## Title
A hierarchical Bayesian model of “memory for when”based on experience sampling data



**Journal**
Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

**Authors**
Dennis, Simon
Yim, Hyungwook
Sreekumar, Vishnu
et al.

**Publication Date**
2017

Peer reviewed

# A hierarchical Bayesian model of "memory for when" based on experience sampling data

**Simon Dennis(simon.dennis@gmail.com)**
**Hyungwook Yim (hyungwook.yim@newcastle.edu.au)**
School of Psychology, University of Newcastle, Australia

**Vishnu Sreekumar (vishnu.sreekumar85@gmail.com)**
Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke
National Institutes of Health, USA

**Nathan J. Evans (nathan.j.evans@uon.edu.au)**
**Paul Garrett (paul.garrett@uon.edu.au)**
School of Psychology, University of Newcastle, Australia

**Per Sederberg (sederberg.1@osu.edu)**
Department of Psychology, The Ohio State University, USA

**Abstract**

Participants wore a smartphone, which collected GPS, audio, accelerometry and image data, in a pouch around their necks for a period of two weeks. After a retention interval of one week, they were asked to judge the specific day on which each of a selection of images was taken. To account for people's judgements, we proposed a mixture model of four processes - uniform guessing, a signal detection process based on decaying memory strength, a week confusion process and a event confusion process in which the sensor streams were used to calculate the similarity of events. A model selection exercise testing all possible subsets of the processes favoured a model that included only the event confusion model. GPS similarities were found to be the most significant predictors, followed by audio and accelerometry similarities and then image similarities.

**Keywords:** memory, experience sampling, hierarchical Bayesian model

## Introduction

Friedman (1993, 2004) argued that people typically employ one of four strategies to identify when events occurred. On some occasions, people can directly retrieve declarative knowledge about the event. For instance, many people can recall that the attacks on the Twin Towers occurred on September 11th 2001. Friedman argues, however, that such declarative knowledge is quite rare and is reserved for events of global or personal significance. On other occasions, people have relative order information that they can use to make a judgement. If one were asked when George W. Bush initiated military action in Afghanistan, one may not know the date, but one can make an inference. The military action in Afghanistan occurred as a consequence of the September 11th attacks and therefore was most likely to have been in late 2001. There is a natural order in which these events occurred and provided someone has the order information and access to the time of the original event they can deduce the timing of the subsequent event. Again, Friedman argues that judgements based on relative order information are rare.

More common, according to Friedman, are judgements made using location-based strategies. Location-based processes rely on the retrieval of information associated with the cues that can be used to draw inferences about the timing of an event. For instance, suppose you are asked when you last saw your friend Mary. You might recall that you share a Psychology-101 class with Mary. Furthermore, you know that Psychology-101 occurs on Mondays and Wednesdays at 2pm. It is now Saturday, so you infer that it was Wednesday at 2pm when you last saw Mary.

Sometimes, however, the necessary knowledge to make an inference is not available. In these circumstances, Friedman argues one resorts to a distance-based strategy. Distance-based strategies rely on some quality of the memory that changes as a function of time. For instance, one might judge strong memories as having occurred more recently.

There is a substantial literature that has asked people to report on the time at which events occurred (see Friedman, 1993 and Thompson, Skowronski, Larsen & Betz, 1996, for reviews). Much of this literature has involved the memory for events that occurred outside of the laboratory, but which can be dated because they are part of the public record or have been recorded in personal diaries (Kemp, 1999). Generally, people are very poor at identifying when events occurred showing a bias to report events as being too recent when they occurred remotely in time - forward telescoping (Huttenlocher, Hedges, & Prohaska, 1988) or too remote when they occurred recently - backward telescoping (Hinrichs & Buschke, 1968).

Early distance-based theories proposed that the psychological representation of time was logarithmically compressed, much as other psychophysical dimensions are

(Ferguson & Martin, 1983). These theories are able to account for the decrease in accuracy that occurs with retention interval, but have been discounted because well-known remote events are often dated accurately. If it were the time axis itself that was compressed, dating accuracy should always be directly related to recency (Huttenlocher et al., 1988). Alternatively, the accessibility of the memory trace could be used to infer the time of occurrence. Some evidence suggests that better known events are dated more recently (Brown, Rips & Shevell, 1985). However, there is substantial subsequent evidence that this does not occur (Thompson et al. 1988). Again, better known events are reported more accurately. Furthermore, people are often capable of accurately reporting the day of the week an event occurred, while struggling to faithfully retrieve the month or year (Friedman & Wilkins, 1985). If people were using distance-based strategies, accuracy at smaller temporal scales ought to be worse than at larger temporal scales. The existing literature has tended to conclude that location-based strategies are far more commonly employed than distance-based strategies (Friedman, 1993; Thompson et. al. 1996).

