brainSCANr: Mapping brain structure, function, and disease relationships with PubMed

Jessica B. Voytek[1] & Bradley Voytek[2]

1. School of Information, and

2. Helen Wills Neuroscience Institute, University of California, Berkeley, USA

Correspondence should be addressed to J.V. (jessica.voytek@gmail.com) or B.V. (bradley.voytek@gmail.com).

Currently, PubMed contains more than 18 million peer-reviewed articles with approximately 40,000-50,000 more added monthly. Modern scientific research stands on the shoulders of countless giants, yet integrating millions of articles is impossible for any researcher. In an attempt to simplify research and knowledge discovery, we populated a database of associations between neuroscientific concepts built from scanning over 3.5 million scientific abstracts indexed in PubMed. From these associations we recreated a whole-brain connectivity graph that enabled us examine relationships between brain function, structure, and disease. Furthermore, we created a website to visualize these relationships, synthesize the breadth of neuroscientific research, and provide a resource for computational modeling. The future of science will rely upon automated, intelligent algorithms to aid scientists in integrating the overwhelming volume of scientific literature. Here we take an important step toward that goal in a manner that is broadly generalizable to any scientific field.

The scientific method begins with a hypothesis about our world that can then be experimentally tested. Hypothesis formation is iterative, always building off prior knowledge. Thus, testing a hypothesis requires a thorough understanding of previous research to ensure that the path of inquiry is founded upon a stable base of established facts. But how can a researcher perform thorough background research when over one million scientific articles are published annually[1]? The rate of scientific discovery has quickly outpaced our ability to integrate even the newest findings in an unbiased, principled fashion. One solution may be via automated information aggregation[2]. In this manuscript we show how scanning for associations in the peer-reviewed scientific literature can discover scientific relationships in neuroscience.

Neuroscience is a complex discipline that integrates biophysics, chemistry, genetics, engineering and computer science, mathematics, psychology, philosophy, medicine, and many other fields[3]. The aim of neuroscience is to understand relationships between brain, behavior, and disease. And yet, despite millions of scientific publications, much about the brain remains unknown. No one researcher or research group can hope to unify all of the relevant information into a coherent framework; understanding the brain will require the integration of massive amounts of facts and data. In this paper we show that the literature contains a vast amount of connected facts that, by definition, recapitulate known neuroscientific relationships. Neuroanatomical, behavior, and disease associations can be rapidly quantified and visualized to speed research and education or to find understudied research paths[4]. Rather than allowing our limited ability to thoroughly review the entire scientific literature bias our hypotheses, we can algorithmically integrate across millions of scientific research papers.

To facilitate this process, we have created a new resource to aid neuroscientists in visualizing the relationships between brain structure, function, and disease: the Brain

Systems, Connections, Associations, and Network Relationships engine (brainSCANr: http://www.brainscanr.com). We use a simple algorithm to calculate the pair-wise association index (AI) between neuroscientific terms (see **Methods**). To accomplish this, we scanned more than 3.5 million papers indexed in PubMed (http://pubmed.gov) for the co-occurrence of pairs of neuroscientific terms (and their synonyms) with the assumption that the more frequently terms appeared together across the titles or abstracts of manuscripts, the stronger their association was likely to be. That is, we assumed an underlying structure within the peer-reviewed neuroscientific literature.

Based only upon co-occurrences in the literature, we show that we can recreate known neuroanatomical connectivity using blind, automated clustering algorithms. These algorithms identify physically diffuse neuroanatomical circuits such as subcortical-cortical visual pathways, brainstem auditory nuclei, and even cognitive/behavioral circuits involved in speech and other higher-level cognitive processes. We can analyze the growth of knowledge across decades of research by limiting the time windows across which we examine published data. Furthermore, by clustering cognitive functions, we can quantify the relationship between a variety of cognitive tasks commonly used in neuroscientific research, such as the relationships between tasks used to study executive functioning or cognitive control. Finally, clustering across all brain structures, functions, and diseases, we can simplify the complexity of the neuroscientific literature to highlight important relationships, or, conversely, to discover less common relationships to aid in scientific discovery. Despite certain limitations and assumptions of our method, our resource represents a novel, intuitive tool that will speed research and education and may allow for the discovery of new, previously unforeseen connections between brain, behavior, and disease. Although we only analyzed relationships within the neuroscientific literature, our method can be used for any scientific or medical field.

