Are mental properties supervenient on brain properties?

Joshua T. Vogelstein[1*], R. Jacob Vogelstein[2], Carey E. Priebe[1]
[1]Department of Applied Mathematics & Statistics,
Johns Hopkins University, Baltimore, MD, 21218,
[2]National Security Technology Department,
Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723

The"mind-brain supervenience" conjecture suggests that all mental properties are derived from the physical properties of the brain. To address the question of whether the mind supervenes on the brain, we frame a supervenience hypothesis in rigorous statistical terms. Specifically, we propose a modified version of supervenience (called ε-supervenience) that is amenable to experimental investigation and statistical analysis. To illustrate this approach, we perform a thought experiment that illustrates how the probabilistic theory of pattern recognition can be used to make a *one*-sided determination of ε-supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the brain-graph or connectome). ε-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given positive misclassification rate, regardless of the relationship between the two. This may provide motivation for cross-disciplinary research between neuroscientists and statisticians.

Questions and assumptions about mind-brain supervenience go back at least as far as Plato's dialogues in circa 400 BCE [1]. While there are many different notions of supervenience, we find Davidson's canonical description particularly illustrative [2]:

> [mind-brain] supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect.

Colloquially, supervenience means "there cannot be a mind-difference without a physical-difference." This philosophical conjecture has potentially widespread implications. For example, neural network theory and artificial intelligence often implicitly assume a local version mind-brain supervenience [3, 4]. Cognitive neuroscience similarly seems to operate under such assumptions [5]. Philosophers continue to debate and refine notions of supervenience [6]. Yet, to date, relatively scant attention has been paid to what might be empirically learned about supervenience.

In this work we attempt to bridge the gap between philosophical conjecture and empirical investigations by casting supervenience in a probabilistic framework amenable to hypothesis testing. We then use the probabilistic theory of pattern recognition to determine the limits of what one can and cannot learn about supervenience through data analysis. The implications of this work are varied. It provides a probabilistic framework for converting philosophical conjectures into statistical hypotheses that are amenable to experimental investigation, which allows the philosopher to gain empirical support for her rational arguments. This leads to the construction of the first explicit proof (to our knowledge) of a universally consistent classifier on graphs, and the first demonstration of the tractability of answering supervenience questions. Supervenience therefore seems to perhaps be a useful but under-utilized concept for neuroscientific investigations. This work should provide further motivation for cross-disciplinary efforts across three fields—philosophy, statistics, and neuroscience—with shared goals but mostly disjoint jargon and methods of analysis.

Results

Statistical supervenience: a definition

Let $\mathcal{M} = \{m_1, m_2, \ldots\}$ be the space of all possible minds and let $\mathcal{B} = \{b_1, b_2, \ldots\}$ be the set of all possible brains. $\mathcal{M}$ includes a mind for each possible collection of thoughts, memories, beliefs, etc. $\mathcal{B}$ includes a brain for each possible position and momentum of all subatomic particles within the skull. Given these definitions, Davidson's conjecture may be concisely and formally stated thusly: $m \neq m' \implies b \neq b'$, where $(m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$ are mind-brain pairs. This mind-brain supervenience relation does not imply an injective relation, a causal relation, or an identity relation (see Appendix 1 for more details and some examples). To facilitate both statistical analysis and empirical investigation, we convert this local supervenience relation from a logical to a probabilistic relation.

Let F_{MB} indicate a joint distribution of minds and brains. Statistical supervenience can then be defined as follows:

Definition 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $\mathbb{P}[m \neq m' | b = b'] = 0$, or equivalently $\mathbb{P}[m = m' | b = b'] = 1$.*

Statistical supervenience is therefore a probabilistic relation on sets which could be considered a generalization of correlation (see Appendix 1 for details).