There are two main models of time reporting that have been proposed (Huttenlocher, Hedges & Prohaska, 1988; Kemp, 1999). The first of these, by Huttenlocher, Hedges, and Prohaska (1988), proposes that events are associated with time information, which is unbiased but subject to error that increases with time. Bias is introduced because answers are constrained to lie within the reference period either implicitly or explicitly defined by the memory query thus generating forward and backward telescoping. In addition, the theory posits that memory units are organized hierarchically - (e.g. day, month, year) and that events may be associated with any of these levels. The model provides a good quantitative fit to data they collected on judgements of when movies that were part of a campus initiative were shown.

Kemp's (1999) theory is similar to Huttenlocher's in that the representation of time is not systematically distorted and that reconstruction of this time information is the basis of memory judgements. Rather than suggesting that temporal information is distorted with age, Kemp (1999) proposed that when time information is available it is accurate regardless of age. However, only a small number of memories have stored time information. Events of a similar kind are associated with each other and retrieval proceeds by retrieving similar events until one is found for which time information has been stored. An inference is then made on the basis of this information.

Both the Huttenlocher and Kemp theories are typically construed as location-based theories because they rely on the retrieval of time information (i.e., a location in time) and inference proceeds on the basis of that information. However, another possibility is that the time information to which they refer evolves in a continuous fashion on multiple time scales simultaneously. This kind of model is commonly employed to account for grouping effects in short term serial recall (e.g., Henson, 1998). Furthermore, it is possible that what is retrieved from memory is a combination of specific content on which conscious inferences can be drawn (location-based) and this more graded hierarchical form of context (distance-based).

Although it has long been argued that memory research that is focused solely on laboratory work is futile (Neisser, 1976), the difficulty has been how to proceed when the experience of the participant before they enter the laboratory cannot be rigorously captured. Today, however, technology provides us with entirely new options. Easy to carry and able to monitor multiple sensor streams, smartphones can provide a convenient and ubiquitous window into the events of the life of an individual. We had participants wear a phone around their necks for two weeks and collected image, audio, GPS and accelerometry data. We then developed a hierarchical Bayesian model to capture distance and location based processes.

## Method

### Participants

A total of 18 adult participants were recruited from flyers posted at the University of Newcastle and received $100 compensation.

### Procedure

In prior work, we built a system that consists of an Android app, server infrastructure and user interfaces. The app acquires image, time, audio (obfuscated), GPS, accelerometer and orientation information at approximately five minute time intervals. The app runs in the background as a service and users carried the phone in a pouch attached to a neck strap from morning till evening (see Figure 1). Participants could turn off the app anytime they needed privacy. When the phone detects WiFi and is charged, it sends the stored data automatically to a remote server. This usually happened once per day at the end of the day when users charge the phone overnight.

Participants were instructed to wear the phone for two weeks. They returned to the laboratory on the Friday of the 3rd week and were presented with images one at a time and were required to determine on which of the week days each image was taken (participants were informed that the images only came from the week days). Each participant's test was based on images drawn from their own lifelogs. We selected images that came from distinct episodes as much as possible, and also avoided using images that were blurred due to excessive motion. The number of stimuli varied between participants since the available data depended on individual lifestyles. A presented image remained on the screen while they made the day judgment and they could use as much time as they needed to respond.

Figure 1. Android phone worn by a participant during the experience sampling phase of the study.

