Computational Psychiatry: New Perspectives on Mental Illness

A. David Redish (ed.), Joshua A. Gordon (ed.)

https://doi.org/10.7551/mitpress/9780262035422.001.0001

Published online: 18 May 2017 **Published in print:** 18 November 2016 **Online ISBN:** 9780262337854 **Print ISBN:** 9780262035422

CHAPTER

5 Computational Approaches for Studying Mechanisms of Psychiatric Disorders

Zeb Kurth-Nelson, John P. O'Doherty, Deanna M. Barch, Sophie Denève, Daniel Durstewitz, Michael J. Frank, Joshua A. Gordon, Sanjay J. Mathew, Yael Niv, Kerry Ressler ... more

https://doi.org/10.7551/mitpress/9780262035422.003.0005 Pages 77–100

Published: November 2016

Abstract

Vast spectra of biological and psychological processes are potentially involved in the mechanisms of psychiatric illness. Computational neuroscience brings a diverse toolkit to bear on understanding these processes. This chapter begins by organizing the many ways in which computational neuroscience may provide insight to the mechanisms of psychiatric illness. It then contextualizes the quest for deep mechanistic understanding through the perspective that even partial or nonmechanistic understanding can be applied productively. Finally, it questions the standards by which these approaches should be evaluated. If computational psychiatry hopes to go beyond traditional psychiatry, it cannot be judged solely on the basis of how closely it reproduces the diagnoses and prognoses of traditional psychiatry, but must also be judged against more fundamental measures such as patient outcomes.

Keywords: Strüngmann Forum Report, computational psychiatry, computational neuroscience, psychiatry, modeling, decision making, mechanisms of psychiatric illness

Subject: Cognitive Neuroscience

Group photos (top left to bottom right) Zeb Kurth-Nelson, John O'Doherty, Yael Niv, Sophie Denève, Kerry Ressler, Josh Gordon, Heike Tost, Zeb Kurth-Nelson, Deanna Barch, Daniel Durstewitz, Josh Gordon, Sanjay Mathew, Kerry Ressler, Yael Niv, Daniel Durstewitz, Sophie Denève, Heike Tost, Michael Frank, Sanjay Mathew, Deanna Barch, John O'Doherty

Introduction

The human mind is the most complex known phenomenon. Mental illness is, almost certainly, correspondingly complex. To treat mental illnesses effectively, we need a deep understanding of its mechanisms, not just a description of its surface properties. Yet these mechanisms likely can only be described
p. 78 as ↳ sophisticated interactions between many moving parts that both function and fail in complex, nonlinear ways. A promise of computational psychiatry is to provide the tools that are naturally suited for describing this complexity, so as to capture the essence of both function and failure.

Although computational methods address complexity, they should be viewed as a way of making things simpler, not more complex, by providing insight and understanding that transcends what can be gleaned from experimental observation alone. In the end, this understanding should be communicable without equations.

A recurring theme in this chapter is that computational tools can be applied at many different points within the science and clinical practice of psychiatry. A crucial mechanism may be at the level of protein folding and ion channel kinetics, but may equally be at the level of structure hidden within patterns of treatment response across patients. Computational psychiatry does not require reductionism but seeks to apply computational tools wherever they might yield insight, often with the result of linking domains of knowledge in a parsimonious way.

The Space of Computational Psychiatry

At play in psychiatric illness is the whole organism: from genes to molecules, cells, circuits, brain systems, behavior, and social influences. To begin to attack this host of processes with computational methods, first we need to organize both the processes and the computational methods. On the computational side, the way that we think about the problem can be divided into levels of analysis, called "computational," "algorithmic," and "implementation," after Marr (1982). In this section we outline the levels of the organism and levels of analysis, and then, within this organizational framework, show where computational methods can be applied to psychiatry.

Levels of Biology and Psychology

Biological processes relevant to psychiatry have been studied extensively across many levels of scale (Figure 5.1). The smallest level of scale relevant for biology is arguably the molecular. (Genes are often situated below molecules because their code is unpacked into molecules.) Proteins and protein complexes are the building blocks of biology, carrying out the various processes that permit neurons, the cells of the nervous system, to develop, survive, signal, learn, grow, and senesce at the appropriate times and places. The cell, most centrally the neuron, is the next level of organization. The neuron (and perhaps the glial cell) forms the elemental information-processing unit, integrating synaptic and other signals in its dendrites and soma and passing this
p. 79 information, transformed according to its own special calculations, downstream ↳

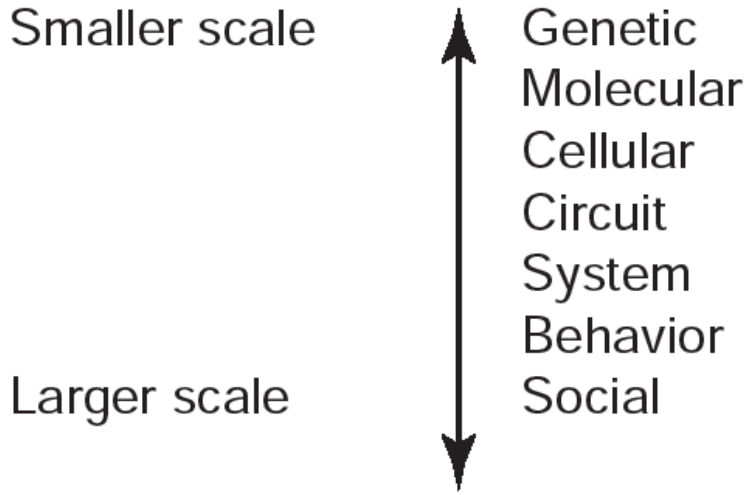

Figure 5.1 Levels of biology and psychology.

via an axon to its various partners. Multiple neurons and their connections form circuits, which act cooperatively and dynamically to increase the complexity and robustness of these neuron-based computations. Circuits combine to form distributed neural systems; activity in these systems guides behavior.

Psychiatric disorders can perhaps best be understood as spanning multiple levels of scale. For example, consider the case of the 22q11.2 microdeletion syndrome. Individuals with this microdeletion have cognitive and neuropsychiatric deficits and often meet criterion for schizophrenia (Karayiorgou et al. 2010). At the genetic level is the root cause of the disorder, a 1.5 to 3 Mb deletion knocking out 2–3 dozen genes. In mice, at least, one of the many consequences of the microdeletion at the molecular level is the mislocalization of an enzyme regulator in the axons (Mukai et al. 2015). At the cellular level, axon-branching deficits can be seen in cortical projection neurons; at the circuit level, neural transmission is reduced between the hippocampus and the prefrontal cortex (PFC). At the systems level, neural synchrony between the hippocampus and PFC is reduced (Sigurdsson et al. 2010), while at the behavioral level, mice carrying the microdeletion have deficits in spatial working memory (Stark et al. 2008).

