 doi:10.20944/preprints202008.0428.v1

On Statistical Analysis of Brain Variability

Oliver Y. Chén[1,2*], Huy Phan[3], Guy Nagels[4], and Maarten de Vos[5,6]

[1]Department of Engineering, University of Oxford, Oxford OX1 3PJ, UK.

[2]Division of Biosciences, University College London, London WC1E 6DE, UK.

[3]Department of Computer Science, Queen Mary University of London, London E1 4NS, UK.

[4]Department of Neurology, Universitair Ziekenhuis Brussel, Brussel 1050, Belgium.

[5]Faculty of Engineering Science, KU Leuven, Leuven 3001, Belgium.

[6]Faculty of Medicine, KU Leuven, Leuven 3001, Belgium.

*Correspondence to: O.Y. Chén, Institute of Biomedical Engineering, University of Oxford, Oxford OX3 7DQ, UK. yibing.chen@seh.ox.ac.uk.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

Abstract

A century ago, Sir R.A. Fisher introduced for the first time the concept of variance in biological studies. In this paper, we present a few new, modified, or integrated perspectives of variance that we feel would contribute to future thinking and practice of data science. We do so by focusing on brain and behavioral data, through which we hope one could extrapolate the discussions to other fields and data. Specifically: (1) We define different types of variation. (2) We demonstrate that both classic regression models and advanced statistical methods can be viewed as variance-decomposition methods. (3) We make a distinction between innate and acquired variability, linked through Bayesian updating. (4) We review and illustrate how to extract information from high-dimensional data and how to visualize them. Additionally, we introduce the Neural Law of Large Numbers. (5) We discuss the statistical basis for association, explanation, prediction, and causation, and recommend a strategy that may be useful to check if association-based findings can be raised to causal discoveries. Taken together, to understand the variation of data, one needs creative statistical thinking. Meanwhile, by incorporating insights learned from data, one can begin to design better statistical apparatuses.

Significance Statement

A key task in data science is to extract succinct, simple, and useful information from large-scale, complex, and variable data to (1) form constant knowledge, rules, and principles to which future research can conveniently and reliably refer; and (2) when constant knowledge is unattainable, describe the variability and unveil its sources, associates, and consequences. Here, we introduce statistical foundations that are useful to study the variability of functional organization, structural topography, and longitudinal trajectory of the brain, arguably the most varying organ. Through the glimpse of brain studies, we hope that the discussed statistical thinking and strategies can be extended to general topics in science to transfer data into knowledge.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

Prologue

The pioneers of scientific discoveries had their origins looking for systematic patterns and irregular signals in nature. Their inspections spawned questions and, through hypotheses and careful verification, established laws, breaking the barriers between ignorance, observations, and knowledge.

James Maxwell established the connection between electromagnetism and light by investigating waves of oscillating electric and magnetic particles disturbed by light, advancing the studies of differential equations. Dmitri Mendeleev studied the periodicity of relative atomic mass and valence of elements and predicted the existence of gallium and germanium based on missing elements on the Periodic Table. Observing the recessional velocity of galaxies, Edwin Hubble and Georges Lemaître provided first evidence for a hitherto philosophical question, universe expansion, anticipated the Big Bang theory, and opened a new chapter for modern cosmology.

In biology, Charles Darwin discussed the importance of variability in *On the Origin of Species* and argued that it is greatest in structures that evolve fastest (Darwin, 1859). R.A. Fisher, the father of modern statistics, introduced the concept of *variance* in his 1918 paper entitled *The Correlation between Relatives on the Supposition of Mendelian Inheritance* (Fisher, 1918). The inspection of varying cytoarchitecture in the brain by Alfred Campbell and Korbinian Brodmann marked the beginning of unveiling the structural architecture and functional organization of the brain (Brodmann, 1909; Campbell, 1905; Toro and Burnod, 2005). Subsequently, by linking (co-)varying neural and behavioural apparatuses, scholars begun to chart the brain's functional organization and structure architecture (Broca, 1861; Croxson et al., 2018; Friston, 2012; Fritsch and Hitzig, 1870; Gordon et al., 2017; Halliday et al., 2017; Seghier and Price, 2018; Smith et al., 2019; Wernicke, 1874; Zeki et al., 1991; Zeki, 1993), unfold the neural origin of cognition and behaviour (Corbetta and Shulman, 2002; De Felice and Holland, 2018; Goldman-Rakic, 1988), and trace potential causes for brain disorders (Cao et al., 2018; Christ et al., 2018).

The major theme of this paper is to connect variability present in brain and behaviour with statistical foundations that are useful to study it. Although the discussions are made using neural and behaviour data, the statistical methods and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

applications we present herein could be modified or extended to investigate the variation of data in a broad range of fields, such as the study of variability in aesthetic perception. Although we cannot list every statistical device or its derivatives for studying variation, we hope what we examine here may stir further discussions about general data-science principles and practices, and that our ever-developing knowledge in statistical, data, and biological sciences will someday allow us to assemble a more comprehensive list of statistical foundations of variation.

As a preamble, we outline the topics covered in this paper:

i. We define different types of variability.

ii. We argue that a useful way to study variation of data is to decompose the total variance, via classic statistical theory and advanced modelling, into elements that can be attributed to internal or external factors. These factors offer a descriptive statement, can explain an associative or predictive phenomenon, or can test a causal claim either about the outcome or about the relationships between those factors and the outcome.

iii. We argue that a distinction needs to be made between two classes of variability: innate variability and acquired variability.

iv. We suggest statistical devices to analyze very large-scale data and to reduce and visualize high-dimensional data. To uncover constant knowledge from very large-scale brain data, we propose the Neural Law of Large Numbers and present empirical evidence for it.

v. Many discoveries in biological sciences are based upon associative and predictive analysis. We discuss the statistical basis for association, explanation, prediction, and causation. We suggest a strategy that may be useful to check if association-based findings can be raised to causal discoveries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

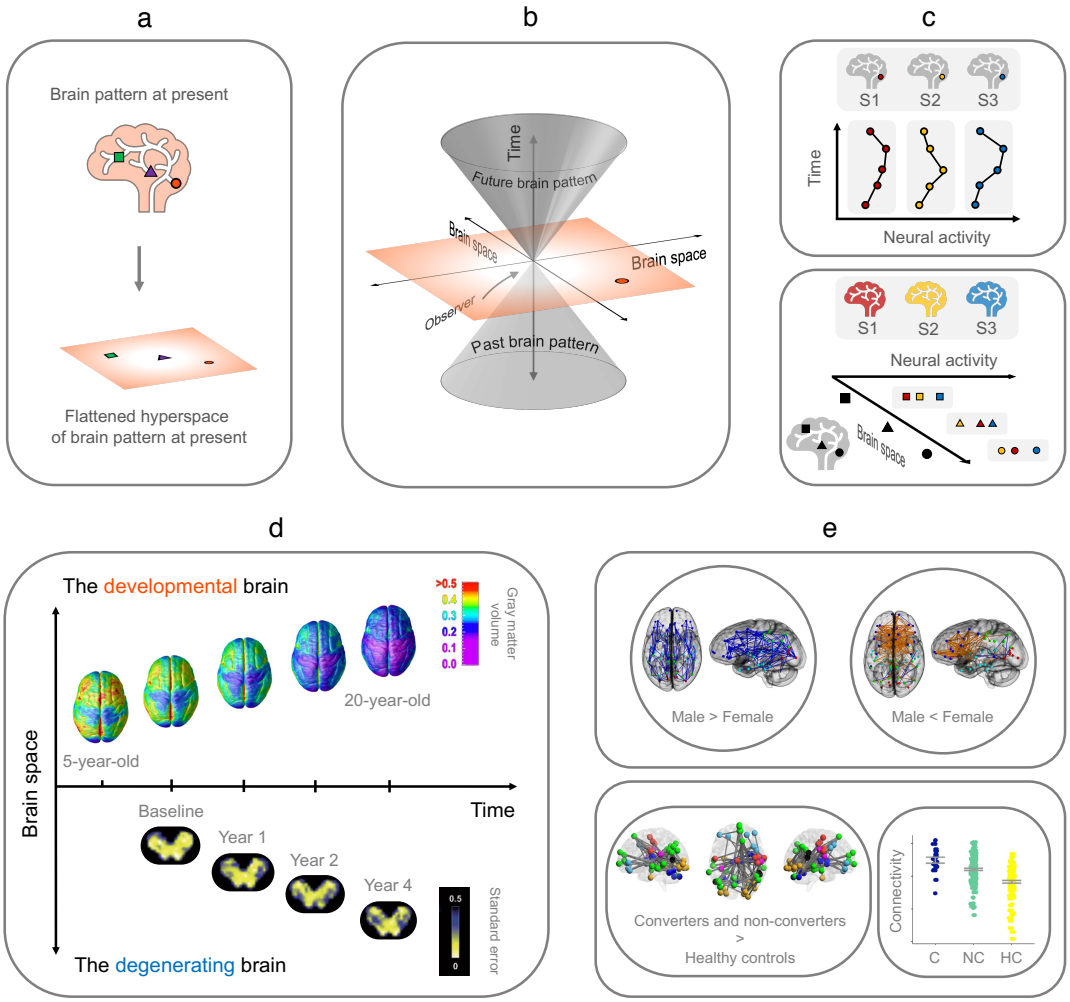

Figure 1. Defining variation in the brain. **(a) A 3-D brain without time.** Consider a 3-D brain space (shown as a 2-D slice of the brain) where each point (green square, purple triangle, or the red dot) can be pinpointed by a 3-D Euclidean coordinate (x, y, z). The brain space is then flattened into a hyperspace in which any point has a one-to-one correspondence to the original 3-D brain. **(b) A 3-D brain travelling in time.** Suppose the reader is looking at the hyperspace at present. The highlighted hyperspace represents the neural activities across the whole brain space at time 0. Above the hyperspace locates future brain information; below the hyperspace is the past brain information. The farther the one moves away from the hyperspace, the farther into the future (or past) the brain information one shall see. **(c) The temporal, spatial, and within-group variation of the brain.** The top panel consists of the same brain area (the coloured circle) of three individuals (S1, S2, and S3) – each colour indicates data of a specific subject. Each brain area is measured over time. The variability of these data points forms the temporal variation. The bottom panel consists of three distinctive brain areas (square, triangle, and circle) from three individuals (S1, S2, and S3) - each colour indicates data of a specific subject. The distribution of data across different brain areas depicts the spatial variation of the brain. If we fix both space and time, the distribution of data across different individuals in the group forms the within-group variation of data respective to that brain area. **(d) The development and degenerating brain.** By focusing on the temporal variation in the brain during childhood and adolescent, one can uncover how the brain develops (Gogtay et al., 2004). By tracing the course of a disease brain data over time, one can study how a brain disease (*e.g.*, the Parkinson's disease) progresses (Burciu et al., 2017). **(e) The between-group variation of the brain.** The heterogeneous distributions of brain data across different groups (*e.g.*, male *vs.* female (Ingalhalikar et al., 2014), and healthy *vs.* disease (Cao et al., 2018)) present the between-group variation of the brain.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

We hope to make this piece accessible to a broad readership. We therefore keep all mathematical notations in this article to a minimal. Wherever equations appear, they are necessary to clarify and support our arguments. Some sections are inevitably more mathematically heavy than others, so we have tried to make each of the sections a self-contained topic - interlaced by statistical methods and neurobiological findings – therefore skipping one (mathematically involved) section will not affect the reading of another.

1. The architecture of variation in the brain

A powerful way to think of the living brain is to picture it as a four-dimensional (spacetime) manifold, consisting of a three-dimensional brain space and one-dimensional time (see **Figure** 1 a-b). To understand the variation in the brain, one can beginning by distinguishing the very sources of variability attributed to space and time. Next, one needs to distinguish brain variation within an individual and between individuals. For between-individual variation, one needs to further differentiate variations that are relatively homogeneous within a particular group but heterogeneous across groups, such as female group *versus* male group, and health group *versus* disease group (Chén, 2019).

To formally define the different types of variability in brain studies, let's first define $y_i(\boldsymbol{v}, t)$ as the brain activity measured at location $\boldsymbol{v} = (x, y, z)$ at time t from an individual i; the size of y indicates how intense the activity is. For demonstration purpose, in our discussion we state the location and the type of data recorded from it in general terms. The former can be a neuron, a voxel, or a predefined brain region; the latter can be action potential of a single neuron, BOLD fMRI of a brain voxel, or EEG recordings from an area on the scalp.

If we fix the location $\boldsymbol{v}$ in the brain space and follow its activities in time, the trajectory of the data shows the *temporal variation* of the brain (see the top panel of **Figure** 1 c). If we fix a time point t, then the distribution of data across different brain areas depicts the *spatial variation* of the brain (see the bottom panel of **Figure** 1 c). If we fix both space at $\boldsymbol{v}$ and time at t, the distribution of data across different individuals in a group forms the *within-group variation* of brain area $\boldsymbol{v}$ (see **Figure** 1 c). Finally, the heterogeneous distributions of brain data across different groups (*e.g.*, male *vs.* female, and healthy *vs.* disease) present the *between-group variation* (see **Figure** 1 e).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

Each path of each type of variability leads to a specific area of brain study. To understand a healthy adult brain is in its own right a difficult task. To understand how a child's brain becomes an adult brain is perhaps more difficult an enterprise. By tracking the brain activity over time during childhood and adolescence, one can begin to understand the maturation and development of the brain (see top panel of **Figure** 1 d) (Casey et al., 2000; Gogtay et al., 2004). At the other end of the timeline, following the brain activity over time in a degenerating brain, one can begin to understand how the brain ages (Garrett et al., 2017; Yankner et al., 2008) and how neurodegenerative diseases, such as the Parkinson's disease, progress (see bottom panel of **Figure** 1 d)(Burciu et al., 2017). Studying past neural activities, one gains insights into future activities of the brain and may be able to make forecasts (Zhang and Shen, 2012). By studying the (co-)variability across brain space, one paints a picture of the structural topology and functional organization of the brain (Biswal et al., 1995; Bullmore and Sporns, 2009). By comparing brain patterns across multiple carefully arranged groups of individuals, one can not only uncover group-specific characteristics, but also detect differences between groups. The former allows the extraction of common, determining features of a group, for example, that are specific to a female brain and that are to a male brain (DeCasien et al., 2020; Ingalhalikar et al., 2014). The latter advances individual classification (into correct groups); for example, whether an individual brain belongs to a healthy group, an at-risk group, or a disease group (Cao et al., 2018). By associating co-varying features and outcomes, one may begin to find neural markers that acquire knowledge from external stimuli such as pain (Rogachov et al., 2016), that infer knowledge by means of behaviour (Cao et al., 2018; Chén et al., 2019; Gabrieli et al., 2015; Power et al., 2011; Reinen et al., 2018), and that are mediating a stimulus and an outcome (Chén et al., 2018; Koban et al., 2019).

