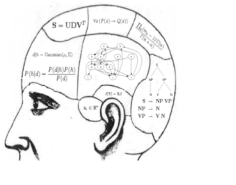

$S = UDV^T$
$\forall x (P(x) \to Q(x))$
$d|h \sim \text{Gaussian}(\mu, \Sigma)$
$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$
$E[Y] = X\beta$
$x_i \in R^n$
S → NP VP
NP → N
VP → V N

Institute for Pure & Applied Mathematics

IPAM

University of California, Los Angeles

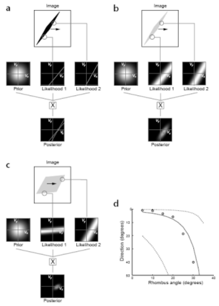
NSF

UCLA

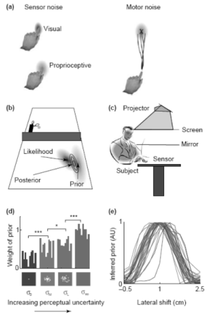
Rational analysis of human memory and prediction

Josh Tenenbaum
MIT

Bayesian inference in perception and motor control



(Weiss, Simoncelli & Adelson 2002)



(Kording & Wolpert 2004)

Today

- To what extent can the “Bayes meets Marr” approach be applied to cognition?
  - Can we measure “true” priors for cognition based on environmental statistics, and assess how well tuned cognition is for these priors?
  - Are there “universal” or “general-purpose” priors that can be used to characterize performance in cognitive tasks? How flexible are cognitive priors?
  - Can we see optimal statistical inference in the neural mechanisms of cognition?

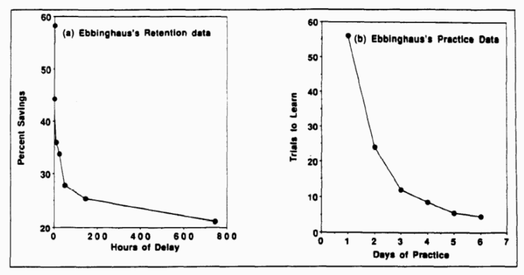
Today

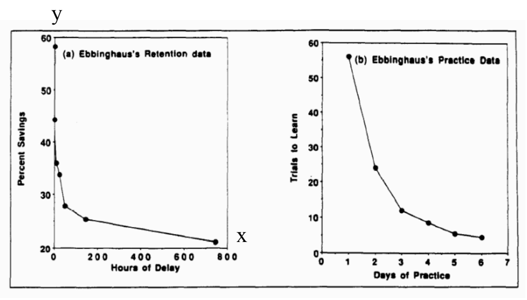
- Two case studies of memory
  - Retrieval (Anderson, 1990)
  - Prediction (Griffiths & Tenenbaum, 2006)

Anderson’s rational analysis of memory retrieval

- Starting point: some items are remembered better than others.
- What determines which items will be remembered better, and why?
  - How do probability and speed of recall depend on amount of study?
  - . . . on delay since the information was studied?
  - . . . on the particular pattern of study, e.g. cramming or steady practice?

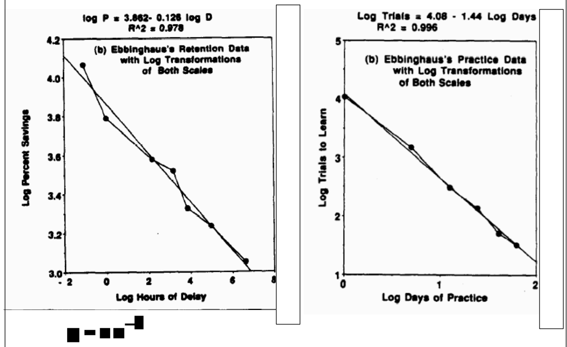
Ebbinghaus’s data (1885)





[redacted]

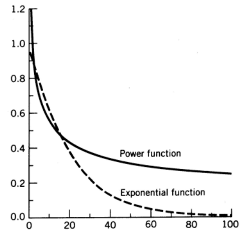
[redacted]



[redacted]

[redacted]

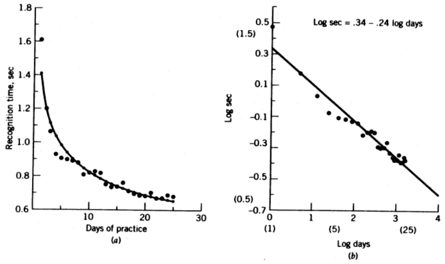
- Many decay functions in nature are exponential:
  - radioactive decay
  - heat diffusion
  - sweeping up dust



[redacted]

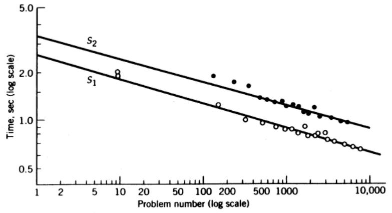
Memorizing sentences

FIGURE 6.3. (*a*) Time to recognize sentences as a function of number of days of practice; (*b*) log–log transformation of data in (*a*) to reveal power function. (From Pirolli & Anderson, 1985).

