Computational Neuroinformatics: Toward Information Theoretic Analyses of Neural Coding

David H. Goldberg[1], Jonathan D. Victor[2], and Daniel Gardner[1]

[1]Lab of Neuroinformatics, Dept. of Physiology, [2]Dept. of Neurology & Neuroscience, Weill Cornell Med. Coll., NYC, NY

1032.9

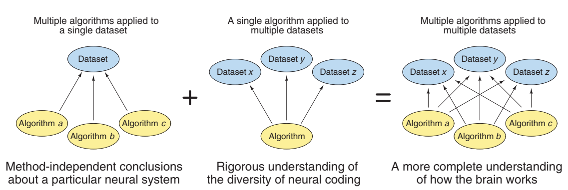
Toward Computational Neuroinformatics

Understanding neural coding—how neurons represent and process information with spike trains—is a central goal of neuroscience. Information theory, originally developed as a means of studying modern communication systems, is now being applied to questions of neural coding by many laboratories. We present a planned computational neuroinformatic resource that combines the neurophysiology data repository at `neurodatabase.org` with a suite of complementary methods for information-theoretic analyses.

Upon completion, this `neurodatabase.org` resource will provide:

- Neurophysiological data from many preparations, modalities, and experimental paradigms, stored in a standard format with detailed metadata.
- A means for experimenters to share their data, encouraging and ensuring compliance with the NIH data sharing mandate.
- Mathematically sophisticated algorithms for the analysis of spike train and other informative neural data.
- A site for researchers to disseminate analytical algorithms, facilitating feedback from, and adoption by, the neuroscience community.
- Access to a dedicated parallel computer for computationally intensive analyses.

To Understand Neural Coding, Apply Multiple Methods to Multiple Datasets



1. Toward an Understanding of Neural Coding

The application of information theoretic concepts to the study of neural coding is non-trivial, because straightforward estimates of information theoretic quantities often require prohibitively large amount of data. Alternative methods reduce the amount of requisite data by making assumptions about the neural system under study. Because we lack *a priori* knowledge about the appropriateness of these assumptions, it is essential that multiple methods be made available to analyze datasets from multiple systems.

The synthesis of neuroinformatics and computational resources ushers in a new era of computational neuroinformatics in which neuroscientists will be able develop a rigorous understanding of neural coding that encompasses a diverse array of preparations and transcends particular methods.

Our Support and Our Standards:

Human Brain Project / Neuroinformatics research in the Laboratory of Neuroinformatics is funded via MH68012 from NIMH, NINDS, NIA, NIBIB & NSF, and MH57153 from NIMH & NINDS.

The Laboratory of Neuroinformatics supports and proposes as standards: Open data, open exchange and open analysis–via open source and open platform–towards open discovery for neuroscience.

LNI Laboratory of Neuroinformatics
Weill Medical College of Cornell University

Modular Computational and Informatic Components Support Standalone Analyses as well as Integrated Resources at Neurodatabase.org

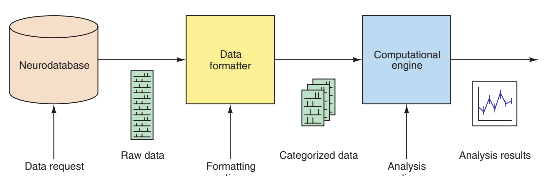
Method	Stimulus applicability	Suitability for multichannel/ multiunit data*	Suitable for continuous signals	Assumptions concerning nature of code#	Provides a description of the code
direct	general	low	no	virtually none	no
reconstruction	time series	favorable	yes	Volterra	yes
metric space	general	very favorable	no	specific models	yes
embedding	general	favorable	yes	continuous	minimal
power series	general	favorable	no	Volterra	yes
spectrotemporal	general	favorable	yes	continuous	yes
bottleneck/codebook	general	very favorable	yes	continuous	yes

* low: exponential growth with number of channels
* favorable: polynomial growth (for low-order models)
* highly favorable: no growth

\# from most to least general: virtually none, continuous, Volterra, specific models

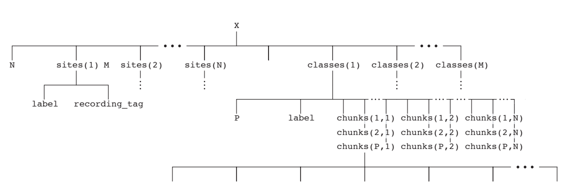
2. Information-Theoretic Analytical Methods

The established algorithms in this suite vary in their computational and informatic requirements, the nature of the assumptions they make about neural coding, and their applicability to multichannel data. Each analytical method will provide a set of criteria to ensure appropriate applicability and meaningful results, including type of signal (continuous/discrete, single/multi-channel) and quantity of data required. Reconstruction is the only method that requires information about the stimulus beyond a categorical classification.



