

Tutorial on Model Comparison Methods **(How to Evaluate Model Performance)**

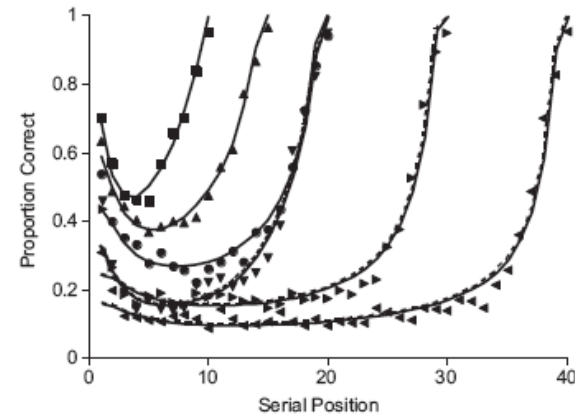
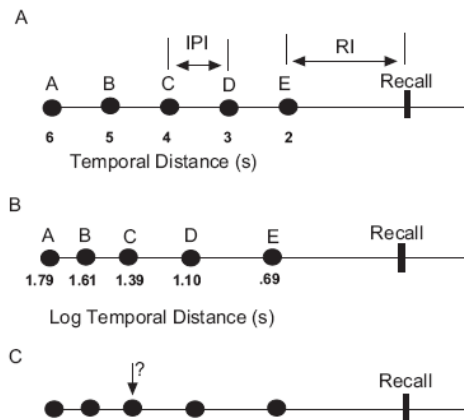
Jay Myung & Mark Pitt
Department of Psychology
Ohio State University

Annual Conference of Cognitive Science Society (July 20 2011: Boston, MA)

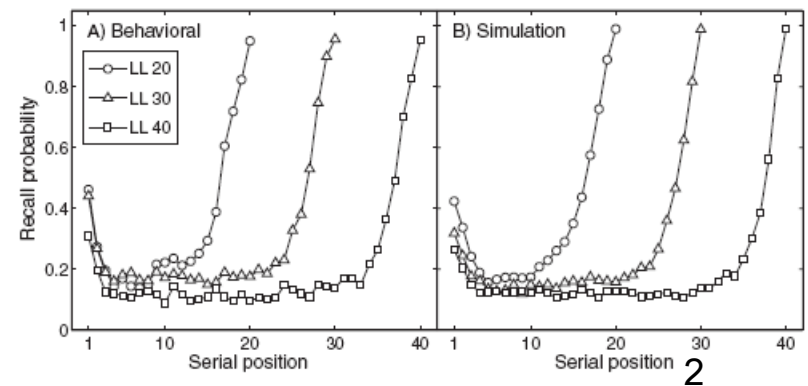
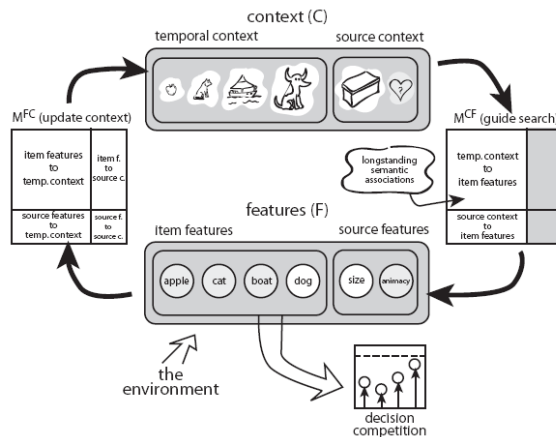
Model Comparison in Cognitive Modeling

How should one decide between competing explanations (models) of data?

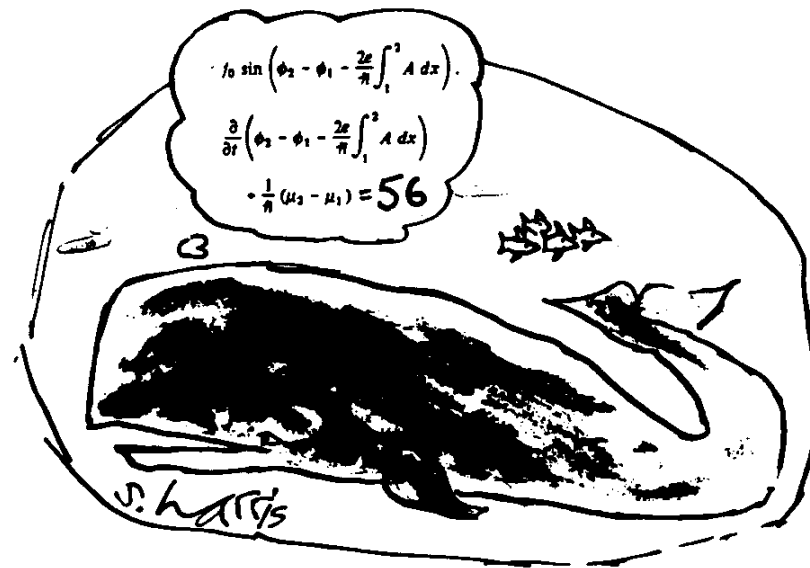
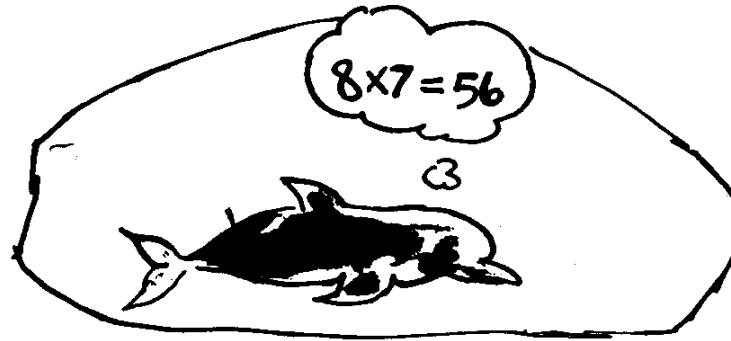
SIMPLE model of memory (Brown et al, 2007, Psy. Rev.)



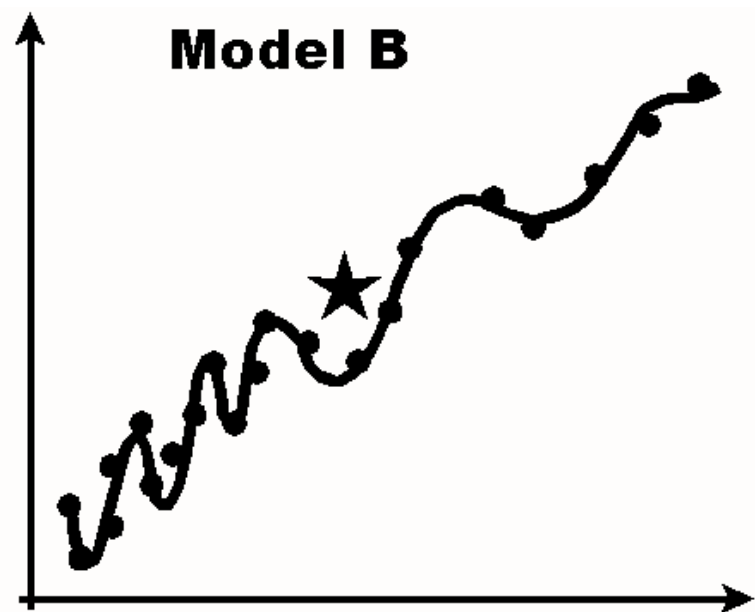
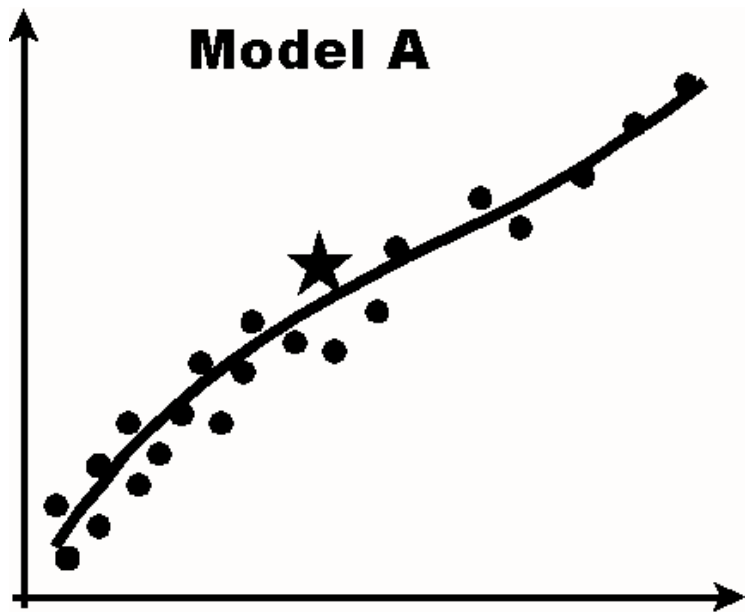
CMR model of memory (Polyn et al, 2009, Psy. Rev.)



Two Whale's Views of Model Comparison



Which one of the two below should we choose?



What we hope to achieve today

- This tutorial is a **first introduction** to model comparison for cognitive scientists
- Our aim is to provide a good **conceptual overview** of the topic and make you aware of some of the fundamental issues and methods in model comparison
- **Not an in-depth, hands-on tutorial** on how to apply model comparison methods to extant models using computing or statistical software tools
- Assume no more than **a year-long course in graduate level statistics**

Outline

- 1. Introduction**
- 2. Evaluating Mathematical Models**
 - 2a. Model selection/comparison methods**
 - 2b. Illustrative examples**
- 3. Evaluating Other Types of Models**
- 4. A New Tool for Model Comparison**
- 5. Final Remarks**

1. Introduction

- Preliminaries
- Formal Modeling
- Model Fitting

Preliminaries

- Models are **quantitative stand-ins** for theories
- Models are **tools** with which to study behavior
 - Increase the precision of prediction
 - Generate novel predictions
 - Provide insight into complex behavior
- Model comparison is a **statistical inference problem**.
Quantitative methods are developed to aid in deciding between models

Preliminaries

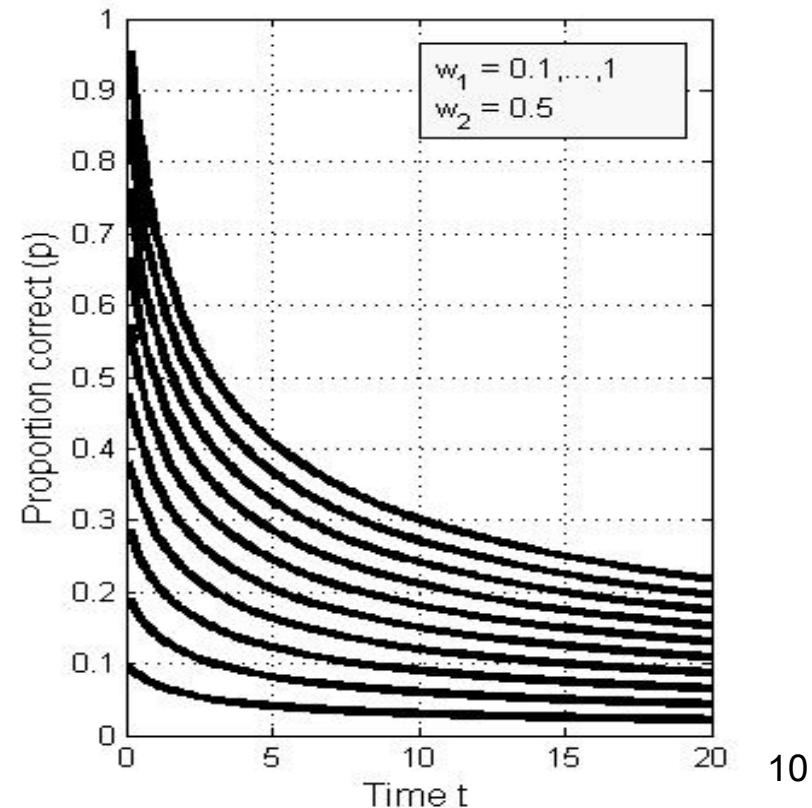
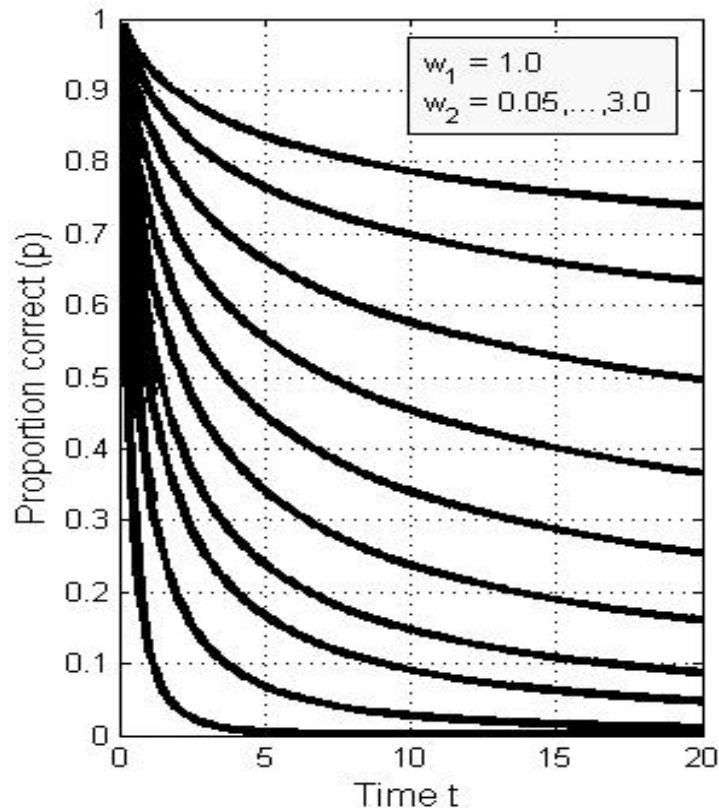
- Diversity of **types of models** in cognitive science makes model comparison challenging
- **Variety of statistical methods** are required
- Discipline would benefit from sharing models and data sets – **Cognitive Modeling Repository (CMR): Thursday night poster (www.osu.edu/cmr)**

Mathematical Modeling

A mathematical model specifies the range of data patterns it can describe by varying the values of its **parameter** (w), for example,

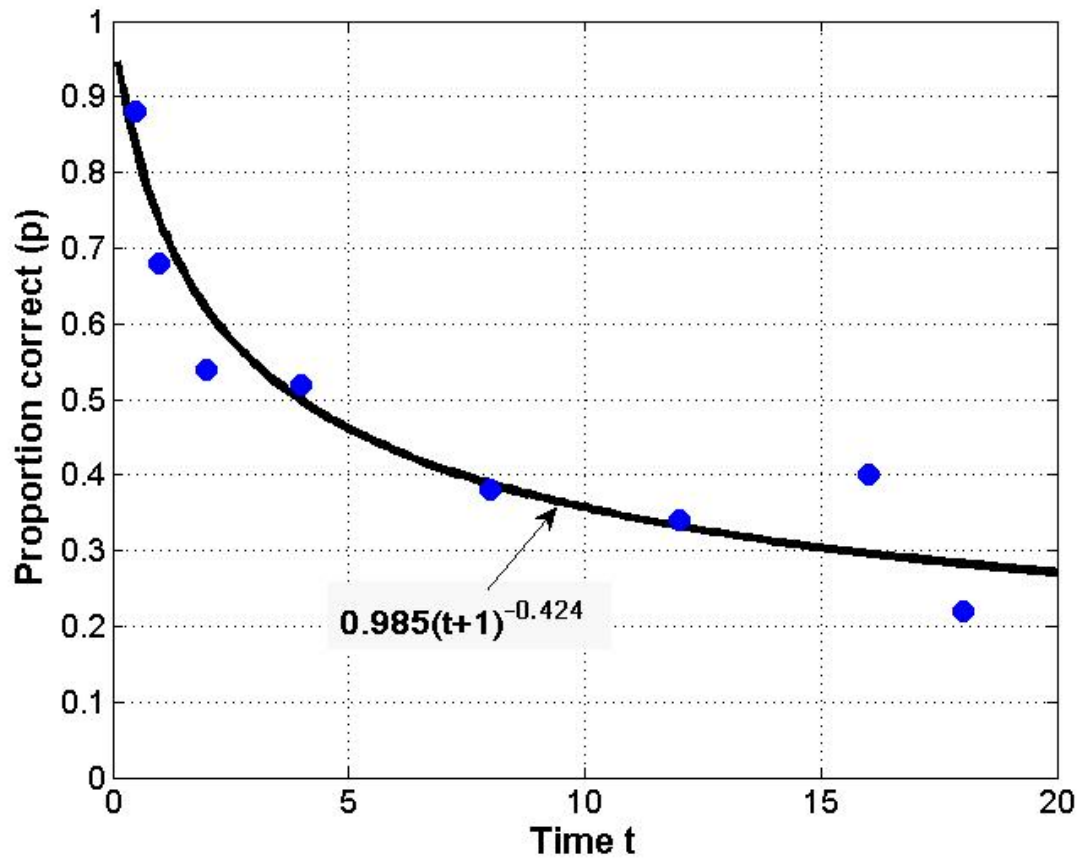
Power model:

$$p = w_1(t + 1)^{-w_2}$$



Model Fitting

Finding the parameter value that **best fits** observed data



Best-fit parameter: $(\hat{w}_1, \hat{w}_2) = (0.985, 0.424)$

Outline

1. Introduction

2. Evaluating Mathematical Models

2a. Model selection/comparison methods

2b. Illustrative examples

3. Evaluating Other Types of Models

4. A New Tool for Model Comparison

5. Final Remarks

2. Evaluating Mathematical Models

Assessing the adequacy of a given model in describing observed data

Goal of Modeling and Approximation to Truth

- The ultimate, ideal goal of modeling is to identify the model that actually generated the observed data
- This is not possible because
 - 1) Never enough observations to pin down the truth exactly
 - 2) The truth may be quite complex, beyond modeler's imagination
- A more realistic goal is to choose among a set of candidate models the one model that provides the “closest approximation” to the truth, **in some defined sense**

Model Evaluation Criteria

- Qualitative criteria

- **Falsifiability:** Do potential observations exist that would be incompatible with the model?
- **Plausibility:** Does the theoretical account of the model make sense of established findings?
- **Interpretability:** Are the components of the model understandable and linked to known processes?

