Part II: How to make a Bayesian model

Questions you can answer…

- What would an ideal learner or observer infer from these data?
- What are the effects of different assumptions or prior knowledge on this inference?
- What kind of constraints on learning are necessary to explain the inferences people make?
- How do people learn a structured representation?

Marr’s three levels

Computation

“What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?”

Representation and algorithm

“What is the representation for the input and output, and the algorithm for the transformation?”

Implementation

“How can the representation and algorithm be realized physically?”

Six easy steps

Step 1: Find an interesting aspect of cognition

Step 2: Identify the underlying computational problem

Step 3: Identify constraints

Step 4: Work out the optimal solution to that problem, given constraints

Step 5: See how well that solution corresponds to human behavior (do some experiments!)

Step 6: Iterate Steps 2-6 until it works

(Anderson, 1990)

A schema for inductive problems

- What are the data?
  - what information are people learning or drawing inferences from?
- What are the hypotheses?
  - what kind of structure is being learned or inferred from these data?

(these questions are shared with other models)

Thinking generatively…

- How do the hypotheses generate the data?
  - defines the likelihood $p(d|h)$
- How are the hypotheses generated?
  - defines the prior $p(h)$
  - while the prior encodes information about knowledge and learning biases, translating this into a probability distribution can be made easier by thinking in terms of a generative process…
- Bayesian inference inverts this generative process

An example: Speech perception

(with thanks to Naomi Feldman)

An example: Speech perception

Speaker chooses
a phonetic category

An example: Speech perception

Speaker chooses
a phonetic category

Speaker articulates a
"target production"

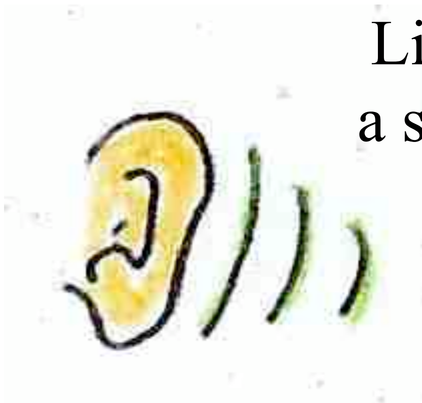
An example: Speech perception

Noise in the
speech signal

Speaker chooses
a phonetic category

Speaker articulates a
"target production"

An example: Speech perception



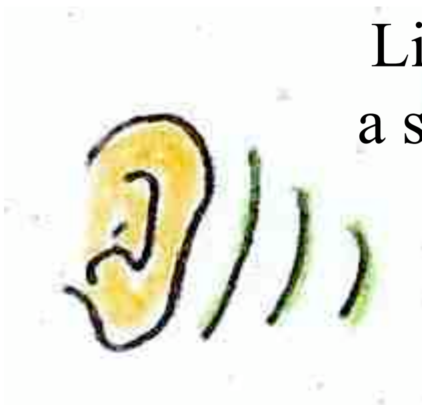
Listener hears
a speech sound

Noise in the
speech signal

Speaker chooses
a phonetic category

Speaker articulates a
"target production"