3. Modular Design Separates Parsing, Grouping, and Analyses of Data

The computational resource must accommodate data from a vast array of preparations and experimental paradigms. Users of the resource will be able analyze data submitted by members of the community, performing analyses that the data submitter may not have anticipated. This can be accomplished by reorganizing the archived data with the aid of accompanying metadata. The computational engine is not concerned with this process; it requires only that the data be segregated into stimulus categories.

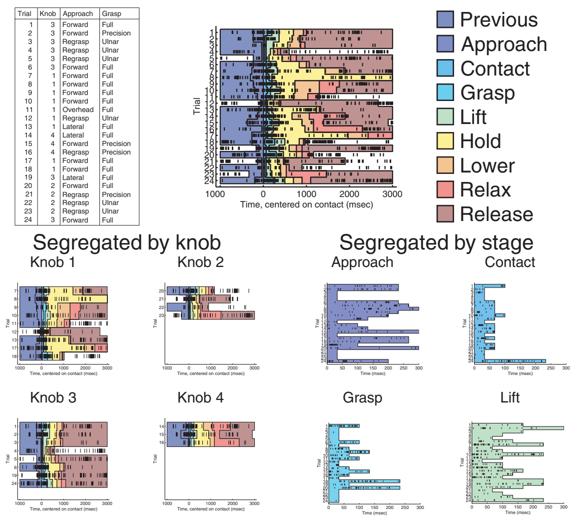


4. An Input Data Container That Facilitates Information-Theoretic Analyses

We have developed a standard categorical data container that is:

- Agnostic to the procedure that generated the data
- Appropriate for both single and multichannel data
- General enough to handle time-stamped spike train data as well as field potentials and other continuously-sampled signals.

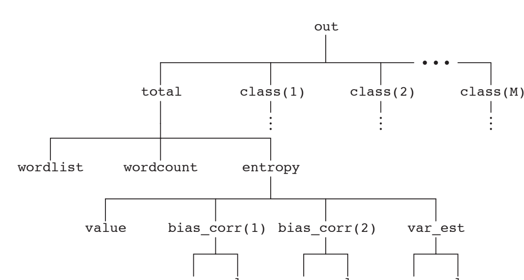
Trial	Knob	Approach	Grasp
1	3	Forward	Full
2	3	Forward	Precision
3	3	Regrasp	Ulnar
4	3	Regrasp	Ulnar
5	3	Regrasp	Ulnar
6	3	Forward	Full
7	1	Forward	Full
8	1	Forward	Full
9	1	Forward	Full
10	1	Forward	Full
11	1	Overhead	Full
12	1	Regrasp	Ulnar
13	1	Lateral	Full
14	4	Lateral	Full
15	4	Forward	Precision
16	4	Regrasp	Precision
17	1	Forward	Full
18	1	Forward	Full
19	3	Lateral	Full
20	2	Forward	Full
21	2	Regrasp	Precision
22	2	Regrasp	Ulnar
23	2	Regrasp	Ulnar
24	3	Forward	Full



5. Neurodatabase.org Design Faciliates Organizing Archived Data for Post-hoc Analyses

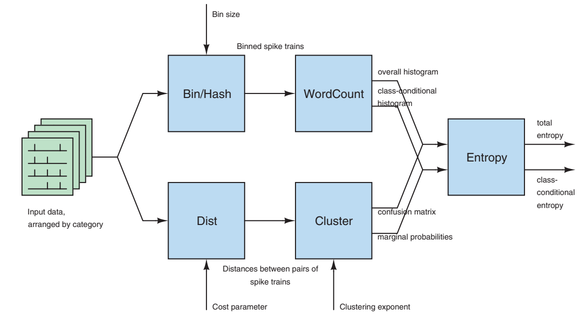
Data categorization is a complex and non-trivial task. The design of `neurodatabase.org` facilitates inclusion of descriptive metadata, stimulus time markers or replicas, and data groupings.

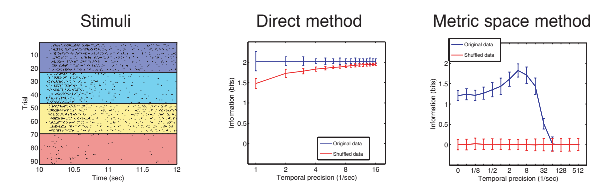
This example, drawn from `neurodatabase.org`, groups spike trains from cortical neurons recorded in an experiment in which a monkey grasped, lifted, held and lowered knobs in response to visual cues. Metadata supplied by the submitter and codified by the database and its BrainML data description language enabled the post-hoc segregations, by knob grasped and by behavioral state, shown in the lower panels (see Debowy et al., in *The Somatosensory System: Deciphering the Brain's Own Body Image* 2002).



