Data Mining and Neurocomputational Modeling in the Neurosciences

Kimberly Kirkpatrick, Professor, Psychological Services, Kansas State University

The era of "big data" and the increasing focus on analytics is impacting most scientific disciplines, including research in cognitive and behavioral neuroscience. The growth of complexity of experimental data sets has led to the need for increased emphasis on data reduction and data mining techniques. An important companion to data mining is neurocomputational modeling, which is increasing in importance in the neurosciences. Such techniques such as data mining and modeling require the use of technical computing applications such as MATLAB, which can create barriers for incorporating students into the research process. The present paper discusses the challenges faced in the big data era of neuroscience and provides some ideas for tools than can promote success by researchers, and their students, in facing such challenges.

Introduction

The overarching mission of modern behavioral and cognitive neuroscience research is to pinpoint the neurobiological mechanisms of that underlie complex cognitive processes and the resulting behaviors. Cognitive neuroscientists typically focus on studying human populations, whereas behavioral neuroscientists typically focus on animal models of human behavior. There have been a number of exciting breakthroughs in the neurosciences that have led to the expansion of the complexity and size of data sets that are now typically collected in experimental studies.

One major trend is the growth and refinement of techniques such as fMRI, MEG, and EEG for cognitive neuroscience and electrophysiology, optogenetics, cyclic voltammetry, and circuit tracing for behavioral neuroscience, just to name a few. Many of the new techniques measure (or regulate) brain activity with high spatial and/or temporal specificity while also studying behavioral responses unfolding in time, thereby resulting in larger and more complex data sets. In relation to these advancements, there has generally been an increased focus on systems and circuits, which also require more complex data to understand. Additional trends relate to the examination of the interaction of complex processes, such as multiple cognitive functions operating together for complex tasks. And, behavioral neuroscience has increasingly come to include different levels of analysis within the same research program from the molecular (cellular) to molar (whole organism functioning) level. All of these trends have resulted in the need for new approaches to data mining and data analysis, which is a focus of the present paper.

In addition, there has been an increased emphasis on computational

modeling, both in terms of process modeling and statistical modeling of complex data sets. Process models provide a means for understanding the computational processes performed by the nervous system that underlie complex behaviors, leading to deeper insights into neural and cognitive mechanisms of behavior. In addition, computational modeling can supply a bridge between the neurobiological (e.g., neuronal firing patterns) and behavioral data, thereby providing a guide for brain-behavior translation.

The new trends in collecting large data sets, coupled with mining and modeling those data sets are heavily mirrored in funding priorities by the major funding agencies such as NIH, NSF, and DOD, and data mining and computational modeling are becoming increasingly necessary tools for incorporation into viable grant applications.

The present paper discusses some techniques and tools that can be utilized for data mining and neurocomputational modeling in the neurosciences, and how to incorporate those techniques into a research environment that involves training graduate and undergraduate students in neuroscience research.

Data mining

Data mining refers to the process of knowledge discovery in data bases. When dealing with large data sets, some degree of data reduction and/or selection is necessary first step in approaching data analysis. This can involve extracting summary measures of the data, smoothing the data, and selecting subsets of the data that directly address experimental hypotheses. As data mining does involve elements of data selection, it is imperative that data mining follow a hypothesis driven approach. Prior to beginning any data mining, researchers should develop a set of target measures dictated by their hypotheses and experimental design and a set of predictions of outcomes for those measures. Data mining should follow a surgical approach, and should ideally involve the use of multiple measures that reveal converging evidence of the true patterns in the raw data.

Three ways of collecting data. As a simple demonstration of the dramatic changes that have occurred in the methods of data collection in the neurosciences, one can look at the evolution of data collection within a basic behavioral paradigm that is used widely in behavioral neuroscience for both behavioral and neurobiological research, classical conditioning (Pavlov, 1927). In a standard conditioning study, a stimulus (e.g., a tone) is delivered for a specified duration (e.g., 10 s), followed by an outcome (e.g., food delivery) and then followed by an intertrial interval. Repeated presentations of the tone-food deliveries result in the emergence of responses during the tone (e.g., food cup checking responses) that indicate learning on the part of the individual. A diagram of this simple procedure is shown in Figure 1.