2. Sherlock: detecting source of variability via variance decomposition

A useful way to extract spatial, temporal, subject- and group-specific information, as well as covariate effect related to an outcome is to strategically decompose the total variance of the outcome.

To understand this, let Y denote an output variable (or dependent variable) of scientific interest. Examples of Y include measured brain activities or observed behaviour measurements. Consider an extreme case where there is no variability in Y: thus, there is little we can or need to learn about Y because all information about it is preserved in Y itself – namely a constant. More practically, Y does vary, and its

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

variation is manifested in spatial, temporal, subject-, and/or group-specific domains, with each domain encodes a portion of the total variability of Y. Similarly, if a covariate X affects Y, then X must also explain a part of the variability of Y.

Analysis of variance (ANOVA) and analysis of covariance (ANCOVA) of repeated measurements across brain space over time among different individuals thus can be regarded as variance decomposition methods of the total variance (of Y). Yet, in addition to noise and measurement error, the error (or residual) term in ANOVA or ANCOVA contains a portion of as-of-yet unexplained total variance (that is, a portion that can neither be attributed to spatiotemporal factors nor covariates) that can be further decomposed and explained.

2.1 Variance decomposition via ANOVA and ANCOVA

The term of variance decomposition is perhaps not too unfamiliar. But in classic analysis of variance or covariance models, the total variance is typically decomposed into components associated with repeated measurements from different treatments and/or covariates (Fisher, 1925; Gelman, 2005). In fact, when Fisher first introduced the concept of *analysis of variance* in his monumental book (Fisher, 1925), he considered it as the study of the variation of populations (aggregates of individuals) and population of measurements (a simple measurement repeated indefinitely), without distinguishing the types of variations – perhaps considering the distinction too trivial. Here, to better understand the variabilities of the brain and behaviour, we argue that there is a need to make an appropriate distinction (see **Figure** 2). The reasons are twofold. Biologically, outcomes in a biological system encode temporal, spatial, individual-, and population-level sources of variation, and/or variation related to covariates. Methodologically, classic and advanced statistical methods can be viewed as variation decomposition approaches that break the total variance into components specific to time, space, individual-specific features, population-level characteristics, and/or covariate effects (see below).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

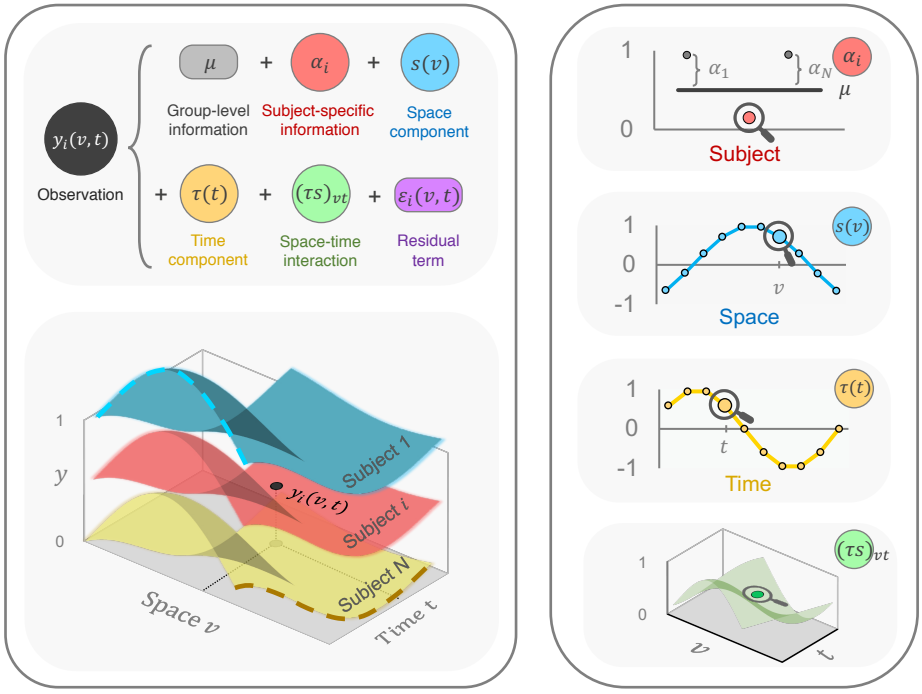

Figure 2. Decomposing the total variance. Top left: A function $y_i(v,t)$ can be decomposed to include a population mean μ, a subject-specific deviation α_i, a temporal fluctuation $\tau(t)$, a spatial shift $s(v)$, a spatial-temporal interaction $(\tau s)_{vt}$, and the residual terms $\varepsilon_i(v,t)$ which contains noise and information not yet explained by the aforementioned components. **Bottom left:** A 3-D construction of $y_i(v,t)$ for three individuals at different time points and brain space. One particular point $y_i(v,t)$ is highlighted which represents the outcome of individual i at location v at time t. **Right:** Cross-sections of each source of variation. From the top to bottom are: (1) population mean μ and a subject-specific deviation α_i, (2) spatial shift $s(v)$, (3) temporal fluctuation $\tau(t)$, and (4) spatial-temporal interaction $(\tau s)_{vt}$.

Suppose there are N individuals, whose V brain regions are measured for T times. Denote $y_i(v,t)$ as the observation corresponding to the i^{th} subject, v^{th} brain region, at time t, for $1 \le i \le N$, $1 \le v \le V$, and $1 \le t \le T$. Consider a bi-variate regression model (see Figure 2) as follows:

In words	Model	Parameter estimates
The observed data $\boldsymbol{y_i(v,t)}$ measured from the $\boldsymbol{v^{th}}$ brain region at time $\boldsymbol{t}$ from subject $\boldsymbol{i}$ contains group-level information $\boldsymbol{\mu}$, subject-specific information $\boldsymbol{\alpha_i}$, space component $\boldsymbol{s(v)}$, a time component $\boldsymbol{\tau(t)}$, a	$y_i(v,t) = \mu + \alpha_i + s(v) + \tau(t) + (s\tau)_{vt} + \varepsilon_i(v,t).$	$\hat{\mu} = \frac{\sum_{i=1}^{N}\sum_{v=1}^{V}\sum_{t=1}^{T} y_i(v,t)}{NVT} = \bar{y}_{...}$, $\hat{\alpha}_i = \frac{\sum_{v=1}^{V}\sum_{t=1}^{T} y_i(v,t)}{VT} - \bar{y}_{...} = \bar{y}_{i..} - \bar{y}_{...}$, $\hat{s}(v) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_i(v,t)}{NT} - \bar{y}_{i..} = \bar{y}_{.v.} - \bar{y}_{i..}$, $\hat{\tau}(t) = \frac{\sum_{i=1}^{N}\sum_{v=1}^{V} y_i(v,t)}{NV} - \bar{y}_{.v.} = \bar{y}_{..t} - \bar{y}_{.v.}$,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

space-time interaction $(s\tau)_{vt}$, and a residual term $\varepsilon_i(v,t) \sim N(0,\sigma^2)$.		$(\widehat{s\tau})_{vt} = \frac{\sum_{i=1}^{N} y_i(v,t)}{N} - \bar{y}_{..t} = \bar{y}_{.vt} - \bar{y}_{..t}$.

One can decompose the total sum of squared residuals or $SST = \sum_{i=1}^{N}\sum_{v=1}^{V}\sum_{t=1}^{T}(y_i(v,t) - \bar{y}_{...})^2$ into variance attributed to (a) individuals (SS_α), (b) difference across brain space (SS_{Space}), (c) fluctuation in time (SS_{Time}), and (d) sum of squared errors (SSE). Specifically, $SS_\alpha = VT\sum_{i=1}^{N}(\bar{y}_{i..} - \bar{y}_{...})^2$ with $N-1$ degrees of freedom (df), $SS_{Space} = NT\sum_{v=1}^{V}(\bar{y}_{.v.} - \bar{y}_{...})^2$with $V-1$ df, $SS_{Time} = NV\sum_{t=1}^{T}(\bar{y}_{..t} - \bar{y}_{...})^2$ with $T-1$ df, $SS_{Space/Time} = N\sum_{t=1}^{T}\sum_{v=1}^{V}(\bar{y}_{.vt} - \bar{y}_{...})^2$ with $(N-1)(V-1)$ df, and $SSE = \sum_{i=1}^{N}\sum_{v=1}^{V}\sum_{t=1}^{T}\left(y_i(v,t) - \hat{y}_i(v,t)\right)^2$, with $[NVT - (N-1)(V-1) - (N-1) - (V-1) - (T-1) - 1]$ df, where $\hat{y}_i(v,t) = \hat{\mu} + \hat{\alpha}_i + \hat{s}(v) + \hat{\tau}(t) + (\widehat{s\tau})_{vt}$. Given each component of the total variance due to individual, space, and time, and their corresponding degrees of freedom, one can subsequently test the individual, spatial, or temporal effect by using the *F*-tests (Fisher, 1925)(Wu and Hamada, 2000).

When necessary, one can choose to discount temporal variation and the interaction terms (if one knows *a priori* that the variability over time is not significant or is not interested in temporal variation) by treating the temporal measurements over time as repeated measures and estimate only the subject-specific information α_i and spatial information $s(v)$. This can be summarized by the following model: $y_i(v,t) = \mu + \alpha_i + s(v) + \varepsilon_i(v,t)$. One can also choose to disregard individual difference (if one's interest is in obtaining only the shared, population-level information) by treating multiple samples of individual data as repeated measurements and estimate only the population-level spatial and temporal information $s(v)$ and $\tau(t)$. This can be summarized by the following model: $y_i(v,t) = \mu + s(v) + \tau(t) + \varepsilon_i(v,t)$. One can furthermore extend the model to include covariates (as in the analysis of covariance (ANCOVA)): $y_i(v,t) = \mu + \alpha_i + s(v) + \tau(t) + (s\tau)_{vt} + \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i(v,t)$, where $\boldsymbol{x}_i$ denotes a vector of covariates (such as age, gender, and BMI) and $\boldsymbol{\beta}$ represents the associated parameter vector.

Whether ANOVA or ANCOVA, and regardless of how many interaction terms each model includes (one can add two- and three-way interactions to the above models), the central point we wish to make here is that: (1) there is a need to separate the total variance of phenotypes (including neural and behavioural patterns) in brain studies into isolated components that can be explained by factors of interests, such as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

individual effect or α_i, temporal fluctuations or $\tau(t)$, spatial patterns or $s(v)$, interactions or $(s\tau)_{vt}$, and covariates such as age and gender or $\boldsymbol{x}_i$; and (2) regression models used to identify, isolate, and quantify sources of variability can be viewed as variance decomposition methods (with regards to the total variability of the outcome).

2.2 Residual variance decomposition

We have so far shown that regression models can be interpreted as variance decomposition approaches, yielding variance components that can explain the total variance (or information). Here we push the concept of variance decomposition a little further by decomposing the residual terms which are otherwise largely overlooked. We use longitudinal statistics models (which track temporal variation) as a gateway. We show that frameworks such as random-effect models and transition (or Markov) models can be seen as two-stage variance decomposition methods where the second-stage decomposition is performed on the residual term. Interested readers can extend this to study variability observed in other domains.

Let's first state what a general longitudinal problem is. Consider n subjects and p features each of which is measured m times (one can relax it to n_i measurements per person; but for simplicity here we assume every individual has the same number of repeated measurements). Consider y_{ij} as j^{th} repeated (outcome) measurement of the i^{th} subject, where $1 \leq i \leq n$ and $1 \leq j \leq m$. Consider x_{ijk} as j^{th} repeated (feature) measurement of the i^{th} subject on the k^{th} feature, where $1 \leq k \leq p$.

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{im})^T$, for $1 \leq i \leq N$. Let $\boldsymbol{y} = (y_1^T, y_2^T, \dots, y_n^T)^T$. Let $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijp})$, for $1 \leq i \leq n$ and $1 \leq j \leq m$. Let $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, \dots, \boldsymbol{x}_{im}^T)^T$, for $1 \leq i \leq n$. Let $\boldsymbol{X} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \dots, \boldsymbol{x}_n^T)^T$. In words, $\boldsymbol{y}$ is a vector containing all y_{ij}'s (where the first m entries are repeated measures for the first subject, the second m entries are repeated measures for the second subject, and so on) and $\boldsymbol{X}$ is a matrix containing all x_{ijk}'s (where the first m rows contain features from the first subject, the second m rows contain features from the second subject, and so on; within each $m \times p$ subject-specific feature matrix $\boldsymbol{x}_i$, each column corresponds to a feature).

A standard longitudinal parametric model can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

where $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \text{blockdiag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$, $\boldsymbol{V}_i$ models the within-subject correlation structure, for $1 \leq i \leq n$. Note that: (a) an equivalent expression for the above model is $\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$; (b) Although linear, the expression can be modified to be a generalized linear model by adding a link function to recapitulate a non-linear relationship.

