1   Introduction

Scientists construct quantitative models to explain observations about natural systems in a coherent and rigorous manner. Today, as the number of models and the quantity of empirical data increases, scientists face a grand challenge: efficiently discovering relevant models and characterizing their validity against a continually growing body of available evidence.

A new model is typically proposed by publishing a description of how it works along with an argument justifying its utility to some targeted scientific community. This argument is made in words, accompanied by a few relevant equations and figures, and judged by peer review. Ideally, reviewers determine whether the model's predictions are consistent with all available and relevant data and compare performance against previously published models. This is increasingly difficult; although authors are may refer to relevant experimental data and review the literature, these citations are incomplete and biased, in that authors will focus on data favorable to their models, or compare them to implausibly simple null hypotheses.

If modeling papers contained figures a) highlighting accordance with all related experiments and b) comparing performance to all related models, publications would be encyclopedic with their main points obscured. A strength of model publication is focused description of how a model works and its conceptual and technical advances. A weakness, however, is that evaluating the scope and validity of a model is intractable using a publication alone. Publications tell us **how** a model works but are less effective at reporting **which** goals it achieves and **how well** it achieves them. This problem is exacerbated as future data is gathered. Although model validity may change in light of new data, there is no systematic process in biology today for re-evaluating existing models. Although new data and its theoretical implications may propagate informally through a scientific community or appear in periodic reviews, the original publications – a resource of first resort for new scientists and onlookers – are cited "as-is" in perpetuity.

The central problem discussed here: the process of model validation is not rigorous, comprehensive, or ongoing. This hampers experimental prediction, simultaneous comparison of models, and identification of outstanding research problems. To overcome these obstacles, we propose formalizing the model validation process by creating software and associated cyberinfrastructure dedicated to the systemization of scientific model validation. This validation framework will exist in parallel to publication, allowing the latter to focus on answering **how**, while the former can answer **how well**.

1.1   Existing Efforts

There are several facilities developed for data and model sharing in biology, but few if any facilitate evaluation of models against data. For example, the Collaborative Research In Computational Neuroscience (CRCNS) data-sharing website [1], and the Open Source Brain (OSB) repository [2] are facilities for data and model sharing in neuroscience, respectively. The CRCNS website hosts several excellent data sets of relevance to computational models; however, the data is not organized or annotated in a way that permits systematic model validation. The OSB is focused on model description and reliable execution; however, it lacks a means to test models. We aims to bridge such resources, strengthening each in turn.

Related efforts in the machine learning community promote model development and validation against publicly-available datasets. Kaggle [3] drives model development by organizing competitions where training data is made public and competitors submit competing algorithms, compared automatically by cross-validated accuracy on an unrevealed test set. The success of Kaggle [4] shows that open competitions are effective, and that this paradigm of "modeling as a competition" attracts data scientists across traditional discipline boundaries. However, models developed in this way fit only a few general cases in machine learning: classification, regression, and clustering. Validation criteria are straightforward in these domains. Discipline-specific competitions for biology have also resulted in technical advances. The quantitative single neuron modeling competition [5] helped us understand the complexity-accuracy tradeoff among reduced models of excitable membranes; the "Hopfield" challenge [6] famously illustrated the challenges of identification facing computational neuroscience; the Diadem challenge is advancing the art of neurite reconstruction [7]; and examples from other subfields of biology abound [8]. Our challenge is to develop a general framework in support of distributed data-driven model validation workflows, but where validation criteria may be complex and discipline-specific. Initially, we will be focus on validation challenges in the

A more recent reference might be better since Kaggle has grown a lot since 2011, but this reference is in Science and other ones I saw are in trade magazines.

biological sciences, particularly neuroscience.

2 Outcomes and Products

Our framework is based on simple validation tests that compute agreement between a model execution result and an experimental observation. Model validity can be identified with the collection of tests passed by the model. This methodology is inspired by the ubiquitous "unit testing" practice of software engineering. A unit test evaluates whether a portion of a computer program (often a single function) meets a simple correctness criterion. A suite of such tests covering desired program behavior validates overall program functionality, and isolates specific causes of failures. Developers often write unit tests before writing the program itself, following a methodology called test-driven development (TDD) [**?**]. TDD asserts that a suite of unit tests serve as a program's specification, guiding its development. This allows developers to measure progress simply by looking at the proportion of tests that pass at any point in development. When modifications are made, developers can ensure that existing functionality is maintained by ensuring that all previously passed tests continue to do so. The success of unit testing and TDD in practice suggests that validation testing is a practical solution to model validation.

[margin note: Fix reference]

2.1 Example

What might a scientific model validation test look like? We begin with a neuroscience example, derived from experiments where a stimulus is delivered to a neuron, in the form of somatically injected current, while the somatic membrane potential of that cell is recorded and stored. A model claiming to capture the membrane potential dynamics of this cell type might be validated in several ways; one simple test of that claim would check for correspondence between the distributions of action potential counts produced by the model and the data, respectively. This test would ask a candidate model to produce a prediction for each provided input current and then check whether each prediction agrees with the data for that input current. The measure and degree of agreement are specified by the test creator. For example, model action potential counts that, for each value of input current, fall with the 95% confidence interval of the empirical data distribution might constitute a passed test. This choice is made explicit in the test specification, so modelers need not guess which criteria testers may be using for validation.

[margin note: Is this whole section useful?]

2.2 Community Workflow

Unit tests are produced by the community of developers involved in the development of a program. Correspondingly, members of a scientific community can collaboratively produce suites of validation tests in common source code repositories. Each test specifies the model requirements, and provides a procedure for determining whether a candidate model (the input to the test) is consistent with a particular summary of experimental observations (the parameters of the test). Model performance on a collaboratively curated suite of validation tests can serve as justification of claims regarding the model's validity as a description of the scientific system characterized by the test suite. This workflow continuously produces a summary of the field indicating which models are capable of taking each test, i.e. whether the test falls within the model's scope, and for each capable model, how well it performed, i.e. its validity. Model usefulness becomes a function of the weight given to each test, determined per investigator or by community consensus. This test-driven scientific workflow leverages the desire of modelers to promote their models' virtues – test suites represent clear *challenges*, issued by experimentalists to the community, and passed tests certify success for the modelers. As more data is gathered and test suites are refined, past models can be tested continuously againt current empirical knowledge, in a fully-automated manner because the interface between tests and models is fixed.

[margin note: Maybe this section should be moved to the end of the introduction]

2.3 *SciUnit*: A Simple Validation Testing Framework

Conceptual and practical simplicity – the absence of "heavy-lifting" – is an essential requirement for community participation. Our design does not require the release of raw data or storage in standard formats. Instead, salient aspects of the data are abstracted away behind a test definition. Similarly, implementation details of a model (e.g. its programming language) need not be public or standardized – the implementation must simply be wrapped to expose a standardized interface, so that tests can access model capabilities.