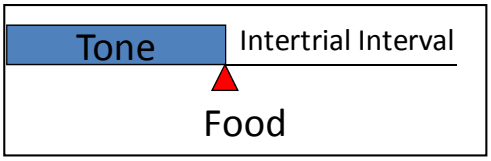

Figure 1. A diagram of a simple classical conditioning procedure that is used in behavioral neuroscience research to study principles of learning. The tone is followed by food and then an intertrial interval. Repeated presentations result in learning that the tone predicts food.

Traditionally, data collection in this simple procedure involved recording the total number of responses during the tone, and then dividing by the total time of tone presentation to obtain a measure of response rate. To accomplish this, one only needed a simple counter which would activate whenever the subject (e.g., the rat) responded while the tone was on. Figure 2 presents an example of typical response rate data from a study by Jennings, Bonardi, and Kirkpatrick (2007). In their study, different groups of rats were given tones of different durations (10, 20 or 40 s) followed by food and the measurement of learning was food cup checking responses. Figure 2 displays the mean response rate in responses per minute during the tone for each group. As seen in the figure, groups with shorter tones responded more than groups with longer tones. This indicates that the rats learned the tone-food connection and that shorter tones were more effective predictors of food delivery. This result has been reported on numerous occasions in other species and with other classical conditioning paradigms (Bitterman, 1964; Black, 1963; Gibbon, Baldock, Locurto, Gold, & Terrace, 1977; Kirkpatrick & Church, 2000; Salafia, Terry, & Daston, 1975; Schneiderman & Gormezano, 1964).

A more recent development in data collection in the same procedure involves obtaining finer-grained measures of responding during the course of the tone stimulus. For this measurement procedure, the tone is divided into several time bins of a specified duration (e.g., 1 s) and responses are collected within each bin comprising the tone and then transformed into a response rate in each bin. This procedure requires a response counter for each bin with a pointer that moves forward as a function of time in the tone. Figure 3 portrays data typical of those collected with this method from Jennings et al. (2007). As seen in the figure, responding during the tone is non-uniform, with a generally low rate at the beginning of the tone and a ramping increase in response rate over time, reaching a maximum near the end of the tone. This pattern indicates that the rats learned more than just a simple connection between the tone and food; they also learned the tone duration. This is now a well-established finding that has been reported in numerous species and other classical conditioning paradigms (e.g., Balsam, Drew, & Yang, 2002; Balsam, Sanchez-Castillo, Taylor, Van Volkinburg, & Ward, 2009; Kirkpatrick & Church, 2000). One factor worth noting is that this observation would not be possible with the first data

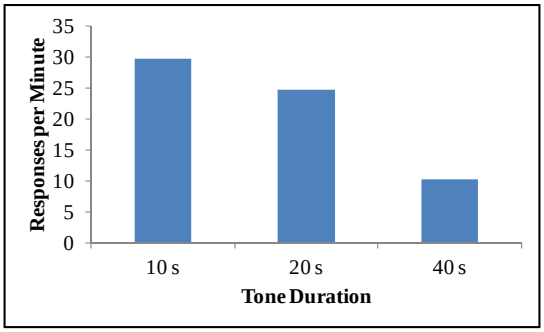

Figure 2. Responses per minute during tone stimuli of different durations (10, 20 or 40 s). Rats responded more during shorter tones. This figure provides an example of typical results obtained using data collection method 1. Adapted from Jennings, Bonardi, and Kirkpatrick (2007).

