Behavior Research Methods (2019) 51:1754–1765
https://doi.org/10.3758/s13428-018-1167-8

A modular, extensible approach to massive ecologically valid behavioral data

Mark VanDam[1] · Paul De Palma[2]

Published online: 16 November 2018

Abstract
We explore here the application of modern computer hardware and software to the collection and analysis of behavioral data. We discuss the issues of ecological validity, storage and processing, data permanence, automation, validity, and algorithmic determinism. Taking the modern landscape into account, we demonstrate several varying projects we have recently undertaken as proofs of concept of the viability and utility of this approach. In particular, we describe four research projects, which involve work on child-directed speech; the application of automatic methods to clinical populations, including children with hearing loss; quality control and the assessment of validity; and the sharing of data in a public database. We conclude by pointing out how the methodology described here can be extended to a wide variety of interdisciplinary and detailed projects that are likely to lead to better science and improved outcomes for populations served by the behavioral, social, and health sciences.

Keywords Big data · Speech · Very large database · Natural language · Automatic speech recognition ASR · Biosensor · Ecological validity

Natural observations have the potential to reveal human behaviors that are not evident in a controlled laboratory setting. Recognizing this, but constrained by human and technical considerations such as the investment of human labor, financial cost, and computational limitations, social scientists have historically collected behavioral data in very small samples or under conditions that have presented threats to their ecological validity (Narayanan & Georgiou, 2013; Neisser, 1967). Many well-documented deleterious or undesirable consequences have followed from this methodological approach, including the Hawthorne effect, in which study participants change behavior due to their own knowledge of being observed (McCambridge, Witton, & Elbourne, 2014); observer biases related to biosocial factors (e.g., a physically larger male researcher interacting with a female participant) and psychosocial factors (e.g., differential effects of undiagnosed disorders in males versus females; Rosenthal & Rosnow, 1991); demand characteristics, in which participants' assumptions about the goals of the research affect their behavior (Orne, 2009); and potential questions about the generalizability of results (Brewer & Crano, 2000; Cook & Campbell, 1979). Even in the best of studies, the sample sizes have been very small. A well-known example from the literature is representative of these criticisms. In the mid-1990s, Hart and Risley (1995) argued that the number of hours that parents converse with children is the single strongest predictor of academic success. Hart and Risley's work was constrained by factors familiar to any social scientist: namely, the labor-intensive nature of collecting, transcribing, and tagging audio-recorded speech and language samples. They studied 42 children for one hour per month over a two-and-a-half-year period, averaging about 28 total observations per family. This produced about 1,200 h of total collected audio, resulting in over 30,000 pages of transcripts. It took "six years of painstaking effort before we saw the first results" and "none of us took a vacation for more than 3 years" (Hart & Risley, 1995, pp. 46, 41). To get a sense of how things have changed in just two decades, the authors of the present article published a study of child-directed speech (De Palma & VanDam, 2017) whose conclusions were based on an analysis of 7,000 h of recorded speech, the overwhelming majority of which was processed automatically by a computer. Similarly, we currently have over 20,000 h of in situ family conversations awaiting analysis.

✉ Mark VanDam
mark.vandam@wsu.edu

[1] Speech & Hearing Sciences, Elson S. Floyd College of Medicine, Washington State University, Spokane, WA, USA

[2] Computer Science, School of Engineering and Applied Science, Gonzaga University, Spokane, WA, USA

Current advances in micro-electronics and the ability to store, process, and interpret vast amounts of data afford scientists new avenues of investigation and resolve some of the earlier problems, even problems that were unseen or unconsidered in previous eras of the research enterprise. These advances include (a) commercially available, wearable biotechnology capable of capturing rich data on human behavior, (b) archival databases of raw and processed data, and (c) computational techniques, such as advances in automatic speech processing and recognition (ASP/R). These have been further enhanced by the discovery and rediscovery of computational and statistical models to interpret the rich, complex data that are generated. Computational and statistical advances and rediscoveries have included Bayesian inference, so-called "robust statistics" (Box, 1953; Gelman & Hennig, 2017; Huber, 1981), and the recognition that assumptions such as normality and homoscedasticity may not be appropriate for very large datasets. Also important are the easy availability and general familiarity of statistical and numerical software packages such as R (R Core Team, 2013), Python's SciPy (Jones, Oliphant, & Peterson, 2014), commercial and third-party quantitative software packages for MATLAB (MathWorks, 2018), and user-interface-driven applications such as SPSS (IBM Corp., 2013) or RStudio (RStudio Team, 2015) that increase the accessibility of analytic techniques (McGrayne, 2012; Wilcox, 2016; Wilcox & Keselman, 2003). The importance of these discoveries and tools cannot be overestimated.

With the rapid rise in the abilities of modern computing, the collection and storage of vast amounts of data are fairly straightforward and, in many applications, pose few new technological challenges. Commercial computing began handling large datasets in a straightforward manner within a decade after IBM researcher E. F. Codd proposed separating the data model from the underlying hardware (Codd, 1970). But, in the era of very large datasets, a number of extratechnological challenges remain. For example, wearable sensors for a disabled or disordered population may pose significant challenges to obtaining data or dealing with noisy, messy, and missing data (Milliken & Johnson, 2001, 2009). It is one thing to record and store everything that is heard (for example), and quite another to assess its validity. An especially thorny problem in the social, behavioral, and health sciences is a variant of the issue that accompanied the rise of the personal computer in the 1980s: the administration of complex systems is itself complex. The administrative costs of hardware may be many times the cost of the hardware itself, perhaps by as much as a factor of 5, as the Gartner Group has been pointing out for more than a decade (Cearley, Burke, Searle, & Walker, 2017; van der Meulen, 2008). Who does this administration is contested territory. A research team may pay for it out of an already-stretched budget, or it becomes the de facto territory of graduate students, postdocs, or, in small research teams, the researcher herself—all amateur information technology (IT) administrators whose time in a well-funded world might be better spent doing the content of the science at hand. Our colleagues in the physical sciences share the same issue, leading to an early recommendation for "science data centers" that "would curate one or more massive datasets, . . . the applications that provide access to that dataset, and . . . staff that understands the data" (Gray et al., 2005). Among the examples cited are CERN and Fermilab.

The structure of and access to very large datasets has of course received much attention since the earliest uses of vacuum-tube computers and monolithic databases stored on tape drives. But the advent of massive increases in the use and storage of computing technology throughout the second half of the 20th century has resulted in new and much more massive datasets. Early massive data archives have been described in terms of "warehousing" (Jarke, Jeusfeld, Quix, & Vassiliadis, 1999), but this approach was shown to have the deleterious effect of isolating or siloing data. Warehoused data is to some degree decontextualized data. One approach to addressing this problem was the concept of a data lake (Dixon, 2010; O'Leary, 2014). A *data lake* is a large body of data filled, piped, or ingested from many sources, in which "users of the lake can come to examine, dive in, or take samples" (Dixon, 2010). There are also detailed methods (e.g., Constance [Hai, Geisler, & Quix, 2016]) and technologies (e.g., Hadoop, Hadoop + Spark, Splunk, Storm [Mohanty, Bhuhan, & Chenthati, 2015]) to deal with the resulting data lake. There has been substantial interest and analytic activity (particularly using Hadoop) from Microsoft, Amazon, Google, Oracle, Facebook, Yahoo!, and others (Pasupuleti & Purra, 2015). The approach and resources discussed here, namely the HomeBank resource detailed below, are consistent with a data lake.