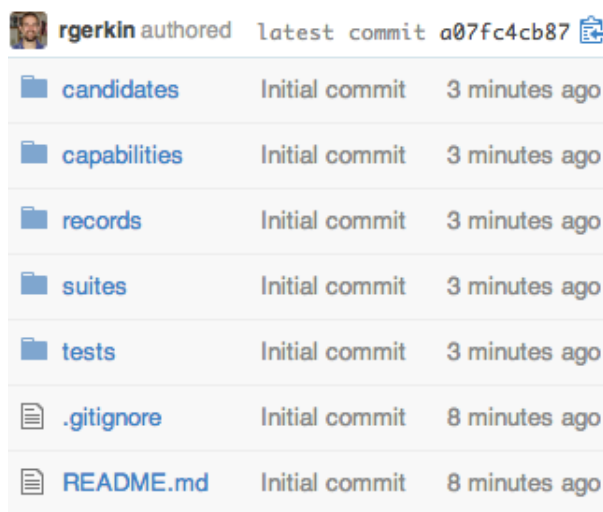*[margin note: Maybe this section should be the start of "Outcomes and Products"]*

2

Each community can collaboratively specify model interfaces capturing high-level operations of their preferred modeling formalisms. Existing modeling standards (e.g. NeuroML [9, 10]) and interface definition languages (IDLs) [?]are leveraged to simplify this process. **Thus, the first product of this proposal is an object-oriented validation testing framework called *SciUnit*, written in Python, that makes test construction (from data) and test-driven model execution easy and flexible.**

Fix reference

2.4 *SciDash*: A Community web application

This framework is most effective when the current state of a research area is represented by a collection of tests and models. This requires coordination among research groups, so community-oriented cyberinfrastructure to support test suite creation and summarization, building upon *SciUnit*, is essentiall. We propose a service, called *SciDash*, utilizing existing infrastructure for coordinated development of software repositories, focusing on GitHub [11] [12]. A suite of related tests will be contained in a GitHub repository with a stereotyped high-level structure that allows the *SciDash* web application to discover, index and summarize it on an ongoing basis (Fig. 2.3). The portal, a web application written using the pythonic Django framework [13] will serve as

Update this figure when SciDash repo structure is finalized.

Figure 1: A *SciDash* repository on GitHub

a central location where scientists can discover relevant test suites, determine test requirements, and summarize the results of test execution on submitted models. Test results can be visualized as a "record matrix" composed of large numbers of model/test combinations (Table 2 provides a toy example). Each row in this matrix will contain results for all tests taken by one model and would serve as clear evidence of that model's scope and validity. Models, tests, and records will be stored on GitHub, pointed to by hyperlinks in the record matrix. *SciDash* will serve as a public record of competition among models, facilitating and encouraging data-driven development among modelers. **Thus, the second product of this proposal is a web application called *SciDash*, whose back-end can index test suite repositories collaboratively developed on GitHub, and whose front-end offers users a simple environment for discovery and evaluation of models, tests, and results.**

2.5 *NeuroUnit*: A Suite of Tests for Neurophysiology

To demonstrate the utility of these tools, an initial case study must be conducted. Because we have substantial experimental and computational neurophysiology training and expertise, we target the work toward this domain. By using machine-readable models from the Open Source Brain repository, and corresponding machine-readable data from resources like The NeuroElectro Project, our efforts will produce a body of useful tests that also serve as a demonstration of the test-driven workflow that other areas in the biological sciences can leverage. **The third product of this proposal is *NeuroUnit*, a large suite of data-driven tests and model interfaces for a representative set of canonical neurophysiology models, all publicly available. A case study examining the implementation of *NeuroUnit* will inform development of domain specific test-suites in other biology sub-fields.**

URL via citation

3 Preliminary Activities

We have developed most of the core testing framework, *SciUnit*, under a free, open-source license at the project development repository [14].

3.1 *SciUnit* Implementation

We implemented *SciUnit* using the Python programming language [**?**] due to its widespread adoption across the quantitative sciences. Python also supports interoperability with other major languages used within science, including C, R [15], and MATLAB [16]. The functionality described below could also be readily translated into any programming language with comparable facilities.

3.2 Validation Tests

Testingin *SciUnit* proceeds in three phases (Figure 2):

1. Capability Checking: The test checks to see that the model is capable of taking the test; i.e., that it exposes the needed functionality for the execution phase to proceed.
2. Test Execution: The model is executed to produce output by interaction of the test with model capabilities. If the model is described according to a standard specification, e.g. NeuroML, this corresponds to loading it into a supported simulator, executing it, and reading simulator output.
3. Output Validation: Model output is compared to empirical data to determine whether the model passes. This comparison can be of several types: for deterministic models, we may extract a matching quantile from a data distribution; for stochastic models, we may compute a measure of agreement between model output and data distributions, e.g. the Kolmogorov-Smirnov statistic. Validation is ultimately pass/fail, but in some cases the result generated in this phase also includes a continuous statistic so that models can be ranked based on their test results in a more fine-grained manner.

> I don't think this abstraction is useful. We make it concrete on the next page anyway.

> Is the workflow figure useful?

A validation test in *SciUnit* is an instance of a Python class implementing the `sciunit.Test` interface. Classes that implement this interface must contain a `run_test` method that receives a model as input and produces a `sciunit.Score` (a normalized score) as output. The capabilities that the model must possess in order for the test to operate correctly are given using the `required_capabilities` attribute. In many cases, several tests will have a similar form, differing only by a choice of parameters or by association with a particular dataset. Such parameterized *test families* correspond to subclasses of `sciunit.Test`. One *test* is an instance of that subclass, with the needed parameters as constructor arguments.

In Figure 3, the `RateTest` class implements a test family that compares the output "firing rate" (action potentials per second) of a neuronal model, driven by a given somatic input current, to an empirical mean and standard deviation (s.d.) of the firing rate derived from a set of replicated experiments. The test produces a boolean score indicating possible agreement within one s.d. of the empirical mean. The input currents, mean, and standard deviation are provided as constructor arguments on line 2 (constructors are named `__init__` in Python.) This parameterization creates a test from the `RateTest` test family. On lines 6-9, the `required_capabilities` are listed: the model must be able to take a current as input and produce firing rates as output. These *capabilities* are used in the `run_test` method, on lines 11 and 12 respectively, to execute the model against the provided current. The resulting
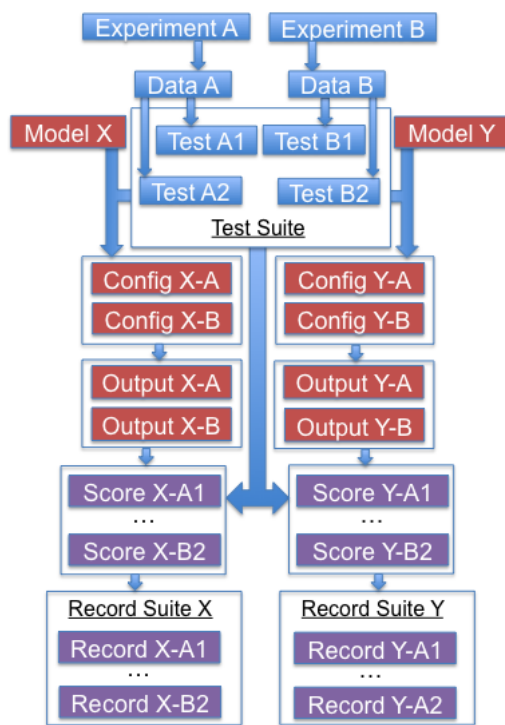

Figure 2: Workflow in *SciUnit*. i) Each experiment generates data, ii) From each set of data, many validation tests can be generated from test families (not shown); an ensemble of tests (from data within or across experiments) forms a test suite. iii) Tests enumerate capabilities required for models to take them, and provide metadata used to configure models for execution. iv) Model execution output is checked against test rules to produce a score. Some test scores are not indicated here (. . .) for clarity. v) The score establishes a validation record for that model+test combination. An ensemble of records for one model is a record suite, and summarizes the scope and validity of the model.

firing rate is compared to the empirical mean and s.d on line
17 and a boolean score is produced on lines 18-22. In addi-
tion to the boolean value itself, the score also contains metadata that may be useful to scientists wishing to
examine the result in detail. In this example, we save the model's output firing rate and the provided mean
and s.d. alongside the result, by specifying the `related_data` parameter.