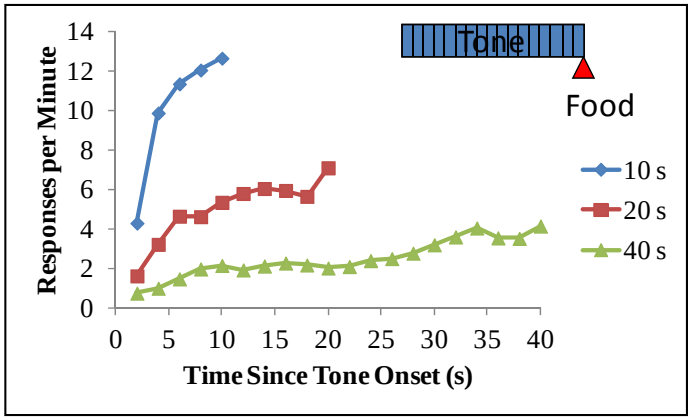

Figure 3. Responses per minute during the tone conditioned stimulus (CS) as a function of time since CS onset for 10, 20 or 40-s tones. The inset displays the division of the tone into time bins. Rats increased their response rates over the course of the tone, reaching a maximum near the end. Adapted from Jennings, Bonardi, and Kirkpatrick (2007).

collection method, where responding is aggregated across the tone duration.

Even more recent advancements in data collection have occurred over the past 10-15 years, with the increased availability of cheap data storage options. This has resulted in even finer-grained data collection methods which allow for more detailed assessments of brain and/or behavioral processes. One method that is used for detailed data collection is time-event codes. An example a time-event code data stream is shown in Figure 4. The numbers to the left of the decimal point are time stamps in milliseconds (ms), which are cumulative during the experimental session. The event codes appear to the right of the decimal point. Different event codes are used to mark different responses and different stimuli.

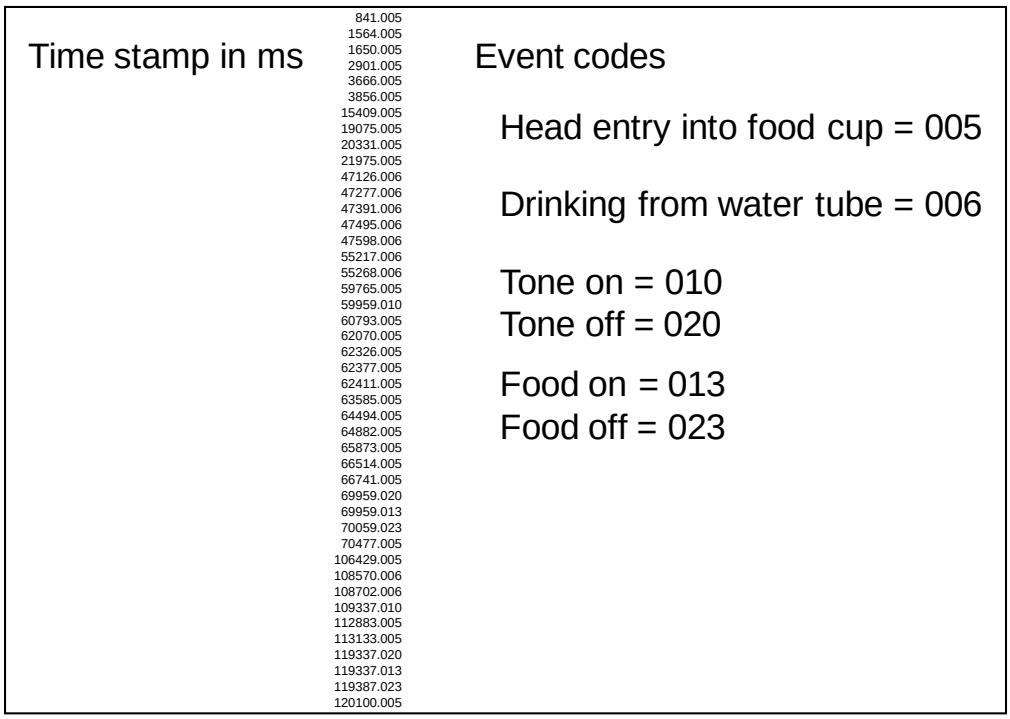

Figure 4. A small segment of a data file collected using time-event codes. The numbers to the left of the decimal point are time stamps in milliseconds (ms) that accumulate over the session, and the numbers to the right of the decimal point are event codes, with event code definitions provided in the figure.

