Just looking: the innocent eye in neuroscience

Samuel J. Gershman
Department of Psychology and Center for Brain Science, Harvard University
Center for Brains, Minds, and Machines, MIT
52 Oxford St., Room 219.40, Cambridge, MA 02138
e-mail: gershman@fas.harvard.edu

April 3, 2021

Abstract

Since the early days of neuroscience, students have been instructed to look at their data and see what it is telling them. The idea that one can derive hypotheses by "just looking" at data implies a kind of innocent eye, apprehending reality in a form untainted by theorizing. More recently, statistical methods have been developed that replace the innocent eye with innocent algorithms. I argue here that neither eye nor algorithm are truly innocent, and that this epistemic attitude obscures the ubiquitous role that theory plays in the practice of data analysis.

"Attention is rarely directed to the space between the leaves of a tree, save when a Keats brings it to our notice."
Norwood Russell Hanson (1958), *Patterns of Discovery*

Introduction

In his 1857 book *The Elements of Drawing*, John Ruskin pointed out that grass sometimes turns yellow when illuminated by sunlight, yet we continue to perceive it as green (what modern vision scientists call "color constancy"). Ruskin argued that perception becomes trained through experience and social convention, so that we "always suppose that we *see* what we only know, and have hardly any consciousness of the real aspect of the signs we have learned to interpret." The painter's task, according to Ruskin, is to recover the infant's *innocent eye*, "a sort of childish perception of these flat stains of color, merely as such, without consciousness of what they signify."

Ruskin's mantra for painters could serve well as a mantra for experimental neuroscientists. If one *just looks* at the data, then facts can be documented and progress can be made. Ramon y Cajal, in his 1897 book *Advice for a Young Investigator*, put it this way:

> A scholar's positive contribution is measured by the sum of the original data that he contributes. Hypotheses come and go but data remain. Theories desert us, while data defend us. They are our true resources, our real estate, and our best pedigree. In the eternal shifting of things, only they will save us from the ravages of time and from the forgetfulness or injustice of men.

Cajal approvingly quoted the great German chemist Justus von Liebig:

> Don't make hypotheses. They will bring the enmity of the wise upon you. Be concerned with the discovery of new facts. They are the only things of merit that no one disregards.

In more recent times, this sentiment has been echoed by neuroscientists striving to shed the "philosophical" descriptors that have historically mediated our understanding of the brain, for example Buzsáki (2020):

> I suggest that neuroscience, as any new discipline, should establish its own vocabulary based on brain mechanisms. It should start with the brain (independent variable) and define descriptors of behavior (dependent variables) that are free from philosophical connotations and can be communicated across laboratories, languages, and cultures.

At the risk of bringing the enmity of the wise upon myself, I will argue against the possibility of data-driven discovery in neuroscience research. There is no innocent eye: our observation reports are inevitably theory-laden. This point has been made repeatedly (and controversially) in the philosophy of science (Duhem, 1954; Feyerabend, 1975; Hanson, 1958; Kuhn, 1962), and intersects with heated debates concerning deduction vs. induction, rationalism vs. empiricism, among others. The main consequence for our purposes is that one cannot maintain a naive distinction between facts and theories. If we cannot define facts in a theory-independent way, then we cannot discover new facts by "just looking" at data. I will examine a number of case studies in systems neuroscience illustrating this point.

In the second part of the paper, I extend this argument to a more sophisticated version of the innocent eye, what I call the *innocent algorithm*: the idea that "hypothesis-free" algorithms can extract meaningful patterns from data, in essence replacing "just looking" with "just analyzing." The challenge is to understand what counts as meaningful without appealing implicitly or explicitly to theoretical concepts. While in principle these algorithms can make fundamental discoveries, we can only recognize them as such by linking them back to what we already know. Thus, neither just looking nor just analyzing on their own suffice to break free from our existing theoretical concepts.

The drunkard's search

Consider the following question: Why do we find so many topographic maps in the brain? One possibility is that topography reflects fundamental organizing principles, such as dimensionality reduction (Durbin and Mitchison, 1990; Obermayer et al., 1990) or wiring length minimization (Koulakov and Chklovskii, 2001). A less obvious possibility is that we find so many topographic maps because *that is what we know how to find*. Space is a salient dimension of our perceptual experience and conceptual understanding of the world. Accordingly, spatial properties of experimental measurements will "jump out" at the neurophysiologist as she moves her electrode gradually through the brain. This is true even in the absence of topography, for example in the case of place cells in the hippocampus and grid cells in the entorhinal cortex, as I discuss further below.