To create a validation test, a scientist instantiates the RateTest class with a particular mean, standard
deviation and input current (I, not shown):

```
CA1_cell_rate_test = RateTest(mean=18.0, std=40.0, input_current=I)
```

The test is then executed against a model instance (described below), using the `sciunit.run` function:

```
score = sciunit.run(CA1_cell_rate_test, CA1_cell_model)
```

The `sciunit.run` function operates in three phases, implementing the abstractions from section 3.2:

1. Capability Checking: The model is verified as capable of taking the test by checking each capability
   in the test's `required_capabilities` attribute.
2. Test Execution: The test's `run_test` method is called to execute the model and cache output.
3. Output Validation: `run_test` returns an instance of a `sciunit.Score` subclass containing a
   goodness-of-fit metric between the model output and the data used to instantiate the test. Here the
   goodness-of-fit is all-or-none, i.e. pass/fail.

Any errors that occur during this process are reported by raising an appropriate exception.

3.3 Test Suites

A test suite is a collection of tests designed to validate a model against several mutually coherent require-
ments:

```
spike_tests = sciunit.TestSuite([SpikeWidthTest, SpikeHeightTest])
```

When a test suite is executed against a model, it produces summary data that can be shown on the console
or visualized by other tools, such as the web application described in the next section. A test suite can also
be used to compare multiple models (described below), producing a table as in Table 2.

3.4 Candidates and Capabilities

A *candidate* is the entity to be tested. In practice, this will usually implementat a particular scientific model,
so *candidate* and *model* can be used interchangeably; we distinguish them to allow for the possibility that
a candidate itself be a dataset subject to testing. For clarity, we will refer to candidates as models herein.
A model is represented by an instance of a class inheriting from `sciunit.Candidate`. In many cases,
models can be grouped into parameterized *model families*, as with tests, with parameters passed to an anal-
ogous subclass constructor.

In order for a model to be tested, it must have certain *capabilities* that the test requires. That is, the
model must be capable of taking certain kinds of input and generating certain kinds of output that the test
will use to evaluate its validity. A *capability* is a contract guaranteeing that a model is able to produce results
of a particular form. We model capabilities as Python classes containing unimplemented methods. In other
object-oriented languages, this corresponds to an *interface* or an abstract class.

Let us consider the capabilities of a model publicly available on the Open Source Brain website [17]. In
the system being modeled, recorded data consists of the somatic membrane potential of a pyramidal neuron
from area CA1 of the hippocampus, and the stimulus consists of somatically-injected current. This somatic
membrane potential includes action potentials, the rate of which is known in neurophysiology as the "firing
rate". Corresponding model capabilities might include `ReceivesCurrent`, `ProducesFiringRate`,
and `TimeIntegrable`, the last of these indicating that the model can be integrated over time.

Many capabilities must be enumerated to span a community's modeling formalisms. In the simple exam-
ple we present here, these three capabilities suffice. As two of them are domain-specific (to neuroscience),
we place them not in *SciUnit* itself but in *NeuroUnit* in the module `neurounit.capabilities`. In
contrast, `sciunit.capabilities` contains only domain-independent model capabilities like

Figure 3: A single neuron firing rate test implemented in *SciUnit*

```python
1  class RateTest(sciunit.Test):
2      def __init__(self, mean, std, input_current):
3          self.mean, self.std, self.input_current = \
4          mean, std, input_current
5
6      required_capabilities = (
7          neurounit.Capabilities.ReceivesCurrent,
8          neurounit.Capabilities.ProducesFiringRate
9      )
10
11     def run_test(self, model):
12         model.set_current(self.input_current)
13         # Implementation guaranteed by capability check.
14         rate = model.get_firing_rate()
15
16         mean, std = self.mean, self.std
17         result = (mean - std) <= rate <= (mean - std)
18         return sciunit.scores.BooleanScore(result, related_data={
19             "rate": rate,
20             "mean": mean,
21             "std": std
22         })
```

TimeIntegrable.

A model is represented by an instance of a class inheriting from `sciunit.Candidate`. In Figure 4, the CA1 cell model family implements the above capabilities, which both permits testing and importantly determines which tests are within its scope. A models implemented at a different level of detail may have different capabilities and thus be eligible to take a different set of tests.

In Figure 4, the model produces a firing rate after being provided a current. The capabilities a model must implement are determined by its base classes; it multiply inherits some `sciunit.Candidate` subclass (here NeuroConstruct.Model, described in 4.1.1), and some number of `sciunit.Capability` subclasses. The model then expresses methods (such as `set_current`) that override (and implement) the abstract, unimplemented methods provided in the corresponding `sciunit.Capability` subclasses. Any capability the model inherits must have all if its associated methods implemented. Failure to do so will result in a `NotImplementedError`, indicating a model scope less than claimed by its inheritance.

Figure 4: A candidate class corresponding to a CA1 Pyramidal Cell model from Open Source Brain

```python
1  class CA1PyramidalCellModel(NeuroConstructModel,
2                              capabilities.ReceivesCurrent):
3      """CA1 Pyramidal Cell model from /neuroConstruct/osb/hippocampus/
4      CA1_pyramidal_neuron/CA1PyramidalCell"""
5      def __init__(self,**kwargs):
6          # Put any other initialization here.
7          project_path = os.path.join(candidates.putils.OSB_MODELS,
8                                      "hippocampus",
9                                      "CA1_pyramidal_neuron",
10                                     "CA1PyramidalCell",
11                                     "neuroConstruct")
12         candidates.NeuroConstructModel.__init__(self,project_path,**kwargs)
13         super(CA1PyramidalCellModel,self).__init__(**kwargs)
```

A *model family* is initialized with parameters (`i_leak` in the case of the model in Fig. 4) to produce a *model*. A test does not specify these parameters, however a tester may nonetheless subject a range of models from a model family to the same test. This is done by repeatedly instantiating the model with different parameters (for example, in a loop). If a particular parameter value in the model is required for a test to be sensible, this can be expressed in the list of `required_capabilities`. For example, if the test is only passable by models corresponding to a particular pharamacological milieu (e.g. with some fraction of ion channels of a certain type blocked), then `required_capabilities` might include `ModifiesConductance`, exposing a method `set_conductance`. The test specification could also include run-time arguments for models, which could correspond to a waveform of injected somatic current, or alternatively to training data for models implementing completing algorithms. Figure 2 illustrates the testing workflow, and Table 1 provides definitions of terms.

