
Lecture 10: Computational Cognitive Modeling

Mechanistic vs. Rational models, Bounded Rationality

email address for instructors:
instructors-ccm-spring2018@nyucll.org

course website:
<https://brendenlake.github.io/CCM-site/>

Today: A variety of topics!

- **Mechanism versus Rational modeling** - hopefully a bit of a debate!
- **Boundedly Rational Models** - sampling as a theory of cognition
- **Model fitting** - how to tell if a model provides a good account of some data

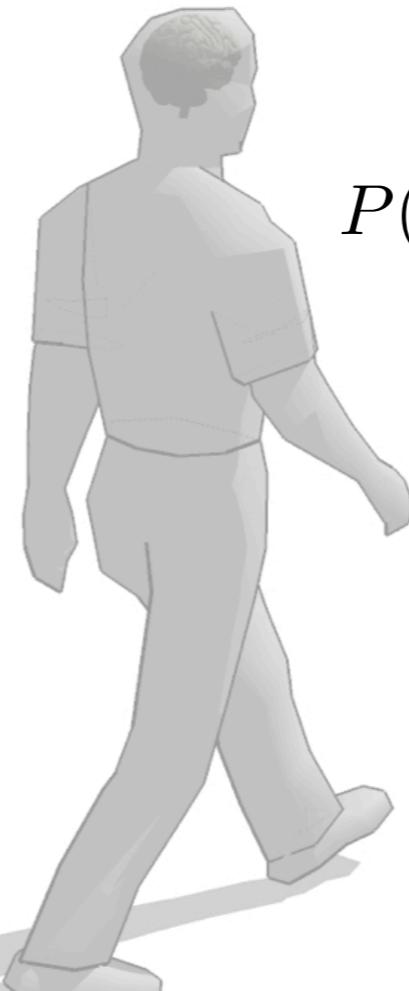
Bayesian modeling is an approach for understanding inductive problems, and it typically takes a strong “top-down” strategy

Three levels of description (*David Marr, 1982*)

Computational

Why do things work the way they do?
What is the goal of the computation?
What are the unifying principles?

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

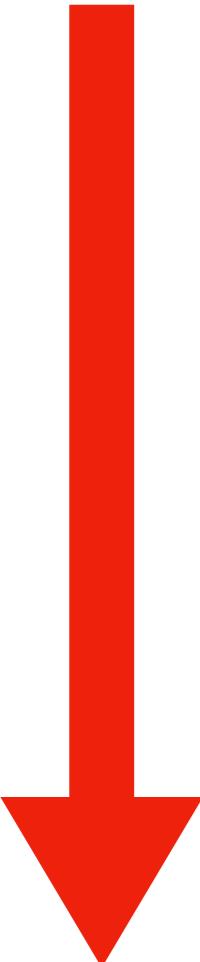


Algorithmic

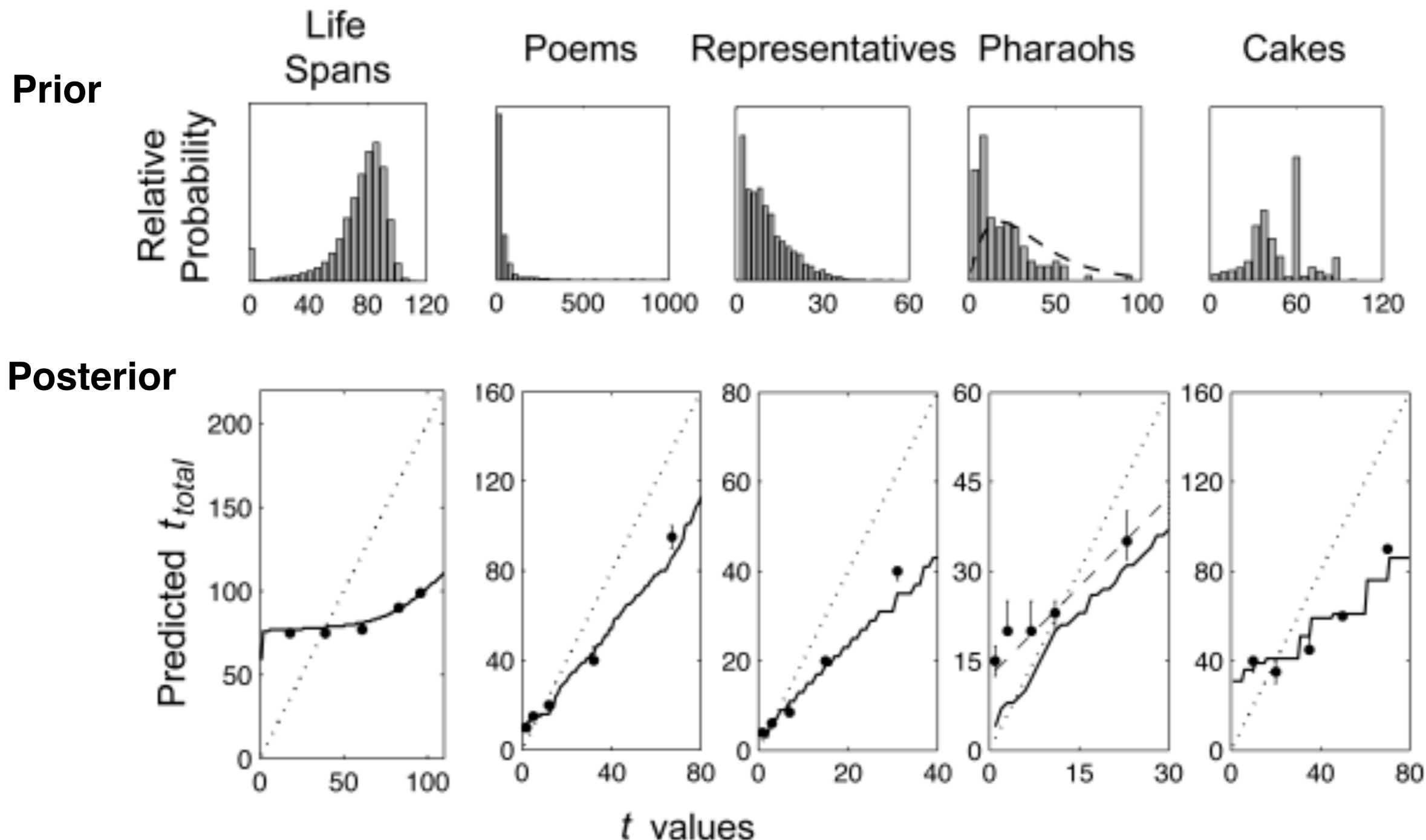
What representations can implement such computations?
How does the choice of representations determine the algorithm?

Implementational

How can such a system be built in hardware?
How can neurons carry out the computations?



What does this mean about cognition?



Black dots are median prediction of participants
Solid lines are optimal Bayesian predictions

Bayesian Models as a Theoretical Framework

- Brought about by mathematical/computational advances in probability and estimation theory
- Generate excitement for several reasons:
 1. Offer a new interpretation of the goal of the cognitive system (behavior is seen as “rational”)
 2. Assumption of Bayesian models are relatively transparent
 3. Can explain very interesting and complex aspects of cognition (e.g., language acquisition or reasoning under uncertainty)

Bayesian Fundamentalism

“Fundamentalist” tenet: human behavior can be explained through rational analysis, given the correct probabilistic interpretation of the task environment. No need to look at mechanism.

Bayesian enlightenment

Elements of a Bayesian model treated as claims about a psychological process and representation, not mathematical conveniences.

Agnostic Bayes

Bayesian inferential methods used to decide among scientific models based on empirical data.

Danger of mistaking metaphors for theories: Connectionism

- Initial excitement (once backpropogation was developed) Backpropogation allowed the training of multi-layer models which could
 - (in theory) encode any function.
 - Reaction to the computer metaphor
 - Avoided the brittleness of propositional models
 - Many studies were done applying neural nets to reproduce a wide gamut of Psychological phenomena.
- Bubble bursts
 - Incapable of capturing compositionality characteristic of language processing
 - Too opaque to offer insight
 - Biologically implausible
- Field was not able to cope with new criticisms

Summary of Bayesian inference

- Two steps:

- ① Specify the hypothesis space. A hypothesis is simply a likelihood function over patterns of observations.
- ② Specify the initial (prior) belief over the various hypotheses.

- Posterior:

$$E_{posterior}(H) = E_{prior}(H) + \sum_i E_{data_i}(H) \quad (1)$$

- Nota bene:

- Hypotheses are simply probability distributions; no relation to Psychological processes.
- Combining prior with observation (the Bayesian step) is formally equivalent to vote counting.
- Elaborate math tends to derive from the difficulty of approximating a solution with a complicated hypothesis space.

Consequences of the denial of mechanism

- Problematic to toss out everything below computational level
 - Mechanistic explanations often make more useful predictions than ad hoc explanations.
 - Optimal decision making is not always possible: how does the brain deal with this problem? (stay tuned!)
 - E.g., mechanistic (connectionist) accounts of semantic memory have inspired useful models of brain damage and even treatments.
 - Manipulations unrelated to rational form, such as making information part of a cue or an outcome within a learning trial can yield surprising results. Or manipulating the order of items in learning strongly influences what is learned by Bayes is “exchangable”
- Also, non-Bayesian applications of cognitive models tend to rely on mechanistic descriptions: e.g., neurology or psychopharmacology

Bayes and Evolutionary Theory

- Animal behavior is adapted by natural selection to increase fitness
 - Thus an animal's behavior should reflect whatever was adaptive in its ancestral habitat.
 - Focus on environment can sometimes come at the cost of mechanism.
 - Not any function is optimized; only those relevant to fitness (more constrained than Bayesian fundamentalism).
 - Proliferation of modules; e.g., snake avoidance module.
- Like the modules in evolutionary Biology, Bayesian analyses are often ad hoc solutions for a given behavior.
 - E.g., children's ability to give the number of items requested.
 - This comes at the expense of identifying general mechanisms and architectural characteristics (e.g., working memory) that are applicable across tasks.
- Bayesian models are in a sense less constrained than evolutionary theories, which are restricted to functions that are relevant to fitness.

Remember the dutch book from last week!

Agent 1 Proposition	Belief	Agent 2		Outcome for Agent 1			
		Bet	Stakes	$a \wedge b$	$a \wedge \neg b$	$\neg a \wedge b$	$\neg a \wedge \neg b$
a	0.4	a	4 to 6	-6	-6	4	4
b	0.3	b	3 to 7	-7	3	-7	3
$a \vee b$	0.8	$\neg(a \vee b)$	2 to 8	2	2	2	-8
				-11	-1	-1	-1

Figure 13.2 Because Agent 1 has inconsistent beliefs, Agent 2 is able to devise a set of bets that guarantees a loss for Agent 1, no matter what the outcome of a and b .

Bayes and Evolutionary Theory

- “Just so” stories; Bayesian fundamentalism labels a behavior as “rational” simply because a probabilistic model can fit it.
 - Both kinds of explanations can predict any outcome post hoc by changing the formulation of the problem.
- Both approaches deemphasize mechanisms to their detriment. constraints on their assumptions.
 - “Completely sidestepping mechanistic considerations when considering optimality leads to absurd conclusions. To illustrate, it may not be optimal or evolutionarily advantageous to ever age, become infertile, and die, but these outcomes are universal and follow from biological constraints.”
 - “Evolution is survival of the best current design, not survival of the globally optimal design.” (Marcus, 2008)

An example - Cognitive Development

- A key question of development: what develops?
 - **In rational models, nothing develops, only more information is accumulated.**
 - Some aspects of development clearly reflect maturational changes in mechanism.
 - Some measures of child development are indexed to prefrontal development
- A fundamentalist Bayesian approach:
 - Posit rational theories that are collections of discrepant causal models (hypothesis spaces)
 - Each model is intended to correspond to a different stage This does not explain how or why the transitions occur
- A satisfying approach needs to consider how selection mechanisms change throughout development, or how the mechanisms of cognition shift.

Bayesian Enlightenment!

- In a Psychological model, the prior and hypothesis space should be beliefs of the subject.
 - Information in brain \neq ground truth.
 - Accuracy of information encoded can be separated from whether the participant acts on that information rationally.
 - Constructing the generative model can be the most substantial part of learning.
- Priors could be another area of inquiry; conjugate priors might be easier for the brain to work with.
- **Approximation algorithms could also be taken seriously as mechanistic models.**
- An appeal of Bayesian models is the transparent view of their operation (as opposed to, say, connectionism); but design assumptions made available are rarely viewed critically in practice.

Bayesian inference is hard

Rich, realistic behavior

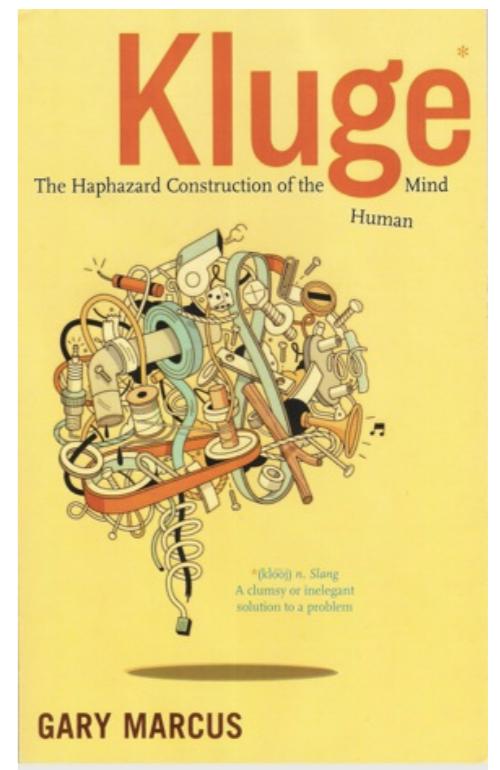
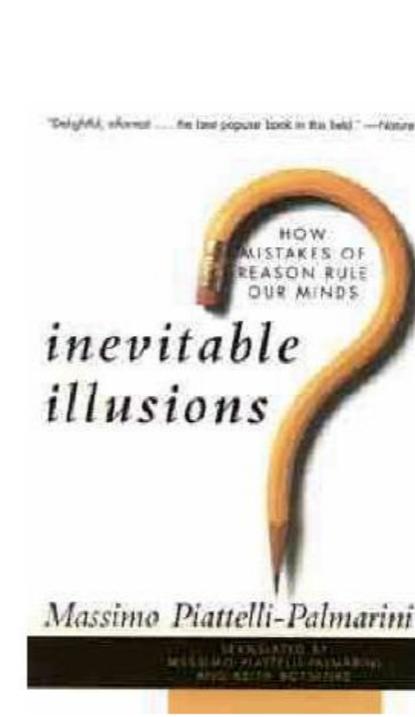
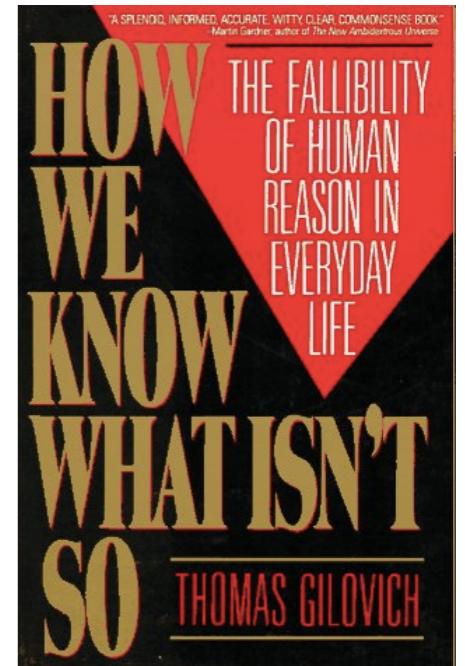
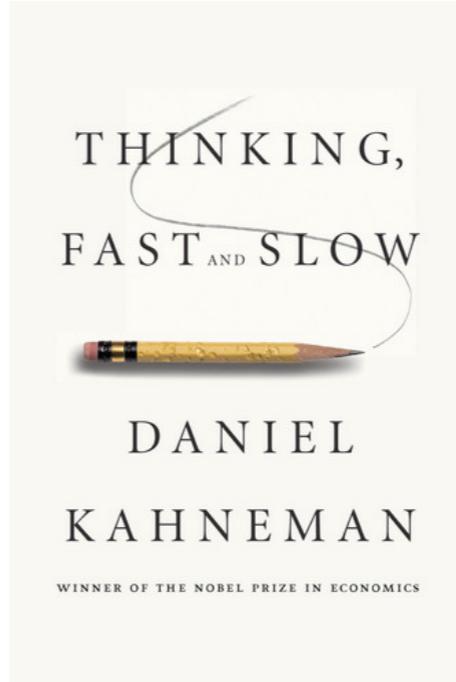


The goal: A *predictive* account of behavior that can *emulate* both the general successes and the occasional failures of human performance, and predict which will happen when.

Why *predict*? Why *emulate*? Why emulate *failures*?

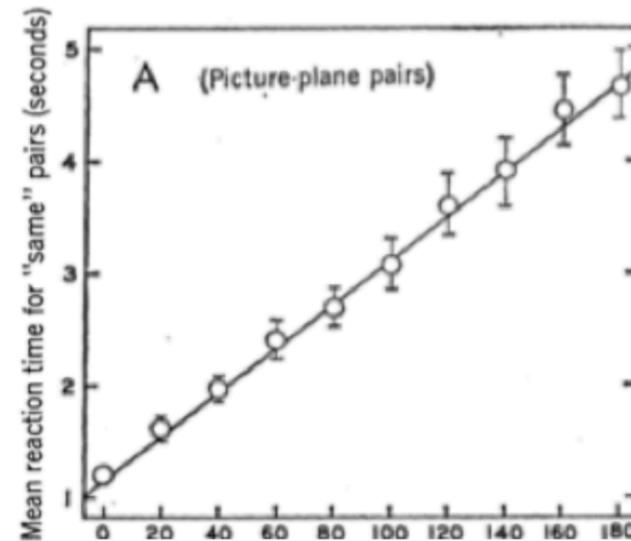
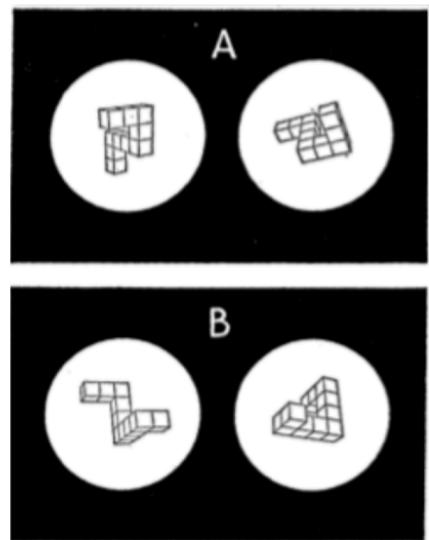
- In realistic problems, the number of possible hypotheses can be huge
 - e.g., more than 100,000 clusterings of 10 objects
- In the worst case, the time required to perform exact Bayesian inference increases linearly in the number of hypotheses

Rationality vs. Heuristics



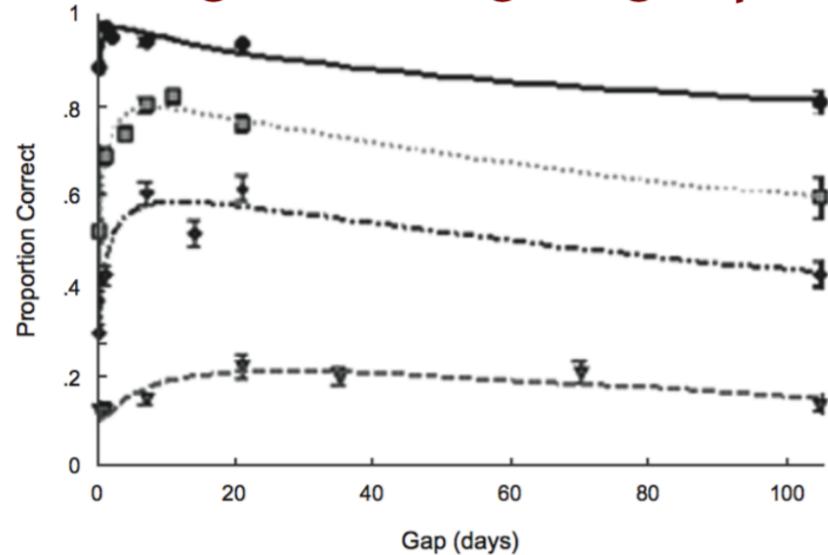
Resource constraints on cognition

Thinking takes time



Shepard & Metzler (1971)

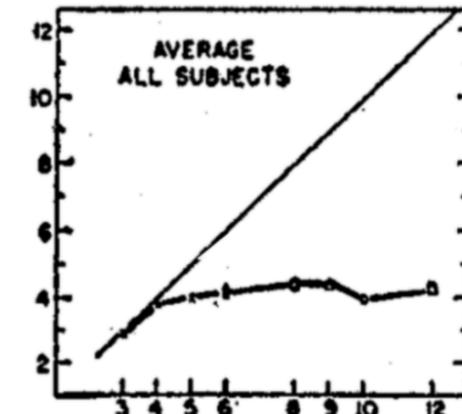
Learning is slow, forgetting is quick



Cepeda, Vul, Rohrer, Wixted, Pahler, 2008

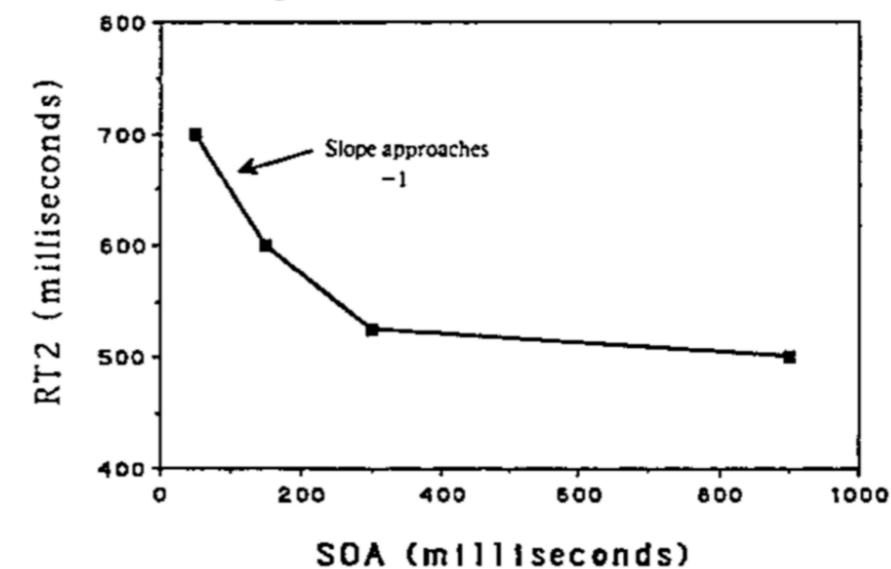
Memory stores are limited

T D R
S R N
F Z R



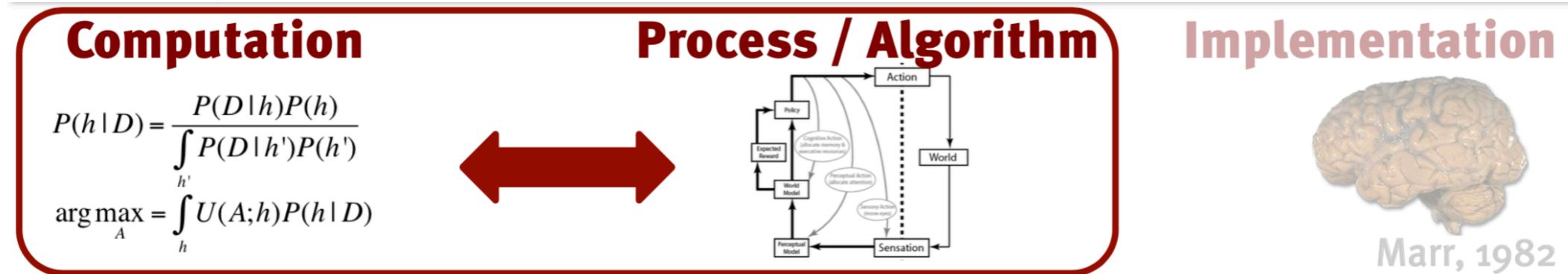
Sperling (1960)