Each of these individual deficits exists within a given level of scale. Yet a thorough understanding of the syndrome requires integration across scales. Are these observations connected? Does the molecular level deficit in enzyme localization cause the cellular level deficit in axon branching? A computational approach has the potential to enable integration by demonstrating, in a mathematically rigorous way, how phenomena on one level impact phenomena on another. Thus, one can construct a formal, mathematically quantifiable model of the axon-branching process, incorporating the location of the enzyme in question as a variable, and then test the effect of mislocalization on branching. The model predictions can be used to predict the results of experimental manipulations to provide further evidence in support of causal connections across scales.

It should be noted that the linear organization of levels presented here is simplified. There are fuzzy boundaries: between molecules and cells, there are organelles like mitochondria, macromolecular complexes such as the postsynaptic density, and signal processing pathways. There are also backward causal connections:
p. 80 neuronal activity shapes gene expression, circuit connectivity ↳ induces cellular specializations (such as synapses that recruit further changes in gene expression), and behavior itself strongly influences activity and plasticity across multiple levels. There are clear instances of level jumping: molecular changes can affect circuit formation and behavior, and systems activity can alter gene expression patterns. Finally, environmental influences appear at every level. Environment affects gene expression, neuronal growth and integrity, circuit plasticity, functional connectivity, and, of course, behavior.

These complexities notwithstanding, the differentiation of biological and psychological processes into levels of analysis can be useful as a means to identify how approaching a phenomenon at one spatial and temporal scale might inform understanding at other spatiotemporal scales. This differentiation can also serve as a starting point for the computational modeler in guiding the selection of a modeling approach that is appropriate to the type of phenomenon being studied.

Levels of Analysis in Computational Modeling

Next, we seek to organize the problem space of psychiatry from the point of view of how it can be described with computational models. David Marr famously described three levels of analysis for computational neuroscience (Figure 5.2; Marr 1982). The "computational level" specifies the goal of a system. For example, a computer program might be charged with sorting a list of numbers in a descending fashion. There are, however, many strategies for tackling this problem, such as sequentially searching for the n^{th}-largest number or repeatedly swapping pairs of numbers that are out of order. These strategies live at the "algorithmic level." Finally, a given algorithm must ultimately be realized in software and hardware, with details such as where in memory to store the array. This is the "implementation level." Computational models of brain and behavior can potentially link to psychiatry at all three levels.

Figure 5.2 Levels of analysis.

Take, for example, reward learning in the brain: A reinforcement learner's computational goal is to maximize their sum of future rewards. One (of several) algorithm that attempts to achieve this goal is the actor-critic algorithm. Here, the critic learns to predict the expected value of particular environmental states, and errors in these predictions (i.e., reward prediction errors) are used in two ways: (a) to improve future estimates and (b) to adjust weights in the actor, which selects among available actions (for further discussion, see chapters by Frank, Huys, and Montague, this volume). Those actions which yield the largest reward prediction errors in the critic are more likely to be selected, increasing experienced rewards. Implementation of this algorithm has been linked to the basal ganglia and dopaminergic system, with the ventral striatum learning the predicted values of states, the dopaminergic neurons signaling reward prediction errors (together with the ventral striatum, this is the critic), and the dorsal striatum playing the role of the actor (O'Doherty et al. 2004). One
p. 81 distinguishing feature of the actor-critic model is that it predicts no preference ↳

for actions that have yielded unexpected rewards over actions that avoid potential losses but have yielded no reward, since both have elicited positive prediction errors relative to a baseline of expected outcomes. Healthy controls and schizophrenia patients with high negative symptoms are well characterized by this pattern, whereas schizophrenia patients with low negative symptoms are better described by an alternative reinforcement-learning model (Gold et al. 2012; Palminteri et al. 2015).

A few words of caution are warranted. As with levels of biology, Marr's levels are inevitably fuzzy. One person's computational goal could be another person's algorithm in service of a broader computational goal. Also, the use of the word "computational" in Marr's levels is sometimes confusing. (A better name might be the "problem" level; that is, which problem the system is designed to solve.) Throughout the rest of this chapter, we will use "computational" much more generally to mean the application of sophisticated mathematical and theoretical tools to complex biological systems. Finally, we note that Flagel et al. (this volume) distinguishes between "normative" models and "process" models: the former is roughly equivalent to the computational level and the latter is roughly equivalent to algorithmic and implementation levels.

Any level of analysis can be applied to any level of biology. For example, we can ask what "goals" a gene network is set up to accomplish (e.g., maintaining balance in expression levels of two proteins), what algorithm

it uses (e.g., feedback inhibition), and how it is biologically implemented (e.g., binding of protein products to promotor regions).

One of the strengths of computational modeling is that it naturally draws out connections between levels of analysis by forcing us to think in detail about what algorithm achieves a computation, and how that algorithm may be implemented. These connections can lead to insights in psychiatry.

Different levels of analysis can and do inform each other. Sometimes biological implementation informs understanding of the algorithmic level. Biophysically detailed models can simulate the differential contributions of D1 and D2 striatal neurons and predict both neurophysiological recordings and effects of dopamine manipulations on behavior. The properties of these models can be summarized at the algorithmic level by a modified actor-critic called OpAL ("opponent actor learning") in which the actor is divided into two components that, via nonlinear learning rules, come to specialize on representing the benefits of alternative actions (in D1 neurons) and the costs of these actions (in D2 neurons) (Collins and Frank 2014). This framework
p. 82 allows the model to simultaneously capture the effects of a variety of manipulations across levels ↳ (including pharmacology, genetics, optogenetics, and behavior). This provides a clue as to the computational function of these opponent processes; indeed a normative analysis provided some evidence that they improve performance compared to classical reinforcement-learning algorithms. Thus, this example illustrates how algorithmic and computational considerations have informed interpretation of the biology and allowed for testable predictions, but also how mechanistic constraints can reciprocally inform the algorithmic level.