In addition to the general data storage and processing described immediately above, a number of projects or tools seek to aggregate or serve as a repository for data, often in the form of actively ongoing, open data submission archives in which users continually submit new data to the repository or data lake. These also provide (open or restricted) access to that data for the purpose of facilitating new empirical discovery for researchers who may not have collected the data. The advantages of this approach are that the costs, broadly construed, are reduced and data permanence increases; the disadvantages include new management responsibilities for the curators and potential breaches of personal or other data unintended for public consumption. Of particular interest to behavioral scientists are the Open Science Framework (Nosek et al., 2015), CLARIN (Hinrichs & Krauwer, 2014), Databrary (Adolph, Gilmore, Freeman, Sanderson, & Millman, 2012), and TalkBank (MacWhinney, 2007). In addition, there are projects whose details are only partially available to the public or to the research community, including massive ongoing behavioral data collection and archiving within Google, Amazon, Microsoft, Netflix, and other tech-based enterprises. At least some data in the private sector are specifically oriented toward language and speech analysis, detection, perception, or other

similar areas of interest that generate massive datasets. Not least of these are the Amazon Alexa, Apple Siri, Microsoft Cortana, and Google Assistant speech recognition and language-processing projects (Chung, Iorga, Voas, & Lee, 2017).

Along with large datasets comes the related issue of data permanence, an issue recognized nearly two decades ago in the context of digital libraries (Rothenberg, 1999). In this case, the rapidity of software and hardware change becomes a liability. The researcher who does not maintain updates is condemned to live on a technological island whose landmass will shrink to zero as hardware and software become unusable, through obsolescence, reduced interoperability, or simple mechanical failure. Recognizing the problem, of course, points to a relatively straightforward solution: maintain software updates and data backups, and copy data to new hardware periodically. These tasks will be performed as a matter of course by professional IT administrators, and perhaps not so routinely by (social, behavioral, and health) researchers who, with their postdocs and graduate students, frequently maintain their own hardware and software.

As technology changes, it is not at all clear what the best practices will be. To offer a simple example, the fairly recent introduction of Amazon Web Services has the capacity to put social scientists out of the hardware business. But technology comes with unintended consequences, and as Edward Tenner and others have observed, in the context of very-large-database computing, not the least of these relate to data privacy (Raghupathi & Raghupathi, 2014; Tenner, 1996). Though the problems are widely recognized and potential solutions exist, one might be forgiven if he or she claimed that we have simply substituted a software problem (how to maintain secure data stored remotely) for a hardware problem (how to maintain large fast computing clusters and data repositories locally while sharing with other researchers).

At a level beyond the collection, storage, and maintenance of data is interpretation of the data. With very large datasets, this has taken on an increasing urgency. That is to say, data collection and all its attendant processes are often less challenging than being able to generalize meaningful trends in the data. Without substantial tools to organize and analyze the collected data, the advantages of ecological validity, the ability to look longitudinally at the time course of human behavior, and the ability to generalize would be diminished. It is the case that algorithms that perform efficiently on small datasets can require days or weeks to run on very large datasets, or may never complete their functions due to hardware limitations such as available memory. Matrix multiplication is a simple example. The number of computations necessary to multiply matrices grows in a nonlinear fashion as the cube of the size of the dataset. Furthermore, as anyone who has tried to send a dataset over the Internet surely knows, network bandwidth has not kept pace with storage capacity.

A related theoretical problem is the necessity of automation in the processing of very large datasets. The demand for automaticity is in fact one of the hallmark definitions of very large datasets, or so-called *big data*: a dataset that is larger than traditional means can accommodate (De Mauro, Greco, & Grimaldi, 2016). In any analysis that depends on automatic, deterministic, or algorithmic methods of computation, there is the possibility of systematic error being introduced into the analysis. Systematic errors or biases in machine learning and artificial intelligence (of which ASP/R is a subtype) have been demonstrated in large datasets, with biases documented in such disparate domains as insects, flowers, race, gender, careers, and first names (Caliskan, Bryson, & Narayanan, 2017); clearly, the potential for bias is great. There are traditional methods to confirm the performance of the automation by quantifying precision, accuracy, sensitivity, and specificity, but those methods can be challenging to implement reliably in large datasets. Importantly, the structural correspondence between the measure and the description, the *validity*, must also be considered. In practice, for sufficiently small datasets, a common assessment method is to examine some subset (or a holdout set) of the data by hand, often 10%–20% of the overall quantity of data, and compare that hand-derived gold standard to the results obtained from the automatic method.[1] If the two results converge, they are considered valid, the results are generally trusted, and the automatic method is accepted (possibly with some conditions, such as higher error rate). But with very large datasets, validation is more challenging, due to the inability to compute by hand an appreciably large gold standard set with which to compare the automatic methods. The generality of the subsample increases as its proportion to the whole increases, but given the samples discussed in the present work, no reasonable effort could attain a hand-derived subsample of even 1%, when 20%–40% is common in the literature (see Kohavi, 1995, and Kim, 2009). For example, in some of our own work we have performed automated analyses on more than 10,000 h of audio collected in home environments. Hand transcription to the level of detail required for these analyses (here, a careful orthographic transcription—but a phonetic transcription would be substantially more labor-intensive) is roughly a ten-to-one time investment for an experienced transcriber. So, to transcribe all 10,000 h by hand would take approximately 100,000 h of labor, or roughly 50 person-years. A holdout or subsample of even 1% would consume an estimated six months of full-time labor from an experienced transcriber. Instead, to assess the validity of automatic methods, we rely not only on gold-standard comparison, but also on content, criterion, and construct validity—all

[1] Other methods, such as leave-one-out cross-validation, are alternative frameworks for the validation of datasets, but other methods are computationally heavy or have other trade-offs.

theoretical measures used to argue that the empirical quantities reflect the things we claim they reflect.

