


The data science future of neuroscience theory

Bradley Voytek

 Check for updates

An approach for integrating the wealth of heterogeneous brain data – from gene expression and neurotransmitter receptor density to structure and function – allows neuroscientists to easily place their data within the broader neuroscientific context.

Imagine we conduct a human neuroimaging experiment that shows that neural activity in Brodmann area 46 scales with the demands of the task the participants performed, and that the activity is statistically higher in older than in younger adults. What do we do with this information? An expert might guess that this was a working memory or attention task, on the basis of their familiarity with the relevant neuroimaging literature. But that expert, like myself, may know relatively little about neurogenetics, neurochemistry or cortical evolution and, because of their (and my) own knowledge gaps, will be unable to see how those results fit into the broader neuroscientific context. If we are to develop anything resembling a set of theories for how the brain gives rise to behavior, cognition and disease, we need a way for placing

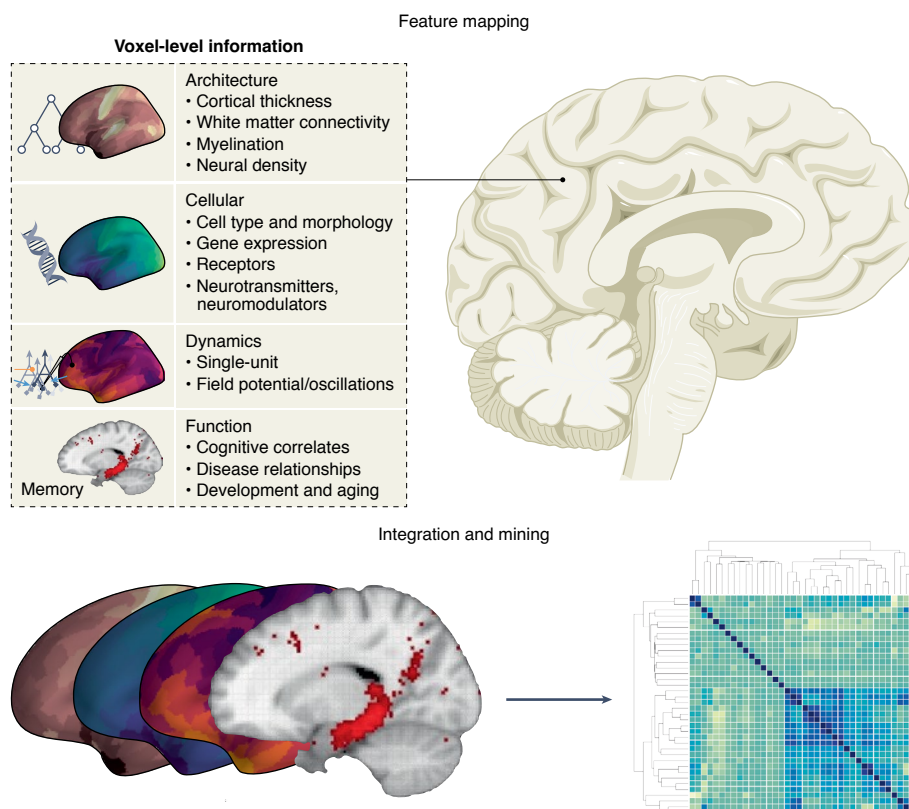
human neuroimaging data within that greater context. The open-source Python package *neuromaps*, introduced by Markello, Hansen and colleagues in this issue of *Nature Methods*¹, provides an invaluable tool for doing just that.

Given the wealth of information published in neuroscience, it is impossible for a research group – let alone a single scientist – to understand how age-related changes in neurogenetics, receptor densities, cell types, structural connectivity and so on might all contribute to the neural activity differences observed in our above hypothetical experiment. Thus, our inability to integrate these many different facets of the brain is understandable and even excusable. However, this separation of data from knowledge has often led to neuroscience being described as data rich but theory poor: every research group is collecting more and more data, unable to fit all these different pieces together as a unified whole.

To accomplish this feat of contextualization, Markello and Hansen et al. solved several technical problems to allow easy data integration. First, *neuromaps* offers functions for easily transforming between four standard human brain template atlases: fsaverage, fsLR, CIVET and MNI-152. Armed with this ability to transform between different coordinate systems, Markello, Hansen et al. then sought out numerous heterogeneous datasets, placing them into the same coordinate

Fig. 1 | The potential of the neuromaps package.

For any arbitrary region of the human brain there exists a wealth of knowledge about its various features: architecture, cellular composition and genetics, neural dynamics, and functions (top). The *neuromaps* package collates and aligns these features to facilitate comparisons across domains, opening new avenues for data mining across features and potentially allowing for novel forms of data-driven, semiautomated hypothesis generation (bottom).



system. We can group the included datasets into four broad categories: architecture, cellular, dynamics and function (Fig. 1):

- Architecture: cortical thickness; MRI T1-weighted/T2-weighted ratio; functional connectivity; intersubject variability
- Cellular: gene expression from the Allen Human Brain Atlas; tracer images from neurotransmitter receptor positron emission tomography
- Dynamics: canonical magnetoencephalography frequency bands; intrinsic neuronal timescale; glucose and oxygen metabolism; cerebral blood flow and volume
- Function: evolutionary expansion; developmental expansion; Neurosynth-derived functional maps

Placing all these different datasets in the same spatial framework then allows researchers to statistically compare their spatial profiles to ask novel research questions. Importantly, however, these spatial maps cannot be compared using simple correlations, because the topography of neural data is spatially autocorrelated. That is, features that are closer together in space are inherently likely to look more similar than features that are farther apart. The fact that nearly all neural data are spatially autocorrelated practically guarantees that correlations will be found between neural data maps, just as maps of human population density are strongly correlated with maps of economic activity – technically correct but not mechanistically insightful. To address this, Markello, Hansen et al. include several approaches for performing spatial permutations for significance testing in neuromaps, to help minimize the inflated *P*-values that result from spatial autocorrelations.

