The Role of Statistics in Contemporary Brain Science

Oliver Y. Chén

Department of Biostatistics
Johns Hopkins University

Understanding how the brain works is arguably one of the greatest scientific challenges of our time (Alivisatos et al., 2012). On one hand, understanding how the brain works helps us understand who we are; on the other hand, millions of people are suffering from enormous mental, physical, and economic burdens caused by the brain disorders[1]. For example, 1 in 4 American adults suffer from a diagnosable mental disorder in any given year[2]; and 5.4 million people in the US have Alzheimer's disease (AD)[3]. Therefore, understanding how the brain works would help us make scientific progress on diagnostics, treatments, cures and management of brain disorders. The multi-faceted nature of neuroscience requires that its reach is broad and multidisciplinary. Our scientific and industrial community ought to play a leading role in helping achieve these goals.

Recent years have seen a rapid advancement in quantitative scientific research and a movement towards multidisciplinary study in neuroscience and statistics, due to the emergence of complicated problems, the production of an explosive amount of data, and the development of high-performance computers. There has been a massive amount of investment in research in neuroscience from both the federal government and private sectors. In 2015, NIH invested $32.3

[1] https://mcgovern.mit.edu/brain-disorders/by-the-numbers
[2] http://www.nimh.nih.gov/health/statistics/index.shtml
[3] http://www.alz.org/downloads/facts_figures_2012.pdf

billion[4] in medical research. Among those, $5.7 billion went into neuroscience, $3.9 billions in brain disorder, $1.7 bilions in neurodegenerative, and $298 millions in brain cancer. 16 major pharmaceutical and biotechnology companies established 20 venture funds[5] - 10 of 20 with funds each ranging from $100 million to $250 million - funding innovative technologies in these areas. In 2014, the United States launched the BRAIN Initiative project, a ten-year project with $3 billion investment from NIH to create a dynamic understanding of brain function; map every neuron in the human brain, based upon the Human Genome Project; and uncover the mysteries of brain disorders, such as Alzheimer's and Parkinson's diseases, depression, and traumatic brain injury (TBI). Internationally, European Union initiated the Human Brain Project (HBP) in 2013, a €1.19 billion 10-year project to simulate the human brain with supercomputers. One of the outcomes, BigBrain, is a high-resolution 3D brain atlas; Japan started the Brain/Minds project in 2014, introducing a 10-year project (with ¥3 billion investment during the first year) to map the primate brain so as to understand human disorders such as Alzheimer's disease and schizophrenia; and China followed in 2015 proposing the China Brain Project. Other completed and ongoing projects include the Human Connectome Project ($30 million, 2009+); Allen Brain Atlas ($100 million, 2003+); Blue Brain Project (EPFL, 2005+), CONNECT (EU); BrainMaps (NIH); NeuroNames (UW); Decade of the Brain; Talairach Atlas; Harvard Whole Brain Atlas; and MNI Template. Our science community needs to take advantage of this exciting time and the momentum to work collaboratively to apply our expertise to address some of the fundemental problems in neuroscience. Universities and industrial companies are organizations with leadning roles in many disciplinary fields, including basic science, medicine, and statistics, that could support essentially all aspects of neuroscience study. Many faculty members and students across different disciplines share strong interest in the interdisciplinary fields between statistics

[4] https://report.nih.gov/categorical_spending.aspx

[5] http://www.genengnews.com/insight-and-intelligence/top-20-corporate-venture-funds/77899832/

and neuroscience, and have a compelling demand for developing their knowledge in both areas to enhance multidisciplinary research. Unfortunately, currently there is a lack of any formal introduction to facilitate statistics research in neuroscience, and many programs in these two areas are working in relative isolation of one another. As researchers working in the interdisciplinary area between statistics, computer science, and neuroscience, we write this article to help define the role of statistics in contemporary neuroscience study, in honor of Sir R. A. Fisher's classical philosophy in statistics, which provides valuable guidance to readers who are interested in pursuing a career in statistical neuroscience.