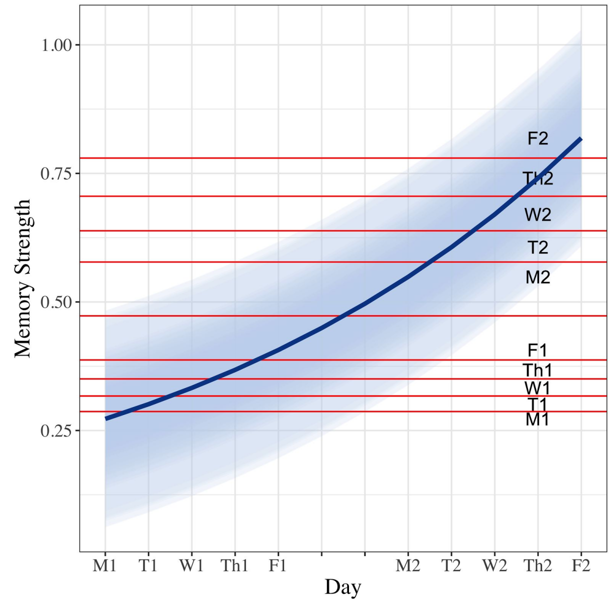


Figure 2. The distance based model.

## Modeling

To account for people's judgements, we proposed a mixture model of four processes - (1) random (uniform) guessing, (2) a signal detection process based on decaying memory strength (distance), (3) a week confusion process (location) and (4) a event confusion process (location) in which the sensor streams were used to calculate the similarity of events. We start by describing the distance and location (sensor) models and then outline the mixture model incorporating all four processes.

### Distance model

Figure 2 depicts the distance based model we employed. Mean memory strength ($\mu$) elicited at retrieval was assumed to decay exponentially with scale, $\alpha$, asymptote, $\beta$, and rate, $\lambda$. Variability around this mean was assumed to be Gaussian with standard deviation, $\sigma$. The probability of a response occurring given that the presented stimulus was taken on a given day, is given by the probability density that falls between criteria that separate it from the neighbouring days.

The nine criteria that determined the response probabilities for each day were fixed to the mid-point of $\mu$ values of each neighbouring day (to alleviate sampling issues that resulting from attempting to estimate these as free parameters). We used Bayesian hierarchical modeling to fit the model, with each individual's parameters being constrained by a group level distribution. All parameters were sampled on a double infinite scale, meaning that we sampled the inverse Probit of $\alpha$ and $\beta$, and the natural logarithm of $\lambda$ and $\sigma$, and that all group level distributions were, therefore, normal.

### Location (sensor) model

The location (sensor) model assumed that events were stored in memory and that the likelihood of confusing the representation of the correct event with the stored representation of another event is determined by the similarity of those events. Each day was divided into hour periods and image, GPS, audio and accelerometry representations of those events were calculated. For a given sensor stream the distance of an image's event to a given day for a given stream was taken to be the minimum Jenson Shannon distance from the event to the events of that day (see Figure 3).

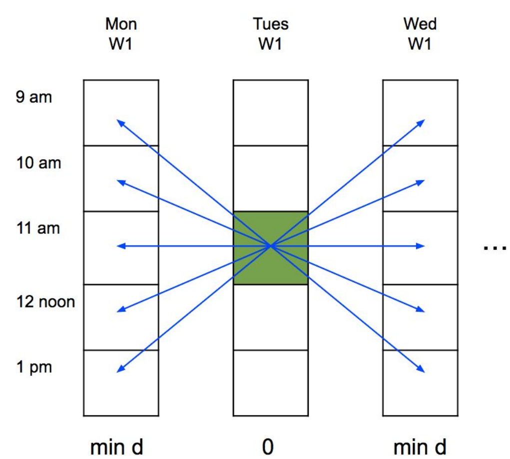


Figure 3. The sensor model.

These distance scores for each of the streams entered into a conditional logit model to determine the probability that the participant would respond with a given day. Missing data were assumed to have a prior of a truncated normal distribution.

Like the distance model, the parameters were estimated in a hierarchical fashion, with the natural logarithm of the weights being estimated, making the group-level distributions normal.

## Mixture model

To estimate the probability of a participant's response we assumed a mixture model of the distance and location (sensor) models described previously as well as a random (uniform) guessing model and a location (week) model that assumed that participants correctly inferred the day of the week on which the event occurred, but had a certain probability of incorrectly determining the week (the kind of model that one might assume if people are relying on their schedules to make judgements).

A model selection exercise testing all possible subsets of the processes was conducted using the common model selection metric WAIC, which attempts to weigh both the goodness of fit to the data and the complexity of the model, in order to approximate the leave-one-out cross validation metric. The preferred model was the location (sensor) model although the location (sensor) + random model also performed well (see Table 1).

