Chapter 1

Developing Cognitive Theory by Mining Large-Scale Naturalistic Data

Michael N. Jones, *Indiana University*

Abstract

Cognitive research is increasingly coming out of the laboratory. It is becoming much more common to see research that repurposes large-scale and naturalistic data sources to develop and evaluate cognitive theories at a scale not previously possible. We now have unprecedented availability of massive digital data sources that are the product of human behavior and offer clues to understand basic principles of cognition. A key challenge for the field is to properly interrogate these data in a theory-driven way to reverse engineer the cognitive forces that generated them; this necessitates advances in both our theoretical models and our methodological techniques. The arrival of big data has been met with healthy skepticism by the field, but has also been seen as a genuine opportunity to advance our understanding of cognition. In addition, theoretical advancements from big data are heavily intertwined with new methodological developments—new techniques to answer questions from big data also give us new questions that could not previously have been asked. The goal of this volume is to present emerging examples from across the field that use large and naturalistic data to advance theories of cognition that would not be possible in the traditional laboratory setting.

1.1. Introduction

While laboratory research is still the backbone of tracking causation among behavioral variables, more and more cognitive research is now letting experimental control go in favor of mining large-scale and real-world datasets. We are seeing an exponential[1] expansion of data available to us that is the product of human behavior: Social media, mobile device sensors, images, RFID tags, linguistic corpora, web search logs, and consumer product reviews, just to name a few streams. Since 2012, about 2.5 exabytes of digital data are created every day (McAfee et al., 2012). Each little piece of data is a trace of human behavior and offers us a potential clue to understand basic cognitive principles; but we have to be able to put all those pieces together in a reasonable way. This

approach necessitates both advances in our theoretical models and development of new methodological techniques adapted from the information sciences.

Big data sources are now allowing cognitive scientists to evaluate theoretical models and make new discoveries at a resolution not previously possible. For example, we can now use online services like Netflix, Amazon, and Yelp to evaluate theories of decision-making in the real world and at an unprecedented scale. Wikipedia edit histories can be analyzed to explore information transmission and problem solving across groups. Linguistic corpora allow us to quantitatively evaluate theories of language adaptation over time and generations (Lupyan and Dale, 2010) and models of linguistic entrainment (Fusaroli et al., 2015). Massive image repositories are being used to advance models of vision and perception based on natural scene statistics (Griffiths, Abbott, and Hsu, 2016; Khosla, Raju, Torralba, and Oliva, 2015). Twitter and Google search trends can be used to track the outbreak and spread of "infectious" ideas, memory contagion, and information transmission (Chen and Sakamoto, 2013; Masicampo and Ambady, 2014; Wu, Hofman, Mason, and Watts, 2011). Facebook feeds can be manipulated[2] to explore information diffusion in social networks (Bakshy, Rosenn, Marlow, and Adamic, 2012; Kramer, Guillory, and Hancock, 2014). Theories of learning can be tested at large scales and in real classroom settings (Carvalho et al., 2016; Fox, Hearst, and Chi, 2014). Speech logs afford both theoretical advancements in auditory speech processing, and practical advancements in automatic speech comprehension systems.

The primary goal of this volume is to present cutting-edge examples that use large and naturalistic data to uncover fundamental principles of cognition and evaluate theories that would not be possible without such scale. A more general aim of the volume is to take a very careful and critical look at the role of big data in our field. Hence contributions to this volume were handpicked to be examples of advancing theory development with large and naturalistic data.

1.2. What is Big Data?

Before trying to evaluate whether big data could be used to benefit cognitive science, a very fair question is simply what is big data? Big data is a very popular buzzword in the contemporary media, producing much hype and many misconceptions. Whatever big data is, it is having a revolutionary impact on a wide range of sciences, is a "game-changer," transforming the way we ask and answer questions, and is a must-have for any new age scientist's toolbox. But when pressed for a definition, there seems to be no solid consensus, particularly among cognitive scientists. We know it probably doesn't fit in a spreadsheet, but opinions diverge beyond that. The issue is now almost humorous, with Dan Ariely's popular quip comparing big data to teenage sex, in that "everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."