Yet another issue in working with very large datasets is that, of necessity, some processing software is probabilistic. For example, much of our work relies upon speech-processing software from the LENA Foundation (LENA, 2018). The advances in automatic speech recognition in the past two decades have been due to the introduction of probabilistic machine-learning models, and, like the humans they are intended to simulate, they make mistakes. Although Google's use of very large datasets along with yet another machine-learning technique, neural networks, has brought ASP/R to near-human performance, LENA is a commercial product (Hinton et al., 2012; Kepuska & Bouthota, 2017), so its details are opaque. Our approach has been to test the accuracy of LENA against the gold standard in automatic speech recognition—namely, human subjects (Silbert, Linck, & VanDam, 2013; VanDam & Silbert, 2016). Though LENA performs well, it does not perform perfectly. And the errors that any computer or automatic method makes are not the same kinds of errors that humans make in language perception. The fact that computers and humans are known to use different storage, analysis, and production mechanisms (Woods, Dekker, Cook, Johannesen, & Sarter, 2017; Woods & Roth, 1988) may introduce problems, both in fine-tuning the output and in a deeper understanding of the mechanisms themselves, whether the mechanisms are the product of human cognition or the often opaque models used to specify these mechanisms (such as in hidden Markov models or computational neural networks).

Outline of the present work

What follows in this report is a description of the technology we use, followed by a series of case studies of several recent empirical projects undertaken by our research group. We use the methodology to explore new scientific questions by capitalizing on the confluence of the technological advances and capabilities set forth above. We use our own empirical work as an example and proof of concept. We demonstrate how the modularity, flexibility, and extensibility of the modern landscape accommodates new areas of scientific inquiry in a wide range of disciplines with both theoretical and practical importance, by implementing techniques that can be shifted or substituted to accommodate different problems. For example, we demonstrate the use of a particular spectral feature from acoustic speech processing, but the same general procedure could be adapted to other spectral or temporal features, or indeed, to any raw, high-density data, including images and data from sensors or sensor networks. We conclude with short discussion of the extensibility of the framework described here.

Description of the technology, its extension, and its application

The LENA system was developed in the early 2000s in an attempt to use automatic methods to capture the linguistic input to children and their auditory environment (Gilkerson & Richards, 2009). The commercially available system consists of a small, body-worn audio recorder to collect audio from the child's perspective and software to process the collected daylong audio. Audio is collected in units of one day, with audio ideally being recorded continuously from when the child awakes in the morning until the child goes to sleep in the evening. After a raw recording is collected, the daylong audio file is transferred to a computer where it is processed using the proprietary LENA software. The goal of the processing is to segment the audio and assign one of about 60 a priori labels to each segment. The segment size depends on the auditory signal and is not of a predetermined duration. Segments ultimately bearing labels corresponding to live human vocal events are generally the size of an uninterrupted utterance, and not at the level of individual words or syllables. The goal of many artificial speech recognition systems is to map audio onto an orthographic representation of lexical items, with the output being a readable transcript of the audio event. This fairly specific, granular goal has obvious utility, but performance declines substantially as noise increases, a serious problem for daylong recordings in which the environment is overtly uncontrolled. Ecological validity and environmental noise are, in the context of naturalistic audio recordings, proportional to one another. To address this problem, the LENA algorithms assign labels at a relatively coarse level, such as *adult male*, *adult female*, *TV/electronic media*, *silence*, *noise*, *target-child*, and so on. The output of the processing is a raw audio file (16-bit, 16-kHz, lossless pulse-code modulated WAV format) and an XML-coded record of the onset–offset times at centisecond resolution bearing the label of each segment. There is also some additional information in the output, such as the mean amplitude of the signal for the duration of that segment. The audio files are as long as 16 h (the limit of the internal memory of the recorder) but are typically less. In a database maintained by the present authors of several thousand recordings, the average daylong recording is just under 11 h. In our case, that database consists of roughly 2,500 recordings from about 120 families. Some families have contributed one recording, and one family has contributed more than 600 recordings from three children, with one child wearing the recorder in about 300 of her first 400 days of life. The original design of the database was to collect one recording per month from each family for one year in early, preschool development. This goal was largely realized, but many factors have influenced the database since its inception in the mid-2000s. For example, families enrolled in the data collection study with a target preschooler might have a school-age sibling who volunteered

to contribute recordings. Similarly, families with certain disorders not targeted in the original design might be included (and documented) if it were convenient to collect such recordings. In principle, we have taken the viewpoint that recordings have a fairly high cost to collect, but we have an established protocol and mechanism, so unplanned opportunities to collect recordings are generally accommodated in the hope that those recordings may facilitate future scientific discovery. The time-aligned record of segment labels depends on the content of the recording, but typically it runs to roughly 25,000 lines of output.

The output of the LENA system can be interpreted directly, in a limited fashion, via the interface software provided with the processing module. Summary reports quantifying child and adult utterances, conversational exchanges, TV and electronic media exposure, noise, and silence are available via the entailed software. Options to output the data into various interpreted formats, such as CLAN/CHAT or Transcriber files, are also available. Using the output of the ASP/R software to provide time-aligned labels and the (raw) audio can be accomplished with third-party or custom-written software, such as the collection of user scripts at the HomeBankCode repository (https://github.com/homebankcode/) or the ADEX parser, provided by the LENA Research Foundation. For example, a researcher may want to quantify certain spectral characteristics of all audio segments labeled *noise* by the ASP/R routines. The researcher might use custom scripts to extract the onset and offset times of those segments, read and process the corresponding raw audio from the daylong file, and return the acoustic features of interest.

Our approach to collecting and analyzing large amounts of ecologically valid data can be decomposed into four principal modules. We describe each in terms of our own work, but argue later in this work that there is wide flexibility and extensibility in the modules described here. First, ecologically valid recordings of human behavior are collected in real time through a body-worn recorder. In our case, we collect daylong audio recordings from the LENA system described above. Second, the raw data are processed by commercially or publicly available software. For us, the daylong audio is processed with LENA ASP/R, taking the daylong audio file as input and outputting a tagged, time-aligned record of segment boundaries and labels known to be important for speech and the development of speech and language in children. Third, we use custom software to process the results for features of interest. We use MATLAB, C, and Python code to process for speech and language features such as utterance duration or fundamental frequencies of child and adult speech. The ability to use the output of commercially available processes as the input for specialized analyses affords the unique possibility of interdisciplinary applications and specializations not previously considered or not available directly within the LENA software. Fourth, massive data are subjected to structural data organization and computational modeling techniques, to gain insight into the structure of the data. The goal of some empirical work is to describe; a loftier goal of empirical work is to use the data to draw larger conclusions and to generalize from the sample to the population with some degree of confidence. In the case of the massive data described here, we are ultimately interested in offering the potential to shed light on human behavior, the structure of the brain, social cognition, or other understanding of complex systems at play, in such a way that the conclusions carry forward, the observations generalize, and the results guide predictive hypotheses. The general structure of these modules, issues important to consider, and examples at each stage are given in Fig. 1. An important feature of this design and implementation is the extensibility of this approach. In particular, each module and the processes contained therein are easily substituted, rearranged, manipulated, added to, or otherwise tailored to a wide array of raw input data, so as to gain empirical insight. So, for example, at present there are no real alternatives to the LENA system for processing daylong audio files, but efforts are underway to achieve a similar level of ASP/R, including the open source Virtual Speech Kitchen project (Metze, Fosler-Lussier, & Bates, 2013). Furthermore, any given single database ingested into such as system could be subjected to many alternative investigations by tailoring the method to the questions of interest. Several examples of such projects are given in the following section from our own work.