RESULTS

Structural associations. As can be seen in **Figure 1** and **Supplementary Figure 1**, we can build a full connectivity graph for the entire brain, limited only by the dictionary used to define the search terms (see **Supplementary List 1** for the full list of terms and their synonyms used in this manuscript). Although the morphology of this graph changes slightly depending on the term used to define the "center" of the graph, the association weights do not change (for **Figure 1**, the first term in our alphabetically-sorted database, "abducens nucleus", forms the center). Within our original database of 124 brain regions, we find associations between all terms except for a few seemingly orphan regions. Although these regions are, of course, not truly unconnected, the strength of their pair-wise associations is relatively weak. For visualization purposes, we classified each brain region as belonging to one of 8 pre-defined macroscale clusters and colored each node according to its group membership. This coloring highlights the clustering of certain groups of brain regions, wherein cortical regions form distinct groups farther from the central brainstem structures while thalamic and basal ganglia structures cluster together nearer the brainstem.

While the graph in **Figure 1** represents the inferred brain connectivity network based upon PubMed co-occurrences, we can also examine the change in relative network associations over time by limiting our search query to specific time windows. In **Figure 2** we show how the early literature (up to 1905) showed only 8 nodes with a few edges connecting the related brain regions. By 1930 the number of nodes had tripled to 24, had increased again to 44 nodes by 1955, and hit the maximum of about 121 nodes by 1985. The number of nodes added to the graph began to rise earlier and faster than the number of edges connecting them until about 1970, at which point there was a rapid increase in both the number of brain regions and edges added to the network

(as can be seen in the rate of change plot in **Figure 2**). Interestingly, since its peak sometime around 1995, the number of edges has been slowly *decreasing* while the number of nodes has remained stable, suggesting a "pruning" of the connectivity network.

Given the association weights between each brain region, we can also blindly cluster structures based upon their interconnectivity (see **Figure 3**). In **Supplementary Table 1**, we outline the clusters identified in this manner, as well as the individual members that form each cluster. These clusters—defined only by their PubMed associations—recapitulate known anatomical circuits. Several of these circuits are quite anatomically diffuse; for example, one circuit, which we identify as a "visual" circuit, associates the lateral geniculate nucleus and pulvinar, the superior colliculus, and primary visual cortex and visual extrastriate. We also observe clusters of brainstem auditory and prosencephalic auditory circuits (cochlear nuclei, superior olive, and trapezoid body; inferior colliculus, medial geniculate body, and primary auditory cortex) as well as oculomotor nuclei (abducens nucleus, interstitial nucleus of Cajal, oculomotor nucleus, and trochlear nucleus). These results show that there is an inherent structure to the peer-reviewed literature that can be blindly recovered.

Functional and disease associations. Just as we can indentify clusters of associated brain regions, we can also cluster functions or diseases (see **Figure 3**). Conceptually, identifying clusters of cognitive and behavioral functions provides a quantification of how closely related two cognitive tasks or behavioral states may be[5]. For example, "visual working memory" is related to "delayed match to sample" tasks[6,7] (AI = 0.0085) but very weakly related with "stress" (AI = 0.0000061). Interestingly, as can be seen in **Supplementary Table 2**, there are two clusters of diverse tasks that we identify as an "executive functions" cluster and a "monitoring and control" cluster. The former contains 9 tasks such as the Stroop, Wisconsin card sorting, and digit span tasks, as well

as verbal memory and visual memory. The latter cluster contains tasks such as the go/no-go, stop signal, and antisaccade tasks.

In **Supplementary Table 3** we show clusters of diseases identified from their associations. This creates an apparent cluster of psychiatric disorders, including bipolar disorder, schizophrenia, and obsessive-compulsive disorder, as well as a cluster of agnosias such as Broca's and Wernicke's aphasia, apraxia, and prosopagnosia. While these classifications provide important data on within-category clustering, by including all structural, functional, and disease terms in one, large association matrix, we can calculate cross-category clusters to identify structure/function/disease relationships. For this larger analysis we also included 7 neuroimaging methodological terms in order to examine whether there are any methodological biases in neuroscientific research. These terms were: diffusion imaging, EEG, fMRI, MEG, PET, single- and multi-unit electrophysiology, and TMS (see **Supplementary List 1** for the full list of terms and their synonyms).

In **Supplementary Table 4** we show clusters of cross-category associations. For example, we observe cross-category relationships for language functioning that includes "language comprehension", "Wernicke's area", and "Wernicke's aphasia", along with other related terms. We also find a Parkinson's disease cluster containing "Parkinson's disease", "caudate nucleus", and "substantia nigra", and other associated terms. There is also a learning and reward cluster that groups the prefrontal cortex, ventral tegmental area, reward, association learning, and other related terms. Although exploratory, this cross-category clustering is useful for integrating and unifying the complex interrelationships across a broad range of neuroscientific results.