Time-event codes allow for detailed analysis of the data that extend beyond the capabilities of the two previous methods. For example, analyses of behaviors other than the target behavior are possible (Reid, Bacha, & Morán, 1993), analyses of behavior during the intertrial interval can be conducted (Kirkpatrick & Church, 2000), multiple measures of timing behavior during the tone can be extracted (Guilhardi & Church, 2004), and trial-by-trial response dynamics can be examined (Church, Meck, & Gibbon, 1994; Gibbon & Church, 1990). In addition, it is possible to produce the previously described summary measures from the time-event code data. Indeed, Figures 2 and 3 were created from time-event code data.

Tools for data mining. As the size and complexity of data sets grows, this presents new challenges for including students, particularly undergraduates and early career graduate students, in the data analysis process. One means of mitigating this problem is to develop multi-use data mining applications that can be access through a graphical user interface (GUI). Technical computing languages such as MATLAB (The Mathworks, Natick, MA) allow for development of custom GUIs for data mining. MATLAB offers excellent tools for GUI development

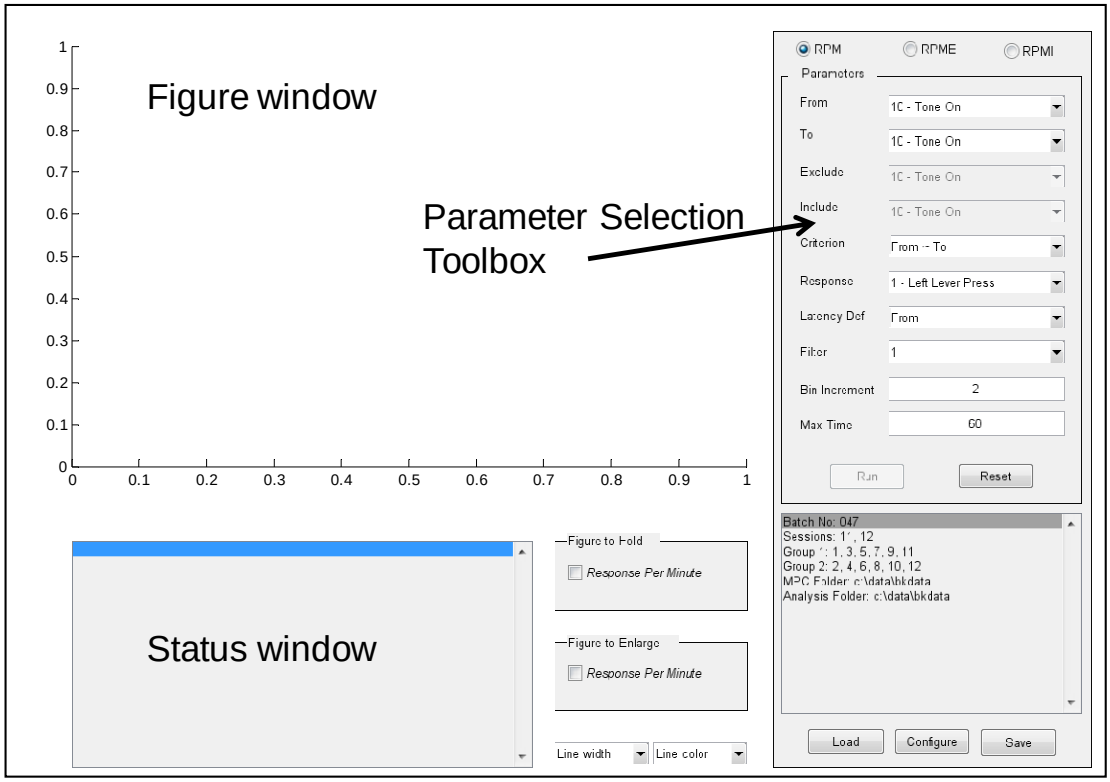

Figure 5. An example graphical user interface (GUI) for data mining applications, created in MatLab. This GUI contains a parameter selection toolbox for entering information about the experimental components and desired stimuli and responses to include in the analysis. The status window provides updates on the analysis progress during execution of the GUI. The figure window provides an editable plot of the output from the data analysis. GUIs can provide an excellent route for promoting the ability of undergraduate and graduate students to engage in data mining.