Detailed examples of recent work

Just as the development of core memory, the transistor, and the high-level programming languages in the 1950s ushered in the cognitive revolution, the formalist approach to the study of language has made available to researchers the use of computers to transcribe speech, now known as ASP/R. ASP/R, along with current developments in computing and computational techniques, has permitted the investigations we have undertaken. These developments include not only inexpensive storage and recording devices, but also the rediscovery of Bayesian inference, the application of probabilistic machine-learning techniques—the hidden Markov model, in particular—to speech processing, and the expansion of Internet bandwidth. The latter has allowed both world-wide collaboration and remotely accessible data repositories. In the following four subsections, we describe using this collection of methodologies, tools, and techniques to explore child-directed speech; the application of automatic methods to clinical populations, including children with hearing loss; quality control and assessment of validity; and the advantage of sharing data in a public database.

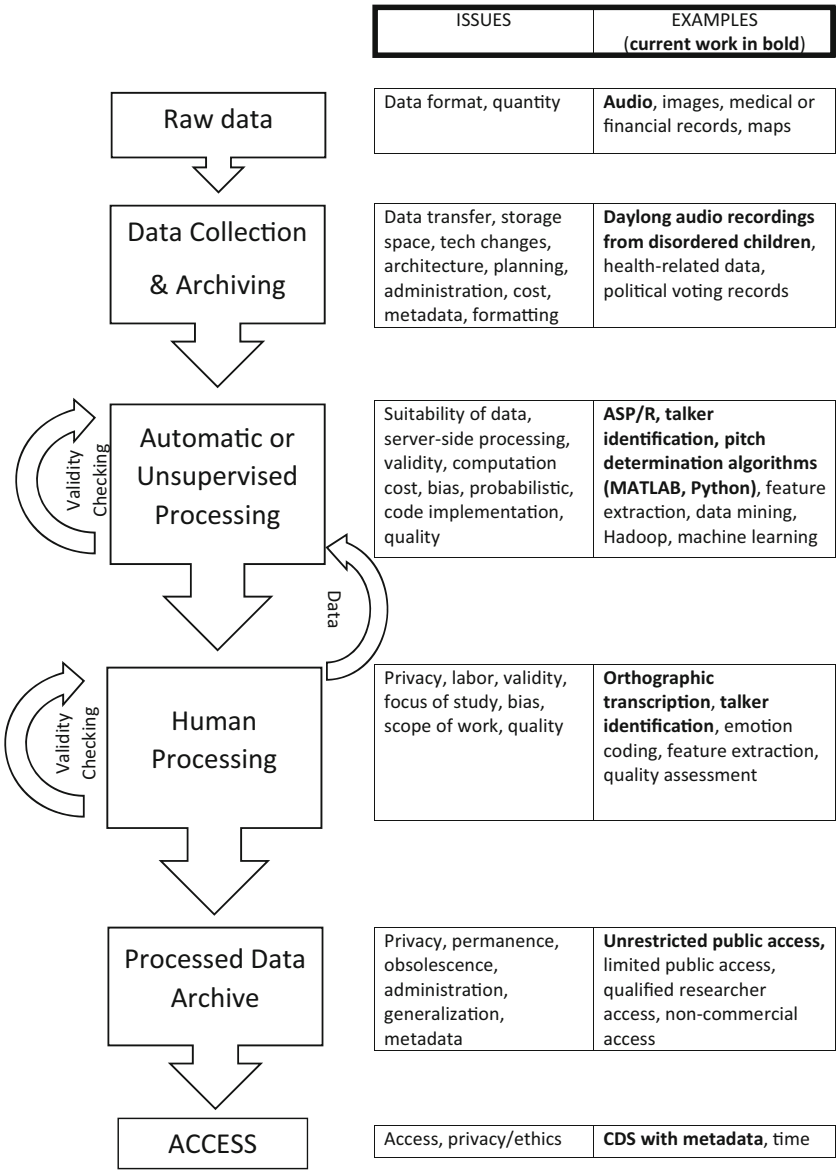

Fig. 1 Flow chart, issues, and examples for the collection and processing of behavioral data from a very large database. ASP/R, automatic speech processing and recognition; CDS, child-directed speech

Child-directed speech In two studies we used the phenomenon of child-directed speech (CDS) as a proof of concept—namely, as evidence that very large speech corpora could be analyzed in a reasonable time to produce results unavailable to other techniques (De Palma & VanDam, 2017; VanDam & De Palma, 2014). CDS is the manner in which adults speak to their infants and toddlers. Characteristics of this speech are shorter utterances, hyperarticulation, decreased syntactic complexity, and raised fundamental frequency (f_0). The phenomenon has been attested in several European languages, Mexican Spanish, and Japanese, leading some researchers to suggest that not only might CDS be a language universal but also may be implicated in language acquisition (Masataka, 1999). Though CDS has been investigated since the 1970s, the availability of automatically processed daylong recordings has afforded a deeper look at the phenomenon.

LENA, as we noted above, is similar to a conventional automatic speech recognizer of the SIRI variety. The system transforms a time-stamped audio stream into a collection of feature vectors that it segments and labels. In essence, LENA, like any other recognizer, is a classifier, but instead of mapping audio segments to word classes, it maps them to the origin of the sound. Possibilities are *target-child*, *other child*, *male adult*, *female adult*, *TV/electronic sound*, and others. These labels permit automated judgments of CDS: a speech segment labeled *adult* and found adjacent to another adult segment is defined as non-CDS, and an adult segment adjacent to a segment labeled *target-child* is defined as CDS. Both studies described below used raised f_0 as a proxy for raised

pitch, since pitch and f_0 are correlated (Hess, 1983). Raised pitch is one of the most recognizable features of CDS. More importantly, f_0 can be extracted from WAV files and analyzed computationally. Interestingly, the automated methods also allowed us to explore whether CDS was different when performed by mothers versus fathers. This was a straightforward extension of the software: All that was required was for mothers and fathers to be identified separately by the automatic methods.

In one study we used over 490 h of daylong audio from 33 families as a preliminary investigation based on a small subset of our corpus (VanDam & De Palma, 2014). The study was done before we had developed software capable of traversing a file system with thousands of hours of recorded speech. There were two hypotheses:

> Mothers and fathers will produce a higher mean f_0 during CDS than during non-CDS.
> Mothers and fathers will produce comparable f_0 regardless of the hearing status of children; for this analysis, families with typically developing (TD) children were compared to families with a child who is hard of hearing (HH).

The results were surprising in two ways. First, during the CDS condition, mothers significantly increased f_0. Fathers, on the other hand, did not show an increase in f_0 in CDS as compared with non-CDS speech. Second, when comparing between the TD and HH groups, the fathers of HH children had a higher overall f_0, indicating that fathers might be sensitive to the hearing status of their children. By contrast, mothers' speech in this study did not differentiate whether they had an HH or TD child.