The residual term can be decomposed into three parts: random effects, serial correlation, and measurement error. Specifically (see, for example, p.83 in (Diggle et al., 2013)), for the j^{th} time point of subject i:

In words	Model
The residual term $\boldsymbol{\epsilon_{ij}}$ regarding subject $\boldsymbol{i}$ for the $\boldsymbol{j^{th}}$ time point can be decomposed to an random effect component ($\boldsymbol{d_{ij}^T U_i}$), a serial correlation component ($\boldsymbol{W_i(t_{ij})}$), and measurement error ($\boldsymbol{Z_{ij}}$).	$$\epsilon_{ij} = d_{ij}^T U_i + W_i(t_{ij}) + Z_{ij},$$ where U_i represents the random effect, which consists of a subject-specific Gaussian random vector with mean 0 and covariance G (d_{ij} are explanatory variables corresponding to individual measurements), $W_i(t_{ij})$ represents serial correlation consisting of a sample from a stationary Gaussian process with mean zero and correlation function $\gamma(u) = \rho(u; \phi)$, and Z_{ij} represents measurement error with Gaussian distribution of mean 0 and variance τ^2 . The correlation function needs to be further specified; popular choices are exponential correlation $\rho(u) = \exp(-\phi u)$ and Gaussian correlation $\rho(u) = \exp(-\phi u^2)$.

The decomposed results of the residual term, however, do not always have to contain all three parts as listed above. Depending on the research interests (and assumptions, for example, if one assumes that there are no random individual differences between subjects), then the error term needs not to contain a random subject-specific component). More specifically, models whose residual term only contains measurement error are called marginal models; models whose error term

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

contains both random effect and measurement error are called random effect models; and models whose error term contains both serial effect and measurement error are called transition (or Markov) models.

3. Innate variability, acquired variability, and the Bayesian brain

Speaking of variability in neural and behaviour outcomes, one needs to make a distinction between its innate and acquired sources. The innate *vs.* acquired distinction has long fascinated ethologists, ecologists, and sociologists (Brigandt, 2005; Burkhardt, 2005; Griffiths, 2004; Lorenz, 1957; Lorenz and Tinbergen, 1957; Tinbergen, 1942, 1951). The former includes both genetic and epigenetic signals (such as methylation patterns on the DNA or RNAs) (Griffiths, 2020). For example, in mammals, the Y-chromosome determines the gender; having three copies of the genes on chromosome 21 would cause Trisomy 21 (Down's syndrome). The acquired information contains environmental factors such as climate (*e.g.*, excessive exposure to sun may increase the chance of having non-melanocytic skin cancer (Kricker et al., 1994)), nutrition and diet (*e.g.*, exposure to Bisphenol A from contaminated food or drinks may increase one's risk of diabetes and cardiovascular disease (Bertoli et al., 2015)), and disease (*e.g.*, sexual transmitted disease such as syphilis may change a patient' appearance and cause neurologic and cardiovascular disease (Hook and Marra, 1992)). In addition to the *main effects* from the innate and acquired factors, there is a third source that may influence the total variability of phenotypes, that is, the interactions between factors. The interaction effects modulate phenotypes in two ways. The first is through gene-gene interaction or environment-environment interaction (see Waddington's developmental landscape (Waddington, 2014)). The second is through gene-environment interaction. For example, in bluehead wrasse (*thalassoma bifasciatum*), when a terminal phase males were removed from a shoal, many females changed into terminal phase males (Piper, 2007). Temperature (environmental signal) may also interact with genetic information. For example, some reptiles and turtles use incubation temperatures to determine sex and sex reversal (Ge et al., 2018; Quinn et al., 1983).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

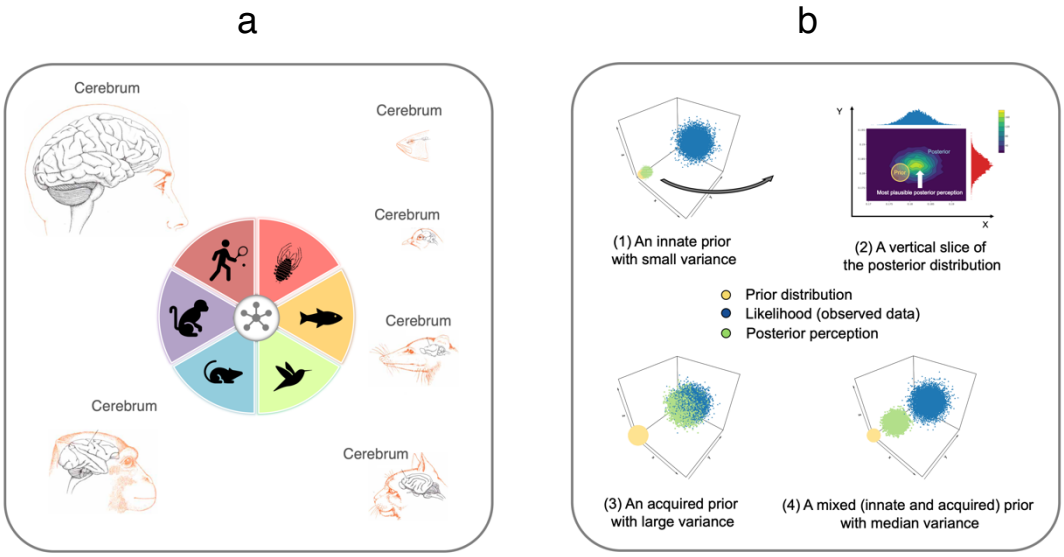

Figure 3. Two aspects of the Bayesian Brain: brain science and data science. (a) Bayesian brain. Empirical evidence for Bayesian updating has been found in animals and humans highlighted in the coloured circle. They are, clock-wise, amphipods (*Gammarus lawrencianus Bousfiel*), fish (*e.g.*, peacock wrasse (*Symphodus tinaca*)), birds (*e.g.*, Inca doves (*Columbina inca*), common cranes (*Grus grus*), and woodpeckers (*Dendrocopos mino*), small mammals (*e.g.*, Merriam's kangaroo rat (*Dipodomys merriami*), Arizona pocket mouse (*Perognathus amplus*) and the round-tailed ground squirrel (*Spermophilus tereticaudus*), monkeys, and humans. Cerebra of animals and human are shown to indicate the size difference of brains. Cerebra of animals are shown to indicate the size difference of brains in different animals. Cerebra images are redrawn from "The Brain" by David H. Hubel. Copyright Patricia J. Wynne 1979. **(b) Bayesian data.** If prior beliefs have greater precision (*i.e.*, have low variability), they will attract posterior beliefs which are not readily modifiable by experience and experimentation. Conversely, if priors are held with less confidence and precision, then experience and experimentation will modify the prior to produce a new posterior. This is illustrated in terms of probability distributions in different colours in the figure. This ubiquitous aspect of Bayesian belief updating underlies the necessity to distinguish two different sorts of priors. Acquired priors are obtained postnatally and therefore are hospitable to modification with learning and experience and are therefore adaptable. Conversely, innate priors are specified innately (possibly genetically) over evolutionary time scales. In consequence, they are much more resistant to belief updating, thereby providing a certain “stability” for inferring the sensorium. Priors are represented in yellow, observed data (experience) in blue, and the posterior produced from observation in green. Size of circles indicates the extent of variability. For comparison purposes, the observed data (blue dots) in (1), (3) and (4) are simulated using the same (multivariate Gaussian) distribution. (1) When the prior distribution (indicated by the yellow dots (see figure (2) for a zoomed-in snapshot) has very small variability (*i.e.*, high precision), the posterior distribution is derived mainly from the prior, and not much from the observed data. This constitutes the concept of an innate prior. (2) shows that, with innate priors, the centre of the posterior distribution is very close to that of the prior distribution. (3) shows that when the prior distribution (yellow dots) has a very large variability, the posterior distribution (indicated by the green squares) is modified mainly by the observed data (blue dots). This constitutes the concept of an acquired prior. Note that the centre of the posterior distribution in (3) is very close to the centre of the observed data (blue), under an acquired prior. (4) when the prior distribution (yellow) has moderate variability, the posterior distribution is derived from both the prior and the observed data; the centre of the posterior distribution is between the prior distribution (yellow circle) and the distribution of the data (blue). Image is adapted by permission from *The European Journal of Neuroscience* “The Bayesian-Laplacian Brain” by S. Zeki and O.Y. Chén. Copyright 2019.

The variance decomposition (see **Section** 2) opens a methodological door to separate the innate and acquired information and their interactions with regards to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

brain functions and structures. In other words, via variance decomposition, one can isolate and quantify the amount of variability in phenotypes for which is accounted by genetic information, environmental factors, and their interactions.

A direct application of variance decomposition is the heritability analysis in population genetics. Let $P = G + E$ denote the relationship between a phenotype (P) and its genetic information (G) and environmental factor (E). The total (phenotype) variance is therefore the sum of (a) genetic variance, (b) variance due to environmental factors, and (c) twice the covariance between genetic information and environmental factors; the heritability (or H^2) is quantified as the fraction of genetic variance over the total (phenotype) variance, namely $H^2 = \frac{V(G)}{V(P)}$.

If a great deal of the total variance of a trait (P) is explained by environmental factors, then: (1) only a small fraction of the total variance of P is explained by the genes, and (2) consequently there is little individual difference with regards to the trait (P) that is due to genetic information. Using a similar deduction, if the total variance of a trait is very small across subjects, then the chance that the trait is modified postnatally is small (as the variance due to the environmental factors can be no larger than the – very small – total variance).

By incorporating previous knowledge (obtained biologically through gene-encoding and acquired postnatally through learning) with new information, perception and behaviour are updated with higher precision (*i.e.*, lower variation). Empirical studies have suggested that both animals and humans are capable of performing Bayesian updating, integrating prior information and postnatal training (see **Figure** 3 (a)). When making mating decisions, male amphipods (*Gammarus lawrencianus Bousfield*) may use their prior information about (the distribution of) the quality of female amphipods to guide their amplexus (the behaviour that the male amphipods hold onto females prior to copulation) (Hunte et al., 1985). Female peacock wrasse (*Symphodus tinaca*) update their probability of finding a nesting male based on a prior probability distribution of successfully finding a nesting male each day; they begin with a noninformative prior and updated its distribution based on the outcome during each day and employ an updated informative prior at the start of each following day (Luttbeg, 1999). Given different prior distribution of food quality across patches, various of birds such as black-chinned hummingbirds (*Archilochus alexandri*) and budgerigars (*Melopsittacus undulates*) modulate their actions by their probability estimates of the food patch quality (Valone, 1992)(Valone and Giraldeau, 1993). Similar foraging behaviour was found in small mammals including Merriam's kangaroo rat (*Dipodomys*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

merriami), Arizona pocket mouse (*Perognathus amplus*), and the round-tailed ground squirrel (*Spermophilus tereticaudus*) (Valone and Brown, 1989). Monkeys estimate time by integrating sensory evidence with prior beliefs (Sohn et al., 2019). Humans use probability updating to modify their perception (Knill and Richards, 1996), cognition (Griffiths et al., 2008), and sensorimotor function (Körding and Wolpert, 2004).

But what does the theory of variance decomposition and empirical evidence of Bayesian updating inform us about the brain and behaviour studies? A direct lesson we have learned is the need to distinguish two types of variabilities: the innate variability and acquired variability (Zeki and Chén, 2019) (see **Figure** 3 (b)). The former is associated with innate properties of a biological organization, likely dictated by the genes, and are small. The latter is developed postnatally, due to environmental factors or a combination of environmental and genetic factors, and are large (relative to the innate variability). For example, colour perception is *a priori*, and varies little in humans. Perceiving the white colour (*i.e.*, achromatic) after seeing a white flag is independent of culture and education, because "(colour has) a definite structure (*i.e.*, concept, emphasis mine); it is natural to see them in certain ways and to conceive of them accordingly" (Kant, 1787; Pears, 1953). Perceiving truce, ceasefire, or surrender when seeing a white flag, however, is *a posteriori*, because it is dependent upon (postnatal) exposures associating and reinforcing the two; it is more variable across different ages and culture groups. Similarly, face recognition is *a priori*, while associating faces of different races with social categorization linked to stereotyping, prejudice, and discrimination is *posteriori* (Bar-Haim et al., 2006).

There is, however, one outstanding question about Bayesian updating, namely, whether it indicates causation. Neither the theoretical hypothesis of the Bayesian brain nor the empirical evidence for which a Bayesian model fits or predicts behaviours endorses that the organization and the functions of the brain *is* Bayesian. In **Section** 5, we shall discuss some aspects regarding the neural basis for association, explanation, prediction, and causation.

4. The neural law of large numbers and the very large-scale data

How does the varying brain, living in a continuously changing world, extract and retain constant and essential information from its environment, such as perceiving colour constancy of a surface despite different wavelength-energy composition of the light reflected from a surface? How do individuals studying the brain derive constant, converging knowledge from the varying brain data? One way to answer these

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

questions is to refer to the Neural Law of Large Numbers (NLLN). Specifically, the NLLN states that the average of a feature (over a large sample) will converge towards its expectation in a population and will tend to become closer to the expectation as more samples are collected. It is a special case of the mathematical law of large numbers (LLN, also known as the Bernoulli's Theorem)(Bernoulli, 1713) (Poisson, 1837) applied to biological feature variables.