Statistical supervenience is equivalent to perfect classification accuracy

If minds statistically supervene on brains, then if two minds differ, there must be some brain-based difference to account for the mental difference. This means that there must exist a deterministic function g^* mapping each brain to its supervening mind. One could therefore, in principle, know this function. When the space of all possible minds is finite—that is, $|\mathcal{M}| < \infty$—any function $g \colon \mathcal{B} \to \mathcal{M}$ mapping from minds to brains is called a *classifier*. Define misclassification rate, the probability that g misclassifies b under distribution $F = F_{MB}$, as

$$L_F(g) = \mathbb{P}[g(B) \neq M] = \sum_{(m,b) \in \mathcal{M} \times \mathcal{B}} \mathbb{I}\{g(b) \neq m\} \mathbb{P}[B = b, M = m], \tag{1}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function taking value unity whenever its argument is true and zero otherwise. The Bayes optimal classifier g^* minimizes $L_F(g)$ over all classifiers: $g^* = \operatorname{argmin}_g L_F(g)$. The *Bayes error*, or Bayes risk, $L^* = L_F(g^*)$, is the minimum possible misclassification rate.

The primary result of casting supervenience in a statistical framework is the below theorem, which follows immediately from Definition 1 and Eq. (1):

Theorem 1. $\mathcal{M} \overset{S}{\sim}_F \mathcal{B} \Leftrightarrow L^* = 0.$

The above argument shows (for the first time to our knowledge) that statistical supervenience and zero Bayes error are equivalent. Statistical supervenience can therefore be thought of as a constraint on the possible distributions on minds and brains. Specifically, let $\mathcal{F}$ indicate the set of all possible joint distributions on minds and brains, and let $\mathcal{F}_s = \{F_{MB} \in \mathcal{F} : L^* = 0\}$ be the subset of distributions for which supervenience holds. Theorem 1 implies that $\mathcal{F}_s \subsetneq \mathcal{F}$. Mind-brain supervenience is therefore an extremely restrictive assumption about the possible relationships between minds and brains. It seems that such a restrictive assumption begs for empirical evaluation, vis-á-vis, for instance, a hypothesis test.

The non-existence of a viable statistical test for supervenience

The above theorem implies that if we desire to know whether minds supervene on brains, we can check whether $L^* = 0$. Unfortunately, L^* is typically unknown. Fortunately, we can approximate L^* using training data.

Assume that training data $\mathcal{T}_n = \{(M_1, B_1), \ldots, (M_n, B_n)\}$ are each sampled identically and independently (iid) from the true (but unknown) joint distribution $F = F_{MB}$. Let g_n be a classifier induced by the training data, $g_n : \mathcal{B} \times (\mathcal{M} \times \mathcal{B})^n \mapsto \mathcal{M}$. The misclassification rate of such a classifier is given by

$$L_F(g_n) = \sum_{(m,b) \in \mathcal{M} \times \mathcal{B}} \mathbb{I}\{g_n(b; \mathcal{T}_n) \neq m\} \mathbb{P}[B = b, M = m], \tag{2}$$

which is a random variable due to the dependence on a randomly sampled training set $\mathcal{T}_n$. Calculating the expected misclassification rate $\mathbb{E}[L_F(g_n)]$ is often intractable in practice because it requires a sum over all possible training sets. Instead, expected misclassification rate can be approximated by "hold-out" error. Let $\mathcal{H}_{n'} = \{(M_{n+1}, B_{n+1}), \ldots, (M_{n+n'}, B_{n+n'})\}$ be a set of n' hold-out samples, each sampled iid from F_{MB}. The hold-out approximation to the misclassification rate is given by

$$\widehat{L}_F^{n'}(g_n) = \sum_{(M_i, B_i) \in \mathcal{H}_{n'}} \mathbb{I}\{g_n(B_i; \mathcal{T}_n) \neq M_i\} \approx \mathbb{E}[L_F(g_n)] \geq L^*. \tag{3}$$

By definition of g^*, the expectation of $\widehat{L}_F^{n'}(g_n)$ (with respect to both $\mathcal{T}_n$ and $\mathcal{H}_{n'}$) is greater than or equal to L^* for any g_n and all n. Thus, we can construct a hypothesis test for L^* using the surrogate $\widehat{L}_F^{n'}(g_n)$.

A statistical test proceeds by specifying the allowable Type I error rate $\alpha > 0$ and then calculating a test statistic. The p-value—the probability of rejecting the least favorable null hypothesis (the simple hypothesis within the potentially composite null which is closest to the boundary with the alternative hypothesis)—is the probability of observing a result at least as extreme as the observed. In other words, the p-value is the cumulative distribution function of the test statistic evaluated at the observed test statistic with parameter given by the least favorable null distribution. We reject if the p-value is less than α. A test is *consistent* whenever its power (the probability of rejecting the null when it is indeed false) goes to unity as $n \to \infty$. For any statistical test, if the p-value converges in distribution to δ_0 (point mass at zero), then whenever $\alpha > 0$, power goes to unity.