Given that *known* relationships can be captured automatically lends support to the idea that there may be *unknown* or understudied relationships that can be recovered

from the literature. For example, in **Figure 4** we show that "hippocampus" is strongly related to "spatial memory" (AI = 0.019; 1966 co-occurrences in PubMed) and "striatum" (AI = 0.060; 8159 co-occurrences). In contrast, "spatial memory" and "striatum" are weakly related (AI = 0.0019; 101 co-occurrences). Given that "hippocampus" and "spatial memory" are strongly associated, and "hippocampus" and "striatum" are strongly associated, it is surprising that there is a scarcity of literature on the association between "spatial memory" and "striatum". This lack of association may be due to a publication bias wherein null results go unreported, and as such spatial memory is unrelated to the striatum, or there may be a true association between the two that is under-examined. Given our association matrix, the process of uncovering potentially new research paths could be semi-automated to speed knowledge discovery.

Website. The previous analyses were performed offline and the database used to create these analyses is available for download. However, we also have created a website to allow users to create simple structure/function/disease association visualizations based upon the terms currently in our database. In **Figure 4** we show an example of the website's search results page. In this example, we show the results obtained by searching for "hippocampus". By default, the top 20 strongest associations are graphed. Blue lines illustrate the associations between the search term and its 20 strongest associations; grey lines illustrate relationships between all other terms. Line width represents the relative strength of the association. Below this main visualization the user can also see a quantification of each association between the search term and all other terms in the database. The user can also view a selection of manuscripts that associate the search term with any other selected terms.

Other applications. We remark upon the associations between neuroimaging research methodologies and the other terms in our database. We observe the strongest mutual relationship for fMRI is with frontal cortical regions, the insula, and the cingulate, as

well as with working memory, all of which are described thoroughly in the literature[8,9]. For single- and multi-unit recording techniques, we find a strong bias toward research on primary cortices (auditory, motor, and visual), the subthalamic nucleus and inferior colliculus, and the ventral tegmental area and substantia nigra, as well as studies of eye movements and saccades. Such analyses provide quantification of how each neuroimaging method is used, and offers insight into their relative biases, strengths, and weaknesses.

Although the current analysis is restricted to a limited dictionary of structure, function, and disease terms, the association and visualization methods we use are applicable to any search term or phrase found in PubMed. Thus, future upgrades to the website will allow users to search for many more neuroscientifically relevant terms such as neurochemicals, proteins, genes, white matter tracts, drugs, or even specific research scientists.

DISCUSSION

In this manuscript we demonstrate that, by mining the scientific literature indexed via PubMed for associations between neuroscientific terms, we can recapitulate known neuroanatomical and structure/function/disease relationships, as well as highlight potentially under-examined research paths. We have incorporated our method and results into a simple, web-based user interface that provides researchers, educators, and the lay public with an opportunity to easily and intuitively explore the complexities of the neuroscientific literature. While other projects have been conducted for genetics and protein analyses[10-13], as well as neuroscientific projects to scan the literature for mapping terms onto brain regions (http://pubbrain.org/), ours is the first site to integrate topics and provide cross-domain relationships.

There is currently a massive scientific effort underway to identify the human connectome[14-16]. While connectomics aims to map the interconnectivity of every neuron or neuronal group, it is unclear as to what the preferred "granularity" is for cognition. That is, what is the scale of the functional groups that operate to give rise to cognition: sub-cellular, neuronal, columnar, or diffuse networks[17,18]? Even at the relatively macroscopic scale of systems and networks, the intricacies of neuroanatomical interconnectivity and how specific brain regions give rise to cognition and relate to disease are difficult to comprehend and visualize. Often these connectivity data are spread across dozens of research manuscripts, brain atlases, websites, and other repositories. While fields such as genetics have put great effort into ontological projects[10], the adoption of ontologies for neuroanatomy and cognition has been slow (but see Bowden, Dubach, & Park[19] and CoCoMac[20]). Nevertheless, we can leverage the power of millions of publications to bootstrap informative relationships[21]. By mining existing relationships between neuroscientific concepts from the published literature we can recover the assumed connectome as well as brain/function/disease relationships. That is, we can add another layer of intelligent automation to the scientific method workflow as has already been show possible for the data modeling stage[22].

Understanding the underlying anatomy of the human brain is critical for answering fundamental questions of human cognition. That is, our neuroanatomy acts as a real world constraint on our neuroscientific models, and any biologically valid model must adhere to neuroanatomical rules. Of course, there are limitations to the method we used to create these association networks. For example, relationships are inherently based upon the existing literature, and thus the observed associations may reflect publication biases rather than necessarily true associations. Also, our method does not differentiate negative from positive results such that, if a paper's title or abstract states that the amygdala does *not* relate to fear, that paper is weighted as strongly as a paper that finds a positive relationship between the amygdala and fear. Nevertheless, there is a

well-described publication bias in the literature such that negative results are under-reported[23-26], and thus the relationships we discover are more likely to be biased toward positive associations, though that bias cannot be quantified.