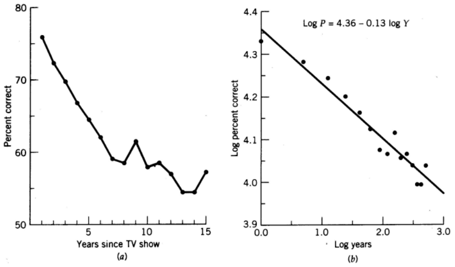


Adding numbers

FIGURE 6.4. The improvement in adding two numbers as a function of practice. Data are plotted seperately for two subjects. (Plot by Crossman, 1959, using data from Blackburn, 1936). *Source:* Figure 6.4 from E. R. F. W. Crossman. *Ergonomics* Volume 2. A theory of the acquisition of speed-skill. Copyright © 1976 by Taylor & Francis Ltd. Reprinted by permission.



Memory for TV characters



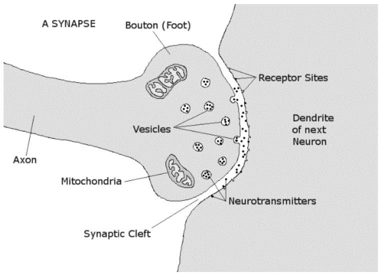
Why power laws?

- Mechanistic explanations in terms of symbolic cognitive architectures:
  e.g., "chunking" dynamics.

- But this mechanism was designed to produce power laws of practice -- it does not provide an independent explanation.

Marr's three levels

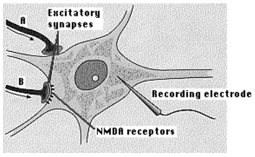
- Level 1: Computational theory
  - What is the goal of the computation, and what is the logic by which it is carried out?
- Level 2: Representation and algorithm
  - How is information represented and processed to achieve the computational goal?
- Level 3: Hardware implementation
  - How is the computation realized in physical or biological hardware?



Excitatory synapse: pre-synaptic cell firing makes post-synaptic cell more likely to fire.

Synapses vary in strength or "weight".

Long-term potentiation (LTP)



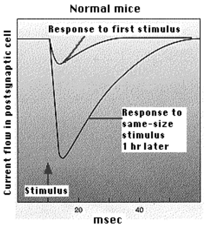
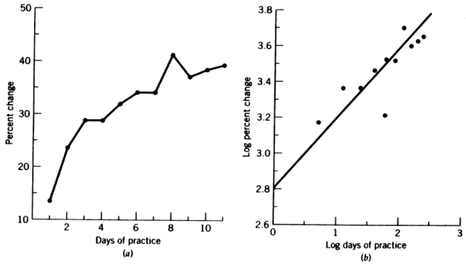
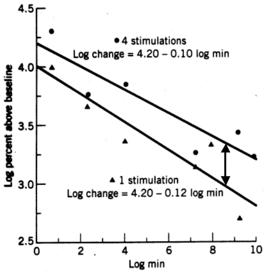


FIGURE 6.9. Growth in LTP as a function of number of days of practice: (*a*) in normal scale; (*b*) in log–log scale. (From Barnes, 1979).

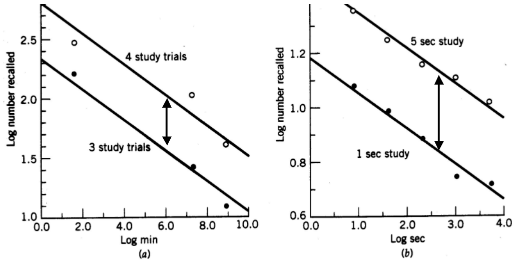


Decay of LTP



Independent effects of delay and learning time.

Independent effects of delay and learning time



Evaluating the LTP account of power laws in memory

- Provides a physical mechanism.
- Makes several nontrivial predictions confirmed in behavioral data.
- But why does LTP work this way?