Based on the above considerations, we might consider the following hypothesis test: $H_0 : L^* > 0$ and $H_A : L^* = 0$; rejecting the null indicates that $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$. Unfortunately, the alternative hypothesis lies on the boundary, so the p-value is always equal to unity [7]. From this, Theorem 2 follows immediately:

Theorem 2. *There does not exist a viable test of* $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$.

In other words, we can *never* reject $L^* > 0$ in favor of supervenience, no matter how much data we obtain.

Conditions for a consistent statistical test for ε-supervenience

To proceed, therefore, we introduce a relaxed notion of supervenience:

Definition 2. *$\mathcal{M}$ is said to ε-supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$, if and only if $L^* < \varepsilon$ for some $\varepsilon > 0$.*

Given this relaxation, consider the problem of testing for ε-supervenience:

$$H_0^\varepsilon : L^* \geq \varepsilon$$
$$H_A^\varepsilon : L^* < \varepsilon.$$

Let $\widehat{n} = n' \widehat{L}_F^{n'}(g_n)$ be the *test statistic*. The distribution of $\widehat{n}$ is available under the least favorable null distribution. For the above hypothesis test, the p-value is therefore the binomial cumulative distribution function with parameter ε; that is, p-value $= \mathbb{B}(\widehat{n}; n', \varepsilon) = \sum_{k \in [\widehat{n}]_0} \text{Binomial}(k; n'; \varepsilon)$, where $[\widehat{n}]_0 = \{0, 1, \ldots, \widehat{n}\}$. We reject whenever this p-value is less than α; rejection implies that we are $100(1-\alpha)$% confident that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$.

For the above ε-supervenience statistical test, if $g_n \to g^*$ as $n \to \infty$, then $\widehat{L}_F^{n'}(g_n) \to L^*$ as $n, n' \to \infty$. Thus, if $L^* < \varepsilon$, power goes to unity. The definition of ε-supervenience therefore admits, for the first time to our knowledge, a viable statistical test of supervenience, given a specified ε and α. Moreover, this test is consistent whenever g_n converges to the Bayes classifier g^*.

The existence and construction of a consistent statistical test for ε-supervenience

The above considerations indicate the existence of a consistent test for ε-supervenience whenever the classifier used is consistent. To actually implement such a test, one must be able to (i) measure mind/brain pairs and (ii) have a consistent classifier g_n. Unfortunately, we do not know how to measure the entirety of one's brain, much less one's mind. We therefore must restrict our interest to a mind/brain *property* pair. A mind (mental) property might be a person's intelligence, psychological state, current thought, gender identity, etc. A brain property might be the number of cells in a person's brain at some time t, or the collection of spike trains of all neurons in the brain during some time period t to t'. Regardless of the details of the specifications of the mental property and the brain property, given such specifications, one can assume a model, $\mathcal{F}$. We desire a classifier g_n that is guaranteed to be consistent, no matter which of the possible distributions $F_{MB} \in \mathcal{F}$ is the true distribution. A classifier with such a property is called a *universally consistent classifier*. Below, under a very general mind-brain model $\mathcal{F}$, we construct a universally consistent classifier.

***Gedankenexperiment* 1.** *Let the physical property under consideration be brain connectivity structure, so b is a brain-graph ("connectome") with vertices representing neurons (or collections thereof) and edges representing synapses (or collections thereof). Further let $\mathcal{B}$, the brain observation space, be the collection of all graphs on a given finite number of vertices, and let $\mathcal{M}$, the mental property observation space, be finite. Now, imagine collecting very large amounts of very accurate identically and independently sampled brain-graph data and associated mental property indicators from F_{MB}. A k_n-nearest neighbor classifier using a Frobenius norm is universally consistent (see Methods for details). The existence of a universally consistent classifier guarantees that eventually (in n, n') we will be able to conclude $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ for this mind-brain property pair, if indeed ε-supervenience holds. This logic holds for directed graphs or multigraphs or hypergraphs with discrete edge weights and vertex attributes, as well as unlabeled graphs (see [8] for details). Furthermore, the proof holds for other matrix norms (which might speed up convergence and hence reduce the required n), and the regression scenario where $|\mathcal{M}|$ is infinite (again, see Methods for details).*

Thus, under the conditions stated in the above *Gedankenexperiment*, universal consistency yields:

Theorem 3. $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B} \implies \beta \to 1$ *as* $n, n' \to \infty$.