- Quantitative criteria

- **Goodness of fit:** Does the model fit the observed data sufficiently well?
- **Complexity/simplicity:** Is the model's description of the data achieved in the simplest possible manner?
- **Generalizability:** How well does the model predict future observations?

Goodness-of-fit (GOF) Measures

- Root mean squared error (**RMSE**)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_{i,obs} - y_{i,prd}(\hat{W}))^2}{n}}$$

- Percent variance accounted for (**PVAF**), or r^2

$$\text{PVAF} = 100 \left(1 - \frac{\sum_{i=1}^n (y_{i,obs} - y_{i,prd}(\hat{W}))^2}{\sum_{i=1}^n (y_{i,obs} - \bar{y}_{obs})^2} \right)$$

Noisy Data

Behavioral data include random noise from a number of sources, such as measurement error, sampling error, and individual differences

$$\text{Data} = \text{Regularity} + \text{Noise}$$

(Cognitive process) (Idiosyncrasies)

Problem with GOF as Model Evaluation Criterion

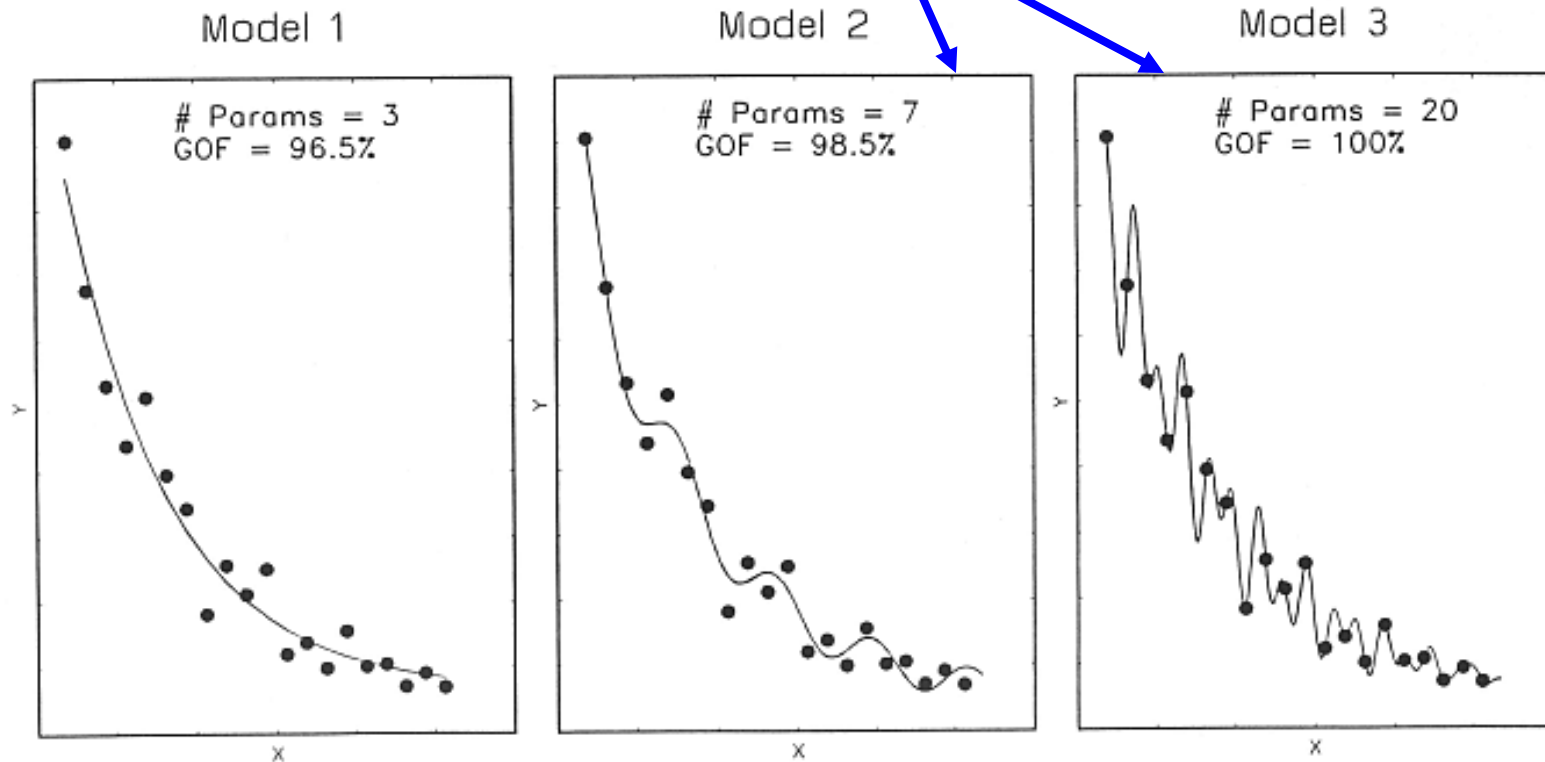
Data = Regularity + Noise
(Cognitive process) (Idiosyncracies)

GOF = Fit to regularity + Fit to noise

Properties of the model that have nothing to do with its ability to capture the underlying regularities can improve GOF

Over-fitting Problem

(Over-fitting)



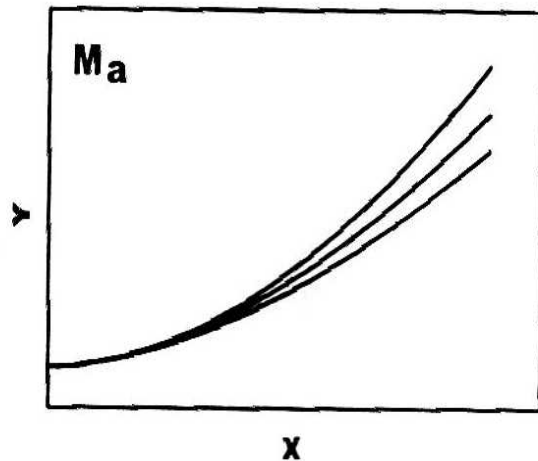
$$\text{Model 1: } Y = ae^{-bX} + c$$

$$\text{Model 2: } Y = ae^{-bX} + c + dX^{-e} \cdot \sin(f \cdot X + g)$$

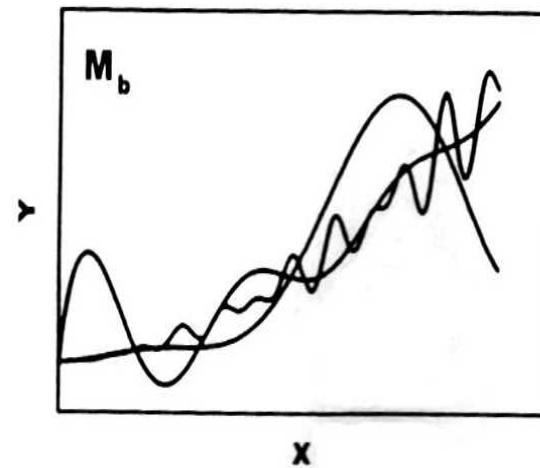
Model Complexity

Complexity: Refers to a model's **inherent flexibility** that enables it to fit a wide range of data patterns

Simple Model



Complex Model



number of model parameters

Complexity: More than Number of Parameters?

Power: $p = a(t + 1)^{-b}$

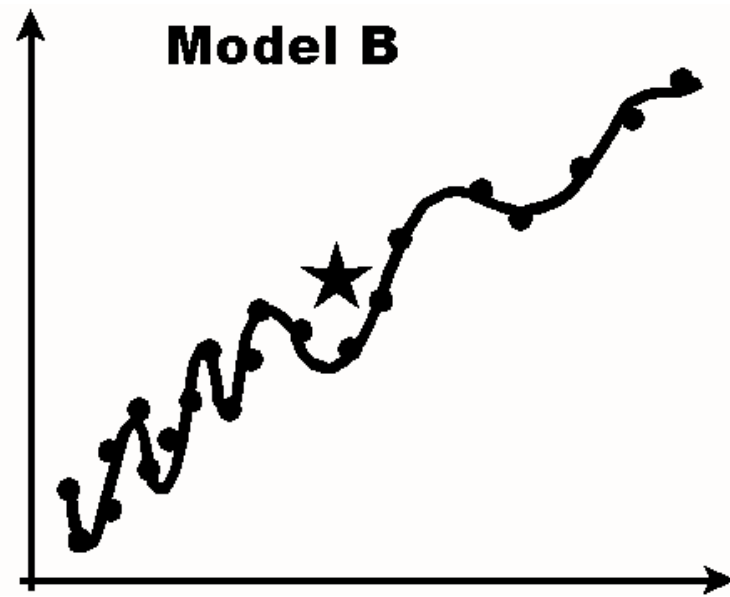
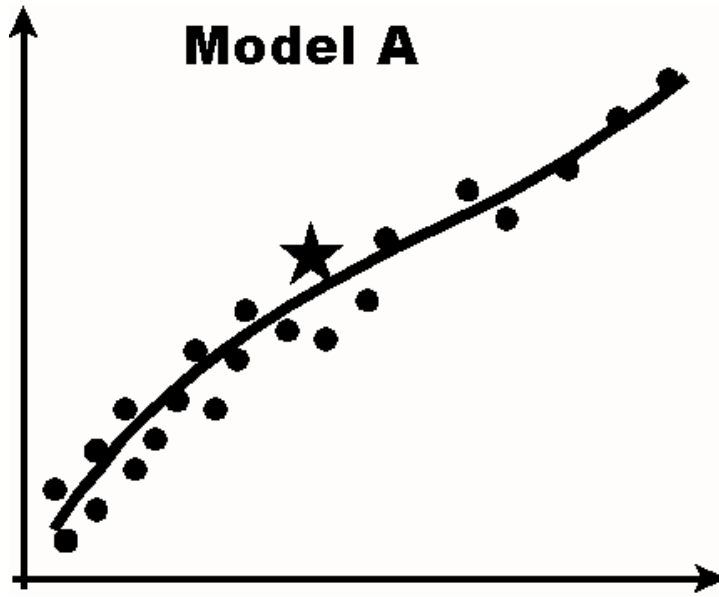
Exponential: $p = ae^{-bt}$

Hyperbolic: $p = 1/(a + bt)$

Are these all equally complex? Maybe not

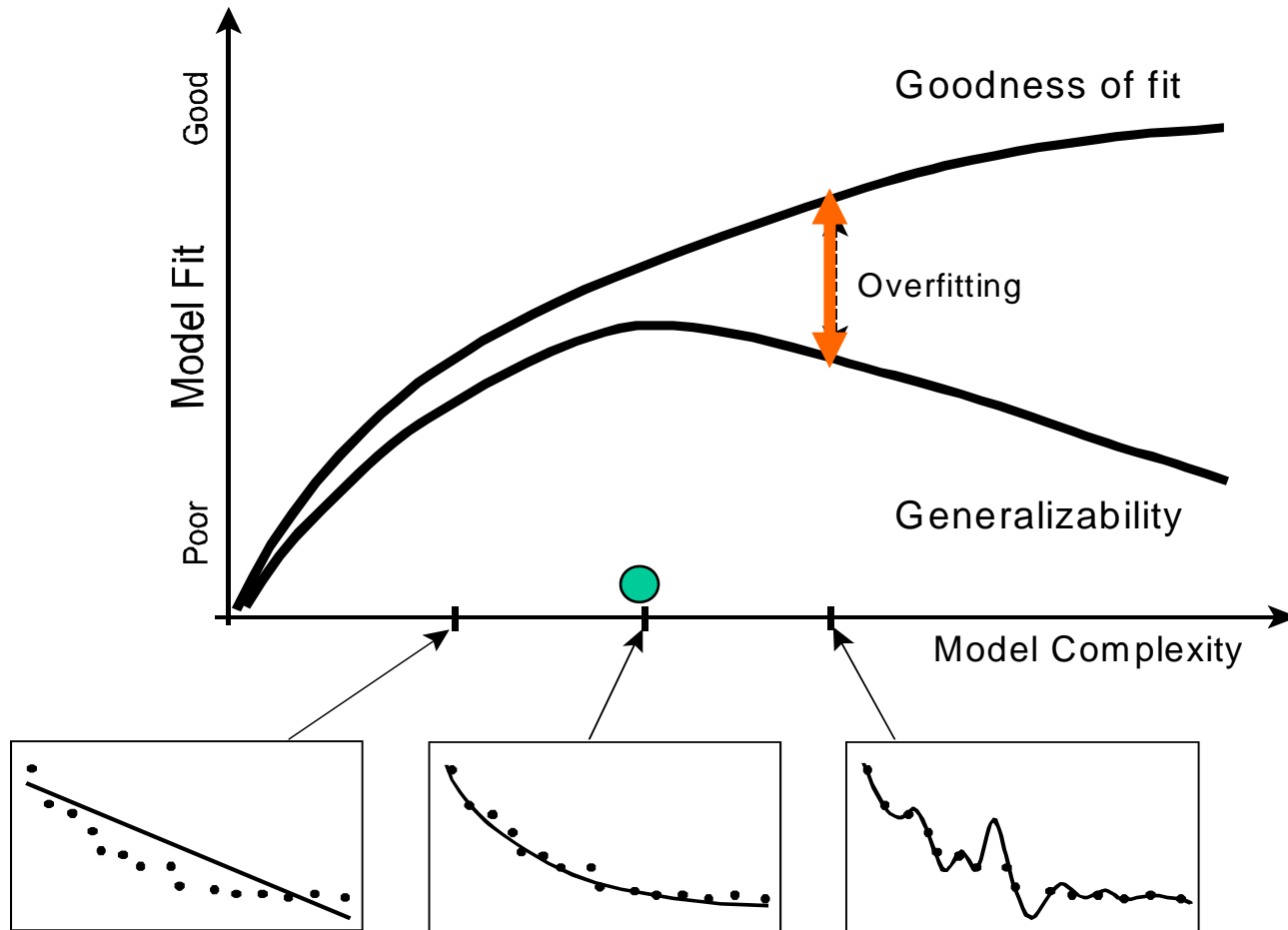
Generalizability: The Yardstick of Model Selection

Generalizability refers to a model's ability to fit all future data samples from the **same** underlying process, not just the currently observed data sample, and thus can be viewed as a measure of **predictive accuracy** or **proximity to the underlying regularity**.



	Model A	Model B
Goodness of fit (PVAF):	80%	99%
Generalizability (PVAF):	70%	50%

Relationship among Goodness of Fit, Model Complexity and Generalizability



Wanted:

A method of model selection that estimates a model's **generalizability** by taking into account effects of its **complexity**

Selection Criterion: Choose the model, among a set of candidate models, that generalizes best

2a. Model Selection Methods

- Occam's Razor
- Likelihood Function
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Minimum Description Length (MDL)
- Bayes Factor (BF)
- Cross-validation (CV)
- Likelihood Ratio Test (LRT)

Occam's Razor: The Economy of Explanation

“Entities should not be multiplied beyond necessity.”

- William of Occam (1288 – 1348)



Likelihood Function

Formally speaking, a mathematical model is defined in terms of the **likelihood function (LF)** that specifies the likelihood of observed data as a function of model parameter:

Likelihood function (LF): $f(y | w)$

(e.g.)