As scientists, we are quite fond of careful operational definitions. However, big data and data science are still-evolving concepts, and are moving targets for formal definition. Definitions tend to vary depending on the field of study. A strict interpretation of big data from the computational sciences typically refers to datasets that are so massive and rapidly changing that our current data processing methods are inadequate. Hence, it

is a drive for the development of distributed storage platforms and algorithms to analyze data sets that are currently out of reach. The term extends to challenges inherent in data capture, storage, transfer and predictive analytics. As a loose quantification, data under this interpretation currently becomes "big" at scales north of the exabyte.

Under this strict interpretation, work with true big data is by definition quite rare in the sciences; it is more development of architectures and algorithms to manage these rapidly approaching scale challenges that are still for the most part on the horizon (NIST Big Data Working Group, 2014). At this scale, it isn't clear that there are any problems in cognitive science that are true big data problems yet. Perhaps the largest data project in the cognitive and neural sciences is the Human Connectome Project (Van Essen, et al. 2012), an ambitious project aiming to construct a network map of anatomical and functional connectivity in the human brain, linked with batteries of behavioral task performance. Currently, the project is approaching a petabyte of data. By comparison, the Large Hadron Collider project at CERN records and stores over 30 petabytes of data from experiments each year.[3]

More commonly, the *Gartner 3 Vs* definition of big data is used across multiple fields: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Laney, 2012). *Volume* is often indicative of the fact that big data records and observes everything within a recording register, in contrast to our commonly used methods of sampling in the behavioral sciences. *Velocity* refers to the characteristic that big data is often a real-time stream of rapidly captured data. The final characteristic, *variety*, denotes that big data draws from multiple qualitatively different information sources (text, audio, images, GPS, etc.), and uses joint inference or fusion to answer questions that are not possible by any source alone. But far from being expensive to collect, big data is usually a natural byproduct of digital interaction.

So while a strict interpretation of big data puts it currently out of reach, it is simultaneously everywhere by more liberal interpretations. Predictive analytics based on machine learning has been hugely successful in many applied settings (see Hu, Wen, and Chua, 2014, for a review). Newer definitions of big data summarize it as more focused on repurposing naturalistic digital footprints; the size of "big" is relative across different fields (NIST Big Data Working Group, 2014). The NIH BD2K (Big Data to Knowledge) program is explicit that a big data approach is best defined by what is large and naturalistic to specific subfields, not an absolute value in bytes. In addition, BD2K notes that a core big data problem involves joint inference across multiple databases. Such combinatorial problems are clearly big data, and are perfectly suited for theoretically driven cognitive models—many answers to current theoretical and practical questions may be hidden in the complimentary relationship between data sources.

1.3. What is Big Data to Cognitive Science?

Much of the publicity surrounding big data has focused on its insight power for business analytics. Within the Cognitive Sciences, we have been considerably more skeptical of big data's promise, largely because we place such a high value on explanation over prediction. A core goal of any cognitive scientist is to fully understand the system under investigation,

rather than being satisfied with a simple descriptive or predictive theory.

Understanding the mind is what makes an explanatory cognitive model distinct from a statistical predictive model—our parameters often reflect hypothesized cognitive processes or representations (e.g., attention, memory capacity, decision thresholds, etc.) as opposed to the abstract predictive parameters of, say, weights in a regression model. Predictive models are able to make predictions of new data provided they are of the same sort as the data on which the model was trained (e.g., predicting a new point on a forgetting curve). Cognitive models go a step further: An explanatory model should be able to make predictions of how the human will behave in situations and paradigms that are novel and different from the situations on which the model was built but that recruit the same putative mechanism(s) (e.g., explaining the process of forgetting).

Marcus and Davis (2014) have argued rather convincingly that big data is a scientific idea that should be retired. While it is clear that large datasets are useful in discovering correlations and predicting common patterns, more data does not on its own yield explanatory causal relationships. Big data and machine learning techniques are excellent bedfellows to make predictions with greater fidelity and accuracy. But the match between big data and cognitive models is less clear; because most cognitive models strive to explain causal relationships, they may be much better paired with experimental data, which shares the same goal. Marcus and Davis note several ways in which paying attention to big data may actually lead the scientist astray, compared to a much smaller amount of data from a well-controlled laboratory scenario.

