Are mental properties supervenient on brain properties?

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

August 16, 2010

Abstract

The "mind-brain supervenience" conjecture suggests that all mental properties (e.g. consciousness, intelligence, personality, etc.) are derived from the physical properties of the brain. The validity of this conjecture has been argued in philosophical terms for over 2,500 years. Alternative conjectures, including various non-physical causes of mental properties, seem rather implausible to many, but proving or disproving these alternatives has remained elusive.

To address the question of whether the mind supervenes on the brain through empirical means, here we frame a supervenience hypothesis in rigorous mathematical terms and propose a modified version of supervenience (called ε-supervenience) that is amenable to scientific methods and statistical analysis. To elucidate this approach, we posit 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 *connectome*), and ε-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given misclassification rate $\varepsilon > 0$, regardless of the relationship between the two. In addition to the theoretical results, we show via simulation that given reasonable assumptions about class conditional probabilities and the amount of data available, the thought experiment can actually be conducted on a simple organism, *Caenorhabditis elegans*, with currently available technology.

The potential significance of this work can be divided into distinct disciplines. To the philosopher, this work demonstrates that philosophical conjectures can be morphed into statistical hypotheses, amenable to experimental investigations, allowing the philosopher to add empirical support to their rational arguments. To the statistician, herein lies the first proof to our knowledge of the existence of a universally consistent classifier on graphs, and a constructivist one at that. To the neuroscientist, a theoretically possible experiment is proposed to garnish support for a hypothesis that is widely believed: that mental properties supervene on brain properties.

1 Introduction

Questioning the relationship between the mind (thoughts, beliefs, preferences, emotions, intelligence, etc.) and the brain (the physical structure inside our skulls) dates back at least as far as 400 BCE, when Plato wrote the dialogues, in which he posited immateriality of the soul [**?**]. Approximately two millennia passed before these ideas reached their canonical form through Descartes's discussion of mind-body dualism [**?**]. Then, in the 20th century, Donald Davidson stated and popularized the mind-brain supervenience conjecture, which claims that an agent cannot alter in some mental property without altering in some physical property [**?**]. Contemporary fields of neural network theory and neuroscientific inquiry often assume mind-brain supervenience, or an even stronger assumption about mind-brain causality, but no previously proposed notion of supervenience seems amenable to empirical investigation. Here we define new versions of supervenience that formulate the conjecture in rigorous mathematical terms and that can be experimentally tested as a hypothesis.

The primary contributions of this work are as follows. First, a notion of supervenience amenable to empirical investigation is formally introduced. This renders the mind-brain dualism debate a hypothesis, rather than an assumption. Second, in addition to expanding the space of questions amenable to hypothesis testing, we also demonstrate the limits of hypothesis testing. Third we posit a very general model of brains and their associated mental properties that admits statistical analysis in a graph theoretical and statistical framework. Fourth, we prove that this formulation admits a universally consistent classifier that is guaranteed to find the relationship between minds and brains, if one exists. Fifth we demonstrate through simulation that the proposed universally consistent classifier has reasonable convergence properties on simulated brain-graph data.

2 Preliminaries

The intention in this work is to develop greater insight regarding the relationship between minds and brains, using statistical methods, with particular interest in notions of supervenience. We therefore first define the objects of interest, that is, minds and brains.

Let b correspond to an agent's brain, which is a particular element from the set of all possible brains, $\mathcal{B}$. The set of possible brains $\mathcal{B}$ is completely unrestricted, meaning that $\mathcal{B}$ could be an infinite set, with arbitrarily complexity. In particular, b might represent the position, momentum, and type of each subatomic particle residing within the skull some agent. Thus, each different $b \in \mathcal{B}$ corresponds to some difference in the position, momentum, or type of at least one subatomic particle composing the brain.

Similarly, let m correspond to an agent's mind, which is a particular element from the set of all possible minds, $\mathcal{M}$. The set of possible minds $\mathcal{M}$ is also unrestricted. In particular, m might represent all an agent's thoughts, beliefs, and preferences. Thus, each different $m \in \mathcal{M}$ corresponds to a difference in at least one thought, belief, or preference.