Power model: $p = w_1(t + 1)^{w_2}$ $(0 < p < 1)$

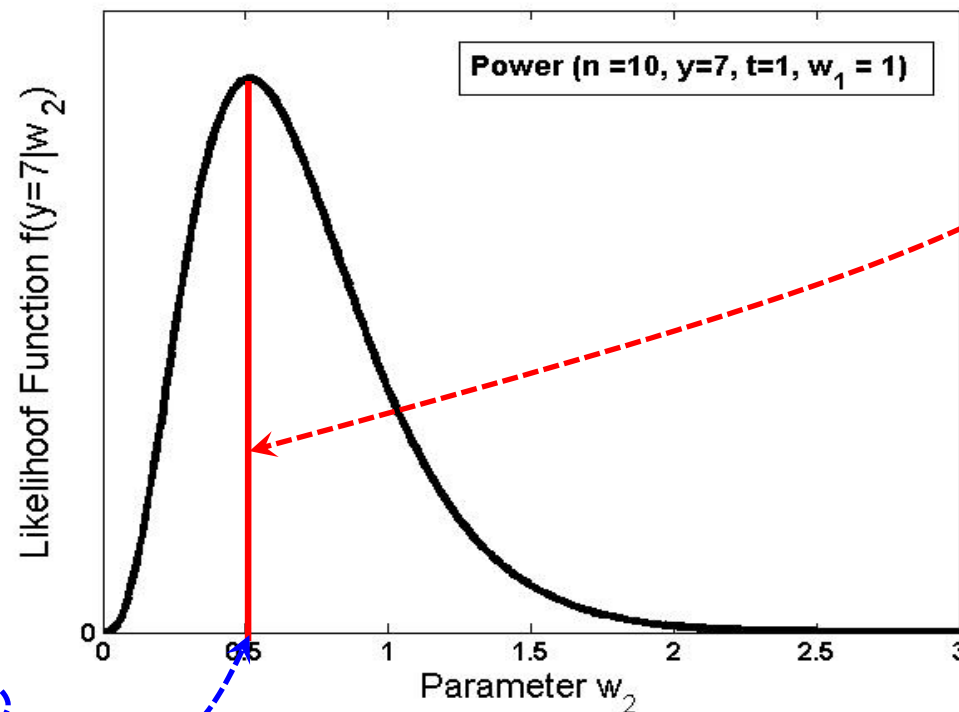
Data: $y \sim \text{Binomial}(n, p)$ $(y = 0, 1, \dots, n)$

LF: $f(y | w) = \frac{n!}{(n - y)! y!} p^y (1 - p)^{n - y}$

Maximum Likelihood

In model fitting, we are interested in finding the parameter value that is most likely to have generated the observed data -- the one that maximizes the likelihood function:

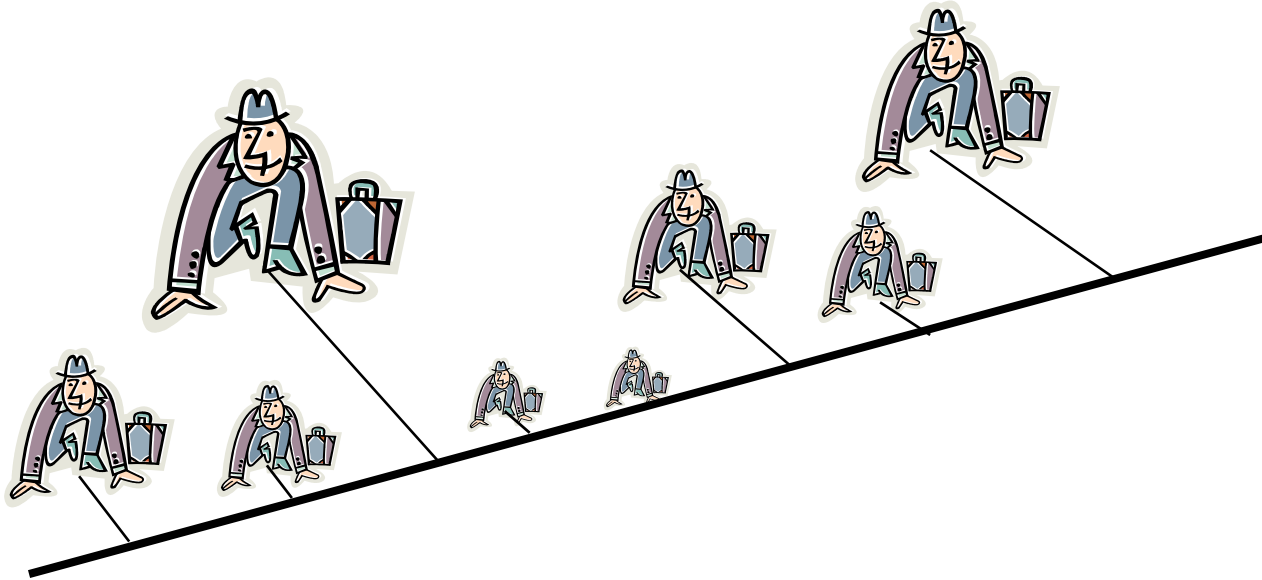
Maximum likelihood (ML): $f(y | \hat{w})$



$$\hat{w}_2 = 0.507$$

Penalized Likelihood Methods

- Formalization of Occam's razor
- Estimate a model's generalizability by **penalizing** it for excess complexity (i.e., more complexity than is needed to fit the regularity in the data)
- Puts models on equal footing



Reminder

- (Generalizability) = - (Goodness of fit) + (Model complexity)

Akaike Information Criterion (AIC)

Akaike (1973):

$$AIC = \underbrace{-2 \ln f(y | \hat{w})}_{\text{Goodness of fit (ML)}} + \underbrace{2k}_{\text{Model Complexity}}$$

of parameters
↓

- The smaller AIC value of a model, the greater generalizability of the model
- The model with smallest AIC is the best, among a set of competing models and thus should be preferred

Bayesian Information Criterion (BIC)

Schwarz (1978):

$$BIC = \underbrace{-2 \ln f(y | \hat{w})}_{\text{Goodness of fit (ML)}} + \underbrace{k \ln n}_{\text{Model Complexity}}$$

Sample size
↓

The model that minimizes BIC should be preferred

Minimum Description Length (MDL)

Rissanen (1996):

$$MDL = \underbrace{-\ln f(y | \hat{w})}_{\text{Goodness of fit (ML)}} + \underbrace{\frac{k}{2} \ln \frac{n}{2\pi} + \ln \int \sqrt{|I(w)|} dw}_{\text{Model Complexity}}$$

functional form
↙

The model that minimizes MDL should be preferred

Bayes Factor (BF)

(Kass & Raftery, 1995)

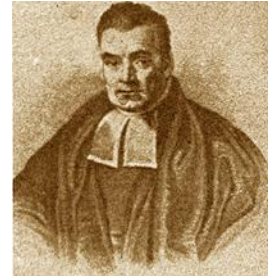
- In Bayesian model selection, each model is evaluated based on its **marginal likelihood** defined as

$$p(y | M) = \int f(y | w, M) \pi(w | M) dw$$

or equivalently, an **average likelihood** (i.e., how well the model fits the data on average, across the range of its parameters)

- **Bayes factor (BF)** between two models is defined as the ratio of two marginal likelihoods

$$BF_{ij} \equiv \frac{p(y | M_i)}{p(y | M_j)}$$



- Under the assumption of equal model priors, BF is reduced to the *posterior odds*:

$$BF_{ij} = \frac{p(M_i | y)}{p(M_j | y)} \quad (\text{from Bayes rule})$$

- Therefore, the model that maximizes marginal likelihood is the one with highest probability of being “true” given observed data

Features of Bayes Factor

- Pros
 - No optimization (i.e., no maximum likelihood)
 - No explicit measure of model complexity
 - No overfitting, by averaging likelihood function across parameters
- Cons
 - Issue of choosing parameter priors (virtue or vice?)
 - Non-trivial computations requiring numerical integration

BIC as an approximation of BF

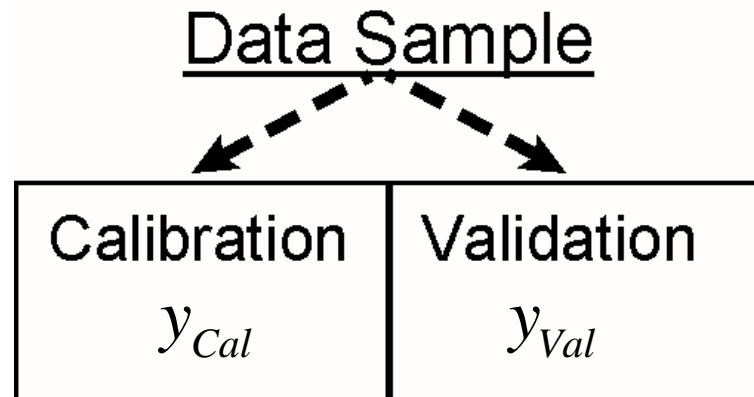
A large sample approximation of the marginal likelihood yields the easily-computable *Bayesian Information Criterion (BIC)*:

$$-2 \log \text{marginal likelihood} \approx BIC$$

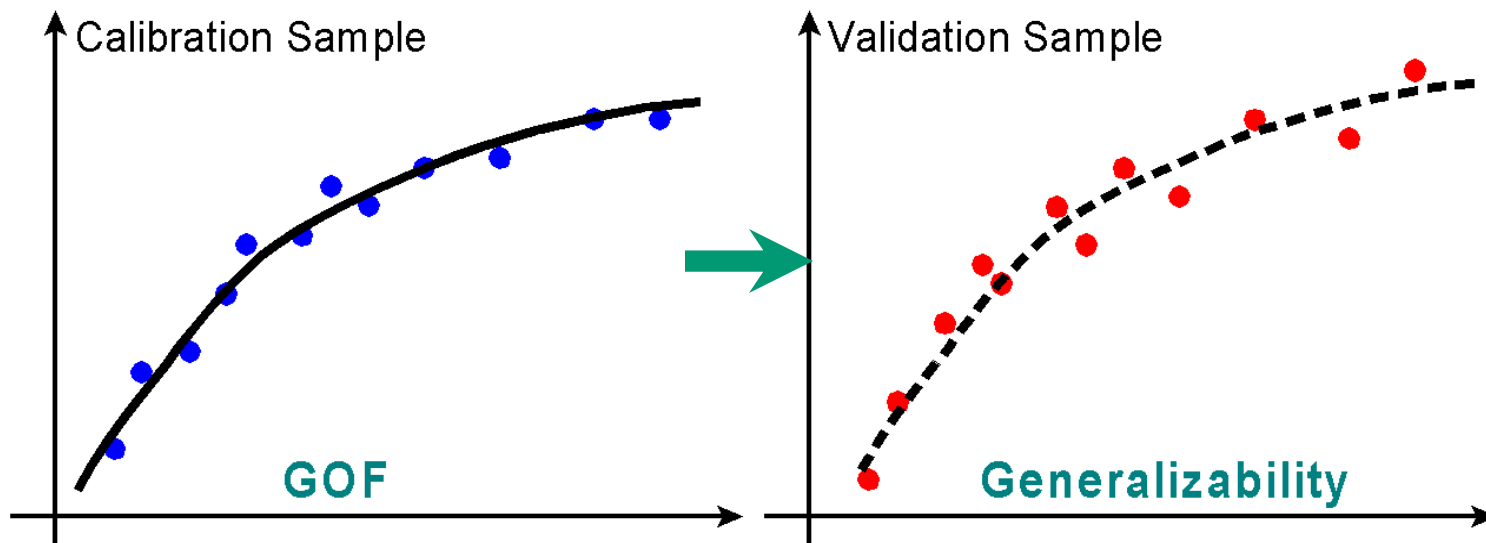
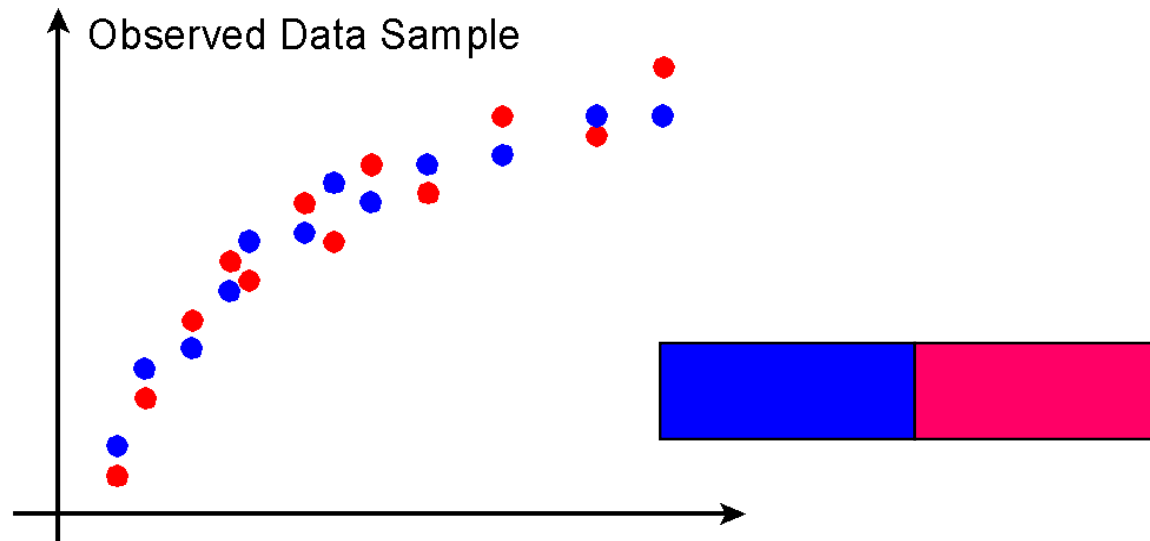
Cross-validation (CV)

(Stone, 1974; Geisser, 1975)

Sampling-based method of estimating generalizability



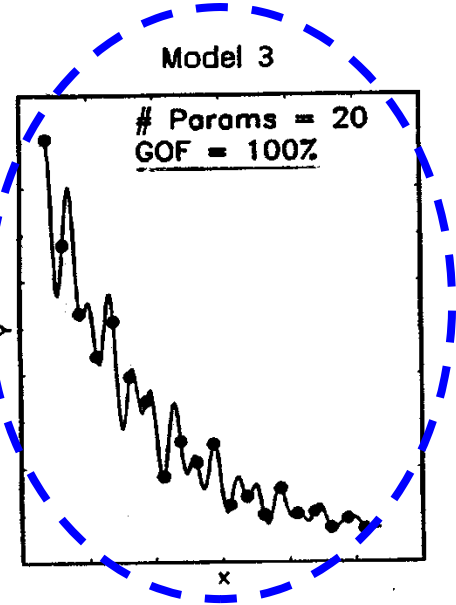
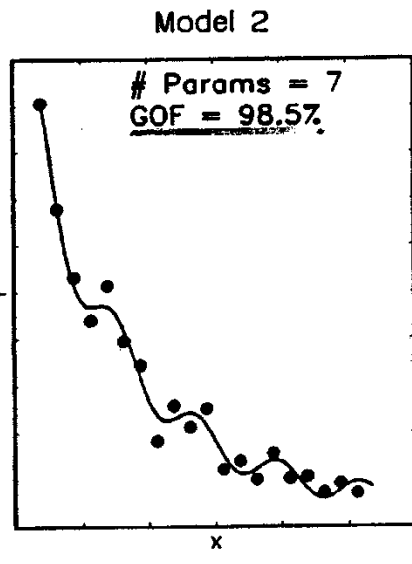
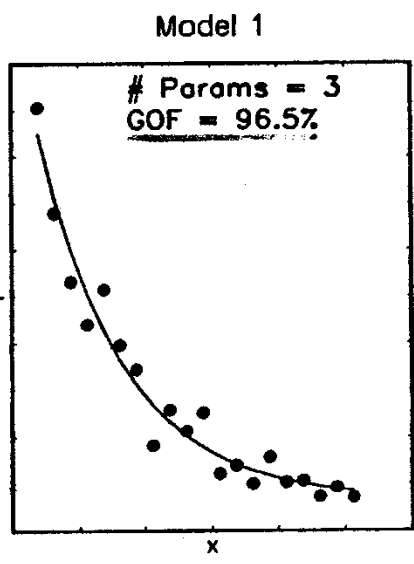
$$CV = SSE(y_{Val} | \hat{w}(y_{Cal}))$$



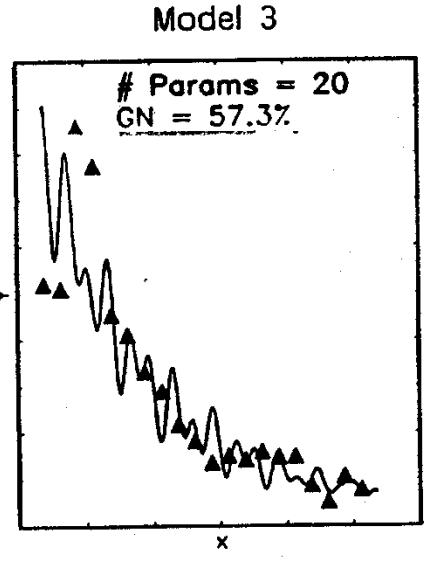
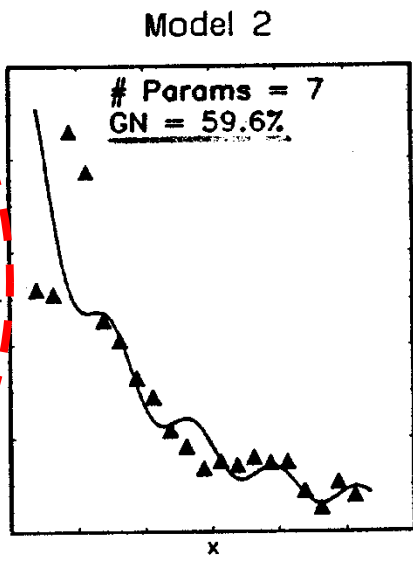
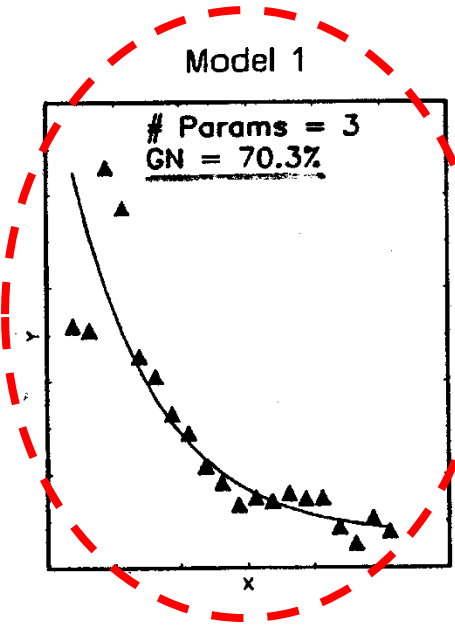
Features of CV

- Pros
 - Easy to use
 - Sensitive to functional form as well as number of parameters
 - Asymptotically equivalent to AIC
- Cons
 - Sensitive to the partitioning used
 - Averaging over multiple partitions
 - *Leave-one-out CV (LOOCV)*, instead of *split-half CV*
 - Instability of the estimate due to “loss” of data

Calibration Sample



Validation Sample



Why not Likelihood Ratio Test (LRT)?