The second study used over 7,500 h of a corpus that has now grown to over 10,000 h (De Palma & VanDam, 2017). The number of families studied grew from 33 in the first study, to 62 in the second. This time we had developed software capable of traversing a complex Linux file system, building nearly four million 1- to 2-s WAV files from the much larger daylong recordings, and extracting f_0 from files designated as CDS and non-CDS. Software constructed for this study is available in the HomeBankCode repository (VanDam et al., 2016), about which we will say more below.

Curious about the strong finding from the 2014 study that fathers do not raise f_0 when speaking to TD children, we began with a single hypothesis: Mothers and fathers will produce higher mean f_0 during CDS than during non-CDS. To test this, we extracted the acoustic record of those segments automatically labeled as the mother or father by the LENA software, segregated into CDS and non-CDS. We processed those acoustic files according to a pitch determination algorithm, and then obtained the mean fundamental frequency estimates for each talker in each social condition. We then performed a paired *t*-test to test for within-talker difference between social conditions. We also performed an ordinary least squares regression to examine the trend. This time mothers remained sensitive to the CDS condition. Fathers also raised f_0 significantly, but not as much as mothers, showing a weaker correlation. The results are interesting in two ways. First, we demonstrated that not only is it possible to analyze a very large corpus, but also that the larger corpus corrected an incorrect inference drawn from a smaller corpus, namely that mothers but not fathers alter speech to their children in specific ways. Second, the differential speech behavior of fathers in the presence of children and adults has not been thoroughly addressed in the literature. We found that fathers' speech patterns were similar to those of mothers, but far from identical, over and above the general differences between male and female speech. Follow-up work will consider TD and HH populations separately, look at the potential role of siblings in the family communicative environment, consider other acoustic–phonetic features known to be important for speech and language, and examine the role of linguistic complexity in language development. The overarching goal here was to demonstrate that very large quantities of in situ speech could be recorded, stored systematically, analyzed through several layers of software (automatic speech processing, finely grained segmentation of WAV files by timestamp, f_0 extraction, and statistical analysis) and yield generalizable inferences.

Application to clinical populations Several studies have used the automatic methods described here to investigate speech production and language development and to describe the auditory environment of children with hearing loss. In addition to our own work looking at children with hearing loss, researchers have used automatic methods and the daylong recordings to look at preterm infants (Caskey, Stephens, Tucker, & Vohr, 2011; Caskey & Vohr, 2013), autism spectrum disorder (Rankine et al., 2017; Warren et al., 2010), Down syndrome (Thiemann-Bourque, Warren, Brady, Gilkerson, & Richards, 2014), language delay (Oller et al., 2010), and others. Here we will briefly describe three studies, to demonstrate the utility and productive nature of using automatic methods to describe and better understand hearing loss in children.

In one study we asked whether two-year-olds who were hard of hearing were exposed to the same quantity of talk from their parents as typically developing children from our own database and from a published sample (VanDam, Ambrose, & Moeller, 2012). We used daylong recordings processed by the LENA software to estimate adult words and conversational turns (normalized by the total hours in the recorded day). We also looked at possible covariates, such as hearing level, sex, socio-economic status, child language skills, and child age. We found that child hearing status had a limited influence

on the quantity of parents' talk to these toddlers, but within those families with a child with hearing loss, parents produced more talk as children's hearing was better.

In another study we looked at the role that exposure to television and electronic media has on the language development of children with hearing loss (Ambrose, VanDam, & Moeller, 2014). Here we showed that toddlers' exposure to more electronic media was negatively associated with receptive language abilities, but that increasing conversational turns might mitigate that disadvantage.

Finally, in another study we compared four groups: typically developing, autistic, language delayed, and hard-of-hearing children (VanDam et al., 2015). Here we analyzed 1,913 daylong recordings from 273 children in four groups of known diagnoses: typically developing, autism spectrum disorder, language delayed, and hard of hearing. Using automatic methods based on LENA recordings, we collected parameters from the acoustic signal known to be important for speech and language development and built statistical models to classify children into groups and examine their vocal development. We were able to show successful classification of children using automated vocal analysis. This result is a step toward using large-sample automated methods to distinguish among disorders and improve treatment to a differentiated population.

The three empirical research projects described above demonstrate a concrete approach to using objective, automated methods and very large sampling to gain insight into at-risk or clinical populations that are currently not well understood.

Quality of automatic procedures One of the issues raised above in the context of investigations of data using probabilistic machine-learning techniques is that these techniques are just that—probabilistic. The conditions under which our data were recorded pose some of the greatest problems for ASP/R, namely open-set vocabulary, continuous speech uttered by many speakers. The consequences of these factors are that the data are both complex and noisy. A natural question to ask, then, is *how accurately do the automatic methods perform*? Said more precisely, to what degree do the segment labels—*target-child*, *adult male*, and so on—conform to those of human judges? Previous empirical studies of children learning English, Spanish, and French have been conducted. In these, the agreement between human coders and LENA has varied from 82% to 64%, depending on the label (Canault, Le Normand, Foudil, Loundon, & Thai-Van, 2015; Soderstrom & Wittebolle, 2013; Weisleder & Fernald, 2013).

A more recent study looked at the fidelity of the automatic LENA methods using many excised tokens of specific labels and a cadre of judges (VanDam & Silbert, 2016). In this study, 26 daylong recordings were used, one from each of 26 families. The ages of the participating children averaged 29.1 months. After the daylong recordings had been collected using the LENA system, segments bearing specific talker labels (*male adult*, *female adult*, and *target-child*) were excised, labeled and cataloged, and saved to a file. The files were sampled evenly throughout the day from the original recordings, using a custom MATLAB script that computed the total number of utterances from one category label (e.g., *father*), *n*, and collected each *n*/30th segment with that label in order to achieve exactly 30 samples evenly spaced throughout the day. A sampling procedure emphasizing randomness, sparser sampling, or other techniques would be trivial to produce. These excised segments then served as the stimuli for 23 human judges who listened to a total of 53,820 stimulus presentations, 17,940 from each of the categories of interest. The judges performed a four-alternative forced choice task, choosing *mother*, *father*, *child*, or *other* for each audio presentation. In a tenfold cross-validation procedure, the error rate stood at .353, consistent with the performance of ASP/R in a large-vocabulary continuous-speech setting (VanDam & Silbert, 2016). Others have investigated the quality of the tools in typically developing (Xu, Richards, & Gilkerson, 2014) and special (Woynaroski et al., 2017) populations. This is an example of using the commercially available LENA system as a first step to data collection, with custom software being used to postprocess for features of interest.

Since this study was completed, researchers in ASP/R have been reporting spectacular decreases in error rate though the introduction of neural networks in their acoustic and language models, and some have looked specifically at the LENA system (Richards et al., 2017). Thus, Xiong and colleagues (2017) report an error rate of .062. There are two caveats. Xiong and colleagues at Microsoft tested on the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992), a corpus of telephone conversations, which are considerably more constrained than our own daylong recordings of in situ child speech. Furthermore, the error rate posted by the Microsoft team is the word error rate, whereas the study cited above reported labeling error. The two can be compared in order to get a rough measure, but, of course, they are not identical.