Table 1: Glossary of Terms

Data	The measurements yielded by an experiment.
Model	An executable function that aims to explain, reproduce, or predict experimental data. More generally, a candidate.
Model Family	A class that generates a set of related model instances. Each model instance from one family differs according to the set of parameters (corresponding to a particular experimental design, or set of training data) with which it was initialized.
Validation	The process of assessing the agreement between model output and experimental data.
(Validation) Test	A function that executes a model and scores the model's output against a summary of experimental data. One experiment can inform the development of many tests.
Test Family	A class that generates a set of related test instances. Test instances from within one family differ only quantitatively, according to the data used to initialize them. All tests from one family attempt to validate the same property of a model.
Test Suite	A collection of tests, each of which attempts to validate a different property of a model.
Score	A value summarizing the performance of a model on a test. This can be a measure of the goodness of fit between model output and experimental data.
Scope	A list of tests that a model is eligible to take, indicating the range of experimental data that the model might be able to explain or replicate.
Capability	A contract that states that a model is able to produce results of a particular form. For example, a Hodgkin-Huxley neuron model is capable of producing a membrane potential time-series, whereas a model without a notion of voltage is not.
Record	An indication of the test score achieved by a model, along with any initialization conditions.
Record Suite	A list of records associated with a model. This summarizes the scope and validity of the model.

3.5 *SciDash* Development

We have also begun work on the *SciDash* web application, which point to models and tests and aggregate test results for scientific communities. A development version is currently live [18]. It is populated with models and tests from the educational illustration in Table 2. This application has several key features. **First**, it provides a "dashboard" view of the state of a field by displaying a matrix of records for selected model/test combinations, i.e. those stored in a given GitHub repository (Fig. 2.3. **Second**, each record will link to the tests results (stored in the repository), displaying line-by-line the executed code and intermediate output statements, as well as the underlying models and tests. One possibility is the use of the Sage notebook [21], whose "worksheets" facilitate visualization and storage of executed code. While Sage was written for Python, the notebook also supports the execution of code and visualization of outputs from other programming languages, making possible worksheets based on MATLAB, Mathematica, and other popular environments for modeling. **Third**, a community-moderated comment section will allow test-makers and test-takers to discuss issues associated with each record. Thus, disagreements about the appropriateness of a test can be openly aired and in many cases resolved. We support open authentication via existing web applications (Google, Yahoo, Twitter, etc.), lowering the barrier to participation. The *SciDash* backend will consist of a MySQL relational database, updated regularly using the GitHub API to populate record matrices.

Is this the appropriate level of detail?

4 Research and Development Plan

In order to 1) accelerate development and testing of *SciUnit* and 2) populate *SciDash* with models and tests, we will focus during the funded period on development of *NeuroUnit*, to serve one discipline (neurophysiology), without compromising generalizability. This will facilitate rapid feedback from an experimental community we know well as we iterate through the software development cycle.

4.1 Leverage of existing resources for the development of *NeuroUnit*

Here we describe standard neuroinformatics tools we have adopted to develop *NeuroUnit* [22].

4.1.1 Candidates from NeuroML Models

NeuroML is a standardized model description language for neuroscience [10]. It permits many neurophysiological/neuroanatomical models to be described in a simulator-independent fashion, and these models can be run across many popular simulators due to emerging inter-conversion capabilities of the NeuroML API. Because NeuroML is an XML specification, model descriptions can be validated for correctness and queried for model properties and components, exposing potential capabilities. It is ideal for model sharing and curation, as well as for answering both *what* and *how* programatically.

NeuroConstruct [23, 24] is a simulation manager that starts with NeuroML models and does what is required to hand off simulation to any of several neural simulators. To this end, *NeuroUnit* offers a `sciunit.Candidate` subclass called `NeuroConstructModel`. `NeuroConstructModel` is instantiated with the path to any NeuroML model created or previously imported into NeuroConstruct. Because the set of models that can be described by NeuroML is vast, the `NeuroConstructModel` class makes limited assumptions about candidate models: that they each is `TimeIntegrable`, and `HasMembranePotential`. It is then subclassed to test specific NeuroML models.

The Open Source Brain project (OSB, [2]) curates many such models described in NeuroML. In contrast to previous curation efforts such as ModelDB [25, 26]), OSB projects are converted from their native format into NeuroML, and run on major neural simulators such as NEURON, Genesis, PyNN, Moose, and others. The degree of concordance between model output (beginning with the NeuroML description) and reference output (from native simulator source files) is reported for each model. Thus, OSB is an excellent source of models that, in addition to being open source, are described completely enough to enable validation. The hippocampal CA1 pyramidal cell is commonly modeled, and we demonstrate testing using a `CA1PyramidalCellModel` class (Figure 4), deriving from `NeuroConstructModel` and implementing a CA1 cell model on OSB [17]. Importantly, this implementation consists of little more than "wrapping" the functionality of the existing model; The code shown in Fig. 4 is all that is required for a basic implementation. All OSB models, and indeed any NeuroML model via NeuroConstruct, can be tested similarly. Working together with OSB is part of our **first collaboration**, and our integration efforts can be publicly tracked in our fork of the NeuroConstruct code repository [27].

Although they span a range of scales, and regardless of their original implementation, all OSB models are formally described using NeuroML, as are each of the model components and sub-components, such as cells, ion channels, calcium stores, etc. These models are regularly executed on OSB servers to ensure that, as they are updated, their output is consistent with previous versions. Therefore, OSB can confirm that they *do* work, while linked journal articles, on-site wiki, and inspection of native code or NeuroML source can establish *how* they work. However, currently there is no mechanism for establishing *how well* they work, i.e. how well the models accord with data. *SciUnit* fills this gap by providing OSB (and the larger biology community) with a means to assess models using data-driven tests in the *NeuroUnit* library. A similar process can be used to apply *SciUnit* to other biology sub-disciplines using *NeuroUnit* analogues developed for those disciplines.

4.1.2 Capabilities from NeuroTools

NeuroTools [28] is a Python library supporting tasks associated with analysis of neural data (or model output). It implements the extraction and analysis of simulation output, such as membrane potential time series, spike trains, etc. NeuroTools is an open source and actively developed project, containing reliable algorithms on which neurophysiology tests can be based.

We use NeuroTools to implement *SciUnit* capabilities in *NeuroUnit* (Figure **??**). For example, a NeuroTools `AnalogSignal` object (e.g. a membrane potential time series) has a threshold detection method that returns a NeuroTools `SpikeTrain` object. A *NeuroUnit* `HasSpikeTrain` Capability requires that a method getSpikeTrain be implemented. NeuroConstructModel does so by making `getSpikeTrain` wrap `AnalogSignal.threshold_detection`. This is one of many examples in which NeuroTools objects are exchanged between `NeuroConstructModel` methods. This greatly simplifies test writing, since many basic properties of model output are obtained trivially by using NeuroTools object methods, and these NeuroTools objects are easily extracted from model output using candidate models derived from the `NeuroConstructModel` base class.