What does it mean when one says "understand how the brain works"? (Understanding how the brain works is to) use neurobiologically plausible approaches and fully explicit computational models, and perform real world complex cognitive tasks to explain neural activity patterns and behavioural data, in human (Kriegeskorte, 2015). In (Fisher, 1925), Sir R. A. Fisher stated that "Statistics may be regarded (i.) as the study of populations, (ii.) as the study of variation, (iii.) as the study of methods of the reduction of data." Inspired by Sir Fisher's philosophy and Kriegeskorte's definition of brain study, the usefulness of statistics in neuroscience can be divided into three areas: (1) we investigate how the brain works neurobiologically by studying the brain in **populations**; (2) the function and structure of the brain vary from subject to subject and from time to time, we perform experiemental tasks and build computational models to study the **variations** of brain measurements in populations so as to provide confidence of our estimates while addressing uncertainty; and (3) to efficiently and effectively study large neural activity patterns and behavioural data, we need to study methods to **reduce large brain data** to a convenient amount that retains relevant information that our human minds and our computer memories are able to grasp yet sufficient to shed light upon original scientific questions. In the following, we shall further expatiate these three areas by including statistical approaches with regards to data science development in neuroscience.

First, the study of brain data is to gain insights to understanding how the brain perceives, processes, stores, and output information, in populations, or aggregates of individuals, rather than of individuals. The term population in brain science refers not only to an aggregate of brain activity measurements from multiple subjects, but also to an aggregate of a single brain measurement repeated multiple times for one subject. The former indicates our recognition of variations of brain activities amongst different individuals, whereas the latter represents our appreciation that the object of studying single subject brain activities is not to attempt to achieve an individual result, but rather, we make our best effort to ensure our findings representative. There are significant merits in studying data containing measurements of multiple subjects and those containing multiple measurements of single subjects. One of the end goals of brain science is to make scientific progress on diagnostics, treatments, cures and management of brain disorders. In order to raise the findings we have about the brain to the rank of science, we shall make statistical arguments about properties of the brain in a large aggregates of individuals. In order to produce treatments that target at a particular individual, we shall make statistical arguments about properties of the brain for that individual, based upon a large aggregates of measurements of his/her brain. Understanding how the brain works across subjects allows us to apply these principles at the individual level, and to advance applications that achieve artificial intelligence by mimicking the way an average brain performs, such as neural networks computers (Silver et al., 2016). Understanding how the brain works at the individual level would assist us in understanding how a specific brain and its activities deviate from the average. It hence leads to scientific progress such as an introduction of personalized medical plans, and a usage of brain signals to identify a subject (e.g. Finn et al. (2015) and Wachinger et al. (2014)). With an advancement of data acquisition technology and the popularization of high-performance computers, we are obtaining brain data in an unprecedentedly high-resolution, rapid, and accurate manner. Yet, there are strides to make. We shall advance our understanding of how the brain

works in different types of populations: infants V.S. adults, females V.S. males, etc., how the brain signals change across time, and how brain signals change according to different (visual, auditory, sensory, etc.) inputs. Furthermore, we shall reduce the errors caused by measurement and data processing, via improving and developing proper statistical and computing techniques. Additionally, we shall aim to increase the sensitivity of our study. It allows pharmaceutical companies to develop affordable medicine that would treat specific brain disorders for the majority of patients.