An example: Speech perception



Listener hears
a speech sound

S

Noise in the
speech signal

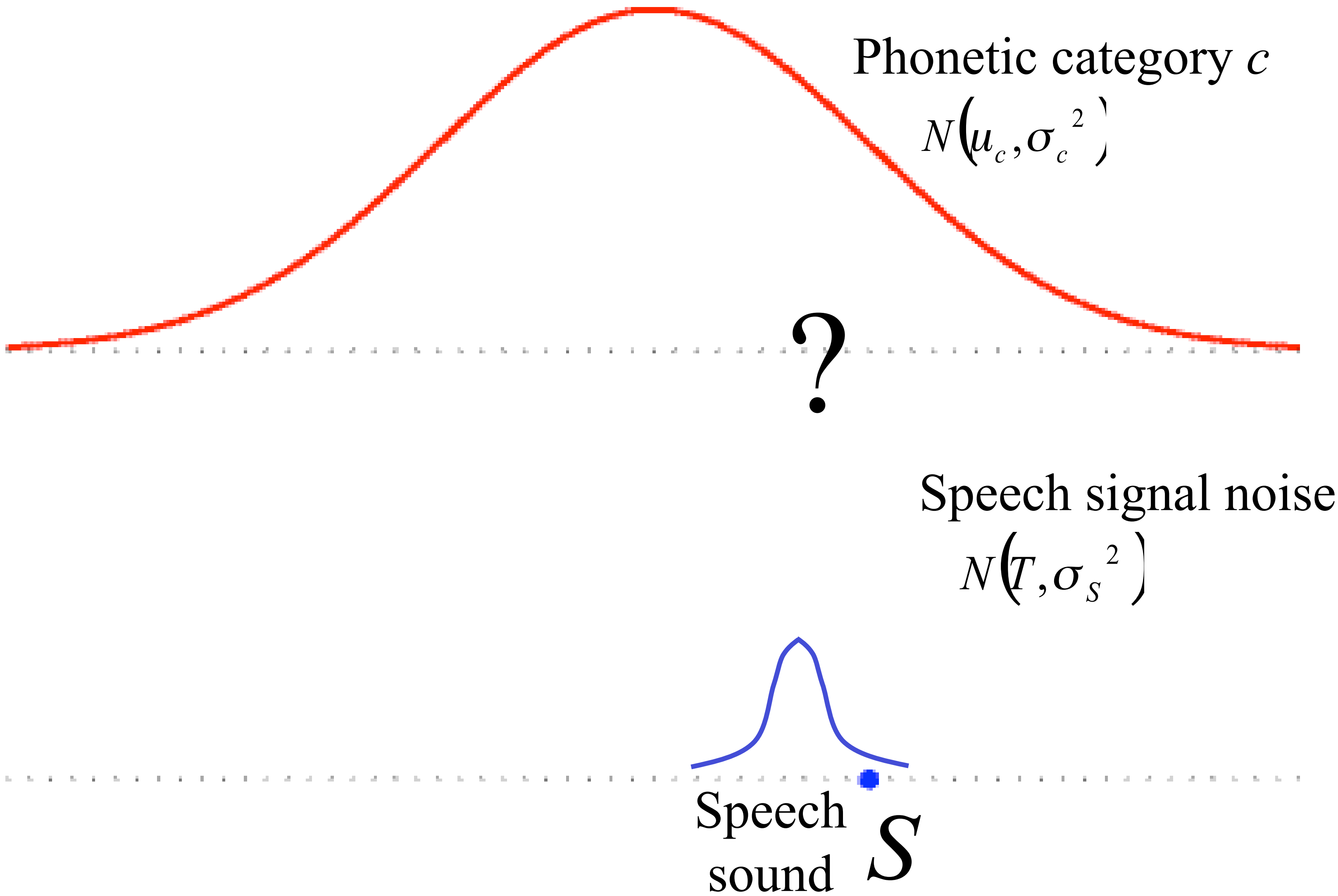
c

Speaker chooses
a phonetic category