A computational analysis

- Goal of memory retrieval:
  - For each item in memory, estimate its *need probability*, the probability that it will be useful in the present context.
  - Retrieve all items for which the expected utility exceeds the cost of retrieval.
- The critical question becomes, how does the mind estimate need probability?
- Failure to retrieve a memory is about prioritization, not loss of information.

A computational analysis

- Factors determining the probability that an item will be useful in the present context:
  - Match to contextual cues
  - History of prior use:
    - Time since last use
    - Number of times used previously
- Model of library book access (Burrell+)

A computational analysis

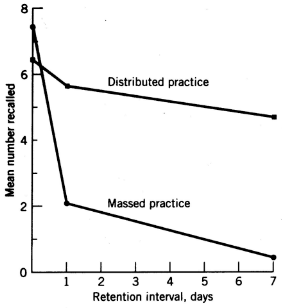
- Factors determining the probability that an item will be useful in the present context:
  - Match to contextual cues
  - History of prior use:
    - Time since last use
    - Number of times used previously

Predicts power laws, spacing effects, ….

- Model of library book access (Burrell+)
- Analogous factors in Google:
  - Text match between query and web page.
  - PageRank(tm) of web page.

Spacing effect

(a/k/a Cramming effect)



2 sessions each over 4 days

8 sessions on 1 day

c.f. motor adaptation

Model of library book access (Burrell+)

Observed retrieval history: [redacted]
(*n* retrievals between $t = 0$ and $t = T$).

Probability of need at $t = T$: [redacted]
where [redacted]

A latent variable model based on three assumptions:

1. There is a distribution of popularity over items, where popularity controls the rate at which an item is accessed.
2. There is an aging process for items and their rate of use decays over time. The rate of decay varies across items.
3. Items undergo random revivals of interest in which their rate of use returns to their original level of popularity.

Model of library book access (Burrell+)

Observed retrieval history: [redacted]
(*n* retrievals between $t = 0$ and $t = T$).

Probability of need at $t = T$: [redacted]
where [redacted]

Popularity [redacted]
Decay rate [redacted]
Revival history [redacted]
Retrieval history [redacted]

Latent variables

where [redacted] is the most recent revival before *t*.

Statistics of information usage in the natural environment

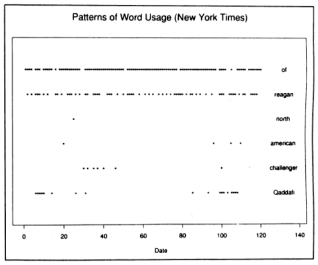
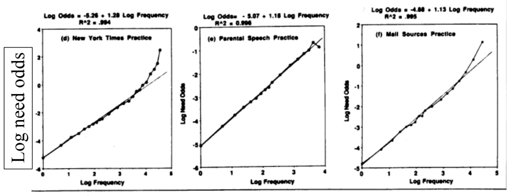


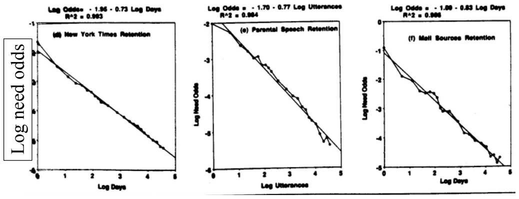
Fig. 5. Patterns of usage of various words in the *New York Times* data base over a 100-day period.

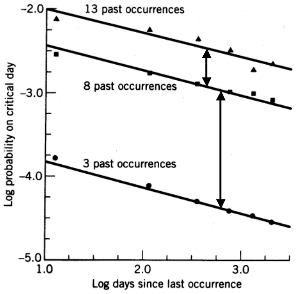
New York Times (practice) Parental speech (practice) Mail sources (practice)



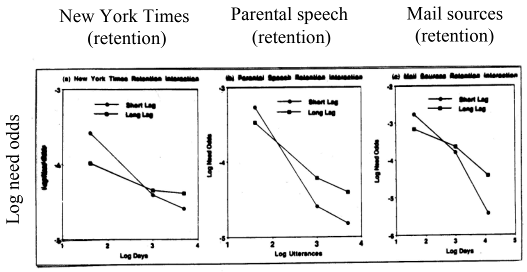
. 6. (a) Probability of a word occurring in a headline of the *New York Times* on Day 101 as a function of number of times it occurred in the previous 100 days; (b) probability of a word occurring in the 101st :rance from a parent as a function of the number of times it occurred in the previous 100 days; (c) bability of receiving a message on the 101st day from a source as a function of the number of times ssages were received from that source in the previous 100 days. Panels (d–f) provide transformation of :) plotting log needs against log frequency.