Thinking has a central bottleneck



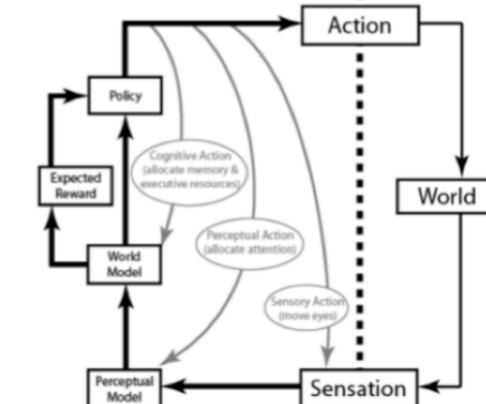
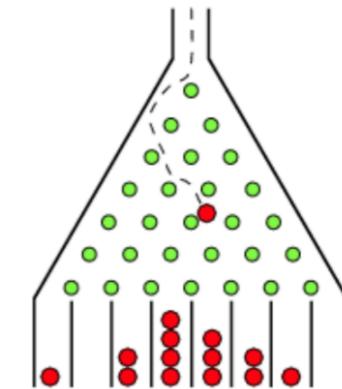
Pashler (1994)

Computation-algorithm intersection



Challenges at the intersection of probabilistic reasoning and cognitive resource constraints:

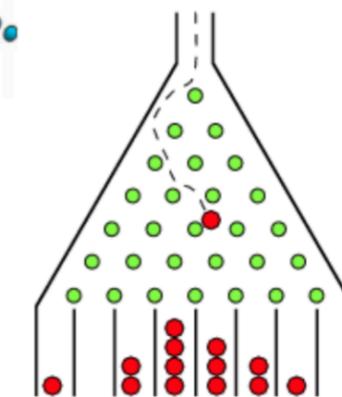
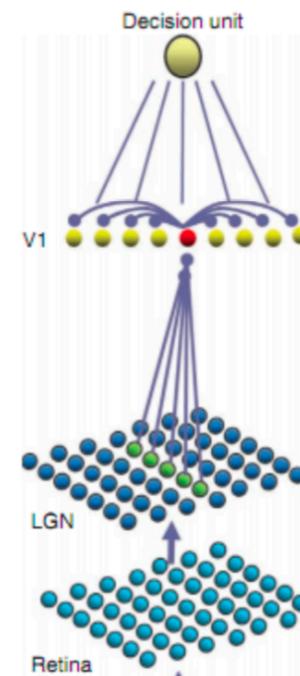
- How does the mind represent uncertainty?
- How do we use limited cognitive resources?



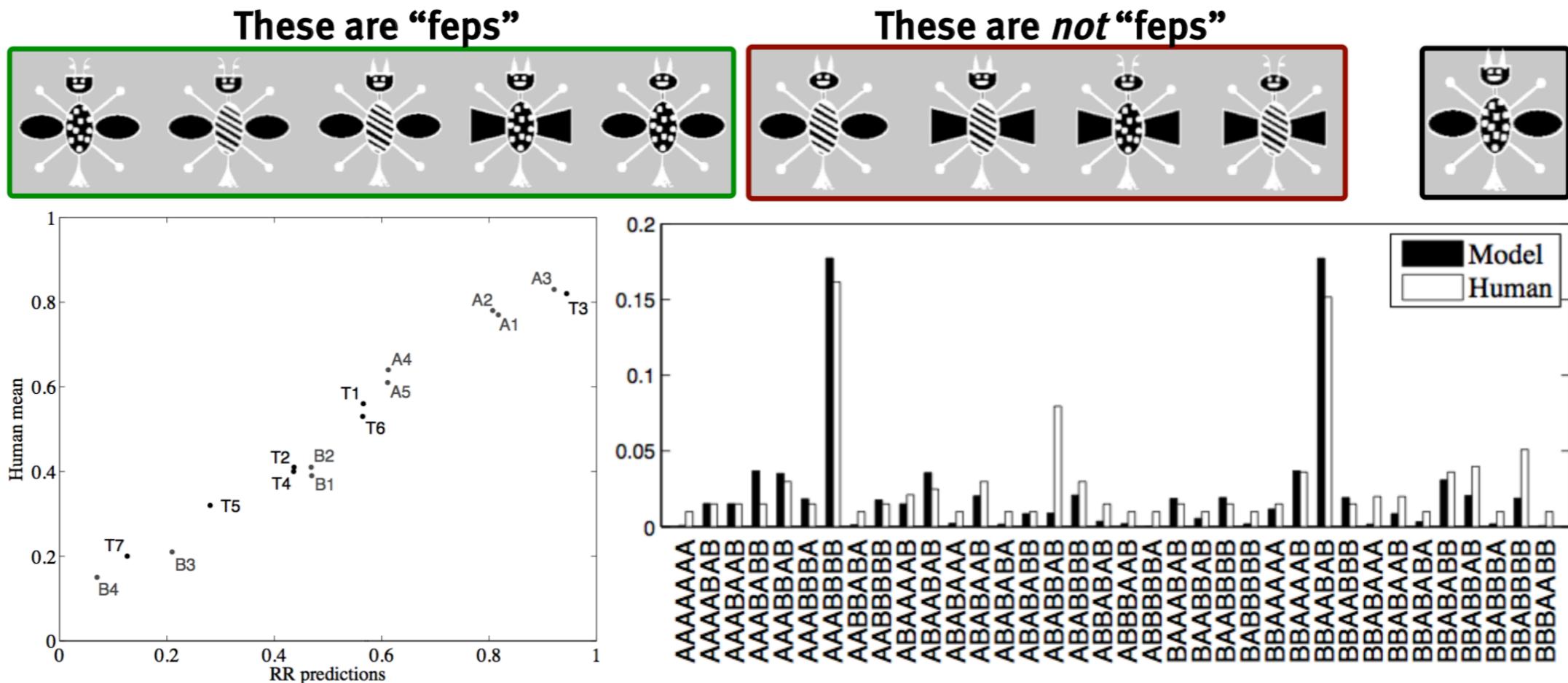
Computation-algorithm intersection

- **Analytical**
i.e., equations, including variational methods
(Friston)
- **Tabular / grids**
e.g., probabilistic population codes
(Pouget/Ma/Beck)
- **Sampling**
(Goodman/Griffiths/Sanborn/Tenenbaum/Vul, etc.)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



Sampling categorization rules

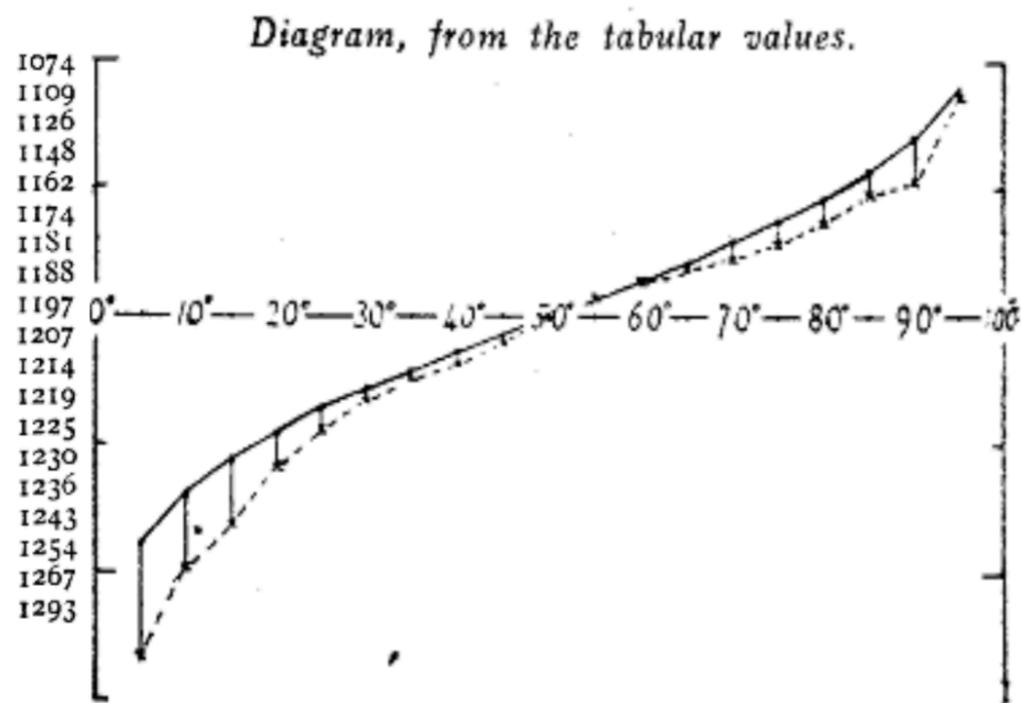


Generalization patterns suggest that each individual was not using a probability distribution over rules, but just one sampled rule.

Goodman, Tenenbaum, Feldman, & Griffiths, 2008

Wisdom of Crowds

Galton, 1907: Vox Populi
How much does an ox weigh?



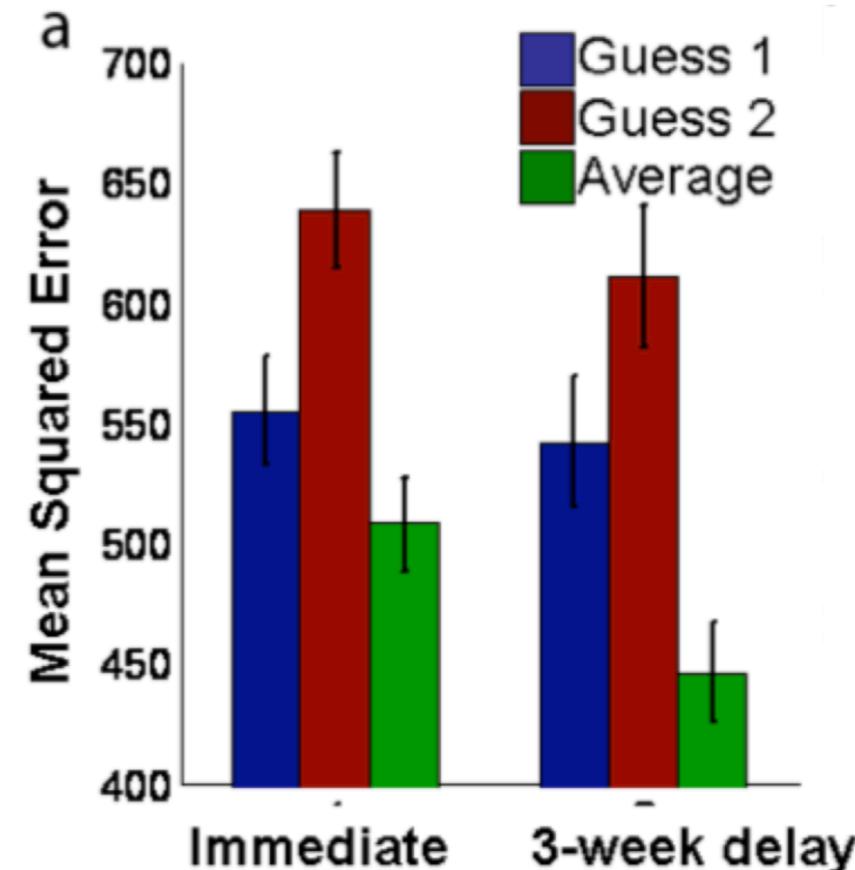
Mean was (1206) was closer to the correct answer (1207) than any one guess.

Benefit of averaging multiple guesses holds so long as errors are independent samples.

Do we get the same effect *within* individuals?

Wisdom of Crowd Within

- What percent of the world's airports are in the United States?
- Saudi Arabia consumes what percentage of the oil it produces?
- What percentage of the world's countries have a higher life expectancy than the United States?



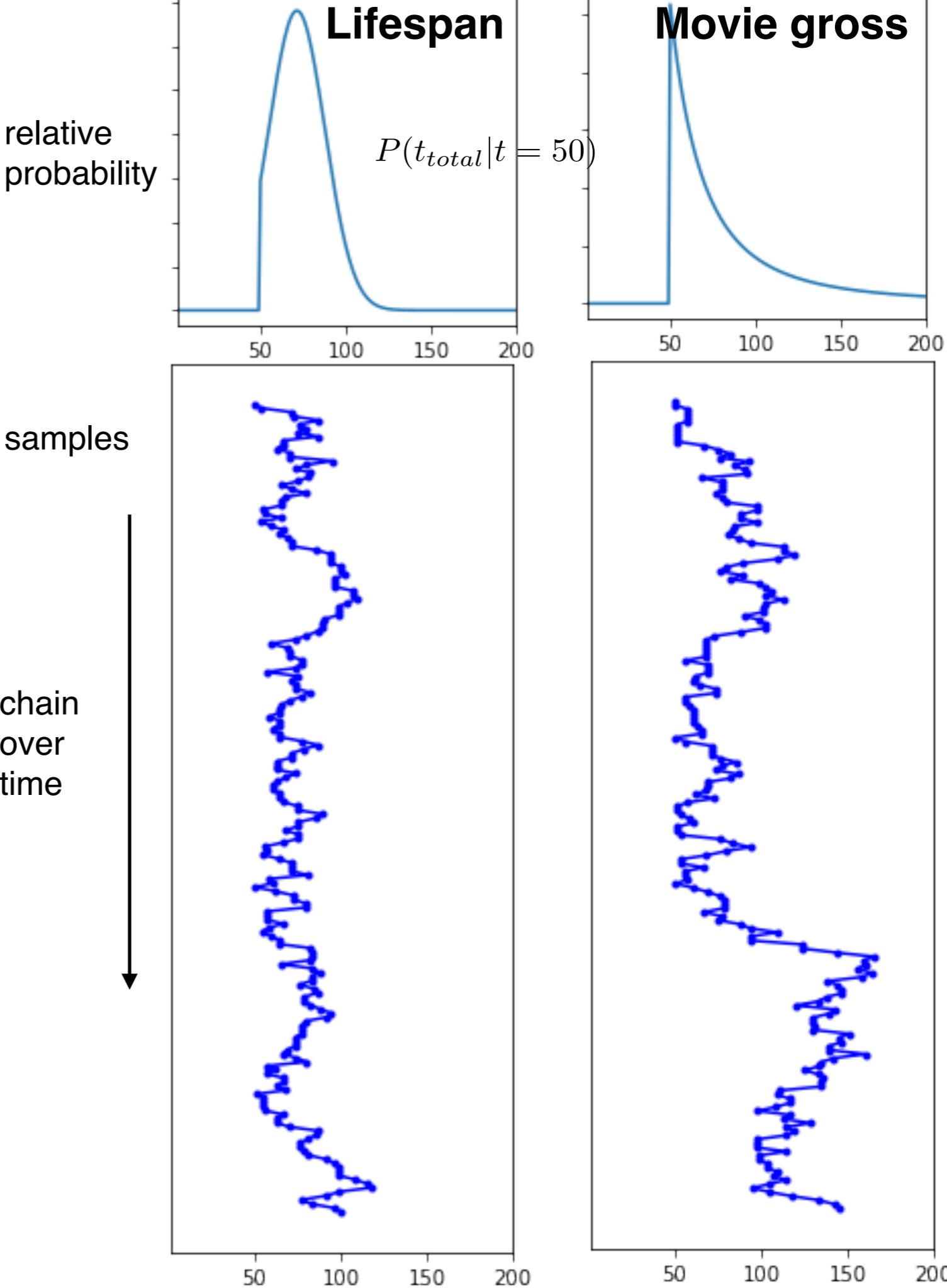
Benefit of averaging multiple guesses from a single individual:
Estimation errors do not arise only from individual biases,
but reflect *sampling* under uncertainty.

The Monte Carlo principle

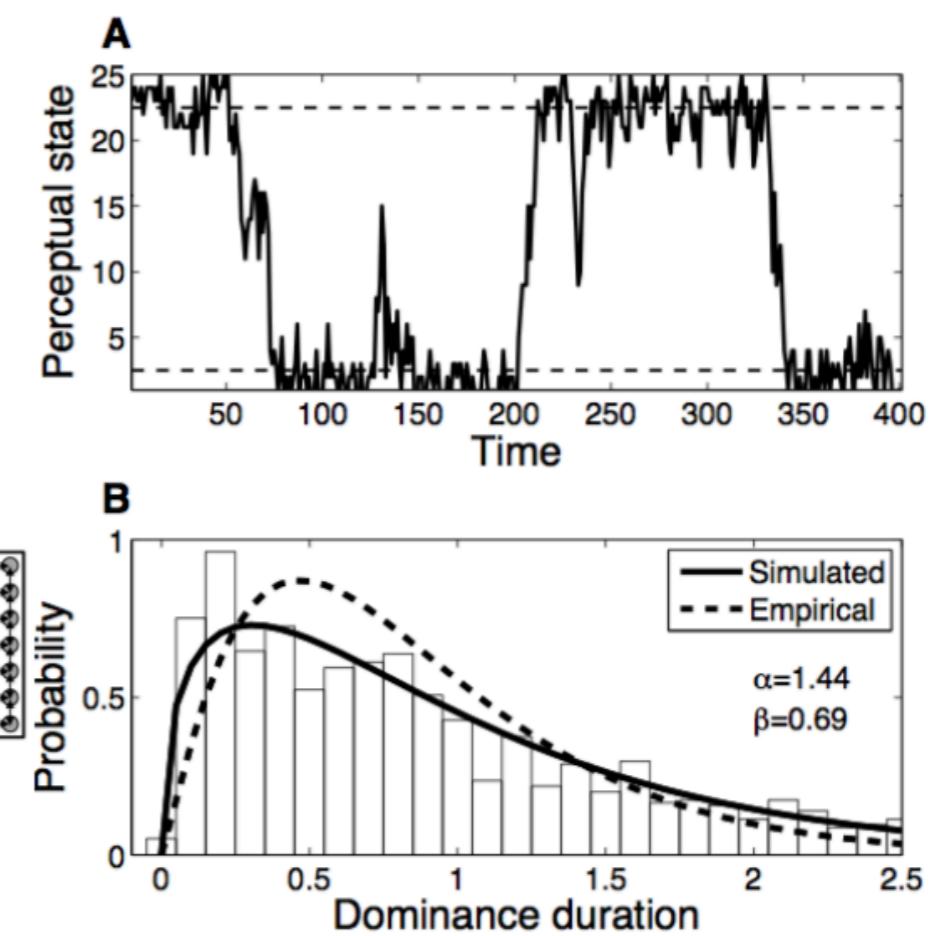
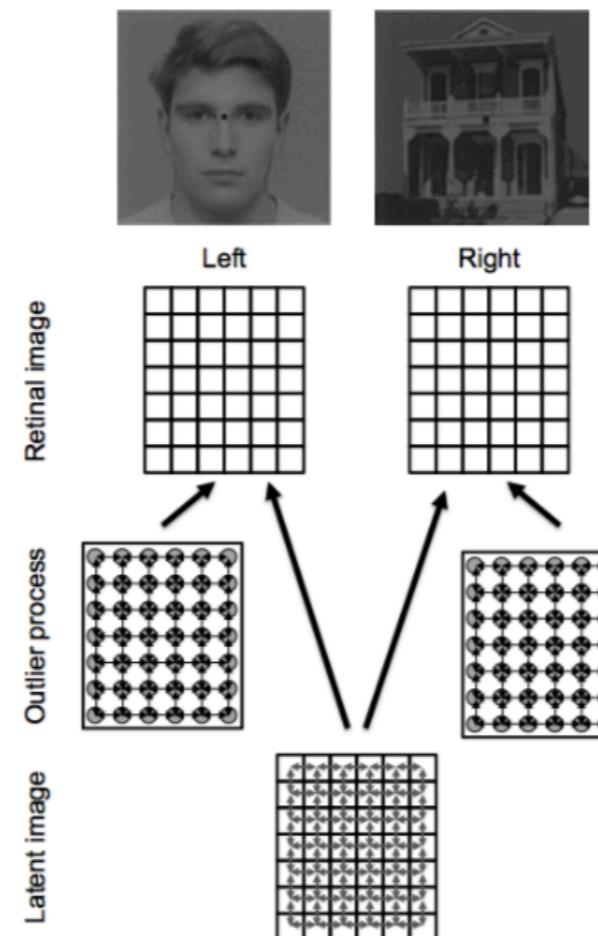
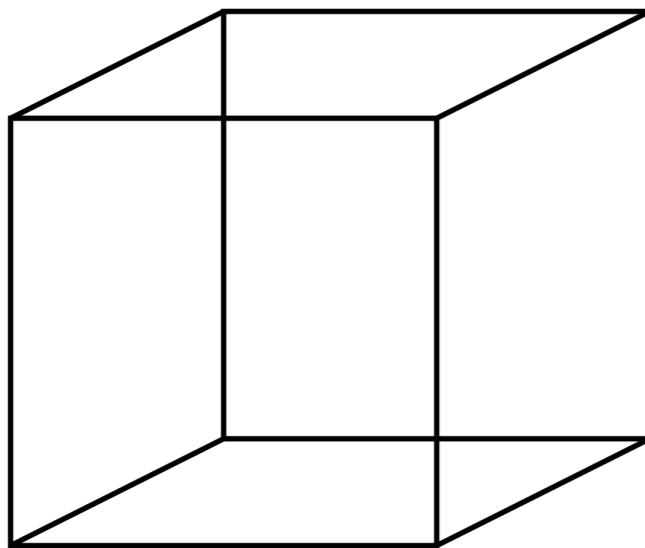
The expectation of f with respect to P can be approximated by

$$E_{P(x)}[f(x)] \approx \frac{1}{n} \sum_{i=1}^n f(x_i)$$

where the x_i are sampled from $P(x)$



The Monte Carlo principle



Gershman, Vul, & Tenenbaum; 2011

How many samples?

$$A^* = \arg \max_A \sum_S U(A; S)P(S|D).$$

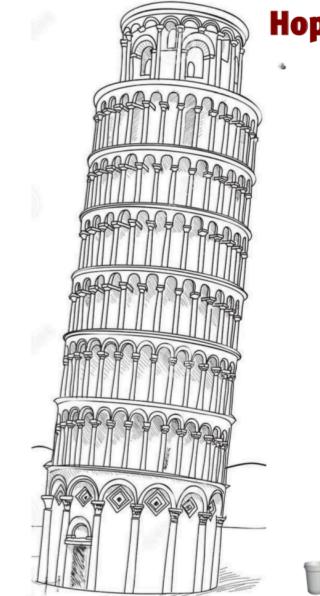
Trivial – no benefit from precision



Just right – aim carefully.