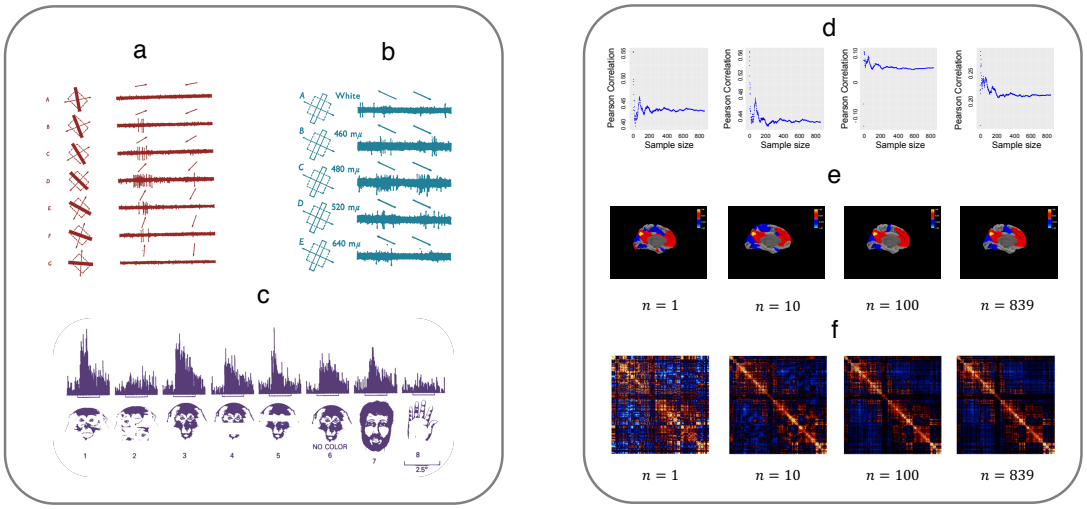

Figure 4. The biological insight of average signals and the Neural Law of Large Numbers (NLLN). **(a) The properties of a motion-selective neuron.** When a black bar moves at different orientation in front of a macaque monkey, single neurons from the striate cortex of the monkey are excited. This (orientation selective) neuron responds, on average, best to up-right motion. **(b) The properties of a colour-selective neuron.** When white and monochromatic light of different wavelength is present to a macaque monkey, single neuron from the layer 2 of the striate cortex of the monkey are excited. Under the same motion (so the variability obtained from the neuron is not due to motion), this (colour selective) neuron responds, on average, the best to wavelength of 450 $m\mu$. **(c) The properties of neurons for face recognition.** Neurons in the inferior temporal cortex of the monkey are involved in face recognition. Here, response distributions in the inferior temporal cortex are shown when the monkey viewed a normal monkey face, a normal human face, fractions of a monkey face, an abnormal monkey face, and non-face object (*e.g.*, a hand). These neurons respond, on average, the best to normal monkey and human faces. **(d) The NLLN for connectivity between two brain regions.** From left to right, each figure considers a specific edge (ROI to ROI Pearson correlation). Correlation between time-course from two ROIs is computed for each individual; the aggregated correlations are then averaged over the sample size. **(e) The NLLN for seed connectivity**. Maps of correlations between a seed (in the posterior cingulate (PCC) or the coloured bright yellow area) and the rest brain-regions are shown for n = 1, 10, 100, and 839 individuals. The colour map is obtained by averaging correlation between every other region and the seed region. As sample size increases, the average seed connectivity converges, tending to what is known as the brain's default mode. **(f) The NLLN for whole brain connectivity.** Whole brain connectivity maps (pair-wise Pearson correlations between 400 sub-brain areas) are shown for n = 1, 10, 100, and 839 individuals. As sample size increases, the average connectivity matrix converges. Images in (a) and (b) are reproduced by permission from RightsLink *Journal of Physiology* "Receptive fields and functional architecture of monkey striate cortex" by T. N. Wiesel and D. H. Hubel (Copyright 1968). Image in (c) is reproduced from "Stimulus-selective properties of inferior temporal neurons in the macaque" by Desimone *et al.* (Copyright 1984 *Society for Neuroscience*). Data in figure (d)-(f) are from the Human Connectome Project (Van Essen et al., 2013).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

4.1 The neural law of large numbers (NLLN)

The average of a feature, deceptively simple, offers important biological insights. Recall the pioneering physiological experiments on motion- and colour-selective cells carried out by Hubel and Wiesel (see **Figure** 4 a-b) (Hubel and Wiesel, 1968). When a black bar moved at different orientation in front of a macaque monkey, a single neuron from the striate cortex of the monkey fired. Apparently, the neuron responded to motions from several directions (see **Figure** 4 a). How, then, can one claim that this neuron was an orientation selective cell that selectively responded to the up-right motion? The mean activation turned out to be a simple, yet powerful indication: the neuron responded on average most strongly with regard to the up-right motion, suggesting that it was wired to be motion-selective for that specific direction. In parallel, when white and monochromatic light of different wavelength was present to a macaque monkey, a single neuron from the layer 2 of the striate cortex of the monkey fired. Under the same motion (so the variability obtained from the neuron was not due to motion), the neuron responded to colour of different wavelengths; but it responded, on average, most excitingly to wavelength of 450 mμ, suggesting that it was wired to be colour-selective for that specific wavelength (see **Figure** 4 b). Similarly, when the monkey viewed a normal monkey face, a normal human face, fractions of a monkey face, an abnormal monkey face, and non-face object (*e.g.*, a hand), neurons in the inferior temporal cortex of the monkey showed responses to all stimuli. But the average responses were the largest when the monkey viewed normal monkey faces and a normal human face, suggesting that these neurons were involved in face recognition (Desimone et al., 1984).

Returning to the NLLN, empirical data analyses showed that the connectivity strength for a fixed edge (connections between two brain areas), seed connectivity, and whole brain functional connectivity all obey the law (see **Figure** 4 d-f). This sends a powerful message about quantitative brain studies, that the average signals of a feature (which shed population information about the feature) one uncovers may still be fluctuating, even under a range of relatively large sample sizes, before they begin to stabilize. Thus, it is important to consider sample size calculation; when obtaining a suitably large sample is difficult, one may consider aggregating multiple datasets and investigate the appropriate sample under which the signals converge (but see **Section** 4.2 below).

The phenomenon that average signals converge asymptotically to the expectation does not endorse that the asymptotic mean equals to the truth, where the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

truth here means the true population average were there no noise, systematic error, or measurement error. What the NLLN unveils is that given the presence of noise, systematic or measurement error, a large number of samples will provide a more stable, and reliable estimate about the quantity of interests. Without a large number of samples, the results not only contain noise and error, but also may tend to be unreliable. In fact, assuming that the expectation of noise and measurement error are around zero, one can consider the asymptotic value under NLLN as the expected population value plus or minus an error that is inherent in the system (*e.g.*, systematic error due to a scanner); it would be difficult for a small data set to achieve this.

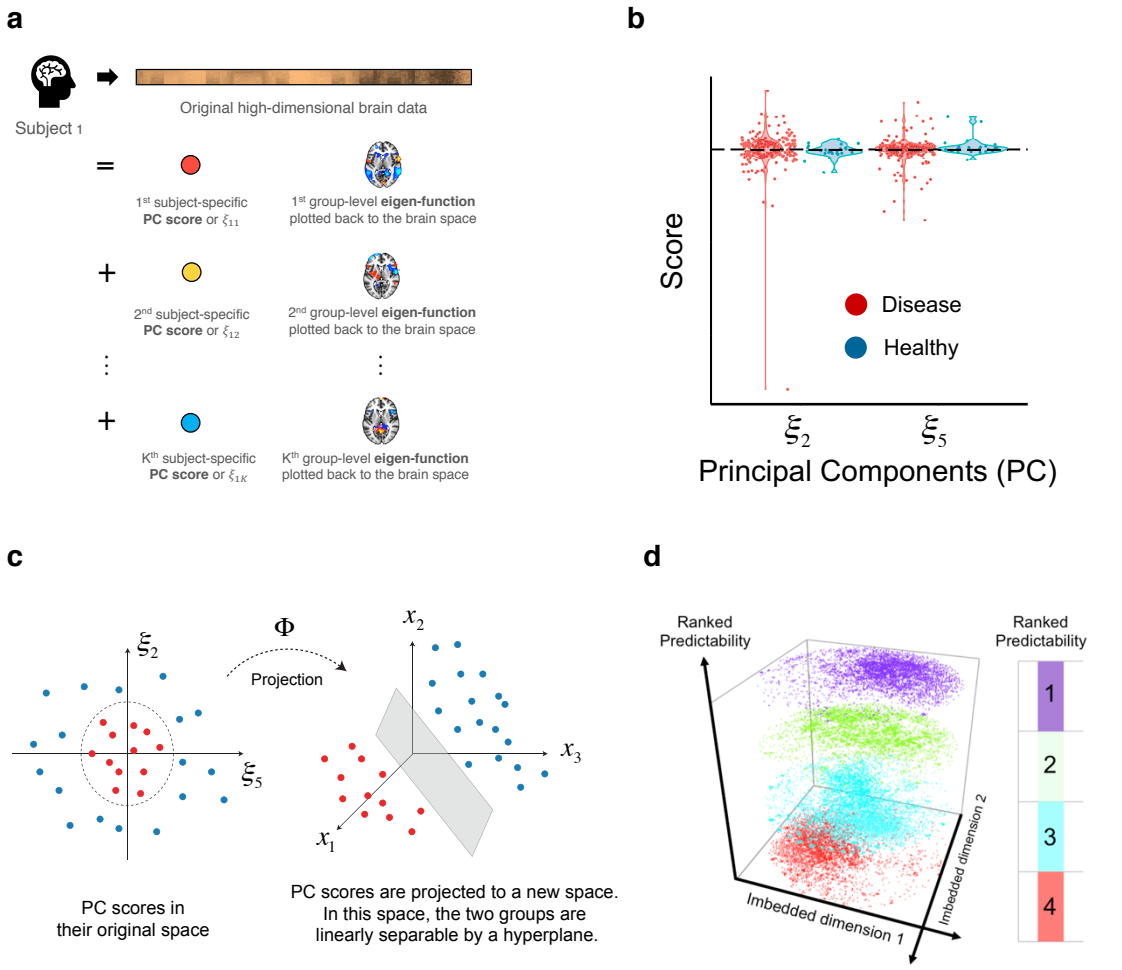

Figure 5. Dimension reduction for, feature extraction from, visualization of, and making predictions using large-scale brain data. (a) Using principal component analysis (PCA) to extract low-dimensional biomarkers from high-dimensional brain data. Low-dimensional principal component (PC) scores (coloured circles) and eigen-functions (brain maps) are extracted from each individual's raw features of length *p*. Here, for demonstration purpose we show the first, the second, and the Kth PC score as well as their corresponding eigen-functions plotted on the brain space. The PCs are subject-specific and can be used as subject-specific transformed features. The eigen-functions represents population-level patterns, indicating shared information across individuals. Here, each set of subject-specific PCs correspond to one brain atlas (indicated by an eigen-function), quantifying the departure each individual is from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

the group-level atlas. **(b) Using PC scores as extracted low-dimensional features.** In this dataset, the disease group in general have large second and fifth PCs that depart from zero (indicated by the dash line) corresponding PCs are closer to zero with smaller a variance. This indicates that these two PCs are potential biomarkers for distinguishing the two groups. **(c) Separating extracted features via a hyperplane.** Specifically, lower dimensional PC scores are trained via a support vector machine with radial basis kernel. The original spatial orientation of the PCs is delineated by (ξ_2, ξ_5). Notice the two groups cannot be separated linearly. Employing a mapping function $\Phi: \boldsymbol{\xi} \in \mathbb{R}^2 \mapsto \boldsymbol{x} \in \mathbb{R}^3, \Phi(\boldsymbol{\xi}) = (\xi_2^2, \sqrt{2}\xi_2\xi_5, \xi_5^2)$, projects the PCs onto a new space. The two groups are now linearly separable by a hyperplane in the new space, whose spacial orientation is delineated by (x_1, x_2, x_3). **(d) Visualizing high-dimensional brain data in 3-D space and using them to predict human behaviour.** First, a sliding window analysis is conducted on BOLD fMRI data from 839 subjects. Sliding windows were created by dividing data measured at 1,200 time points into chunks each consists of uniform temporal resolution (width = 43.2 s, 60 time points) with a Gaussian window (window = 6 s). This is repeated successively along the time course in steps of 3 repetition time (TR = 720 mm; 2.16 s), resulting in 381 windowed connectivity matrices for each subject. Each windowed connectivity matrix is of 400 × 400. The resulting 319,659 400 × 400 are then clustered (using *k*-means clustering) into four groups. Prediction of individual fluid intelligence (Gf) is then conducted using data from each of the four clusters. Each cluster of 400 × 400 matrices is then projected onto a 3-D space, using t-SNE. Each point in the figure represents a 400 × 400 windowed connectivity matrix; in other words, t-SNE embeds each 400 × 400 matrix into a point in a 2-D space. Each colour specifies a cluster to which a point belongs. The height of each coloured layer indicates prediction power of data from each cluster. For example, the purple cluster receives a ranking score of 1, and the red cluster has a ranking score of 4, indicating data from these two clusters are the most and the least predictive of fluid intelligence, respectively. Data in (d) are from the Human Connectome Project (Van Essen et al., 2013).

4.2 From the neural law of large numbers to very large-scale data

The relationship between the neural law of large numbers (NLLN) and very large-scale data (VLSD) resembles that of a fast car and a large quantity of gasoline. On the one-hand, a fast car needs a large quantity of gasoline to showcase its optimal speed; the NLLN needs VLSD to demonstrate its convergence (see above). On the other hand, adding a large quantity of gasoline may cause issue to a fast car. For example, aggregating different types of gasolines (some have more tetraethyl lead than others) increases the quantity but may cause damage to the optimal performance of the car. Similarly, VLSD may also cause issues to the NLLN. As a result of data aggregation from multiple sites, paradigms, and sessions, unprecedented amount of noise is introduced. For example, in an fMRI study where one million voxels are concerned, even if each voxel contains a small amount of noise, the aggregation of the noise from all can become sizable, confounding meaningful signals.

To address the above issue, namely noise aggregation due to a large number of (*e.g.*, high-dimensional) features, one can rely on sparse assumption. This means that only a small number of the features or networks are relevant to underpin a biological system (see Occam's razor (Baker, 2004; Jefferys and Berger, 1992; Sober,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

1990)). In other words, we select a subset of features or networks via variable selection or regularization prior to further analysis. The sparse assumption is not made categorically for technical convenience; it is also supported by biological evidence. For example, small number of network hubs and many poorly connected nodes have been found in cells (Barabási and Oltvai, 2004), the brain (van den Heuvel and Sporns, 2013), and the genes (Leclerc, 2008), suggesting the existence of sparse networks and shedding biological light on the sparse assumption.