Unfortunately, the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ depends on the (unknown) distribution $F = F_{MB}$ [9]. Furthermore, arbitrarily slow convergence theorems regarding the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ demonstrate that there is no universal n, n' which will guarantee that the test has power greater than any specified target $\beta > \alpha$ [10]. For this reason, the test outlined above can provide only a one-sided conclusion: if we reject

we can be $100(1-\alpha)$% confident that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ holds, but we can never be confident in its negation; rather, it may be the case that the evidence in favor of $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ is insufficient because we simply have not yet collected enough data. This leads immediately to the following theorem:

Theorem 4. *For any target power $\beta_{min} > \alpha$, there is no universal n, n' that guarantees $\beta \geq \beta_{min}$.*

Therefore, even ε-supervenience does not satisfy Popper's falsifiability criterion [11].

The feasibility of a consistent statistical test for ε-supervenience

Theorem 3 demonstrates the availability of a consistent test under certain restrictions. Theorem 4, however, demonstrates that convergence rates might be unbearably slow. We therefore provide an illustrative example of the feasibility of such a test on synthetic data.

Caenorhabditis elegans is a species whose nervous system is believed to consist of the same 279 labeled neurons for each organism [12]. Moreover, these animals exhibit a rich behavioral repertoire that seemingly depends on circuit properties [13]. These findings motivate the use of C. elegans for a synthetic data analysis [14]. Conducting such an experiment requires specifying a joint distribution F_{MB} over brain-graphs and behaviors. The joint distribution decomposes into the product of a class-conditional distribution (likelihood) and a prior, $F_{MB} = F_{B|M} F_M$. The prior specifies the probability of any particular organism exhibiting the behavior. The class-conditional distribution specifies the brain-graph distribution given that the organism does (or does not) exhibit the behavior.

Let A_{uv} be the number of chemical synapses between neuron u and neuron v according to [15]. Then, let $\mathcal{S}$ be the set of edges deemed responsible for odor-evoked behavior according to [16]. If odor-evoked behavior is supervenient on this signal subgraph $\mathcal{S}$, then the distribution of edges in $\mathcal{S}$ must differ between the two classes of odor evoked behavior [17]. Let $E_{uv|j}$ denote the expected number of edges from vertex v to vertex u in class j. For class m_0, let $E_{uv|0} = A_{uv} + \eta$, where $\eta = 0.05$ is a small noise parameter (it is believed that the C. elegans connectome is similar across organisms [12]). For class m_1, let $E_{uv|1} = A_{uv} + z_{uv}$, where the signal parameter $z_{uv} = \eta$ for all edges not in $\mathcal{S}$, and z_{uv} is uniformly sampled from $[-5, 5]$ for all edges within $\mathcal{S}$. For both classes, let each edge be Poisson distributed, $F_{A_{uv}|M=m_j}$ = Poisson$(E_{uv|j})$.

We consider k_n-nearest neighbor classification of labeled multigraphs (directed, with loops) on 279 vertices, under Frobenius norm. The k_n-nearest neighbor classifier used here satisfies $k_n \to \infty$ as $n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, ensuring universal consistency. (Better classifiers can be constructed for the joint distribution F_{MB} used here; however, we demand universal consistency.) Figure 1 shows that for this simulation, rejecting $(\varepsilon = 0.1)$-supervenience at $\alpha = 0.01$ requires only a few hundred training samples.

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D superresolution imaging [18] combined with neurite tracing algorithms [19, 20, 21] allow the collection of a C. elegans brain-graph within a day. Genetic manipulations, laser ablations, and training paradigms can each be used to obtain a non-wild type population for use as $M = m_1$ [13], and the class of each organism (m_0 vs. m_1) can also be determined automatically [22].