Our method is also susceptible to occasional spurious relationships introduced by English homographs. For example, one of our functional terms ("rhythm") is most strongly associated with the structure "suprachiasmatic nucleus" and method "electroencephalography". Considered as a behavioral or cognitive function, "rhythm" refers to the perception of musical rhythm. However, instead of highlighting the musical relationships of rhythm, our method exalts the co-occurrence of the terms "suprachiasmatic nucleus" and *circadian* rhythm, and "electroencephalography" with *oscillatory electrophysiological brain rhythms*. We also are unable to differentiate animal species in our search results, and thus we are aggregating results across research from all species for which our dictionary terms apply. Because the full association matrix is required for visualization and analysis, brainSCANr searches are limited solely to terms currently in the database. However we provide all of our data for researchers interested in exploring putative neuroanatomical relationships, and our resource would prove very useful for modeling information flow across macroscale brain regions. Despite the limitations of our approach, our associations map onto known relationships with remarkable accuracy, and the simplicity of our design allows us to expand upon and improve the website as needed.

METHODS

Starting dictionary. We initially populated the dictionary with phrases for 124 brain regions, 291 cognitive functions, and 47 diseases. Brain region names and associated synonyms were selected from BrainInfo (2007), Neuroscience Division, National

Primate Research Center, University of Washington (http://www.braininfo.org)[27]. Cognitive functions we obtained from (http://www.cognitiveatlas.org/)[3]. Disease names are from (http://www.ninds.nih.gov/). The initial population of the dictionary was meant to represent the broadest, most plausibly common search terms that are also relatively unique (and thus likely not to lead to spurious connections). The full list of terms and their synonyms are included in **Supplementary List 1**.

Association index. We quantified the association between two terms using a weighted calculation that highlights the unique relationship between term pairs; a co-occurrence algorithm[11]. For any given pair of terms *i* and *j*, we define the association $U_{i,j} = c_{i,j} / d_{i,j}$, where $c_{i,j}$ (the conjunction) is the number of papers containing both terms (and their synonyms) and $d_{i,j}$ (the exclusive disjunction) is the total number of papers containing (*i* not *j*) and (*j* not *i*). For each pair of words the number of papers for the conjunction and each disjunction was separately queried from the PubMed database using the ESearch utility and the *count* return type. For example, to calculate *U* for the terms "prefrontal cortex" and "striatum", the following searches were performed:

- http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&field=word&term=("prefrontal+cortex"+OR+"prefrontal+cortices")+AND+("striatum"+OR+"neostriatum"+OR+"corpus+striatum")&rettype=count

- http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&field=word&term=("prefrontal+cortex"+OR+"prefrontal+cortices")+NOT+("striatum"+OR+"neostriatum"+OR+"corpus+striatum")&rettype=count

- http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&field=word&term=("striatum"+OR+"neostriatum"+OR+"corpus+striatum")+NOT+("prefrontal+cortex"+OR+"prefrontal+cortices")&rettype=count

Note that in these searches that all synonyms for a given term are included within the parentheses and that each term is individually surrounded by quotation marks. This limits the search to each exact phrase, without any intelligent parsing by PubMed. Furthermore, the "field=word" modifier limits the search to just the article text. This reduces instances of false associations due to name or journal title homographs (*e.g.*, author name "Fear" or journal "Language" as opposed to the functions "fear" and "language").

Website creation. The brainSCANr website was created using the Google App Engine (Google, Inc.) framework. Graph connectivity plotting was performed using the JavaScript InfoVis Toolkit (Nicolas Garcia Belmonte, http://thejit.org/). Website users can download the full association database included in our study for offline analyses.

Graphical visualization and clustering. For **Figures 1** and **2**, graphs were plotted using the GraphViz (AT&T Research Labs) radial plot function. Clustering was performed using an iterative (*k*-means) clustering algorithm (MATLAB® R2009b, Natick, MA; *kmeans.m*). For the brain structure and functions analyses, we used 20 clusters, and for the disease analysis we used 5 clusters.

1. Björk, B., Roos, A., & Lauri, M. Global annual volume of peer reviewed scholarly articles and the share available via different Open Access options. *Proceedings of the Conference on Electronic Publishing–Toronto* **2**, (2008).

2. Chen, C., Ibekwe‐SanJuan, F., & Hou, J. The structure and dynamics of cocitation clusters: A multiple‐perspective cocitation analysis. *J. Am. Soc. Inf. Sci. Technol.* **61**, 1386-1409 (2010).

3. Yarkoni, T., Poldrack, R.A., Essen, D.C.V., & Wager, T.D. Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends Cogn. Sci.* **14,** 489-496 (2010).