through the GUIDE programming environment. This can be supplemented with back-door programming of custom functions and scripts for data extraction and data reduction that are accessed through the GUI. MATLAB also provides functions for statistical analyses through various toolboxes such as the Statistics, Curve Fitting, and Optimization toolboxes. In addition, there is a well-developed community surrounding MATLAB from which have emerged several freely available toolboxes such as the MATLAB to R toolbox which provides an interface for running statistical analyses in R through the MATLAB environment. GUIs developed in MATLAB can be used to run analyses in R as well.

An example GUI is shown in Figure 5. This GUI is designed to analyze the timing of responses during a window of time defined by the From and To event codes selected from drop-down menus in the parameter selection toolbox. The target response is also selected from this menu as well as the bin sizes. Information regarding the location of data files, and the experimental details is supplied in the lower half of the parameter selection toolbox. The status window provides continuous progress updates during GUI execution. The figure widow plots the final results from the analysis, and the formatting of the figure can be edited using the normal MATLAB figure editing tools. This GUI is used to produce data such as those in Figure 3.

Neurocomputational modeling

Neurocomputational modeling is a relative new approach to producing process models that explain known phenomena and provide predictions that motivate future research.

Approaches to modeling. Computational modeling has a long and rich history in the neurosciences. The traditional approach to computational modeling involves defining a domain of phenomena to model. These phenomena will be defined by a set of relationships between environmental inputs and behavioral outputs. The goal of computational models is to explain the intervening processes that produce the behavioral outputs with reference to the environmental inputs. Such models often rely on metaphors. For example, scalar timing theory, a predominant model of timing behavior that could be used to model the data in Figure 3, was developed around the metaphor of a stop watch (Gibbon & Church, 1984; Gibbon, Church, & Meck, 1984). This model proposes that a timing signal (e.g., the tone in Figure 1) activates a switch and this results in the transfer of pulses from a clock into an accumulator. The accumulator accrues pulses over time. When an outcome such as food occurs, the contents of the accumulator are stored in memory for future reference and the contents of the accumulator are reset to zero. Thus, the clock-accumulator component of the model functions in an identical fashion to a stop watch. Scalar timing theory has been successful in predicting a wide range of phenomena in the timing field, although the model is not without its criticisms (Wearden & Lejeune, 2007). One criticism that applies to many models developed through the metaphor route is a lack of neural plausibility (Bhattacharjee, 2006). For example, scalar timing theory

assumes an infinite capacity for memory storage in the nervous system as every time interval of importance experienced in the lifetime of the individual is stored as s separate sample in memory.

Neurocomputational modeling extends on the computational modeling approach by incorporating neurobiological processes into the modeling environment. Specifically, neurocomputational models aim to develop computational process models that are guided and constrained by the known properties of the relevant neural circuitry. This can include using information such as the neural pathways (and their directionality), the firing dynamics of cells, and the neurotransmitter dynamics within each pathway. This information is used to assist in determining the likely computational processes performed by each pathway within a larger circuit. Figure 6 displays the circuitry relevant to explaining the results in Figures 2 and 3. One sub-circuit is responsible for reward prediction learning and includes the ventral tegmental area (VTA), the nucleus accumbens (NA) and the basolateral amygdala (BLA). Neurocomputational models of this sub-circuit have been developed and refined and are probably the best example of neurocomputational applications within this domain (e.g., Schultz, 2006). Another sub-circuit is the timing circuit which involves the basal ganglia pathways (including the thalamus, TH, the sub-thalamic nucleus, STN, and the substantia nigra pars reticula, SNr/internal segment of the globus pallidus, GPi and the external segment of the globus pallidus) coupled with the substantia nigra pars compacta, SNc, and the dorsal striatum, DS. There have been attempts at producing neurocomputational models of timing (e.g., Matell & Meck, 2004), but these models are still under development. The integration of reward prediction and timing is accomplished by multiple cortical