New York Times (retention) Parental speech (retention) Mail sources (retention)





Independent effects of delay and learning time in the New York Times.



Spacing (a/k/a cramming) effects.

Summary: Bayes meets Marr in memory retrieval

- Level 1: Computational theory
  - What is the goal of the computation, and what is the logic by which it is carried out?
- Level 2: Representation and algorithm
  - How is information represented and processed to achieve the computational goal?
- Level 3: Hardware implementation
  - How is the computation realized in physical or biological hardware?

Today

- Two case studies of memory
  - Retrieval (Anderson, 1990)
  - Prediction (Griffiths & Tenenbaum, 2006)

Everyday prediction problems

(Griffiths & Tenenbaum, 2006)

- You read about a movie that has made $60 million to date. How much money will it make in total?
- You see that something has been baking in the oven for 34 minutes. How long until it's ready?
- You meet someone who is 78 years old. How long will they live?
- Your friend quotes to you from line 17 of his favorite poem. How long is the poem?
- You meet a US congressman who has served for 11 years. How long will he serve in total?
- You encounter a phenomenon or event with an unknown extent or duration, t_{total}, at a random time or value of $t < t_{total}$. What is the total extent or duration t_{total}?

Bayesian analysis

$$P(t_{total}|t) \propto P(t|t_{total})\ P(t_{total})$$

$$\propto 1/t_{total} \quad P(t_{total})$$

Assume random sample (for $0 < t < t_{total}$ else = 0)

Form of $P(t_{total})$?
e.g., uninformative (Jeffreys) prior $\propto 1/t_{total}$

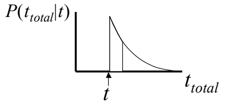
Bayesian analysis

$$P(t_{total}|t) \propto 1/t_{total} \quad 1/t_{total}$$

posterior probability

Random sampling

"Uninformative" prior



Best guess for t_{total}:
t^* such that $P(t_{total} > t^*|t) = 0.5$

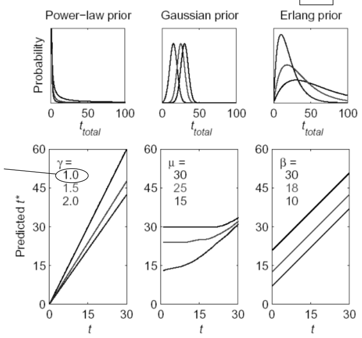
Yields Gott's Rule:
Guess $t^* = 2t$

Evaluating Gott's Rule

- You read about a movie that has made \$78 million to date. How much money will it make in total?
  - "\$156 million" seems reasonable.
- You meet someone who is 35 years old. How long will they live?
  - "70 years" seems reasonable.
- Not so simple:
  - You meet someone who is 78 years old. How long will they live?
  - You meet someone who is 6 years old. How long will they live?

The importance of priors

Different kinds of priors $P(t_{total})$ are appropriate in different domains.

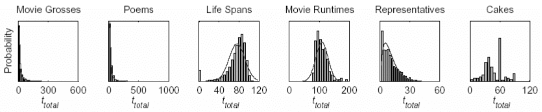


Gott's rule $P(t_{total}) \propto t_{total}^{-1}$

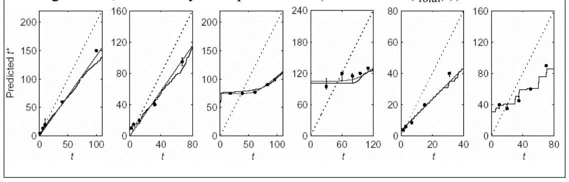
Evaluating human predictions

- Different domains with different priors:
  - A movie has made \$60 million
  - Your friend quotes from line 17 of a poem
  - You meet a 78 year old man
  - A move has been running for 55 minutes
  - A U.S. congressman has served for 11 years
  - A cake has been in the oven for 34 minutes
- Use 5 values of t for each.
- People predict t_{total} .

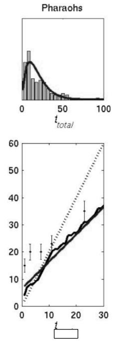
Priors $P(t_{total})$ based on empirically measured durations or magnitudes for many real-world events in each class:



Median human judgments of the total duration or magnitude t_{total} of events in each class, given that they are first observed at a duration or magnitude t, versus Bayesian predictions (median of $P(t_{total}|t)$).