Next, collecting VLSD may increase the likelihood of observing spurious correlations. For example, consider 100 individuals each of whose brain consists of a network formed by p edges (region-to-region correlations). If p is very large, then it is very likely that a few irrelevant edges are *spuriously* [i] associated with an outcome. This may introduce an erroneous scientific discovery that brain regions associated with these edges are the neurobiological underpinnings of the outcome. Out-of-sample prediction (training one group of individuals and confirming the discovery, without further model fitting, on a group of completely independent subjects), cross-validation (*e.g.,* leave-*k*-subject-out cross-validation), and repeated sampling test (*e.g.,* bootstrap test and permutation test) may alleviate this issue.

Related to spurious correlation is incidental endogeneity, where some features are coincidentally correlated with the residual term in an analytical model. The validity of most statistical models rests on the assumption that the predictors (*i.e.*, features) are uncorrelated with the residual term. Using the same argument as spurious correlation, the high dimensionality of the data is likely to introduce [ii] coincidental association between some features and the residual term, thereby violating the vital model assumption. The treatment for incidental endogeneity, despite its difficulty, is currently an actively-pursued area in statistical science and econometrics (see (Fan and Liao, 2014) for solutions under minor assumptions).

[i] Another source of spurious correlation is due to confounding (or lurking) variable(s). For example, nucleus accumbens (NAc) plays a pivotal role in rewarding and reinforcing (*e.g.*, sports, drugs, and sex). Blood oxygen level dependent signal (BOLD) increases in voxels within the NAc when a reward (*e.g.*, delicious food and a beautiful image) is at present. The fMRI data obtained from those voxels can thus be used as biomarkers for studying happiness. The neighboring voxels, despite functionally irrelevant, also show activation due to relatively poor resolution of fMRI. The predictability of neighboring voxels on happiness is thus in fact confounded by the voxels in the NAc.

[ii] Similar to spurious correlation, endogeneity can also be caused by (uncontrolled) confounding variables, or by simultaneity (*i.e.*, a looped causal effect between the features and the outcome. For example, taking addictive drug activates NAc; meanwhile activation in NAc encourages more drug consumption).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

Collecting high-dimensional features introduces extensive computational challenges as well as interpretation difficulties. (a) Many high-dimensional problems are computationally intractable (*i.e.,* it is solvable in theory but cannot be solved in practice). (b) High-dimensional data sometimes generate even higher-dimensional intermediate data. For example, calculating and inverting a correlation matrix is an integral step in some models (*e.g.*, the generalized estimating equations (Liang and Zeger, 1986)); but a correlation matrix based on features from *p* brain regions, where *p* is one million, is difficult to store and analyse.

High-dimensional data with large sample size (*n*) may yield an effect size that is extremely small yet with a significant *P-value*. For example, a correlation of 0.1 in a sample of 500 has a *P-value* around 0.025; a correlation of 0.01 in a sample of 100,000 has a *P-value* around 0.002. The *P-value* in these cases may have little inference value (Chén et al., 2020). Additionally, if the effect size is from a clinical trial or pathological studies, the finding, despite significant, may not be clinically or pathologically meaningful, and is oftentimes difficult to interpret and reproduce (Miller et al., 2016).

Finally, directly visualizing high-dimensional data is difficult, thereby hindering exploratory data analysis and *post hoc* interpretation. This, however, can be in part addressed by using projection pursuit (Huber, 1985). For example, one can use *principal component analysis* [iii] (PCA, which projects the high-dimensional data onto the lower-dimensional space while preserving the majority of total variance of the raw data), and *t-distributed stochastic neighbour embedding* (t-SNE, which projects the high-dimensional data onto two- or three- dimensional space) to reduce dimension of the data and to plot them onto space that is convenient to visualize (Van Der Maaten and Hinton, 2008; Pearson, 1901) (see **Figure** 5).

4.3 Statistical methods for dealing with high-dimensional data

A type of VLSD that requires special statistical treatments is the so-called high-dimensional data. A set of high-dimensional data is VLSD with a large number of (high-dimensional or ultra-high dimensional) features but not necessary a large sample size (at least the sample size grows no quicker than the number of features). Thanks to the advancement in data collection, storage, and computing, one can now begin to record neuroimaging data measured from a million voxels, and sequencing data generated

[iii] Another useful dimension reduction approach is *random projection*(Johnson and Lindenstrauss, 1984). It approximates the architecture of raw high-dimensional data in lower-dimensional space and is computationally more efficient than PCA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

from the whole genome consisting of billions of nucleotides. High-dimensional data analysis is useful for two reasons. First, by analyzing high-dimensional data, one can unveil hidden structures of subpopulations of the data; such information is not easily attainable from small-size data, as underlying structures may be treated as "outliers". Second, one can extract important common features across many subpopulations even when there are large individual variations (Fan et al., 2014).

As high-dimensional data begin to break the barriers between old and new data paradigms, they spawn new challenges. Let p denote the number of features and n be the number of subjects (or experimental units) on which the features are measured. The chief obstacles one faces with high-dimensional data are the so-called "large p, small n" problem (where p goes to infinity faster than n) and "large p, large n" problem (where p and n go to infinity at the same rate). Under either circumstance, the classic statistical theories (developed based upon "small p, large n" paradigm) collapse.

Classic large sample theory in statistics, which is typically developed based on the Law of Large Numbers and the Central Limit Theorem, has focused on the scenario where p is smaller than n. Under such setting, one can develop asymptotic properties of estimators of parameters, such as consistency of the estimators and their asymptotic (*e.g.*, Gaussian) distributions; with theoretical supports, one can then assign values and confidence intervals (or bands) to the estimated parameters to make statistical inference.

The challenge brought by high-dimensional data, however, can be alleviated if the following assumption holds, namely there exists a sparse representation of the features. In other words, there are only k $(k < n)$ features from all p features that are influential. Such an assumption, of course, is not purely hypothetical; rather, many empirical evidences have been discovered that suggest the existence of sparse data structures in biological systems. For example, it is known that there exist a small number of network hubs (a class of highly connected nodes or dictating features) and many poorly connected nodes in cells(Barabási and Oltvai, 2004), the brain(van den Heuvel and Sporns, 2013), and the genes(Leclerc, 2008). Under the sparse assumption, no matter how large the dimensionality is, one can apply feature selection techniques to reduce the number of features to that is smaller than n.

The two simplest classes of feature selection approaches are perhaps regularization (*e.g.*, SCAD, Dantzig selector, Lasso, elastic net, and adaptive Lasso) and stepwise feature selections (Chén, 2019). In the following, we introduce a few arrays of methodological frameworks proposed during recent years to deal with high-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

dimensional data, although, in our view, the fundamental principles of these novel methods do not depart from performing regularization and feature selection (or dimension reduction). In the following, we shall treat feature selection or dimension reduction ambiguously; but one should note that dimension reduction includes both feature selection and feature transformation (which converts high-dimensional variables to lower dimensional ones).

The first class of feature selection methods is suitable for linear models with ultra-high dimensional features. When dimension p grows fast with sample size n, regularized models may not pick up the correct features. The sure independence screening (SIS) (Fan and Lv, 2008) address this issue by first reducing the dimensionality from p to relatively large number $d < n$, and subsequently applying a common regularized method for feature selection. For nonlinear parametric models in which the mean function is defined explicitly as a nonlinear function of parameters, and various types of semiparametric or nonparametric approach, one can consider sparse additive models (Ravikumar et al., 2009) and Bayesian additive regression trees (Chipman et al., 2012). In essence, the former bears the form $y_i = \sum_{j=1}^{p} \beta_j f_j(x_{ij}) + \varepsilon_i$ and chooses (arbitrary) functions to minimize the residual sum of squares under constraints that both encourage sparsity and render the optimization problem convex. The latter treats the regression as a sum of trees $y_i = \sum_{j=1}^{p} f_j(x_i; T_j, \beta_j) + \varepsilon_i$ and uses each T_j to represent a tree structure involving a small number of the features. Finally, there are two roads leading to territories where regression models cannot lead: one that explores the classic statistic devices based on eigenanalysis of sample covariance matrices, and one that looks for a nonlinear manifold of low dimension (Johnstone and Titterington, 2009). The former class includes principal components analysis, canonical correlation analysis, multi-variate analysis of variance, and discriminant analysis, but newer developments such as direct thresholding (Bickel and Levina, 2008) and sparse covariance models (El Karoui, 2008) are increasingly useful to explore the sparsity nature of the covariance matrix when the covariance matrix is particularly large. The latter include ISOMAP (Tenenbaum et al., 2000), local linear embedding (Roweis and Saul, 2000), Laplacian eigenmaps (Belkin and Niyogi, 2003) and Hessian eigenmaps (Donoho and Grimes, 2003); the lower dimensional nonlinear manifold better represents the original data than lower dimensional linear subspace into which the projections of the original data have the largest variance (say from principal component analysis and multi-dimensional scaling).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

5 The statistical basis for correlation, association, explanation, and causation of the variation in the brain

Aristotle said in *Metaphysics* that "[w]e do not have knowledge of a thing until we have grasped its why, that is to say, its cause". We do not fully understand the variation in the brain and behaviour until we discover the causes. But what are the causes of brain variability? How does the variability of one brain area affect it of another? How do such neural causal connections give rise to perception, cognition, and behaviour?

5.1 Effective connectivity and predictive modelling

Chief to the pursuit of causation in the brain is the effective connectivity, "a circuit diagram that would replicate the observed timing relationships between the recorded neurons" (Aertsen and Preissl, 1991). Effective connectivity depends on a model of interactions or coupling (Friston, 2011). The idea of effective connectivity originates from the analysis of individual spike trains obtained from multiunit electrode recordings (Gerstein and Perkel, 1969, Gerstein et al 1989; Gochin et al 1991; Aertsen and Preissl 1991) but has since extended to study general neural systems. Compared to directionless functional connectivity, effective connectivity *mediates the influence that one neuronal system (*either at a synaptic or cortical level*) exerts on another and, therefore, discounts other influences* (Friston et al 1993).

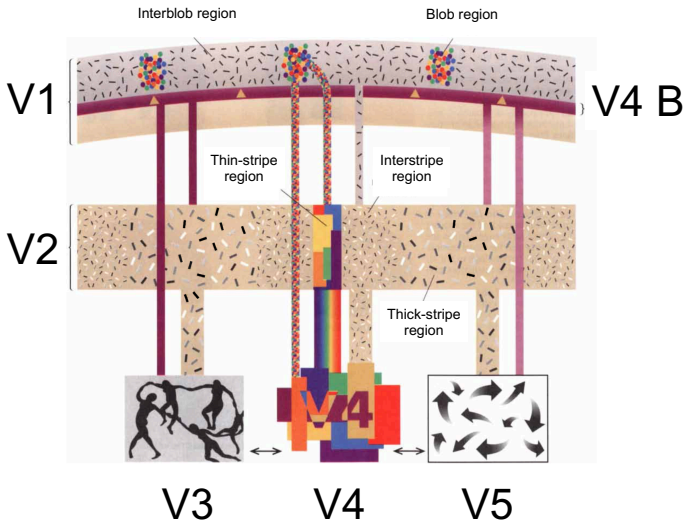

Figure 6. Four directed perceptual pathways within the visual cortex. Color is seen when wavelength-selective cells in the blob regions of V1 send signals to specialized area V4 and also to the thin stripes of V2, which connect with V4. Form in association with color depends on connections between the interblobs of V1, the interstripes of V2 and area V4. Cells in layer 4B of V1 send signals to specialized areas V3 and V5 directly and also through the thick stripes

 doi:10.20944/preprints202008.0428.v1

of V2; these connections give rise to the perception of motion and dynamic form. Redrawn from "The Visual Image in Mind and Brain" by Semir Zeki. Copyright Guilbert Gates and Jared Schneidman 1992.

The visual pathways in the cerebral cortex can be categorized into two major classes: The magnocellular (or M) pathway and the parvocellular (or P) pathway. The former is thought to be concerned with spatial relations and with “where” an object is; it consists of layer 4B of V1, the output from layer 4B to area V5, directly and through V2, and then the further output from area V5 to the parietal cortex. The latter is thought to be concerned with “what” an object is; it consists of layer 2 and 3 of V1, the output from these two layers to area V4, directly and through V2, and from V4 to the inferior temporal cortex (Zeki, 1993). More specifically, for the P pathway, to see color, wavelength-selective cells in the blob regions of V1 will first send signals to specialized area V4 and also to the thin stripes of V2, which connect with V4. To form the perception of motion and dynamic form, cells in layer 4B of VI send signals to specialized areas V3 and VS directly and also through the thick stripes of V2 (see **Figure 6**) (Zeki, 2010).

An enquiry into the causal couplings that underpin brain function and human behavior can be made in two steps. First, one charts a diagram formed by effective connectivities between brain areas. Subsequently, one associates each directed edge with human behavior or brain disease output to reduce the likelihood that the directed edge is spurious and to make further neurobiological enquires about their association

Yet it remains possible that some of the identified effective edges are mere numerical coincidence. A simple example can be found in Granger-analysis of the brain, where signals recorded from a brain region A that Granger-cause (causally antecede in the Granger-sense) signals recorded from anther region B does not prove that area A sends signals to area B in order to produce a particular motion or perception. A useful way to guard a method from yielding spurious causal claims on effective connectivity is to perform out-of-sample prediction modeling using directed edges. For example, by linking directed edges with brain disease status or behavior metric, one can select directed edges that are associated with the disease or behavior; by validating selected edges on novel subjects, one can further verify if the directed edges can be extrapolated.

Although the search for causation and their validation via predictive modeling are beginning to shed light upon potential directed neural communication underpinning perception, motion, and cognition, the causal neural basis of features remains unclear.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

In other words, if a selected feature is associated with or predictive of an outcome (even in novel subjects), is this feature causally linked to the outcome? The short answer is no. In the following, we shall discuss some perspectives regarding the statistical basis for association, explanation, prediction, and causation in neurobiological studies.