Discussion

This work makes the following contributions. First, we define statistical supervenience based on Davidson's canonical statement (Definition 1). This definition makes it apparent that supervenience implies the possibility of perfect classification (Theorem 1). We then prove that there is no viable test against supervenience, so one can *never* reject a null hypothesis in favor of supervenience, regardless of the amount of data (Theorem 2). This motivates the introduction of a relaxed notion called ε-supervenience (Definition 2), against which consistent statistical tests are readily available. Under a very general brain-graph/mental property model (*Gedankenexperiment* 1), a consistent statistical test against ε-supervenience is always available no matter the true distribution F_{MB} (Theorem 3). In other words, the proposed test is guaranteed to reject the null whenever the null is false, given sufficient data, for any possible distribution governing mental property/brain property pairs.

Alas, arbitrary slow convergence theorems demonstrate that there is no universal n, n' for which convergence is guaranteed (Theorem 4). Thus, a failure to reject is ambiguous: even if the data satisfy the above assumptions, the failure to reject may be due to either (i) an insufficient amount of data or (ii) $\mathcal{M}$ may not be ε-supervenient on

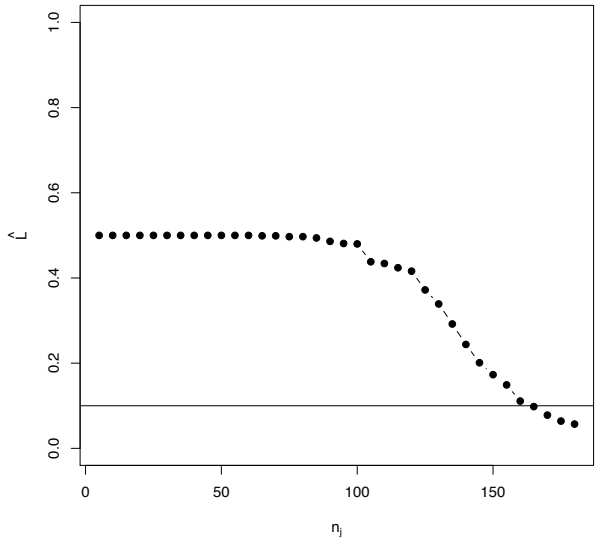

Figure 1: C. elegans graph classification simulation results. The estimated hold-out misclassification rate $\widehat{L}_F^{n'}(g_n)$ (with $n' = 1000$ testing samples) is plotted as a function of class-conditional training sample size $n_j = n/2$, suggesting that for $\varepsilon = 0.1$ we can determine that $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ holds with 99% confidence with just a few hundred training samples generated from F_{MB}. Each dot depicts $\widehat{L}_F^{n'}(g_n)$ for some n; standard errors are $(\widehat{L}_F^{n'}(g_n)(1 - \widehat{L}_F^{n'}(g_n))/n')^{1/2}$. For example, at $n_j = 180$ we have $k_n = \lfloor\sqrt{8n}\rfloor = 53$ (where $\lfloor\cdot\rfloor$ indicates the floor operator), $\widehat{L}_F^{n'}(g_n) = 0.057$, and standard error less than 0.01. We reject $H_0^{0.1} : L^* \geq 0.1$ at $\alpha = 0.01$. Note that $L^* \approx 0$ for this simulation.

$\mathcal{B}$. Moreover, the data will not, in general, satisfy the above assumptions. In addition to dependence (because each human does not exist in a vacuum), the mental property measurements will often be "noisy" (for example, accurately diagnosing psychiatric disorders is a sticky wicket [23]). Nonetheless, synthetic data analysis suggests that under somewhat realistic assumptions, convergence obtains with an amount of data one might conceivably collect (Figure 1 and ensuing discussion).

Thus, given measurements of mental and brain properties that we believe reflect the properties of interest, and given a sufficient amount of data satisfying the independent and identically sampled assumption, a rejection of $H_0^\varepsilon : L^* \geq \varepsilon$ in favor of $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ entails that we are $100(1-a)\%$ confident that the mental property under investigation is ε-supervenient on the brain property under investigation. Unfortunately, failure to reject is more ambiguous.