Principles for Applying Computation to Psychiatry

With this multidimensional spectrum of biology and computational approaches, what principles can guide our attack on psychiatric illness? The first principle, sometimes overlooked (Markram 2012), is that computational models must be targeted carefully to the questions we want to answer. A model can, *by design*, answer some questions but not others. For example, an architectural model of a building that is made of cardboard can be used to ask questions about general aesthetics, aspect ratios, and visual impact of the building, but not structural questions such as whether the roof will sustain a pool. The questions one would like to answer with the model should prescribe the level of description at which the model is designed, and what levels of description it can safely abstract over. For instance, if one would like to ask whether different methods for detoxification (inpatient, outpatient, etc.) might be more or less effective in preventing relapse to the addictive substance, a model at the level of systems and behavior might be more useful than one specified at a biophysical level of detail. In contrast, if the goal is to develop targeted gene therapy, detail at the biophysical level may well be needed.

The second principle is that models can provide insight into complex systems, but the insight itself need not be complex. For instance, consider the hypothesis that dopamine neurons implement a reward prediction error which is used for reward-related learning (Montague et al. 1996; Sutton and Barto 1998). This insight arose from theoretical models of reinforcement learning, yet in the end, the principle that it revealed is quite simple. This is a strength of computational models, not a weakness. The gleaning of simple principles from complex neural and behavioral data is the goal. The model is a tool for sharpening our thinking and for formalizing hypotheses, not an end unto itself.

The third principle is that disease mechanisms and mechanisms of healthy function are both productive targets for computational methods in psychiatry. Directly seeking to explain behavioral and genetic abnormalities in disease are obvious targets. However, applying computational methods to enrich our understanding of healthy function can also lead to translational results. For example, electrophysiological work
p. 83 that described dopamine effects on persistent neural activity in PFC (Seamans and Yang 2004) and the role of ↳ dopamine on working memory-related persistent activity (Goldman-Rakic et al. 2004) informed

computational models (Durstewitz et al. 2000; Durstewitz and Seamans 2002). These models, in turn, helped shape expectations about what D1 agonist pro-cognitive drugs might achieve (Rosell et al. 2015) and why PFC D1 deficits in schizophrenia might be associated with working memory dysfunction (Abi-Dargham et al. 2002).

The fourth principle is that both postdiction (i.e., explaining existing data) and prediction are useful. In some cases, models show how a simple set of principles can explain a broad range of existing data, which is valuable in clarifying our understanding and setting future directions. In other cases, models make novel, testable predictions that were difficult or impossible to make without the model. It is not uncommon to start by explaining a body of existing data and then make new predictions.

What Is the Toolkit That Computation Brings to Psychiatry?

In this section we outline the range of computational tools available to attack this problem space. It is useful to organize these tools both from the perspective of the biological or psychological systems to which they directly pertain as well as in terms of the mathematical framework from which they originate.

Organizing Computational Tools by Level of Biological Scale

Computational tools may be distinguished by the level of abstraction and biological scale they address (outlined in Table 5.1), moving from models that address detailed biophysical or biochemical processes to those that describe behavioral principles at an abstract level.

Table 5.1 Organizing models by level of biological scale. The left column lists some broad classes of computational models that target different features of biological systems. The right column maps the levels of biology and psychology to which these models are most often linked.

Type of model	Levels of biology
Biophysical models	Molecular
	Cellular
	Circuit
Connectionist models	Circuit
	System
Reinforcement learning	System
	Behavior
Bayesian inference	Circuit
	System
	Behavior

p. 84

Biophysical Models

Biophysical models attempt to faithfully capture real biological details, such as the temporal evolution of membrane potentials or the temporal and voltage-dependent behavior of ionic conductances. The advantage of these models for psychiatry is that they provide a close link to pharmacology and genetics by explicitly describing drug targets and gene products. Biophysical models exist at many levels of abstraction. At the most detailed level, a biophysical model may capture the whole spatial extent of neurons with all their axons and dendrites (compartmental models), with the gating behavior of a large array of ionic conductances, and perhaps even intracellular molecular cascades; in short, any biochemical or biophysical process that can be expressed in terms of differential equations (Koch and Laurent 1999). Intermediate-level models may reduce this structure to just a few spatial compartments (e.g., one for soma and one for dendrite) and retain just a few ionic currents essential for the questions at hand (e.g., Durstewitz and Gabriel 2007). At the most abstract level, a biophysical model may consist simply of one or a few differential equations for the membrane voltage and for variables which capture the lump effect of many ionic currents (e.g., Hertäg et al. 2012). In general, the most abstract is therefore the class of models most suitable for addressing questions about how, for instance, specific pharmacological agents or, more generally, genetic, molecular, or physiological factors impact on network dynamics, as variables measured experimentally at this level can be translated into the models with none or only few additional assumptions or simplifications. For instance, in Durstewitz et al. (2000), changes in several currents due to D1-or D2-class receptor activation as measured *in vitro* were implemented in compartmental models, which then were used in a level-bridging approach to investigate the implications of these current changes for network dynamics, and ultimately working memory and cognitive symptoms in schizophrenia (Durstewitz and Seamans 2008; Frank 2015).

Connectionist Models

One key approach to understanding mechanism has been the use of neural network models, also referred to as connectionist or deep learning models (McClelland et al. 2010). These models bridge between cells and behavior by explaining how groups of cells encode information in ways meaningful to behavior, and are often applied to circuit-or systems-level neural phenomena. A first generation of neural network modeling, beginning as early as the 1960s but flourishing in the 1990s and 2000s, provided powerful insights into the computational mechanisms underlying complex patterns of clinical dysfunction in language disorders and certain forms of dementia (Plaut and Shallice 1993; McClelland and Rogers 2003). Neural networks were also
p. 85 applied, beginning in this period, to account for impairments of cognitive control and ↳ working memory function in schizophrenia and other disorders, implementing some of the first computationally explicit proposals concerning the role of dopamine in psychiatric pathophysiology (Cohen and Servan-Schreiber 1992). Innovations in neural network methods combined with the advent of faster computers, capable of running large-scale simulations, have recently triggered a new wave of neural network research, allowing more direct validation of these models as accurate representations of neural information processing (Afraz et al. 2014), and application to richer bodies of data. Another feature of this new wave of research is the development of neural network models that also include constraints from biological data, while retaining close contact with behavioral phenomena (Hoffman and Cavus 2002).

A key feature of connectionist models is learning representations of the inputs that are useful for generating outputs. Even very complex representations learned by these models can be strikingly similar to real neurons (Yamins et al. 2014), suggesting relevance for psychiatry. This may include examples such as disordered perceptual representations as in schizophrenia as well as disordered representations of abstract decision-related variables that could underlie many phenomena including posttraumatic stress disorder (PTSD). These models can also potentially explore the time evolution in psychiatric disorders and model the effects of treatments at a network level.