5.2 A comparison between correlation, association, explanation, and causation

Let us begin by defining what the problem of causal inquiry is regarding a biological organization in general and a neurobiological system in particular (Shmueli, 2010). Let $\boldsymbol{\xi}$ denote a class of neurobiological random variables (where we use the bold font to indicate that $\boldsymbol{\xi}$ can be multivariate) and let ω denote an outcome variable (for simplicity consider that ω is univariate). If the properties of $\boldsymbol{\xi}$ affect the expression of ω, then we can use a link function ϕ to denote the relationship between ξ and ω, namely $\omega = \phi(\boldsymbol{\xi})$. In words, the feature variable $\boldsymbol{\xi}$ cause the outcome variable ω by means of ϕ. The link function ϕ can be a complex mathematical map, edge(s) in a graphic model, or a (set of qualitative) statements.

As $\boldsymbol{\xi}$ and ω are random (biological) variables, one can only aim at uncovering the underlying (causal) relationship by analyzing the observed values of $\boldsymbol{\xi}$ and ω. Let's denote them as $\boldsymbol{X}$ and o, respectively. Since observed data contain noise and measurement error, the statistical model used to uncover the potential causal relationship can be written as $o = f(\boldsymbol{X}) + \varepsilon$, where ε denotes the residual term following a specific distribution (say Gaussian). Here, f specifies the model, which delineates the relationship between the observed $\boldsymbol{X}$ and o but not necessarily between $\boldsymbol{\xi}$ and ω. If $\boldsymbol{X}$ and o contain meaningful signals of $\boldsymbol{\xi}$ and ω and the noise and error are small, then one would hope that estimated f is a reasonable proxy of the unknown causal map ϕ.

Suppose that we find a model f that depicts the relationship between one or several features, say physical activities, and an outcome, say, having Parkinson's disease. We would certainly be pleased if those features are the *only* ones (among all candidate features) that have an effect on the outcome. We would be equally pleased if we can use the trained model f to make forecasts about future outcomes of a specific subject based on his/her past features or make forecasts about outcomes of new subjects based on subjects we have studied. In spite of advances, for example the discovery of GBA1 being present among 5% of Parkinson's patients (Stoker et al., 2018), definitive or direct causal relationships are in general difficult to identify and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

hard to verify (sometimes one has to wait until *post-mortem* for testing the causal links). Such difficulties, fortunately, can be partially addressed by performing associative or predictive analysis, which aim at unveiling features and relationship that, in a weaker sense, can explain the underlying biological system. Subsequently, one can test whether the discovered relationship can be strengthened into potential causation by performing further statistical analysis and evaluating *post hoc* empirical evidence.

Association *vs.* causation. Because the underlying causation ϕ is unknown, the estimated relationship $\hat{f}$ (or $\hat{f}(\boldsymbol{X}, o)$) can only suggest that there is association between $\boldsymbol{X}$ and o and can *at most* say that there is a potential causal relationship between feature variables $\boldsymbol{\xi}$ and outcome variable ω.

The reason is twofold. First, the observed features $\boldsymbol{X}$ and outcome o, due to noise and measurement error, may not accurately describe the properties of the underlying $\boldsymbol{\xi}$ and ω. Second, even if features $\boldsymbol{X}$ and outcome o are accurately describing the properties of random biological entities $\boldsymbol{\xi}$ and their outcome ω, as f is an arbitrary statistical model, there may exist another model g that better depicts the relationship between observed features $\boldsymbol{X}$ and outcome o. By *better depicts*, we mean that the model g (where $o = g(\boldsymbol{X}) + e$) may produce a smaller error e than ε (as in $o = f(\boldsymbol{X}) + \varepsilon$) or yield better out-of-sample prediction than f.

But the fact that there exists an alternative model g that is better than f in terms producing smaller an error or yield better a prediction does not conclude that g is more suitable than f to describe the causal link of ϕ. First, a model g that overfits the data will produce a smaller error e. Second, due to measurement error and noise, a model g that yields higher prediction power may be one that is better at predicting signal plus noise (rather than predicting signals). Finally, sometimes an alternative model g may suggest causation when there is not. Let's consider an extreme case when there is no underlying causation (say, between temporal signals obtained from two nodes ξ and ω); in other words, suppose ϕ is not causal. Due to (say, delayed) measurement error, there is a spurious temporal relationship between features $\boldsymbol{X}$ and outcome o which is captured by g. Therefore, one cannot say because g exists, causation exists.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

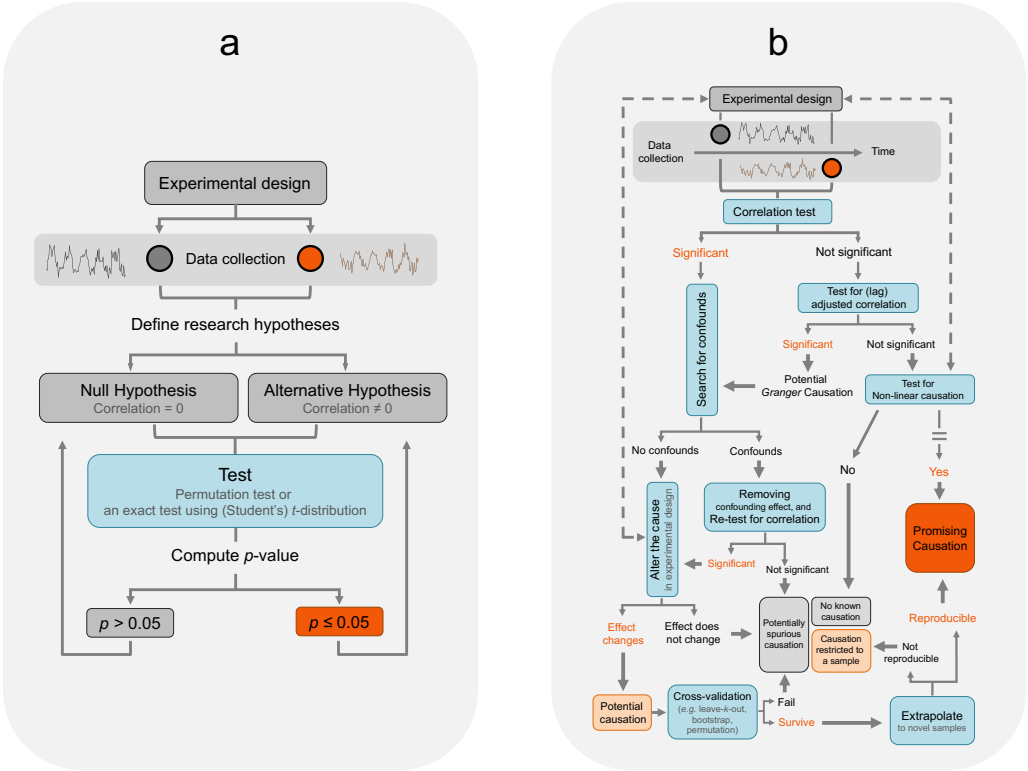

Figure 7. From association to potential causation. (a) A typical flowchart for conducting hypothesis-led testing on whether correlation between two random variables is significantly different from zero. A significant correlation, however, does not equate causation. Note that this framework forms the first part of the flowchart in figure (b). Figure **(b)** provides a more rigorous flowchart for making causal enquires. Note that the correlation test is used as an example; it can be replaced with other models or tests. It extends to cases where more than two variables are concerned. For the purpose of demonstration, we focus mainly on testing linear causation, and abbreviate the procedure for testing non-linear causation (which is marked with two parallel bars in the figure) - interested readers can refer to (Bai et al., 2010; Hiemstra and Jones, 1994). We do not claim nor advocate that this is the only procedure to raise correlation to causation; rather, it serves as an example to remove confounding effects, avoid over-fitting, and conduct reproducible research. We emphasize that a careful experimental design, appropriate data processing, and contextual scientific interpretation are equally important, but are not shown in the figure. The illustration demonstrates that even simple analysis needs additional caution when causal inference and reproducibility are concerned.

Because of these reasons, it seems suitable to claim association during statistical modeling of neural and behavioral data before one obtains further (*e.g.*, clinical or medical) confirmation. One can, however, consider more stringent a procedure to conduct testing aimed at making better causal claims in brain studies (see a pipeline in **Figure** 7). We restrict our recommendation to studies that conduct (null) statistical hypothesis tests. We do not claim nor advocate that this is the only way to perform statistical data analysis. Rather, the pipeline serves as an example where rigorous statistical thinking and analysis may reduce confounding effects, avoid over-fitting, and render reproducible research. We stress that experimental design,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

data processing, and scientific interpretation are equally important, but are not shown in the figure or discussed in detail, as they are not the main focus of the article. We take a correlation test between two random variables (*e.g.*, the edge strength between two brain regions over time) as an example. The flowchart in **Figure** 7 (b) can be extended or modified to suit other models or tests; it can also extend to cases involving multivariate variables.

Prediction *vs.* Causation. The difficulty in isolating causal relationships with regard to the brain, mind, body, behavior and pathology and a lack of knowledge of ϕ should not prevent us from looking for features that are able to explain a neurological or behavioral phenomenon. A beginning can be made by looking for features that are predictive of the outcome and considering them as superset of causal features. A causal feature, such as a gene mutation, is almost surely predictive of the outcome (to a certain extent) and can therefore explain a portion of the variance of the outcome. The reverse is not always true. For example, a brain area whose signals are predictive of an outcome does not mean the former is causally related to the latter.

Suppose that observations $\boldsymbol{X}$ and outcome o have been properly preprocessed and nothing additional can be done to make them better represent the signals of the observed values of $\boldsymbol{\xi}$ and ω. The value of predicative modelling can be explained, in part, by the bias-variance trade-off. Following the above notation, the expected predictive square error is (Friedman et al., 2001):

$$\text{Expected predictive square error } = E\left(o - \hat{f}(\boldsymbol{x})\right)^2$$

$$= \underbrace{E\left(o - f(\boldsymbol{x})\right)^2}_{\text{True variance}} + \underbrace{\left(E\left(\hat{f}(\boldsymbol{x})\right) - f(\boldsymbol{x})\right)^2}_{\text{Bias}} + \underbrace{E\left(\hat{f}(\boldsymbol{x}) - E\left(\hat{f}(\boldsymbol{x})\right)\right)^2}_{\text{Estimation variance}}$$

where the first term represents the variance even if the model is correctly specified and estimated, the second term denotes the bias, and the last term is the estimation variance.

The causal inquiry, in essence, is to minimize the bias between the estimated model ($\hat{f}(\boldsymbol{x})$) and the underlying theory about the causal relationship f. Meanwhile, predictive modeling aims at minimizing both bias and the estimation variance, even if at a cost of theoretical accuracy for improved empirical precision; in other words, one

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

would prefer a "wrong" (or less realistic) model that yields better predictions (Shmueli, 2010).

The usefulness of predictive modeling lies in that it reduces the likelihood of overfitting and *may* (our emphasis) help raise association to potential causation. This is done via out-of-sample prediction, where the estimated function $\hat{f}$ trained from a sample is used to predict outcomes in previously unseen samples. Thus, if the features and estimated function can be validated in novel subjects, it suggests that $\hat{f}$ may not be restricted to a training sample and is representing some general properties shared in a broad population. Further, if the relationship (indicated by $\hat{f}$) exists in a wide population, it hints that there exists a stronger sense of causal relationship between the features and the outcome.

Explanation *vs.* causation. Both predictive modeling and association analysis shed useful light upon the relationship between features and the outcome. Thus, until a definitive causal map is charted, one can rely on features showing predictive or associative merits to make sense of the observed data. For example, dexterity data are known to be associated with Parkinson's disease (PD). Trained in a predictive model, dexterity data can be used to predict PD in novel samples (Chén et al., 2020). Naturally, however, irregularity in dexterity data is not the cause for PD, for PD is a neurodegenerative disease; rather, irregular dexterity is an early effect of the underlying neurodegenerative disease. Nonetheless, building an associative or predictive model between dexterity and PD is still useful for patient identification and disease severity estimation. Further, it may offer insights about behavioral characteristics of the disease. For example, tremor, bradykinesia, rigidity, and postural instability are significantly associated with (but are not necessarily the causes for) the onset of PD, suggesting that motor functions are hindered in PD patients. Thus, observing that many PD patients have motor issues raises the question of the neural origin of the PD in the brain, for the motor disruption must be in one way or another linked to cortical abnormalities in the motor cortex of the brain.

Prediction *vs.* explanation. The aim of predicative modeling is to achieve accurate predictions via modeling techniques. It is useful when one is interested in extrapolating model representations learned from a sample into new observations. But predictive models are oftentimes built at the cost of their explanation capability. For example, regularization methods (*e.g.*, RIDGE, Lasso, and elastic-net regularizations) reduce

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

estimation variance (by shrinking some feature parameters towards zero) introduces bias into the model, thereby making the model less explainable. Similarly, ensemble methods such as bagging, boosting, and random forest improve overall prediction accuracy by averaging predictions from each model meanwhile making it difficult to explain the ensembles (Shmueli, 2010). Finally, neural networks may capture complicated associations hidden in the features space and between features and the outcome and yield accurate predictions, but the models may be neither able to accurately represent the underlying causal mechanism ϕ nor easy to interpret the model f itself. On the other hand, including multiple highly correlated features in a model or including an insignificant feature would reduce the prediction power of the model, but the correlated features, such as different types of motor symptoms in Parkinson's studies, or insignificant (in terms of predictability) features, such as age, gender, and smoking status in medical research, may shed meaningful scientific and medical lights upon the research question; sometimes a research protocols may ask to include them mandatorily (such as to include smoking history in a cancer study).

Taken together, it is important to clarify what the research goals are during study design, model development, and data analysis and to balance predictability and explainability by carefully evaluating the bias-variance trade-off (Douglas, 2009) (Yarkoni and Westfall, 2017).

Statistical causal inference methods. Various statistical devices are useful to evaluate if association can be raised to (promising) causation. They can be classified into five categories: randomization, discovery validation and reproduction, causal reasoning, causal alternation, and instrumental variables (IV). In the following, we highlight a few important statistical devices with a brief explanation or an example.