Interestingly, much of contemporary research in neuroscience and cognitive science could be cast as mind-brain supervenience investigations. Specifically, searches for "engrams" of memory traces [24] or "neural correlates" of various behaviors or mental properties (for example, consciousness [25]), may be more aptly called searches for the "neural supervenia" of such properties. Letting the brain property be a brain-graph is perhaps especially pertinent in light of the advent of "connectomics" [26, 27], a field devoted to estimating whole organism brain-graphs and relating them to function. Testing supervenience of various mental properties on these brain-graphs will perhaps therefore become increasingly compelling; the framework developed herein could be fundamental to these investigations. For example, questions about whether connectivity structure alone is sufficient to explain a particular mental property is one possible mind-brain ε-supervenience investigation. The above synthetic data analysis demonstrates the feasibility of ε-supervenience on small brain-graphs. Note that ε-supervenience tests need not investigate seemingly intractable problems, like consciousness. For example, aspects of visual perception appear to supervene on visual cortical activity (for example, binocular rivalry [28]). Moreover, an inability to reject ε-supervenience for small ε is also potentially meaningful. For example, perhaps auditory localization precision supervenes on a rate code only to some $\varepsilon > c$, the rest supervening on a spike

timing code [29]. Similar supervenience tests on increasingly complex mental properties will potentially benefit from either higher-throughput imaging modalities [30, 31], more coarse brain-graphs [32, 33], or both.

Methods

The 1-nearest neighbor (1-NN) classifier works as follows. Compute the distance between the test brain b and all n training brains, $d_i = d(b, b_i)$ for all $i \in [n]$, where $[n] = 1, 2, \ldots, n$. Then, sort these distances, $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$, and consider their corresponding minds, $m_{(1)}, m_{(2)}, \ldots, m_{(n)}$, where parenthetical indices indicate rank order among $\{d_i\}_{i \in [n]}$. The 1-NN algorithm predicts that the unobserved mind is of the same class as the closest brain's class: $\widehat{m} = m_{(1)}$. The k_n nearest neighbor is a straightforward generalization of this approach. It says that the test mind is in the same class as whichever class is the plurality class among the k_n nearest neighbors, $\widehat{m} = \text{argmax}_{m'} \mathbb{I}\{\sum_{i=1}^{k_n} m_{(i)} = m'\}$. Given a particular choice of k_n (the number of nearest neighbors to consider) and a choice of $d(\cdot, \cdot)$ (the distance metric used to compare the test datum and training data), one has a relatively simple and intuitive algorithm.

Let g_n be the k_n nearest neighbor (k_nNN) classifier when there are n training samples. A collection of such classifiers $\{g_n\}$, with k_n increasing with n, is called a classifier sequence. A universally consistent classifier sequence is any classifier sequence that is guaranteed to converge to the Bayes optimal classifier regardless of the true distribution from which the data were sampled; that is, a universally consistent classifier sequence satisfies $L_F(g_n) \to L_F(g^*)$ as $n \to \infty$ for all F_{MB}. In the main text, we refer to the whole sequence as a classifier.

The k_nNN classifier is consistent if (i) $k_n \to \infty$ as $n \to \infty$ and (ii) $k_n/n \to 0$ as $n \to \infty$ [34]. In Stone's original proof [34], b was assumed to be a q-dimensional vector, and the L_2 norm ($d(b, b') = \sum_{j=1}^{q} (b_j - b'_j)^2$, where j indexes elements of the q-dimensional vector) was shown to satisfy the constraints on a distance metric for this collection of classifiers to be universally consistent. Later, others extended these results to apply to any L_p norm [9]. When brain-graphs are represented by their adjacency matrices, one can stack the columns of the adjacency matrices, effectively embedding graphs into a vector space, in which case Stone's theorem applies. Stone's original proof also applied to the scenario when $|\mathcal{M}|$ was infinite, resulting in a universally consistent regression algorithm as well.

Note that the above extension of Stone's original theorem to the graph domain implicitly assumed that vertices were labeled, such that elements of the adjacency matrices could easily be compared across graphs. In theory, when vertices are unlabeled, one could first map each graph to a quotient space invariant to isomorphisms, and then proceed as before. Unfortunately, there is no known polynomial time complexity algorithm for graph isomorphism [35], so in practice, dealing with unlabeled vertices will likely be computationally challenging [8].

References

[1] Plato. *Plato: complete works*. Hackett Pub Co, (1997).

[2] Davidson, D. *Experience and Theory*, chapter Mental Eve. Duckworth (1970).

[3] Haykin, S. *Neural Networks and Learning Machines*. Prentice Hall, 3rd edition, (2008).

[4] Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, (2008).

[5] Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. *Cognitive Neuroscience: The Biology of the Mind (Third Edition)*. W. W. Norton & Company, (2008).