Database management and shared data The question of how to catalog and store thousands of hours of recorded speech in such a way that the speech can be made available to research groups other than our own, while protecting the privacy of the subjects, is a nontrivial problem. After a corpus reaches a certain size, it becomes unreasonable for investigators to become their own systems administrators and, in the case where software is to be shared, part-time librarians. This prompted the initiation of HomeBank (https://homebank.talkbank.org), a permanent, online database of daylong audio recordings made in naturalistic settings along with an associated GitHub code repository (https://github.com/homebankcode/) to analyze, process, and store the data. The HomeBank

database currently has eight corpora representing greater than 100 talkers and 1,000 recordings. HomeBank is administered by a professional technology team. The recordings consist of vetted and unvetted data. In the vetted portion, the recordings and associated metadata have been examined by trained listeners to ensure that no private or personally identifying material is present. The vetted database is unrestricted and available for public download and analysis. The unvetted database has not been examined by human listeners for content. It is password-protected and available only to registered HomeBank members who have agreed to confidentiality and who have passed through recognized ethics training on how to handle human data (VanDam et al., 2016).

Extensions, adaptations, and future applications

The characteristic feature of the work detailed here is the high quantity of behavioral data available to the researcher. We believe that this is a recent advance and fertile ground for research. That is to say, big data have been considered and collected for much longer than *behavioral* big data have been available. The advent of increases in computing, not least of which are storage and miniaturization, have afforded interested researchers the ability to collect behavioral data from biological organisms (i.e., humans) as never before. The work shown here is particular to research programs primarily interested in utilizing automatic methods to explore a wide variety of open questions about speech production, language development, and disorders with associated risk of communication dysfunction, but the general data acquisition and analytic approach can be and has been straightforwardly extended or adapted.

One area of current interest is the development of an alternative to the LENA system (Metze et al., 2013; Schuller et al., 2017). An alternative has several challenges and several advantages. Among the challenges are standardizing hardware, cost, coordination of effort, lack of validation, lack of continuity in the literature for a new, undescribed system, and the varying interests of research teams leading to different or conflicting paths of development. The advantages of an alternative may afford the research community with transparent processing (which is not a feature of the LENA system), flexibility in coding and applications, such as models that may directly apply to variable problems to address specific populations of interest, and interoperability with other hardware and software systems. The advantages will allow an alternative to be research-driven, with development undertaken to address problems of interest within the research community. We expect and hope that the development of this alternative will continue to receive generous support in proportion to the increasing interest in the international scientific community. Researchers around the world using daylong audio recording technology have made discoveries in science and engineering, anthropology, human health disorders, language development, and many disparate fields. With this broad scientific attention and demonstrated results, the development of an alternative to the LENA system is fast approaching viability. Finally, an alternative may accelerate the course of scientific discovery, expressly because of the extensibility of such a system. Compared with the present landscape, dominated by a single proprietary provider, a collaborative enterprise (as is currently being undertaken) will most likely be transparent, open-source, and accessible to a wider range of interested scientists, who can take advantage of this flexibility and extensibility in order to gain new insights and promote new discoveries. There is also potential for additional collaboration between the scientific community and industry as more participants enter the research space.

In the behavioral and health sciences domain, researchers are exploring a breathtaking array of parameters. Just to name a few from disparate fields, there is extant research on processing radiological images (Shin et al., 2015), exploring the genotypology of disease (Nalls et al., 2014), dermatological diagnoses (Esteva et al., 2017), the onset and time course of disease epidemics (Brownstein, Freifeld, & Madoff, 2009; Young, 2015), neuroimaging (Beaton, Dunlop, & Abdi, 2016; Biswal et al., 2010), understanding personality traits (Bleidorn, Hopwood, & Wright, 2017), economic behavior and trends (Einav & Levin, 2014), the nature of human long-term memory (Stanley & Byrne, 2016), theoretical developmental phonology (Bergmann, Tsuji, & Cristia, 2017), and many more.

The framework described above takes in large quantities of behavioral data and uses modern computer hardware and software to store, organize, and analyze the data. Instead of acoustic sensors such as the microphones described above, researchers are using piezoelectric sensors, gyroscopes, implanted electrodes, accelerometers, and pressure devices (among many others) to obtain data about human behavior in domains from the physiological to the cognitive (Gowers et al., 2015; Imani et al., 2016; Klucken et al., 2013; Soh, Vandenbosch, Mercuri, & Schreurs, 2015; Staudenmayer, He, Hickey, Sasaki, & Freedson, 2015). In principle, sensors that collect behavioral data in another domain—such as the electrochemical response to a stimulus or in a veterinary setting, for example—can be treated in largely the same way as the data we have collected from microphones. These data can be collected at a particular sampling frequency, stored in a database, processed (or mined) for features of interest, and (possibly) shared for future or collaborative analysis.

This modular approach also has certain drawbacks. First, the computational requirements of a third-party provider are likely to be fairly substantial. It is not clear who will fund offsite data storage, how it will be managed, who will make

management decisions, and how the data will be maintained in perpetuity. Second, access and throughput are continuing concerns. As data volume increases, the need for increased capacity also increases. This might restrict overall access to the data, which would reduce the very reason it is put in place. Third, access is a double-edged sword that is particularly risky for behavioral data, since it critically provides details about the humans it represents. Despite a robust appreciation for the responsible handling of human data in the academic research setting, it is unclear what potential harm to individuals, other ethical violations, or alternative uses might arise out of these data being accessible in ways they have not been previously (Zhou et al., 2012; Krutz &Vines, 2010).

Conclusions

This article describes the use of large-scale, daylong ecologically valid audio recordings to examine human behavioral variables. The daylong audio recordings are processed with ASP/R software, then postprocessed to focus on variables of interest. The resulting data are rich and complex and afford the opportunity to generalize observations beyond a descriptive case alone. This practical approach lends itself to a great variety of extensions, expansions, and cross-discipline collaborations. Here we have focused on our own research interest, acoustic phonetic analyses to gain insight into human speech and language development, but the extensibility of the methodological approach is ripe for a wide array of other scientific questions.

Author note This research was supported by the Washington Research Foundation; Grants NSF-SBE RIDIR 1539133, NIH/NIDCD R01DC009569, and DC009560-01S1; WSU Seed Grant 124172-001; and sabbatical support from the School of Engineering and Applied Science, Gonzaga University.

References

Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry*, *23*, 244–247. https://doi.org/10.1080/1047840X.2012.705133

Ambrose, S. E., VanDam, M., & Moeller, M. P. (2014). Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear and Hearing*, *35*, 139–147. https://doi.org/10.1097/AUD.0b013e3182a76768

Beaton, D., Dunlop, J., & Abdi, H. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods*, *21*, 621–651.