- The LRT is often used to test the adequacy of a model of interest (M_R) in the context of another, fuller model (M_F):

$$LRT = -\ln \frac{f(y | \hat{w}_R)}{f(y | \hat{w}_F)}$$

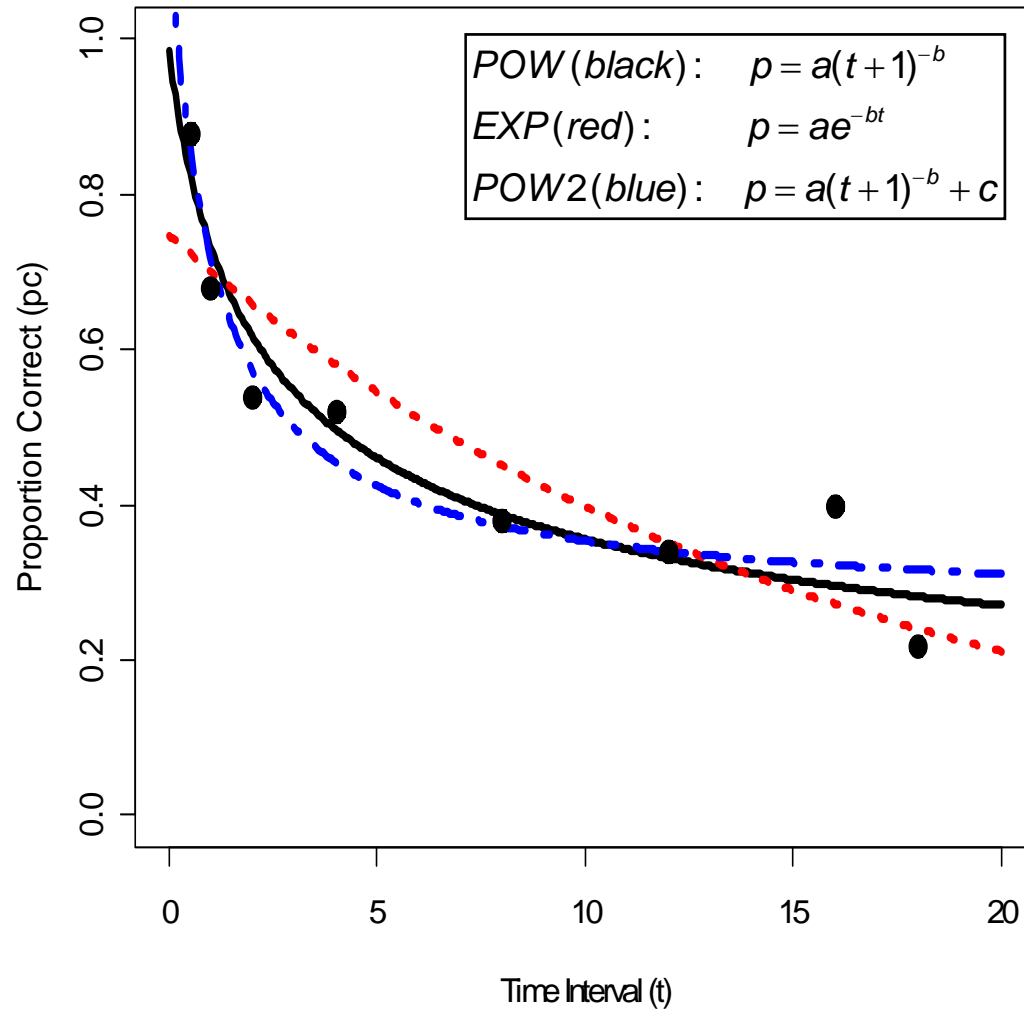
- The **acceptance/rejection** decision on M_R is then made based on a **p-value** obtained from Chi-square distribution with $df = k_F - k_R$.
- **Acceptance** \rightarrow “Model M_R provides a sufficiently good fit to observed data, and therefore the extra parameters of model M_F are judged to be unnecessary”

LRT is NOT a model selection method

- LRT does not access a model's generalizability, which is the hallmark of model selection.
- LRT is a **null hypothesis significance testing** (NHST) method that simply judges descriptive adequacy of a given model (*cf.* contemporary criticisms of NHST).
- LRT requires the setting of **pair-wise & nested** comparison (i.e., can compare only two models at a time).
- LRT is developed in the context of **linear models with normal error**; Sampling distributions of the LRT statistic under nonlinear and non-normal error conditions are generally unknown.
- At best, LRT is a poor substitute for current model selection methods, such as AIC and BIC.

2b. Illustrative Examples

Example #1



Model Selection Results

	POW	EXP	POW2
# Params	2	2	3
PVAF	91.2	79.0	92.6
AIC	498.67	508.11	499.35
BIC	502.50	511.93	505.09
LOOCV _{loglik}	-31.409	-32.529	-31.644

Example #2

Selection method	Model fitted	Model the data were generated from		
		M_1	M_2	M_3
PVAf	M_1	0	0	0
	M_2	38	97	30
	M_3	62	3	70
AIC	M_1	79	0	0
	M_2	9	97	30
	M_3	12	3	70
MDL	M_1	86	0	0
	M_2	1	92	8
	M_3	13	8	92

$$M1: p = (t + 1)^{-a}$$

$$M2: p = (t + b)^{-a}$$

$$M3: p = (bt + 1)^{-a}$$

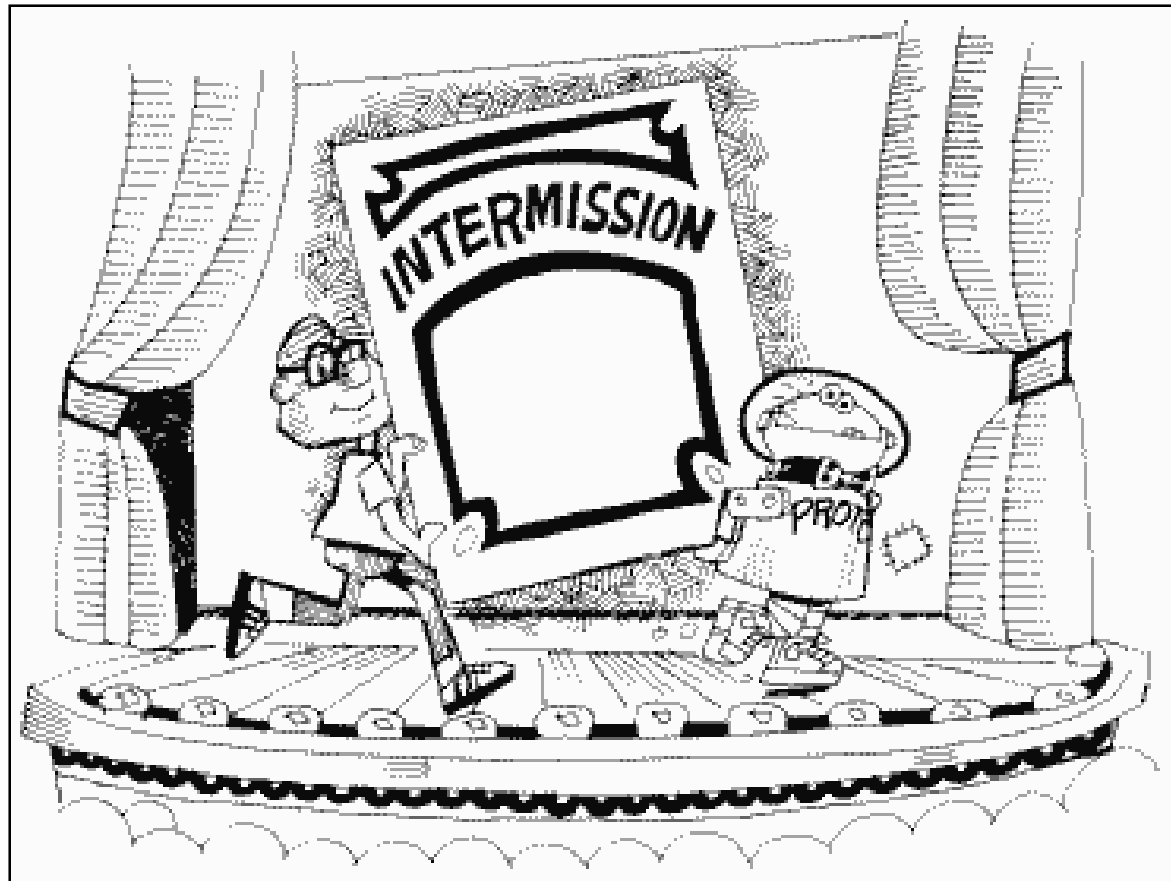
Interim Conclusion

- Models should be evaluated based on **generalizability**, not on **goodness of fit**



“Thou shall not select the **best-fitting** model but shall select the **best-predicting** model.”

- Other non-statistical but *very important* selection criteria
 - Plausibility
 - Interpretability
 - Explanatory adequacy
 - Falsifiability



How should one decide between competing models of data?

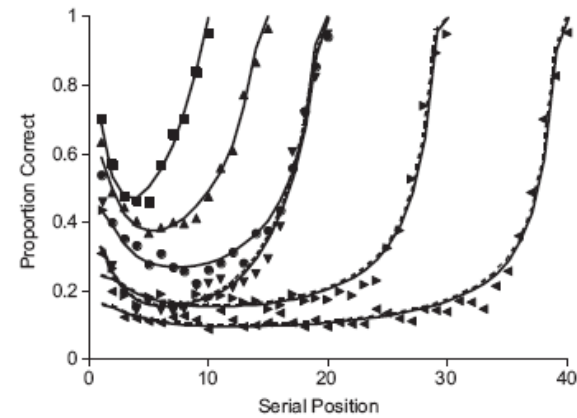
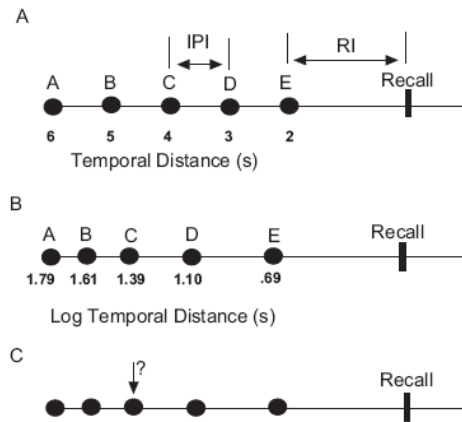
- Evaluate a model's fit to data (descriptive adequacy)
- Consider a model's fit to other possible data sets (complexity/flexibility)
- Normalize model fit to measure generalizability (MDL, Bayes Factor, etc.)

Restricted Scope of the Methods

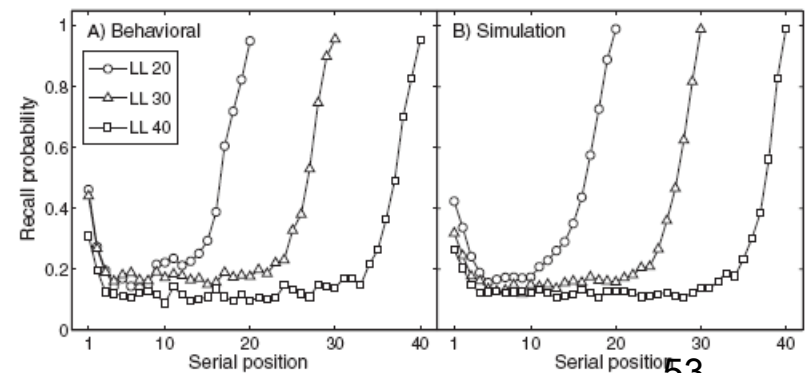
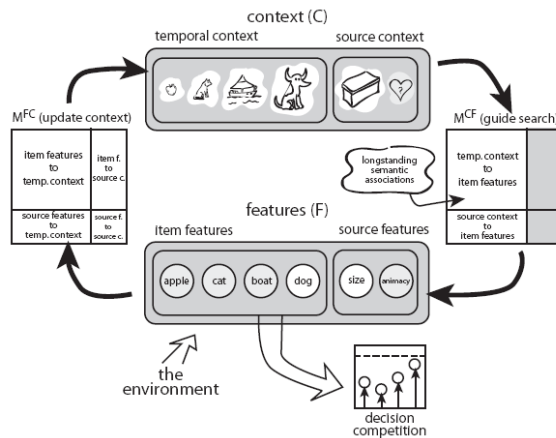
- Models must generate likelihood functions (distribution of fits across parameters)
- Not all models can do this (without simplification)
 - Connectionist
 - Simulation-based (CMR, Diffusion)
 - Cognitive architectures
- Diversity of types of models in cognitive science makes model comparison challenging

Model Comparison in Cognitive Science

SIMPLE model of memory (Brown et al, 2007)



CMR model of memory (Polyn et al, 2009)



Outline

1. Introduction
2. Evaluating Mathematical Models
 - 2a. Model selection/comparison methods
 - 2b. Illustrative examples
- 3. Evaluating Other Types of Models**
4. A New Tool for Model Comparison

Broader Framework

- **Local Model Analysis**
 - Can the model simulate (or fit) the particular data pattern observed in an experimental setting?
 - Evaluate success of model across diverse experiments
 - Difficult to synthesize results to obtain a picture of why the model consistently succeeds in simulating human performance
 - Is it the correct model or an overly flexible model?
 - Is evidence for the model similar across settings?
- To interpret the behavior of a model's local performance, the global behavior of the model must also be understood (i.e., descriptive adequacy must be balanced by model complexity)

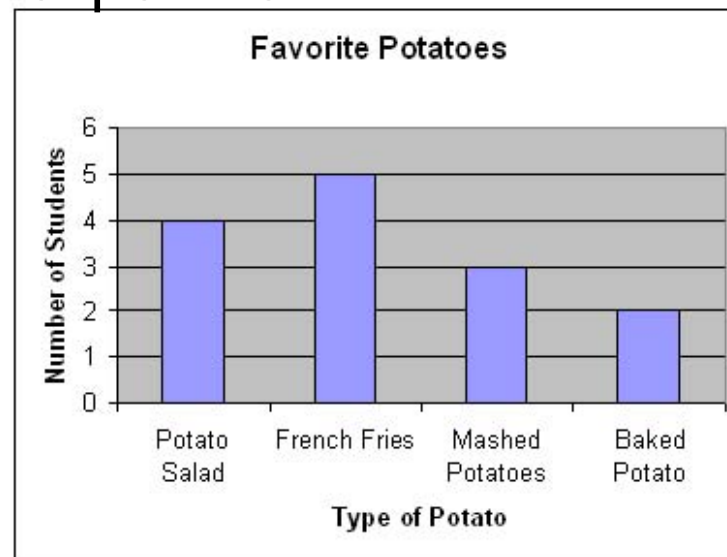
Broader Framework

- **Global Model Analysis**

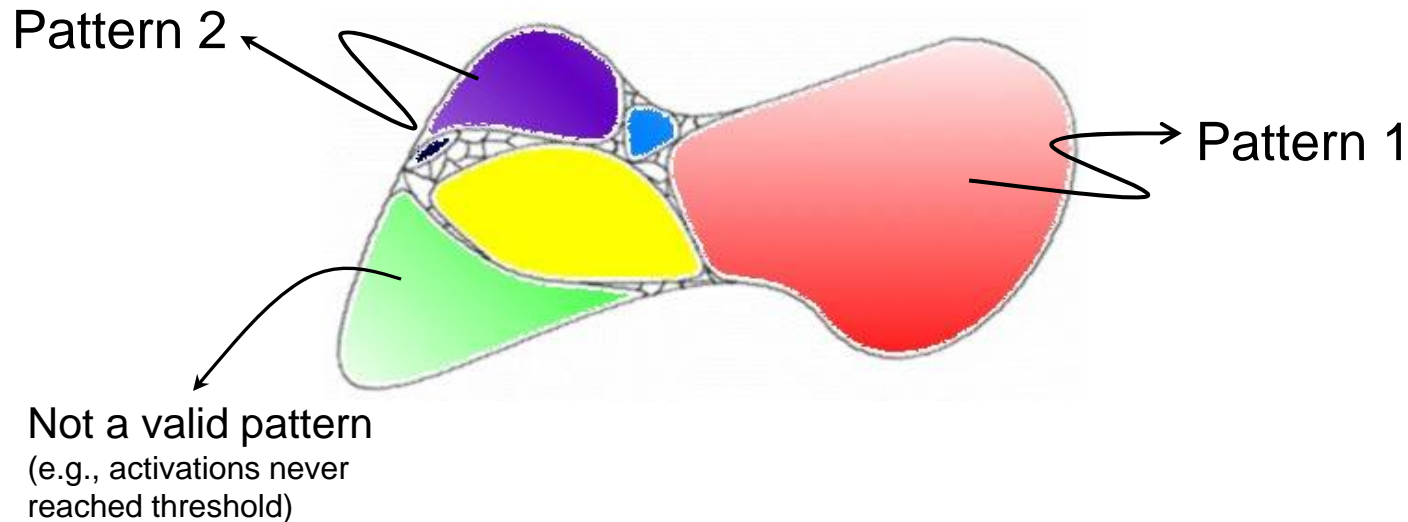
- What **other** data patterns does the model produce besides the empirical one?
- Learn what a model is and is not capable of doing in a particular experimental setting (i.e., experimental design) with the goal being to obtain a more comprehensive picture of its behavior