The first category is randomization. Randomized experiments are used to study, for example, whether a drug is effective. Critically, researchers have to compare the result after an individual has taken the drug to it had the same individual not taken the drug. Unless the condition is relatively stable and one does a crossover study, only one of the two is observable; even when a crossover study is possible, there is likely a carry-over effect or an order effect. With randomization (when randomization is not

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

available, see Propensity Score Matching (PSM)[iv]), individuals in a treatment group (where each individual receives a medication), and a placebo group (where each individual receives a placebo) are on average similar (Ott and Longnecker, 2015). Randomly assigning children to attend public *versus* private schools to study educational effect, however, is impossible[v]. To solve this, potential outcomes frameworks or Neyman-Rubin causal model considers a non-random assignment mechanism to make certain groups "comparable" when there are unobserved outcomes (Neyman, 1935) (Rubin, 1974).

The second category is discovery validation and reproduction. For example, we can repeat the same experiment to verify if the result can be replicated or reproduced (Vaux et al., 2012). Another approach is to conduct out-of-sample testing. For example, we first perform model fitting on one sample, and then test the fitted model on a completely different sample, without further modelling (Woo et al., 2017).

The third category is (probabilistic) causal reasoning[vi]. More concretely, it studies whether an event (*e.g.*, high alcohol consumption) causes another event (*e.g.*, Korsakoff's syndrome), by comparing the probability of having Korsakoff's syndrome

[iv] There are times when randomization becomes impossible. For example, it is unethical to assign a group of 45-year-old heathy subjects to take a new Levodopa-based drug to investigate whether the drug reduces one's PD symptoms at 50. Additionally, it is likely that there is another source, say, the socioeconomic status (which may be related to the affordability of new drugs) or genetics (if there is a family history of PD, one may be more willing to take the drug), that is both associated with taking the drug and developing PD at 50. Similarly, it would be difficult to estimate the effect of taking the drug on reducing PD symptoms by comparing the PD symptoms of an individual at 50 who had taken the drug with his or her PD symptoms at 50 had he or she not taken the drug. To solve these issues, Propensity Score Matching (PSM) estimates the treatment effect by comparing the outcomes of the subjects under treatment (*e.g.*, taking the drug) with those of a different set of "matched" subjects without treatment (*e.g.,* having not taken the drug) (Caliendo and Kopeinig, 2008; Dehejia and Wahba, 1999, 2002; Rosenbaum and Rubin, 1983). More concretely, one could first compute the propensity score of A taking the drug based on his or her gender, economic, social, genetic, and demographic backgrounds, and choose an individual from a group of 50-year-old who had not taken the drug but has a propensity score (of taking the drug during his or her younger years) closest to A's. Then we can compare the PD symptoms between these two individuals and estimate the effect of taking the drug on reducing PD symptoms at age 50.

[v] In rare cases, one could offer parents who do not have the funds for the more expensive schools a randomisation: either the child stays where they are, or they get a better spot. This is the principle for waiting list control studies.

[vi] Its modern development is based on Reichenbach's macrostatistical theory (Reichenbach, 1991) and Suppes' probabilist theory (Suppes, 1970). Interested readers can refer to the books edited by Sosa and Tooley for a thorough review (Sosa, 1975) (Sosa and Tooley, 1993).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

given one drinks with the probability of having the syndrome given one does not drink, after controlling all other factors (Pearl, 2011); also see Simpson's paradox (Gardner, 1976). Related to probabilistic causal reasoning are graphical models (*e.g.*, directed acyclic graphs of brain networks), which use vertices to represent events (*e.g.*, activation from two neurons A and B), (directed) edges (*e.g.*, the arrow in 'A → B') to represent causal relationship between each pair of vertices, and edge width (which can be quantified using the probability of A causes B) to indicate the strength of causation (Pearl, 1993)(Pearl et al., 1999)(Hinton, 2005).

The fourth category is cause alteration. It examines if modifying a hypothesised cause would result in a change of the hypothesised effect. For example, using transcranial magnetic stimulation (TMS), one can use a magnetic field generator (or coil) to produce electric current, which modifies the magnetic field of neurons in a small surface region of the brain (Romei et al., 2012)(Lipton and Pearlman, 2010). When confounders are controlled, comparing the behaviour (or the symptom) when these neurons are "on" with the behaviour (or the symptom) when they are "off", one can assess if these neurons are responsible for the behavioural (or symptom) change.

The fifth class is the method of instrumental variables (IV)[vii]. The logic of IV is as follows. Suppose we are interested in investigating if alcohol consumption causes deterioration of overall health. Associative analysis and randomized control study here may not be suitable, as we present below. Large correlation between alcohol consumption and overall health does not imply drinking causes deterioration of overall health. On the contrary, deteriorating healthy may make one start drinking or increase alcohol consumption. Furthermore, drinking may first affect an intermediating variable, such as mental health, which then affects overall health. Mental health, in return, could also cause changes in alcohol consumption. Randomized control experiment on alcohol consumption is unethical and perhaps illegal. When neither direct association analysis nor randomized control experiment is suitable, the method of IV may play a big role in causal inquires. Using an IV, for example, tax rate on alcohol, one can study the causal effect of drinking on overall health, because tax rate on alcohol is correlated with alcohol consumption, and can only be correlated with overall health through drinking (Epstein, 1989) (Stock and Trebbi, 2003).

[vii] A suitable instrumental variable (IV) is one that is correlated with an endogenous explanatory variable, such as alcohol consumption, but is not correlated with the error term (for example, in a regression). An endogenous variable is a covariate that is correlated with the error term.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

Epilogue

"Anyone who writes about 'Darwin's theory of evolution' in singular, without segregating the theories of gradual evolution, common descent, speciation, and the mechanism of natural selection, will be quite unable to discuss the subject competently" (Mayr, 1982). Anyone in statistical science who performs "analysis of variance" in singular without concerning the variability from temporal, spatial, individual, and group components, and their causes and consequences, is likely not to gain a full picture of the topic.

When laying the foundations for the subject of statistics and introducing the analysis of variance (Fisher, 1918, 1925), Fisher seemed to have not distinguished the different types of variations, perhaps deeming this too trivial an exercise. Today, standing on his shoulders, we argue that it is important, and beneficial, to differentiate specific classes of variability. By isolating variance components and investigating their causes and effects, one may begin to paint a better picture of the organization of the data and gain new insights about the biological underpinning of variation. The enquiry of each type of variance component, in its own right, forms a different lineage of research problems and establishes important building blocks of statistical, biological, and data science.

In this paper, reflecting on Fisher's philosophy, we presented a few new, modified, and integrated perspectives of variance. We used brain and behavioral studies as a gateway, because we know slightly more about them, but we hope that, through the lenses of neural and behavioral data, one can peer into statistical analysis of variance in other fields and datasets.

We began our journey by defining and distinguishing temporal, spatial, within-group, and between-group variations. We then argued that both classic statistics methods, such as ANOVA and ANCOVA, and advanced statistical models, such as random-effect models and transition (or Markov) models, could both be viewed as variance-decomposition methods. Specifically, they aimed at segregating the total variation of the outcome into components attributed to internal or external factors that offered a descriptive statement, an associative or predictive explanation, or a causal claim either about the outcome or about the relationships between those factors and the outcome. To better trace the potential causes of neural and behavioural variability, we suggested that a distinction needed to be made between innate variability and acquired variability. The former was likely dictated by genetic factors; the latter was likely due to environmental factors or a combination of environmental and genetic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0428.v1

factors. The modelling of variability of phenotypes could then be made by connecting the innate variability and acquired variability via Bayesian updating. Next, we suggested statistical devices to analyse very large-scale data (VLSD) and to reduce and visualize high-dimensional data. To uncover constant knowledge from very large-scale brain data, we proposed the Neural Law of Large Numbers and present empirical evidence for it. Finally, we provided a discussion regarding the statistical basis for association, explanation, prediction, and causation and suggested a strategy that may be useful to check if association-based findings were able to be raised to causal discoveries.

Despite advances, our understanding of the variability in data and biological basis for variation is imperfect, owing in part to our ignorance in biology and in part to a poverty of statistical tools. We shall, however, not be deterred from making scientific inquiries using existing data while collecting new. Nor should we be discouraged from improving old methodological frameworks and inventing new, as Maxwell wrote, “[t]horoughly conscious ignorance ... is a prelude to every real advance in knowledge”.

To sum up, the introduction of the concept of variance by R.A. Fisher a century ago demonstrated that variation among phenotypic traits could be due to Mendelian inheritance (Fisher, 1918). During the past century, the concept has advanced time and time again scientific and methodological studies (Gelman, 2005), injecting new insights into epidemiology, population genetics, experimental design, and statistical inference. Moving forward, careful studies of variabilities will allow us to continue improving our understanding of the genetic, environmental, and neural bases of variability in structure and functioning of biological organizations, and the underpinnings of variability in behaviours, perception, cooperation, and, perhaps above all, our varying selves.

Truly, to understand the science of data, one needs creative statistical thinking; by incorporating knowledge learned from data, one may begin to develop a better statistical enterprise.

Acknowledgements

Funding: This research received no external funding. Author contributions: O.Y.C. wrote the paper with comments from all other authors. The authors thank Tianchen Qian, Paul Matthews, and the Editor for helpful comments on an earlier version of this paper. Competing interests: Authors declare no competing interests. Data and materials availability: All data is available in the main text.

References

Aertsen, A., and Preissl, H. (1991). Dynamics of activity and connectivity in physiological neuronal networks. *Nonlinear Dyn. Neuronal Networks* **9**, 1303-1350.

Bai, Z., Wong, W.K., and Zhang, B. (2010). Multivariate linear and nonlinear causality tests. *Math. Comput. Simul.* **81**, 5–17.

Baker, A. (2004). Simplicity. In *The Stanford Encyclopedia of Philosophy (Spring 2017 Edition)*, edited by Edward N. Zalta.

Bar-Haim, Y., Ziv, T., Lamy, D., and Hodes, R.M. (2006). Nature and nurture in own-race face processing. *Psychol. Sci.* **17**, 159–163.

Barabási, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113.

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373–1396.

Bernoulli, J. (1713). Ars conjectandi: usum & applicationem praecedentis doctrinae in civilibus, In *Moralibus et Oeconomicis*, translated in to English by Oscar Sheynin.

Bertoli, S., Leone, A., and Battezzati, A. (2015). Human bisphenol A exposure and the "diabesity phenotype". *Dose-Response*, pp. 1–12. doi: 10.1177/1559325815599173.

Bickel, P.J., and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Stat.* **36**, 2577–2604.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., and Hyde, J.S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34**, 537–541.

Brigandt, I. (2005). The instinct concept of the early Konrad Lorenz. *J. Hist. Biol.* **38**, 571–608.

Broca, P. (1861). Remarques sur le siége de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin et Memoires de la Societe anatomique de Paris*. **6**, 330–357.

Brodmann, K. (1909). Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues (Leipzig: Barth).

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198.

Burciu, R.G., Ofori, E., Archer, D.B., Wu, S.S., Pasternak, O., McFarland, N.R., Okun, M.S., and Vaillancourt, D.E. (2017). Progression marker of Parkinson's disease: A 4-year multi-site imaging study. *Brain* **140**, 2183–2192.

Burkhardt, R.W. (2005). Patterns of Behavior: Konrad Lorenz, Niko Tinbergen and the Founding of Ethology (Chicago, USA: University of Chicago Press).

Caliendo, M., and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* **22**, 31–72.

Campbell, A.W. (1905). Histological studies on the localisation of cerebral function. *Journal of Mental Science* **50**, 651–662.

Cao, H., Chén, O.Y., Chung, Y., Forsyth, J.K., McEwen, S.C., Gee, D.G., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., et al. (2018). Cerebello-thalamo-cortical hyperconnectivity as a state-independent functional neural signature for psychosis prediction and characterization. *Nat. Commun.* **9**, 1–9.

Casey, B.J., Giedd, J.N., and Thomas, K.M. (2000). Structural and functional brain development and its relation to cognitive development. *Biol. Psychol.* **54**, 241–257.

Chén, O.Y, Lipsmeier, F., Phan, H., Prince, J., Taylor, K., Gossens, C., Lindemann, M., and De Vos, M (2020). Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. *IEEE Trans Biomed Eng.* doi: 10.1109/TBME.2020.2988942.

Chén, O.Y. (2019). The roles of statistics in human neuroscience. *Brain Sci.* **9**, 194.

Chén, O.Y., Crainiceanu, C., Ogburn, E.L., Caffo, B.S., Wager, T.D., and Lindquist, M.A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19**, 121–136.

Chén, O.Y., Cao, H., Reinen, J.M., Qian, T., Gou, J., Phan, H., De Vos, M., and Cannon, T.D. (2019). Resting-state brain information flow predicts cognitive flexibility in humans. *Sci. Rep.* **9**, 1–16.

Chén, O.Y., Saraiva, R.G., Nagels, G., Phan, H., Schwantje, T., Cao, H., Gou, J., Reinen, J.M., Xiong, B., and Vos, M. de (2020). Thou shalt not reject the *P*-value. **ArXiv *2002.07270***.

Chipman, H.A., George, E.I., and McCulloch, R.E. (2012). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298.

Christ, B.U., Combrinck, M.I., and Thomas, K.G.F. (2018). Both reaction time and accuracy measures of intraindividual variability predict cognitive performance in Alzheimer's disease. *Front. Hum. Neurosci.* **12**, 124.

Corbetta, M., and Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **3**, 201–215.

Croxson, P.L., Forkel, S.J., Cerliani, L., and Thiebaut De Schotten, M. (2018). Structural Variability Across the Primate Brain: A Cross-Species Comparison. *Cereb. Cortex.* **28**, 3829–3841.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. (London, UK: John Murray).

DeCasien, A.R., Sherwood, C.C., Schapiro, S.J., and Higham, J.P. (2020). Greater variability in chimpanzee (Pan troglodytes) brain structure among males. *Proc. R. Soc. B* **287**, p.20192858.