Table 1: Models tested and corresponding WAICs

| Model | WAIC |
|---|---|
| Location (sensor) | -1544 |
| Location (sensor) + Random | -1546 |
| Location (sensor) + Distance | -1557 |
| Location (sensor) + Distance + Random | -1565 |
| Distance + Random | -1649 |
| Location (week) + Random | -1958 |
| Distance + Location (week) | -2161 |
| Random | -2544 |
| Distance | -2688 |
| Distance + Location (week) + Random | -2812 |
| Location (week) | $-\infty$ |

To understand the performance of the models, it is useful to compare the posterior confusion matrices they produce to the data. Figure 4 shows the confusion matrix of responses accumulated over participants. The x-axis show the days on which the events actually occurred, and the y-axis shows the participants' responses. The diagonal represents correct responses, while responses off the diagonal are errors. The matrix is dominated by correct responses, with cells close to the major diagonal (representing adjacent days) showing significant mass particularly in week one.

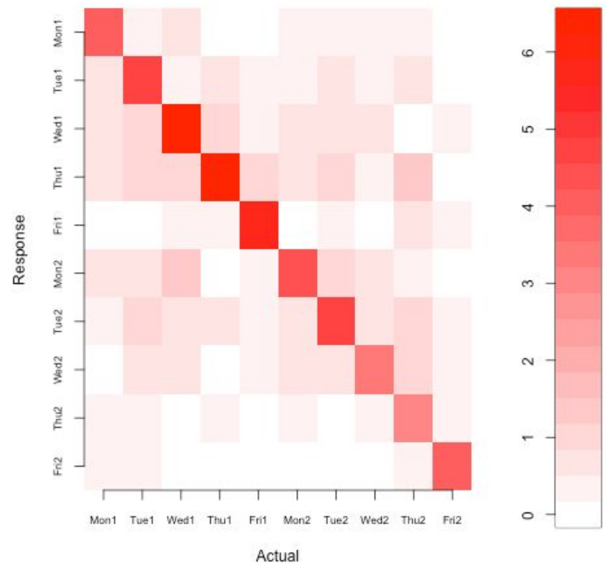


Figure 4. Data confusion matrix

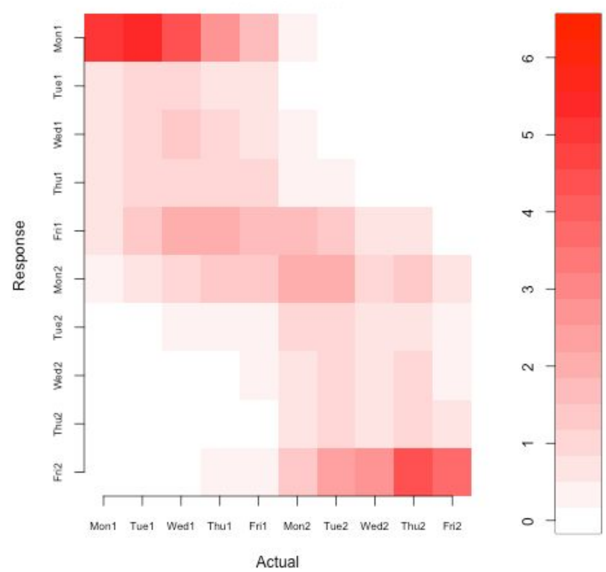


Figure 5. Distance only confusion matrix

The distance only model is able to explain the structure off the diagonal by estimating a large standard deviation for the strength distributions (see Figure 5). However, a large standard deviation prevents the model from capturing the proportion of correct responses on the diagonal.

The distance model performs much better when it is mixed with the uniform distribution (see Figure 6). The structure off the diagonal is captured by the uniform component, while the structure on and adjacent to the diagonal is captured by the distance model. The observation that counts adjacent to the diagonal are larger in week one is accommodated naturally by the model because in the first week the gradient of the strength is small, which makes it more difficult to distinguish between adjacent days.