The mind-brain supervenience conjecture, is a relation between these two sets, the set of mental states and the set of brain states. Donald Davidson canonized this conjecture in 1970 with the following quote: [**?**]

> 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,

can be concisely and formally stated: $m \neq m' \implies b \neq b' \, \forall (b,m),(b',m') \in \mathcal{B} \times \mathcal{M}$. While mind-brain supervenience is a relatively strong claim, importantly, one can imagine far stronger relations, such as the following.

First, it may be the case that minds supervene on brains, but one cannot form an *injective* relation from brains to minds. An injective relation is any relation that preserves distinctness. Thus if minds are injective on brains, then $b \neq b' \implies m \neq m' \, \forall (b,m),(b',m') \in \mathcal{B} \times \mathcal{M}$ (note that the directionality of the implication has been switched relative to supervenience). For instance, it might be the case that a brain could change without the mind changing. Consider the case that a single subatomic particle shifts its position by a Plank length, changing brain state from b to b'. It is possible that the mental state supervening on brain state b remains m, even after b changes to b'. In such a scenario, the mind might still supervene on the brain, but the relation from brains to minds is not injective. This argument also shows that supervenience is not necessarily a *symmetric* relation. Minds supervening on brains does not imply that brains supervene on minds.

Second, it may be the case that minds supervene on brains, but that brains do not cause minds. For instance, consider an analogy where M and B correspond to two coins being flipped, each possibly landing on heads or tails. Further assume that every time one lands on heads so does the other, and every time one lands on tails, so do the other. This implies that M supervenes on B, but assumes nothing about whether M causes B, or B causes M, or some exogenous force causes both.

Third, supervenience does not imply *identity*. Consider, for example, acceleration and velocity. Clearly, acceleration supervenes on velocity, as acceleration cannot change without velocity changing (assuming one does not consider gravity as acceleration). Similarly, velocity supervenes on position, as velocity cannot change without position changing. Therefore, acceleration supervenes on position, by the transitive property of supervenience, but it is not the case that a change in acceleration is equal to a change in position. Rather, position can change with constant velocity, meaning without acceleration changing.

What supervenience does imply, however, is the following. Imagine finding two different minds. If $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$, then the brains subvening under those two minds must be different. In other words, there cannot be two different minds,

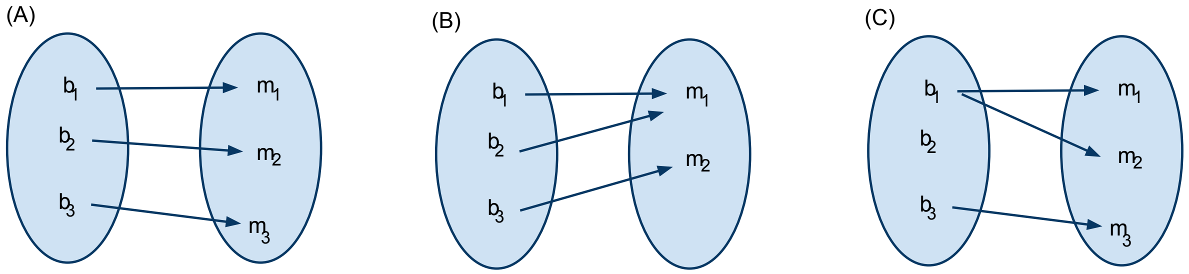

Figure 1: Possible relations between minds and brains. (A) Minds supervene on brains, and it so happens that there is a bijective relation from brains to minds. (B) Minds supervene on brains, and it so happens that there is a surjective relation from brains to minds. (C) Minds are *not* supervenient on brains, because two different minds supervene on the same brain.

either of which could supervene on a single brain. Figure 1 shows several possible relations between the sets of minds and brains.