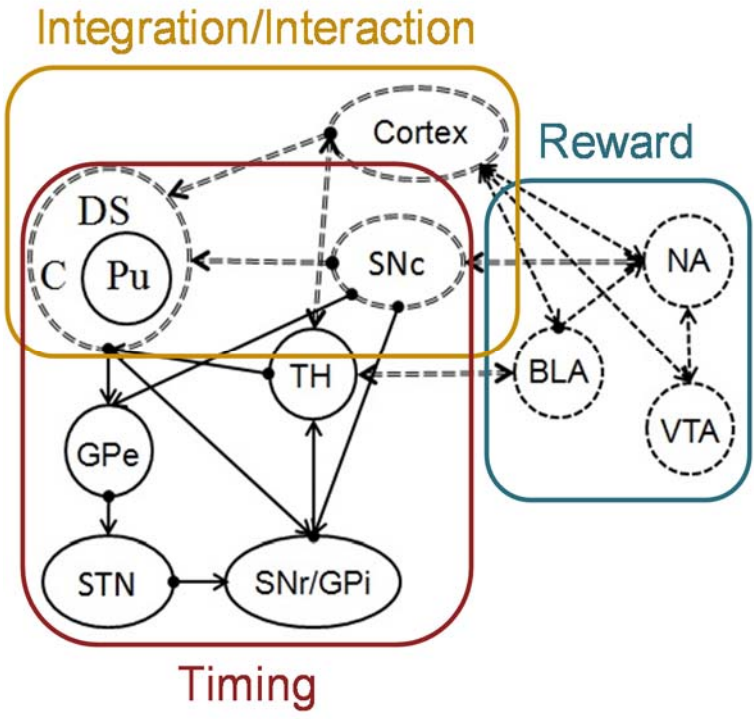

Figure 6. A neural circuitry diagram of the components of the timing and reward systems and the circuits responsible for their integration. These circuits play a key role in the classical conditioning task shown in Figure 1 and can form the basis of development of future neurocomputational models of learning. NA = nucleus accumbens, VTA = ventral tegmental area, BLA = basolateral amygdala, SNc = substantia nigra pars compacta, DS = dorsal striatum, C = caudate, Pu = Putamen, TH = thalamus, GPe = globus pallidus external segment, GPi = globus pallidus internal segment, SNr = substatia nigra pars reticula, STN = subthalamic nucleus. Adapted from Kirkpatrick (2013).

regions coupled with the SNc, DS and their connections with the NA. There are no current neurocomputational models that deal with the interaction of reward and timing processes, so this is a clear area for future development (see Galtress, Marshall, & Kirkpatrick, 2012; Kirkpatrick, 2013).

Techniques and tools for modeling. One excellent approach for modeling involves the development of model simulations in MATLAB for specific tasks and behaviors. Model simulations can be conducted using custom scripts and functions written in MATLAB. The model output can be produced in the form of time-event codes so that the model data can be analyzed in the same fashion as the data from experimental participants. Formal comparison of the model with the data can then be undertaken (see Church & Guilhardi, 2005). As with data mining, computational modeling applications present challenges for integrating students into the research program. This concern can be mitigated by developing robust tools for modeling using MATLAB GUIs where the model configurations can be selected using menus created in the GUIDE environment.

Summary and Conclusions

The growth of the collection of increasingly large and more complex data sets in the neurosciences is leading to the need for the development of new tools to promote capabilities for data mining. Technical languages such as MATLAB can serve as an excellent source for developing customized scripts and functions, and these can be made accessible to students involved in research through the use of GUIs. The future of neuroscientific research would be greatly benefited by increased availability of archived data for mining and computational modeling, increased sharing of tools for analysis, and the development of standards for approaches to mining neuroscientific data. An important companion to data mining is computational modeling, which provides a means of understanding complex patterns in data. Computational modeling is increasingly informed by neurobiology and this is leading to increased developments in neurocomputational modeling, which explicitly incorporate neurobiological evidence in the development of process models of behavior. Here, too, the use of technical computing languages coupled with GUIs can provide powerful tools for model development and implementation.