Hopeless – don't bother.



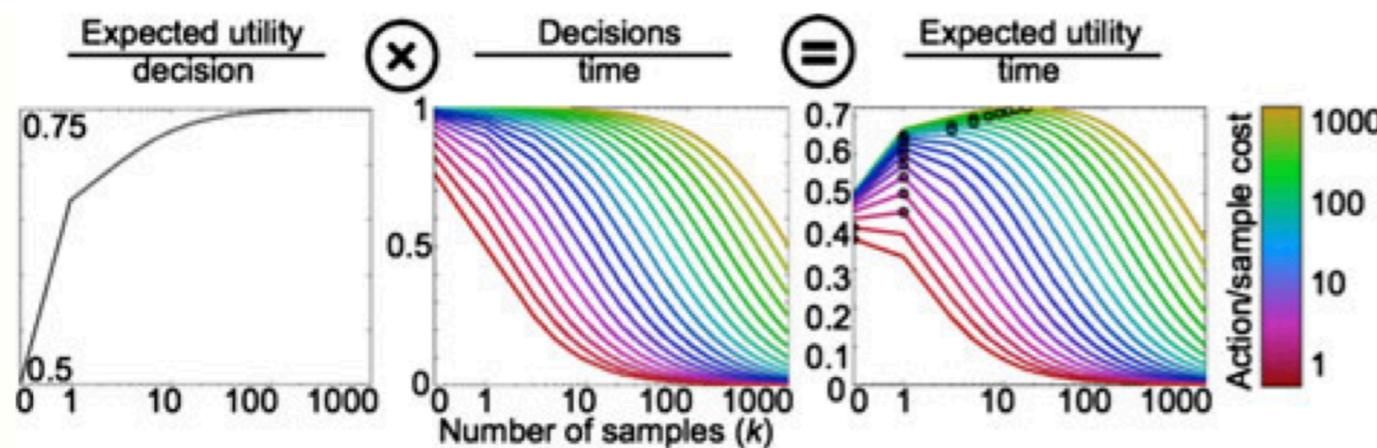
Uncertainty

Reward

Uncertainty

Reward

What does adaptive behavior look like if there is a cost (at least of time) to drawing new samples (computation)?



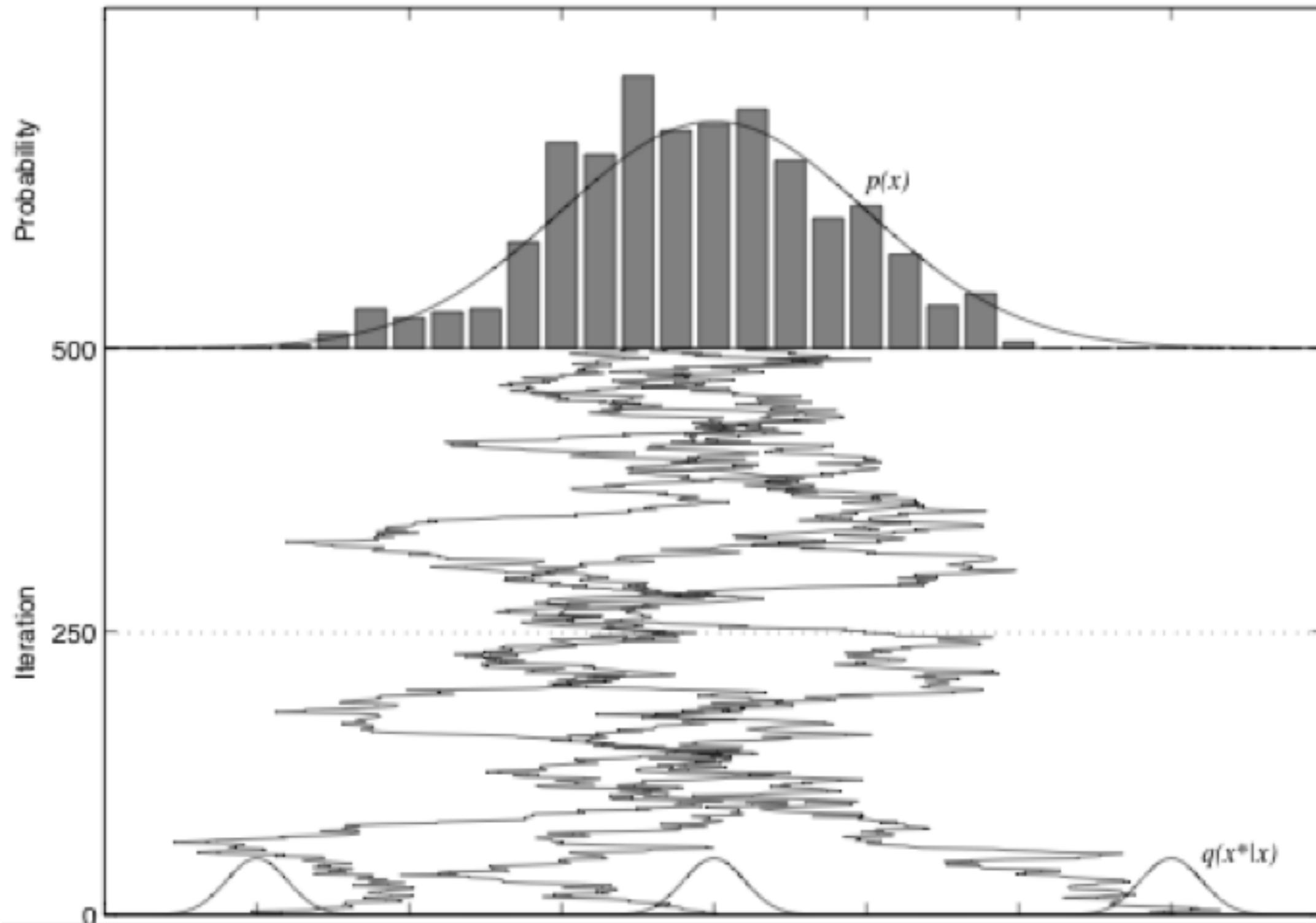
**cost of computation
= cost of thinking**



Anchoring and adjustment

- Is the population of Chicago greater or less than 200,000 people?
- Now guess the population of Chicago
- People give lower estimates when given a lower anchor (200,000) than a higher anchor (5 million) (Jacowitz & Kahneman, 2005)

Metropolis-Hastings



Sampling as mechanism

- Each sampling solution yields different samples:
Variability across decisions;
Variability across trials;
Variability across participants
- Solutions outside of the “convergent” regime:
Systematic deviations from optimal decisions that will reflect the biases of the sampling algorithm.

Biases

- Probability matching
Vul et al. 2009
- Anchoring and adjustment
Lieder et al. 2013
- Sequential effects
Sanborn et al. 2010
- Garden-pathing
Levy et al. 2009
- Dynamics of belief change
Gershman et al. 2012
- Memory reproduction
Shi et al. 2010

We now face tradeoffs between the effort of computation, variability, and bias of answers. These tradeoffs yield tricky meta-cognitive decisions.

Three levels of description (*David Marr, 1982*)

Computational

Why do things work the way they do?
What is the goal of the computation?
What are the unifying principles?



Algorithmic

What representations can implement such computations?
How does the choice of representations determine the algorithm?

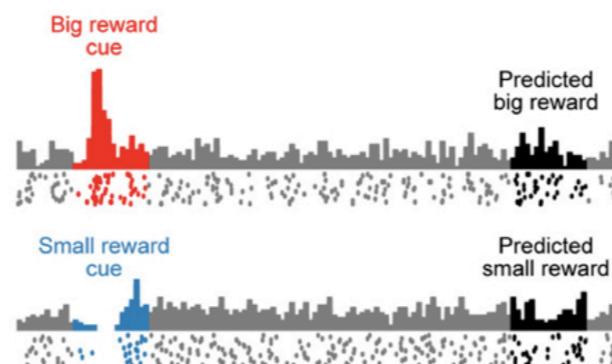
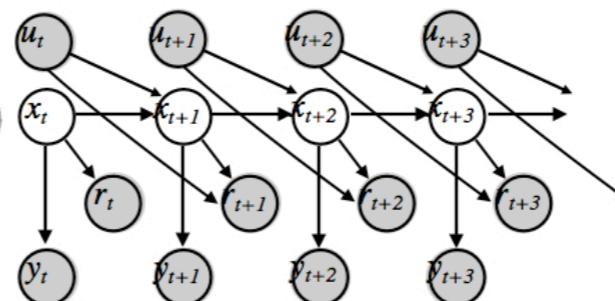
Implementational

How can such a system be built in hardware?
How can neurons carry out the computations?

maximize:

$$R_t = r_{t+1} + r_{t+2} + \dots + r_T$$

Bellman



Dynamic programming,
TD methods, Monte
Carlo

Neural firing patterns,
prediction errors,
system level
neuroscience

Slide Credits

Tom Griffiths (bounded rationality)
Ed Vul (bounded rationality)
John McDonnell (Love/Jones paper)

What makes a good model?

Qualitative Criteria

- Explanatory adequacy
 - The assumptions of the model are plausible and consistent with other findings (assumption are not ad-hoc)
- The model does more than just re-describe the data
 - “curve fitting” versus theoretical principals
 - e.g., “power law of practice” versus “instance theory of automaticity”
- Interpretability
 - The model makes sense
 - The model components and parameters link to psychological or neural processes and constructs

Qualitative Criteria

- Faithful
 - model depends on “core” theoretical features, not implementation details
 - e.g., neural network shouldn’t depend on relative number of “hidden units” unless a strong commitment is being made
- Parsimony?
 - Occam’s razor?

Quantitative Criteria

- Goodness-of-fit
 - SSE, $\log(L)$, RMSE, etc...
 - A good fit merely qualifies the model as one of the candidate models for further consideration... necessary but not sufficient.

Quantitative Criteria

Table I. Results of a model recovery simulation in which a GOF measure (RMSE) was used to discriminate models when the source of the error was varied.

Condition (sources of variation)	Model the data were generated from			Model fitted	
	M_A $a = 0.4$	M_A $a = 0.6$	M_B	M_A	M_B
(1) Sampling error	100	–	–	0.040 (0%)	0.029 (100%)
(2) Sampling error + individual differences	50	50	–	0.041 (0%)	0.029 (100%)
(3) Different models	–	50	50	0.075 (0%)	0.029 (100%)
(4) Sampling error	–	–	100	0.079 (0%)	0.029 (100%)

The severity of the problem is shown in Table I, which contains the results of a model recovery simulation using RMSE. Four datasets were generated from a combination of the two models (M_A and M_B), defined as follows:
 $M_A: y = (1+t)^{-a}$, $M_B: y = (b+ct)^{-a}$ where $a, b, c > 0$. Datasets

Quantitative Criteria

- Goodness-of-fit
 - SSE, $\log(L)$, RMSE, etc...
 - A good fit merely qualifies the model as one of the candidate models for further consideration... necessary but not sufficient.
- Parsimony?
 - The “simplest” model that does not fit significantly worse than possibly more complex models
 - via Hierarchical model testing, G^2 statistics, etc...
- Generalizability
 - model can account for new unseen data

Quantitative Criteria

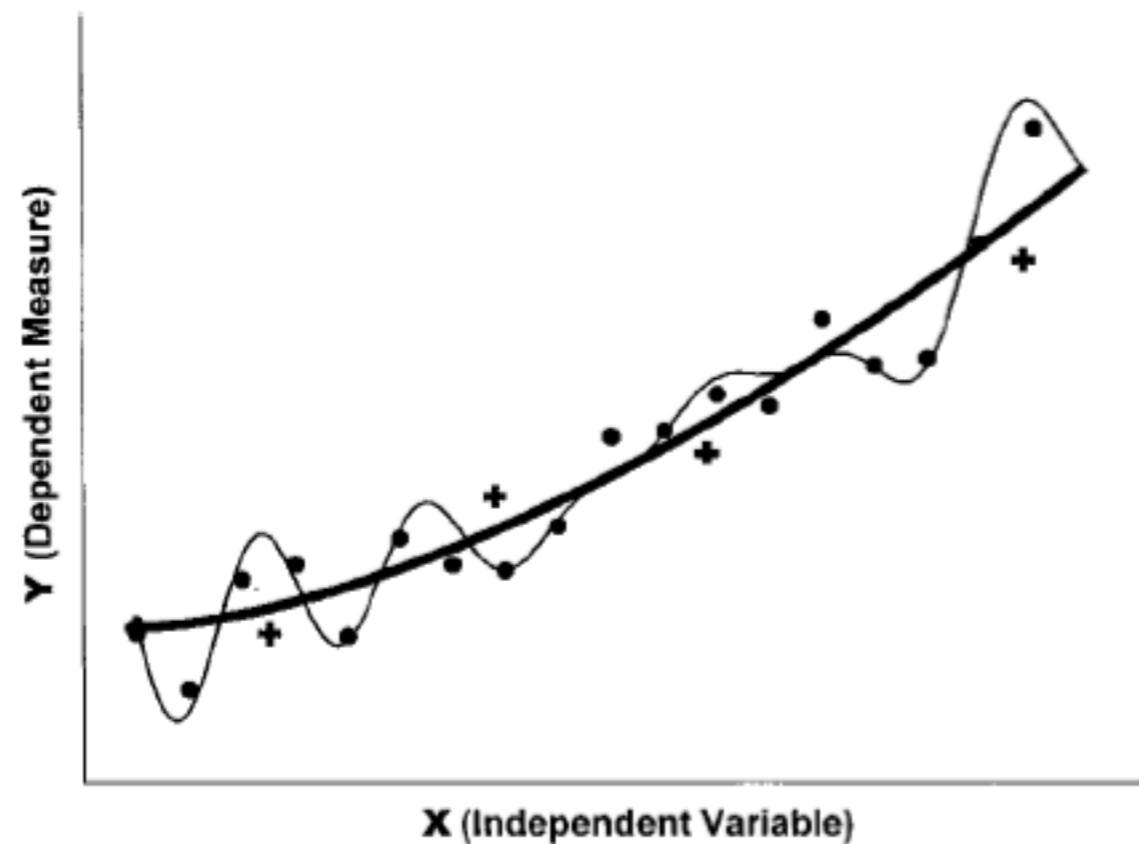


Figure 1. Illustration of the trade-off between goodness of fit and generalizability. An observed data set (dots) was fitted to a simple model (thick line) and a complex model (thin line). New observations are shown by the plus symbol.

Quantitative Criteria

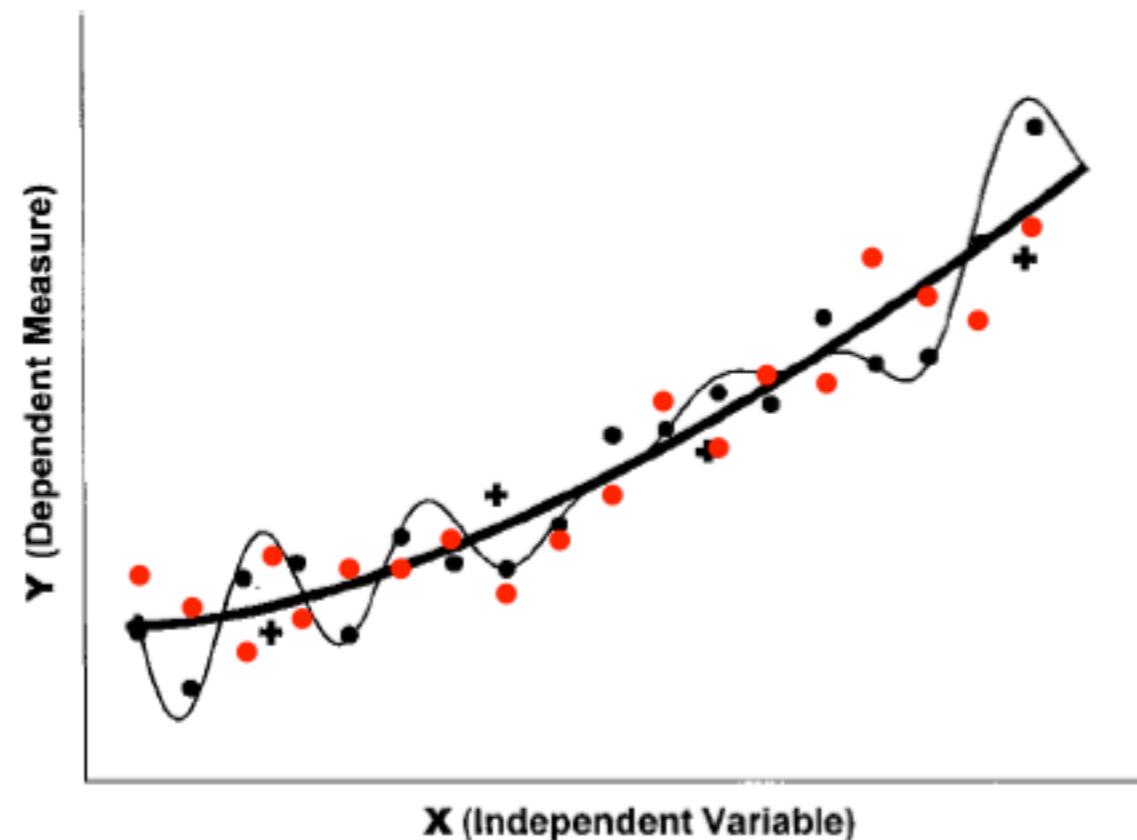


Figure 1. Illustration of the trade-off between goodness of fit and generalizability. An observed data set (dots) was fitted to a simple model (thick line) and a complex model (thin line). New observations are shown by the plus symbol.

Quantitative Criteria

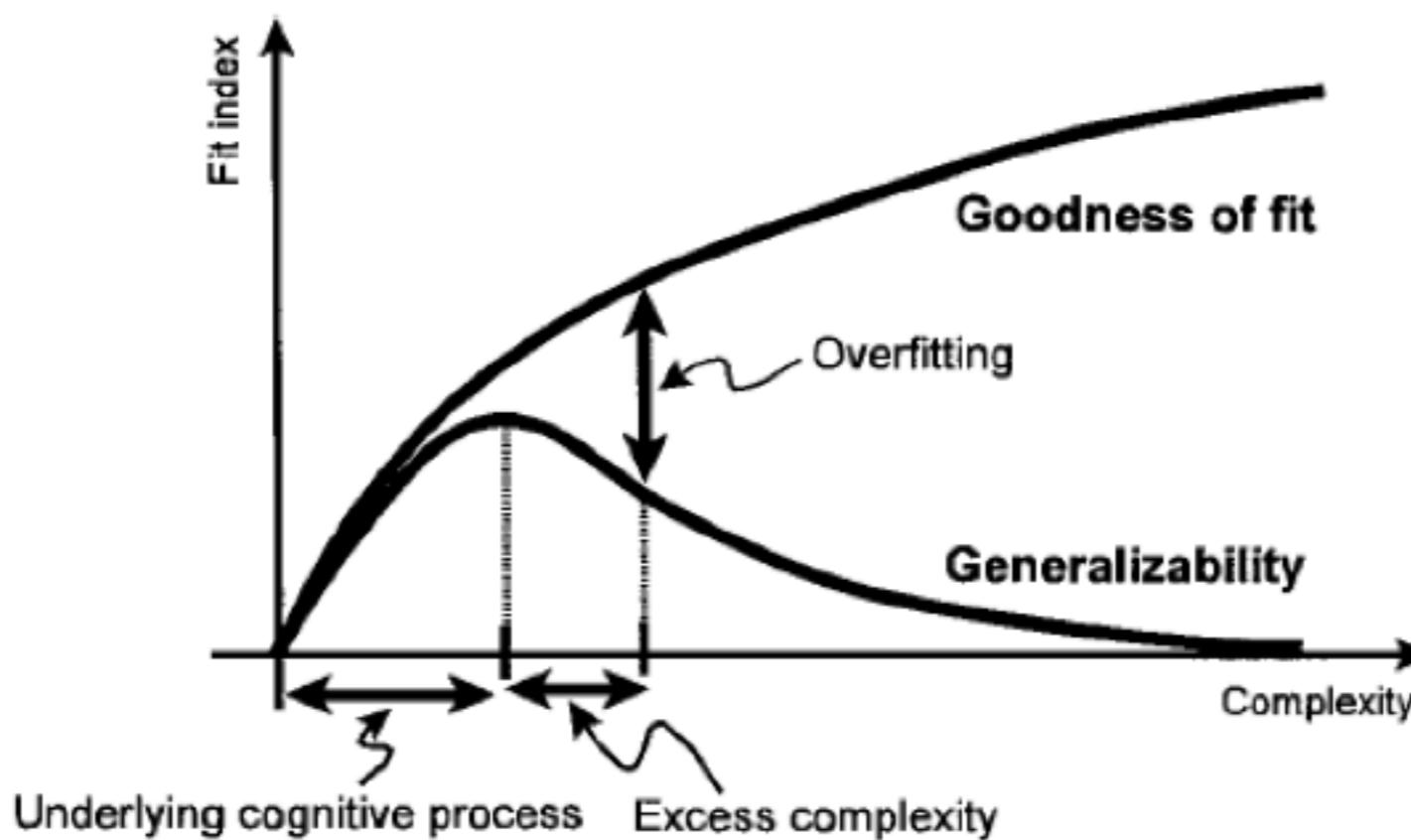


Figure 2. Illustration of the relationship between goodness of fit and generalizability as a function of model complexity (Myung & Pitt, 2001). From *Stevens' Handbook of Experimental Psychology* (p. 449, Figure 11.4), by J. Wixted (Editor), 2001, New York: Wiley. Copyright 2001 by Wiley. Adapted with permission.