All of the above relations are *logical* relations, not probabilistic relations. To facilitate both statistical analysis and empirical investigation, we project this supervenience notion into a statistical setting. To proceed, we first define a *model*, $\mathbb{P}[B, M] = F_{BM}$, which specifies the probability of any element (b, m) occurring from the space of all possible elements $\mathcal{B} \times \mathcal{M}$ (also called the *sample space*). The model could in theory come from any possible joint distribution defined on the sample space, $F_{BM} \in \mathcal{F}$, meaning that we only assume that it conforms to the *probability axioms*:

1. $0 \leq \mathbb{P}[B = b, M = m] \leq 1$ for all $(b, m) \in \mathcal{M} \times \mathcal{B}$
2. $\mathbb{P}[\Omega] = 1$ and $\mathbb{P}[\emptyset] = 0$, where Ω is the whole sample sample, that is, all elements, $(b, m) \in \mathcal{B} \times \mathcal{M}$
3. any countable sequence of pairwise disjoint elements, $(b_1, m_1), (b_2, m_2), \ldots$ satisfies $\mathbb{P}[(b_1, m_1) \cup (b_2, m_2) \cup \ldots] = \sum_i \mathbb{P}[(b_i, m_i)]$.

Given a model, one can then calculate any marginal or conditional distributions as a function of the model. For instance, $\mathbb{P}[B] = \int_{m \in \mathcal{M}} \mathbb{P}[B, M] \mathrm{d}m$, or $\mathbb{P}[M|B] = \mathbb{P}[B, M]/\mathbb{P}[B]$. Given the definition of a model, statistical supervenience can be defined as follows:

Definition 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{BM}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $\mathbb{P}[m \neq m'|b = b'] = 0 \, \forall (b, m), (b', m') \in \mathcal{B} \times \mathcal{M}$. Alternately, $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ can be written as $\mathbb{P}[m = m'|b = b'] = 1 \, \forall (b, m), (b', m') \in \mathcal{B} \times \mathcal{M}$.*

Statistical supervenience is therefore a probabilistic relation on sets. Note that statistical supervenience is distinct from statistical correlation. *Statistical correlation* between brain states and mental states is defined as $\rho_{BM} = \mathbb{E}[(B - \mu_B)(M - \mu_M)]/(\sigma_B \sigma_M)$, where μ_X and σ_X are the mean and variance of X, and $\mathbb{E}[X]$ is the expected value of X. If $\rho_{BM} = 1$, then both $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ and $\mathcal{B} \overset{S}{\sim}_F \mathcal{M}$. Thus, correlation implies supervenience, but supervenience does not imply correlation.

3 Results

3.1 Theoretical results

If minds statistically supervene on brains, $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, then two different minds must supervene on two different brains. This means that there exists a unique mapping from each brain to a single mind. In other words, one can in principle construct a function $g(b) : \mathcal{B} \mapsto \mathcal{M}$, that is a deterministic mapping from brains to minds. It may be the case that subsets of brains from equivalence classes, such that any brain in that subset is mapped to the same mind (see, for example, b_1 and b_2 in Figure 1(A)). Assuming for the moment that the space of all possible minds is finite, that is $|\mathcal{M}| < \infty$, then we call any such function a *classifier* (this assumption will later be relaxed). Let $\widehat{m}$ denote the output of a classifier, $g(b) = \widehat{m}$. Define misclassification rate as:

$$L_F(g) = P_F[g(B) \neq M] = \frac{1}{|\mathcal{B}||\mathcal{M}|} \iint \mathbb{I}\{g(b) \neq m\} \mathrm{d}b \mathrm{d}m \qquad (1)$$

where $\mathbb{I}\{\cdot\}$ indicates the indicator function, taking value one if its argument is true, and zero otherwise. $L_F(g)$ therefore effectively counts the fraction of time g misclassifies b. The Bayes optimal classifier g^* minimizes $L_F(g)$ over all classifiers, that is

$$g^* = \operatorname*{argmin}_{g \in \mathcal{G}} L_F(g) \qquad (2)$$

where $\mathcal{G}$ is the set of all possible classifiers. Thus, the *Bayes error*, or Bayes risk, $L_F(g^*)$ is the minimum possible misclassification rate. The primary result of casting supervenience is a statistical framework is the following theorem:

Theorem 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{BM}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $L_F(g^*) = 0$.*

If minds supervene on brains, then, by the definition of supervenience, there exists a function that maps each brain deterministically to a particular mind. This means that one could draw a decision boundary between all equivalence classes of brains, each class corresponding to a different mind, and no mind will reside within two different equivalence classes. Thus, the optimal classifier would correctly find these decision boundaries, and therefore have no opportunity to err. □