Bergmann, C., Tsuji, S., & Cristia, A. (2017). Top-down versus bottom-up theories of phonological acquisition: A big data approach. In *Proceedings of Interspeech 2017* (pp. 2103–2107). Baixas, France: International Speech Communication Association. https://doi.org/10.21437/Interspeech.2017-1443

Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., ... Dogonowski, A. M. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, *107,* 4734–4739.

Bleidorn, W., Hopwood, C. J., & Wright, A. G. (2017). Using big data to advance personality theory. *Current Opinion in Behavioral Sciences*, *18*, 79–82.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, *40*, 318–335.

Brewer, M. B., & Crano, W. D. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 3–16). London, UK: Cambridge University Press.

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection—harnessing the Web for public health surveillance. *New England Journal of Medicine*, *360*, 2153–2157.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*, 183–186. https://doi.org/10.1126/science.aal4230

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2015). Reliability of the Language Environment Analysis system (LENA TM) in European French. *Behavior Research Methods*, *48*, 1109–1124. https://doi.org/10.3758/s13428-015-0634-8

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2011). Importance of parent talk on the development of preterm infant vocalizations. *Pediatrics*, *128*, 910–916.

Caskey, M., & Vohr, B. (2013). Assessing language and language environment of high-risk infants and children: A new approach. *Acta Paediatrica*, *102*, 451–461.

Cearley, D. W., Burke, B., Searle, S., & Walker, M. J. (2017, October 3). Top 10 strategic technology trends for 2018 (). Retrieved from https://www.gartner.com/doc/3811368?srcId=1-6595640781

Chung, H., Iorga, M., Voas, J., & Lee, S. (2017). Alexa, can I trust you? *Computer*, *50*, 100–104.

Codd, E.F. (1970). A relational model of data for large shared databanks. *Communications of the ACM*, *13*, 377–387.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand-McNally.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, *65*, 122–135. https://doi.org/10.1108/LR-06-2015-0061

De Palma, P., & VanDam, M. (2017). Using automatic speech processing to analyze fundamental frequency of child-directed speech stored in a very large audio corpus. In *Proceedings of the Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)* (pp. 1–6). Piscataway, NJ: IEEE Press. https://doi.org/10.1109/IFSA-SCIS.2017.8023224

Dixon, J. (2010). Pentaho, hadoop, and data lakes. Available at http://jamesdixon.wordpress.com/2010/10/14/pentahohadoop-and-data-lakes

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*, 1243089.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A*, *180*, 967–1033.

Gilkerson J., & Richards, J. A. (2009). *The power of talk: Impact of adult talk, conversational turns, and TV during the critical 0–4 years of child development* (Technical Report LTR-01-2, 2nd ed.). Boulder, CO: LENA Foundation. Retrieved from www.lenafoundation.org/wp-content/uploads/2014/10/LTR-01-2_PowerOfTalk.pdf.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92* (Vol. 1, pp. 517–520). Piscataway, NJ: IEEE Press.

Gowers, S. A., Curto, V. F., Seneci, C. A., Wang, C., Anastasova, S., Vadgama, P., ... Boutelle, M. G. (2015). 3D printed microfluidic device with integrated biosensors for online analysis of subcutaneous human microdialysate. *Analytical Chemistry, 87,* 7763–7770.

Gray, J., Liu, D., Nieto-Santisteban, M., Szalay, A., Dewitt, D., & Heber, G. (2005). Scientific Data Management in the Coming Decade. *SIGMOD Record*, *34*, 34–41.

Hai, R., Geisler, S., & Quix, C. (2016). Constance: An intelligent data lake system. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2097–2100). New York, NY: ACM Press.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.

Hess, W. (1983). *Pitch determination of speech signals*. Berlin, Germany: Springer.

Hinrichs, E., & Krauwer, S. (2014). The CLARIN research infrastructure: Resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp. 1525–1531.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*, 82–97.

Huber, P. J. (1981). *Robust statistics*. New York, NY: Wiley.

IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0 [Software]. Armonk, NY. Available from https://www.ibm.com/products/spss-statistics

Imani, S., Bandodkar, A. J., Mohan, A. V., Kumar, R., Yu, S., Wang, J., & Mercier, P. P. (2016). A wearable chemical–electrophysiological hybrid biosensing system for real-time health and fitness monitoring. *Nature Communications*, *7*, 11650.

Jarke, M., Jeusfeld, M. A., Quix, C., & Vassiliadis, P. (1999). Architecture and quality in data warehouses: an extended repository approach. *Information Systems*, *24*, 229–253.

Jones, E., Oliphant, T., & Peterson, P. (2014). {SciPy}: Open source scientific tools for {Python}. Available from https://www.scipy.org/

Kepuska, V., & Bouthota, G. (2017). Comparing speech recognition systems. *Journal of Engineering Research and Application*, *7*, 20–24.

Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, 53, 3735–3745.

Klucken, J., Barth, J., Kugler, P., Schlachetzki, J., Henze, T., Marxreiter, F., ... Winkler, J. (2013). Unbiased and mobile gait analysis detects motor impairment in Parkinson's disease. *PLoS ONE, 8,* e56956. https://doi.org/10.1371/journal.pone.0056956

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*, 1137–1145.

Krutz, R., & Vines, R. (2010). *Cloud security: A comprehensive guide to secure cloud computing*. Hoboken, NJ: Wiley.

LENA. (2018). LENA Research Foundation [Software]. Boulder, CO. Available from www.lenafoundation.org

MacWhinney, B. (2007) The TalkBank Project. In K. P. Corrigan, J. C. Beal, & H. L. Hoisl (Eds.), *Creating and digitizing language corpora: Vol. 1* (pp. 163–180). Houndmills, UK: Palgrave Macmillan. https://doi.org/10.1057/9780230223936

Masataka, N. (1999). The role of modality and input in the earliest stage of language acquisition: Studies in Japanese sign language. In J. Morford & R. Mayberry (Eds.), *Language acquisition by eye* (pp. 3–24). New York, NY: Psychology Press.

MathWorks. (2018). Matlab, version R2018a [Software]. Natick, MA. Available from www.mathworks.com

McCambridge, J., Witton, J., & Elbourne D. R. (2014). Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*, 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

McGrayne, S. (2012). *The theory that would not die*. New Haven, CT: Yale University Press.

Metze, F., Fosler-Lussier, E., & Bates, R. (2013). The speech recognition virtual kitchen. In *Proceedings of INTERSPEECH 2013* (pp. 1858–1860). Baixas, France: International Speech Communication Association.

Milliken, G. A., & Johnson, D. E. (2001). *Analysis of messy data: Vol. III. Analysis of covariance*. Boca Raton: Chapman & Hall/CRC.

Milliken, G. A., & Johnson, D. E. (2009). *Analysis of messy data: Vol. I. Designed experiments* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Mohanty, H., Bhuhan, P., & Chenthati, D. (eds.). (2015). *Big data: A primer* (Studies in Big Data 11). New Delhi, India: Springer India. https://doi.org/10.1007/978-81-322-2494-5

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., ... Schulte, C. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics, 46,* 989.