Second, the brain is an extremely complicated organ stored in a blackbox. Despite the advance in brain science, little do we know about how information is processed in the box. For example, does the brain process information linearly, or more plausibly, non-linearly (but in which form)?; (b) there is a tremendous amount of variations amongst different brains in terms of sizes, volumes, shapes, etc; and (c) there is much variation in measuring brain signals. Whilst the first challenge is extensively tackled by physicists and computer scientists via the studying of dynamic systems, spiking neural networks, and other neural networks (e.g. recurrent neural networks (Medsker and Jain, 2001); Boltzmann neural network (Aarts and Korst, 1988); deep neural networks (Schmidhuber, 2015); adaptive neural networks (Ghiassi et al., 2005); radial basis networks (Broomhead and Lowe, 1988)), statisticians working on neuroscience are actively seeking to solve the latter two. Once we have identified our goal in studying the brain in populations, it is a natural follow-up to study variations because the brains in populations display variation in one or more aspects. We, nevertheless, are not interested in variation of the brain *per se*; rather we recognize that variation is an inevitably troublesome by-product delineating circumstances where repeated measurements of the brain deviate from the average. Therefore, while describing the absolute properties of the brain (via parameters, e.g. mean activation intensity of a region of the brain), we encompass them with variances to address their uncertainty (and confidence). The introduction of variances leads to two further areas of statis-

tical studies in brain science: the study of frequency distributions, and the study of correlations. The frequency distribution may be expressed as a mathematical function of the variate (e.g. voxel-specific t-statistic), either (i.) the proportion of the population (regions, voxels, neurons, etc.) for which the variate is less than a given value, or (ii.) by differentiating this function, the infinitesimal proportion of the population for which the variate falls within any infinitesimal element of its range. In brain science, many frequency distributions are heavy-tailed - hence the study of them has some implication in studying certain financial models, e.g. Liu et al. (1999)), and *vice versa*. On the other hand, we are not only interested in studying the variations of the parameters of interests at present, but also interested in estimating the quality and types of these variations. Especially, we are interested in examining the simultaneous variation among multiple variates. It, therefore, gives rise to the correlation analysis. For ultra-highdimensional brain data, however, a voxel-wise correlation analysis could be unmanageably troublesome. This leads to the following section.

The third usefulness of statistics in brain science is due to the practical need of reducing large bulks of data to a convenient amount that retains relevant information in the original data that our human minds (and our computer memories) are able to grasp, by means of a manageable amount of numerical values. How much data reduction, however, should we conduct? In all cases, it is possible to reduce data to a simple numerical form, or to an amount that our computers are able to efficiently handle, where, the reduced data are sufficient to shed light upon scientific questions the investigator has original in mind. In brain science, two useful approaches in conducting data reduction are (I) principal component analysis (PCA) and its variants and (II) assuming sparsity. The PCA method captures the majority of the variation of the data. However, in brain science, oftentimes we have data with $p \sim n$ or $p \gg n$, under which its estimates are inconsistent. There are a few papers on sparse PCA that have demonstrated subspace consistency. For example, Ma et al. (2013) and Jung et al. (2009). The sparsity assumption makes neuro-

biological sense in the following manner: the working brain consumes energy. At any given time, be it resting state or task state, only a portion of the neurons are activated so as to reserve energy. We had an amiable conversation with Professor Pien-Chien Huang[6], during which he mentioned that we human beings do not dream in color (or at least have dreams less vivid and coloful). Schwitzgebel et al. (2006) has an article discussing this. We conjecture (with absence of scientific evidence) that a part of the reason is the brain attempts to reserve energy while sleeping (so only the minimal amount of information is processed: e.g. the brain only recalls the outlines, orientations, movements, etc. of objects. But they are sufficient to distinguish one from another and form visual events) - statistically a natural way of conducting data reduction! Recent years have seen a tremendous amount of exisiting published and on-going projects with regards to whole-brain connectivity (e.g. Allen et al. (2012)), high-dimensional mediation (e.g. Chén et al. (2015)), and brain decoding, such as facial recognition and dream decoding, (see decoding simple pictures: (Haxby et al., 2001), decoding objects with edges and orientations: (Haynes and Rees, 2005) and (Kamitani and Tong, 2005), decoding complex picutres: (Kamitani and Tong, 2005), decoding movies[7], decoding intentions: (Haynes, 2011), and decoding dreams: (Horikawa et al., 2013)); when the size of data becomes massive, it is considerably helpful and necessarily to conduct data reduction prior to further analysis.