4.1.3 Reference Data for Tests from NeuroElectro

Answering *how well* requires validation testing against data. The NeuroElectro project [29] is an effort to curate all published single cell neurophysiology data [30]. Currently, up to 27 electrophysiological properties are reported for 93 cell types, spanning over 2000 single pieces of published data extracted from journal article tables. *NeuroUnit* makes it easy to construct tests using values reported by NeuroElectro as reference data. Tests can be based upon data from single journal articles, or from ensembles of articles with a common theme (e.g. about a particular neuron type). The former is illustrated in Figure **??**.

Data about specific neuron types can be accessed using the NeuroElectro API and the statistics of that data (mean, standard error, and sample size are typically reported) enable judgement of model output. While NeuroElectro alone is not sufficient to judge all model aspects, it can nonetheless serve to validate basic features of most neurophysiology models, such as resting membrane potential, action potential width, afterhyperpolarization amplitude, etc. While the data may reflect artifacts of experimental preparation, such as electrode type or recording temperature, this metadata is increasingly visible on NeuroElectro, and test writers may identify the subset of data appropriate to the system being modeled. As NeuroElectro is the only publicly curated source of such data, it represents a key component for *NeuroUnit* test constuction. Continued development of the NeuroElectro API, through which data are systematically exposed to test authors, represents our **second collaboration** [31].

> Should there actually be an example of workflow with NeuroElectro here?

4.1.4 A Complete Pipeline

Although NeuroConstruct, NeuroTools, and NeuroElectro do not represent the only possible sources of models, capabilities, and test data, they provide an immediate point of entry into the neurophysiology community and powerful demonstration of our proposal. In the *NeuroUnit* repository [22] is a runnable script (*examples.py*) which demonstrates a complete testing pipeline. It selects a CA1 Pyramidal cell model from OSB, simulates it using NeuroConstruct, and tests the widths of the resulting action potentials, extracted and computed using NeuroTools, against NeuroElectro data obtained on-the-fly via API, using a *NeuroUnit* test class called `SpikeWidthTest`. The script instantiates `SpikeWidthTest` using the NeuroElectro data, tests the model, and finally computes and prints a test score.

4.1.5 Creating New Candidates, Capabilities and Tests

NeuroUnit provides base classes to enable rapid generation of candidates, capabilities, and tests for neurophysiology data. However these objects can also be created using few or none of the resources described here. All that is required is adherence to the *SciUnit* interface. For example, a Candidate could implement an `integrate` capability method by wrapping execution of a MATLAB script and a `get_spikes` capability method by parsing a .csv file on disk; a test could be initalized using empirical spike rates collected in the lab. From there, the basic workflow is the same as above.

4.2 Public Dissemination

Concurrent with these efforts, we are creating a community portal (*SciDash*) written using Django, a Python web framework. The source code for the web application and the framework will be developed openly and released under a permissive open source license continuously. An open release, rather than one implemented solely behind our portal, supports labs or groups that want to test models on their own machines as they are being developed, or to compare competing models on unreleased data. So long as a GitHub repository has been forked from the main SciDash repository [**?**], and is publicly visible (the default in GitHub), it will

be automatically indexed by and visible on SciDash. The community portal will thus point to models and tests as they become available, and display corresponding test results. Thus, there will exist a dedicated resource for evaluating the validity of models, and the progress of the SciUnit initiative, including the rate of community adoption, will be transparent.

4.3 Project Milestones

Year 1: We will continue *SciUnit* core development, and implement *NeuroUnit* tests using NeuroElectro data and capabilities using NeuroTools functions. At the end of this period we will have a manuscript in review, describing the idea and preliminary results, and have released a stable *SciUnit* Python module for general use.

Year 2: We will automating model testing for OSB, greatly increasing the number and scope of models tested. We will continue to define *SciUnit* model capabilities via collaborative development of NeuroTools, and aggregate new community-provided datasets to test the capabilities of OSB models. At the end of this period, OSB will support automated testing and result summaries, and *SciDash* will contain a broad array of models and tests to browse and visualize.

Year 3: We will extensively documentation *SciUnit* and associated tools to encourage adoption. We will actively promote its use at conferences and workshops. We will continue to write tests and specify (existing, published) models in NeuroML for execution and testing. At the end of this period we will have another manuscript in review, describing the results of the research program and promoting the *SciDash* portal. *NeuroUnit* will contain a sufficient array of usable tests to motivate considerable community interest and serve as an example for developers in other biology disciplines.

4.4 Challenges

Here we describe theoretical and practical challenges to implementation, and how they can be overcome.

4.4.1 Participation from Modeling Communities

Modelers may not want to expose model capabilities, a requirement of test-taking. We anticipate four solutions: **First**, interfacing a model to *SciUnit* requires only implementing selected model capabilities. This means identifying model outputs that satisfy a capability, and returning their values. It also means identifying model inputs used for parameterization, or run-time arguments, and providing a means to set them. However, each of these procedures may require only one line of code. Importantly, the modeler is not required to expose or rewrite any model flow control. **Second**, we will support multiple environments automatically by using NeuroML [10], a simulator-independent model description language. As we have shown, models described in NeuroML will be easy to test using *NeuroUnit*. Efforts are also underway to automate generation of NeuroML descriptions from the source code used by models written for popular simulators (Gleeson, personal communication). For the NEURON simulator, this process is largely complete. Thus for a large and growing number of models, modeler effort is already close to zero. **Third**, modelers have a strong incentive to use *SciUnit* to demonstrate in an umambiguous and public way that their models are consistent with data. Participation in public modeling competitions (see section 1.1) illustrates this incentive. Fourth, modelers have a strong incentive to use *SciUnit* during development (see TDD, above) to guarantee that on-going development does not break the correspondence between the model and the system it explains. A suite of popular tests will represent a "gold standard" by which modelers can judge their progress during development.

4.4.2 Participation from Experimental Communities

Experimentalists may not want to write tests derived from their data, or are not comfortable with writing code. We anticipate three solutions: **First**, rather than demand the use of special formats for data, each test will need only a list of required model capabilities (for selecting eligible models), essential experimental metadata (as run-time arguments to models) and a statistical summary of data (for scoring results) be written into each test. By definition each unit test is focused, and does not require the ability to do arbitrary computations on a data set. Suppose that one has evoked by intracellular current injection in a cell 100 action potentials and wishes to write a test concerning the width of these action potentials. Writing the test then consists of selecting ReceivesCurrent and ProducesActionPotentialShape capabilities (one line of

code each), typically computing the mean and variance of action potential widths in the data (one line of code), specifying the parameters of the current injection, e.g. the amplitude and the duration (two lines of code), and finally selecting a scoring mechanism from `sciunit.scores`, e.g. (colloquially) "Must be < 1 standard deviation of the mean value" (one line of code). Most of the heavy-lifting is done by the interface. An example can be found in `neurounit.tests.SpikeWidthTest`. **Second**, as data-sharing becomes more ubiquitous, this task can be distributed across a large number of scientists, including non-experimentalists interested in data analysis or testing their own models. **Third**, a large number of tests can be automatically generated using the NeuroElectro API, and the continued emergence of systematically organized data repositories will expand these possibilities. **Fourth**, a strong incentive to write tests for one's data exists: the ability to identify models that explain one's data, giving the data clear context and impact.