A neuroscientist confronted with the bewildering opacity of the brain will first reach for familiar concepts ready at hand: space, time, size, color, orientation, and so on. Receptive fields organized along these dimensions may have been relatively easy to discover, not because they are fundamental, but because they are obvious. This issue is reminiscent of the parable known as the

drunkard's search. A policeman comes across a drunk man searching for his keys under a streetlight. The drunkard admits that he actually lost his keys in the park. "So why are you searching here?" asks the policeman, to which the drunkard replies: "Because this is where the light is."

Many neuroscientists will understandably bristle at this characterization. Is the history of neuroscience not rich with stories of true discovery? A celebrated example is Hubel and Wiesel's discovery of edge detectors in the cat primary visual cortex. Here is how Hubel and Wiesel describe the story:

> We had been recording in visual cortex from a large, isolated and stable cell for several hours without getting anywhere: none of our retinal stimuli produced any change in the cell's firing. Then we began to sense vague changes in firing as we stimulated one part of the retina. Suddenly there was a vigorous discharge, which occurred as we slid the glass slide into place. It took a while to discover that the firing had nothing to do with turning on or off the dark spot but occurred as we slid the piece of glass into and out of the slot. The stimulus turned out to be the faint but sharp line shadow cast on the retina by the moving edge of the glass (Hubel and Wiesel, 1998, pp. 402–403).

This seems like a clear case of data-driven discovery, with no particular hypothesis being tested. However, consider whether it would have been possible for Hubel and Wiesel to discover edge detectors without knowing about edges! It is easy to forget that edges are obvious because our perception is sensitive to edges. Yet there is no reason to expect that all the principles of brain function will be similarly obvious (see Poeppel and Adolfi, 2020, for further discussion). Indeed, there is reason to think that these neurons are not edge detectors at all. Hubel and Wiesel themselves noted an "endstopping" phenomenon, where extending the length of an edge beyond a neuron's classical receptive field caused an inhibition of firing (Hubel and Wiesel, 1968). It is not clear why a simple edge detector would exhibit this kind of tuning.

Decades later, an influential paper by Rao and Ballard (1999) provided an answer: the neurons are reporting *errors* between bottom-up sensory signals and top-down predictions. Longer edges indicate higher-order stimulus structure that renders the sensory signals predictable. If you know, for example, that you are looking at the edge of a table, then any segment of the table's edge is highly predictable from the table's large-scale geometry. Consistent with this hypothesis, removal of feedback from higher visuals areas strongly attenuates endstopping (Sandell and Schiller, 1982). Later experimental work confirmed that introduction of higher-order stimulus structure (e.g., shape; Murray et al., 2002) can have a suppressive effect on signaling in early visual areas.

Importantly, Rao and Ballard did not build edge detectors into their model. They trained a simple neural network to predict sensory data, and showed that both edge tuning and endstopping (as well as other receptive field properties) emerged from solving the prediction problem. The discovery of this fundamental principle (see also Srinivasan et al., 1982, for earlier work on this principle in the retina) came not from "just looking" at the data, but by thinking about the computational problems that need to be solved by the brain.

A similar story can be told about hippocampal place cells. The discovery of spatial tuning in the hippocampus (O'Keefe and Dostrovsky, 1971) directed much of the subsequent neurophysiology research on this area towards understanding its spatial tuning properties. Is this not an example of data-driven discovery *par excellence*? Recapitulating the argument about edge detectors, consider whether O'Keefe and Dostrovsky would have been able to identify spatial tuning

without a concept of space. This counterfactual is almost too unintuitive to contemplate, since space is so fundamental to human perception. Nonetheless, it highlights the point that O'Keefe and Dostrovsky were not "just looking"—their observations were embedded within the spatial framework of perception. We therefore have no way of knowing whether space is truly a fundamental organizing principle of the hippocampus or if space is a fundamental organizing principle of perception that constrains the kinds of principles we can discover through observation.

The hippocampus is an instructive example because many forms of non-spatial tuning have been discovered since (and even before) the discovery of place cells, suggesting possibly more abstract coding principles (Lisman et al., 2017). Recently, some of these principles have been formalized in computational models (e.g., Mok and Love, 2019; Stachenfeld et al., 2017; Whittington et al., 2020). These models make certain patterns in the data visible that were previously invisible. They are instruments not only of *understanding* the data but also of *describing* the data.