6. An Output Data Container That Provides an Intuitive Structure for Results of Analyses

A specialized output container stores discrete histograms and associated statistics, commonly used for estimation of entropy. The container accommodates multiple bias corrections and variance estimates with each estimated entropy.





Taste stimuli representing four basic taste qualities are presented, and spike trains are recorded from the NTS of anesthetized rats (Di Lorenzo & Victor, *J. Neurophys.* 2003).

The direct method, which makes no assumption of the underlying code, often requires more data than are easily obtainable. Here, results are not meaningful, even for large bin sizes, as demonstrated by the large amount of (apparent) information in shuffled surrogate datasets.

The metric space method assumes that neural codes can be parametrized by the importance of spike timing. It provides a more useful estimate of information from the available data. It also characterizes *how* the information is transmitted.

7. An Initial Modular Implementation of Two Information-Theoretic Analytical Methods

Methods are implemented as a sequence of basic modules that can easily be replaced with more sophisticated modules as they become available. Here, methods share a module that estimates the entropy of a discrete histogram and provides bias and confidence limits.

The initial two methods are :

- Direct method (Strong et al., *Phys. Rev. Lett.* 1998).
Spike trains are divided into non-overlapping bins and the spikes in each bin are counted. The number of spikes in the bins are considered to be letters in a word. Entropies are estimated from the histogram of the total and class-conditional word probabilities.

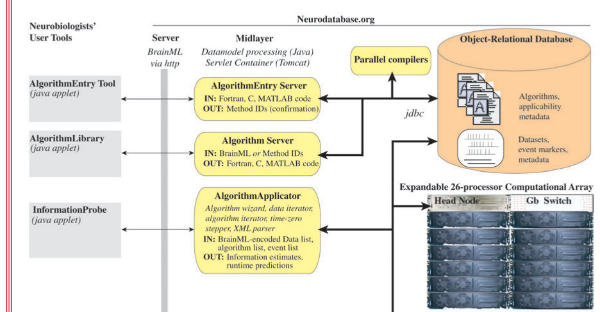
- Metric space method (Victor & Purpura, *Network: Comput. in Neural Sys.* 1997).
Distances between pairs of spike trains are computed under an assumption of the neural code (e.g., dependence on spike count, spike time, etc.). The spike trains are clustered according to these distances and mutual information is computed based on the extent of the stimulus-dependent clustering.

Two Computational Neuroinformatic Resources for Analysis of Neural Coding

- We Have Begun Development of a Portable Spike-train Analysis Toolkit.

Modules will be supplied as an integrated portable computational suite for information-based analyses. This toolkit will be suitable for use on local desktop workstations and individual parallel clusters.

Users will have the option of interfacing with the data repository and data formatter at `neurodatabase.org`, or with local data (if stored in a valid data container). The open source code is written in C for portability and supplied with MEX wrappers so that modules can be called from the Matlab command line.

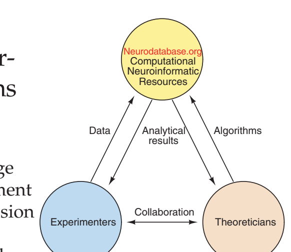


8. An Integrated Computational Neuroinformatic Resource at Neurodatabase.org

Enhancing `neurodatabase.org`, the modules will also be used to form an open, Web-accessible comprehensive resource for computational neuroinformatics. In-development Java tools, midlayer servers, and an algorithm database will integrate modules and provide a graphical user interface for acquiring, supplying, and applying algorithms as well as data.

These resources include a parallel array for the computationally intensive analyses required by several of the algorithms we are implementing. With the advent of multielectrode recording technology, researchers are exploring population codes that exploit the concerted activity of several neurons, requiring even more computational power, exceeding the capabilities of a conventional desktop computer. We have dedicated a 26-processor Beowulf cluster as a parallel computational resource for community analysis of databased data.

9. Towards Experimenter-Theoretician Collaborations



Using these resources:

- Experimenters will be able to analyze their data with cutting-edge analytical techniques, an enhancement that will further encourage submission to the database and data sharing.
- Theoreticians will be able to apply newly-developed analytical techniques to a diverse collection of datasets and compare the results to those derived from established methods.

By eliminating technical barriers such as incompatible data formats, platform-dependent software, and lack of computational power, `neurodatabase.org` will serve as a catalyst for collaboration between these two complementary groups.