Reinforcement Learning

Reinforcement-learning models quantify the dynamics of learning and decision making over time. These models, often cast at the level of systems and behavior, can be used to specify precisely hypotheses about how information obtained at one point in time affects beliefs and behavior at another. Reinforcement-learning models are concerned with learning to predict future rewards and punishments (as in Pavlovian or classical conditioning) and learning to select actions that would maximize future reward (as in operant or instrumental conditioning). The name "reinforcement learning" suggests an emphasis on learning dynamics; however, the models can also be used in steady state, after learning has achieved equilibrium, to make predictions and test hypotheses about decision making and action selection in different situations. Importantly, by fitting free parameters of these models to time series of behavioral data, one can precisely quantify different aspects of learning and decision making in individual patients. Relevant parameters might include the learning rate for appetitive and aversive outcomes, the degree of exploration versus exploitation, and the extent to which patients generalize across instances and stimuli. These parameter values can, in principle, be used as a diagnostic tool to characterize the different ways in which decision making can break down and to quantify individual differences (Moran et al., this volume). Indeed, reinforcement-learning parameters systematically
p. 86 vary as a function of disease, ↳ genetics, and pharmacology in ways that match predictions from decades of systems neuroscience (Frank and Fossella 2011).

In terms of Marr's levels, one of the strengths of reinforcement models is that they link from the computational level of optimizing future reward, through the algorithmic level of temporal difference learning, to the implementational level of dopamine-dependent plasticity in corticostriatal synapses (Barto 1995; Montague et al. 1996; Schultz et al. 1997). These neural and behavioral systems are also largely preserved in phylogeny, and thus reinforcement-learning models can be applied to humans and animals alike, even insects (Montague et al. 1995). The simplicity and transparency of these models allows one to give semantic interpretation to every construct of the model. However, one might argue that this is at the expense of allowing properties to "emerge" from the model in a way that sometimes occurs with models that embody more complex dynamics. This aspect of reinforcement-learning models can be seen as a feature rather than a flaw. The fact that these models do not have many moving parts allows one to easily form an intuitive understanding of the behavior of the model even from simply observing the model equations. Thus these models are most useful for specifying and sharpening hypotheses regarding learning dynamics in both healthy and clinical populations.

It is important to note that although reinforcement learning is highly prominent and promising for psychiatry, there are other classes of models, which we do not discuss here, that similarly provide process models for behavior that are linked to neural substrates. These include sequential sampling models, of which drift diffusion models are the most familiar (see Frank, this volume; Gold and Shadlen 2007).

Organizing Computational Tools by Mathematical Framework

Computational tools could also be organized according to mathematical frameworks and methodological toolkits. Specific computational models may, for instance, rely on mathematical tools from areas such as probability and statistical theory, nonlinear dynamics, or information theory. These areas of mathematics provide general frameworks for addressing computational questions at any level of abstractness or biological organization. Other mathematical frameworks that are applied across many levels of biology and psychology include statistics and machine learning, dynamical systems theory, and Bayesian methods as well as, to a lesser degree, information theory and optimal control theory.

Machine Learning

Machine-learning tools come out of statistics and computer science and were originally used mainly in the context of pattern recognition applications. In general, they use various types of mathematical principles and
p. 87 specific algorithms to analyze data so as to make predictions about existing and future data. ↳ Machine-learning tools can be either supervised (e.g., you have a specific categorical or dimensional variable guiding analysis) or unsupervised methods for characterizing data. Examples of supervised tools include support vector machines (Cortes and Vapnik 1995), which try to learn optimal decision boundaries for predicting class labels. Some examples of unsupervised tools include factor analysis, independent components analysis, and clustering approaches. Common to these unsupervised methods is that they attempt to detect structure within or suitable reductions of the data space without explicit advance knowledge of what that structure may be. All these approaches can be augmented by Bayesian methods to incorporate prior knowledge. These tools can either be used with a single type of data (e.g., behavior, neuroimaging, or genetic) or can be expanded to include several types of data or levels of data, such as in multimodal fusion approaches (Sui et al. 2012). Machine-learning tools can be used to identify novel structures in psychopathology, whether they might be dimensional, categorical, or a hybrid. For example, these tools could help identify categories or dimensions in high-dimensional data. They have already been used in this way in the psychopathology field, as a means to develop new models of the meta-structure of psychopathology based on phenomenological data (Krueger and Markon 2006; Wright et al. 2012, 2013). They can also help to integrate from one level of analysis to another (e.g., imaging to behavior, gene to imaging). In such cases, one could train on level A and predict on level B, without starting from strong hypotheses about how these transformations happen. For example, if we can identify structure directly in complex patterns of brain activity, this structure could inform our theories of the computations underlying behavior. As such, these general-purpose tools may help us identify additional types of data needed to understand the nature or mechanisms of such transformations. They can also help us to predict risk (Paulus et al., this volume) for the development of various forms of psychopathology based on different types of biomarkers. Further, they might even be able to be used to identify biomarkers that predict the success of different treatments based on similar cases, where similarity metrics are determined by the specific machine-learning method that is used. These tools can also be used more generally as data analytic tools for a variety of types of data (and are actively being used in this way), such as analysis of fMRI or connectivity data

For example, many risk factors for mental illness involve complex interactions between genes, the brain, and environmental influences which develop dynamically in naturalistic contexts (Kaddurah-Daouk and Weinshilboum 2015; Michino et al. 2015). To address this, we need data that is not only “big” but also multimodal. Ongoing studies are currently collecting functional neuroimaging, real-time/real-life data collected via smartphone, and geographical mapping. Mapping can link real-time smartphone data to specific locations where we have data about neuropsychiatric risk factors, such as urbanicity, pollution, sociodemographics, etc. Machine-learning techniques may aid the identification of patterns that predict risk-
p. 88 related neuroimaging markers. ↳ Further extensions of this concept in longitudinal study designs (accelerated longitudinal data acquisitions covering critical age ranges of neurodevelopmental disorders) may aid the identification of the dynamics of the neural correlates over time (discussed further below).