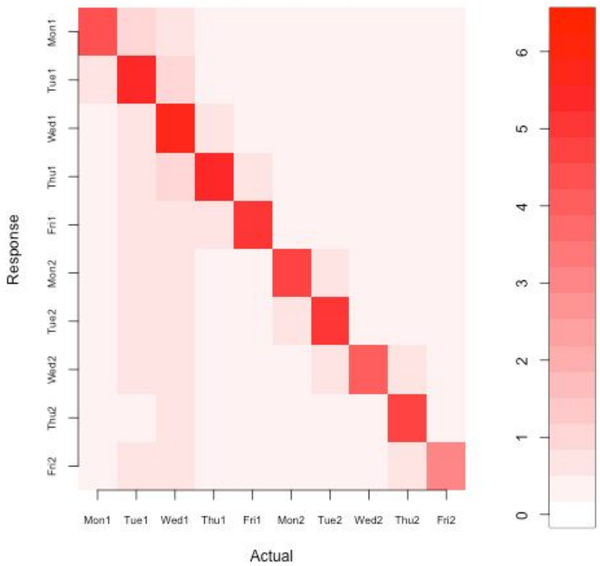


Figure 6. Distance + Random confusion matrix

The week only model does poorly. The model assigns no probability to cells that are neither on the diagonal nor exactly a week out (the off diagonals five above and below the main diagonal). As there are observations in those cells, the WAIC is negative infinity. When mixed with the random model, the model does better, but still has a tendency to predict more week out responses than appears in the data (see Figure 7).

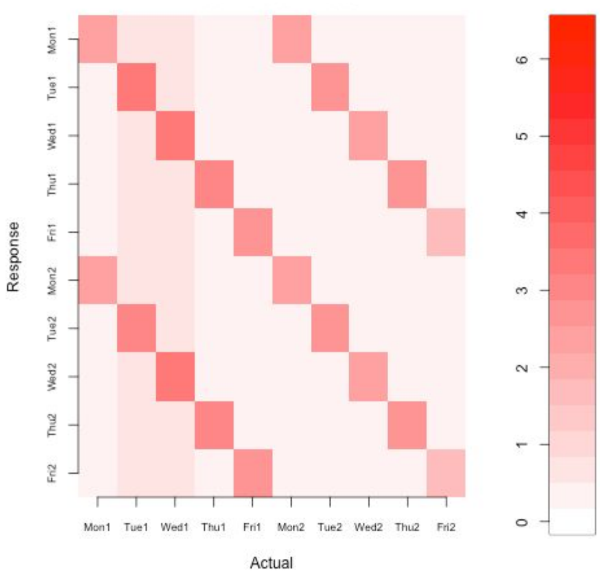


Figure 7.: Week + Random confusion matrix

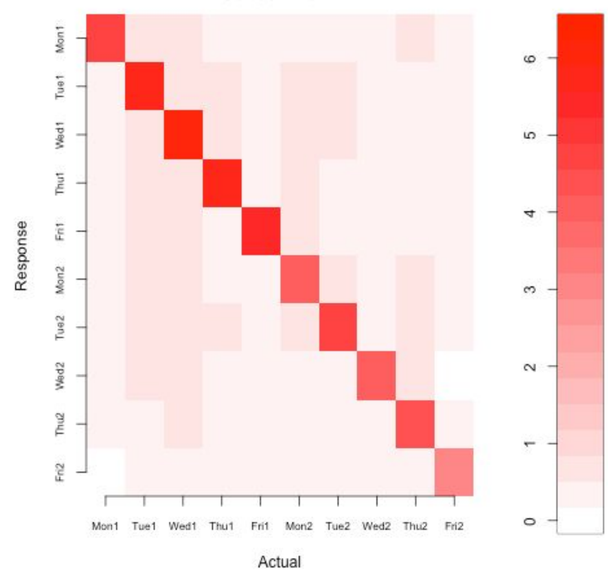


Figure 8. Sensor confusion matrix

The best model is the sensor model (see Figure 8). Unlike the distance and week models the sensor model did not require the random component in order to provide a good fit to the data. In fact, adding the random component decreases the WAIC slightly as the model is penalized for additional complexity (i.e. the mixture probability). While the distance and week models are informed only by the day on which the event occurred, the sensor model constructs a representation of the event that captures where the participant was (GPS), what the participant was hearing (audio), what the participant was doing (accelerometry) and what the participant was seeing (images) and compares it with representations of all other events. The importance of these features can be inferred from the weights associated with each of the data streams.