In addition, popular media headlines are chock-full of statements about how theory is obsolete now that big data has arrived. But theory is a simplified model of empirical phenomena—theory explains data. If anything, cognitive theory is more necessary to help us understand big data in principled way given that much of the data were generated by the cognitive systems that we have carefully studied in the laboratory, and cognitive models help us to know where to search and what to search for as the data magnitude grows.

Despite initial skepticism, big data has also been embraced by Cognitive Science as a genuine opportunity to develop and refine cognitive theory (Griffiths, 2015). Criticism of research using big data in an atheoretic way is a fair critique of the way some scientists (and many outside academia) are currently using big data. However, there are also scientists making use of large datasets to test theory-driven questions—questions that would be unanswerable without access to large naturalistic datasets and new machine learning approaches. Cognitive scientists are, by training, [experimental] control freaks. But the methods used by the field to achieve laboratory control also serve to distract it from exploring cognitive mechanisms through data mining methods applied big data.

Certainly, big data is considerably more information than we typically collect in a laboratory experiment. But it is also naturalistic, and a footprint of cognitive mechanisms operating in the wild (see Goldstone and Lupyan, 2016, for a recent survey). There is a genuine concern in the Cognitive Sciences that many models we are developing may be overfit to specific laboratory phenomena that neither exist nor generalize beyond the walls of the lab. The standard cognitive experiment takes place in one hour in a well-controlled setting with variables that normally covary in the real world held constant. This allows us to determine conclusively that the flow of causation is from our manipulated variable(s)

to the dependent variable, and often by testing discrete settings ("factorology;" Balota, Yap, Hutchison, and Cortese, 2012).

It is essential to remember that the cognitive mechanisms we study in the laboratory evolved to handle real information-processing problems in the real world. By "capturing" and studying a mechanism in a controlled environment, we risk discovering experiment- or paradigm-specific strategies that are a response to the experimental factors that the mechanism did not evolve to handle, and in a situation that does not exist the real world. While deconfounding factors is an essential part of an experiment, the mechanism may well have evolved to thrive in a rich statistically redundant environment. In this sense, cognitive experiments in the lab may be somewhat analogous to studying captive animals in the zoo and then extrapolating to behavior in the wild.

The field has been warned about over-reliance on experiments several times in the past. Even four decades ago Estes (1975) raised concern in mathematical psychology that we may be accidentally positing mechanisms that apply only to artificial situations, and that our experiments may unknowingly hold constant factors that may covary to produce very different behavior in the real world. More recently, Miller (1990) reminded cognitive scientists of Estes' reductionism caution:

"I have observed over the years that there is a tendency for even the best cognitive scientists to lose sight of large issues in their devotion to particular methodologies, their pursuit of the null hypothesis, and their rigorous efforts to reduce anything that seems interesting to something else that is not. An occasional reminder of why we flash those stimuli and measure those reaction times is sometimes useful." (Miller, 1990; p. 7)

Furthermore, we are now discovering that much of the behavior we want to use to make inferences about cognitive mechanisms is heavy-tail distributed (exponential and power-law distributions are very common in cognitive research). Sampling behavior in a one-hour lab setting is simply insufficient to ever observe the rare events that allow us to discriminate among competing theoretical accounts. And building a model from the center of a behavioral distribution may fail horribly to generalize if the tail of the distribution is the important characteristic that the cognitive mechanism evolved to deal with.

So while skepticism about big data in Cognitive Science is both welcome and warranted, the above points are just a few reasons that big data could be a genuine opportunity to advance our understanding of human cognition. If dealt with in a careful and theoretically-driven way, big data offers us a completely new set of eyes to understand cognitive phenomena, to constrain among theories that are currently deadlocked with laboratory data, to evaluate generalizability of our models, and to have a impact on the real world situations that our models are meant to explain (e.g., by optimizing medical and consumer decisions, information discovery, education, etc.). And embracing big data brings with it development of new analytic tools that also allow us to ask new theoretical questions that we had not even considered previously.