To conclude, the prevalance and severity of brain diseases call for scientific collaboration and progress on diagnostics, treatments, cures and management of brain disorders; the emergence of complicated problems, the production of an explosive amount of data, the development of high-performance computers, and the heavy government and organizational funding investment provide us adequate resources and tools to study the brain in modern time. Due to the nature of the challenges in neuroscience, statisticians and data scientists play a growingly important role in addressing these issues, by means of studying the brain in populations, un-

[6] http://www.jhsph.edu/faculty/directory/profile/323/pien-chien-huang
[7] https://www.youtube.com/watch?v=nsjDnYxJ0bo

derstanding the variations of brain measurement to make scientific findings representative and reproducible, and finding convenient scientific paths to extract relavant and succinct information from explosively massive brain data. Finally, we encourage our readers to join us to raise our society's awareness of diseases caused by the human brain and related public health problems, and to help make scientific progress on diagnoses, treatments, cures and management of these diseases.

Acknowledgement

I would like to thank Professors Martin Lindquist and Brian Caffo of Johns Hopkins Bloomberg School of Public Health (JHSPH) for their generous funding support and thought-provoking teaching and guidance in both statistics and neuroscience that allow myself to passionately work on this article. I would also like to thank Professor Charles (Chuck) Rohde for his introduction of Sir Fisher's "Statistical Trilogy" books, and for his teaching in foundations of statistics. I would like to dedicate this article to my late grandfather, who always encouraged me to dedicate my career to serve for society and people.

References

Aarts, E. and J. Korst (1988). Simulated annealing and boltzmann machines.

Alivisatos, A. P., M. Chun, G. M. Church, R. J. Greenspan, M. L. Roukes, and R. Yuste (2012). The brain activity map project and the challenge of functional connectomics. *Neuron 74*(6), 970–974.

Allen, E. A., E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, bhs352.

Broomhead, D. S. and D. Lowe (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, DTIC Document.

Chén, Y., C. Crainiceanu, E. Ogburn, B. Caffo, T. Wager, and M. Lindquist (2015). High-dimensional Multivariate Mediation with Application to Neuroimaging Data. *arXiv:1511/09354v1/stat-ME*.

Finn, E. S., X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Ghiassi, M., H. Saidane, and D. Zimbra (2005). A dynamic artificial neural network model for forecasting time series events. *International Journal of Forecasting 21*(2), 341–362.

Haxby, J. V., M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science 293*(5539), 2425–2430.

Haynes, J.-D. (2011). Decoding and predicting intentions. *Annals of the New York Academy of Sciences 1224*(1), 9–21.

Haynes, J.-D. and G. Rees (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience 8*(5), 686–691.

Horikawa, T., M. Tamaki, Y. Miyawaki, and Y. Kamitani (2013). Neural decoding of visual imagery during sleep. *Science 340*(6132), 639–642.

Jung, S., J. Marron, et al. (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics 37*(6B), 4104–4130.

Kamitani, Y. and F. Tong (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience 8*(5), 679–685.

Kriegeskorte, N. (2015). A new framework for modeling brain information processing.

Liu, Y., P. Gopikrishnan, H. E. Stanley, et al. (1999). Statistical properties of the volatility of price fluctuations. *Physical Review E 60*(2), 1390.

Ma, Z. et al. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41*(2), 772–801.

Medsker, L. and L. Jain (2001). Recurrent neural networks. *Design and Applications*.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks 61*, 85–117.

Schwitzgebel, E., C. Huang, and Y. Zhou (2006). Do we dream in color? cultural variations and skepticism. *Dreaming 16*(1), 36.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature 529*(7587), 484–489.

Wachinger, C., P. Golland, and M. Reuter (2014). Brainprint: identifying subjects by their brain. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pp. 41–48. Springer.