The current fields of genetics, genomics, and epigenetics provide abundant data for machine learning. Publicly available databases provide hundreds of thousands of control and patient subjects throughout the world,[1] and these genetic data are paired with categorical and continuous phenotypic data. For a smaller subset of individuals, there are also physiology and neuroimaging data. Computational methods across available big datasets will almost certainly allow deeper understanding of connections across levels of analyses from genetics, to epigenetics, circuits and behavior.

Dynamical Systems

Dynamical systems theory is a field in mathematics that addresses systems described by sets of equations which dictate the evolution of variables over time and/or space. It specifically addresses nonlinear systems for which some of the more conventional mathematical techniques (analytical approaches to equation solving) break down. A central concept in dynamical systems theory is that of a state-space (i.e., the space spanned by all dynamical variables of the system). A point in this space captures the current state of the system, and the evolution of this state across time is given by a trajectory meandering through this space. The course of this trajectory is determined by various geometrical properties of this space, like for instance the existence of attractor states, such as stable orbits (limit cycles), which give rise to nonlinear oscillations. Nonlinear dynamics provides a set of tools to characterize the flow of these trajectories (and thus the system's evolution in time and space) and analyze their behavior. Since essentially all neural and behavioral phenomena can be cast in terms of variables that evolve dynamically in time, nonlinear dynamics provides a very general framework for describing and analyzing computational models.

Dynamical systems may also provide ways of capturing phenomena that may be central to understanding the mechanisms of breakdown in psychiatric conditions. For example, NMDA receptor dysfunction has been implicated in schizophrenia (Barch, this volume). We can begin to understand the mechanics of this dysfunction with dynamical systems theory. As NMDA conductance steadily increases, both in real cells and in biophysically plausible models, the system suddenly jumps from quiescence into a bursting mode, then jumps again from regular bursting into chaotic irregular activity, and finally from chaos into regular steady single spiking (Durstewitz and Gabriel 2007). Although the changes in the underlying system parameter (NMDA
p. 89 conductance) are ↳ gradual, the neuron's spiking modes change abruptly. These abrupt jumps between different operating regimes are called "phase transitions." There are also numerous examples of phase transitions at the level of neural populations (Durstewitz et al. 2010).

The idea of abrupt or critical transitions between operating regimes can also arise at very different levels of scale. For example, there may be enough resilience in the brain and in behavioral or social coping mechanisms to allow underlying biological changes to occur without obvious psychiatric symptoms, until some critical point is reached and there is an abrupt shift to a different regime such as depression or psychosis.

Bayesian Methods

Bayesian inference describes how one can use probability theory to infer the state of variables we are interested in, given prior knowledge and noisy observations. Bayesian approaches start with a hypothesis H, and some observed data O. For example, H could be the hypothesis that a patient has lung cancer, and O could be a positive blood test. It is relatively easy to measure P(O|H) (i.e., the probability of getting a positive blood test given that one has lung cancer), by measuring the proportion of the population with cancer that have positive blood tests. This can be used to compute, using Bayes's rule, the more important quantity of P(H|O); that is, the probability that a patient with a certain blood test result has lung cancer. According to Bayes's rule, P(H|O) is proportional to P(O|H) · P(H), where P(H) is the prior probability of having lung cancer (i.e., the prevalence of lung cancer in the general population). This general framework can also be used to include multiple observations or to predict new observations. It can also be used to build hierarchical representations: "H" can play the role of the observation, "O," for another, higher-order model. Finally, Bayes's rule can capture temporal prediction or temporal evolution of a state, with P(H) corresponding to knowledge from the past and P(H|O) representing the new knowledge updated by observations. The utility of this framework for psychiatry is highlighted by Flagel et al. (this volume).

Bayesian models can be used as "normative" descriptions of brain function. Bayesian belief updating represents the optimal solution to many problems under a very broadly applicable set of constraints. Thus, it is reasonable to posit that in many cases the brain may be attempting to calculate P(H|O) or some good

approximation thereof. This may be true of the computations of individual neurons (e.g., if their firing rates encode beliefs about perceptual features, and their synaptic inputs encode new evidence about these features), circuits, systems, or the entire organism (e.g., how an individual reaches their beliefs about others' intentions). This view is very powerful because it then allows us to explore the mechanism by which the brain calculates beliefs based on observations, or how the calculation might go wrong (Huys et al. 2015b).

p. 90 The brain constructs highly hierarchical representations of its environment. For example, visual areas construct increasingly "meaningful" representations (from local contrast to contour, basic shapes, and objects) of the visual world. To do so, both feedforward connections (sending information from sensory area to "higher level" area) and feedback connections (sending information prior expectations to the sensory area) are essential. For instance, detecting a tree requires integrating feedforward sensory information ("green," "tall,"etc.) with prior knowledge ("I am in a forest"). Since both feedforward and feedback connections are excitatory in the human brain, such highly recurrent excitatory circuits could not work properly on their own. Sensory information would be sent up the hierarchy, generating expectation, then reverberated back, combined with themselves, then sent back up as if they were new sensory evidence, in an endless cycle. Such a system would suffer from an extreme amount of "circular inference," making us "see what we expect" or "expect what we see." It would also learn "fake" causal relationships between completely uncorrelated events, simply because their neural representations are correlated through the network dynamics. To function properly and generate an accurate belief system, the brain needs to combine excitatory (E) feedforward and feedback connections with strong, balancing inhibition (I), whose goal is to cancel all predictable (reverberated) excitatory inputs in the network. Such tight E/I balance is a widely observed phenomena in cortex. It could be that imbalances in E/I (involved in a wide range of mental illness such as bipolar disorder, schizophrenia, or autism) causes circular inference, leading to the formation of aberrant beliefs (overconfidence, hallucinations, delusions, alien control). New experiments confirm that the behavior of schizophrenic patients in probabilistic inference tasks was well described by such a "circular inference" hypothesis (Jardri and Denève 2013).

This method can also be used as a tool to organize and generalize from complex, high-dimensional and noisy data in any domain. As such, Bayesian inference can provide a useful tool for diagnosis and treatment of mental illness. A well-known example is "Bayesian causal models," widely used to interpret imaging data, but applicable, in general, to any type of data (for further detail, see Moran et al., this volume).