4.4.3   Diversity of Levels and Kinds of Models and Data

How can one framework deal with so many topics in biology? **First**, we solve this by providing an interface that allows modelers to express the capabilities which their model possesses. The set of all capabilities so described determines the range of tests that can be executed, and the set implemented by one model determines the range of tests that the model can take. Hierarchies of scale are embedded in the inheritance of capability classes. For example, an answer to the question "Does this model have action potentials" requires a "yes" answer to the question "Does this model have neurons". Consequently, the incompatibility of a test-requiring-action-potentials for a model-lacking-neurons is known without explicit tagging. Conversely, a model with a `HodgkinHuxley` capability also inherits a `VoltageGated` capability, because the former implies the latter. **Second**, expressing models in NeuroML naturally addresses diversity of levels (i.e. scales) because NeuroML is developed in "levels", with a hierarchical organization. Thus, models can be sub- or supersets of other models. Analogous hierarchies exist in SED-ML [40, 39], a general systems biology markup language. **Third**, testing across levels could also be implemented using the "Representational Similarity Analysis" (RSA) framework [34], requiring only that a model be capable of responding to a defined set of inputs (e.g. stimuli). A "similarity matrix" for responses within a model and across inputs defines a unique signature for that model, and can be the intermediate output of a test. Model scale becomes irrelevant, because test scores are based on goodness-of-fit between similarity matrices for model and experiment – these matrices can be compared directly no matter the model scale because their size depends only on the number of test inputs, not on system detail.

4.4.4   Appropriateness of Models for Validation

Not all models attempt to reproduce experimental findings. Some models serve as proof of concept that a dynamical system with have certain properties. We do not intend to challenge these proofs; however, tests that abstract away most experimental details can inform and illuminate such models. For example, rather than encoding specific experimental stimulus and response values, a test could simply evaluate a mapping between sets of numbers. Some abstract models may have unexpected homology to that mapping, thus highlighting the relevance of such models where it may otherwise have been missed. Alternatively, a model may make specific experimental predictions, but require significant contextual information not provided by a test. Rather than fail such models, we give them an "incomplete", i.e. no record of the test result is generated for such a model, in accordance with the model's scope.

4.4.5   Arbitrary Scoring Criteria for Tests

A raw test score is computed from goodness-of-fit to data, and a pass/fail mark from that score. At both stages there is room for arbitrary design choices that will benefit some models at the expense of others. **First**, many goodness-of-fit functions have, for a given input set, identically rank-ordered outputs, meaning that such design choices will rarely cause an otherwise passing model to fail and vice versa. For example, the function mapping Z-scores to p-values in monotonic. Indeed, one may ignore quantitative differences between model scores, with focus given instead to the rank ordering of those scores. **Second**, since test repositories are open (e.g. Fig. 2.3), it is straightforward to clone a test and change the statistical choices used for scoring. With each test transparent, a community can decide which test version is most appropriate. This process will be open and documented on *SciDash* via GitHub. Thus, the community decides on what

models should do, and the framework then determine whether it does those things.

4.4.6 Reliability of Data Underlying Tests

Unreliable data will lead to tests that even perfect models will not pass. **First**, it is incumbent upon the community to evaluate experimental design and techniques involved in producing data, and to discount data produced using questionable methods. *SciDash* will support community moderation, permitting users to rate and comment on tests, indicating their concerns. **Second**, models cannot fit perfectly to data when that data is a random draw of finite sample size from a "true" distribution. This can be addressed by making uncertainty in the data explicit, or by asking how well a data set can validate its own experimental replications [34]. Since a model cannot validate better (in expectation) than an identical experiment, test scores may be modified to reflect that goodness-of-fit matching that of experimental replication represents a "perfect" score.

4.4.7 Computational Efficiency

Large models may execute slowly, and subjecting such models to many tests could be computationally intensive. To lighten the burden, we prioritize minimal re-execution of the same model in *SciUnit*'s design. Sets of tests requiring as input the same model output only require the model to be executed once. Model output, generated by execution of the first test, can be cached for use by similar tests. For example, model execution instances could be stored as Sage worksheets in the corresponding repositories. We have also implemented caching in the `sciunit.utils` module, by storing candidate instances and associated execution data in an optional database backend.

4.4.8 Occam's Razor

A model's ability to explain data is typically weighed against its complexity – simpler models being better, *ceteris paribus*. Model complexity has many definitions, so *SciUnit* will report several complexity metrics, including: 1) the number of lines, instructions, or operations in the model; 2) memory use during model execution; 3) average CPU load during model execution; 4) number of model parameters. Larger, longer, and more expressive models may be considered more *complex* ([35]). The tradeoff between model validity and complexity can be reported in tabular form (e.g. Table 2); alternatively, a scatter plot, with the "best" models being in the high validity / low complexity corner of the plot, may be informative. The set of models which *dominate* all others, i.e. that have higher validity and lower complexity than their rivals, can be represented as a "validity vs. complexity" front, showing those models with the highest validity for a given level of complexity, a visualization familiar from the symbolic regression package Eureqa [36]. One then weighs the relative importance of validity vs. complexity for one's application.

4.4.9 Expansion Into Other Areas of Biology

After proving its utility in neurophysiology, we would like *SciUnit* to expand across neuroscience and then into other biological sciences. Since the core framework is discipline-agnostic, the only obstacles are community participation and model description. As with neurophysiology, community participation begins with enumerating the capabilities relevant to a sub-discipline, and then writing tests. Model description can expand within NeuroML (which already covers multiple levels and approaches within neuroscience) and *NeuroUnit* tools can begin to incorporate libraries for neuroimaging (NiBabel [37]), neuroanatomy (Neuro-HDF, [38]) and other sub-disciplines. SED-ML [39, 40] will facilitate expansion outside of neuroscience. This transition will be facilitated by parallel efforts in the NeuroML community to interface with SED-ML (Crook, unpublished).

5 Community and Educational Outreach

5.1 K-12 Education

The cornerstone of basic science education is learning the scientific method. However, the process by which the scientific method is applied in practice may be too informal to be recognizable to students. *SciDash* will provide a visible example of the scientific method in practice, determining which hypotheses (models) can withstand the scrutiny of evidence (can pass tests). Revision of hypotheses to match evidence will also be transparent, as *SciDash* will show and optionally group test results for model variants. Indeed, OSB already

tracks and exposes model revisions, so addition of a testing framework to that project would both implement and present the scientific method for selected models in neuroscience. The ability to visualize the scientific method at work from any computer in the world will represent a major step forward in science education.