4. Wren, J., Bekeredjian, R., Stewart, J., Shohet, R., & Garner, H. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* 4211 (2004).

5. Poldrack, R., Halchenko, Y., & Hanson, S. Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychol. Sci.* **20**, 1364-1372 (2009).

6. Voytek, B., & Knight, R.T. Prefrontal cortex and basal ganglia contributions to visual working memory. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 18167-18172 (2010).

7. Voytek, B., *et al*. Dynamic Neuroplasticity after Human Prefrontal Cortex Damage. *Neuron* **68,** 401-408 (2010).

8. Curtis, C.E., & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* **7,** 415-423 (2003).

9. Gazzaley, A., Rissman, J., & D'Esposito, M. Functional connectivity during working memory maintenance. *Cogn. Affect. Behav. Neurosci.* **4,** 580-599 (2004).

10. Zhang, Y., *et al*. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Med. Genomics.* **3,** 1 (2010).

11. Alako, B.T., *et al*. CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, **6,** 51 (2005).

12. Frijters, R., *et al*. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput. Biol.* **6,** e1000943 (2010).

13. Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., & Leser, U. AliBaba: PubMed as a graph. *Bioinformatics* **22,** 2444-2445 (2006).

14. Sporns, O., Tononi, G., & Kötter, R. The Human Connectome: A Structural Description of the Human Brain. *PLoS Comp. Biol.* **1,** e42 (2005).

15. Modha, D., & Singh, R. Network architecture of the long-distance pathways in the macaque brain. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 13485-13490 (2010).

16. A critical look at connectomics. *Nat. Neurosci.* **13**, 1441 (2010).

17. Bohland, J., *et al*. A Proposal for a Coordinated Effort for the Determination of Brainwide Neuroanatomical Connectivity in Model Organisms at a Mesoscopic Scale. *PLoS Comp. Biol.* **5,** e1000334 (2009).

18. Varel, F., Lachaux, J.P., Rodriguez E., & Martinerie, J. The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* **2,** 229-39 (2001).

19. Bowden, D.M., Dubach, M., & Park, J. Creating neuroscience ontologies. *Methods Mol. Biol.* **401,** 67-87 (2007).

20. Stephan, K.E., Hilgetag, C.C., Burns, G.A., O'Neill, M.A., Young, M.P., & Kötter R. Computational analysis of functional connectivity between areas of

primate cerebral cortex. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **355**, 111-126 (2000).

21. Michel, J.B., *et al*. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* (in press).

22. Schmidt, M., & Lipson, H. Distilling Free-Form Natural Laws from Experimental Data. *Science* **324**, 81-85 (2009).

23. Begg, C.B., & Berlin, J.A. Publication bias and dissemination of clinical research. *J. Natl. Cancer Inst.* **81,** 107-115 (1989).

24. Dirnagl, U., & Lauritzen, M. Fighting publication bias: introducing the Negative Results section. *J. Cereb. Blood Flow. Metab.* **30,** 1263-1264 (2010).

25. Ioannidis, J.P., Cappelleri, J.C., Sacks, H.S., & Lau, J. The relationship between study design, results, and reporting of randomized clinical trials of HIV infection. *Control. Clin. Trials* **18,** 431-444 (1997).

26. Stern, J.M., & Simes, R.J. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *B. M. J.* **315,** 640-645 (1997).

27. Bowden, D.M., & Dubach, M.F. NeuroNames 2002. *Neuroinformatics* **1,** 43-59 (2003).

ACKNOWLEDGEMENTS

We thank Curtis Chambers for technical assistance, and Leon Deouell, Amitai Shenhav, Avgusta Shestyuk, Kirstie Whitaker, and many brainSCANr beta testers for technical discussions. B.V. is funded by the American Psychological Association Diversity Program in Neuroscience (5-T32-MH18882) and NINDS grant NS21135-S1.

AUTHOR CONTRIBUTIONS

B.V. conceived of the methods and analyzed the data; J.V. analyzed the data and created the website; both co-authors wrote the manuscript.

COMPETING INTERESTS STATEMENTS

The authors declare that they have no competing financial interests.