Quantitative Criteria

- Complexity
 - Inherent flexibility in model
 - AIC, BIC - # of free parameters (basically goodness of fit+ penalty)
 - insensitive to the “functional form” of the model

Quantitative Criteria

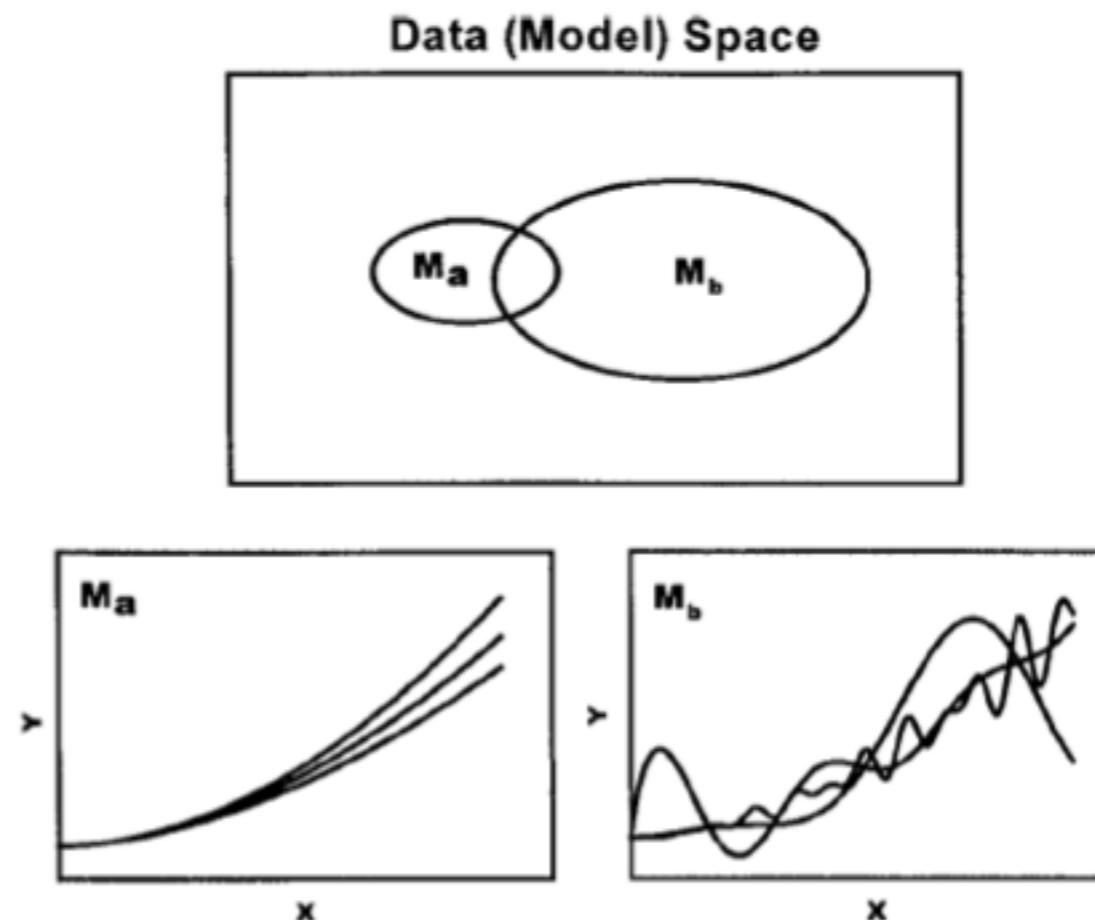


Figure 3. The top panel depicts regions in data space occupied by two models, M_a (simple model) and M_b (complex model), with the range of data patterns that can be generated by each model in the lower panels.

Quantitative Criteria

simple
model
(easy to falsify)

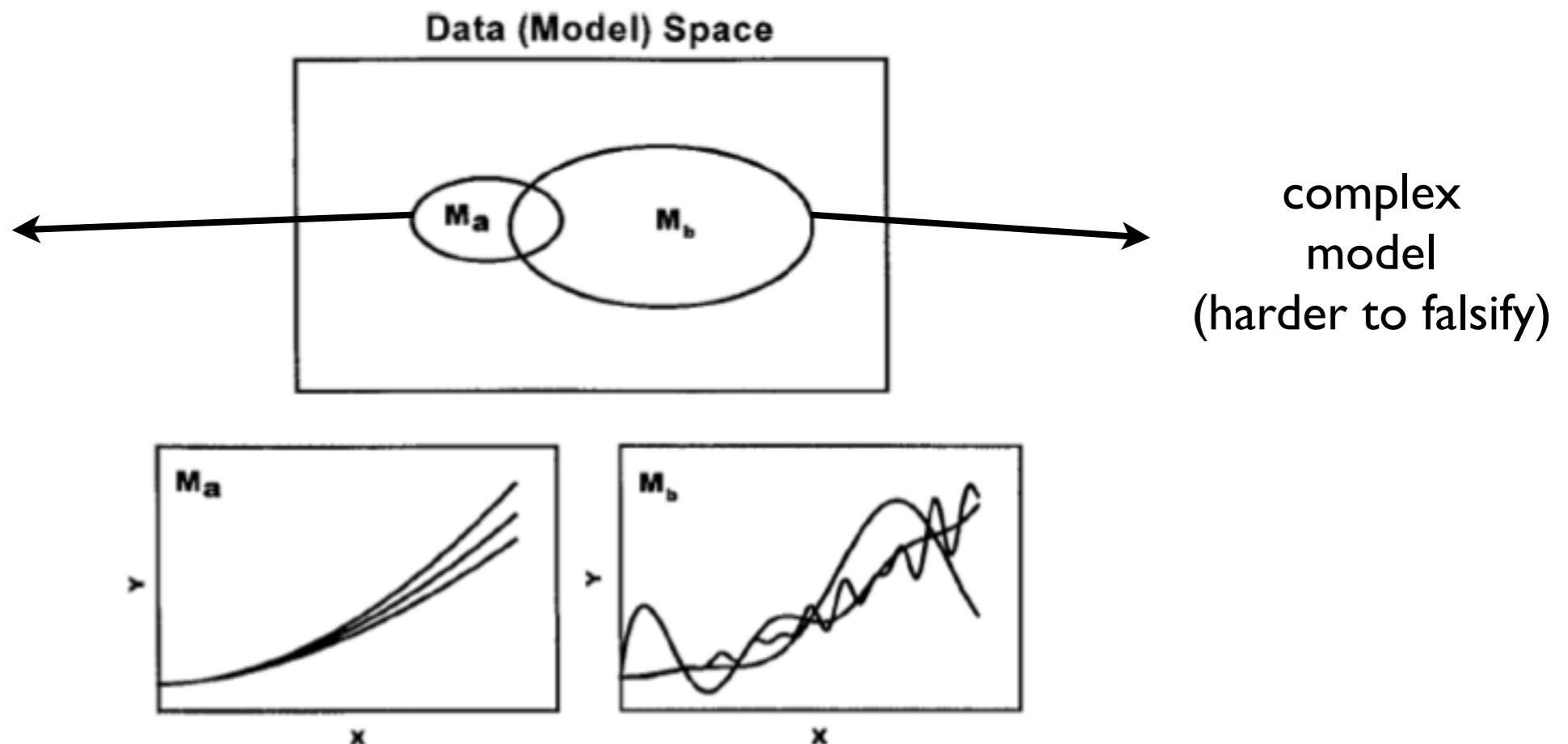


Figure 3. The top panel depicts regions in data space occupied by two models, M_a (simple model) and M_b (complex model), with the range of data patterns that can be generated by each model in the lower panels.

Quantitative Criteria

Table II. Two GOF Measures, four generalizability measures, and the dimensions of complexity to which each is sensitive

Selection method	Criterion equation	Dimensions of complexity considered
Root Mean Squared Error	$RMSE = (SSE/N)^{1/2}$	None
Percent Variance Accounted For	$PVAF=100(1-SSE/SST)$	None
Akaike Information Criterion	$AIC = -2 \ln(f(y \theta_0)) + 2k$	Number of parameters
Bayesian Information Criterion	$BIC = -2 \ln(f(y \theta_0)) + k \cdot \ln(n)$	Number of parameters, sample size
Bayesian Model Selection	$BMS = -\ln \int f(y \theta) \pi(\theta) d\theta$	Number of parameters, sample size, functional form
Minimum Description Length	$MDL = -\ln(f(y \theta_0)) + (k/2)\ln(n/2\pi) + \ln \sqrt{\det(I(\theta))} d\theta$	Number of parameters, sample size, functional form

In the equations above, y denotes observed data, θ is the model's parameter, θ_0 is the parameter value that maximizes the likelihood function $f(y|\theta)$, k is the number of parameters, n is the sample size, N is the number of data points fitted, SSE is the minimized sum of the squared errors between observations and predictions, SST is the sum of the squares total, $\pi(\theta)$ is the parameter prior density, $I(\theta)$ is the Fisher information matrix in mathematical statistics [a], \det denotes the determinant of a matrix, and \ln denotes the natural logarithm of base e.

Quantitative Criteria

Cognitive Science 34 (2010) 10–50
Copyright © 2009 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01076.x

Direct Associations or Internal Transformations? Exploring the Mechanisms Underlying Sequential Learning Behavior

Todd M. Gureckis,^a Bradley C. Love^b

^a*Department of Psychology, New York University*
^b*Department of Psychology, The University of Texas at Austin*

Quantitative Criteria

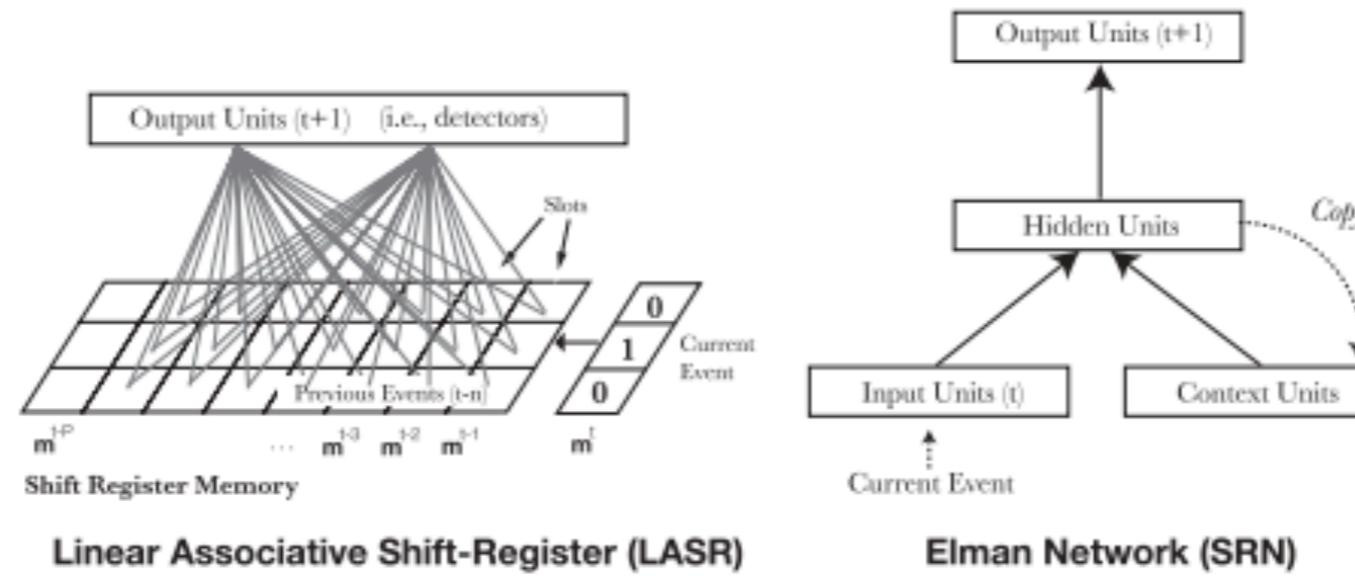


Fig. 1. The schematic architecture of the LASR (left) and SRN (right) networks. In LASR, memory takes the form of a shift-register. New events enter the register on the right and all previous register contents are shifted left by one position. A single layer of detector units learns to predict the next sequence element given the current contents of the register. Each detector is connected to all outcomes at all memory slots in the register. The model is composed of N detectors corresponding to the N event outcomes to be predicted (the weights for only two detectors is shown). In contrast, in the SRN, new inputs are presented on a bank of input units and combine with input from the context units to activate the hidden layer (here, solid arrows reflect fully connected layers, and dashed arrows are one-to-one layers). On each trial, the last activation values of the hidden units are copied back to the context units, giving the model a recurrent memory for recent processing. In both models, learning is accomplished via incremental error-driven adaptation of learning weights.

Quantitative Criteria

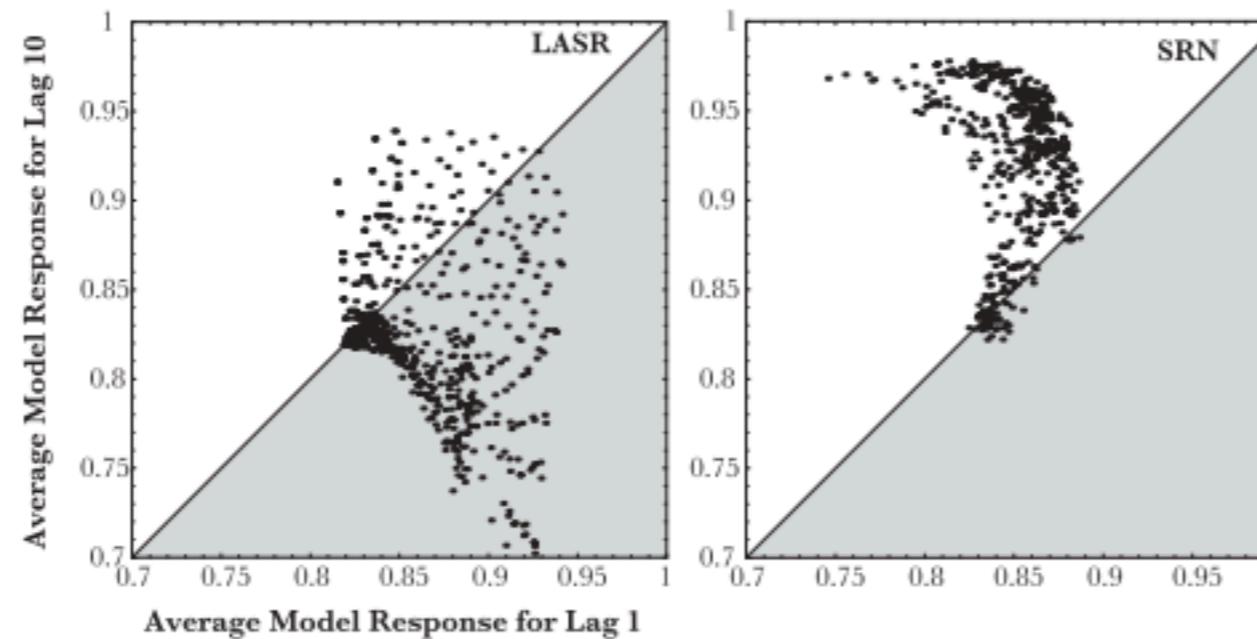


Fig. 9. Explorations of the parameter space for LASR and the SRN in Experiment 1. Each model's average response for lag-1 is plotted against the average response for lag-10. The division between the grey and white regions represents the line $y = x$. Each point in the plot represents the performance of the respective model with a particular setting of the parameters. If the point appears below the $y = x$ line in the grey area, it means the model predicts faster responding to lag-10 events than to lag-1 (the correct qualitative pattern). Note, however, that accounting for the full pattern of human results requires a monotonically decreasing function of predicted RT across all 10 event lags, while this figure only illustrates the two end points (lag-1 and lag-10). Thus, the few instances where the SRN appears to predict the correct pattern are not in general support for the model (see main text).

Quantitative Criteria

people show learning only in this condition (and so does LASR)

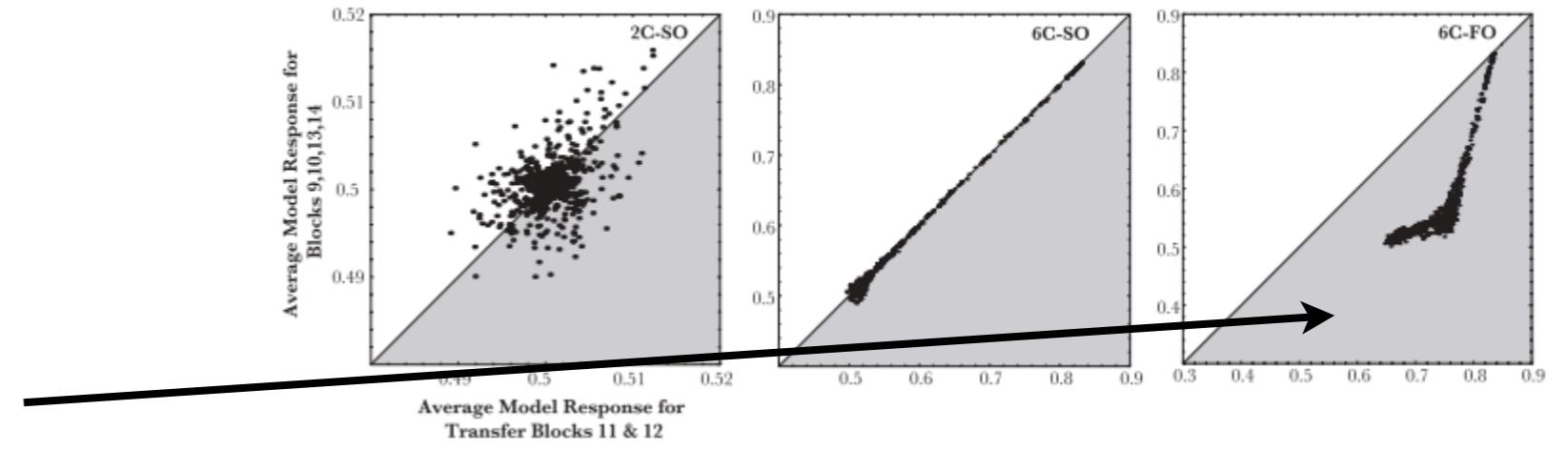


Fig. 11. Explorations of the parameter space of the LASR in the three conditions (2C-SO, 6C-SO, 6C-FO) of Experiment 2. The key behavioral pattern from Experiment 2 was that subjects only responded differently during transfer block in the 6C-FO condition. The model's average response for the transfer blocks 11 and 12 are plotted against the average response for the surrounding learning block (9, 10, 13, and 14). Also plotted is the line $y = x$. Each point in the plot represents the performance of the LASR with a particular setting of the parameters. If the point appears below the $y = x$ line, it means the model predicts slower responding during the transfer blocks (and thus evidence of learning). LASR predicts a systematic learning effect only in the 6C-FO condition (i.e., first-order, linear learning), like human subjects.

SRN shows learning in all conditions!

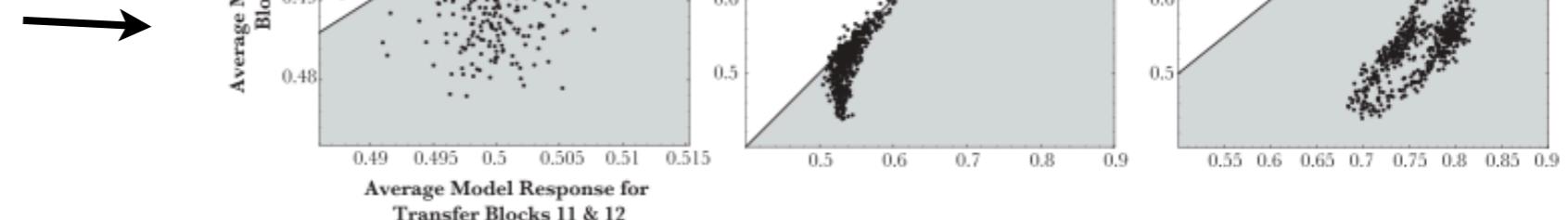


Fig. 12. Explorations of the parameter space of the SRN in the three conditions (2C-SO, 6C-SO, 6C-FO) of Experiment 2. The model's average response for the transfer blocks (11 and 12) are plotted against the average response for the surrounding learning block (9, 10, 13, and 14). Also plotted is the line $y = x$. Each point in the plot represents the performance of the SRN with a particular setting of the parameters. If the point appears below the $y = x$ line, it means the model predicts slower responding during the transfer blocks (and thus a learning effect). The SRN, unlike human subjects, shows a learning effect in all three conditions.