To drive this point home, recall the common practice in electrophysiology of focusing on a subset of neurons acquired during recording. These are the neurons that the experimenter can make sense of; the rest are ignored, at least temporarily. It's not the case that the experimenter has a ready description of what these cells are doing, and they're ignored simply because they're not interesting for whatever reason. On the contrary, the cells are often ignored precisely because the experimenter lacks such a description. Our descriptive powers increase with our theoretical powers.

Some of the most important discoveries in neuroscience arrived when experimenters brought to bear new theoretical ideas. For example, the discovery that dopamine neurons appear to signal reward prediction errors was driven by theorists who already knew about reward prediction errors (Schultz et al., 1997). It was not the case that neuroscientists discovered reward prediction errors from just looking at their data. To echo the theme of this section: how could they? Similarly, the discovery that neurons in the lateral intraparietal area (LIP) appear to signal accumulated evidence for perceptual decisions was enabled by the fact that the experimenters already knew about the concept of integrators (Shadlen and Newsome, 2001). It was not the case that neuroscientists discovered integrators from just looking at their data. How could they? Today, the ideas about reward prediction errors and evidence accumulators are being challenged (Hamid et al., 2016; Latimer et al., 2015). The challenges come not from just looking at the data, but from formalized alternative hypotheses which permit rigorous testing.

At this point, one might reasonably object that my examples are rather trivial; referring to visual features like edges as theoretical objects is elevating them above their station. Surely no one would hold that we begin experiments *completely* empty-headed? In that case, why not say that we begin with a repertoire of relatively theory-neutral (and possibly even innate) "perceptual input analyzers" (Carey, 2009). If one accepts that *some* theory-neutral observation is possible, then one rejects the strong form of the theory-ladenness argument (see Hacking, 1983). So where does one draw the line between theory-laden and theory-neutral observation? Without trying to resolve this question here, I think it's worth keeping our eye on the prize, which is not to catalog tuning curves but ultimately to understand the principles of brain function. Can we get from tuning curves to principles without any theory-laden observation along the way? It seems unlikely that we are equipped with the appropriate perceptual input analyzers that would make a theory of the brain pop out from experimental data.

The innocent algorithm

One response to the issues raised in the last section is to augment our powers of observation with better analysis algorithms. Can these algorithms reveal patterns that lead to substantive discoveries? I argue, on the one hand, that they can in principle output hypotheses that no one has yet conceived. On the other hand, our interpretations of these outputs are still limited by our current theoretical concepts.

To illustrate this point, I'll discuss a few examples. Norman-Haignere et al. (2015) took on the challenge of understanding representations in human auditory cortex using functional magnetic resonance imaging. They used a latent variable model (a variant of independent component analysis) to decompose voxel-wise tuning of auditory cortex voxels into a linear combination of response profiles (the latent components). The response profiles derived in this way were intuitive: four were selective for acoustic features such as frequency and pitch, one was selective for speech, and one was selective for music. The authors emphasize that these results were "hypothesis-free" in the sense that they didn't constrain the response profiles in any way that would bias them to recover these particular patterns of selectivity. This is true, but what if the response profiles reflected patterns that were unintelligible to the experimenters? Could they discover frequency and pitch tuning if they lacked these concepts coming into the experiment? In that case they would have no way of knowing whether the profiles were meaningful or garbage. At the end of the day, the hypothesis-free analysis can only be validated by appeal to prior hypotheses.

A similar issue applies to virtually all other papers proposing (relatively) unbiased data analysis algorithms (e.g., Hirokawa et al., 2019; Rubin et al., 2019; Williams et al., 2018). These algorithms are advertised as being unconstrained by prior hypotheses, but they inevitably fall back on prior hypotheses in order to certify their validity. For example, Kobak et al. (2016) proposed *demixed principal components analysis* of neural population data. They showed that the recovered components map onto features of the data identified by previous studies (e.g., responses time-locked to stimulus, memory and choice). Even though in principle the algorithm could (and probably does) recover components that don't map onto such intuitive features, it's not clear what to do with those components. Again, how do we know if they're meaningful or garbage? Only by appeal to prior hypotheses. There is no innocent algorithm, because ultimately we rely on our own eyes to judge the algorithm's usefulness, and our eyes are not innocent.

In a subversive study, Jonas and Kording (2017) asked whether sophisticated data analysis algorithms could be used to reveal how a microprocessor works. In short, their answer was no: these algorithms could identify some patterns, but not the underlying functional principles. This result should be a sobering reminder that we cannot completely rely on algorithms to guide us towards theories of the brain. They can in fact seriously mislead us.