This relationship between statistical supervenience and Bayes error can therefore be described concisely: $\mathcal{M} \overset{S}{\sim}_F \mathcal{B} \Leftrightarrow L_F(g^*) = 0$. Thus, the above arguments shows (for the first time to our knowledge) that statistical supervenience and zero Bayes error are equivalent. Further, statistical supervenience can be thought of as a constraint on the possible models. Specifically, let $\mathcal{F}_s \subset \mathcal{F}$ be subset of models for which supervenience holds. Then, $\mathcal{F}_s = \{F | L_F(g^*) = 0\} \subset \mathcal{F}$.

3.2 Hypothesis testing

While the above theorem is of potential theoretical interest, because the arguments rest on knowing F_{BM} and g^*, which are typically unknown, they are pragmatically useless. However, both F_{BM} and g^* could be estimated from data. Let $(b_1, m_1), (b_2, m_2), \ldots, (b_n, m_n)$ be random samples taking their values in $\mathcal{B} \times \mathcal{M}$, independently and identically distributed according to model F_{BM}. Generalizing the concept of a classifier g to allow incorporation of training data, consider $g_n : \mathcal{B} \times (\mathcal{B} \times \mathcal{M})^n \mapsto \mathcal{M}$ which takes as input an observed brain connectivity structure b and n training pairs $\mathcal{T}_n = \{(b_1, m_1), \cdots, (b_n, m_n)\}$ and produces a classification $g_n(b; \mathcal{T}_n) = \widehat{m}$. Misclassification rate for this classifier will therefore be a random variable, because the training data $\mathcal{T}_n$ are random samples. Therefore, instead of calculating misclassification rate for g_n, we compute the expected misclassification rate:

$$\mathbb{E}[L_F(g_n)] = \mathbb{E}[P_F[g_n(B; \mathcal{T}_n) \neq M | \mathcal{T}_n]] = \int \mathbb{P}_F[g_n(B) = M | \mathcal{T}_n] \mathbb{P}[\mathcal{T}_n] d\mathcal{T}_n. \qquad (3)$$

Unfortunately, in practice, computing $\mathbb{E}[L_F(g_n)]$, requires integrating over all possible training data corpuses, and by definition, we only have access to a single training data corpus. We therefore define "hold out" misclassification performance:

$$\mathbb{E}[L_F(g_n)] \approx \widehat{L}_F^{n'}(g_n) = \sum_{\mathcal{T}_{n-n'}} \mathbb{P}_F[g_n(B) = M | \mathcal{T}_{n-n'}] \mathbb{P}[\mathcal{T}_{n-n'}], \qquad (4)$$

where $n' < n$ is the number of held-out training samples, and the sum is taken over a sufficiently large number of subsets, $\mathcal{T}_{n-n'}$, such that $\widehat{L}_F^{n'}(g_n)$ converges. $n' \widehat{L}_F^{n'}(g_n)$ is the expected number of misclassified minds. Simplifying further, assuming no sum, rather, just one "hold-out" set, then $n' \widehat{L}_F^{n'}(g_n)$ and has a binomial distribution, because for any of the n' held-out samples, the classifier could be either correct or incorrect, thus $n' \widehat{L}_F^{n'}(g_n) \sim \text{Binomial}(n', L_F(g_n))$.

Before explicitly considering the problem of testing for statistical supervenience, we define a relaxed notion of supervenience:

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

Given this relaxation, consider the problem of testing for $\varepsilon-$supervenience. First, specify a significance level, α, such that if the p-value is less than α, then the null is rejected. Because we hope to reject the null, in favor of the alternative, let the null hypothesis be H_0: $L_F(g_n) \geq \varepsilon$, and the alternative hypothesis be H_A: $L_F(g_n) < \varepsilon$. We reject for low values of the test-statistic, $n'\widehat{L}_F^{n'}(g_n)$. Specifically, if $n'\widehat{L}_F^{n'}(g_n)$ is less than the critical value, $c_\alpha(n', \varepsilon)$, then we reject. The critical value is available under the least favorable distribution Binomial(n', ε). Thus, if $n'\widehat{L}_F^{n'}(g_n) < c_\alpha(n', \varepsilon)$, we can conclude with $100(1-\alpha)\%$ confidence that $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$. The definition of ε-supervenience therefore admits, for the first time to our knowledge, a statistical test of supervenience, given a specified ε and α. Similar to the above, one can define the set of ε-supervenience models as the set of models under which ε-supervenience holds, that is: $\mathcal{F}_\varepsilon = \{F | L_F(g^*) < \varepsilon\}$. One could then sort ε-supervenience subsets, $\mathcal{F}_s \subseteq \mathcal{F}_\varepsilon \subseteq \mathcal{F}_{\varepsilon'} \subseteq \mathcal{F}$, for any $\varepsilon < \varepsilon'$.

3.3 Power and consistency

Importantly, the utility of any statistical test depends both on the p-value, the probability of obtaining a test statistic at least as extreme as the observed value (under the assumed model), and its power, the probability that the test will reject a false null hypothesis (in other words, the probability that it will not make a Type II error). Ideally, the power of this test would go to unity, as $n, n' \to \infty$. A sufficient condition for power to approach unity is that g_n is a *consistent* classifier. A classifier is consistent if and only if its expected misclassification rate converges to the Bayes optimal limit with sufficient data, that is $\mathbb{E}[L_F(g_n)] \to L_F(g^*)$ as $n \to \infty$. As the notation suggests, consistency of a classifier is a function of the true model, F. Without any prior knowledge of what the model might be, one desires a *universally consistent* classifier, that is a classifier that is consistent for all $F \in \mathcal{F}$.

Unfortunately, the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ depends on the (unknown) distribution $F = F_{BM}$ [**?**]. 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 $\beta > \alpha$ [**?**]. 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. This means that we can never be confident that $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ does *not* hold; rather, it may be the case that the evidence in favor of $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ is insufficient for any number of reasons, including that we simply have not yet collected enough data. Unfortunately, arbitrarily slow convergence theorems inform us that no matter how much data we collect, we cannot disambiguate between not yet having enough data, and $\mathcal{M}\overset{\varepsilon}{\sim}_F\mathcal{B}$ not holding. Thus, without restrictions on F_{BM}, arbitrarily slow convergence theorems imply that our theorem of ε-supervenience does not strictly satisfy Popper's *falsifiability* requirement [**?**]. Given these limitations on even universal consistency, it is still the best one can hope for. Therefore, we hope to obtain a universally consistent classifier to test for ε-supervenience.

In 1977, Stone proved that a certain collection of k_n nearest neighbor algorithms is universally consistent, assuming b was a random d-dimensional vector $\mathcal{B} \subseteq \mathbb{R}^d$ [**?**] (see Appendix A for an explanation of k_n nearest neighbor algorithms, and constraints to ensure universal consistency) (since then, other algorithms have also been shown to be universally consistent, most notably neural networks [**?**]). This beautiful result would be directly applicable to mind-brain supervenience, if we were satisfied representing brains by random vectors. However, brains are rich with structure, and therefore, other, more sophisticated data structures can better capture aspects of the brain that many of us believe to be true. Therefore, in the next section, we introduce the concept of a *brain-graph*, which can have a much richer structure than a vector. We then generalize the k_n nearest neighbor algorithm to operate in this domain, and extend the universal consistency proof as well.

3.4 Brain-graphs

In 1891, H Waldeyer-Hartz first formally proposed the "neuron doctrine" [**?**], which states that the nervous system is a complex *network* of "neurons" (a term invented in the above review), largely based on Ramon y Cajal's work using the Golgi stain [**?**]. This doctrine has been central to much of the development of neuroscience and artificial intelligence for over 100 years, including the development of neural network theory [**?**] and cognitive science [**?**]. In parallel with the development and refinement of concepts of how collections of neurons may collectively operate to subserve various brain functions, a simultaneous development has taken place in the statistical analysis of networks, or graphs [**?**, **?**]. We therefore propose to represent the brain as a random graph, where nodes represent neurons and edges represent synapses. Formally, a brain-graph $G = (V, A)$ is characterized by a set of vertices (or nodes), $\{V_i\} = \{V_1, \ldots, V_n\}$, where n is the number of neurons, and arcs (or edges) $\{A_{ij}\}$, where A_{ij} represents the synapse from neuron V_j to V_i. For simplicity (to be generalized below), assume that brain-graphs are simple graphs, meaning that (i) there are no self loops (no autapses, or self-connections), (ii) all edges are binary, meaning that $a_{ij} \in \{0, 1\}$, and (iii) all edges are symmetric, that is, $a_{ij} = a_{ji}$. Further assuming that the vertices are *labeled*, meaning that there is a known one-to-one mapping