[6] Kim, J. *Physicalism, or Something Near Enough (Princeton Monographs in Philosophy)*. Princeton University Press, (2007).

[7] Bickel, P. J. and Doksum, K. A. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I (2nd Edition)*. Prentice Hall, (2000).

[8] Vogelstein, J. T. and Priebe, C. E. *Submitted for publication* (2011).

[9] Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, (1996).

[10] Devroye, L. *Utilitas Mathematica* **483**, 475–483 (1983).

[11] Popper, K. R. *The logic of scientific discovery*. Routledge, (1959).

[12] Durbin, R. M. *Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans*. PhD thesis, University of Cambridge, (1987).

[13] de Bono, M. and Maricq, A. V. *Annu Rev Neurosci* **28**, 451–501 (2005).

[14] Gelman, A. and Shalizi, C. R. *Submitted for publication* , 1–36 (2011).

[15] Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., Chklovskii, D. B., Spring, C., and Farm, J. *World Wide Web Internet And Web Information Systems* , 1–41.

[16] Chalasani, S. H., Chronis, N., Tsunozaki, M., Gray, J. M., Ramot, D., Goodman, M. B., and Bargmann, C. I. *Nature* **450**(7166), 63–70 November (2007).

[17] Vogelstein, J. T., Gray, W. R., Vogelstein, R. J., and Priebe, C. E. *Submitted for publication* (2011).

[18] Vaziri, A., Tang, J., Shroff, H., and Shank, C. V. *Proceedings of the National Academy of Sciences of the United States of America* **105**(51), 20221–6 December (2008).

[19] Helmstaedter, M., Briggman, K. L., and Denk, W. *Current opinion in neurobiology* **18**(6), 633–41 December (2008).

[20] Mishchenko, Y. *J Neurosci Methods* **176**(2), 276–289 January (2009).

[21] Lu, J., Fiala, J. C., and Lichtman, J. W. *PLoS ONE* **4**(5), e5655 (2009).

[22] Buckingham, S. D. and Sattelle, D. B. *Invertebrate neuroscience : IN* **8**(3), 121–31 September (2008).

[23] Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., and Walters, E. E. *Archives of general psychiatry* **62**(6), 593–602 June (2005).

[24] Lashley, K. S. *Symposia of the society for experimental biology* **4**(454-482), 30 (1950).

[25] Koch, C. *The Quest for Consciousness*. Roberts and Company Publishers, (2010).

[26] Sporns, O., Tononi, G., and Kotter, R. *PLoS Computational Biology* **1**(4), e42 (2005).

[27] Hagmann, P. *From diffusion MRI to brain connectomics*. PhD thesis, Institut de traitement des signaux, (2005).

[28] Tong, F., Meng, M., and Blake, R. *Trends in Cognitive Sciences* **10**(11), 502–511 (2006).

[29] Chase, S. M. and Young, E. D. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **26**(15), 3889–98 April (2006).

[30] Hayworth, K. J., Kasthuri, N., Schalek, R., Lichtman, J. W., Program, N., Angeles, L., and Biology, C. *World* **12**(Supp 2), 86–87 (2006).

[31] Bock, D. D., Lee, W.-C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzel, A. W., Yurgenson, S., Soucy, E. R., Kim, H. S., and Reid, R. C. *Nature* **471**(7337), 177–182 March (2011).

[32] Palm, C., Axer, M., Gräß el, D., Dammers, J., Lindemeyer, J., Zilles, K., Pietrzyk, U., and Amunts, K. *Frontiers in Human Neuroscience* **4** (2010).

[33] Johansen-Berg, H. and Behrens, T. E. *Diffusion MRI: From quantitative measurement to in-vivo neuroanatomy*. Academic Press, (2009).

[34] Stone, C. J. *The Annals of Statistics* **5**(4), 595–620 July (1977).

[35] Garey, M. R. and Johnson, D. S. *Computers and intractability. A guide to the theory of NP-completeness. A Series of Books in the Mathematical Sciences*. WH Freeman and Company, San Francisco, Calif, (1979).

Acknowledgments

The authors would like to acknowledge helpful discussions with J. Lande, B. Vogelstein, S. Seung, and K. Kording.

Author Contributions

JTV, RJV, and CEP conceived of the manuscript. JTV and CEP wrote it. CEP ran the experiment.

Additional Information

The authors have no competing financial interests to declare.