Acknowledging that there is no innocent algorithm does not mean that the kinds of algorithms mentioned above are useless. On the contrary, they may be indispensable for linking theory and data (Linderman and Gershman, 2017). When a theorist invents a concept like prediction error, it may not be straightforward to map that concept onto brain activity. Is it encoded by a single neuron or a population, linearly or non-linearly, in firing rates or spike times? Statistical algorithms can address these questions.

Conclusion

I once had a conversation with a neuroscientist that went roughly as follows. He told me that we didn't need psychology to understand the brain; all we had to do was measure what's going on in the brain, and from those measurements we could derive everything we wanted to know about cognition. I think this point of view is common among neuroscientists, who yearn to free themselves from antiquated notions of mental representations once brain measurement and analysis technology become sufficiently advanced. This, I argue, is a pipe dream born from a naive belief in the innocent eye (see also Krakauer et al., 2017, for further arguments that amplify this point).

Discovery is possible, not by looking but by thinking. The only way to observe new things is to think new thoughts. The greatest leaps in our understanding of the brain have come at moments when new theories rendered the invisible visible, expanding our conceptual vocabulary and our descriptive powers. In Friedrich Nietzsche's words, "The greatest ideas are the greatest events."

Acknowledgments

I am grateful to John Krakauer, Kenny Blum, Venki Murthy, Nancy Kanwisher, Momchil Tomov, and Rick Born for comments on an earlier version of the paper. This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Buzsáki, G. (2020). The brain–cognitive behavior problem: A retrospective. *Eneuro*, 7.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Duhem, P. (1954). *The Aim and Structure of Physical Theory*. Princeton University Press.

Durbin, R. and Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, 343:644–647.

Feyerabend, P. (1975). *Against Method*. Verso.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., Kennedy, R. T., Aragona, B. J., and Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, 19:117–126.

Hanson, N. R. (1958). *Patterns of Discovery*. Cambridge University Press.

Hirokawa, J., Vaughan, A., Masset, P., Ott, T., and Kepecs, A. (2019). Frontal cortex neuron types categorically encode single decision variables. *Nature*, 576:446–451.

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

Hubel, D. H. and Wiesel, T. N. (1998). Early exploration of the visual cortex. *Neuron*, 20:401–412.

Jonas, E. and Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS Computational Biology*, 13:e1005268.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C. K. (2016). Demixed principal component analysis of neural population data. *Elife*, 5:e10989.

Koulakov, A. A. and Chklovskii, D. B. (2001). Orientation preference patterns in mammalian visual cortex: a wire length minimization approach. *Neuron*, 29:519–527.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93:480–490.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Latimer, K. W., Yates, J. L., Meister, M. L., Huk, A. C., and Pillow, J. W. (2015). Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349:184–187.

Linderman, S. W. and Gershman, S. J. (2017). Using computational theory to constrain statistical models of neural data. *Current Opinion in Neurobiology*, 46:14–24.

Lisman, J., Buzsáki, G., Eichenbaum, H., Nadel, L., Ranganath, C., and Redish, A. D. (2017). Viewpoints: how the hippocampus contributes to memory, navigation and cognition. *Nature Neuroscience*, 20:1434–1447.

Mok, R. M. and Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature Communications*, 10:1–9.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., and Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99:15164–15169.

Norman-Haignere, S., Kanwisher, N. G., and McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88:1281–1296.

Obermayer, K., Ritter, H., and Schulten, K. (1990). A principle for the formation of the spatial structure of cortical feature maps. *Proceedings of the National Academy Of Sciences*, 87:8345–8349.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*.

Poeppel, D. and Adolfi, F. (2020). Against the epistemological primacy of the hardware: The brain from inside out, turned upside down. *Eneuro*, 7.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87.

Rubin, A., Sheintuch, L., Brande-Eilat, N., Pinchasof, O., Rechavi, Y., Geva, N., and Ziv, Y. (2019). Revealing neural correlates of behavior without behavioral measurements. *Nature Communications*, 10:1–14.

Sandell, J. H. and Schiller, P. H. (1982). Effect of cooling area 18 on striate cortex cells in the squirrel monkey. *Journal of Neurophysiology*, 48:38–48.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Shadlen, M. N. and Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86:1916–1936.

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216:427–459.

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, 20:1643–1653.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2020). The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183:1249–1263.

Williams, A. H., Kim, T. H., Wang, F., Vyas, S., Ryu, S. I., Shenoy, K. V., Schnitzer, M., Kolda, T. G., and Ganguli, S. (2018). Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, 98:1099–1115.