from vertex V_i in

Formally, assume that each brains, $b \in \mathcal{B}$

4 Old results

Thought experiment 1. *Let the physical property under consideration be brain connectivity structure ("connectome"), so b is a graph (or, network) with vertices representing neurons (or neuroanatomical regions) and edges representing connections between neurons (or white matter tracts). Further let $\mathcal{B}$, the observation space, be the collection of all graphs on a finite number of vertices, and let $|\mathcal{B}|$ be countable. Now, imagine collecting very large amounts of very accurate independent and identically distributed brain-graph data and the associated mental property indicators. A k_n-nearest neighbor classifier using an isomorphism-matching Frobenius norm is universally consistent (see Appendix 1 for proof). Therefore, Theorem 1 applies and 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 mental/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. Furthermore, Appendix 1 also extends the proof to deal with other matrix norms (which might speed up convergence), and the regression scenario, where $|\mathcal{M}|$ is infinite.*

5 Discussion

While the above thought experiment addresses the question of ε-supervenience, it does not address causality. Assuming we have confirmed $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ for a particular mental/brain property pair with confidence level α, then morphing the brain (by altering edges) could be used to determine whether the relation is in fact causal.

Practical issues regarding actually conducting the above thought experiment include: (1) as stated, we must consider the space $\mathcal{B}$ to be the quotient space of graphs mod graph isomorphism, unless the vertices are *labeled*; (2) a more informative and tractable distance on $\mathcal{B}$ may be desired, as the k_n-nearest neighbor classifier under our Frobenius norm (or other norm; see Appendix 1 for details) may have a rate of convergence so slow and a computational demand so high as to be impractical; and (3) collecting enough sufficiently accurate independent and identically distributed brain-graph data and the associated mental property indicators may be beyond current technological capabilities. Regardless, related experimental work includes collecting various types of brain graph data [**?**, **?**, **?**] and various approaches to inference on brain graphs [**?**, **?**, **?**], suggesting feasibility of such an experiment in the near future (see Appendix 2 for a simulated example of a feasible experiment). Nevertheless, our thought experiment suggests that we can hope to determine that a given mental property under consideration ε-supervenes on a brain's connectivity structure. This thought experiment, together with (i) the formal definition of ε-supervenience, (ii) the brain-graph model, and (iii) the universal consistency proof on graphs, is the first *proof* (to our knowledge) that empirically investigating supervenience is at least theoretically possible.

A k_n nearest neighbor algorithm

Assume that b is a real d-dimensional vector, $b \in \mathcal{B} \subseteq \mathbb{R}^d$, and m is a binary indicator, $m \in \mathcal{M} = \{0, 1\}$. Then, further assume that we have observed a collection of training data, $\mathcal{T}_n = \{(b_i, m_i)\}_{i=1}^n$, each sampled identically and independently from some unknown joint distribution, $(b_i, m_i) \overset{iid}{\sim} F_{BM}$. A new brain, b, called the "test brain", is then observed, and one desires to find the most likely class of the new brain, m. The 1-nearest neighbor classifier works as follows. Compute the distance between the test brain and all the training brains, $d_i = d(b, b_i)$ for all $i \in [n]$, where $[n] = 1, 2, \ldots, n$. Then, sort them, $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$, where parenthetical indices, (i), indicate rank order. One can then also obtain a rank order for the training minds, $m_{(1)}, m_{(2)}, \ldots, m_{(n)}$, where $m_{(i)}$ is the class of the i^{th} closest training brain to b. The 1 nearest neighbor algorithm predicts that the unobserved mind is of the same class as the closest brain's class: $m = m_{(1)}$. The k_n nearest neighbor is a straightforward generalization of this approach. It says that the test mind is in the same class as which ever class is the majority class of the k_n nearest neighbors, $m = \mathbb{I}\{\sum_{i=1}^{k_n} m_{(i)} > k_n/2\}$. 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 then has a relatively simple and intuitive algorithm.

