected by private entities for business purposes form a final group. The data collected by private entities about individuals largely concern consumer behavior although health and fitness related data constitute a third category (e.g., Fitbit; `http://www.fitbit.com`). A growing number of smartphone apps enable parents to collect data on their own children for personal purposes (e.g. `https://www.baby-connect.com`). Commercial datasets are sometimes made available to academic researchers, but the policies that govern data access are under the control of the entities that provide the services.

Some large-volume sources of developmental data are collected and managed by private, non-academic or government entities. For example, 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 or 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.); 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. 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. Clearly, the scale of data collected and managed by non-academic entities dwarfs that of other providers. 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 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 (Goodman et al., 2014).

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. These reduce the manual labor involved in preparing a survey and processing completed data for analysis. Developmental research commonly use

behavioral measures involving computer-based tasks, 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. Some researchers have begun to use tools such as Amazon's Mechanical Turk (`http://www.mturk.com`) or Apple's HealthKit (`https://developer.apple.com/healthkit/`) to conduct large-scale behavioral science experiments (e.g., `https://autismandbeyond.researchkit.duke.edu`). These sites deliver data in well-structured electronic formats, sometimes using tools specialized for psychological research (e.g., PsiTurk, `https://psiturk.org`). 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.

Large numbers of developmental researchers collect video and audio recordings. 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 generate large, dense files and come in a diverse formats. With few notable exceptions (e.g., Databrary, `http://databrary.org`, and the MET Project) most existing data archives support the storage and sharing of text files, but not images (including brain images), audio, and video data.

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. Lab-based tools for conducting physiological measurements such as EEG, heart-rate or skin conductance,

produce electronic files. 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`, `https://autismandbeyond.researchkit.duke.edu`), activity patterns, facial expressions, and momentary cognitive, and emotional states broadly available to researchers. For example, a new class of wearable devices for infants and children has arisen (e.g., `https://www.owletcare.com`, `http://mimobaby.com`, `http://www.sproutling.com`), coupled to parent-controlled child tracking apps mentioned previously. These tools enable the collection of data from large numbers of people in short periods of time, significantly enlarging the volume, velocity, and variety of data available for analysis. Whether and how the data can be made available for academic and medical research in developmental science remains an open question.

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`) to build tools that allow researchers to upload session-level video and flat-file data to a secure web-based server as the data are collected, thereby 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 made discoverable and accessible to other 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. Available datasets may lack persistent, citable, searchable, identifiers (e.g., digital object identifiers or DOIs). 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; `https://ndar.nih.gov`) maintain extensive standardized metadata about tasks and participant demographics. This can help investigators to search for specific data sources. But, not all datasets support variable-level search, and supplementing datasets with extensive metadata requires expertise and financial resources many research teams lack. 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 edited and 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 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, thus reducing the curational burden. 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, observations, statistical analyses, and visualizations in human-friendly 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. Still, 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 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 must be supervised by an Institutional Review Board or its equivalent, and many institutions conducting research with human subjects that is not federally-funded follow the same procedures. IRBs 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 that 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 researchers remove or alter data items that could reveal a participant's identity to reduce the likelihood of disclosure. Whether deidentified data of this sort can be shared with researchers outside the original IRB-approved team that collected it depends on several factors. One factor is the sensitivity of the data collected and the likelihood that specific identities or home locations 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 and thus requires no additional approval. Many big datasets in developmental science have restrictions on access either because the data were collected under Federal regulations that prohibit releasing individually 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 secure and document 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 recommends

that permission to share be sought separately from consent to participant in research and after a given data collection episode has ended. 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 remaining 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 not subject to 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 the 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 only be minimized and managed, but not entirely eliminated. The DataTags Project (`http://datatags.org` at Harvard provides example of a practical solution that may help researchers navigate the complexities of sharing data in the future. DataTags seeks to make the process of determining what risks particular datasets pose and provide a practical way of "tagging" datasets based on that level of risk.

Of course, there are unresolved questions about privacy protections in the consumer domain that have the potential to influence public attitudes toward academic research (Meyer, 2014).

Transparency and Reproducibility. Another 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 published research findings are false (Ioannidis, 2005), and true effect sizes 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 replicability problems exist in developmental science and whether they constitute a crisis is unknown. Undoubtedly, 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). Still, 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 scientific practices, as will the widespread adoption of more consistent journal practices related to these issues (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 specifically for this purpose (e.g., OSF,

Databrary, RStudio, Jupyter).

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 in data practices. Many funders require data management plans, 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 are beginning to 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). One problem with data sharing mandates from funders is that there is no specific mechanism to provide ongoing financial support to data archives. Another is that few researchers budget funds to support data management and archiving and with increasing competition for grants, may be reluctant to do so. 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 developmental science.

Conceptual and Theoretical Issues

The increasing availability of big datasets for analysis in developmental research poses significant theoretical and conceptual questions alongside the pragmatic ones already discussed. Big(ger) data may help to overcome limitations with our existing knowledge base. Specifically, big data may help mitigate 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 psychological science, 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 may still yield conclusions that are locked in time.

