Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

# Computational models in the age of large datasets

Timothy O'Leary, Alexander C Sutton and Eve Marder

Technological advances in experimental neuroscience are generating vast quantities of data, from the dynamics of single molecules to the structure and activity patterns of large networks of neurons. How do we make sense of these voluminous, complex, disparate and often incomplete data? How do we find general principles in the morass of detail? Computational models are invaluable and necessary in this task and yield insights that cannot otherwise be obtained. However, building and interpreting good computational models is a substantial challenge, especially so in the era of large datasets. Fitting detailed models to experimental data is difficult and often requires onerous assumptions, while more loosely constrained conceptual models that explore broad hypotheses and principles can yield more useful insights.

**Addresses**
Biology Department and Volen Center, Brandeis University, Waltham, MA 02454, United States

Corresponding author: Marder, Eve (marder@brandeis.edu)

**Current Opinion in Neurobiology** 2015, **32**:87–94

This review comes from a themed issue on **Large-scale recording technology**

Edited by **Francesco Battaglia** and **Mark J Schnitzer**

http://dx.doi.org/10.1016/j.conb.2015.01.006



## Introduction

By nature, experimental biologists collect and revere data, including the myriad details that characterize the particular system they are studying. At the [same time], as the onslaught of data increases, it is clear that we need tools that allow us to crisply extract understanding from the data that we can now generate. How do we find the general principles hiding among the details, and how do we understand which details are critical features of a process, and which details can be approximated or ignored while still permitting insight into an important biological question? Intelligent model building coupled to disciplined data analyses will be required to progress from data collection to understanding.

Computational models differ in their objectives, limitations and requirements. *Conceptual models* examine the consequences of broad assumptions. These kinds of models are useful for conducting rigorous thought experiments: one might ask how noise impacts latency in a forced choice between multiple alternatives [1], or how network topology determines the fusion and rivalry of visual percepts [2]. While conceptual models must be constrained by data in the sense that they cannot violate known facts about the world, they do not strive to assimilate or reproduce detailed experimental measurements. *Phenomenological data-driven models* aim to capture details of empirically observed data in a parsimonious way. For example, reduced models of single neurons [3,4] can often capture the behavior of neurons, but with simplified dynamics and few parameters. These kinds of models are useful for understanding 'higher level' functions of a neural system, be it a dendrite, a neuron or a neural circuit [5••] that, in the appropriate context, are independent of low-level details. Used carefully, they can tell us biologically relevant things about how nervous systems work without needing to constrain large numbers of parameters. *Detailed data-driven or 'realistic' models* attempt to assimilate as much experimental data as are available and account for detailed observations at the same time. Successful examples might include detailed structural models of ion channels that capture voltage-sensing and channel gating [6], or carefully parameterized models of biochemical signaling cascades underlying long-term potentiation [7]. With notable exceptions, models of this kind are often the least satisfying, as they can be most compromised by what has not been measured or characterized [8••].

How should we approach computational modeling in the era of 'big data'? The non-linear and dynamic nature of biological systems is a key obstacle for building detailed models [8••,9••] even when large amounts of data are available. For example, even well-characterized neural circuits such as crustacean CPGs that have full connectivity diagrams have not, to date, been successfully modeled in a level of detail that incorporates all of what is known about the synaptic physiology, intrinsic properties and circuit architecture [10]. As a consequence, there is still a big role for conceptual models that tell investigators what *kinds* of processes may underlie the data [11], or, more importantly, what potential mechanisms one should rule out [12,13•].

## Relating data to models

The Hodgkin-Huxley [14] model stands almost alone in its level of impact and in the way it achieved a more-or-less complete fit of the data. In hindsight their success came from extraordinarily good biological intuition about how action potentials are generated and a clever choice of

experimental preparation. Their model revealed fundamental principles of how a ubiquitous phenomenon — the spike, or action potential — resulted from few processes, namely two voltage-dependent membrane currents mediated by separate ionic species.