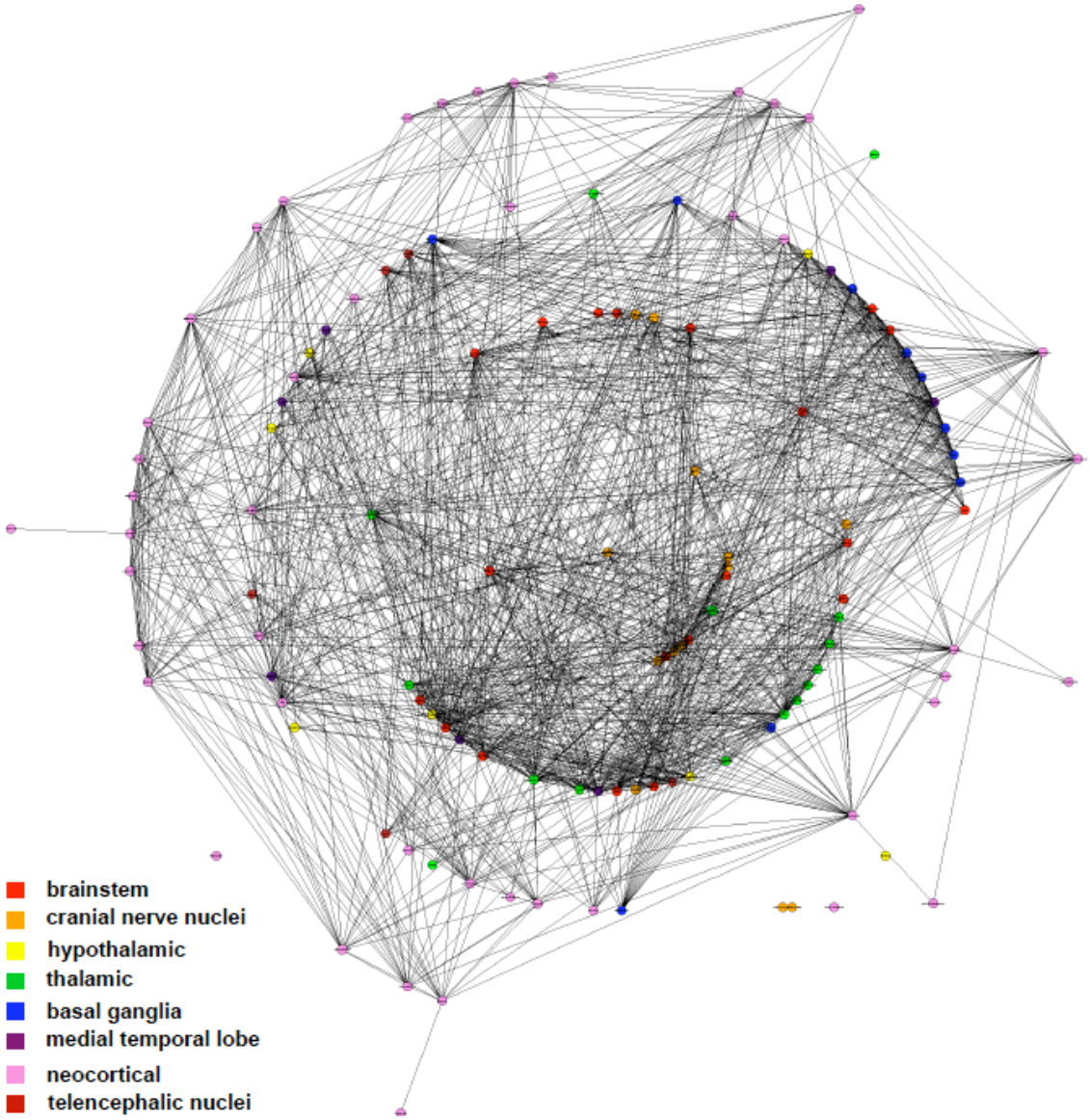

Figure 1 Inferred brain connectivity graph. Based upon a pre-defined dictionary of 124 brain regions and their 703 synonyms, we calculated the probability of association between all pairs of brain regions based upon their co-occurrence in the scientific literature indexed via PubMed. This method recovers known neuroanatomical relationships (see **Supplementary Table 1**). In the center rings, brainstem structures cluster together, with telencephalic/neocortical structures arranged in the outside rings. Note the clustering of thalamic and basal ganglia structures in the middle rings. Graphic visualization was performed using GraphViz (AT&T Research Labs) with a connectivity threshold of 0.095.