Parameter Space Partitioning (PSP)

- Global Model Analysis method
- Partition a model's parameter space (hidden weight space) into distinct regions corresponding to **qualitatively** different data patterns the model could generate in an experiment (Pitt, Kim, Navarro & Myung, 2006)
- PSP interfaces between the continuous behavior of models and the often discrete, qualitative predictions across conditions in an experiment



Parameter Space Partitioning

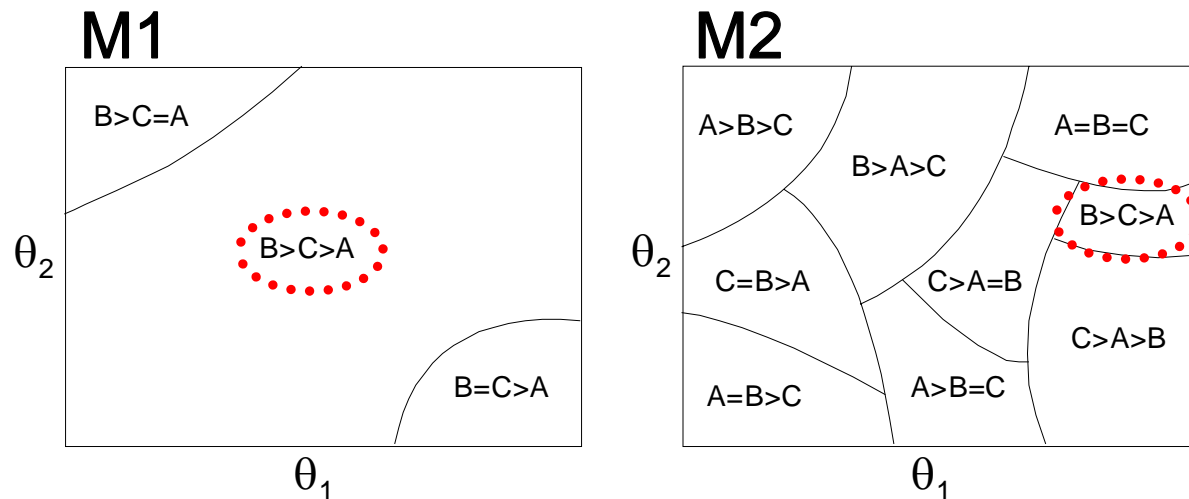


How do you find all data patterns that a model can generate?

- Hard problem
- Potentially huge search space
- Model simulation required for each set of parameter values chosen
- Classify each simulation result (Does the current pattern match others already found or is it new?)

What Can be Learned from PSP?

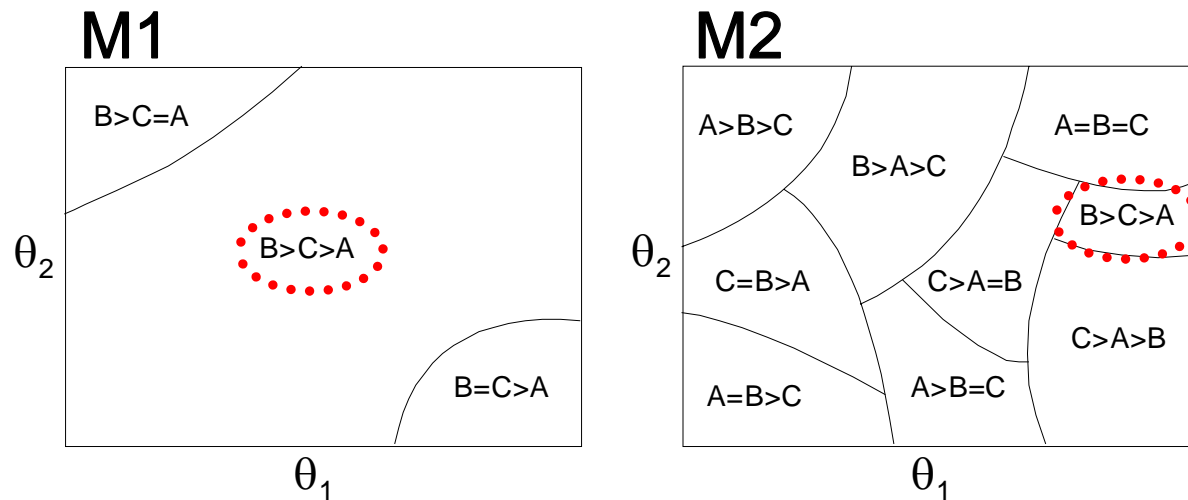
Empirical data pattern across three conditions: **B > C > A**



Questions PSP can answer

- How many of the 13 different data patterns can a model simulate?
- How much of the space is occupied by the empirical pattern? (central/peripheral)
- What is the relationship between these other patterns and the empirical pattern (in terms of volume and similarity)

Is it good or bad if a model's predictions are central or peripheral?

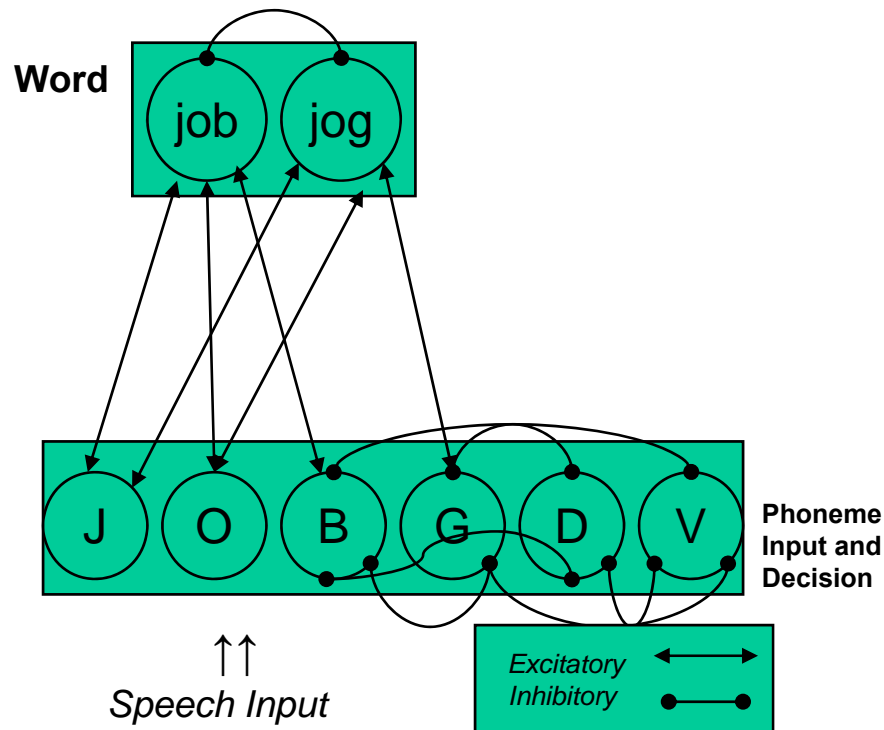


- Simulation success takes on additional meaning with knowledge of other model behaviors
- Model comparison methods do not make decisions for you. They provide you with data that are intended to help you arrive at an informed decision

How does memory for a spoken word influence perception of the word's phonemes?

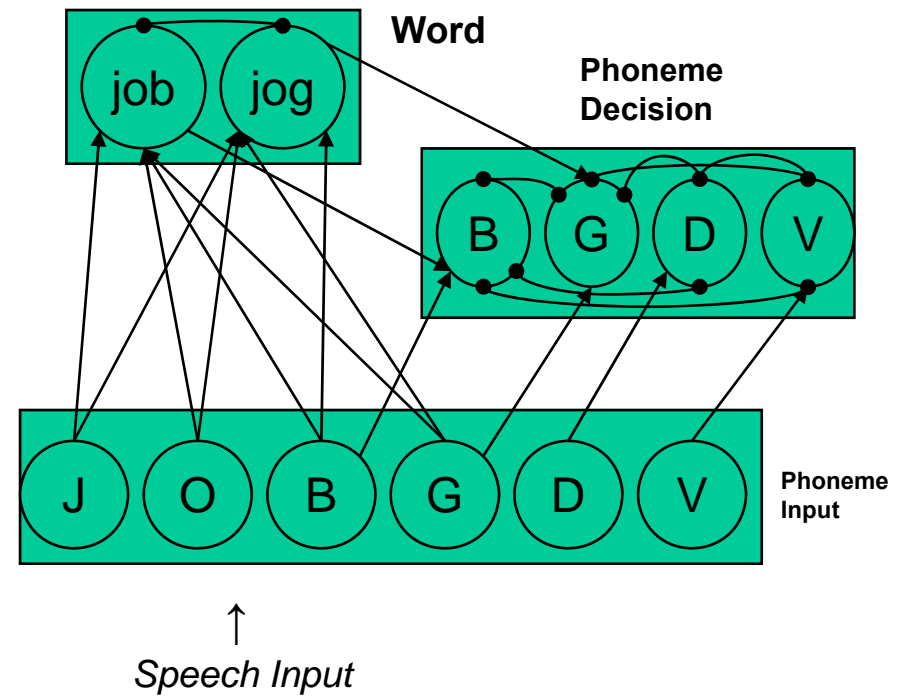
TRACE

McClelland & Elman, 1986



Merge

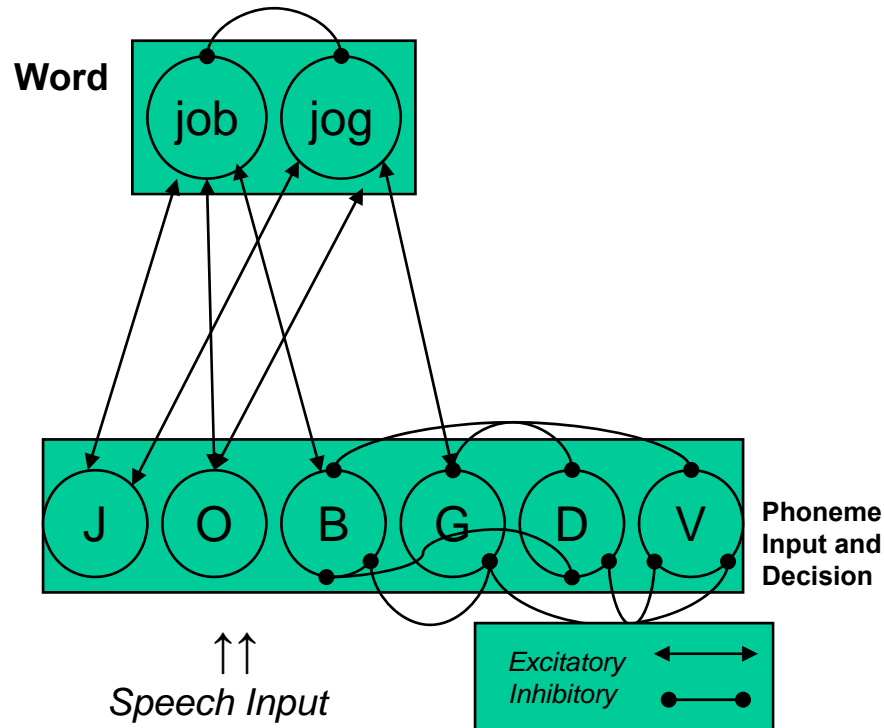
Norris, McQueen, & Cutler, 2001



What are the consequences of splitting the phoneme level in two?

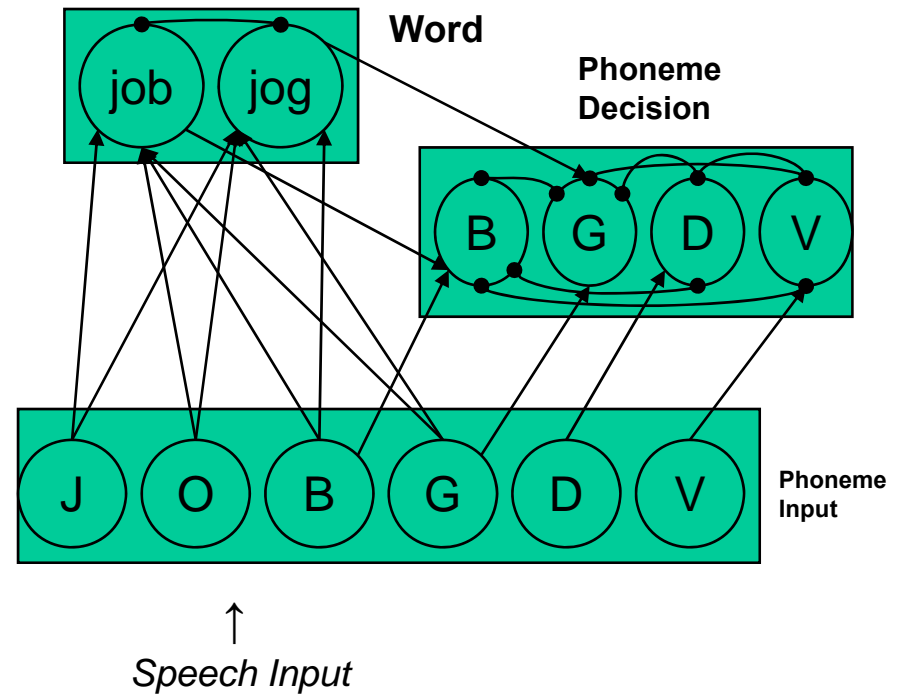
TRACE

McClelland & Elman, 1986



Merge

Norris, McQueen, & Cutler, 2001



PSP Analysis

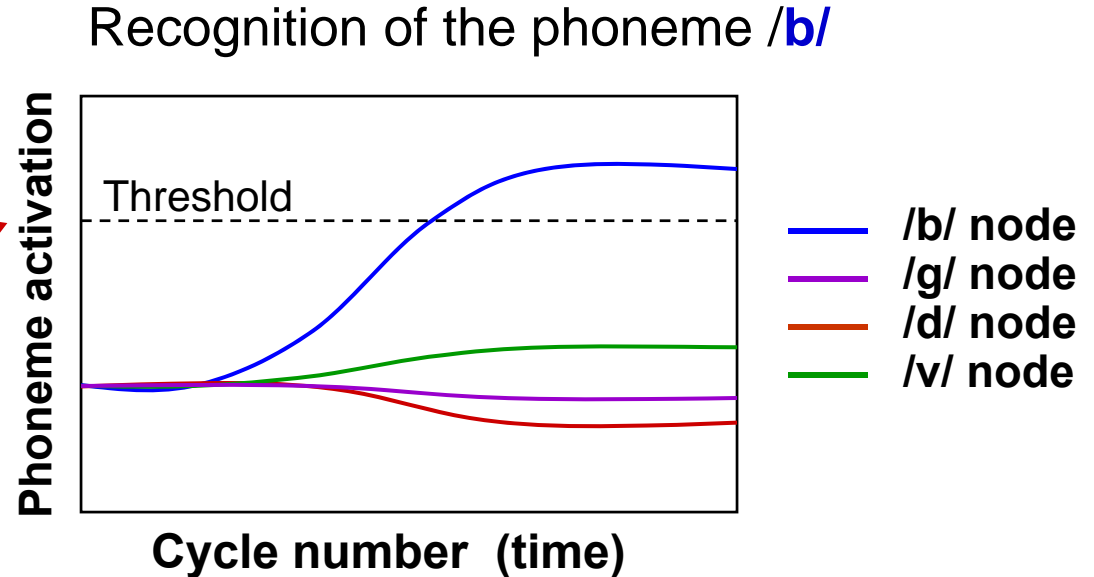
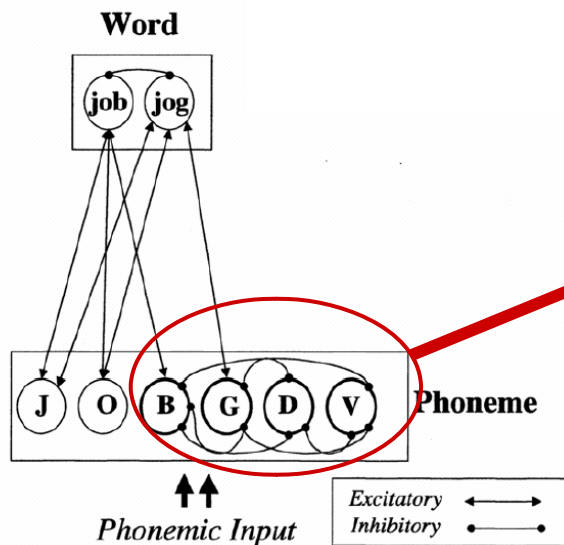
- Compared model performance in a 12-condition experiment
- Empirical data pattern (subcategorical mismatch, Norris et al, 2000)