By contrast, the success of subsequent attempts to fit and model the biophysics of more complex neuronal conductances, neurons and circuits has been less dramatic — although insight into the roles of specific currents in neuronal dynamics has certainly been achieved [6,14,15,16•,17,18]. Understanding why this is the case requires investigators to step back and view the problem in a general setting. Biological systems are assembled from many component enzymes, signaling molecules and cellular structures. Modeling these components and their interactions produces complex nonlinear dynamical systems with multiple parameters for each component. For example, even if one specifies quite rigidly the desired output of a neuronal network, the underlying parameters that can give rise to these properties is weakly constrained as multiple solutions to neuronal and network dynamics are found [19,20]. Subsequent work, informed by this general finding, explored families of models with parameters scattered over plausible ranges [21,22,23,24•]. Although these studies abandoned the idea of finding unique fits to data, they nonetheless revealed important principles about how specific combinations of conductances contribute to neuronal and network behavior [22,23], and how temperature-robust neuronal function might emerge in cold-blooded animals that experience significant changes in temperature [21,24•].

There are fundamental reasons why it is challenging to fit large numbers of parameters in biological models [9••,25]. First, the models are typically nonlinear, so the relation between the parameters and the output can be complicated and many-valued. Averages of measured parameters can give rise to non-observed behavior [26] and models can be exquisitely sensitive to measured parameters [27–30]. The value of averaging as a means of combating experimental noise might thus be obviated by the possibility that the *average values* are not valid parameter combinations themselves. Second, biological systems have degenerate pathways and components, meaning that properties and functions of structurally distinct components overlap. While this confers robustness to the systems themselves, it means that models can be remarkably insensitive to many combinations of parameters [5••,21–23,27,29–31]. This 'sloppy' property of biological systems is well-documented in systems biology [8••] and neuroscientists may benefit from a wider appreciation of the tribulations and successes of model building in this sister field [32].

*Sloppiness* (Figure 1) means that models with large numbers of parameters exhibit relatively few sensitive directions in local regions of parameter space, although these directions are not generically aligned with parameter axes. Instead, the sensitive (and insensitive) directions are comprised of mixtures of parameters (Figure 1c), meaning that performance of a detailed model will be severely compromised by poor measurement, or ignorance of even a single parameter [8••]. A recent, elegant modeling study of oculomotor integration [5••] revealed a handful of sensitive directions in the high-dimensional parameter space of a complex neuronal circuit model (Figure 1d). The model permitted fresh insight into the trade-offs between structural and functional properties of a circuit and did so by constraining model behavior rather than measured parameters. As this study illustrates, useful insight into circuit function can be obtained from phenomenological matching of the overall model behavior to experimental data, provided the non-sloppy, or 'stiff', parameter combinations are identified [33].

A third reason for the difficulty of the 'fitting problem' arises because biological systems are intrinsically variable [34]. This variability is well-appreciated in the context of single neuron parameters, where neurons with highly stereotyped properties exhibit surprisingly large variability in their membrane conductance expression [20,35–38]. High variability is present wherever one looks, whether it is the synaptic connectivity of well-defined neural circuits [39–42] or the behavior of entire animals [43]. As a consequence, the number of valid, distinct parameter sets — should they be accessible — can equal the number of biological repeats of an experiment. This kind of variability is not noise; it represents genuinely different parameter combinations that the biological system has found. For this reason, understanding the *regulatory logic* of the nervous system is of fundamental importance [44••].

In an age when increasingly voluminous and complex datasets are demanding interpretation, these fundamental model-fitting problems are sobering. However, there are direct means of taming these difficulties by exploiting the resolution and high-dimensionality of the data themselves. An elegant analysis of the requirements for fitting a multicompartment model [31] showed that if one could access, at high temporal resolution, the membrane voltage of each compartment in a neuron, then one can recover the densities of multiple voltage-gated conductances — providing the identity and kinetics of the conductances are known. At the time this study was published, such measurements seemed impractical. Nearly 10 years later, we are on the verge of being able to make such measurements thanks to new molecular tools and improved microscopy.

Advances in statistical methods and fitting algorithms are accompanying advances in data collection. Many of these exploit fast computers and numerical methods such as