1.4. How is Cognitive Research Changing with Big Data?

Cognitive scientists have readily integrated new technologies for naturalistic data capture into their research. The classic cognitive experiment typically involved a single subject in a testing booth making two-alternative forced choice responses to stimuli presented on a monitor. To be clear, we have learned a great deal about fundamental principles of human cognition with this basic laboratory approach. But the modern cognitive experiment may involve mobile phone games with multiple individuals competing in resource sharing simultaneously from all over the world (Dufau, et al., 2011; Miller, 2012), or dyads engaged in real-time debate while their attention and gestures are captured with Google Glass (Paxton, Rodriguez, and Dale, 2015).

In addition, modern cognitive research is much more open to mining datasets that were created for a different purpose to evaluate the models we have developed from the laboratory experiments. Although the causal links among variables are murkier, they are still possible with new statistical techniques borrowed from informatics, and the scale of data allows the theorist to paint a more complete and realistic picture of cognitive mechanisms. Furthermore, online labor markets such as Amazon's Mechanical Turk have accelerated the pace of experiments by allowing us to conduct studies that might take years in the laboratory in a single day online (Crump, McDonnell, and Gureckis, 2013; Gureckis et al., 2015).

Examples of new data capture technologies advancing our theoretical innovations are emerging all over the cognitive sciences. Cognitive development is a prime example. While development unfolds over time, the field has traditionally been reliant on evaluating infants and toddlers in the laboratory for short studies at regular intervals across development. Careful experimental and stimulus control is essential, and young children can only provide us with a rather limited range of response variables (e.g., preferential looking and habituation paradigms are very common with infants).

While this approach has yielded very useful information about basic cognitive processes and how they change, we get only a small snapshot of development. In addition, the small scale is potentially problematic because many theoretical models behave in a qualitatively different way depending on the amount and complexity of data (Frank, Tenenbaum, and Gibson, 2013; McClelland, 2009; Qian and Aslin, 2014; Shiffrin, 2010). Aslin (2014) has also noted that stimulus control in developmental studies may actually be problematic. We may be underestimating what children can learn by using oversimplified experimental stimuli: These controlled stimuli deconfound potential sources of statistical information in learning, allowing causal conclusions to be drawn, but this may make the task much more difficult than it is in the real world where multiple correlated factors offer complimentary cues for children to learn the structure of the world (see Shukla, White, and Aslin, 2011). The result is that we may well endorse the wrong learning model because it explains the laboratory data well, but is more complex than is needed to explain learning in the statistically rich real world.

A considerable amount of developmental research has now come out of the laboratory. Infants are now wired with cameras to take regular snapshots of the visual information available to them across development in their real world experiences (Aslin, 2009; Fausey, Jayaraman, and Smith, 2016; Pereira, Smith, and Yu, 2014). LENA^{TM} recording

devices are attached to children to record the richness of their linguistic environments and to evaluate the effect of linguistic environment on vocabulary growth (Bergelson, 2016; Weisleder and Fernald, 2013). In one prominent early example, the SpeechHome project, an entire house was wired to record 200,000+ hours of audio and video from one child's first three years of life (Roy, Frank, DeCamp, Miller, and Roy, 2015). Tablet-based learning games are now being designed to collect theoretically constraining data as children are playing them all over the world (e.g., Frank, Sugarman, Horowitz, Lewis, Yurovsky, 2016; Pelz, Yung, and Kidd, 2015).

A second prime example of both new data capture methods and data scale advancing theory is in visual attention. A core theoretical issue surrounds identification performance as a function of target rarity in visual search, but the number of trials required to get stable estimates in the laboratory is unrealistic. Mitroff et al. (2015) opted to instead take a big data approach to the problem by turning visual search into a mobile phone game called "Airport Scanner." In the game, participants act the part of a TSA baggage screener searching for prohibited items as simulated luggage passes through an x-ray scanner. Participants respond on the touchscreen, and the list of allowed and prohibited items grows as they continue to play.