Bridging Levels

We have considered various levels of modeling and how they can be used to ask different sorts of questions, from biophysical to normative. We have also emphasized that no single level of analysis is sufficient to make the connections between mechanism and behavioral symptoms relevant for psychiatric illness; the complementary values of each level implies that an-all-of-the-above strategy is useful. Informally, one can also interpret modeling endeavors at one level in terms of the other. For example, a biophysical model of dopamine
p. 91 modulation of attractor dynamics and flexibility in PFC can be summarized by ↳ analogous functions in connectionist networks and their application to cognitive tasks (Cohen et al. 2002). More formally, one can also quantitatively map the properties of one model onto another. This affords a richer testable prediction that leverages the utilities of both levels (Frank 2015). Typically, two approaches are taken. The first is to derive exact mappings. For example, Ma et al. (2006) showed how spiking models, including probabilistic population codes, can precisely implement Bayesian inference in a sensory cue combination task, building on work from Zemel et al. (1998), who developed the idea of spikes as encoding probabilistic information. Bogacz and Gurney (2007) showed how an optimal model of evidence accumulation during decision making can be mapped onto the anatomy of the basal ganglia.

The second approach is not to assume that the mapping between levels is exact, but rather that it is approximate, and instead to fit the behavioral output patterns of complex network quantitatively using a

higher-level algorithmic description. This leverages the advantage of the algorithmic models: because the behavioral data can be fit quantitatively with few free parameters and the same strategy (that these models use) can be applied when fitting to empirical data, a determination can be made as to which of several alternative algorithmic models best describes the behavior of the system. Thereafter, an estimate can be made as to the impact of biological manipulations in the network on higher-level algorithmic parameters, which in turn can guide empirical experimentation. For example, Ratcliff and Frank (2012) showed that the outputs of a network model of the basal ganglia are well approximated by drift diffusion models (DDMs). In these models, evidence for each of two or more options is accumulated noisily over time until one option reaches a decision threshold and is chosen. Ratcliff and Frank also found that parametric modulations of the subthalamic nucleus (STN) affect the decision threshold (as opposed to other decision parameters), particularly in the face of choice conflict. This prediction was tested empirically by recording and manipulating STN function and estimating its impact on drift diffusion parameters, based on choices and response time distributions (and EEG data). Indeed, subsequent fMRI studies provided evidence that STN activity is related to decision threshold adjustment during choice conflict (Frank et al. 2015). STN manipulation in Parkinson disease reduced the decision threshold for these choices (Cavanagh et al. 2011; Green et al. 2013), providing a novel interpretation for how impulse control disorders can arise in these patients. This is just one example of how computations at one level can afford analysis at another, allowing falsifiable predictions. One can also further bridge these levels with machine-learning tools to classify or cluster patients based on fitted model parameters as well as to identify which parameters/mechanisms contribute most strongly to classification (Wiecki et al. 2015).

Another example for the scale-bridging approach is provided by the "dual-state model" of PFC dopamine function (Durstewitz and Seamans 2002). In this biophysically anchored theory, slice-electrophysiological
p. 92 observations on ↳ the dopamine D1-and D2-class receptor modulation of a range of different voltage-gated and synaptic currents were linked through dynamical systems tools to alterations in prefrontal attractor dynamics, which in turn could be related to changes in working memory function and cognitive flexibility (see Frank, this volume).

Applying Computational Methods to the Evolution of Systems over Time

A central feature of mental illness is that it is not static; it evolves both developmentally and in adulthood with prodromal stages and subclinical antecedents, through episodic or slowly changing patterns of symptoms, to remission and often relapse. Adding the dimension of time creates significantly more complexity, which provides an entry point for computational methods.

As an example, PTSD is unique among psychiatric disorders in that one key component of its etiology (i.e., the traumatic event) is known. What remains uncertain is how the initial clinical manifestations of acute stress (e.g., hyperarousal, re-experiencing, avoidance) may progress in some individuals to the constellation of symptoms and associated social dysfunction that characterizes the disorder. It is noteworthy that only a relatively small proportion of individuals who are evaluated in the emergency room following a traumatic stressor (e.g., a motor vehicle accident or assault) are diagnosed with PTSD at 6 months to one year following the trauma. A prime application of computational methods would be to enhance prediction of symptom progression from acute stress reactions (as seen in an emergency room setting).

Another example pertains to the long-term course of depressive episodes of major depressive disorder, which can be a highly recurrent illness marked by discrete illness episodes and periods of relative stability (Thase 2013). Depressive episodes can be sorted into categories based on specifiers such as melancholic features (e.g., minimal mood reactivity, early morning awakening, diurnal variation), atypical features (e.g., mood reactivity, hyperphagia, hypersomnia), and psychotic features (e.g., delusions, hallucinations), which have state-

dependent neurobiological correlates. However, individuals often switch unpredictably from one category to another between episodes (Oquendo et al. 2004), which poses challenges in implementing treatment strategies for relapse prevention. Moreover, proper clinical decision making requires predictions. Issues such as how long to continue a medication after a patient achieves remission, or whether to continue with a partially effective treatment or to switch to a new one are crucial for treatment providers. Computational models that take into account dynamics over time would be immensely helpful in making such decisions.

Change over time may be driven by spontaneous internal changes, but also by changing environmental factors,
p. 93 including treatments. Furthermore, internal ↳ and external variables interact in complex ways. Individuals with depression, for example, are especially vulnerable to variations in mood according to seasons as well as to hormonal perturbations related to pregnancy, the postpartum period, menstrual cycle variation, and use of hormonal contraceptives.

The emergence of the complex neurobiology of chronic or recurrent mood disorders may be viewed as having progressed through a number of stages. Before the first episode of depression, vulnerability exists at genetic, physiological, and environmental levels (Lupien et al. 2011). Chronic and acute stress (allostatic load) perturb homeostatic mechanisms at multiple organismal and neural levels, such as mood, sleep, appetite, and motivation (McEwen 2003). Factors contributing to the transition from "having a bad month" to developing "depression" include elevations in circulating glucocorticoid and inflammatory cytokines. These factors have numerous consequences for the brain. One mechanism that has received attention is the compromise of the glial capacity to transport glutamate, resulting in elevations in extrasynaptic glutamate levels. These elevations suppress point-to-point synaptic functional connectivity in circuits regulating mood by inhibiting glutamate release via stimulation of presynaptic mGluR2 receptors and by causing the retraction of dendritic spines. In the long term, this reduces dendritic complexity due to excessive stimulation of extrasynaptic GluN2B-containing NMDA receptors and reductions in the level of trophic factors (reviewed in Krystal et al. 2013). The disruptions in structural and functional connectivity, combined with many other neuroplastic mechanisms (including alterations in reward and social learning), may make it impossible to "bounce back."