To make the educational relevance of a validation testing framework more clear, consider teaching the history of advances in cosmology according to the schematic of Table 2: here we consider 5 astronomical models and 4 validation tests derived from relevant data. This record matrix, showing implied test scores, recapitulates the history of cosmology, showing the scientific method to be an on-going process. The Geocentric model of Claudius Ptolemy ([41]) passes a test derived from Babylonian records of planetary motion. Ptolemy's model fails all other tests shown here, including one constructed from Tycho Brahe's more meticulous measurements of planetary motion ([42]), being inconsistent with Ptolemy's notion of perfectly circular orbits. The Copernican Heliocentric model ([43]) also predicts circular orbits and thus fails Brahe's test, but is far simpler than the Ptolemaic model, dispensing with notions such as "epicycles". Kepler's laws of planetary motion permit and explain these elliptic orbits ([44]), as well as Galileo's unanticipated discovery of moons taking elliptic orbits around Jupiter ([45]). Newton's gravitational model passes these tests and does so by more succinctly by unifying Kepler's laws under one principle: gravity ([46]). Newton's model was challenged by the observation of perihelion precession of Mercury ([47]). General Relativity resolved this problem ([48]). Interestingly, Einstein explicitly proposed 3 tests of his theory, of based on Le Verrier's precession data. While this account may be simplistic compared with testing practice in modern biology, it would be appropriate for teaching the scientific method to a high school or college student. Although teaching is not part of the budget request for this proposal, future proposals focused on using *SciUnit* to teach the process of science will be forthcoming.

5.1.1 Journalistic Media

When reporting on new scientific theories, lay members of the media have no reliable way to determine the importance or quality of such theories. Direct consultation of scientists may be unreliable, due to bias or lack of focused

Table 2: A record matrix illustrating models and data-driven tests from the history of cosmology. Each row is a model; each column is a test.

	Complexity	**Babylon**	**Brahe**	**Galileo**	**Le Verrier**
Ptolemy	Medium	Pass	Fail	Fail	Fail
Copernicus	Low	Pass	Fail	Fail	Fail
Kepler	Medium	Pass	Pass	Pass	Fail
Newton	Low	Pass	Pass	Pass	Fail
Einstein	High	Pass	Pass	Pass	Pass

expertise. However, *SciDash* would – for any model that it covers – provide an unbiased way for media members to identify the scope and validity of a model, and contrast this with previous efforts. With the original sources of models and tests well-documented, contributors could be contacted for comment or explanation. Site-embedded commentary provided by modelers and test-writers should also be a helpful media resource.

The competitive nature of *SciDash* will be appealing both to scientists and to the lay community. Competitions for machine learning, solar cars, and autonomous vehicles already draw considerable media coverage and we believe that the brain is of no less interest to the public. Public competitions also welcome teams who may not yet have academic credentials, such as students, both to learn the craft and possibly to demonstrate key insights not previously recognized in the professional scientific community.

6 Personnel and Coordination

A scientific software framework succeeds in proportion to its rate of adoption, which is driven in part by: **Relevance** to the needs of both experimental and theoretical scientists; **Quality** of architecture and usability; **Conformance** to accepted data and modeling standards in the targeted communities; **Integration** with existing software tools; **Applicability** to outstanding questions in a field; **Community** access (i.e. an open and accessible internet presence).

6.1 Personnel

We have the appropriate team members to meet the criteria above.

[PI: Richard C Gerkin, PhD]
Title: Assistant Research Professor, School of Life Sciences, Arizona State University (ASU)
Expertise: Neurophysiology, computational models, informatics, web development.
Role (Time): Will coordinate all project activities (25 hrs/week).
Detail: As the main PI, he will coordinate all project activities, and be responsible for all output including code, databases, and manuscripts. He will also: coordinate with the NeuroElectro project and solicit other sources to obtain physiology data, format and annotate this data, and construct tests from the data (experience as both an experimentalist and a modeler will facilitate this objective (**Relevance**)); write NeuroML-related bindings (e.g. `NeuroConstructModel`) (**Conformance**); subject models from OSB (and similar models publicly available) to these tests; develop and maintain the *SciDash* website (**Community**); coordinate with the developers of OSB to implement automated *SciUnit* testing.

[Co-PI: Sharon M Crook, PhD]
Title: Associate Professor, School of Mathematical and Statistical Sciences, School of Life Sciences, ASU
Expertise: Neuroinformatics, Computational Modeling
Role (Time): Interface of *SciUnit* with NeuroML and Open Source Brain (2 hrs/week)
Detail: Access to the broader neuroscience community requires the use of an accepted model description standard. We chose NeuroML due to its advanced state and broad coverage of multiple scales. Dr. Crook is funded by NIH to maintain and support NeuroML and will help us use NeuroML for model specification (**Conformance**).

[Co-PI: Jonathan Aldrich, PhD]
Title: Associate Professor, School of Computer Science and Institute for Software Research, CMU
Expertise: Software engineering, software verification and validation, human factors
Role (Time): Guide overall software architecture (2 hrs/week)
Detail: Dr. Aldrich will supervise and train Mr. Omar. His expertise in software design for science and engineering applications will be key in this role, and assure **quality**. While Mr. Omar will be writing code, Dr. Aldrich will be providing guidance on the overall structure of the implementation. Dr. Aldrich is already Mr. Omar's graduate supervisor (funded by other sources), but for this project Dr. Aldrich will take on the additional training responsibilities specific to the focus of this proposal.

[Junior Personnel, Cyrus Omar]
Title: Predoctoral Fellow, School of Computer Science, CMU
Expertise: Computational models and computer science
Role (Time): Core framework design and development (25 hrs/week)
Detail: Mr. Omar is a senior graduate student who has expertise in both computational modeling in neuroscience and software infrastructure for science. He will be responsible for guiding the overall software architecture of the project (**Quality**) and developing and maintaining the core *SciUnit* Python module. He will get feedback from Dr. Gerkin about practicalities of model testing, which will inform revisions to the code.

[Collaborator, R. Angus Silver, PhD]
Title: Professor of Neuroscience and Wellcome Trust Senior Research Fellow, Faculty of Life Sciences, University College, London
Expertise: Neurophysiology, computational neuroscience and neuroinformatics
Role: Interface of *SciUnit* with NeuroML and Open Source Brain
Detail: A frequent collaborator of Crook, Silver maintains Open Source Brain (OSB), the largest standards-driven respository for models in neuroscience, for which he is funded by The Burroughs Wellcome Trust. The OSB team has welcomed us to integrate *SciUnit* into OSB for automated model testing(**Integration**). This exposes a wide range of neuroscience models specified in NeuroML (**Applicability**), and in turn exposes the project to the international neuroscience modeling community, serving as proof of concept for expansion into other areas of biology.

[Collaborator, Shreejoy Tripathy, PhD]
Title: Research Fellow, Dept. of Biological Sciences, CMU
Expertise: Neuroinformatics, computational neuroscience, data mining.
Role: Interface of *SciUnit* with NeuroElectro
Detail: A frequent collaborator of Gerkin, Tripathy maintains The NeuroElectro Project, the largest data repository concerning electrophysiological properties of neurons. This provides us a wide range of data for testing neuroscience models. Tripathy will facilitiate **integration** with NeuroElectro through further development of the API.

6.2 Means of Coordination

Dr. Gerkin trained in Pittsburgh and maintains ongoing collaborations with Dr. Aldrich, Mr. Omar, and Dr. Tripathy. Dr. Gerkin works at ASU and regularly meets with Dr. Crook for "Math Biology" group meetings there. Because the existing projects that form the basis for each collaboration are a) well-documented, and b) available under an open license, there should be no barriers to code sharing and collaborative development. The popular and powerful Github platform will be used for code development and communication between developers.

7 Results from Prior NSF Work

Sharon Crook was PI on an NSF grant that...