Mitroff et al. (2015) analyzed data from the first billion trials of visual search from the game, making new discoveries about how rare targets are processed when they are presented with common foils, something that would never have been possible in the laboratory. Wolfe (1998) had previously analyzed 1 million visual search trials from across 2,500 experimental sessions which took over 10 years to collect. In contrast, Airport Scanner collects over 1 million trials each day, and the rate is increasing as the game gains popularity. In addition to answering theoretically important questions in visual attention and memory, Mitroff et al.'s example has practical implications for visual detection of rare targets in applied settings, such as radiologists searching for malignant tumors on mammograms. Furthermore, data from the game have the potential to give very detailed information about how people become expert in detection tasks

1.5. Intertwined Theory and Methods

Our theoretical advancements from big data and new methodological developments are heavily interdependent. New methodologies to answer questions from big data are giving us new hypotheses to test. But simultaneously, our new theoretical models are helping to focus the new big data methodologies. Big data often flows in as an unstructured stream of information, and our theoretical models are needed to help tease apart the causal influence of factors, often when the data are constantly changing. Big data analyses are not going to replace traditional laboratory experiments. It is more likely that the two will be complimentary, with the field settling on a process of recurring iteration between traditional experiments and data mining methods to progressively zero in on mechanistic accounts of cognition that explain both levels.

In contrast to our records from behavioral experiments, big data is usually unstructured, and requires sophisticated analytical methods to piece together causal effects. Digital behavior is often several steps from the cognitive mechanisms we wish to explore, and these data often confound factors that are carefully teased apart in the laboratory

with experimental control (e.g., the effects of decision, response, and feedback). To infer causal flow in big data, cognitive science has been adopting more techniques from machine learning and network sciences.[4] One concern that accompanies this adoption is that the bulk of current machine learning approaches to big data are primarily concerned with detecting and predicting patterns, but they tend not to explain why patterns exist. Our ultimate goal in cognitive science is to produce explanatory models. Predictive models certainly benefit from more data, but it is questionable whether more data helps to achieve explanatory understanding of a phenomenon more than a well-controlled laboratory experiment.

Hence, development of new methods of inquiry from big data based on cognitive theory is a priority area of research, and has already seen considerable progress leading to new tools. Liberman (2010) has compared the advent of such tools in this century to the inventions of the telescope and microscope in the 17th. But big data and data mining tools on their own are of limited use for establishing explanatory theories; Picasso had famously noted the same issue about computers: "But they are useless. They can only give answers." Big data in no way obviates the need for foundational theories based on careful laboratory experimentation. Data mining and experimentation in cognitive science will continue to be iteratively reinforcing one another, allowing us to generate and answer hypotheses at a greater resolution, and to draw conclusions at a greater scale.

1.6. Acknowledgments

This work was supported by NSF BCS-1056744 and IES R305A150546

1.7. Notes

1. And I don't use the term "exponential" here simply for emphasis—the amount of digital information available currently doubles every two years, following Moore's Law (Gantz and Reinsel, 2012).

2. However, both the Facebook and OKCupid massive experiments resulted in significant backlash and ethical complaints.

3. LHC generates roughly two petabytes of data per second, but only a small amount is captured and stored.

4. "Drawing Causal Inference from Big Data" was the 2015 Sackler Symposium organized by the National Academy of Sciences.

References

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and vision science: official publication of the American Academy of Optometry, 86(6)*, 561.

Aslin, R. N. (2014). Infant learning: historical, conceptual, and methodological challenges. *Infancy, 19*(1), 2-27.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on*

World Wide Web (pp. 519-528). ACM.

Balota, D. A., Yap, M. J., Hutchison, K. A., and Cortese, M. J. (2012). Megastudies. *Visual word recognition volume 1: Models and methods, orthography and phonology,* 90.

Carvalho, P.F., Braithwaite, D.W., de Leeuw, J.R., Motz, B.A., and Goldstone, R.L. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. *PLoS ONE 11*(3): e0152115.