Cross validation

- Fit the model to subset of the data, then test performance on the remaining data (or fit data to one experiment, make prediction in others)
- One approach is leave-one out... fit to all the data minus 1 data point, repeat holding out a different data point each time. Compute average error on the held out point as a measure of model quality!
- “Generalization criterion methodology” (Busemeyer & Wang, 2000) - use calibration design and validation design...
 - basically fit to one part of your experiment that “calibrates the parameters” then make predictions for the interesting part
 - e.g., maybe fit model to learning trials, then evaluate on test trials

Parameter fitting techniques

How do we find the values of model parameters that maximize the fit of a model to observed data?

we need some

“measure of fit”

Examples:

1. Pearson Correlation
2. Sum-squared error
3. Root mean squared error
4. % Variance accounted for
5. Likelihood

we need some

“measure of fit”

Examples:

1. Pearson Correlation
2. Sum-squared error
3. Root mean squared error
4. % Variance accounted for
5. Likelihood ←



goodness of fit measures

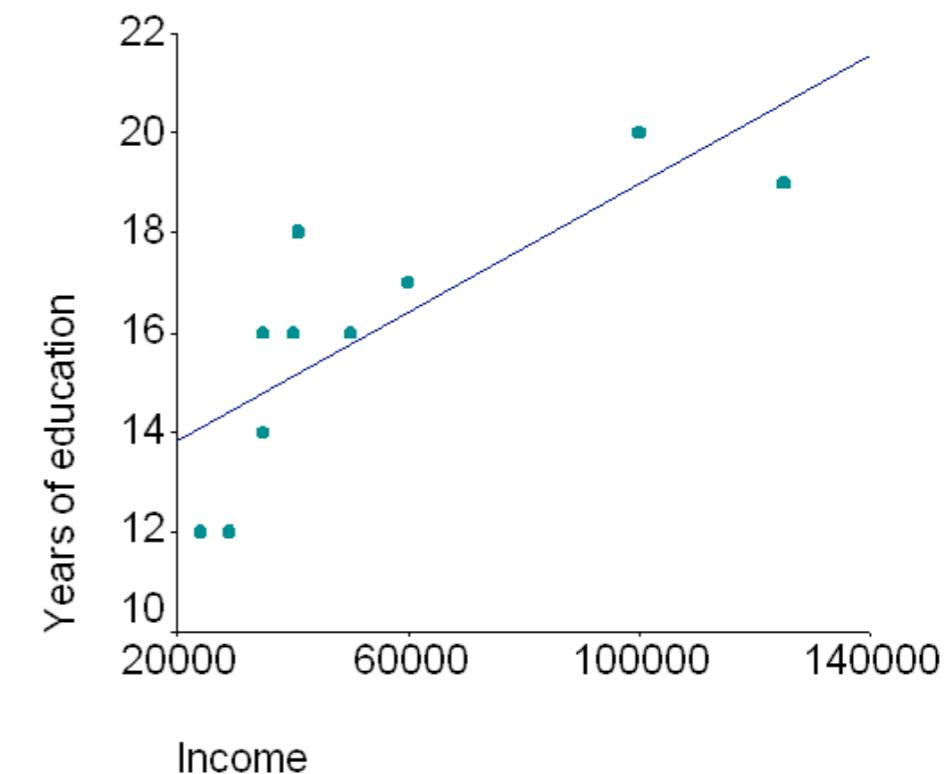
ironically LESS GOOD!

better!

pearson correlation

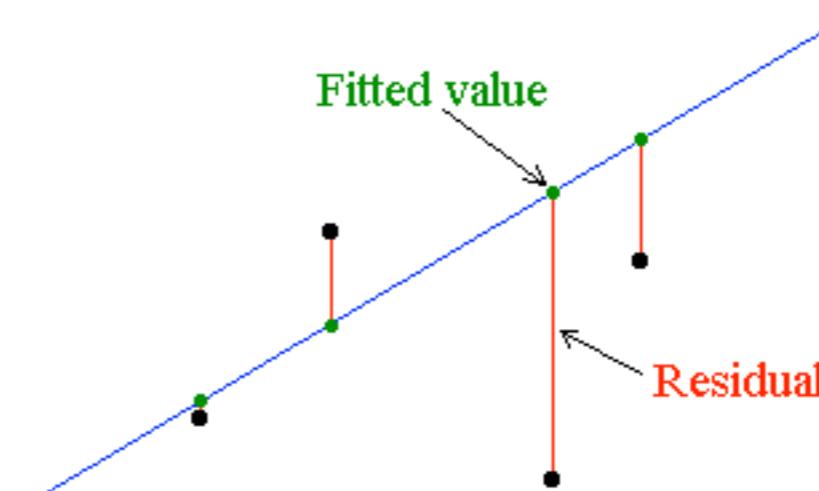
$$r_{obs,prd} = \frac{\sum (obs - \mu_{obs})(prd - \mu_{prd})}{\sqrt{\sum (obs - \mu_{obs})^2 \sum (prd - \mu_{prd})^2}}$$

everyone is likely very familiar with this!



sum of squared error (SSE)

$$SSE_{obs,prd} = \sum (obs - prd)^2$$



nothing new here.

root mean squared error (RMSE)

$$RMSE_{obs,prd} = \sqrt{\frac{\sum (obs - prd)^2}{N}}$$

used this last time.

% variance

$$\%Var = \frac{SSE_{null} - SSE_{model}}{SSE_{null}}$$

$$SSE_{null} = \sum_i (obs_i - \mu_{obs})^2$$

$$SSE_{model} = \sum_i (obs_i - prd_i)^2$$

likelihood

$$p(d|model)$$

more on this in a bit

ok, so how to find good parameters?

1. Calculus
2. Grid search
3. Hill climbing
4. Nelder-meade simplex (used by fmin in python and matlab)
5. Simulated Annealing
6. Genetic algorithms

calculus

d_ij	obs s_ij	prd s_ij
0	1	
1	0.368	
2	0.135	
3	0.05	
4	0.018	
5	0.007	

$$s_{ij} = \alpha + \beta d_{ij}$$

find parameters for regression that
minimize SSE between obs s_ij and
predicted sij

calculus

$$SSE = \sum_k (obs_k - prd_k)^2$$

$$SSE = \sum_k (obs_k - (\alpha + \beta d_k))^2$$

calculus

$$SSE = \sum_k (obs_k - prd_k)^2$$

$$SSE = \sum_k (obs_k - (\alpha + \beta d_k))^2$$

$$\frac{\partial SSE}{\partial \alpha} = \sum_k 2(obs_k - \alpha - \beta d_k)(-1)$$

$$\frac{\partial SSE}{\partial \alpha} = -2 \sum_k (obs_k - \alpha - \beta d_k)$$

$$\frac{\partial SSE}{\partial \beta} = \sum_k 2(obs_k - \alpha - \beta d_k)(-d_k)$$

$$\frac{\partial SSE}{\partial \beta} = -2 \sum_k d_k (obs_k - \alpha - \beta d_k)$$

calculus

$$SSE = \sum_k (obs_k - prd_k)^2$$

$$SSE = \sum_k (obs_k - (\alpha + \beta d_k))^2$$

$$\frac{\partial SSE}{\partial \alpha} = \sum_k 2(obs_k - \alpha - \beta d_k)(-1)$$

$$\frac{\partial SSE}{\partial \alpha} = -2 \sum_k (obs_k - \alpha - \beta d_k)$$

$$\frac{\partial SSE}{\partial \alpha} = 0$$

$$\frac{\partial SSE}{\partial \beta} = \sum_k 2(obs_k - \alpha - \beta d_k)(-d_k)$$

$$\frac{\partial SSE}{\partial \beta} = -2 \sum_k d_k (obs_k - \alpha - \beta d_k)$$

$$\frac{\partial SSE}{\partial \beta} = 0$$

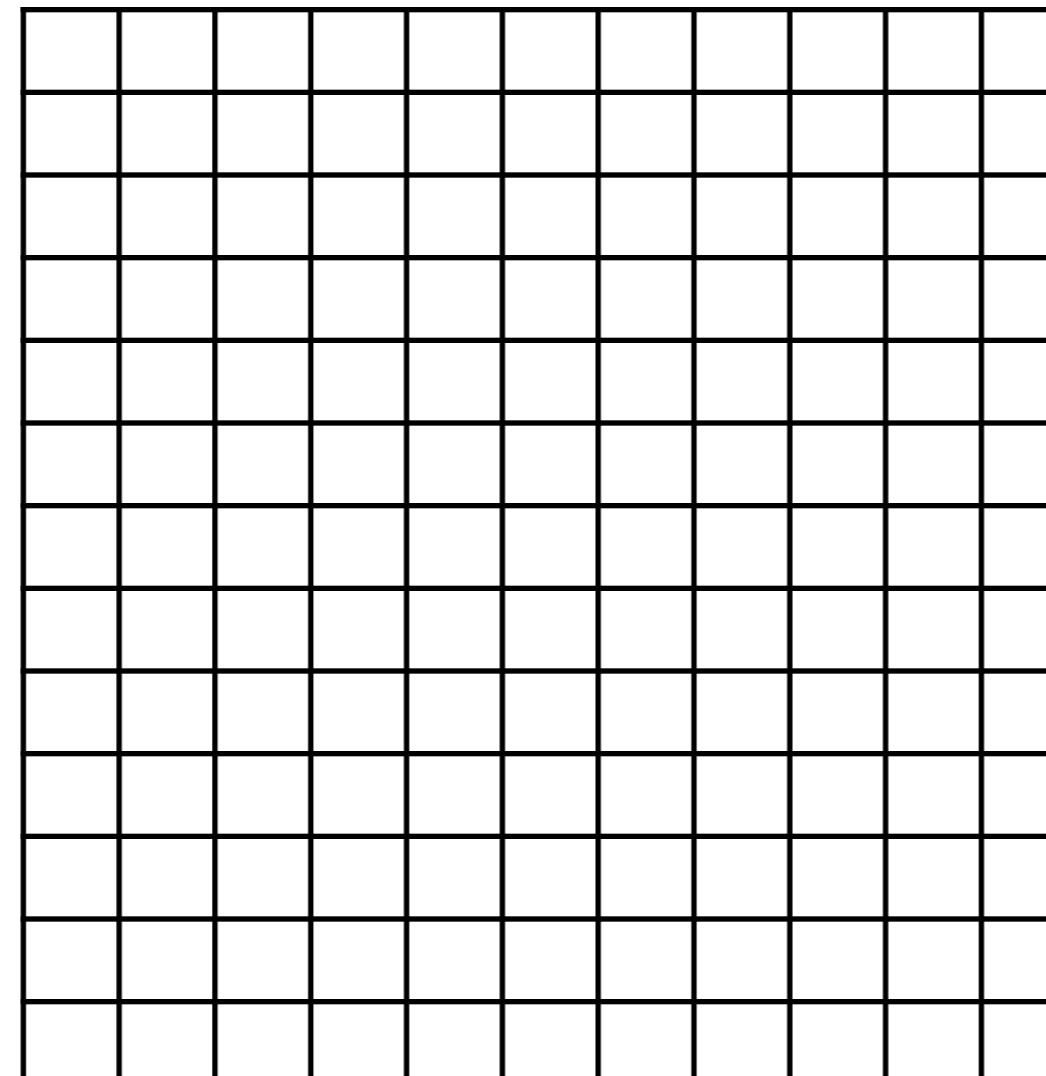
**ok, who want to write this
down for the phonological
loop?**

**ok, who want to write this
down for the phonological
loop?**

in general, when you aren't smart enough, simulate!

grid search

parameter 2



parameter 1

calculated SSE at each combination of parameters and keep the best.

grid search

```
from scipy.optimize import brute  
brute(evalmodel, [[min, max], [min, max]], args, Ns=20, full_output=1)
```

grid search

in general, this sucks.

evaluation time for one set of parameters

x

of evaluations

grid search

in general, this sucks.

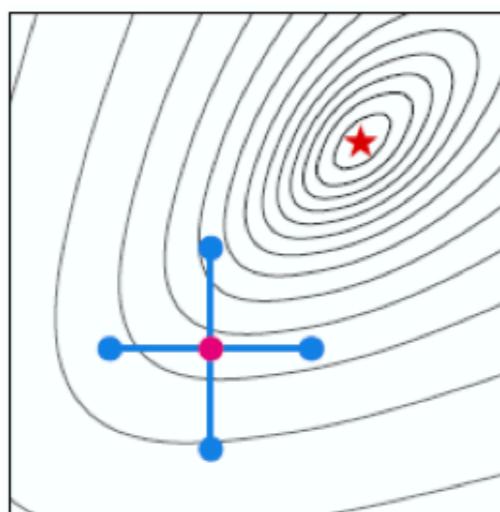
100 seconds per evaluation

x

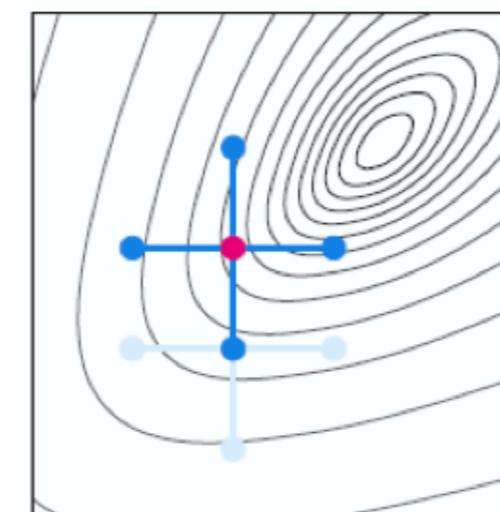
10^{12} of evaluations

10^{14} seconds = 3 million years!

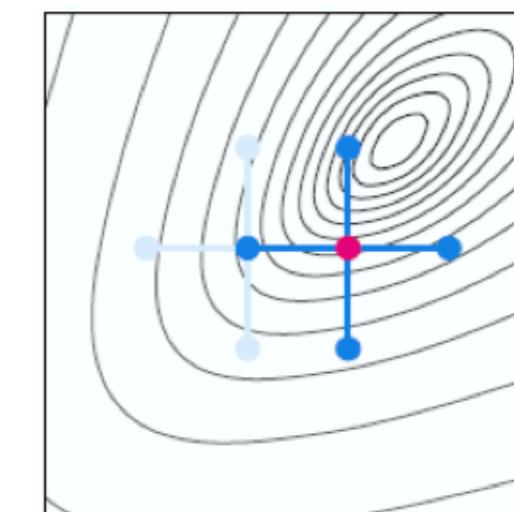
hill climbing



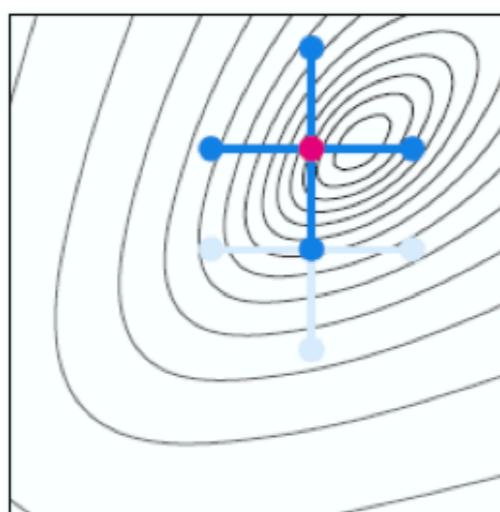
(a) Initial pattern



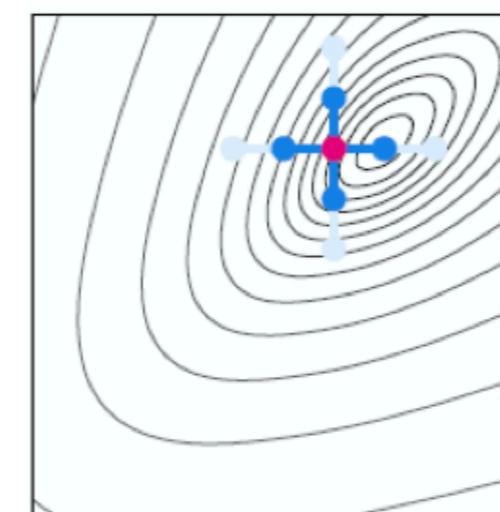
(b) Move North



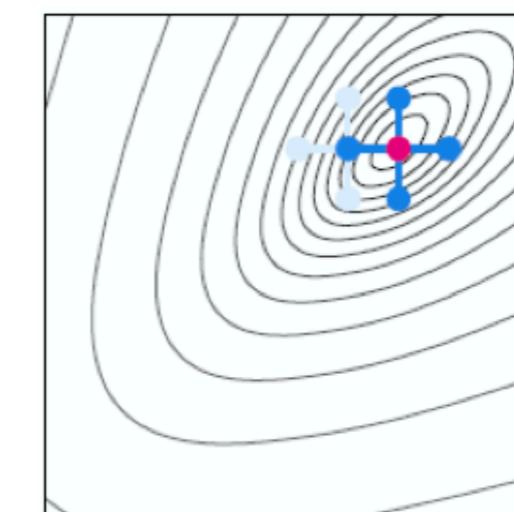
(c) Move West



(d) Move North



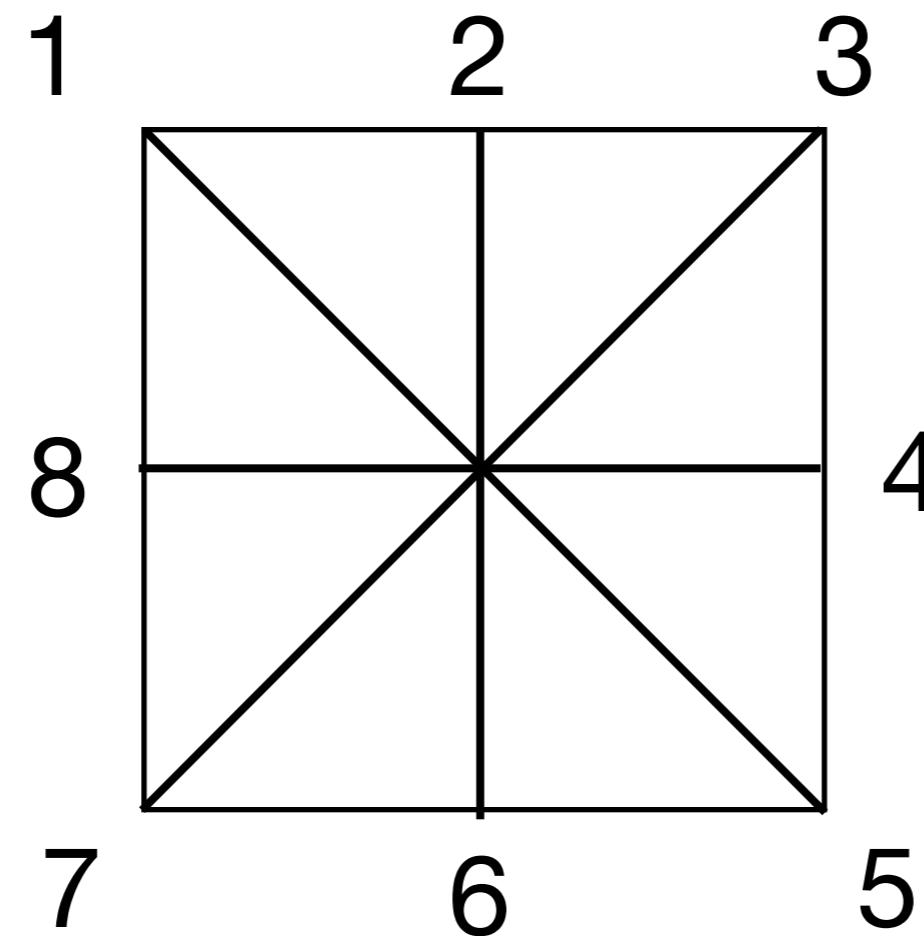
(e) Contract



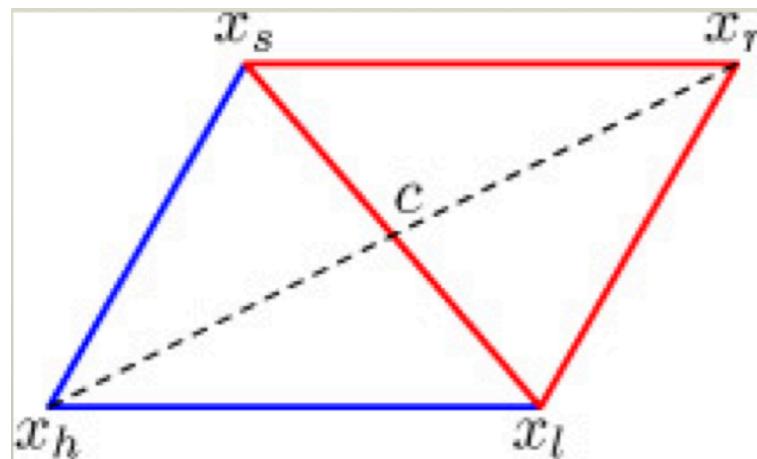
(f) Move West

hill climbing

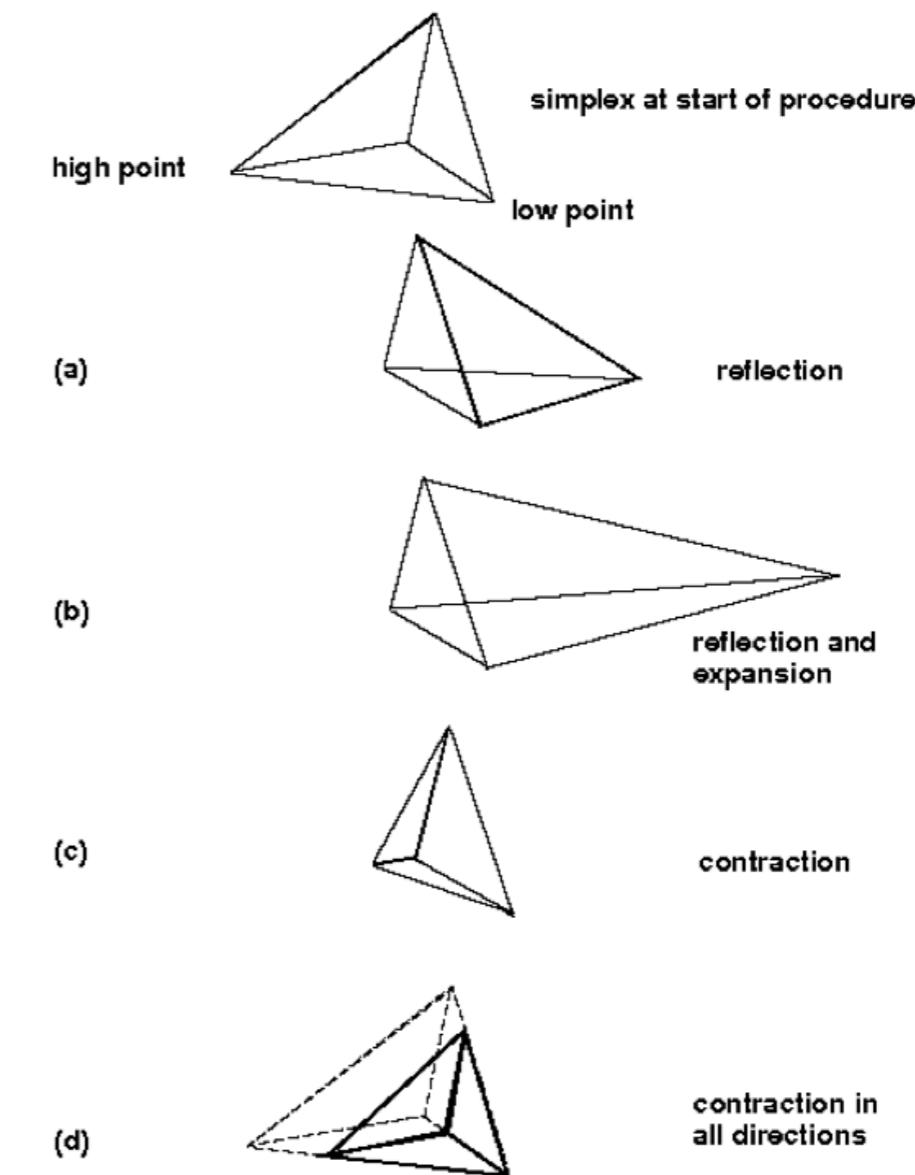
how many points are evaluated each step?