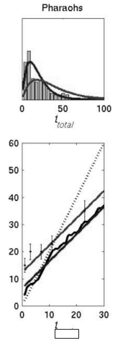


You learn that in ancient Egypt, there was a great flood in the 11th year of a pharaoh's reign. How long did he reign?



You learn that in ancient Egypt, there was a great flood in the 11th year of a pharaoh's reign. How long did he reign?

How long did the typical pharaoh reign in ancient egypt?



Summary: prediction

- Predictions about the extent or magnitude of everyday events follow Bayesian principles.
- Contrast with Bayesian inference in perception, motor control, memory: no "universal priors" here.
- Predictions best explained by priors that are appropriately calibrated for different domains.
  - Form of the prior (e.g., power-law or exponential)
  - Specific distribution given that form (parameters)
  - Non-parametric distribution when necessary.
- In the absence of concrete experience, priors may be generated by qualitative background knowledge.

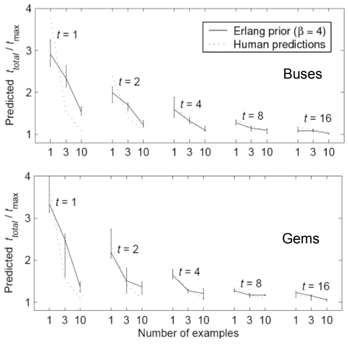
Open questions

- How flexible are memory retrieval and prediction? How quickly and easily can people adapt to new kinds of environmental statistics?
- Can we scale the approach down to analyses of individual subjects, as in perception and motor control?
- Can we scale up this approach to more complex kinds of knowledge?
- Can we find deeper mappings to neural mechanisms?

Predictions in novel environments

Subjects were exposed to a class of novel events with durations ranging from 0.5 to 32 seconds (geom. mean = 4 sec).

Subjects were then asked to predict durations of novel events given 1, 3, or 10 samples of that event.

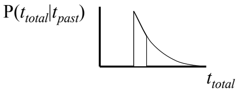


Bayesian prediction

$P(t_{total}|t_{past}) \propto$ $1/t_{total}$ $P(t_{past})$

posterior probability — Random sampling — Domain-dependent prior

What is the best guess for t_{total}?
Compute t such that $P(t_{total} > t|t_{past}) = 0.5$:

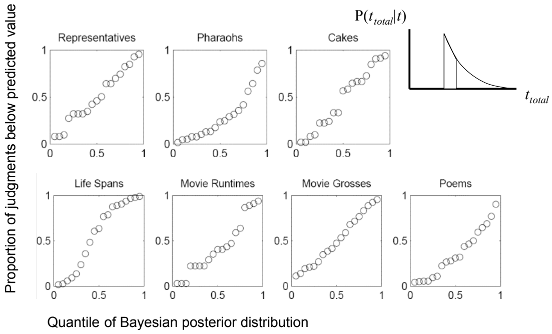


We compared the *median* of the Bayesian posterior with the *median* of subjects' judgments… but what about the distribution of subjects' judgments?

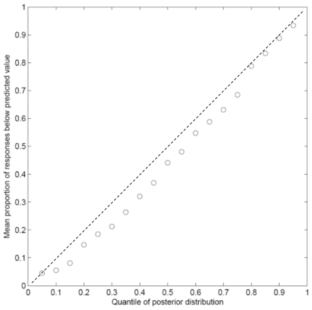
Sources of individual differences

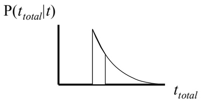
- Individuals' judgments could by noisy.
- Individuals' judgments could be optimal, but with different priors.
  - e.g., each individual has seen only a sparse sample of the relevant population of events.
- Individuals' inferences about the posterior could be optimal, but their judgments could be based on probability (or utility) matching rather than maximizing.

Individual differences in prediction



Individual differences in prediction





Average over all prediction tasks:

- movie run times
- movie grosses
- poem lengths
- life spans
- terms in congress
- cake baking times

Why probability matching?

- Optimal behavior under some (evolutionarily natural) circumstances.
  - Optimal betting theory, portfolio theory
  - Optimal foraging theory
  - Competitive games
  - Dynamic tasks (changing probabilities or utilities)
- Side-effect of algorithms for approximating complex Bayesian computations.
  - Markov chain Monte Carlo (MCMC): instead of integrating over complex hypothesis spaces, construct a sample of high-probability hypotheses.
  - Judgments from individual (independent) samples can on average be almost as good as using the full posterior distribution.