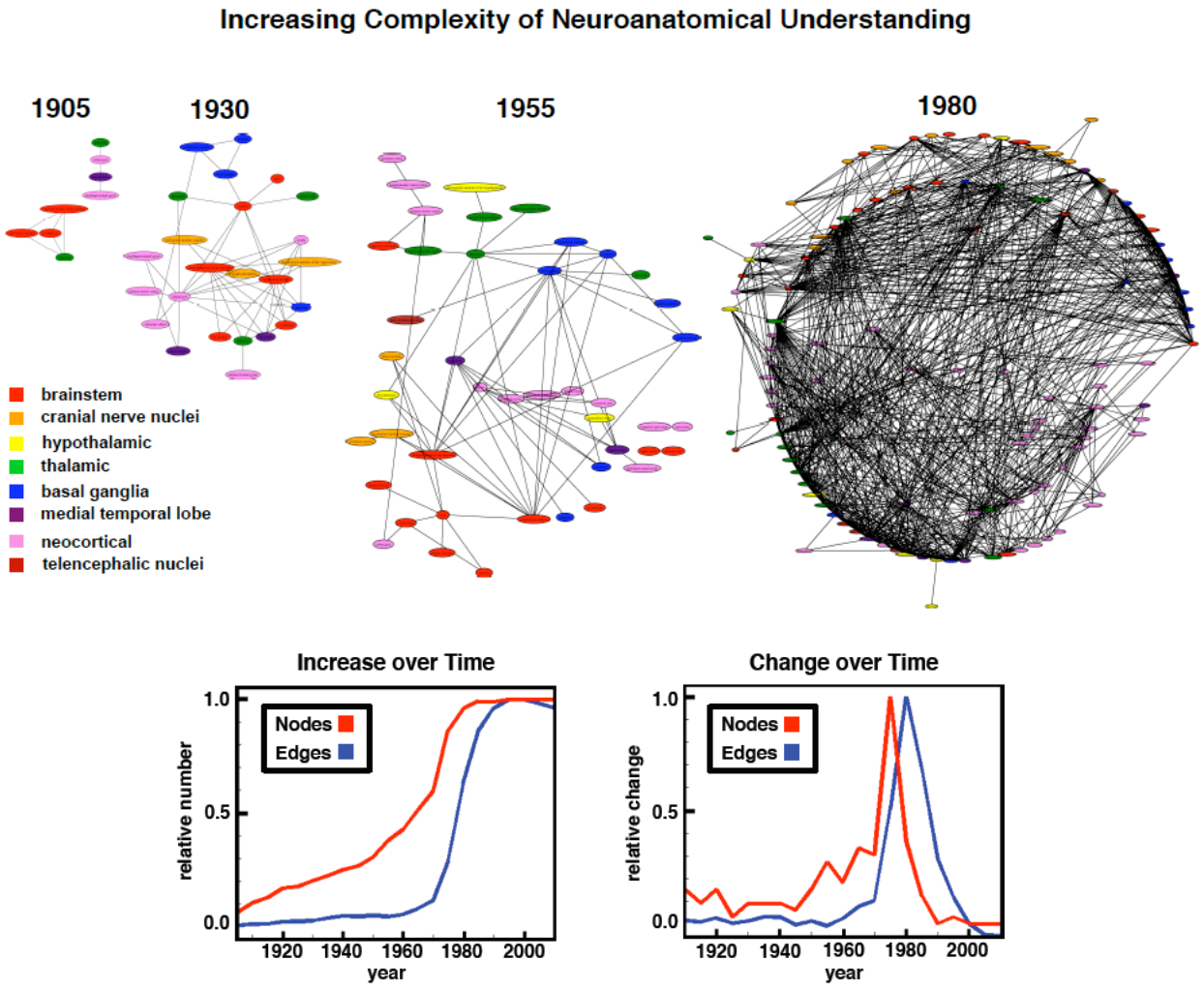

Figure 2 Neuroanatomical knowledge gains over time. These association networks illustrate how scientific understanding of relative neuroanatomical relationships has increased since 1900. The line plots below show that the number of nodes increased steadily from 1900 through approximately 1990, whereas the edges associating those nodes remained stable until about 1970, after which there was a dramatic, 20-year increase. Since 1990 there has been a slow pruning of edges despite the relative stability of the number of nodes. The left plot shows the relative number of nodes and edges in the network integrated across time. The plot on the right represents the rate of change in the number of nodes and edges. Graphic visualization was performed using GraphViz (AT&T Research Labs) with a connectivity threshold of 0.005.