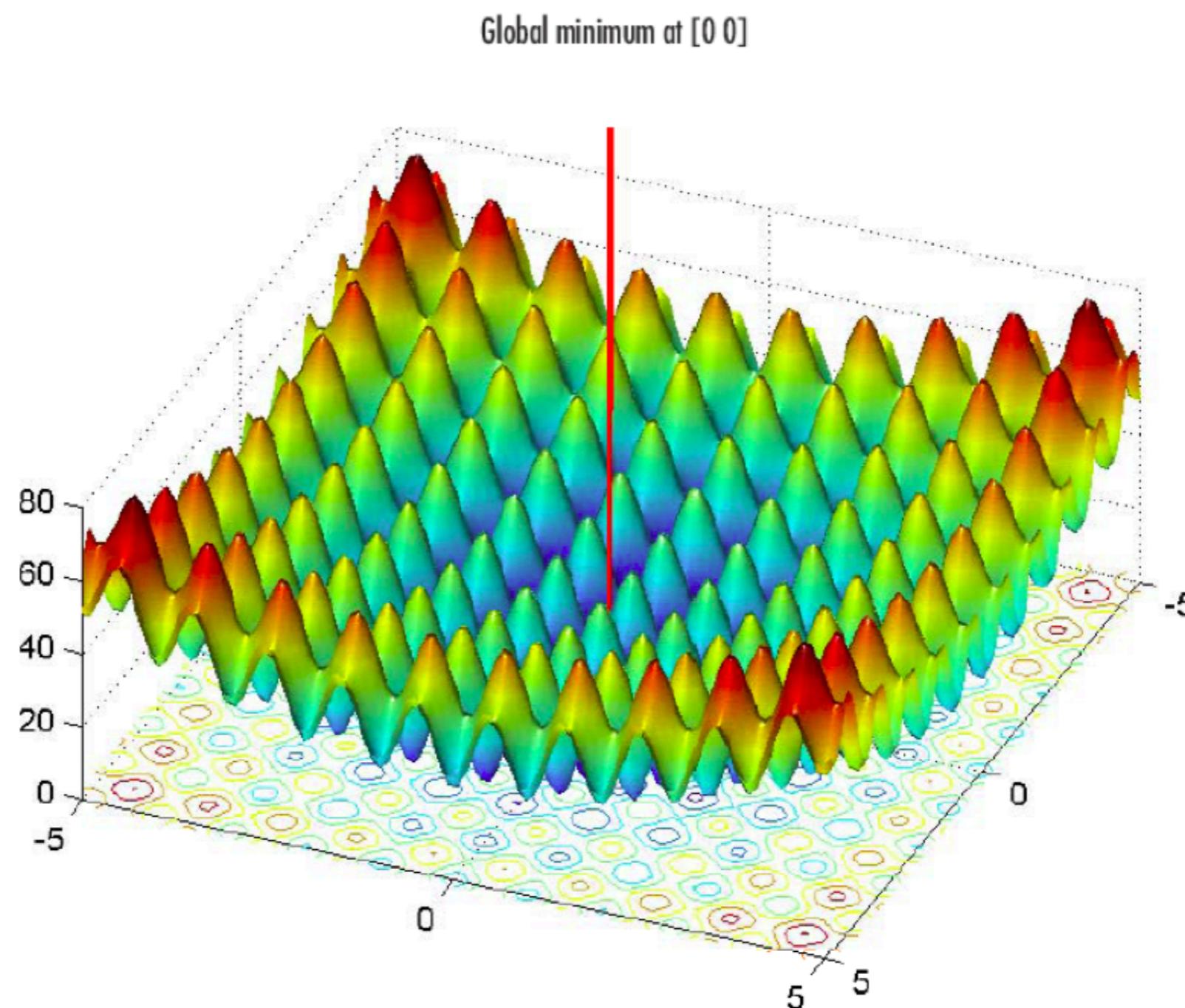
nelder mead simplex



the simplex approach minimizes the number of points you have to evaluate



start multiple places!



does you search code work?

how do you know?

does your search code work?

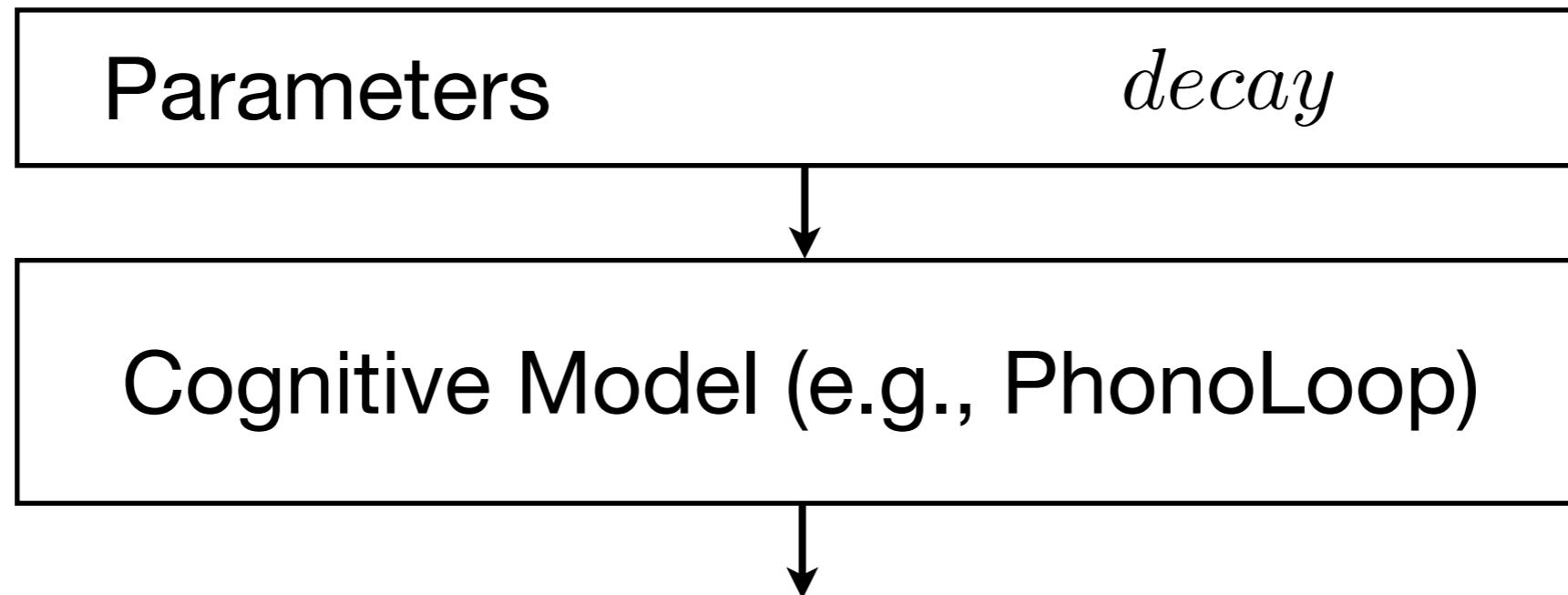
One idea:

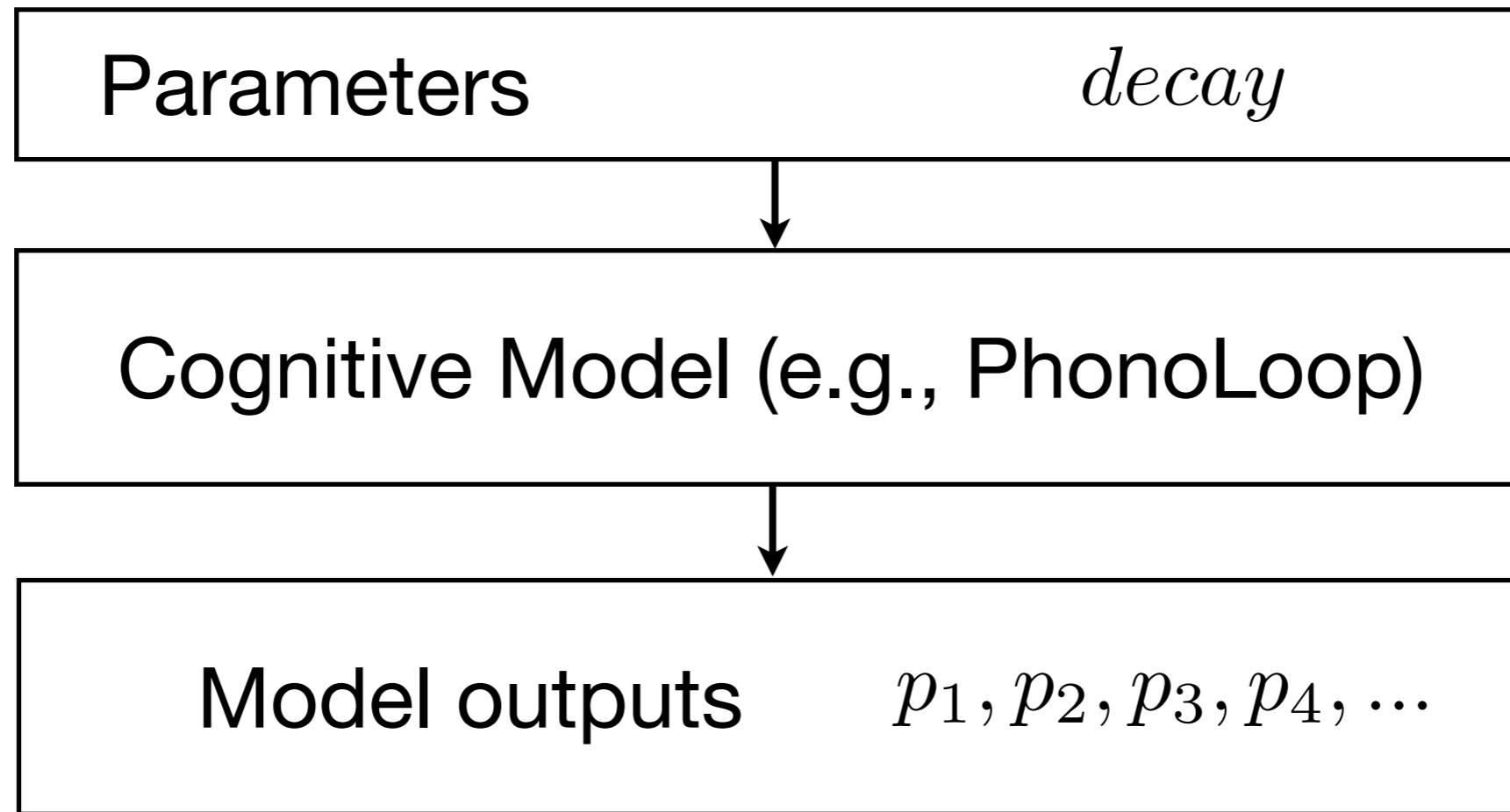
Run your model with known parameters, to GENERATE pretend human data. Then, re-fit the model using your fit measure and parameter search algorithm to see if you can get the same parameters you used to generate the data.

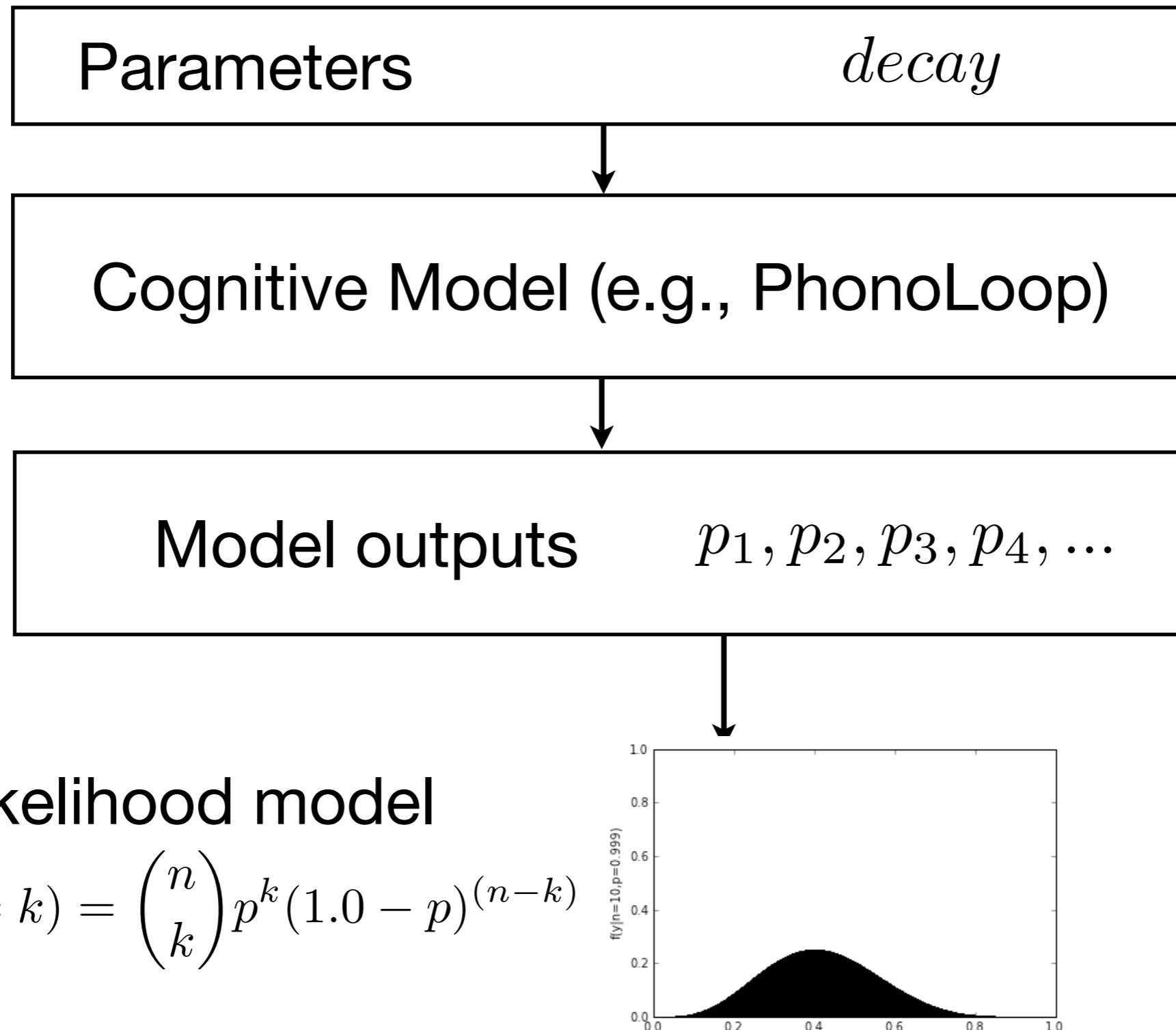
If this doesn't work the model parameters may be non-identifiable (multiple parameters give the same fit).

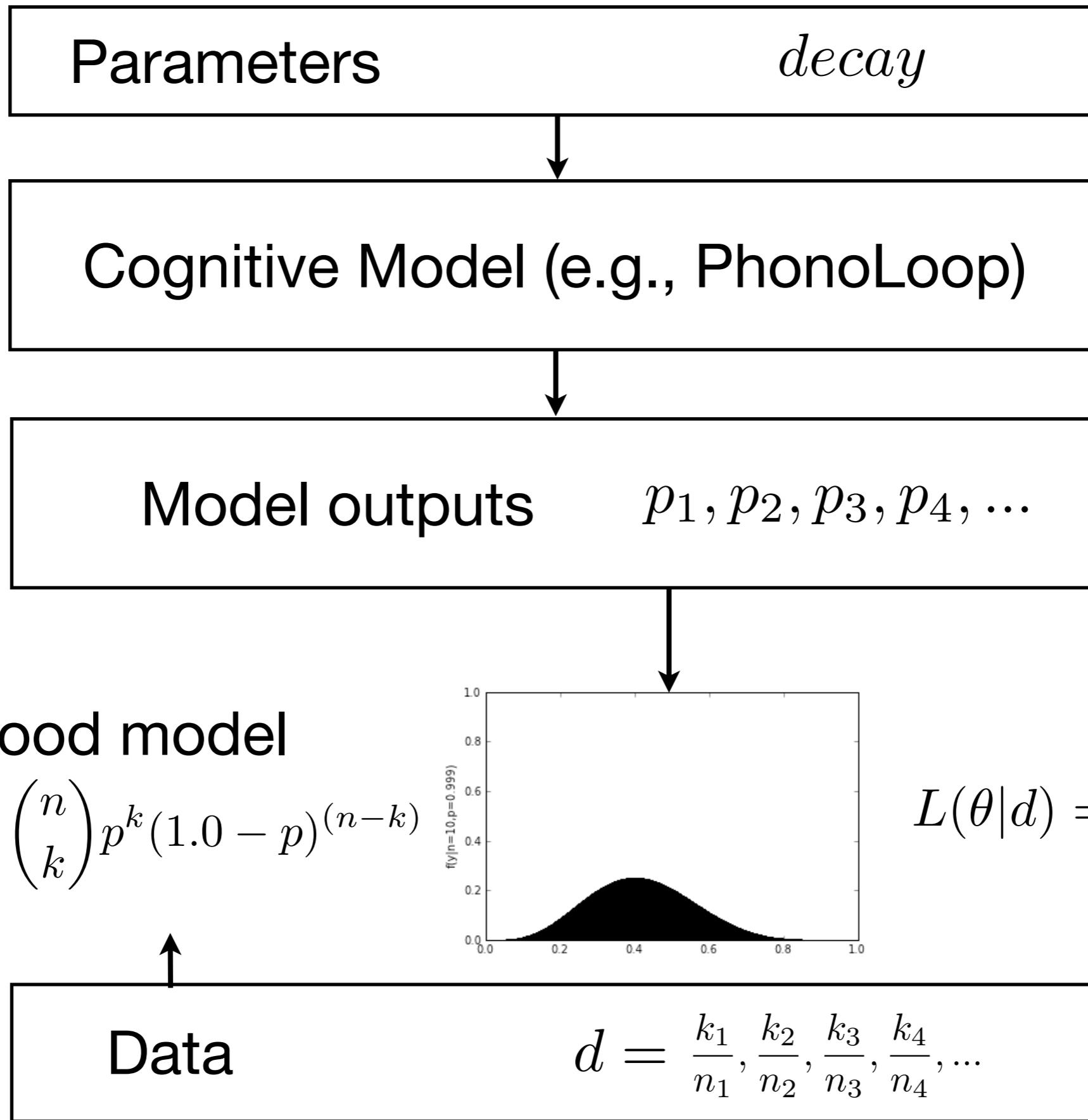
Maximum likelihood estimation

- Myung (2003) - “MLE is not as widely recognized among modelers in psychology, but it is a standard approach to parameter estimation and inference in statistics.”
- Sufficiency - complex information about parameter of interest contained in MLE estimator
- Consistency - true parameter value that generated the data can be recovered asymptotically
- Parameter invariance - same MLE solution obtained independent of way parameterized
- LSE is a hack, but is correct for linear regression. But shouldn't be used all over the place.