Condition	Phonemic categorization				Lexical decision		
Word	/b/	/g/	/z/	/v/	'job'	'jog'	nonword
Matching cues	✓				✓		
Mismatching cues from word	✓				✓		
Mismatching cues from nonword	✓				✓		
Nonword	/b/	/g/	/z/	/v/	'job'	'jog'	nonword
Matching cues			✓				✓
Mismatching cues from word			✓				✓
Mismatching cues from nonword			✓				✓

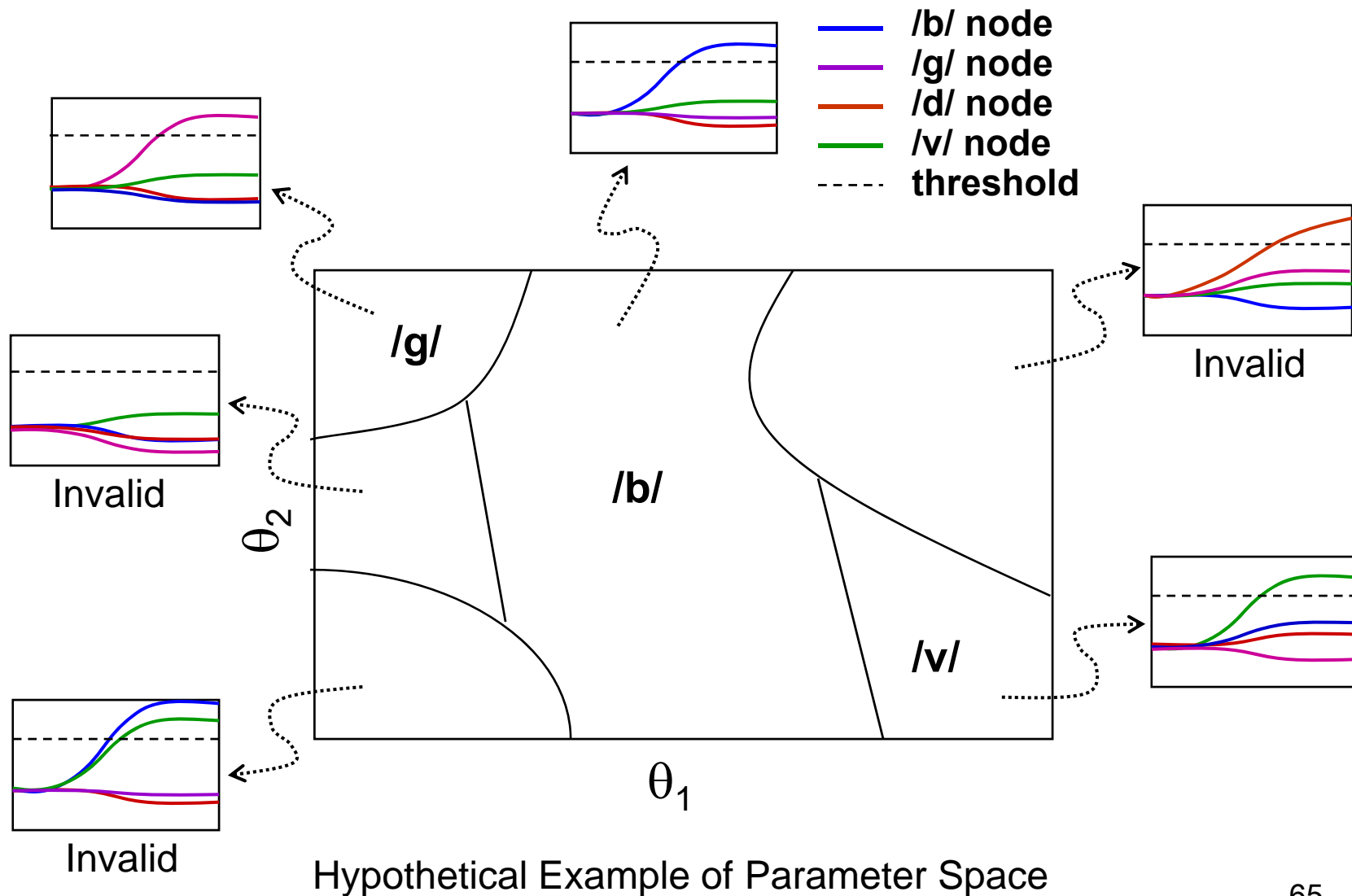
- How many of the possible ~3 million data patterns can each model generate?

PSP Analysis

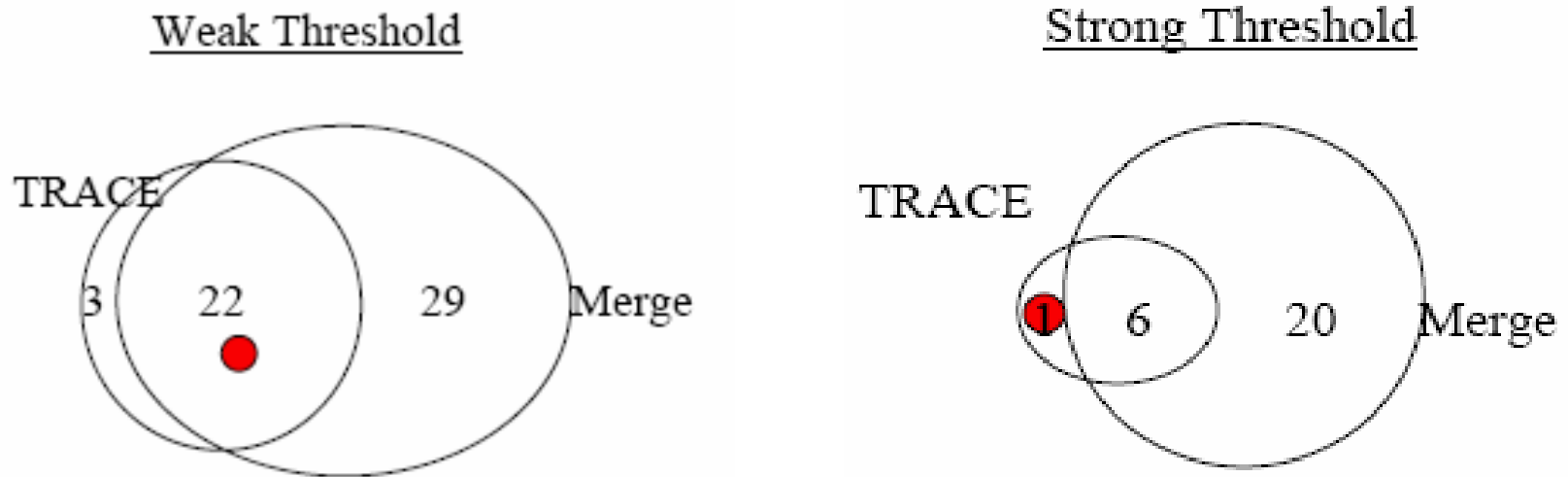
- Phoneme and word recognition defined as node activation reaching threshold



Partitioning of Parameter Space in terms of Phoneme Activation



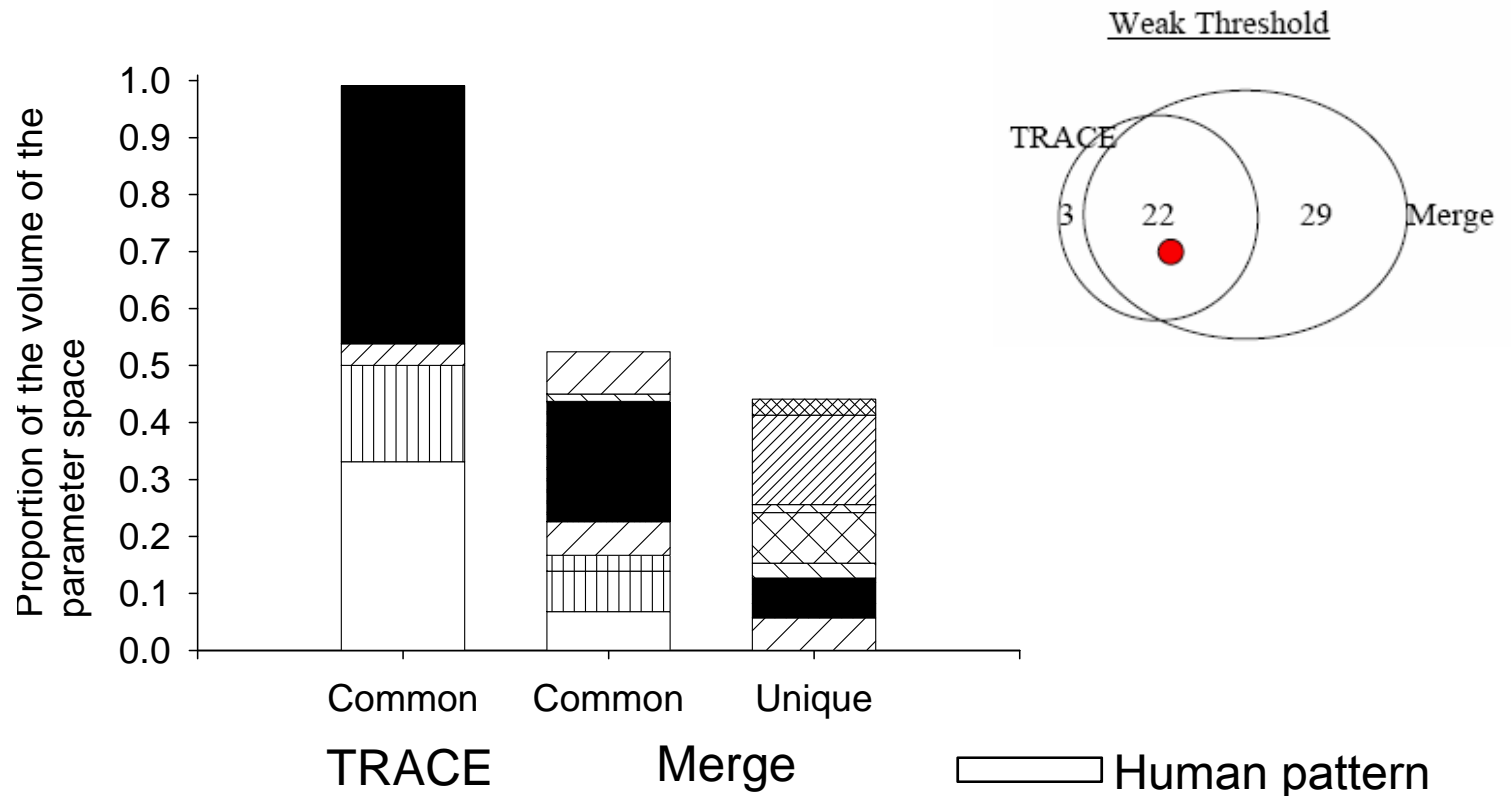
Results of PSP Analysis: Number of Data Patterns



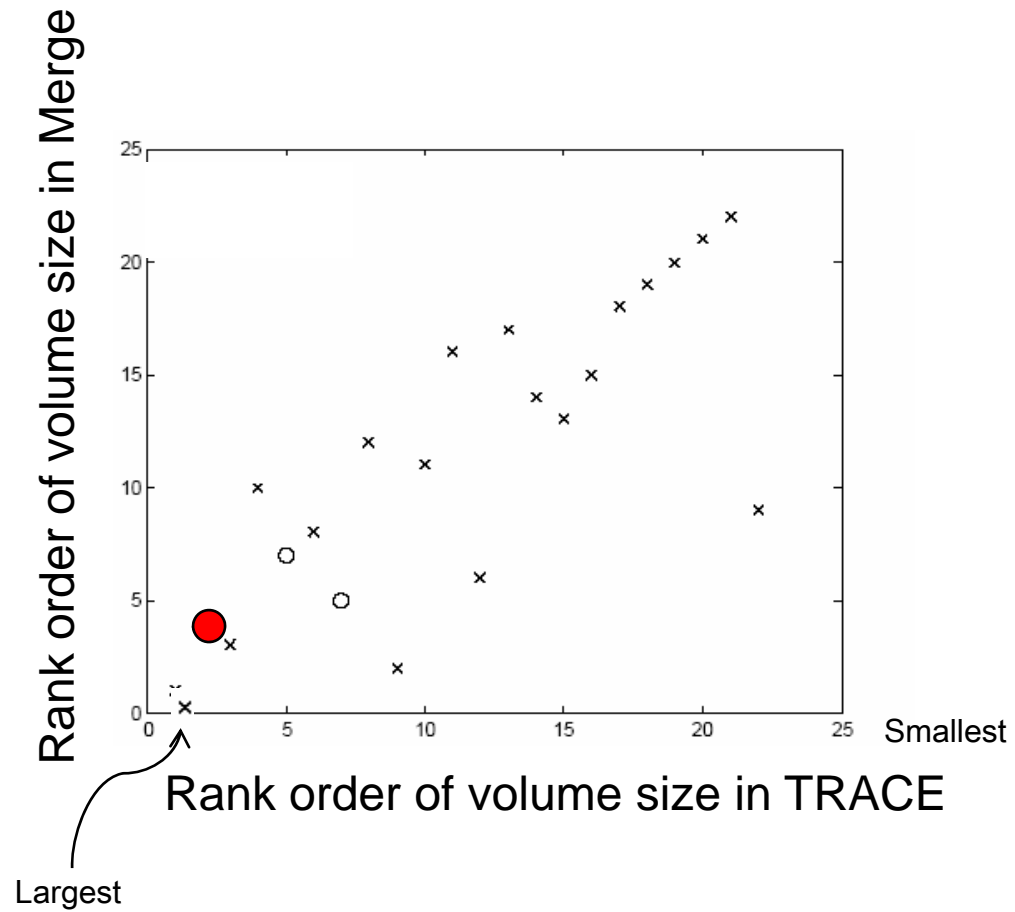
- By splitting the phoneme level in two, Merge is able to generate more data patterns (more flexible)

Volume Analysis (weak threshold only)

Each fill pattern represents a different data pattern
Only patterns that occupy more than 1% of the volume are shown



Correlation of Volumes of Common Regions



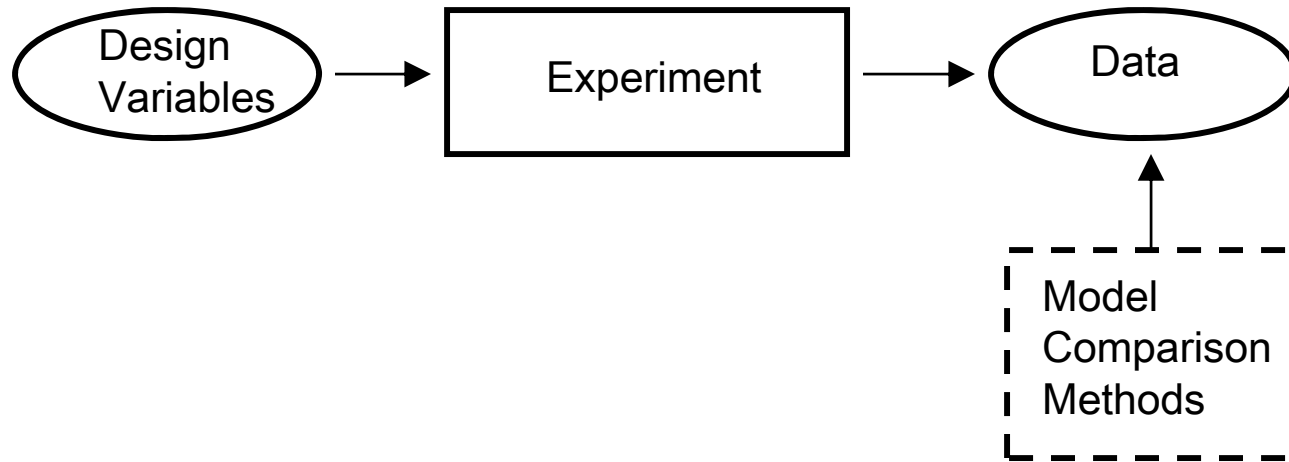
Summary

- Parameter Space Partitioning provides a global perspective on model behavior
- Broader context for interpreting simulation/fit success
- Complements local model analysis
- Assess similarity of models
- Applicable to a wide range of models
- Results are specific to the experimental design
- PSP can be used to evaluate the effectiveness of experiments in distinguishing between models prior to testing (Does only one model generate the predicted pattern?)