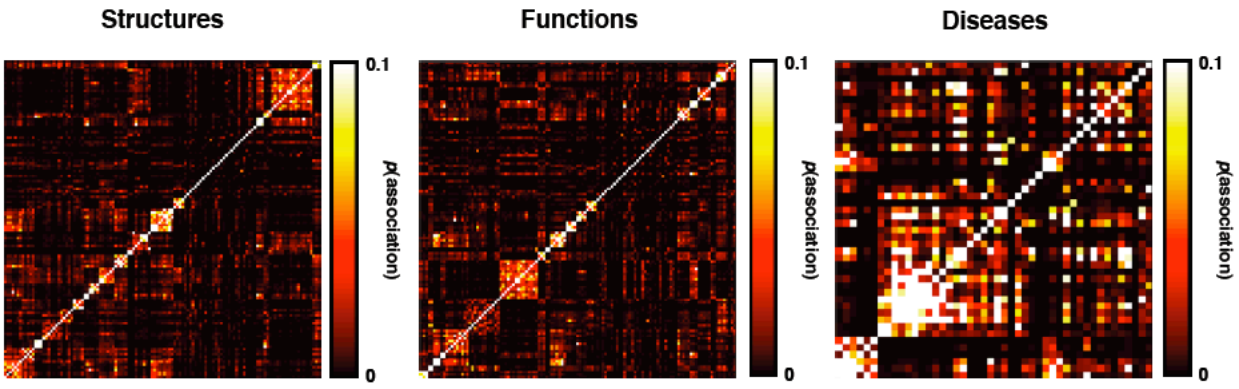

Figure 3 Association matrices for structure, function, and disease relationships. Each row/column index represents the probability of association between two terms as calculated from PubMed. For each matrix, data are sorted according to the clusters identified via *k*-means clustering using 20 structure clusters, 20 function clusters, and 5 disease clusters. This method highlights several within-cluster associations along the diagonal.

p(association): weighted probability of association between two phrases (association index)

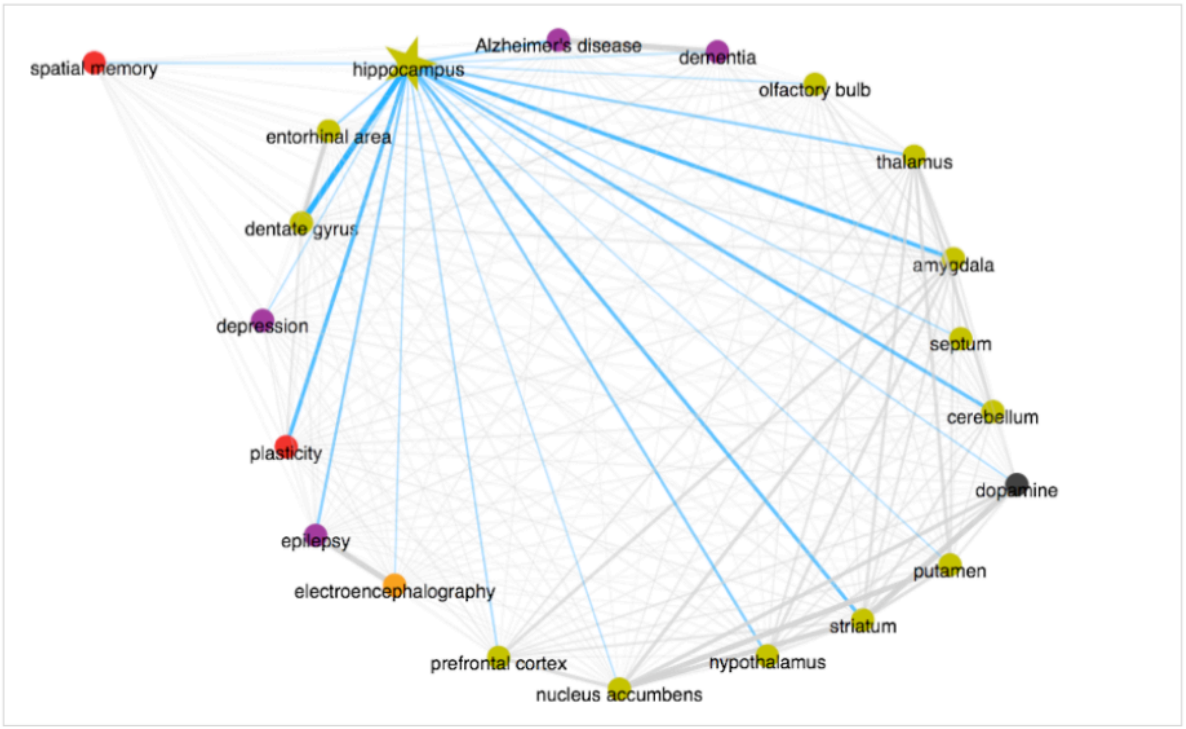

Figure 4 Example search results from brainSCANr.com. The user enters a search term from our database in the search box (in the example above, “hippocampus”). This search returns a graphic visualization of the resulting associations between the hippocampus (and its synonyms) and the top 20 associated terms. The hippocampus is associated with a variety of brain regions, as well as certain pathologies such as epilepsy and Alzheimer’s disease. However, the association with “depression” is likely confounded by the English homograph for depression as a psychological state and “long term depression”, a physiological phenomenon that is heavily studied in hippocampal neurons. Of interest is the fact that “spatial memory” is strongly associated with the hippocampus, but weakly associated with other hippocampal-related terms. This may highlight an understudied research association for a role of the cerebellum, striatum, etc. in spatial memory. Blue lines illustrate connections between the search term and other terms; gray lines are connections between the other (non-searched) terms. Line width represents the relative strength of association.