Narayanan, S., & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, *101*, 1203–1233.

Neisser, U. (1967). *Cognitive psychology.* New York, NY: Appleton-Century.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Contestabile, M. (2015). Promoting an open research culture. *Science, 348,* 1422–1425.

O'Leary, D. E. (2014). Embedding AI and crowdsourcing in the Big Data Lake. *IEEE Intelligent Systems*, *29*, 70–73.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences, 107,* 13354–13359.

Orne, M. T. (2009). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow, *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's classic books* (pp. 110–137). Oxford, UK: Oxford University Press.

Pasupuleti, P., & Purra, B. S. (2015). *Data lake development with big data*. Birmingham, UK: Packt Publishing Ltd.

R Core Team. (2013). R: A language and environment for statistical computing [Software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*, 3.

Rankine, J., Li, E., Lurie, S., Rieger, H., Fourie, E., Siper, P. M., ... Kolevzon, A. (2017). Language ENvironment Analysis (LENA) in Phelan-McDermid syndrome: Validity and suggestions for use in minimally verbal children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders, 47,* 1605–1617.

Richards, J. A., Xu, D., Gilkerson, J., Yapanel, U., Gray, S., & Paul, T. (2017). Automated assessment of child vocalization development using LENA. *Journal of Speech, Language, and Hearing Research*, *60*, 2047–2063.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York, NY: McGraw-Hill.

Rothenberg, J. (1999). *Avoiding technological quicksand: Finding a viable technical foundation for digital preservation: A report to the Council on Library and Information Resources*. Washington, DC: Council on Library and Information Resources.

RStudio Team. (2015). RStudio: integrated development for R [Software]. Boston, MA: RStudio. Retrieved from http://www.rstudio.com

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., ... Warlaumont, A. S. (2017). The Interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proceedings of the Computational Paralinguistics Challenge (ComParE) Conference, Interspeech 2017* (pp. 3442–3446). Baixas, France: International Speech Communication Association.

Shin, H. C., Lu, L., Kim, L., Seff, A., Yao, J., & Summers, R. M. (2015). Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1090–1099). Piscataway, NJ: IEEE Press.

Silbert, N. H., Linck, J. A., & VanDam, M. (2013). Multilevel models, covariates, and controlled factors in experimental speech research: Unified analyses of highly structured data. *Proceedings of Meetings on Acoustics*, *19*, 0600029. https://doi.org/10.1121/1.4790331

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE*, *8*, e80646. https://doi.org/10.1371/journal.pone.0080646

Soh, P. J., Vandenbosch, G. A., Mercuri, M., & Schreurs, D. M. P. (2015). Wearable wireless health monitoring: Current developments, challenges, and future trends. *IEEE Microwave Magazine*, *16*, 55–70.

Stanley, C., & Byrne, M. D. (2016). Comparing vector-based and Bayesian memory models using large-scale datasets: User-generated hashtag and tag prediction on Twitter and stack overflow. *Psychological Methods*, *21*, 542–565.

Staudenmayer, J., He, S., Hickey, A., Sasaki, J., & Freedson, P. (2015). Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *Journal of Applied Physiology*, *119*, 396–403.

Tenner, E. (1996). *Why things bite back: Technology and the revenge of unintended consequences.* New York, NY: Knopf.

Thiemann-Bourque, K. S., Warren, S. F., Brady, N., Gilkerson, J., & Richards, J. A. (2014). Vocal interaction between children with Down syndrome and their parents. *American Journal of Speech-Language Pathology*, *23*, 474–485.

van der Meulen, R. (2008, April 16). Gartner says effective management can cut total cost of ownership for desktop PCs by 42 per cent (Press release). Retrieved from https://www.gartner.com/newsroom/id/636308

VanDam, M., Ambrose, S. E., & Moeller, M. P. (2012). Quantity of parental language in the home environments of hard-of-hearing 2-year-olds. *Journal of Deaf Studies and Deaf Education*, *17*, 402–420. https://doi.org/10.1093/deafed/ens025

VanDam, M., & De Palma, P. (2014). Fundamental frequency of child-directed speech using automatic speech recognition. In *IEEE Proceedings of the Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)* (pp. 1349–1353). Piscataway, NJ: IEEE Press. 10.1109/SCIS-ISIS.2014.7044876

VanDam, M., Oller, D. K., Ambrose, S. E., Gray, S., Richards, J. A., Xu, D., Gilkerson, J., Silbert, N. H., & Moeller, M. P. (2015). Automated vocal analysis of children with hearing loss and their typical and atypical peers. Ear and Hearing, 36(4): e146-e152. https://doi.org/10.1097/AUD.0000000000000138

VanDam, M., & Silbert, N.H. 2016. Fidelity of automatic speech processing for adult and child talker classifications. *PLoS ONE*, *11*, e0160588. https://doi.org/10.1371/journal.pone.0160588

VanDam, M., Warlaumont, A., Bergelson, E., Cristia, A., Sodestrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: A online repository of daylong child-centered audio recordings. *Seminars in Speech and Language*, *37*, 128–140.

Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2010). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*, *40*, 555–569.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152.

Wilcox, R. (2016). *Introduction to robust estimation and hypothesis testing*. Amsterdam, The Netherlands: Elsevier.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254–274. https://doi.org/10.1037/1082-989X.8.3.254

Woods, D., Dekker, S., Cook, R., Johannesen, L., & Sarter, N. (2017). *Behind human error*. London, UK: CRC Press.

Woods, D. D., & Roth, E. M. (1988). Cognitive systems engineering. In M. Helander (Ed.), *Handbook of human–computer interaction* (pp. 3–43). Amsterdam, The Netherlands: Elsevier Science.

Woynaroski, T., Oller, D. K., Keceli-Kaysili, B., Xu, D., Richards, J. A., Gilkerson, J., ... Yoder, P. (2017). The stability and validity of automated vocal analysis in preverbal preschoolers with autism spectrum disorder. *Autism Research, 10,* 508–519.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2017). The Microsoft 2016 conversational speech recognition system. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5255–5259). Piscataway, NJ: IEEE Press. https://doi.org/10.1109/ICASSP.2017.7953159

Xu, D., Richards, J. A., & Gilkerson, J. (2014). Automated analysis of child phonetic production using naturalistic recordings. *Journal of Speech Language Hearing Research*, *57*, 1638–1650. https://doi.org/10.1044/2014_JSLHR-S-13-0037

Young, S. D. (2015). A “big data” approach to HIV epidemiology and prevention. *Preventive Medicine*, *70*, 17–18. https://doi.org/10.1016/j.ypmed.2014.11.002

Zhou, M., Mu, Y., Susilo, W., Yan, J., & Dong, L. (2012). Privacy enhanced data outsourcing in the cloud. *Journal of Network and Computer Applications*, *35*, 1367–1373.