Unfortunately, no such algorithm is universally consistent. Let g_n be the k_n nearest neighbor classifier when there are n training points. Then, a collection of such algorithms, $\{g_n\}$, with k_n increasing with n, can be universally consistent under certain constraints. In particular, as n increases, k_n must also increase, but not quite as quickly. Formally, k_n must satisfy: (i) $k_n \to \infty$ as $n \to \infty$ and (ii) $k_n/n \to 0$ as $n \to \infty$. In Stone' original proof, the L_2 norm ($d(b, b') = \sum_{j=1}^{d}(b_j - b'_j)^2$, where j indexes elements of the d-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 [**?**].

This algorithm also readily generalizes to the C-way classification problem, where there $C \geq 2$ classes. Instead of finding the majority class of the k_n nearest neighbors, one instead finds the plurality class, that is, the class with the most training samples within the k_n nearest neighbors. This algorithm generalizes even further, when $\mathcal{M}$ is no longer discrete, but rather continuous, for instance, $\mathcal{M} = \mathbb{R}$. In such a scenario, this approach is called k_n nearest neighbor *regression*.

B k_n-nearest neighbor universal consistency for graphs

Assume first that all graphs are simple (meaning undirected with no loops and binary edges), on the same set of vertices, and that the graphs are labeled so that we know which vertex in one graph corresponds to which vertex in another. Then the Frobenius distance function $d(b_1, b_2)$ can be written in terms of the associated adjacency matrices A_1 and A_2: $d(b_1, b_2) = ||A_1 - A_2||_F$. If the graphs are identical, then $d(b_1, b_2) = 0$, and if the graphs are different, then $d(b_1, b_2) \geq 1$. Since the space $\mathcal{B}$ is finite, n large enough guarantees that with probability approaching unity at least k_n training samples coincide with each atom, so long as $k_n/n \to 0$. Then $k_n \to \infty$ guarantees that the nearest neighbor vote-winner for each atom will eventually coincide with Bayes' choice, yielding universal consistency.

In the foregoing argument, there exists a smallest non-zero atomic probability p_{min}, and "n large enough" is driven by this probability. Generalizing to countable $\mathcal{B}$ with discrete weights, we see that given $\delta > 0$, there is a finite set S with $\mathbb{P}[S] > 1 - \delta$ and smallest atomic probability p_{min}, so that $L_F(g_n) \to c \leq L_F(g^*) + \delta$, yielding universal consistency.

If the graphs may have different numbers of vertices, and are unlabeled, we consider the isomorphism-matching Frobenius norm. Assume without loss of generality that b_1 has at least as many vertices as b_2, and write A_2^P for the adjacency matrix associated with b_2 "padded" to include extra isolated vertices so that A_2^P is the same size as A_1. Then $d(b_1, b_2) = \min_Q ||QA_1Q^T - A_2^P||_F$ where the minimum is taken over all permutation matrices [**?**]. Under the equivalence relation induced by this isomorphism-matching, the foregoing universal consistency argument holds.

Several points of note: isolated vertices are ignored in our equivalence relation; the class-conditional signal is entirely encompassed by the connectivity structure; the graph isomorphism problem is computationally hard [**?**, **?**]; and the argument employed here does not capture the concept of "nearness implies likelihood of similar class"—we simply rely on atomic behavior.

Finally, the above proof can be straightforwardly generalized to utilize any matrix norm, $\|b_1 - b_2\|_A = \|A(b_1 - b_2)\|$, assuming that AA^{T} is positive definite (see [**?**] pg. 455 for proof when b_i's are vectors). Furthermore, we can relax the constraint that mental properties are finite, that is, we can allow any $|\mathcal{M}| = \infty$ (see [**?**] for proof when b_i's are vectors). The proof for the case when b is an adjacency matrix follows immediately upon first defining a bijective embedding of the matrix into a vector space, and then embedding each adjacency matrix in that space.