References

Balsam, P., Drew, M., & Yang, C. (2002). Timing at the start of associative learning. *Leaning and Motivation, 33*, 141-155.

Balsam, P., Sanchez-Castillo, H., Taylor, K., Van Volkinburg, H., & Ward, R. D. (2009). Timing and anticipation: conceptual and methodological approaches. *European Journal of Neuroscience, 30*, 1749-1755.

Bhattacharjee, Y. (2006). Neuroscience: A timely debate about the brain. *Science (Washington, D. C., 1883-), 311*, 596-598.

Bitterman, M. E. (1964). Classical Conditioning in the Goldfish as a Function of the CS-US Interval. *Journal of Comparative and Physiological Psychology, 58*, 359-366.

Black, A. H. (1963). The effects of CS-US interval on avoidance conditioning in the rat. *Canadian Journal of Psychology, 17*, 174-182.

Church, R. M., & Guilhardi, P. (2005). A Turing test of a timing theory. *Behavioural Processes, 69*(1), 45-58.

Church, R. M., Meck, W. H., & Gibbon, J. (1994). Application of scalar timing theory to individual trials. *Journal of Experimental Psychology: Animal Behavior Processes, 20*(2), 135-155.

Galtress, T., Marshall, A. T., & Kirkpatrick, K. (2012). Motivation and timing: clues for modeling the reward system. *Behavioural Processes, 90*, 142-153. doi: 10.1016/j.beproc.2012.02.014

Gibbon, J., Baldock, C., Locurto, C. M., Gold, L., & Terrace, H. S. (1977). Trial and intertrial durations in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes, 3*, 264-284.

Gibbon, J., & Church, R. M. (1984). Sources of variance in an information processing theory of timing. In H. L. Roitblat, T. G. Bever & H. S. Terrace (Eds.), *Animal cognition* (pp. 465-488). Hillsdale, NJ: Elrbaum.

Gibbon, J., & Church, R. M. (1990). Representation of time. *Cognition, 37*(1-2), 23-54.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. In J. Gibbon & L. Allan (Eds.), *Timing and time perception (Annals of the New York Academy of Sciences)* (Vol. 423, pp. 52-77). New York: New York Academy of Sciences.

Guilhardi, P., & Church, R. M. (2004). Measures of temporal discrimination in fixed-interval performance: A case study in archiving data. *Behavior Research Methods, Instruments & Computers, 36*(4), 661-669.

Jennings, D. J., Bonardi, C., & Kirkpatrick, K. (2007). Overshadowing and stimulus duration. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(4), 464-475.

Kirkpatrick, K. (2013). Interactions of timing and prediction error learning. *Behavioural Processes*.

Kirkpatrick, K., & Church, R. M. (2000). Independent effects of stimulus and cycle duration in conditioning: The role of timing processes. *Animal Learning & Behavior, 28*, 373-388.

Matell, M. S., & Meck, W. H. (2004). Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cognitive Brain Research, 21*(2), 139-170.

Pavlov, I. P. (1927). *Conditioned Reflexes* (G. V. Anrep, Trans.). New York: Dover.

Reid, A. K., Bacha, G., & Morán, C. (1993). The temporal organization of behavior on periodic food schedules. *Journal of the Experimental Analysis of Behavior, 59*(1), 1-27.

Salafia, W. R., Terry, W. S., & Daston, A. P. (1975). Conditioning of the rabbit (Oryctolagus cuniculus) nictitating membrane response as a function of trials per session, ISI, and ITI. *Bulletin of the Psychonomic Society, 6*, 505-508.

Schneiderman, N., & Gormezano, I. (1964). Conditioning of the nictitating membrane of the rabbit as a function of the CS-US interval. *Journal of Comparative and Physiological Psychology, 57*, 188-195.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annu Rev Psychol, 57*, 87-115.

Wearden, J. H., & Lejeune, H. (2007). Scalar properties in human timing: Conformity and violations. *The Quarterly Journal of Experimental Psychology, 61*(4), 569-587. doi: 10.1080/17470210701282576