To test the potential for heterogeneous data integration, Markello, Hansen et al. then leveraged multiple open datasets to run several comparisons. For example, using an open dataset of cortical thinning in patients with chronic schizophrenia, they showed that the spatial topography of cortical thinning is greatest in regions that show the greatest neurodevelopmental expansion, as quantified from an entirely different set of independent data. Using other datasets, they also showed that the brain regions that have the greatest evolutionary expansion also have the greatest interindividual variability in regional functional connectivity.

Markello, Hansen et al. take a pure data science approach to neuroscience with neuromaps, integrating heterogeneous data types and bringing them together into a unified framework. These data types include categorical gene expression data, time-series oscillation and timescale data, functional data derived from text mining of the neuroscience literature, and functional and structural connectivity graphs. A common refrain in what limits the development of a systems-level mechanistic understanding of behavior and cognition is the difficulty in bridging between scales – between genetics and structure² or between dynamics and function³. If we could better bridge these scales, it is argued, we could better reveal key neuroscientific insights. It is exciting that neuromaps allows us to place results from one domain – such as the functional neuroimaging example from above – into the greater context of neural architecture, cellular composition, dynamics and function.

What does neuromaps portend for the future of neuroscience? Are more data and better data integration enough? Algorithmically churning through massive amounts of data is no guarantee of mechanistic understanding⁴, but is it possible to derive a systems-level mechanistic understanding without massive amounts of data⁵? Will more neural data be as “unreasonably effective” for uncovering mechanistic insights, similarly to how more data has been shown to improve the performance of deep learning and artificial intelligence⁶?

Markello, Hansen et al. emphasize that the current state of neuromaps is a beginning, not an end. One criticism of these data-driven approaches is that the data reflect the historical biases of researchers and research trends, just as has been demonstrated using data-driven clustering of functional domains in human neuroimaging⁷. However, in theory, as more data are collected, from more data types and from more people, the fidelity of these maps should only improve, allowing more fine-scaled links across domains. In addition, the future of neuromaps can be expanded to include even more maps, such as single-neuron electrophysiological properties, laminar profiles for neuron and glial cell types, dendritic geometry and more, further improving our ability to place results into the broader context.

To demonstrate its future potential, let us revisit the opening neuroimaging scenario, but from the neuromaps perspective. Here, we begin from a desire to understand the possible neurogenetic underpinnings of Alzheimer’s disease. We can take the maps for the brain regions associated with Alzheimer’s disease, as algorithmically extracted using text-mining via the Neurosynth platform⁸. We can then compare the spatial topography of these maps against maps of gene expression in the human brain to find genes that are statistically over- or underexpressed in these putative Alzheimer’s disease-linked brain regions to see whether there are any potential genes that have been historically overlooked. Or we can look to see whether these putative Alzheimer’s disease brain regions have a greater density of specific neuronal or glial cell types as compared to other brain regions not associated with Alzheimer’s disease, to see whether there is an overlooked cellular driver of Alzheimer’s disease.

This hypothetical scenario demonstrates the powerful potential for data-driven semiautomated hypothesis generation⁹. From this perspective, neuromaps is not something that is just used to answer new questions, but also to form entirely new hypotheses (Fig. 1). This approach complements experimental research, where we are no longer living in a fractured, data rich and theory poor world; rather we find ourselves in a data-rich neuroscience ecosystem that may lead to a theory-rich neuroscientific environment¹⁰.

Bradley Voytek  

Department of Cognitive Science, Halıcıoğlu Data Science Institute, Neurosciences Graduate Program, and Kavli Institute for Brain and Mind, University of California, San Diego, La Jolla, CA, USA.

✉ e-mail: bradley.voytek@gmail.com

Published online: 06 October 2022

References

1. Markello, R. D. et al. *Nat. Methods* <https://doi.org/10.1038/s41592-022-01625-w> (2022).
2. Fornito, A., Arnatkevičiūtė, A. & Fulcher, B. D. *Trends Cogn. Sci.* **23**, 34–50 (2019).
3. Kopell, N. J., Gritton, H. J., Whittington, M. A. & Kramer, M. A. *Neuron* **83**, 1319–1328 (2014).
4. Jonas, E. & Kording, K. *PLOS Comput. Biol.* <https://doi.org/10.1371/journal.pcbi.1005268> (2017).
5. Churchland, P. S. & Sejnowski, T. J. *Nat. Rev. Neurosci.* **17**, 667–668 (2016).
6. Halevy, A., Norvig, P. & Pereira, F. *IEEE Intell. Syst.* **24**, 8–12 (2009).
7. Beam, E., Potts, C., Poldrack, R. A. & Etkin, A. *Nat. Neurosci.* <https://doi.org/10.1038/s41593-021-00948-9> (2021).
8. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. *Nat. Methods* **8**, 665–670 (2011).
9. Voytek, J. B. & Voytek, B. J. *Neurosci. Methods* **208**, 92–100 (2012).
10. Voytek, B. *PLOS Comput. Biol.* **12**, e1005037 (2016).

Acknowledgements

B.V. is supported by National Institute of General Medical Sciences grant R01GM134363.

Competing interests

The author declares no competing interests.