C Simulation

As an example of a feasible experiment, one may consider a species whose nervous system consists of the same (small) number of labeled neurons for each organism. *Caenorhabditis elegans* is believed to be such a species [**?**]. The hermaphroditic C. elegans' somatic nervous system consists of 279 interconnected neurons. While the graph with these neurons as vertices and edges defined by chemical synapses between neurons is not identical across individuals, it is reasonably consistent [**?**]. Furthermore, these animals exhibit a rich behavioral repertoire that depends on circuit properties [**?**]. Thus, one may design an experiment by describing the joint distribution F_{BM} via class-conditional distributions $F_{B|M=m_j}$ for the C. elegans brain-graph for two mental properties of interest, m_0 and m_1, along with the prior probability of class membership $\mathbb{P}[M = m_1]$. Here the mental property corresponds to the C. elegans exhibiting or not exhibiting a particular behavior (e.g., response to an odor).

Simulations suggest that one may build a classifier, practically and with a manageable training sample size n, that demonstrates ε-supervenience with reasonable choices for ε and α and a plausible joint distribution F_{BM} (Figure 2). To generate the data, we let the class-conditional random variable $E_{ij}|M = m_0$ be distributed Poisson$(A_{ij} + \eta)$, where A_{ij} is the number of chemical synapses between neuron i and neuron j according to [**?**], with noise parameter $0 < \eta \ll 1$. The class-conditional random variable $E_{ij}|M = m_1$ is distributed Poisson$(A_{ij} + z_{ij})$ for neurons $i, j \in \mathcal{D}$, where $\mathcal{D}$ is the set of edges deemed responsible for odor-evoked behavior according to [**?**], with signal parameter z_{ij} uniformly sampled from $[-5, 5]$. 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_{BM} used here; however, we demand universal consistency.)

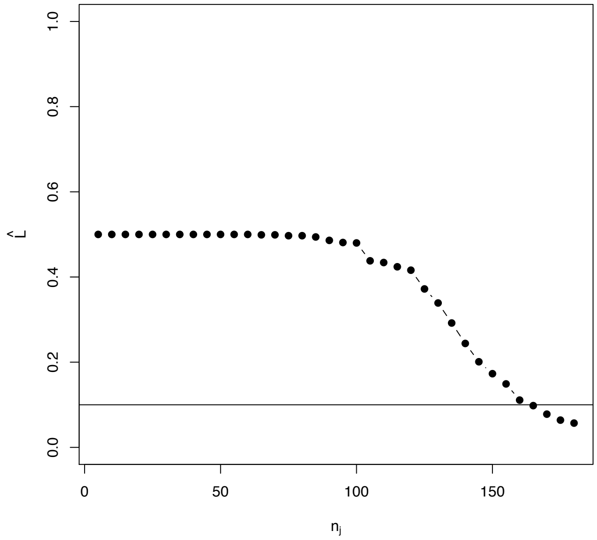

Figure 2: C. elegans graph classification simulation results. $\widehat{L}_F^{1000}(g_n)$ is plotted as a function of class-conditional training sample size n_j, suggesting that for $\varepsilon = 0.1$ we can determine that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ holds with 99% confidence with just a few hundred training samples generated from F_{BM}. Each dot depicts an estimate for $L_F(g_n)$; standard errors are $(L_F(g_n)(1 - L_F(g_n))/1000)^{1/2}$; e.g., $n_j = 180$; $k_n = 53$; $\widehat{L}_F^{1000}(g_n) = 0.057$; standard error less than 0.01. We reject $H_0 : L_F(g^*) \geq 0.10$ at $\alpha = 0.01$. $L_F(g^*) \approx 0$ for this simulation.

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D superresolution imaging [**?**] combined with neurite tracing algorithms [**?**, **?**, **?**] allow the collection of a 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 $M = m_1$ [**?**], and the class of each organism (m_0 vs. m_1) can also be determined automatically [**?**].

D Acknowledgments

The authors would like to acknowledge helpful discussions with J Lande and B Vogelstein.