Jonathan Aldrich has been a PI or co-PI on four prior NSF grants. NSF CAREER grant CCF-0546550, "Lightweight Modeling and Enforcement of Architectural Behavior" (2006-2010), focused on the Scholia approach to extracting a run-time architecture from object-oriented code, as well as analyzing architectural behavior. 12 major journal or conference publications – note that major conferences are more prestigious than journals in CS – as well as 12 workshop articles emerged from the grant, including (Abi-Antoun et al, 2007, Abi-Antoun et al, 2008; Abi-Antoun and Aldich, 2008, Abi-Antoun and Aldrich, 2009). NSF grant CCF-0811592, "CPA-SEL: Practical Typestate Verification with Assume-Guarantee Reasoning" (2008-2011), funded fundamental advances in lightweight typestate verification approaches. These were described in 7 major publications, including (Bierhoff and Aldrich, 2007; Beckman et al, 2008; Bierhoff et al, 2009; Jaspan and Aldrich, 2009; Sunshine et al, 2011). NSF grant CCF-1116907, "SHF:Small:Foundations of Permission-Based Object-Oriented Languages" (2011-2014), is only a year old, but has already yielded a paper in the premier publication venue for programming languages describing a novel type system for modularly reasoning about aliasing in object-oriented programs (Naden et al, 2012). These grants resulted in several open-source software tools for research and education, including ArchJava, Plural, Crystal, and SASyLF (see references cited). Aldrich has just been awarded NSF grant "TUES: Collaborative Research: Teaching Software Modularity through Architectural Review," but only a month has passed since that award.

References

[1] *http://www.crcns.org/*.

[2] *http://www.opensourcebrain.org*.

[3] *http://www.kaggle.com*.

[4] Jennifer Carpenter. May the best analyst win. *Science (New York, N.Y.)*, 331(6018):698–699, February 2011. PMID: 21311005.

[5] Renaud Jolivet, Felix Schürmann, Thomas Berger, Richard Naud, Wulfram Gerstner, and Arnd Roth. The quantitative single-neuron modeling competition. *Biological Cybernetics*, 99(4):417–426, November 2008.

[6] J J Hopfield and C D Brody. What is a moment? "Cortical" sensory integration over a brief interval. *Proceedings of the National Academy of Sciences of the United States of America*, 97(25):13919–13924, December 2000. PMID: 11095747.

Get previous NSF support section from Sharon.

Replace Jonathan's citations here and below with citation codes.

Document officially ends here for page count purposes. Must be under 15 pages.

[7] *http://www.diademchallenge.org.*

[8] *http://www.the-dream-project.org/.*

[9] *http://neuroml.org.*

[10] Padraig Gleeson, Sharon Crook, Robert C Cannon, Michael L Hines, Guy O Billings, Matteo Farinella, Thomas M Morse, Andrew P Davison, Subhasis Ray, Upinder S Bhalla, Simon R Barnes, Yoana D Dimitrova, and R Angus Silver. NeuroML: a language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS computational biology*, 6(6):e1000815, June 2010. PMID: 20585541.

[11] *http://www.neuron.yale.edu/neuron.*

[12] Karthik Ram. Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine*, 8(1):7, 2013. PMID: 23448176.

[13] *http://www.djangoproject.com.*

[14] *http://github.com/cyrus-/sciunit.*

[15] *http://www.r-project.org/.*

[16] *http://www.mathworks.com/products/matlab/.*

[17] *http://opensourcebrain.org/projects/ca1pyramidalcell.*

[18] *http://www.scidash.org.*

[19] W Stein and D Joyner. SAGE: system for algebra and geometry experimentation. *ACM SIGSAM Bulletin*, 39(2), 2005.

[20] *http://www.sagemath.org/.*

[21] *www.sagenb.org.*

[22] *http://github.com/rgerkin/neurounit.*

[23] *http://www.neuroconstruct.org.*

[24] Padraig Gleeson, Volker Steuber, and R Angus Silver. neuroConstruct: a tool for modeling networks of neurons in 3D space. *Neuron*, 54(2):219–235, April 2007. PMID: 17442244.

[25] *http://senselab.med.yale.edu/modeldb.*

[26] Michael L Hines, Thomas Morse, Michele Migliore, Nicholas T Carnevale, and Gordon M Shepherd. ModelDB: a database to support computational neuroscience. *Journal of computational neuroscience*, 17(1):7–11, August 2004. PMID: 15218350.

[27] *http://github.com/rgerkin/neuroConstruct.*

[28] *http://neuralensemble.org.*

[29] *http://www.neuroelectro.org/.*

[30] SJ Tripathy, J Saviskaya, RC Gerkin, and NN Urban. NeuroElectro: a database describing the electro-physiology properties of different neuron types. *Neuroinformatics Meeting*, 2012.

[31] *http://bitbucket.org/rgerkin/neuroelectro_org.*

[32] *http://www.neuron.yale.edu/neuron*.

[33] NT Carnevale and ML Hines. *The NEURON Book*. Cambridge University Press, Cambridge, UK, 2006.

[34] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008. PMID: 19104670.

[35] TJ McCabe. A complexity measure. *IEEE Transactions on Software Engineering*, 2(4):308–320, 1976.

[36] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science (New York, N.Y.)*, 324(5923):81–85, April 2009. PMID: 19342586.

[37] *http://nipy.org/nibabel*.

[38] *http://neurohdf.readthedocs.org*.

[39] M Hucka, A Finney, H M Sauro, H Bolouri, J C Doyle, H Kitano, A P Arkin, B J Bornstein, D Bray, A Cornish-Bowden, A A Cuellar, S Dronov, E D Gilles, M Ginkel, V Gor, I I Goryanin, W J Hedley, T C Hodgman, J-H Hofmeyr, P J Hunter, N S Juty, J L Kasberger, A Kremling, U Kummer, N Le Novère, L M Loew, D Lucio, P Mendes, E Minch, E D Mjolsness, Y Nakayama, M R Nelson, P F Nielsen, T Sakurada, J C Schaff, B E Shapiro, T S Shimizu, H D Spence, J Stelling, K Takahashi, M Tomita, J Wagner, J Wang, and SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, 19(4):524–531, March 2003. PMID: 12611808.

[40] *http://sed-ml.org/*.

[41] Cladius Ptolemy. *The Almagest*. 150.

[42] Johannes Kepler. *The Rudolphine Tables*. 1627.

[43] Nicolas Copernicus. *De revolutionibus orbium coelestium*. 1543.

[44] Johannes Kepler. *Astronomia Nova*. 1609.

[45] Galileo Galilei. *Siderius Nuncius*. 1610.

[46] Isaac Newton. *Philosophiae Naturalis Principia Mathematica*. 1687.

[47] Urbain Le Verrier. Lettre de m. le verrier à m. faye sur la théorie de mercure et sur le mouvement du périhélie de cette planète. *Comptes rendus hebdomadaires des séances de l'Académie des sciences (Paris)*, 49:379–383, 1859.

[48] Albert Einstein. The foundation of the general theory of relativity. *Annalen der Physik*, 49(7):769–822, 1916.

> Remove bibliography with nobibliography tag before submission

Todo list