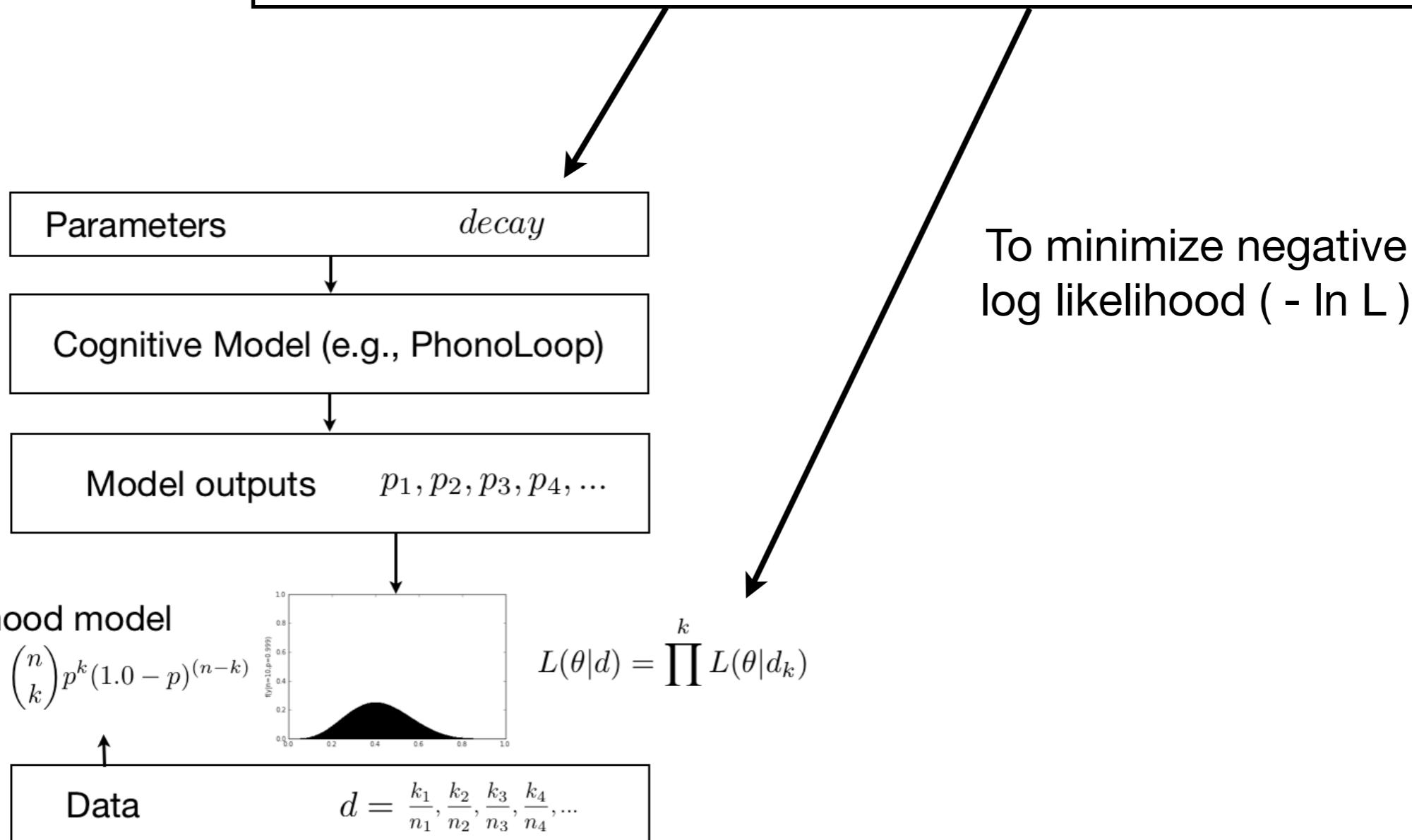




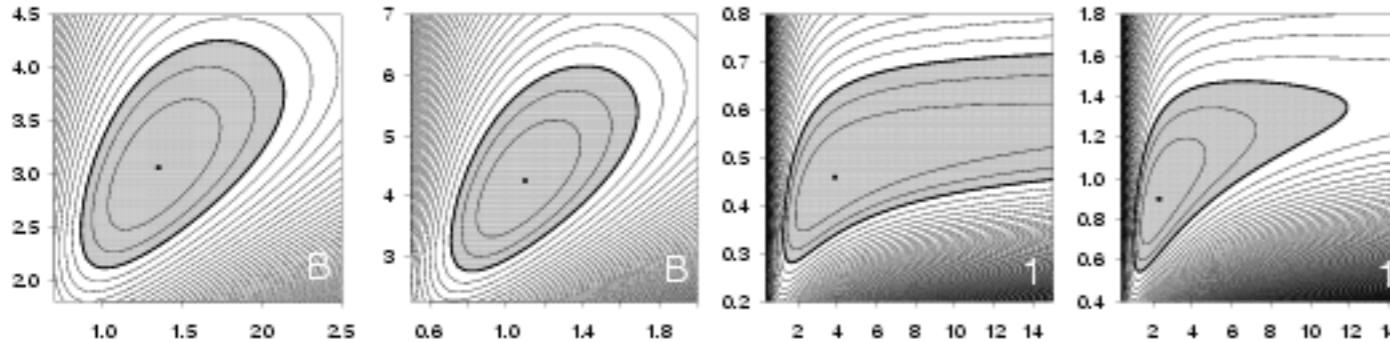


Search algorithm

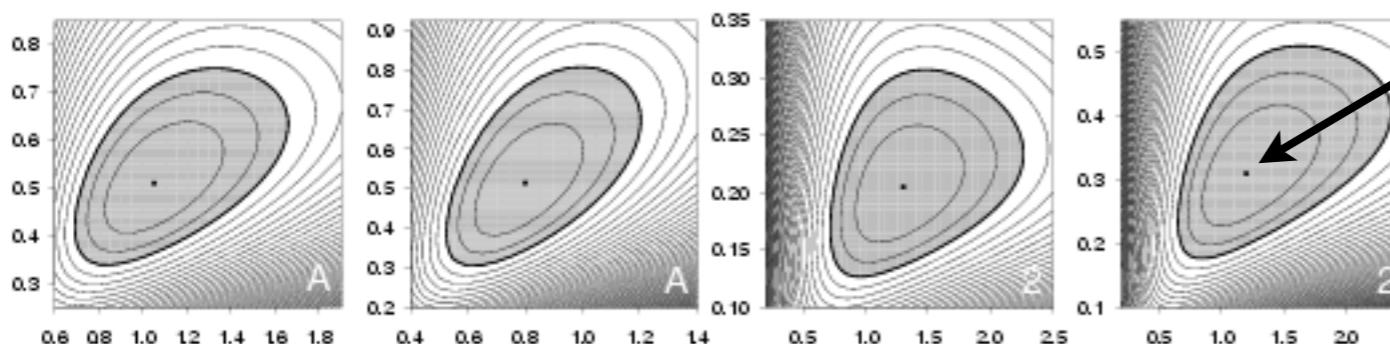
`fmin`



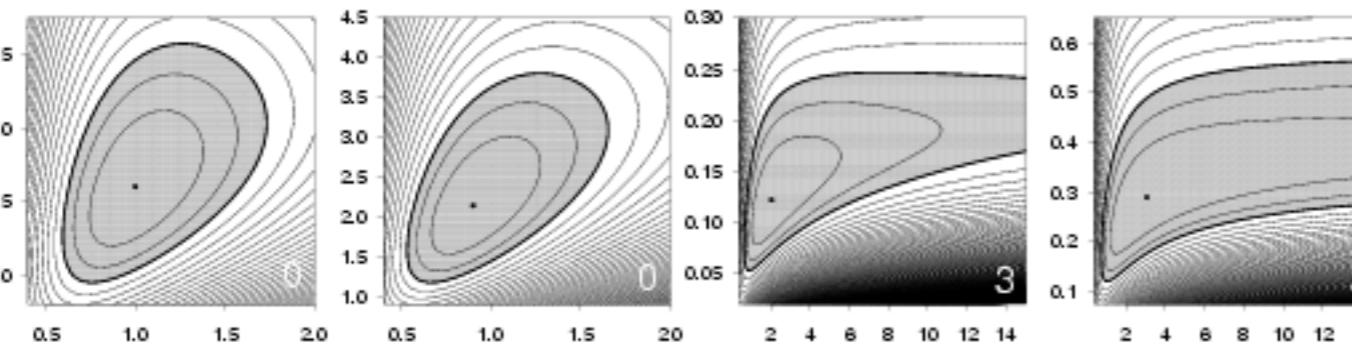
Parameter uncertainty



the process
described on the previous
slides results in a single set of
parameters



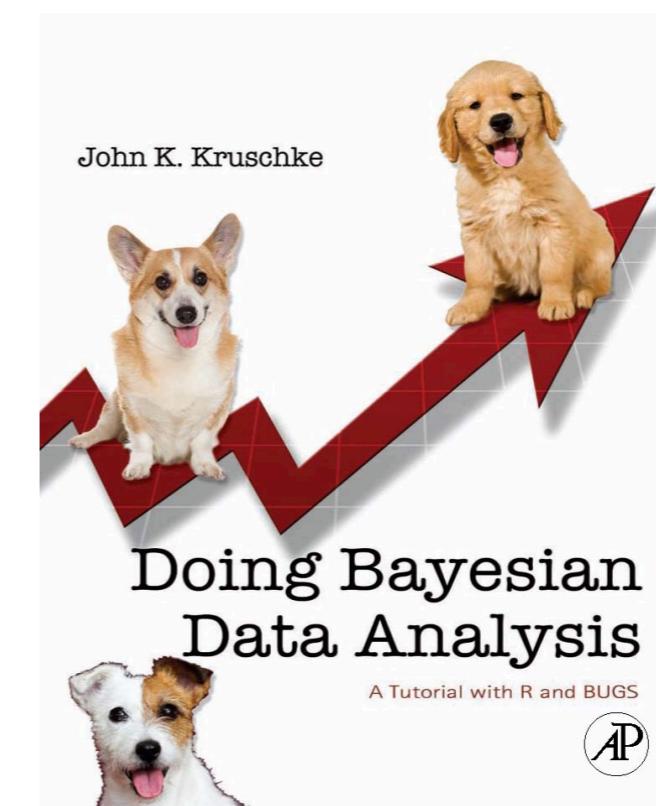
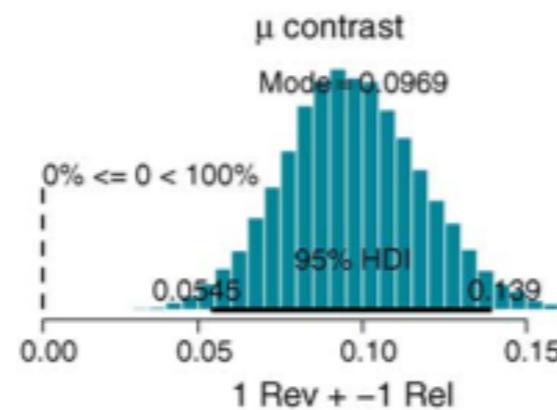
a “point” estimate.



Parameter uncertainty

- If we run the experiment again and get different data, the ML parameter will likely change
- How can we capture some of this uncertainty? In other words, when we tell people the best parameters are X, Y, Z, can we also convey our confidence in those values?

1. Be a Bayesian



Parameter uncertainty

- If we run the experiment again and get different data, the ML parameter will likely change
- How can we capture some of this uncertainty? In other words, when we tell people the best parameters are X, Y, Z, can we also convey our confidence in those values?

2. Analyze your model parameters as “measurements” across subjects

A Contribution of Cognitive Decision Models to Clinical Assessment: Decomposing Performance on the Bechara Gambling Task

Jerome R. Busemeyer and Julie C. Stout
Indiana University Bloomington

BUSEMEYER AND STOUT

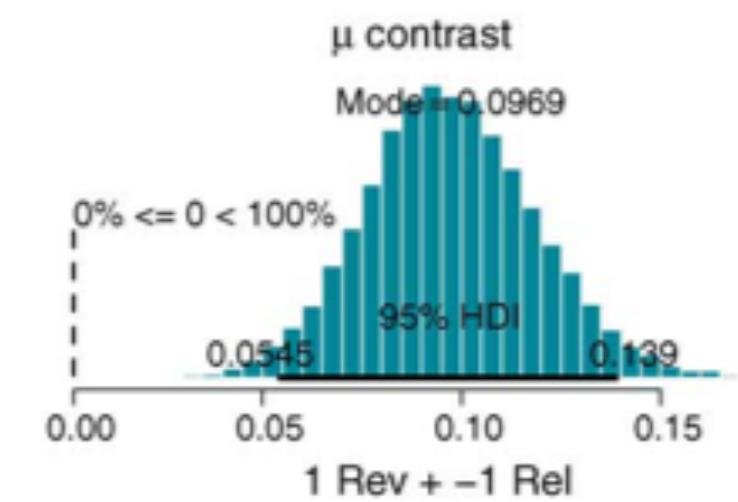
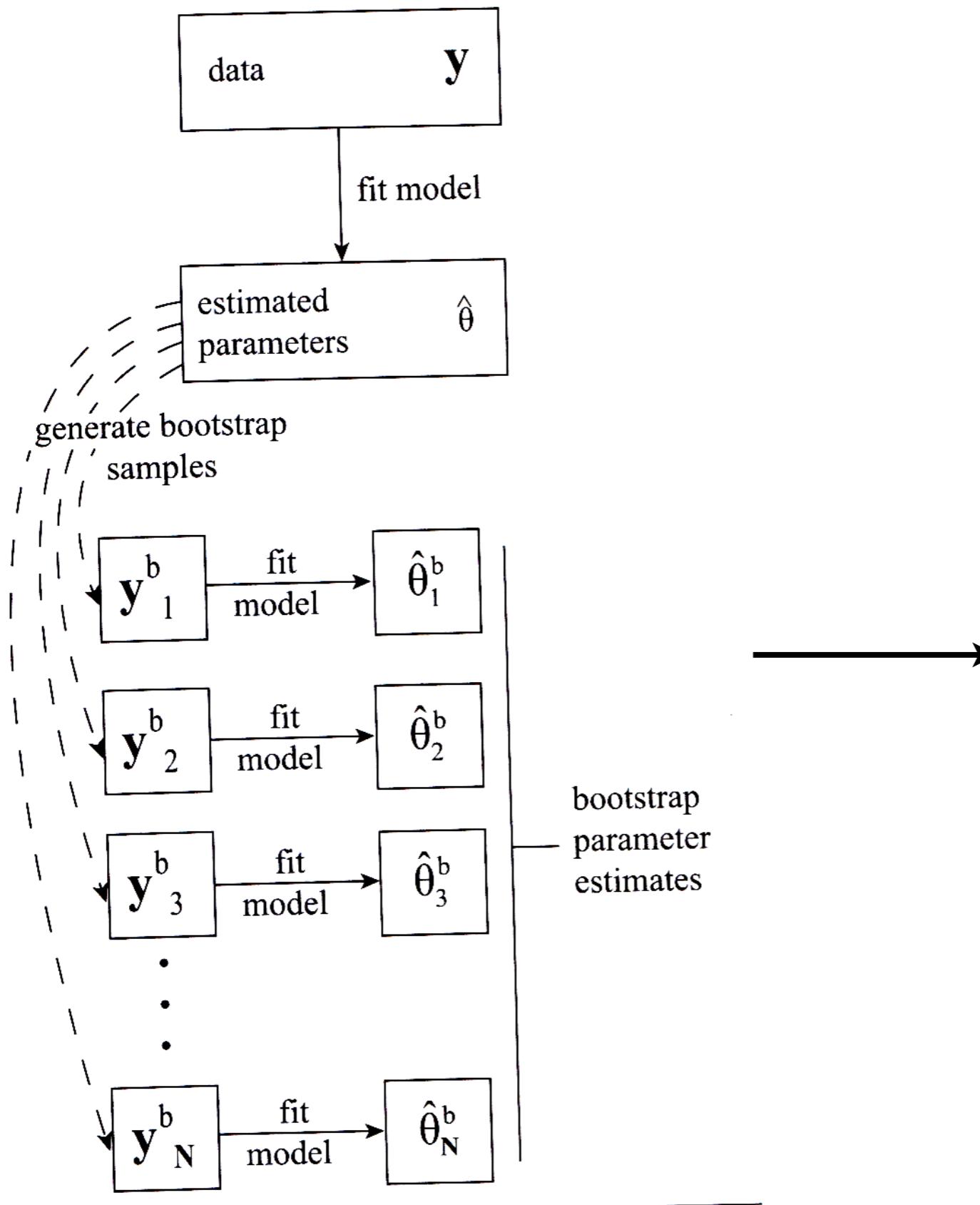
Table 2
Parameter Estimates From Expectancy–Valence Model

Group	Updating rate			Attention weight			Sensitivity		
	<i>M</i>	<i>Mdn</i>	<i>SD</i>	<i>M</i>	<i>Mdn</i>	<i>SD</i>	<i>M</i>	<i>Mdn</i>	<i>SD</i>
Healthy	0.34	0.16	0.39	0.35	0.32	0.30	0.32	0.68	1.58
Huntington	0.58	0.78	0.46	0.40	0.17	0.47	-0.89	-0.78	2.13
Parkinson	0.29	0.09	0.40	0.28	0.20	0.32	0.36	0.01	1.99

Parameter uncertainty

- If we run the experiment again and get different data, the ML parameter will likely change
- How can we capture some of this uncertainty? In other words, when we tell people the best parameters are X, Y, Z, can we also convey our confidence in those values?

3. Bootstrap!

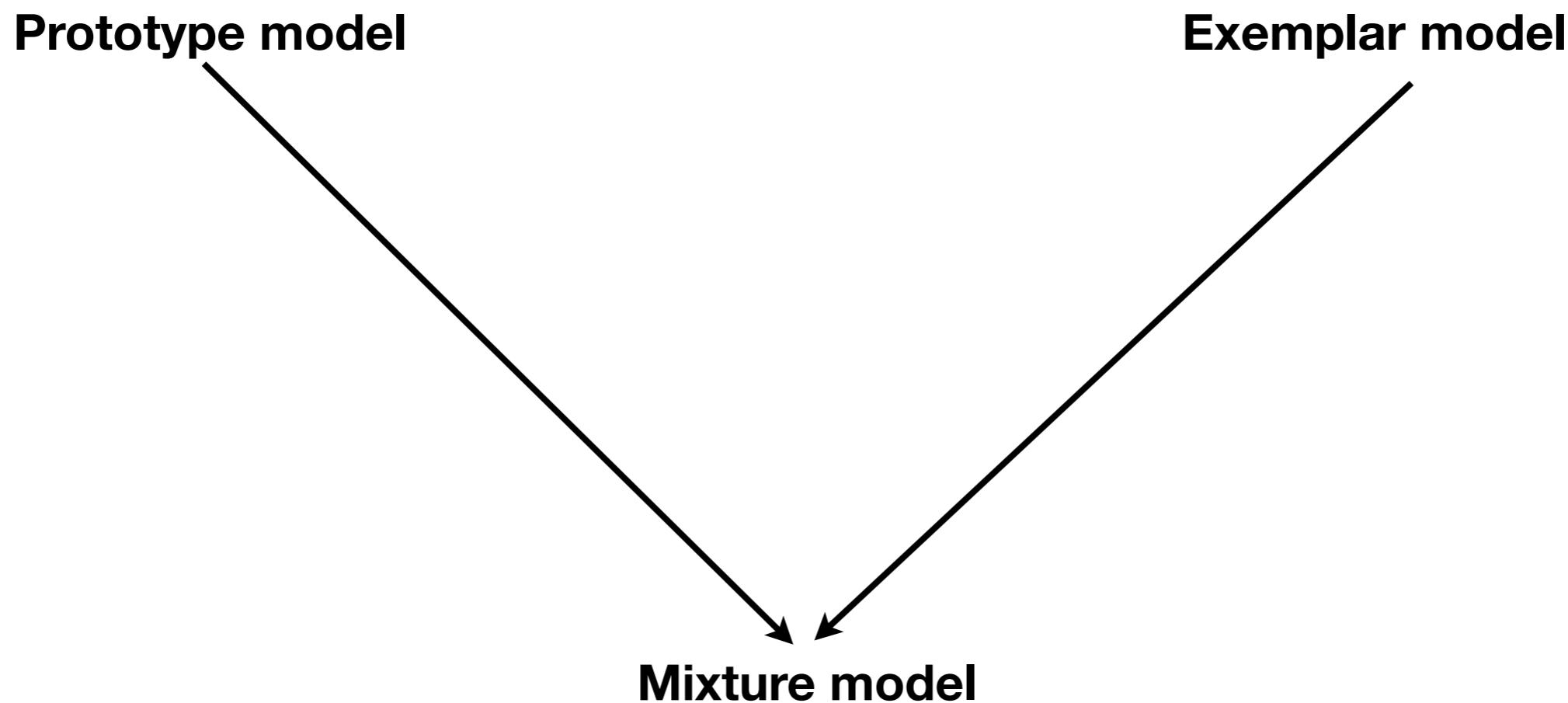


Parameter uncertainty

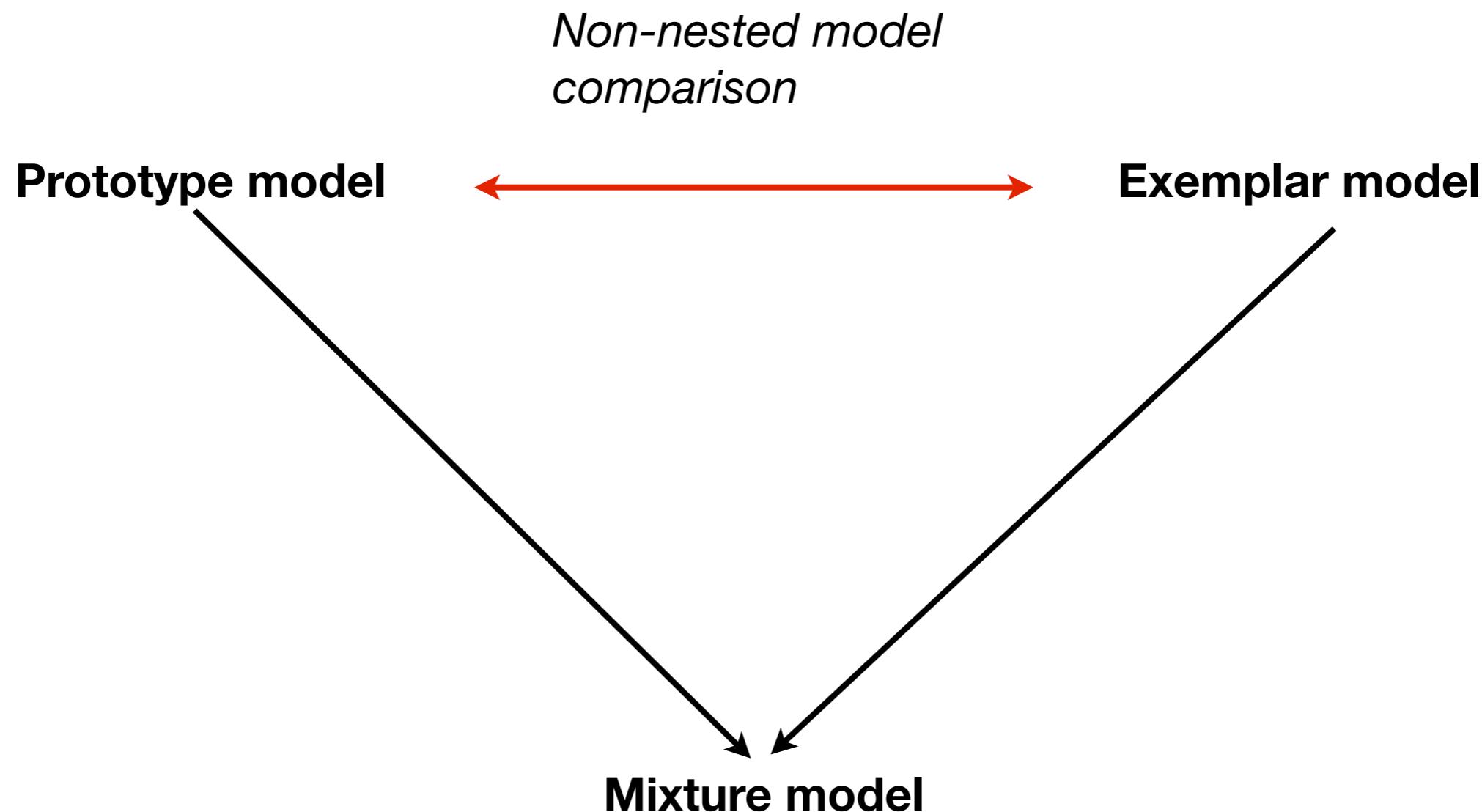
- If we run the experiment again and get different data, the ML parameter will likely change
- How can we capture some of this uncertainty? In other words, when we tell people the best parameters are X, Y, Z, can we also convey our confidence in those values?