Modeling Time

Data which describe trajectories over time come in many forms and include both behavioral assessments and physiological markers. Can this multivariate time series of data be used to learn more about the mechanisms of the disease? Can practical predictions be made about the future disease course or the effects of treatments and interventions?

Formally, there are several useful general-purpose approaches. First, autoregressive–moving-average (ARMA) models express current observations as a weighted linear combination of previous observations plus noise. Here, forward prediction is straightforward using the estimated weight parameters. Nonlinear variants of these models also exist, such as threshold or piecewise linear autoregressive models.

Second, state-space models include latent (i.e., unobserved) as well as observed variables. Latent variables capture underlying causes, such as neural activity, which cannot be directly measured but nonetheless have effects on the observed data. The time evolution of the latent variables can be described mathematically as an
p. 94 ARMA process, or as a discrete set of states with ↳ transition probabilities, or a combination thereof as in the class of switching state-space models (Ghahramani and Hinton 2000).

Third, there is a large toolbox from nonlinear dynamical (NLD) systems theory. NLD methods usually start from a state-space representation of the observed system. This is the space spanned by all the dynamical variables of the system (e.g., the firing rates of a set of recorded units). A point in this space specifies the current state of the system, and the movement of this point through the space, as time passes, yields a trajectory. In physical and biological systems, these trajectories are described by geometrical objects within these spaces like

"attractor states." Based on such representations, NLD theory offers various methods for prediction and assessing the effect of interventions in these spaces (Lapish et al. 2015).

Understanding the dynamics of mental illness is not only crucial for their diagnosis, management, and treatment but also for bringing some light to the underlying mechanisms. For instance, when episodes occur in an approximately periodic fashion (as, e.g., in bipolar disorder), they might be described in dynamical systems terms through an underlying periodic or chaotic oscillator. This, in turn, may offer a way to study when in the cycle it would be best to intervene therapeutically. An acute episode recruits compensatory processes, which persist after the episode is finished. However, these compensatory processes themselves may be regulated through other feedback loops with the environment, as is common in biology, which in turn can cause another episode. In contrast, when episodes occur erratically without prior warning and sudden onset, compensatory processes might be better described through metastable states or bifurcation mechanisms giving rise to an instability. In this case, the healthy state is fragile (e.g., due to weakened homeostasis). The brain state can be temporarily thrown out of this state, jumping to a pathological state.

Is There a Use for Computational Approaches without Understanding "Fundamental" Mechanisms?

Computational neuroscience brings a powerful set of conceptual tools for understanding complex systems. However, we must be cognizant that we may never fully understand every level of mechanistic detail in the path from molecule to behavior. Still, there are many ways in which computational approaches can be used to enhance our understanding of behavior as well as approaches to treatments and interventions.

It is tempting to look to the most detailed or microscopic level for the most "fundamental" understanding, but this is often a mistake. For example, a liquid only exists as the interaction between atoms. Some argue that the fundamental level for a given phenomenon is the most detailed level at which the phenomenon exists. Others argue that it is the level at which the phenomenon is most parsimoniously captured. There is broad agreement
p. 95 that for phenomena which ↳ live mostly at higher levels, it is most useful to study them at these levels. Crucially, we *do not know* at which level most psychiatric illnesses are most usefully described and studied. Even if a gene conferring risk for an illness produces a malformed ion channel, it is possible that behavior at a cellular or circuit level (e.g., synaptic plasticity) might look essentially normal and that pathology may only appear when investigating properties of the brain at a higher level of organization.

Furthermore, the best level of description is related to the use that one makes of the description. If one is interested in etiology, for instance, then genetics may be particularly important, but if one is interested in developing pharmacologic treatments, then an understanding of cells and circuits are important. Even with a single type of use, say medication development, and desired endpoint, to alleviate a disorder, different treatment mechanisms will be developed to target different levels of description (e.g., to correct an abnormal protein, to correct synaptic or circuit dysfunction, or to correct a behavior).

In this section we outline three ways that computation offers a benefit without necessarily reaching the most detailed level of explanation. First, some approaches allow us to characterize some aspects of mechanism without requiring an understanding of fundamentals. Second, whether or not we can achieve mechanistic understanding of the disease, it is useful to obtain mechanistic understanding of other related phenomena (e.g., recovery and resilience). Third, we can eschew mechanism entirely and use computational methods to optimize treatment directly.

Characterizing Mechanism at a High Level

It is possible to extract knowledge and impact treatments using computational approaches, even without understanding the fundamental mechanisms behind the illness. We might have a very useful understanding of how the system behaves and misbehaves that is in some sense mechanistic, but without reference to deeper mechanisms. Neuroimaging, for example, can be used to identify which areas of the brain are activated during hallucinations. These areas could then be targeted with, say, a 1 Hz transcranial magnetic stimulation (TMS), which decreases the activation of the targeted area, to reduce hallucinations over a time period of several weeks. In service of treatment, this leverages a partial understanding of brain regions and disturbances in excitability (balance of excitation and inhibition) that might be corrected, through TMS, without requiring a complete picture of the underlying neural signaling disturbances or the impact of the TMS on these disturbances (Hoffman and Cavus 2002; Hoffman et al. 2007). Likewise, there are very effective treatments at a purely behavioral level that rely on some understanding of mechanism (e.g., from psychology) at this level, without understanding anything about the brain.

p. 96

Mechanisms of Resilience and Recovery

Without fully understanding the original causes of a disease, we may begin to understand mechanisms of resilience to the disease or recovery from it. Resilience may simply constitute lower vulnerability to disease. On the other hand, resilience may be a more active phenomenon. Within a certain range, neural systems can return to their original homeostatic states. But if stretched too far, a system may "break," resulting in a discontinuity, such as a pathological response (e.g., PTSD, anxiety disorder). However, in some cases, the organism achieves a new stable state that not only adapts to the current stressor but can better withstand subsequent stress, thus resulting in enhanced resilience (Friedman et al. 2014). The mechanisms of this reactive resilience are still poorly understood, but represent a prime target for dynamical systems models that capture such multistability.

Likewise, mechanisms of recovery are sometimes quite distinct from mechanisms of pathology. Most psychiatric treatments do not fix the underlying pathology (e.g., depression is not caused by a lack of electroconvulsive seizures). However, models can potentially be used to understand the mechanisms of treatment and recovery. An understanding of learning theory suggests ways to make extinction permanent (e.g., fear in PTSD or phobia, compulsive behavior in obsessive-compulsive disorder, craving in addictive behaviors). Computational learning theory implies ways to optimize behavioral therapy or computer apps without understanding the underlying neural and molecular mechanisms of the original disorder.