Another 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). Developmental research reflects many of these characteristics, but the collection, analysis, and sharing of larger datasets should work to reduce their impact.

Developmental research based on big data faces a specific point of tension related to measurement. Many of the measures for which high volume data are available come from proprietary, expensive instruments such as the Bayley and the WIPPSI for which baseline data about population norms are unavailable. 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 are commonly used.

Although new technologies make it possible to carry out large-scale experimental studies with developmental populations (e.g., LookIt, PsiTurk), big data techniques often invoke some form of correlational analysis. This makes causal inference problematic at best. Indeed, some critics have raised concerns that the rise of 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 do not necessarily lead to deeper insights (Graham, 2012). Some data intensive fields, largely in computer science, have adopted theory-free approaches to discovery. But, developmental science has a rich and rigorous intellectual history in which theory, correlational analyses, and experiments play central, essential roles in scholarly discourse. It's vital that tradition continue.

Conclusion

As boyd and Crawford (2012) observe "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 discoveries have the potential to improve health and maximizing the potential for human achievement.

However, that potential is limited because most developmental science data are hard to find and cumbersome to access, even for researchers. Data that are available have restrictions that largely prohibit analyses at the level of individual participants. Most data

linked to publications are not stored in open data repositories. 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 in some subfields a degree 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.

Still, 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 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. Investigator-initiated projects with the largest impacts have attracted significant intellectual communities around the datasets that extend the beyond the boundaries of the original investigative teams. Thus, the impact of existing big datasets appears tightly linked to the degree to which information from them is widely shared. This suggests that the future of big data approaches in developmental science depends on the extent to which barriers to data sharing can be overcome.

Technical issues about data formats, storage, cleaning, visualization, and provenance remain, but significant progress has been made in addressing them. Developmental researchers have available a growing array of data repositories (CHILDES, Databrary, Dataverse, ICPSR) and new data 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 (Goodman et al., 2014). New ethical procedures

for seeking informed consent to share identifiable data have been developed and are being implemented in diverse research contexts. These promise to accelerate the reuse of data which has previously been difficult or impossible to share widely.

We should remember that Facebook was launched in 2004, Twitter in 2006, and the iPhone in 2007. It would be short-sighted 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 build upon data about the multiple facets of human development, as many are beginning to do, we should 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.

Goldsmith, H., & Rothbart, M. (1993). The laboratory temperament assessment battery (lab-tab). *University of Wisconsin*.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., ... Slavkovic, A. (2014, April). Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput Biol*, *10*(4), e1003542. Retrieved 2015-09-02, from http://dx.doi.org/10.1371/journal.pcbi.1003542 doi: 10.1371/journal.pcbi.1003542

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

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

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.

Pearson, H. (2015, October). Massive UK baby study cancelled. *Nature*, *526*(7575), 620–621. Retrieved 2015-11-28, from http://www.nature.com/doifinder/10.1038/526620a doi: 10.1038/526620a

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

Rothbart, M. K. (1981). Measurement of temperament in infancy. *Child development*, 569–578.

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

Tamis-LeMonda, C. (2013). *Language, cognitive, and socio-emotional skills across the first 6 years in u.s. children from african-american, dominican, mexican, and chinese backgrounds.* Databrary. Retrieved from http://dx.doi.org/10.17910/B7CC74 doi: 10.17910/B7CC74

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.

Table 1

Illustrative big datasets hosted by governmental entities

Dataset	Sample	URL
Current Population Survey (CPS)		http://www.bls.gov/cps/demographics.htm
Danish National Birth Cohort	n=100,000 women recruited 1999-2002, assessed during pregnancy and when children were 6 mos, 18 mos, 7 yrs, and 11 yrs	http://www.ssi.dk/English/RandD/Research%20areas/Epidemiology/DNBC/
Early Childhood Longitudinal Study (ECLS)	Birth cohort (ECLS-B): n=14,000 children born in 2001 from birth through kindergarten; ECLS-K cohort: kindergarten through 8th grade; ECLS-K:2011: kindergarten through 5th grade.	http://nces.ed.gov/ecls/
Japan Environment and Children's Study (JECS)	n=100,000 parent-child pairs	http://www.env.go.jp/en/chemi/hs/jecs/
National Health and Nutrition Examination Survey (NHANES)	nationally-representative sample of n=5,000	http://www.cdc.gov/nchs/nhanes.htm
National Longitudinal Survey of Youth (NLSY)	n=10,000 youth born 1957-1964 surveyed beginning in 1979.	http://www.bls.gov/nls/
National Youth Fitness Survey (NYFS)	n=1,650 3-15 year-olds surveyed in 2012	http://www.cdc.gov/nchs/nnyfs.htm
NICHD Study of Early Child Care and Youth Development (SECCYD)	n=1,364 infants born in 1991	https://www.nichd.nih.gov/research/supported/Pages/seccyd.aspx
NIH MRI Study of Normal Brain Development	n=554 4-18 year-olds	http://pediatricmri.nih.gov/nihpd/info
NIMH National Database for Autism Research (NDAR)	n>85,000 individuals	http://ndar.nih.gov
Norwegian Mother and Child Cohort Study (MoBa)	n>90,000 pregnant women and n>70,000 men, recruited 1999-2008	http://www.fhi.no/eway/
WHO Multicentre Growth Reference Study (MGRS)	n=8,500 0-24 mo-olds and 18-71 mo-olds	http://www.who.int/childgrowth/en/

Table 2

Illustrative big developmental datasets hosted by non-governmental entities.