Outline

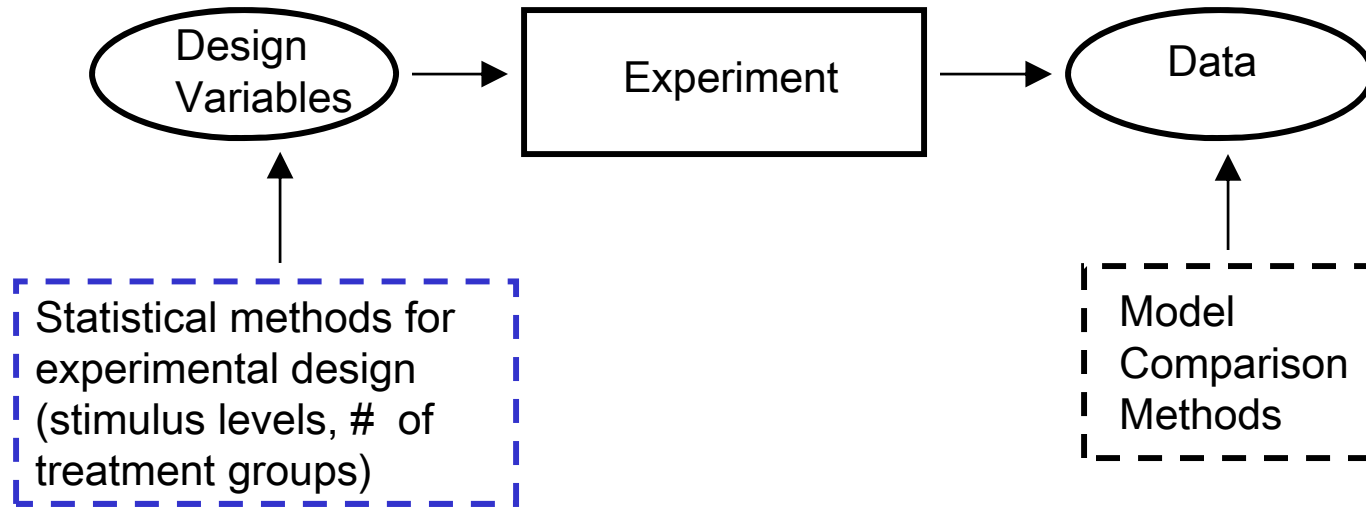
1. Introduction
2. Evaluating Mathematical Models
 - 2a. Model selection/comparison methods
 - 2b. Illustrative examples
3. Evaluating Other Types of Models
- 4. A New Tool for Model Comparison**

Current Model Selection Paradigm



- MDL, Bayes Factor, PSP are, given data from an experiment, tools to assist in making inferences about the models and learning about the models
- Clarity of the answer is limited by the quality of the empirical data

A Complementary Approach to Model Selection



- Improve the quality of the inference (i.e., data) by improving the experimental design
- The clearer the data, the less of a need to rely on model selection methods
- Apply statistical inference methods **before** (or **during**) experimentation

Design Optimization (DO)

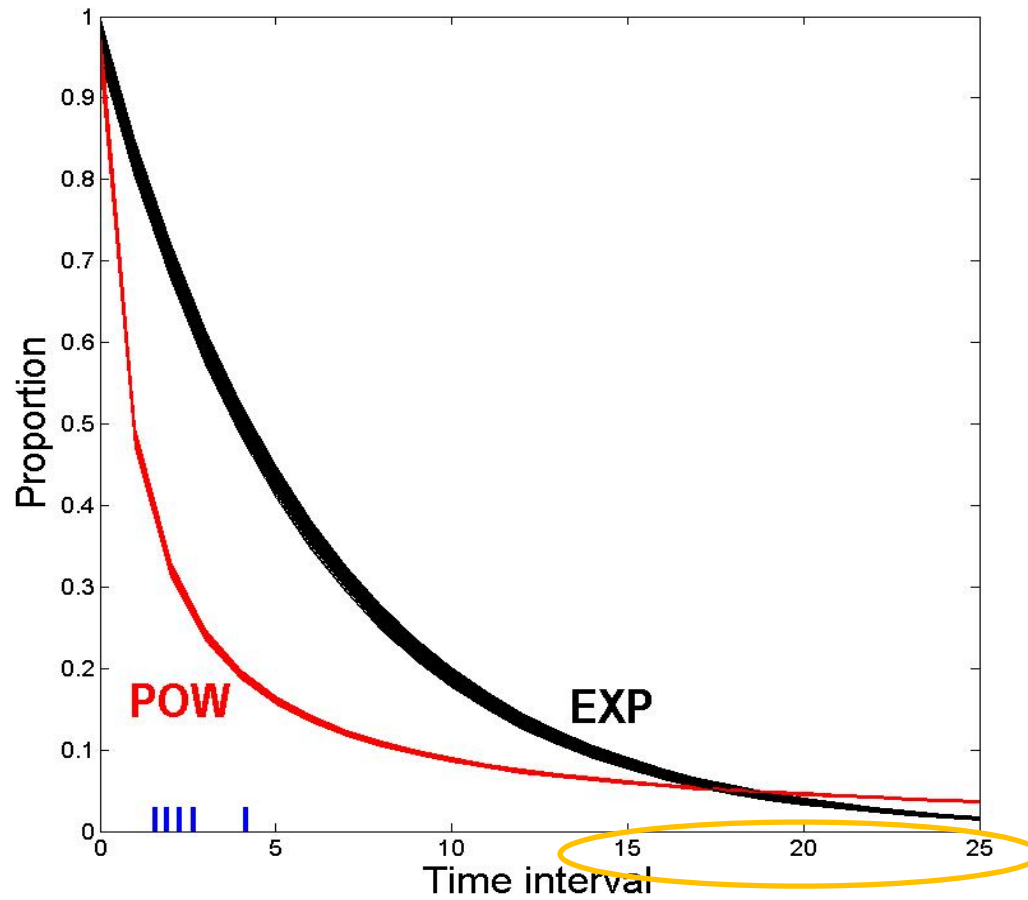
- A method for identifying the optimal experimental design that has the highest likelihood of discriminating between models (Myung & Pitt, 2009)
- An **optimal** experimental design is one that maximizes the informativeness of the experiment while being cost effective for the experimenter (Atkinson & Donev, 1992)



How to find an optimal design?

- As with PSP, difficult search problem
 - Models' parameter spaces must be searched
 - Experimental design space must be searched
 - What is happening during the search?
 - Each design is treated as a gamble whose payoff is determined by the outcome of a hypothetical experiment carried out with that design (e.g., using Bayes Factor)
 - Computationally very time consuming (Sequential Monte Carlo)
 - Choose the most discriminating design
- Model must have a likelihood function

Illustrative Example of Design Optimization Retention Models

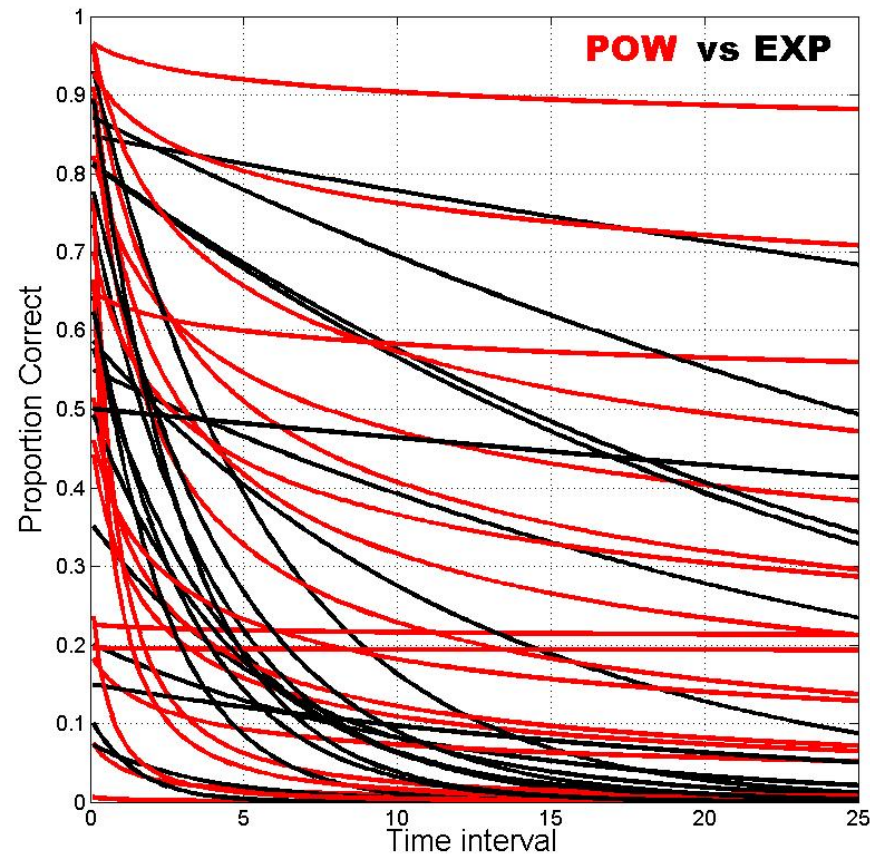


POW: $p = a(t+1)^{-b}$ $(0.95 < a < 1; 1.00 < b < 1.01)$

EXP: $p = ae^{-bt}$ $(0.95 < a < 1; 0.16 < b < 0.17)$

Restricted parameter ranges

True Difficulty of Discriminating the Models



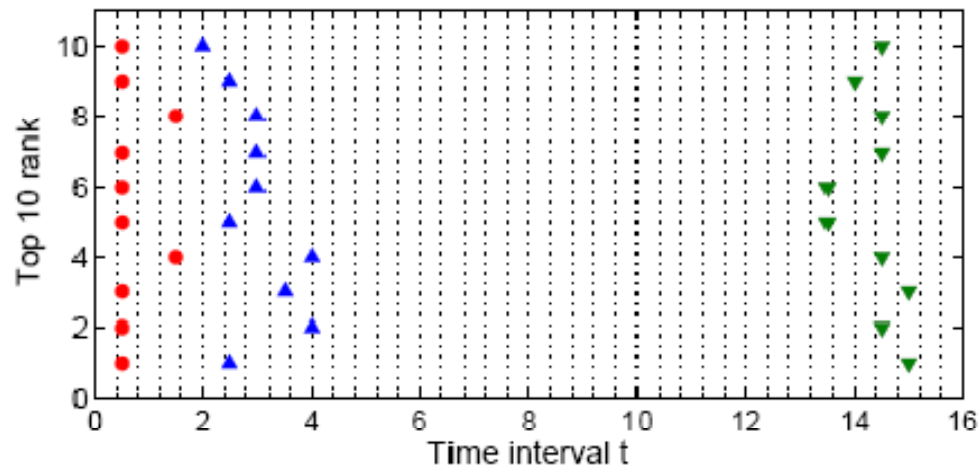
$$\text{POW: } p = a(t+1)^{-b} \quad (0 < a < 1; 0 < b < 3)$$

$$\text{EXP: } p = ae^{-bt} \quad (0 < a < 1; 0 < b < 3)$$

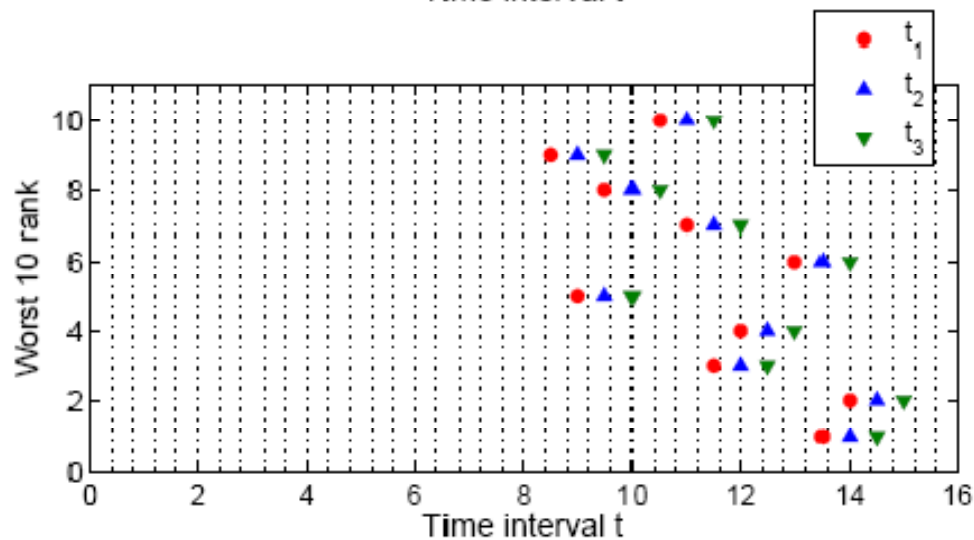
Best and Worst Designs for the two models

Experiment with three retention intervals

Best

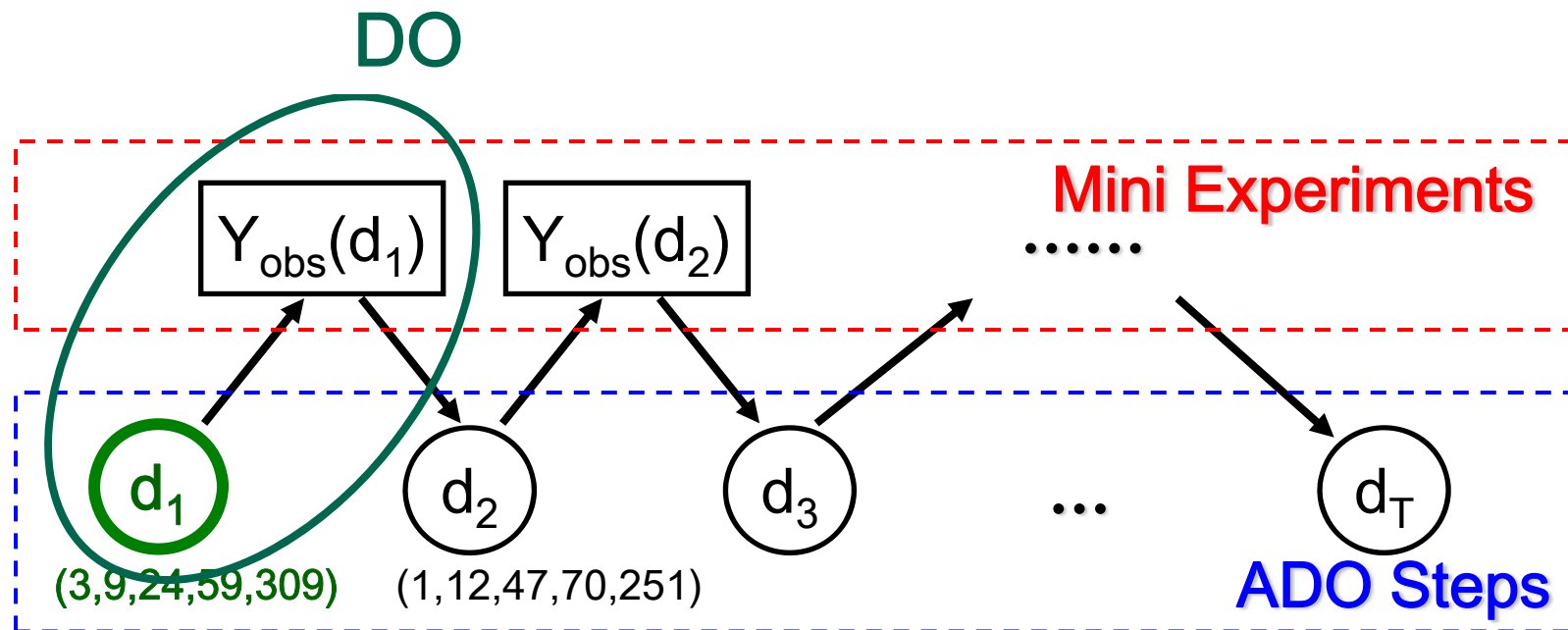


Worst



Adaptive Design Optimization (ADO)

- **Re-optimize the design throughout the experiment**
 - Break the experiment into a series of **mini experiments**
 - Improve the design of the next **mini experiment** using knowledge gained from the previous **mini experiment**



Retention Experiment Using ADO

(Cavagnaro et al, 2011)