Computational Methods to Optimize Treatment Directly

In some cases, we may temporarily abandon the quest for mechanistic understanding of a disease process and use computational methods to analyze data directly and make predictions and recommendations about treatment. Advances here can be in the realm of descriptive nosology (see chapters by First, MacDonald et al., and Flagel et al., this volume), such as clustering of patients using computational algorithms based on current symptom/intermediate phenotype datasets, independent of underlying mechanistic knowledge (Borsboom et al. 2011; Borsboom and Cramer 2013). Computational approaches can also be very helpful in optimizing treatments and understanding of outcomes: from optimal timing and dosing of medication, when to start/stop treatment, and even optimal organization of psychiatric treatment flow in clinics. All can be improved with computational modeling to optimize current processes, agnostic to the underlying mechanisms of functioning of these approaches. This approach is being used, for example, to enhance treatment parameters with electroconvulsive therapy (Deng et al. 2013; McClintock et al. 2014).

p. 97 A striking and non-obvious observation (and perhaps a deep principle in computational psychiatry) is that applying analysis methods to data, without explicitly trying to model the mechanism, may actually help *reveal the mechanism of the disease*. For example, suppose one were to attempt to model the trajectory of disease episodes in a patient with schizophrenia using a hidden Markov model. The goal of this model might be to predict the next episode so as to guide treatment. Yet as part of the process of establishing an optimal model fit to the clinical data, we may infer a number of states in the hidden Markov model. If this parameter is consistent between subjects, or consistently relates to some other important variables that have biological or psychological relevance, we may accidentally reap clues about the mechanism of the disorder itself.

Finally, notwithstanding the above, it is worth striving for a more fundamental mechanistic understanding. Because the mechanisms of hypertension are known, a physician usually will not prescribe another beta blocker if a patient is already on beta blockers; instead, the physician will try adding a drug with a different mechanism of action. Perhaps even more importantly, a deeper understanding of mechanism will help to achieve more complete remission and ultimately lasting recovery. This could be the difference between a treatment that works partially and temporarily versus a cure. Realistically, all of these approaches must be combined to create a versatile armamentarium.

How Can We Measure the Success of Computational Approaches?

In early computational psychiatry, theoretical approaches were sometimes judged by how well they could reproduce traditional approaches (e.g., whether clustering model parameters could reproduce diagnostic categories in the DSM). Since computational psychiatry may soon exceed the usefulness of traditional approaches, this correspondence should not be used as a primary metric. Instead we need to step back and think about how we can gauge, in a more fundamental sense, what is or is not working (Clementz et al. 2016).

Treatment Outcomes

In a clinical sense, the ultimate gold standard is to improve outcomes for patients. In the ideal case, computational psychiatry could come to be very explicitly and directly part of treatment, so that changes in the prevalence or incidence of the disease after the introduction of the techniques could be measured. Here we outline five primary vehicles toward this end:

1. Computational approaches might inform basic research that subsequently leads to improvements in patient outcomes (e.g., by identifying critical neural circuits or components of cognition).

2. Computation might help predict risk status, thus enabling more informed interventions. In bridging
p. 98 levels of biology, models may be ↳ particularly well suited to develop an integrated understanding of risk factors across levels: from the known genetic risk architecture of mental illness (Gottesman and Gould 2003; Preston and Weinberger 2005; Cannon and Keller 2006; Meyer-Lindenberg and Weinberger 2006; Kendler and Neale 2010) to known neurobiological, behavioral, and environmental factors. Relatedly, predicting risk may also serve a role in forensic psychiatry, which includes risk to others—an area that could potentially have a tremendous impact on society (Buchanan 1999; Freedman 2001; Loza and Dhaliwal 2005; Odgers et al. 2005; Dahle 2006; Odeh et al. 2006; Hill et al. 2012; Chu et al. 2013).

3. Computation will reveal new treatments and treatment targets. Biophysical models could be used to identify novel molecular targets, greatly facilitating screening for new drugs. At the circuit level, we can identify neural circuits for brain stimulation interventions based on an understanding of these circuits' role in overall brain function, allowing us to fine-tune stimulation parameters (Gutman et al. 2009; Datta et al. 2012; Rotem et al. 2014; Li et al. 2015; Senço et al. 2015). Similarly, models that identify particular

behavioral variables, and deficits in the same, might suggest novel types of psychotherapies aimed at addressing these behavioral pathways. For example, reinforcement learning and other learning paradigms have already had a great deal of impact at the level of informing certain forms of cognitive behavioral therapy; in particular, prolonged exposure therapy, use of virtual reality-based therapies, and the use of cognitive enhancers (such as d-cycloserine) to enhance the rate of extinction learning in combination with exposure (Conklin and Tiffany 2002; Rothbaum and Davis 2003; McNally 2007; Craske et al. 2008; Abramowitz 2013).

4. By understanding *how* individual treatments work, existing treatments can be repurposed to treat a different disease or to work more effectively. In the models of Parkinson disease discussed above, modeling suggested that the disease involves learned avoidance due to exaggerated learning in D2 neurons, and that this learning component can be rescued by adenosine antagonists which block plasticity in these neurons (Beeler et al. 2012). This may also explain the failure of such antagonists in clinical trials wherein they were administered to Parkinson patients in the advanced stage: it predicts that these treatments will be most effective during very *early* stages of the disease to prevent aberrant learning.

5. A particular strength of computational approaches is identifying exactly what data we need to collect to make effective predictions. We often have the potential to gather a huge array of data modalities about the patient from neuroimaging, cognitive tasks, questionnaires, genetics, hormone levels, etc., but we need better methods to determine which will actually provide the critical information to guide treatment.

p. 99

Making Scientific Progress

Finally, in some views, basic scientific progress is an end unto itself. Psychiatric disorders may even be seen as fortuitous because they shed light on how the healthy brain works. Early progress in neuroscience was accelerated by observing the consequence of gunshot wounds in particular areas of the brain (suddenly plentiful after World War I). Similarly, elucidation of the nature of dysfunction in psychiatric disorders may facilitate progress in the understanding of fundamental brain mechanisms.

We wish, however, to end with a note of caution. The field of computational psychiatry is still in its relative infancy, and the problem to be tackled is immense. Thus, patience must be exercised, as we expect progress in
p. 100 fits and starts. The original promise of computational psychiatry may take decades to be fully realized. ↳

Notes

1 National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/genome/ (accessed June 15, 2016).