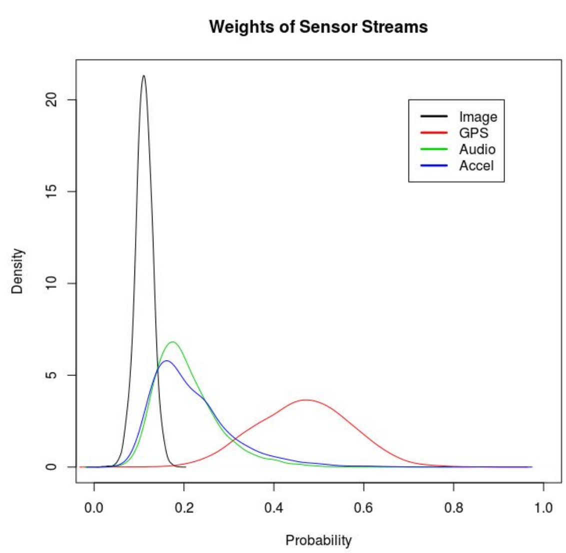


Figure 9. The posterior distributions of the sensor weights.

Figure 9 shows the posterior distributions of these weights. The GPS stream has the strongest weights followed by the audio stream and the accelerometry stream, which are approximately equal. The image stream has the lowest weights. That the image stream should have the lowest weights is counter intuitive. The participants are presented with the image as a retrieval cue, and so one might have expected the visual information to be salient.

There are multiple possible interpretations of this result. It may be that the image representation that we chose (GIST; Oliva & Torralba, 2001) does not carry the information that participants rely upon when making memory judgements. Another possibility is that it is the static nature of the images or the fact that they are not synchronized with the direction of gaze that compromised this stream. While head mounted video technologies exist they are currently difficult to deploy for the duration of recording required for the time scales we explore here. Furthermore, they introduce additional ethical hurdles that need to be considered. A third possibility is that the result is not artifactual, but is a reflection of the information employed by the memory system. While the visual domain seems salient perhaps it is other aspects of experience that drive the retrieval and inferential systems that people employ to make location based judgements.

## Conclusions

When people are asked to determine when an event occurred, Friedman (2004) argues that people use a combination of distance based and location based processes, with location based processes being the most common. The current work supports this assertion.

Furthermore, we have demonstrated that it is possible to predict the responses people will make to images taken from their personal experience in the world outside the laboratory on a stimulus by stimulus basis. We believe this work establishes a new benchmark for what models of episodic memory should achieve and provides the promise of a more quantitatively rigorous, ecologically valid and translationally relevant memory science.

## Acknowledgments

We would like to acknowledge the funding from the Australian Research Council (DP150100272) and Air Force Office of Scientific Research (FA9550-09-1-0614).

## References

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Ferguson, R. P., & Martin, P. (1983). Long-term temporal estimation in humans. *Perception and Psychophysics, 33*, 585-592.

Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin, 113*(1), 44-66.

Friedman, W. J. (2004). Time in autobiographical memory. *Social Cognition, 22*(5), 591-605.

Friedman, W. J., & Wilkins, A. J. (1985). Scale effects in memory for the time of past events. *Memory and Cognition, 13*, 168-175.

Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology, 36,* 73-137.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory, 6*, 287-317.

Huttenlocher J., Hedges L., & Prohaska V. (1988). Hierarchical organization of ordered domains: Estimating the dates of events. *Psychological Review, 95,* 471–484.

Huttenlocher J., Hedges L., & Prohaska V. (1992). Memory for day of the week: A 5 + 2 day cycle. *Journal of Experimental Psychology: General, 121*(3), 313-325.

Kemp, S. (1999). An associative theory of estimating past dates and past prices. *Psychonomics Bulletin and Review, 6*, 41-56.

Matlock, T. (2001). *How real is fictive motion?* (Unpublished doctoral dissertation), Psychology Department, University of California, Santa Cruz.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* New York: Freeman.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145-175.

Shrager, J., & Langley, P. (Eds.) (1990). *Computational models of scientific discovery and theory formation.* San Mateo, CA: Morgan Kaufmann.

Sreekumar, V., Dennis, S., Doxas, I., Zhuang, Y., & Belkin, M. (2014). The geometry and dynamics of lifelogs: discovering the organizational principles of human experience. *PloS one, 9*(5), e97166.

Thompson, C. P., Skowronski, J. J., & Lee, D. J. (1988). Reconstructing the date of a personal event. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues: Vol. 1. Memory in everyday life.* Chichester: Wiley.

Thompson, C. P., Skowronski, J. J., Larsen, S. F., & Betz, A. L. (1996). *Autobiographical memory: Remembering what and remembering when.* Mahwah, NJ: Erlbaum.