Chen, R., and Sakamoto, Y. (2013). Perspective matters: Sharing of crisis information in social media. In *System Sciences (HICSS), 2013 46th Hawaii International Conference* (pp. 2033-2041). IEEE.

Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research.*PloS one, 8*(3), e57410.

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., ... and Ktori, M. (2011). Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. *PloS one,6(*9), e24974.

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology, 12*(3), 263-282.

Fausey, C. M., Jayaraman, S., and Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition, 152*, 101-107.

Fox, A., Hearst, M.A., and Chi, M.T.H. (Eds.), *Proceedings of the First ACM Conference on Learning At Scale*, L@S 2014, March 2014.

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development 17*(1):1–17.

Frank, M. C., Tenenbaum, J. B., and Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. *PloS one, 8*(1), e52500.

Fusaroli, R., Perlman, M., Mislove, A., Paxton, A., Matlock, T., and Dale, R. (2015). Timescales of massive human entrainment. *PloS one, 10*(4), e0122742.

Gantz, J., and Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007, 1-16.

Goldstone, R. L., and Lupyan, G. (2016). Harvesting naturally occurring data to reveal principles of cognition. *Topics in Cognitive Science, 8(*3), 548-588.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition, 135*, 21-23.

Griffiths, T. L., Abbott, J. T., and Hsu, A. S. (2016). Exploring human cognition using large image databases. *Topics in Cognitive Science, 8*(3), 569-588.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... and Chan, P. (2015). psiTurk: An open-source framework for conducting replicable

behavioral experiments online. *Behavior research methods*, 1-14.

Hu, H., Wen, Y., Chua, T. S., and Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *Access, IEEE*, 2, 652-687.

Khosla, A., Raju, A. S., Torralba, A., and Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2390-2398).

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788-8790.

Laney, D. (2012). The importance of 'Big Data': A definition. *Gartner*. Retrieved, 21.

Lupyan, G., and Dale, R. (2010). Language structure is partly determined by social structure. *PloS one, 5*(1), e8559.

Marcus, G., and Davis, E. (2014). Eight (no, nine!) problems with big data. *The New York Times, 6*(04), 2014.

Masicampo, E. J., and Ambady, N. (2014). Predicting fluctuations in widespread interest: Memory decay and goal-related memory accessibility in Internet search trends. J*ournal of Experimental Psychology: General, 143*(1), 205.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., and Barton, D. (2012). Big data. The management revolution. *Harvard Bus Rev, 90*(10), 61-67.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1*(1), 11-38.

Miller, G. A. (1990). The place of language in a scientific psychology.*Psychological Science, 1*(1), 7-14.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*(3), 221-237.

Mitroff, S. R., Biggs, A. T., Adamo, S. H., Dowd, E. W., Winkle, J., and Clark, K. (2015). What can 1 billion trials tell us about visual search?. *Journal of experimental psychology: human perception and performance, 41*(1), 1.

NIST Big Data Working Group (2014). http://bigdatawg.nist.gov/home.php.

Paxton, A., Rodriguez, K. and Dale, R. (2015). PsyGlass: capitalizing on Google Glass for naturalistic data collection. Behavior *Research Methods, 47*, 608-619

Pelz, M., Yung, A., and Kidd, C. (2015). Quantifying Curiosity and Exploratory Play on Touchscreen Tablets. *Proceedings of the IDC 2015 Workshop on Digital Assessment and Promotion of Children's Curiosity.*

Pereira, A. F., Smith, L. B., and Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic bulletin and review, 21*(1), 178-185.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., and Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences, 112*(41),

12663-12668.

Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in cognitive science, 2*(4), 736-750.

Shukla, M., White, K. S., and Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences, 108*(15), 6038-6043.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., ... and Della Penna, S. (2012). The Human Connectome Project: a data acquisition perspective. *Neuroimage, 62*(4), 2222-2231.

Weisleder, A., and Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological science, 24*(11), 2143-2152.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search?. *Psychological Science, 9*(1), 33-39.

Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 705-714). ACM.