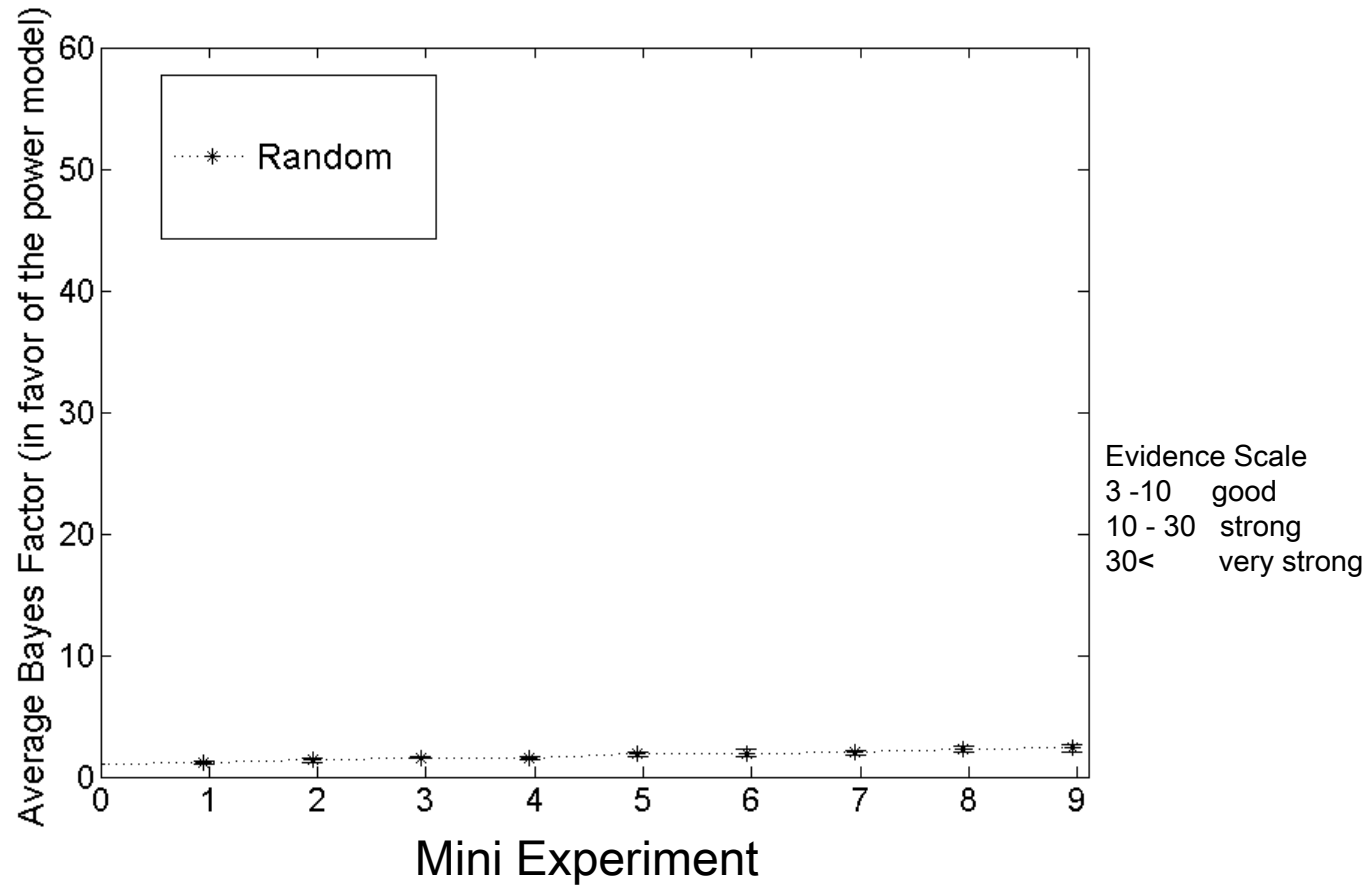
- 8 participants
- Range of retention intervals: 1-40 seconds
- Single retention interval used in each mini experiment
- Three conditions
 - **ADO**: search algorithm chooses each interval
 - **Random**: randomly select each interval
 - **Geometric**: typical spacing of intervals
(1,2,3,5,8,12,18,27,40)
- Does ADO outperform Random and Geometric?

Experiment Set-up

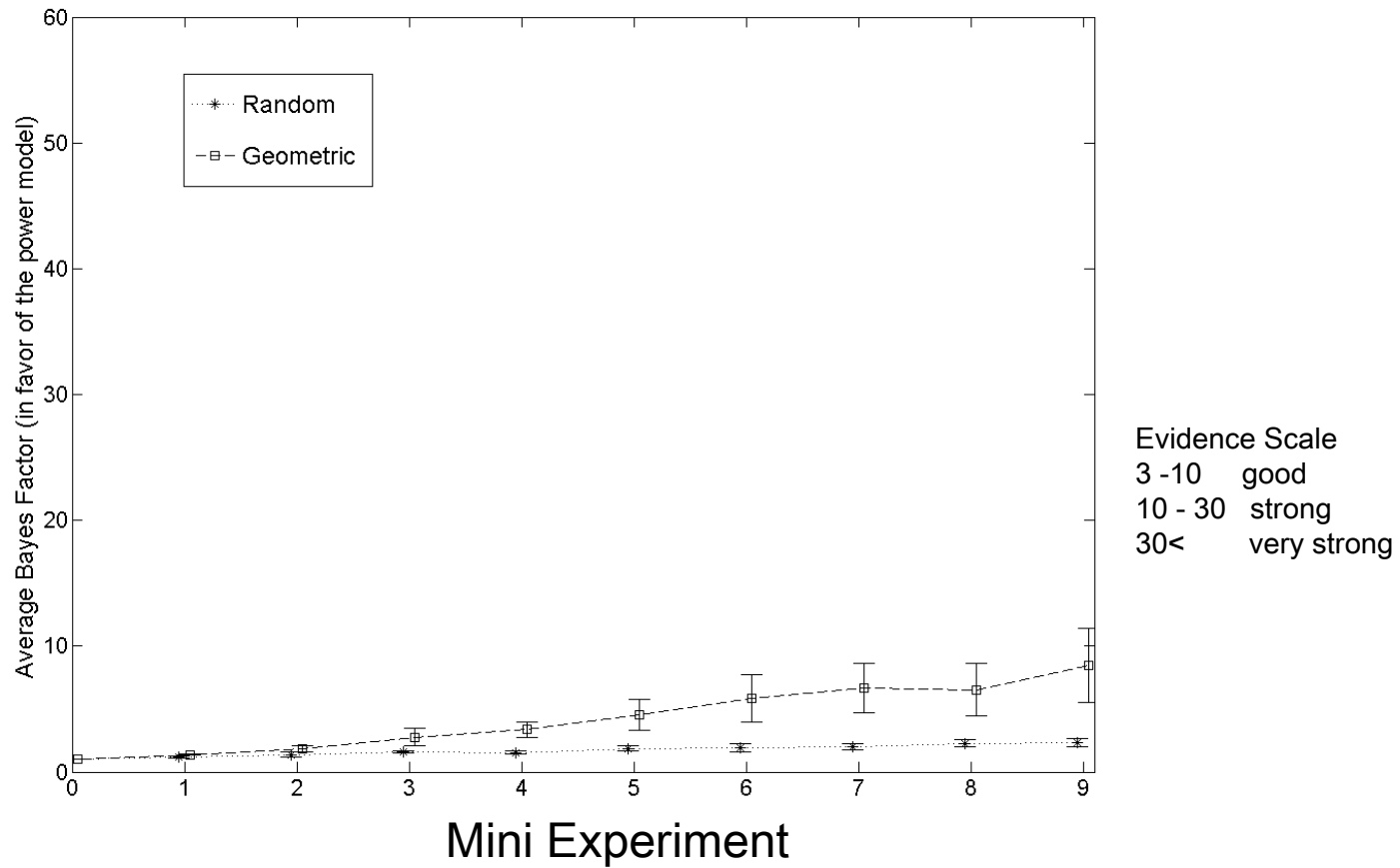
- **Brown-Peterson-type task**
 - Visually presented six monosyllables
 - Retention interval contained a distractor task of reading words aloud
 - Recall words after retention interval
 - 9 mini experiments

Results

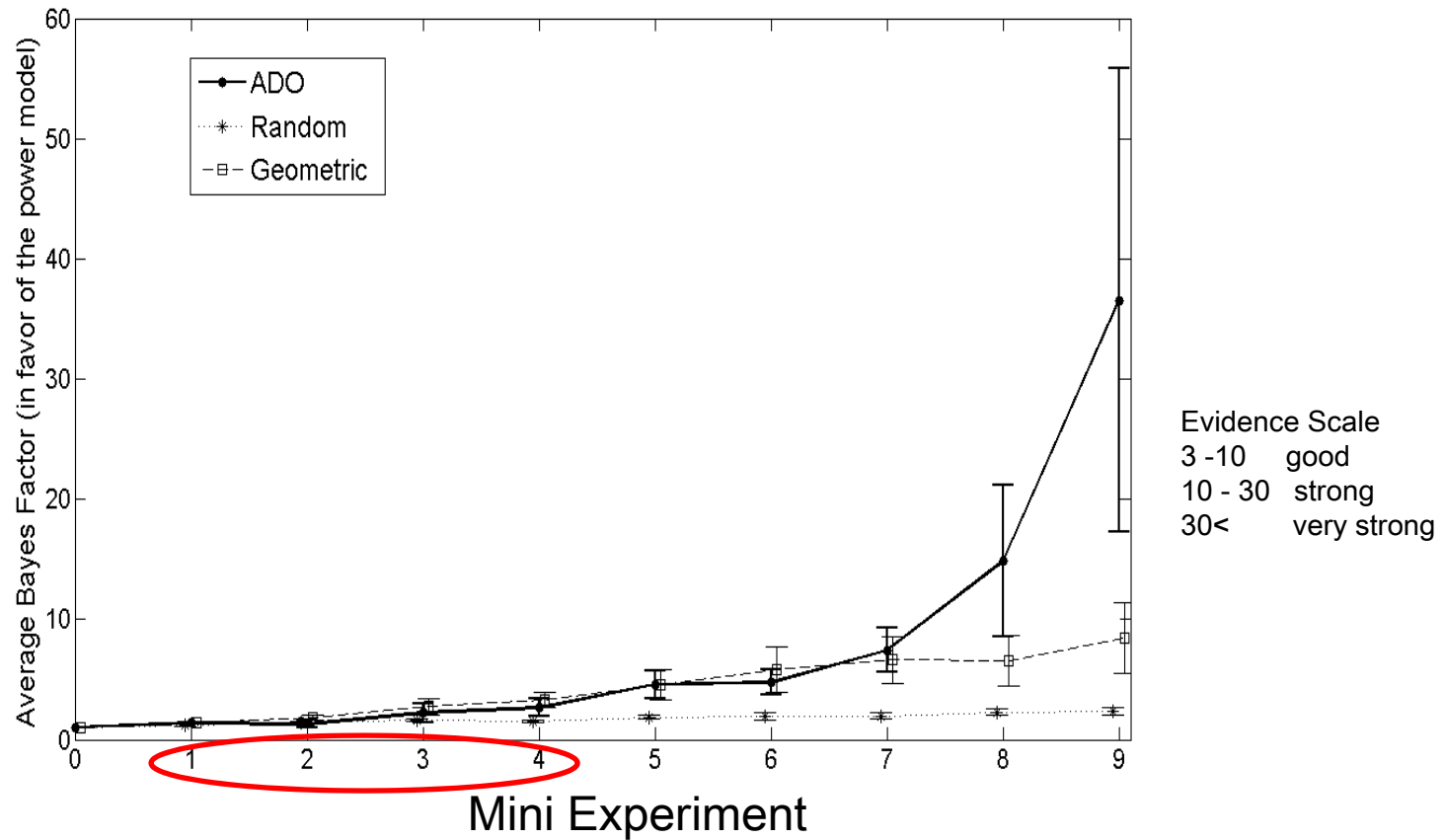
$$BF_{ij} \equiv \frac{p(y | M_i)}{p(y | M_j)} \frac{\text{Power}}{\text{Exponential}}$$



Results

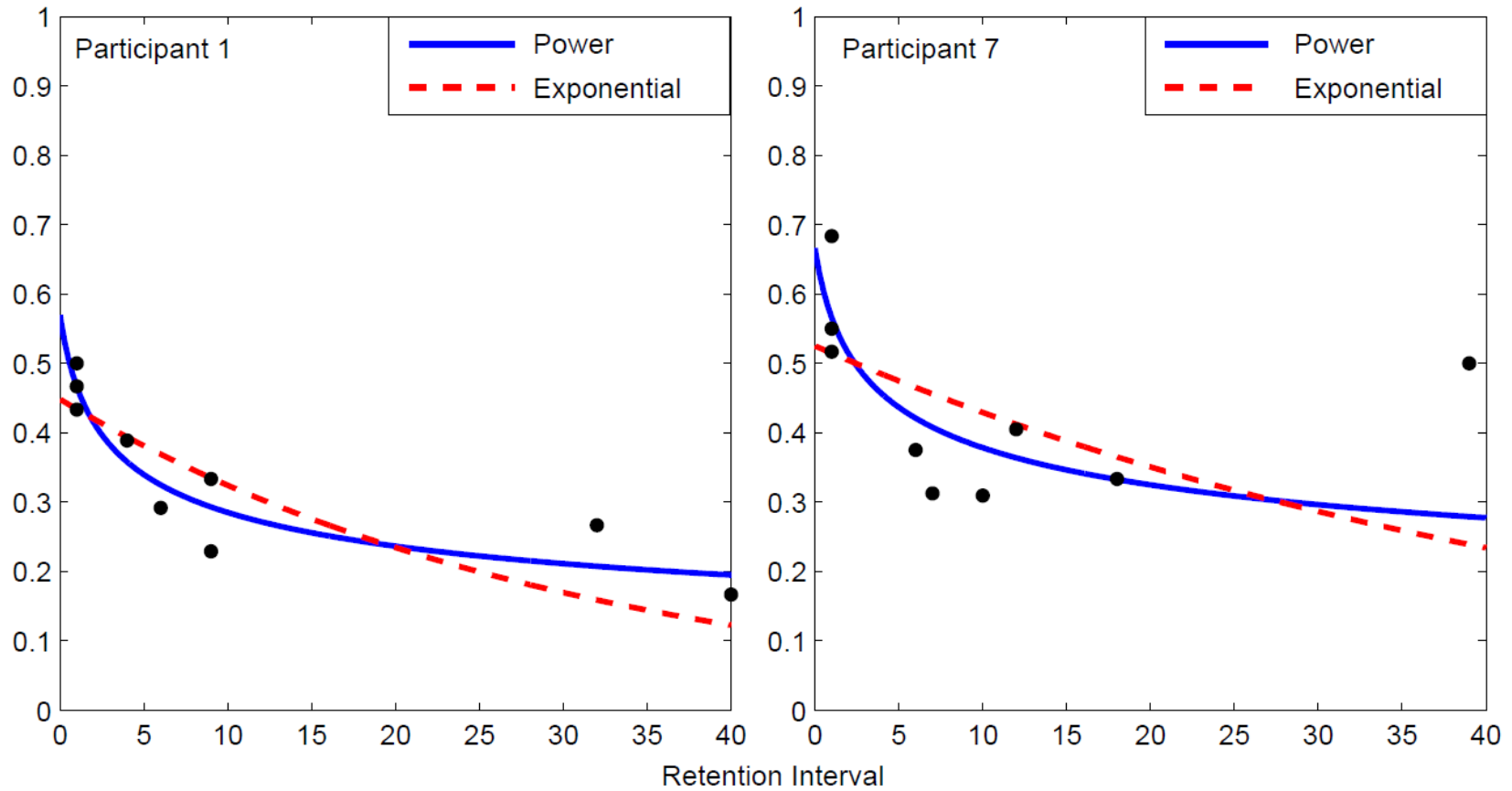


Results



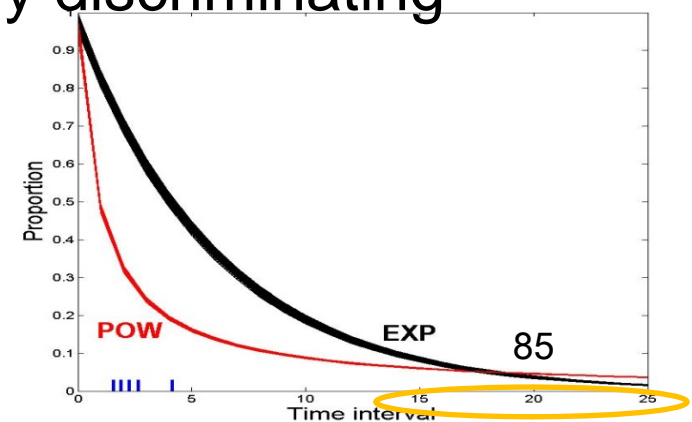
Very strong evidence of the superiority of the power model

ADO Capitalizes on Individual Differences



Summary

- Competing models can be difficult to discriminate because good experimental designs are elusive
- ADO finds and exploits differences between models to maximize the likelihood of discrimination
- Qualifications
 - Not usually a single optimal design, but many
 - Optimal designs are not necessarily discriminating
 - Not all variables can be optimized



5. Final Remarks

- To model behavior, we need to know how models behave
- A model's good fit to a data set is a necessary first step in model evaluation, but not a sufficient, final (confirmatory) step
- To claim that a model deserves credit for good performance (a good fit or simulation) requires understanding **why** the model performed well (e.g., MDL, PSP)
- Design optimization can further improve model discrimination

Further readings on model selection

- Special issues

- Gluck, K., Bello, P., & Busemeyer, J. (2008). Special issue on model comparison. *Cognitive Science*, 32, 1245-1424.
- Myung, I. J., Forster, M., & Browne, M. W. (2000). Special issue on model selection. *Journal of Mathematical Psychology*, 44, 1-231.
- Wagenmakers, E-J., & Waldorp, L. (2006). Special issue on model selection. *Journal of Mathematical Psychology*, 50, 99-213.

- Articles

- Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6, 421-425.
- Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109, 472-491.
- Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248-1284.

Further readings on parameter space partitioning and design optimization

- Parameter space partitioning

- Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, 113, 57-83.

- Design optimization

- Cavagnaro, D. R., Myung, J. I., Pitt, M. A. & Kujala, J. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, 22, 887-905.
- Myung, J. I & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116, 499-518.

The End