Dataset	Sample	URL
National Longitudinal Study of Adolescent Health (AddHealth)	Wave I: n=90,118 adolescents in-school questionnaires, n=20,745 in-home interviews; Wave II: n=14,738 adolescents in-home interviews; Wave III n=15,197 young adults in-home interviews and biomarker collection; Wave IV: 15,701 adult in-Home interviews and biomarker collection.	http://www.cpc.unc.edu/projects/addhealth
Child Language Data Exchange System (CHILDES)	n>3,000 6 mos - 8 yrs	http://childes.psy.cmu.edu/
Colorado Adoption Project (CAP)	n=450 families (>2,400 individuals)	http://ibg.colorado.edu/cap/
Databrary	>3,400 hours of video from n>2,900 participants, 2 mos - 50+ yrs; experimental displays	http://databrary.org
Developing Human Connectome Project	20 to 44 weeks post-conception	http://www.developingconnectome.org
Family Life Project (FLP)	n=800 sampled at 2, 6, 15, 24, and 36 mos	http://flp.fpg.unc.edu
Genome of the Netherlands Project (GoNL)	n=250 familes	http://www.nlgenome.nl
Human Speechome Project	10 hrs video/day from 1 child from 0-3 yrs	http://www.media.mit.edu/cogmac/projects/hsp.html
Maternal Lifestyle Study (MLS)	n=1,388 infants (and mothers/caregivers) assessed at 4, 8, 10, 12, 18, 24, 30, and 36 mos	https://neonatal.rti.org/about/mls_background.cfm
Measures of Effective Teaching Project (MET)	n>3,000 teachers recorded beginning in 2009	http://www.metproject.org
MetroBaby	n>1,000 racially and ethnically diverse children video-recorded at home or in the lab at 14, 24, 36, 52, 64, and 76 months	(http://doi.org/10.17910/B7CC74
Panel Study of Income Dynamics (PSID)	n=5,000 families (>18,000 individuals)	https://psidonline.isr.umich.edu
Pediatric Imaging, Neurocognition, and Genetics (PING)	n=1,000 3-20 year-olds	http://pingstudy.ucsd.edu
Progress in International Reading Literacy (PIRLS)		http://timssandpirls.bc.edu
Psychiatric Genomics Consortium (PGC)	n>900,000	http://www.med.unc.edu/pgc
Trends in International Math and Science Study (TIMSS)	n>20,000 per sample	http://timss.bc.edu
Trends in International Math and Science Study Video (TIMSSVideo)	teachers recorded in n=7 countries	http://www.timssvideo.com
Twins Early Development Study (TEDS)	survey data from n=15,000 twins born in 1994-1996; lab-based behavioral task and DNA samples from a smaller subset	http://www.teds.ac.uk
Twin and Offspring Study in Sweden (TOSS)	target of n=900 parents and families (3,000 individuals)	http://ki.se/en/meb/twin-offspring-study-in-sweden-toss

Table 3

Illustrative measure-based datasets.

Measure	Comments
Bayley Scales of Infant and Toddler Development	(Bayley, 2006). Data underlying published norms not available, but some study-specific data is available: http://www.epicure.ac.uk and http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4091.
Computer-based cognitive tutor data	LearnLab (http://learnlab.org/technologies/datashop/index.php) hosts DataShop repository/data analysis tool.
Infant Behavior Questionnaire (IBQ)	(Rothbart, 1981). Information at http://www.bowdoin.edu/~sputnam/rothbart-temperament-questionnaires/instrument-descriptions/infant-behavior-questionnaire.html. IBQ data on n=1,388 participants archived at ICSPR as part of the Maternal Lifestyle Study (MLS; https://neonatal.rti.org/about/mls_background.cfm)
Laboratory Temperament Assessment Battery (Lab-TAB)	(Goldsmith & Rothbart, 1993). Information available at http://www.uta.edu/faculty/jgagne/labtab
MacArthur Communicative Inventory (CDI)	n>40,000 samples archived at WordBank (http://wordbank.stanford.edu)
Spatial Learning	NSF-funded Spatial Intelligence and Learning Center (SILC; http://spatiallearning.org) stores information about research tests and intruments
Teller Acuity Cards	Publications citing norms for 0-4 year-olds have been published (Teller, McDonald, Preston, Sebris, & Dobson, 1986; Mayer et al., 1995; Salomão & Ventura, 1995)
Wechsler Intelligence Scale for Children-III (WISC-III)	Designed for children ages 6-16. Data underlying norms not available. Published by Pearson (http://www.pearsonclinical.com/psychology/products/100000771/wechsler-intelligence-scale-for-childrensupsupfifth-edition--wisc-v.html)
Wechsler Preschool and Primary Scale of Intelligence-R (WPPSI-R)	Designed for children aged 4-6 1/2 years. Data underling norms not available.