Dehejia, R.H., and Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J. Am. Stat. Assoc.* **94**, 1053–1062.

Dehejia, R.H., and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* **84**, 151–161.

Desimone, R., Albright, T.D., Gross, C.G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* **4**, 2051–2062.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2013). Analysis of Longitudinal Data (Oxford: Oxford University Press).

Donoho, D.L., and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci.* **100**, 5591–5596.

Douglas, H.E. (2009). Reintroducing prediction to explanation. *Philos. Sci.* **76**, 444–463.

Epstein, R.J. (1989). The fall of ols in structural estimation. *Oxf. Econ. Pap.* **41**, 94–107.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn human connectome project: an overview.

Neuroimage **80**, 62–79.

Fan, J., and Liao, Y. (2014). Endogeneity in high dimensions. *Ann. Stat.* **42**, 872–917.

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B.* **70**, 849–911.

Fan, J., Han, F., and Liu, H. (2014). Challenges of Big Data analysis. *Natl. Sci. Rev.* **1**, 293–314.

De Felice, S., and Holland, C.A. (2018). Intra-individual variability across fluid cognition can reveal qualitatively different cognitive styles of the aging brain. *Front. Psychol.* **9**, 1973.

Fisher, R.A. (1918). The Correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* **52**, 399–433.

Fisher, R.A. (1925). Statistical Methods for Research Workers. (Edinburgh: Oliver and Boyd).

Friedman, J., Hastie, T., and Tibshirani, R. (2001). The Elements of Statistical Learning (New York: Springer series in statistics).

Friston, K. (2012). The history of the future of the Bayesian brain. *Neuroimage* **62**, 1230–1233.

Friston, K.J. (2011). Functional and effective connectivity: a review. *Brain Connect.* **1**, 13–36.

Fritsch, G., and Hitzig, E. (1870). Uber die elektrische Erregbarkeit des Grosshirns. *Arch, anat. Physiol. Wiss. Med.* **37**, 300–332.

Gabrieli, J.D.E., Ghosh, S.S., and Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* **85**, 11–26.

Gardner, M. (1976). Mathematical Games: on the fabric of inductive logic, and some probability paradoxes. *Sci. Am.* **234**, 119–125.

Garrett, D.D., Lindenberger, U., Hoge, R.D., and Gauthier, C.J. (2017). Age differences in brain signal variability are robust to multiple vascular controls. *Sci. Rep.* **7**, 1–13.

Ge, C., Ye, J., Weber, C., Sun, W., Zhang, H., Zhou, Y., Cai, C., Qian, G., and Capel, B. (2018). The histone demethylase KDM6B regulates temperature-dependent sex determination in a turtle species. *Science* **360**, 645–648.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *Ann. Stat.* **33**, 1–53.

Gogtay, N., Giedd, J.N., Lusk, L., Hayashi, K.M., Greenstein, D., Vaituzis, A.C., Nugent, T.F., Herman, D.H., Clasen, L.S., Toga, A.W., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proc. Natl. Acad. Sci.* **101**, 8174–8179.

Goldman-Rakic, P.S. (1988). Topography of cognition: parallel distributed networks in primate association cortex. *Annu. Rev. Neurosci.* **11**, 137–156.

Gordon, E.M., Laumann, T.O., Adeyemo, B., and Petersen, S.E. (2017). Individual Variability of the System-Level Organization of the Human Brain. *Cereb. Cortex.* **27**, 386–399.

Griffiths, P. (2020). The Distinction between innate and acquired characteristics. In *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), edited by Edward N. Zalta.

Griffiths, P.E. (2004). Instinct in the ’50s: the British reception of Konrad Lorenz’s theory of instinctive behavior. *Biol. Philos.* **19**, 609-631.

Griffiths, L.T., Kemp, C., and Tenenbaum, B. (2008). Bayesian models of cognition. In *Cambridge Handbook of Computational Cognitive Modeling*, edited by R. Sun (Cambridge, UK: Cambridge University Press), pp. 59–100.

Halliday, D.W.R., Mulligan, B.P., Garrett, D.D., Schmidt, S., Hundza, S.R., Garcia-Barrera, M.A., Stawski, R.S., and MacDonald, S.W.S. (2017). Mean and variability in functional brain activations differentially predict executive function in older adults: an investigation employing functional near-infrared spectroscopy. *Neurophotonics* **5**, p.011013.

van den Heuvel, M.P., and Sporns, O. (2013). Network hubs in the human brain. *Trends Cogn. Sci.* **17**, 683–696.

Hiemstra, C., and Jones, J.D. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. *J. Finance* **49**, 1639–1664.

Hinton, G. (2005). What kind of a graphical model is the brain? *Proc. Intl. Jt. Conf. Artif. Intell.* **5**, 1765–1775.

Hook, E.W., and Marra, C.M. (1992). Acquired syphilis in adults. *N. Engl. J. Med.* **326**, 1060–1069.

Hubel, D.H., and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J.Physiol.* **195**, 215–243.

Huber, P.J. (1985). Projection Pursuit. *Ann. Stat.* **13**, 435–475.

Hunte, W., Myers, R.A., and Doyle, R.W. (1985). Bayesian mating decisions in an amphipod, Gammarus lawrencianus Bousfield. *Anim. Behav.* **33**, 366–372.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Hakonarson, H., Gur, R.E., Gur, R.C., and Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proc. Natl. Acad. Sci.* **111**, 823–828.

Jefferys, W.H., and Berger, Ja.O. (1992). Ockham's razor and Bayesian analysis. *Am. Sci.* **80**, 64–72.

Johnson, W.B., and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* **26**, 189–206.

Johnstone, I.M., and Titterington, D.M. (2009). Statistical challenges of high-dimensional data. *Phil. Trans. R. Soc. A.* **367**, 4237–4253

Kant, I. (1787). Kritik der reinen Vernunft, 2nd edition, translated as Critique of pure reason by W.S. Pluhar (Indianapolis: Hackett).

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Stat.* **36**, 2717–2756.

Knill, D.C., and Richards, W. (edited) (1996). Perception as Bayesian Inference (Cambridge University Press).

Koban, L., Jepma, M., López-Solà, M., and Wager, T.D. (2019). Different brain networks mediate the effects of social and conditioned expectations on pain. *Nat. Commun.* **10**, 1–13.

Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247.

Kricker, A., Armstrong, B.K., and English, D.R. (1994). Sun exposure and non-melanocytic skin cancer. *Cancer Causes Control.* **5**, 367–392.

Leclerc, R.D. (2008). Survival of the sparsest: Robust gene networks are parsimonious. *Mol. Syst. Biol.* **4**. 213.

Liang, K.-Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lipton, R.B., and Pearlman, S.H. (2010). Transcranial Magnetic Simulation in the Treatment of Migraine. *Neurotherapeutics* **7**, 204–212.

Lorenz, K.Z. (1957). The nature of instinct. In *Instinctive Behavior: The Development of a Modern Concept*, edited by C.H. Schiller (New York, USA: International Universities Press).

Lorenz, K.Z., and Tinbergen, N. (1957). Taxis and instinct: taxis and instinctive action in the eggretrieving behavior of the graylag goose. In *Instinctive Behavior: The Development of a Modern Concept*, edited by C.H. Schiller (New York, USA: International Universities Press).

Luttbeg, B. (1999). Reproductive decision-making by female peacock wrasses: flexible versus fixed behavioral rules in variable environments. *Behav. Ecol.* **10**, 666–674.

Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605.

Mayr, E. (1982). The Growth of Biological Thought: Diversity, Evolution, and Inheritance (Cambridge, MA, USA: Belknap Press).

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536.

Neyman, J. (1935). Statistical problems in agricultural experimentation. *J. R. Stat. Soc.* **2**, 107–180.

Ott, R.L., and Longnecker, M.T. (2015). An Introduction to Statistical Methods and Data Analysis (Toronto, Canada: Nelson Education).

Pearl, J. (1993). Graphical Models, Causality and Intervention. https://escholarship.org/uc/item/8d93w51g

Pearl, J. (2011). Causality: Models, Reasoning, and Inference (Cambridge, UK: Cambridge University Press).

Pearl, J., Robins, J.M., and Greenland, S. (1999). Confounding and collapsibility in causal inference. *Stat. Sci.* **14**, 29–46.

Pears, D.F. (1953). Incompatibilities of Colours (Oxford, UK: Blackwell).

Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space*. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**, 559–572.

Piper, R. (2007). Extraordinary Animals : An Encyclopedia of Curious and Unusual Animals (London, UK: Greenwood Press).

Poisson, S.D. (1837). Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés (Paris, France: Bachelier).

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., et al. (2011). Functional network organization of the human brain. *Neuron* **72**, 665–678.

Quinn, A.E., Georges, A., Sarre, S.D., Guarino, F., Ezaz, T., and Marshall Graves, J.A. (1983). Temperature sex reversal implies sex gene dosage in a reptile. *Science*

316, 411.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B.* **71**, 1009–1030.

Reichenbach, H. (1991). The Direction of Time (Berkeley, USA: University of California Press).

Reinen, J.M., Chén, O.Y., Hutchison, R.M., Yeo, B.T.T., Anderson, K.M., Sabuncu, M.R., Öngür, D., Roffman, J.L., Smoller, J.W., Baker, J.T., et al. (2018). The human cortex possesses a reconfigurable dynamic network architecture that is disrupted in psychosis. *Nat. Commun.* **9**, 1–15.

Rogachov, A., Cheng, J.C., Erpelding, N., Hemington, K.S., Crawley, A.P., and Davis, K.D. (2016). Regional brain signal variability: A novel indicator of pain sensitivity and coping. *Pain* **157**, 2483–2492.

Romei, V., Thut, G., Mok, R.M., Schyns, P.G., and Driver, J. (2012). Causal implication by rhythmic transcranial magnetic stimulation of alpha frequency in feature-based local vs. global attention. *Eur. J. Neurosci.* **35**, 968–974.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

Seghier, M.L., and Price, C.J. (2018). Interpreting and Utilising Intersubject Variability in Brain Function. *Trends Cogn. Sci.* **22**, 517–530.

Shmueli, G. (2010). To explain or to predict? *Stat. Sci.* **25**, 289–310.

Smith, S., Duff, E., Groves, A., Nichols, T.E., Jbabdi, S., Westlye, L.T., Tamnes, C.K., Engvig, A., Walhovd, K.B., Fjell, A.M., et al. (2019). Structural Variability in the Human Brain Reflects Fine-Grained Functional Architecture at the Population Level. *J. Neurosci.* **39**, 6136–6149.

Sober, E. (1990). Explanation in biology: let's razor Ockham's razor. *Royal Institute of Philosophy Supplement* **27**, 73–93.

Sohn, H., Narain, D., Meirhaeghe, N., and Jazayeri, M. (2019). Bayesian computation through cortical latent dynamics. *Neuron* **103**, 934–947.

Sosa, E. (edited) (1975). Causation and conditionals (Oxford, UK: Oxford University Press).

Sosa, E., and Tooley, M. (edited) (1993). Causation (Oxford, UK: Oxford University

Press).

Stock, J.H., and Trebbi, F. (2003). Retrospectives: who invented instrumental variable regression? *J. Econ. Perspect.* **17**, 177–194.

Stoker, T., Torsney, K., and Barker, R. (2018). Pathological mechanisms and clinical aspects of GBA1 mutation-associated Parkinson's disease. In *Parkinson's Disease: Pathogenesis and Clinical Aspects*, Edited by B.S. Thomas, and J.C. Greenland (Brisbane, Australia: Codon Publications).

Suppes, P. (1970). A Probabilistic Theory of Causality (Amsterdam, Netherlands: North-Holland Publishing Company).

Tenenbaum, J.B., De Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Tinbergen, N. (1942). An Objectivist Study of the Innate Behaviour of Animals (Leiden, Netherlands: E.J. Brill).

Tinbergen, N. (1951). The Study of Instinct (Oxford, UK: Oxford University Press).

Toro, R., and Burnod, Y. (2005). A morphogenetic model for the development of cortical convolutions. *Cereb. Cortex.* **15**, 1900–1913.

Valone, T.J. (1992). Information for patch assessment: A field investigation with black-chinned hummingbirds. *Behav. Ecol.* **3**, 211–222.

Valone, T.J., and Brown, J.S. (1989). Measuring patch assessment abilities of desert granivores. *Ecology* **70**, 1800–1810.

Valone, T.J., and Giraldeau, L.A. (1993). Patch estimation by group foragers: What information is used? *Anim. Behav.* **45**, 721–728.

Vaux, D.L., Fidler, F., and Cumming, G. (2012). Replicates and repeats-what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Rep.* **13**, 291–296.

Waddington, C.H. (2014). The strategy of the genes: A discussion of some aspects of theoretical biology (Abingdon, UK: Routledge).

Wernicke, C. (1874). Der aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer Basis (Breslau, Cohn & Weigert).

Woo, C.W., Chang, L.J., Lindquist, M.A., and Wager, T.D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377.

Wu, C.F.J., and Hamada, M. (2000). Experiments: Planning, Analysis, and Parameter Design Optimization (New Jersey, USA: John Wiley Sons).

Yankner, B.A., Lu, T., and Loerch, P. (2008). The aging brain. *Annu. Rev. Pathol. Mech. Dis*. **3**, 41–66.

 doi:10.20944/preprints202008.0428.v1

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122.

Zeki, S. (1993). A Vision of the Brain (Oxford: Blackwell Scientific).

Zeki, S. (2010). The Visual Image in Mind and Brain. *Sci. Am.* **267**, 68–77.

Zeki, S., and Chén, O.Y. (2019). The Bayesian-Laplacian brain. *Eur. J. Neurosci.* **51**, 1441–62.

Zeki, S., Watson, J.D., Lueck, C.J., Friston, K.J., Kennard, C., and Frackowiak, R.S. (1991). A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.* **11**, 641–649.

Zhang, D., and Shen, D. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS ONE* **7**, e33182.