4. Look at the curvature
(read the book)

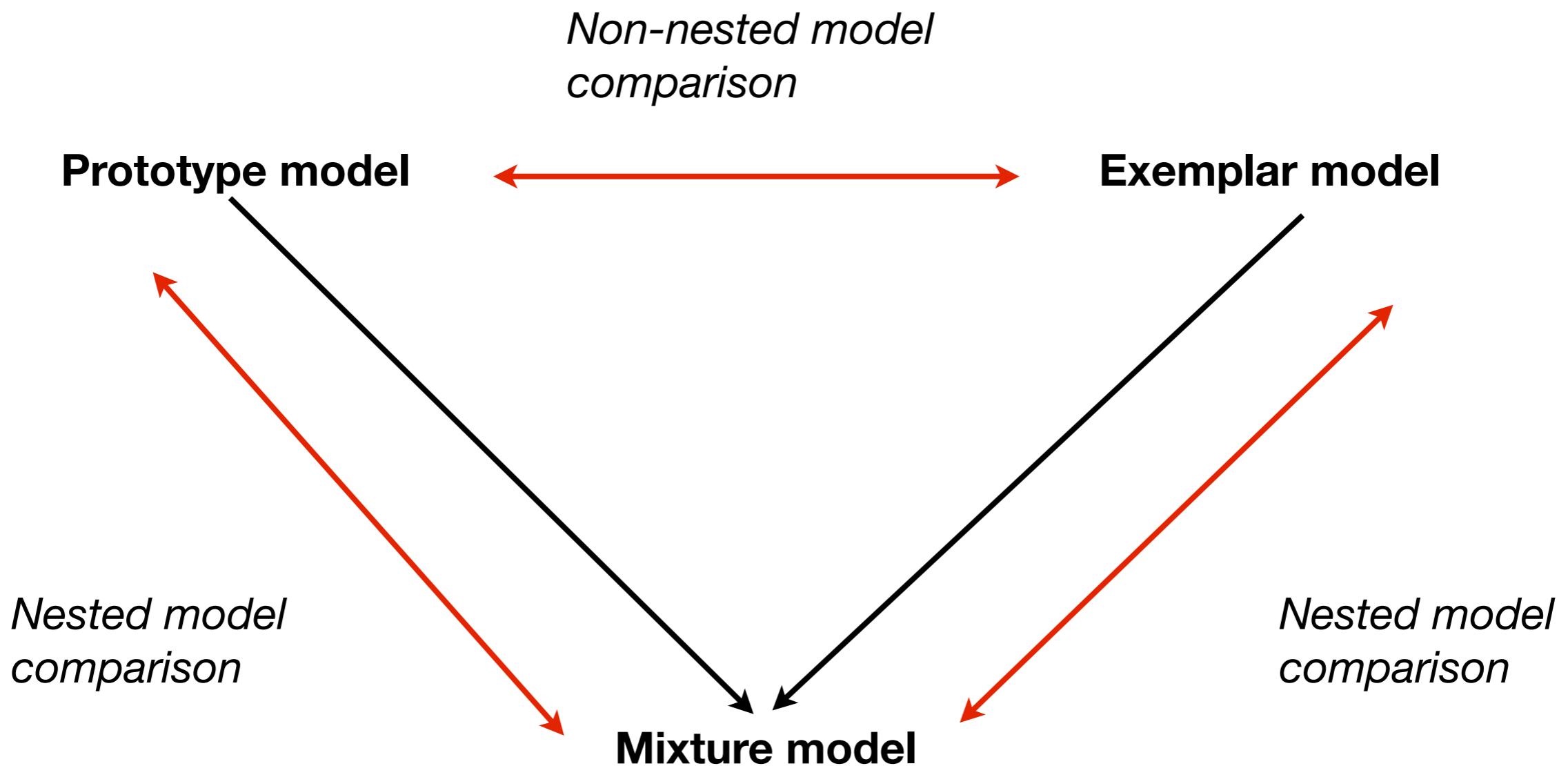
Model comparison



Model comparison



Model comparison



Null model



Exemplar model

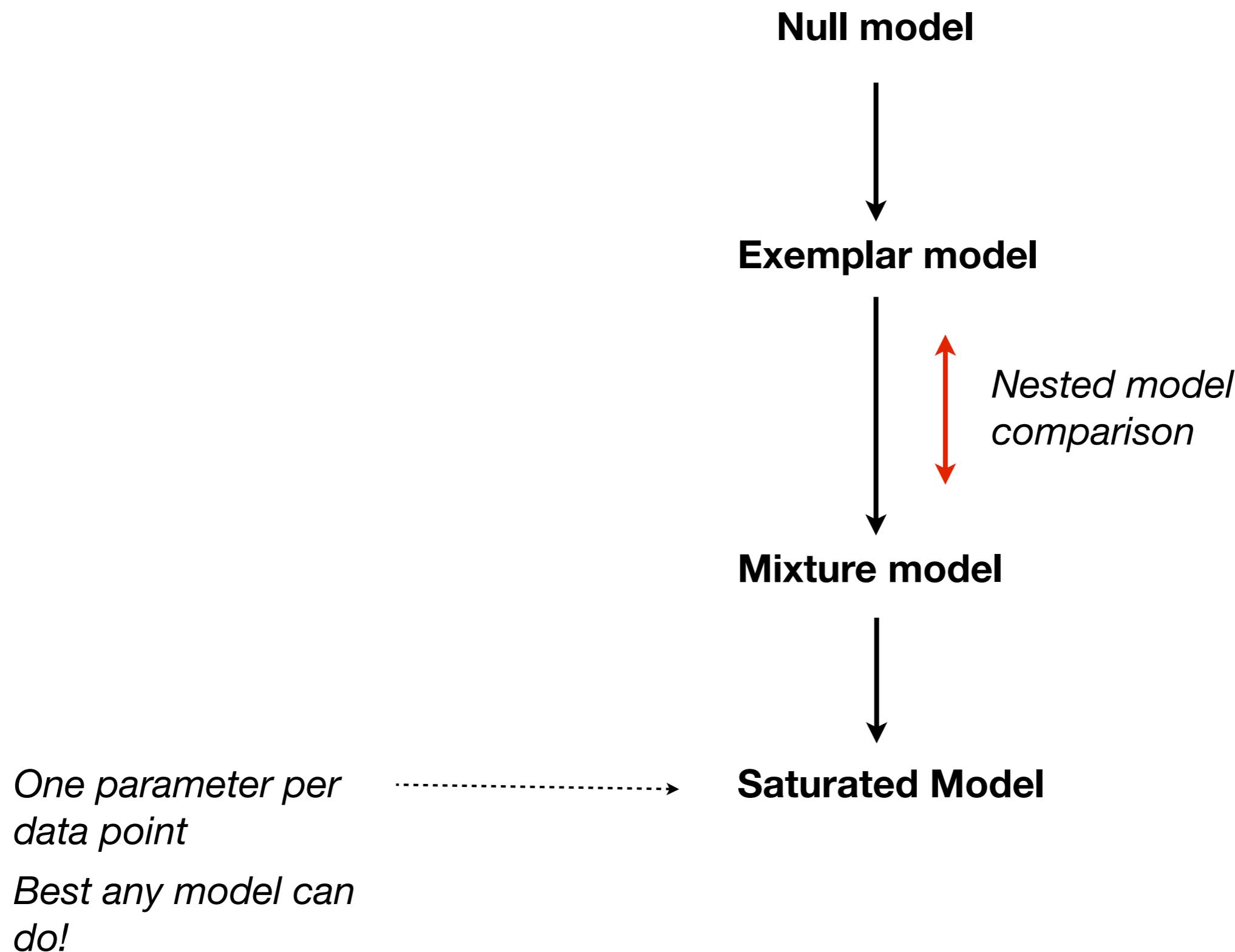


*Nested model
comparison*

Mixture model



Saturated Model



*Not as bad as any
model can do, but a
good floor*



Null model

Exemplar model

Mixture model

Saturated Model

Nested model comparison

One parameter per data point

Best any model can do!

*Not as bad as any
model can do, but a
good floor*

*To be useful this model
has to fit better than
the null!*

*One parameter per
data point*

*Best any model can
do!*



Null model



Exemplar model



*Nested model
comparison*

Mixture model



Saturated Model

*Not as bad as any
model can do, but a
good floor*

*To be useful this model
has to fit better than
the null!*

*Logically this model
has to fit better... but
significantly better?*

*One parameter per
data point*

*Best any model can
do!*



Null model



Exemplar model



Mixture model

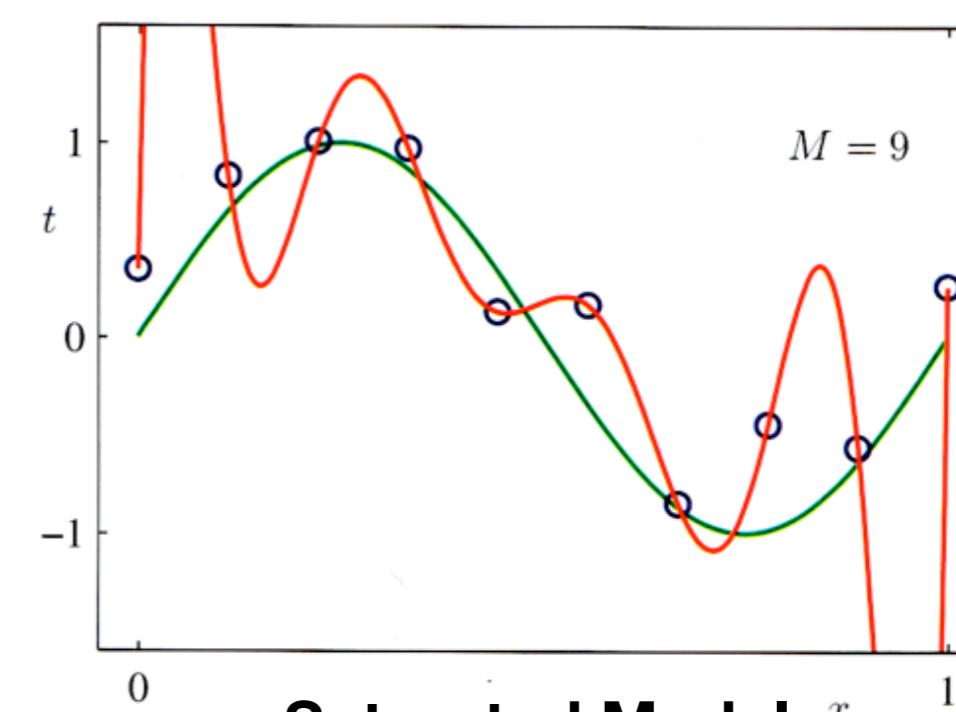
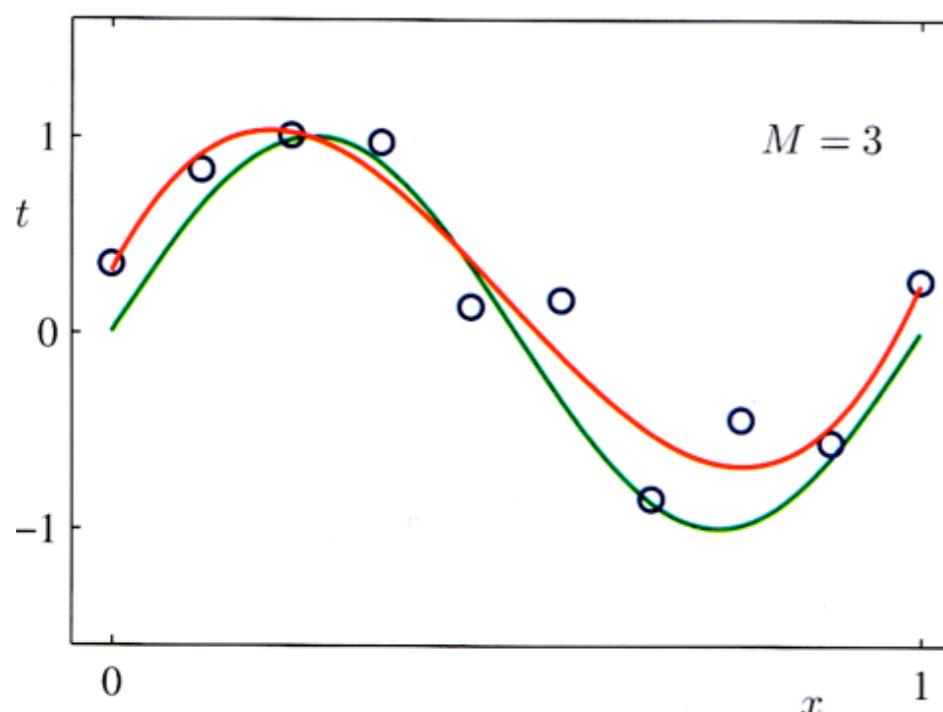
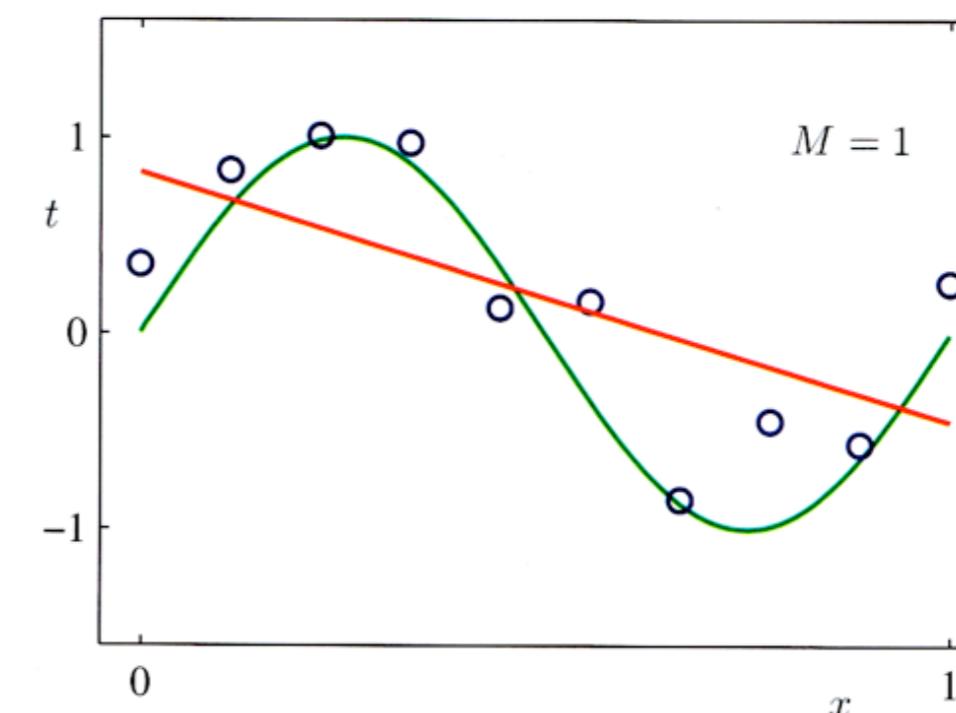
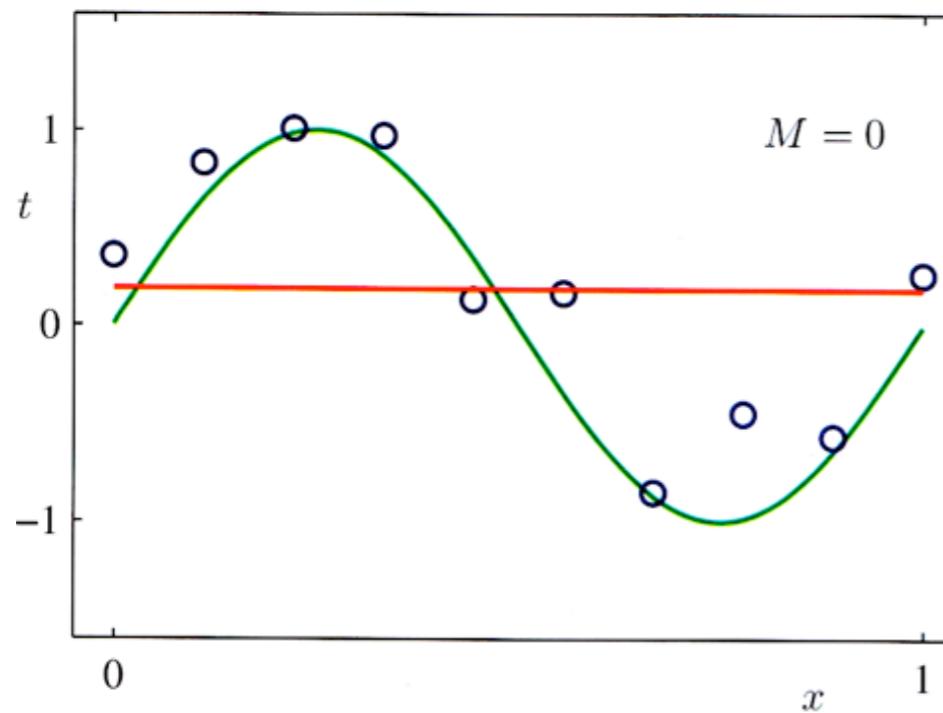


Saturated Model



*Nested model
comparison*

Null Model



Saturated Model

In an ideal world, you want to fit as well as the saturated model, but using no “free” parameters.

More on “nesting”



- Fundamental requirement is that one model is a reduced version of another model
- Reducing means that you clamp the value of some parameter to zero or some null value.
- **The question is if the improvement in fit provided by the extra flexibility is enough to justify the added complexity in the bigger model**

Likelihood of the model given
the data $\mathbf{L} = p(\mathbf{d}|\mathbf{m})$

log Likelihood = ln L

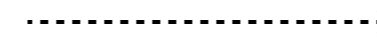
$\ln L_R$

“restricted” model



$\ln L_F$

“full” model



Null model

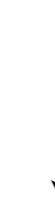


Exemplar model

*Nested model
comparison*



Mixture model



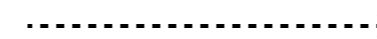
Saturated Model

Likelihood of the model given
the data $\mathbf{L} = p(\mathbf{d}|\mathbf{m})$

log Likelihood = ln L

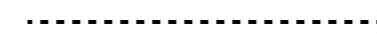
$\ln L_R$

“restricted” model



$\ln L_F$

“full” model



Null model



Exemplar model

*Nested model
comparison*



Mixture model



Saturated Model

$$\ln L_F - \ln L_R$$

difference in fit

$$2(\ln L_F - \ln L_R)$$

Likelihood of the model given
the data $\mathbf{L} = p(\mathbf{d}|\mathbf{m})$

log Likelihood = ln L

$\ln L_R$

“restricted” model

$\ln L_F$

“full” model

$$2(\ln L_F - \ln L_R)$$

god says. (*convenient relationship*
between deviance and χ^2)

Null model

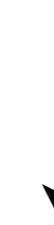
Exemplar model

Mixture model

Saturated Model



*Nested model
comparison*



Likelihood of the model given the data $\mathbf{L} = p(\mathbf{d}|\mathbf{m})$

log Likelihood = ln L

$\ln L_R$

“restricted” model

$\ln L_F$

“full” model

$$G^2 = 2(\ln L_F - \ln L_R) \leftarrow$$

Null model

Exemplar model

Mixture model

Saturated Model

Nested model comparison

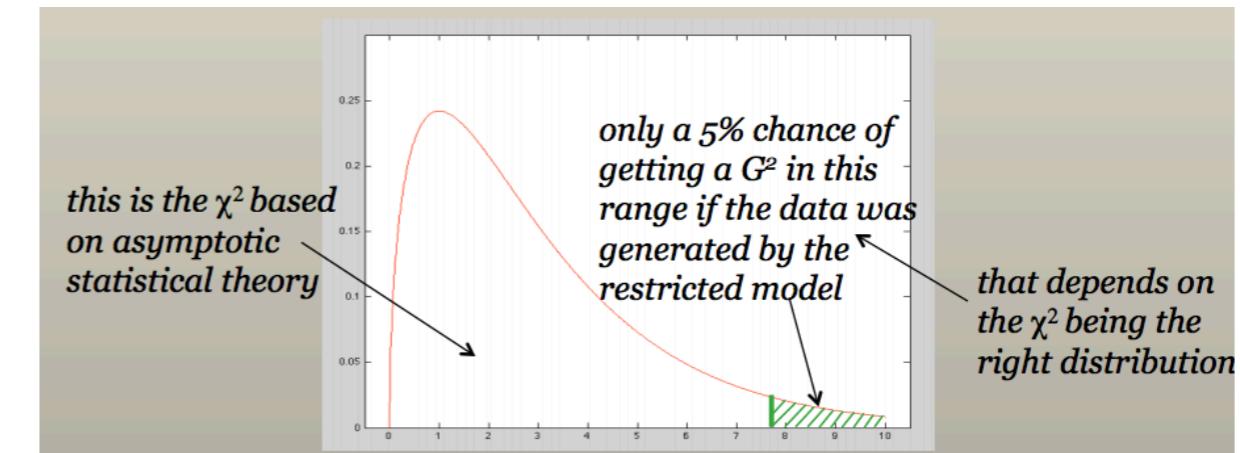
distributed as χ^2
with $N_{paramsFull} - N_{paramsRest}$
degrees of freedom

Likelihood ratio test

$$G^2 = 2(\ln L_F - \ln L_R) \leftarrow$$

*distributed as
with $N_{paramsFull} - N_{paramsRest}$
degrees of freedom*

If G^2 exceeds the critical value in the χ^2 distribution with the specified degrees of freedom and particular alpha-level, then the restricted model is significantly worse than the full model.



Likelihood ratio test

$$G^2 = 2(\ln L_F - \ln L_R) \leftarrow \begin{array}{l} \text{distributed as} \\ \text{with } N\text{paramsFull} - N\text{paramsRest} \\ \text{degrees of freedom} \end{array}$$

Example

$$\ln L_R = -293.12$$

$$df = 240 - 45 = 195$$

$$\ln L_F = -263.12$$

$$G^2 = 2 \times [-263.12 - -293.82] = 61.40$$

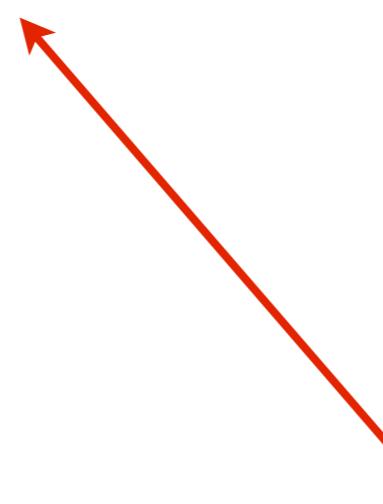
$$X_c^2(df = 195, \alpha = .05) = 228.6$$

the restricted model is NOT significantly worse

Likelihood ratio test

$$G^2 = 2(\ln L_F - \ln L_R) \quad \leftarrow \dots$$

*distributed as
with $NparamsFull - NparamsRest$
degrees of freedom*


$$G^2 = 2 \times \left(\ln \frac{L_F}{L_R} \right)$$

Likelihood ratio test

- **Only applies to nested models**
- Null Hypothesis Testing framework - one model fits better than the other
- Alternatively, we might want to weight the evidence between two models, and to compare non-nested models.

Information criteria

- Akaike's Information Criterion (AIC)

$$AIC = -2\ln L(\theta|u, M) + 2K$$

- Bayesian Information Criterion (BIC)

$$BIC = -2\ln L(\theta|u, M) + K \ln N$$

Information criteria

- Akaike's Information Criterion (AIC)

$$AIC = -2\ln L(\theta|u, M) + 2K$$



Better fit



More parameters

> 0 restricted model fit better

$= 0$... equivalence

< 0 full model fit better

$$AIC_{full} - AIC_{rest}$$

Information criteria

- Akaike's Information Criterion (AIC)

$$AIC_{full} - AIC_{rest}$$

if change is in 0-2 range no support for distinguishing the models

4-7 = “considerably less” support for the model with larger AIC

>10 = no support for the larger AIC model

Information criteria

- Bayesian Information Criterion (BIC)

$$BIC = -2\ln L(\theta|u, M) + K \ln N$$

Related to “Bayes Factor”

$$B = \frac{p(M_1|y)}{p(M_2|y)} = \exp(-\frac{1}{2} \Delta BIC)$$

Similar “range” of values that are considered to be strong support (see Text).

Information criteria

- If you fit both AIC and BIC likely that which model is preferred can actually **conflict!**
- **Then what!?**
- BIC is generally more conservative in that it penalizes models more for extra parameters. Some people prefer it on that basis. Sometimes people report both.

Search algorithm

`fmin`