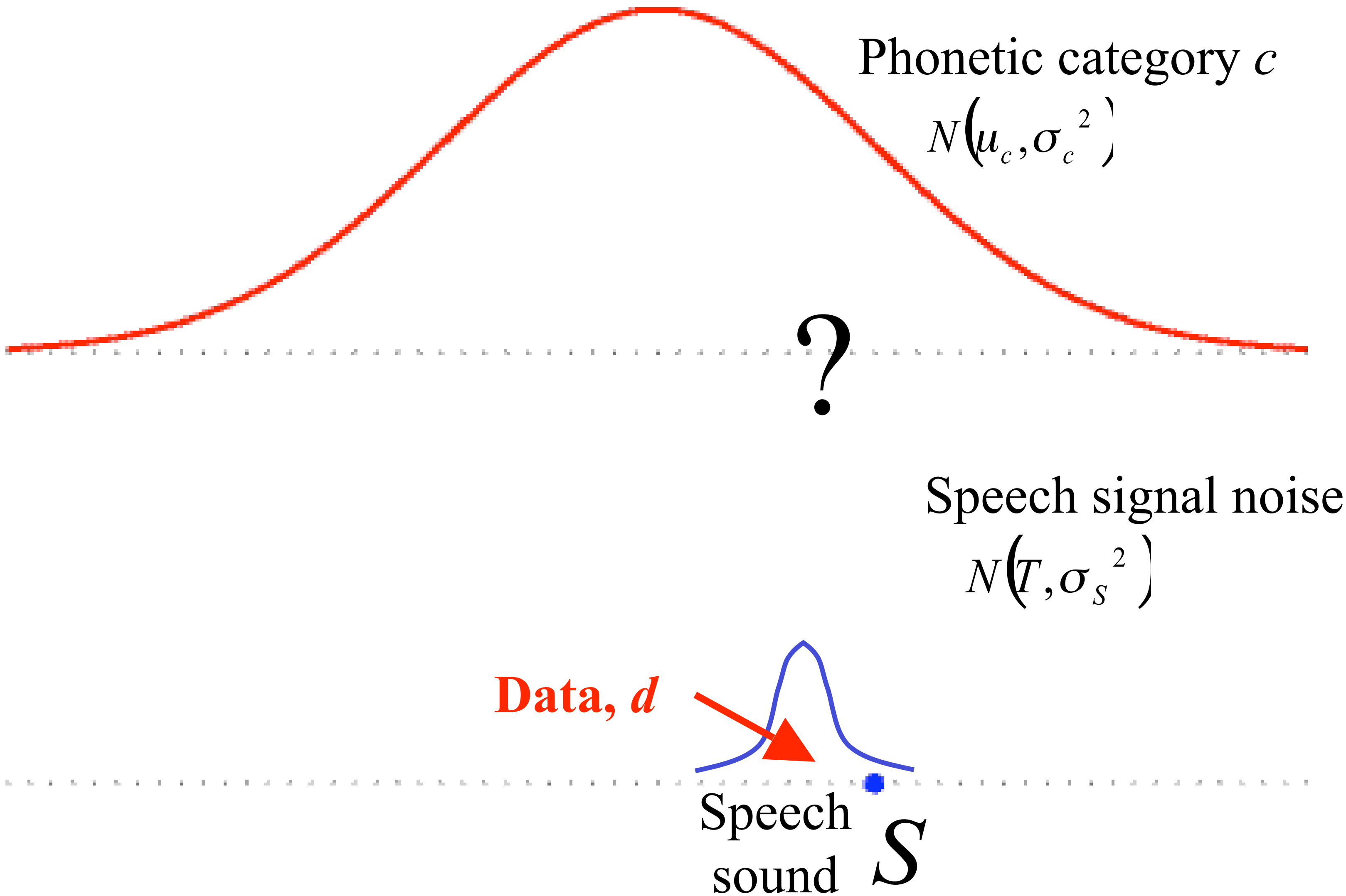
T

Speaker articulates a
"target production"

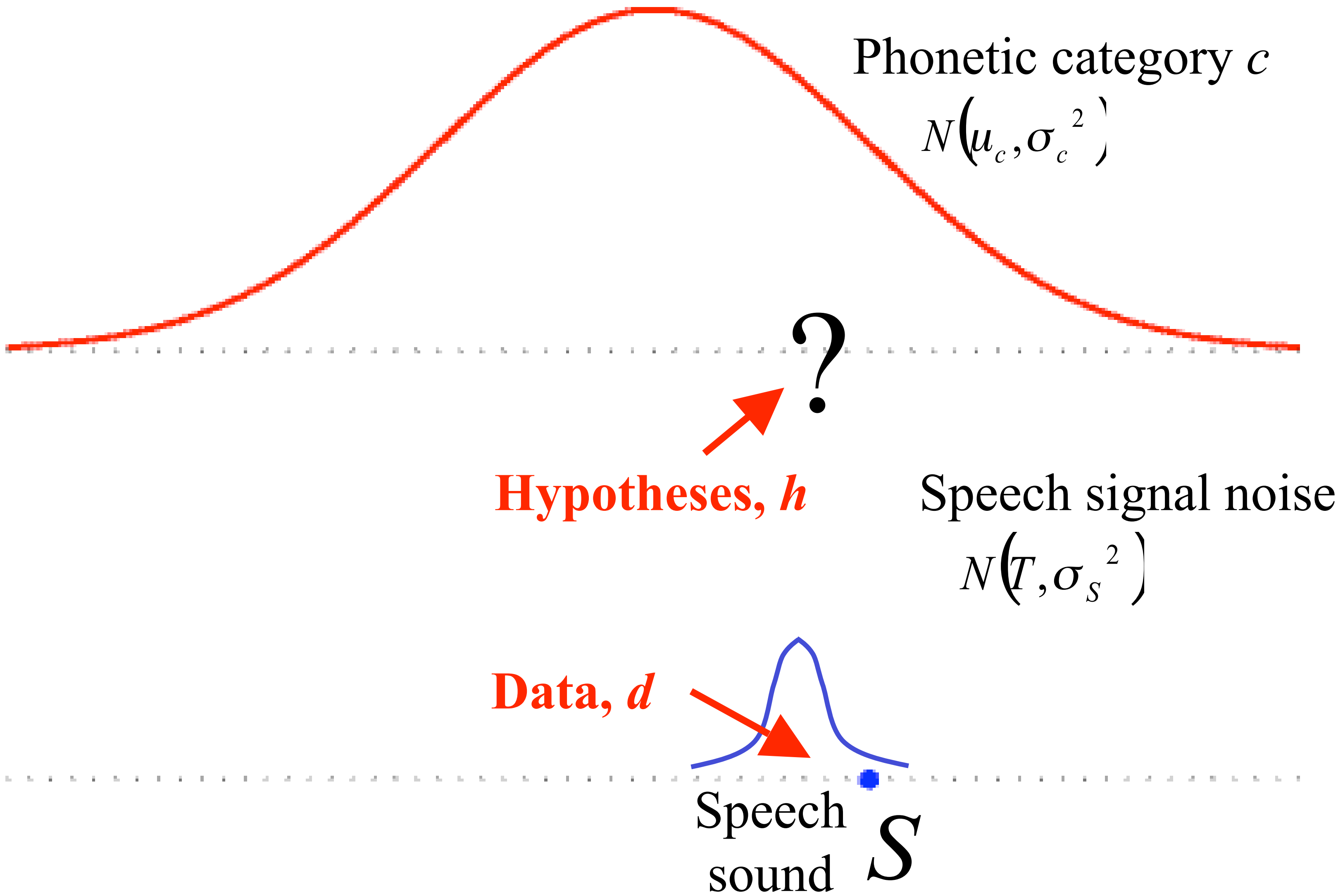
Bayes for speech perception



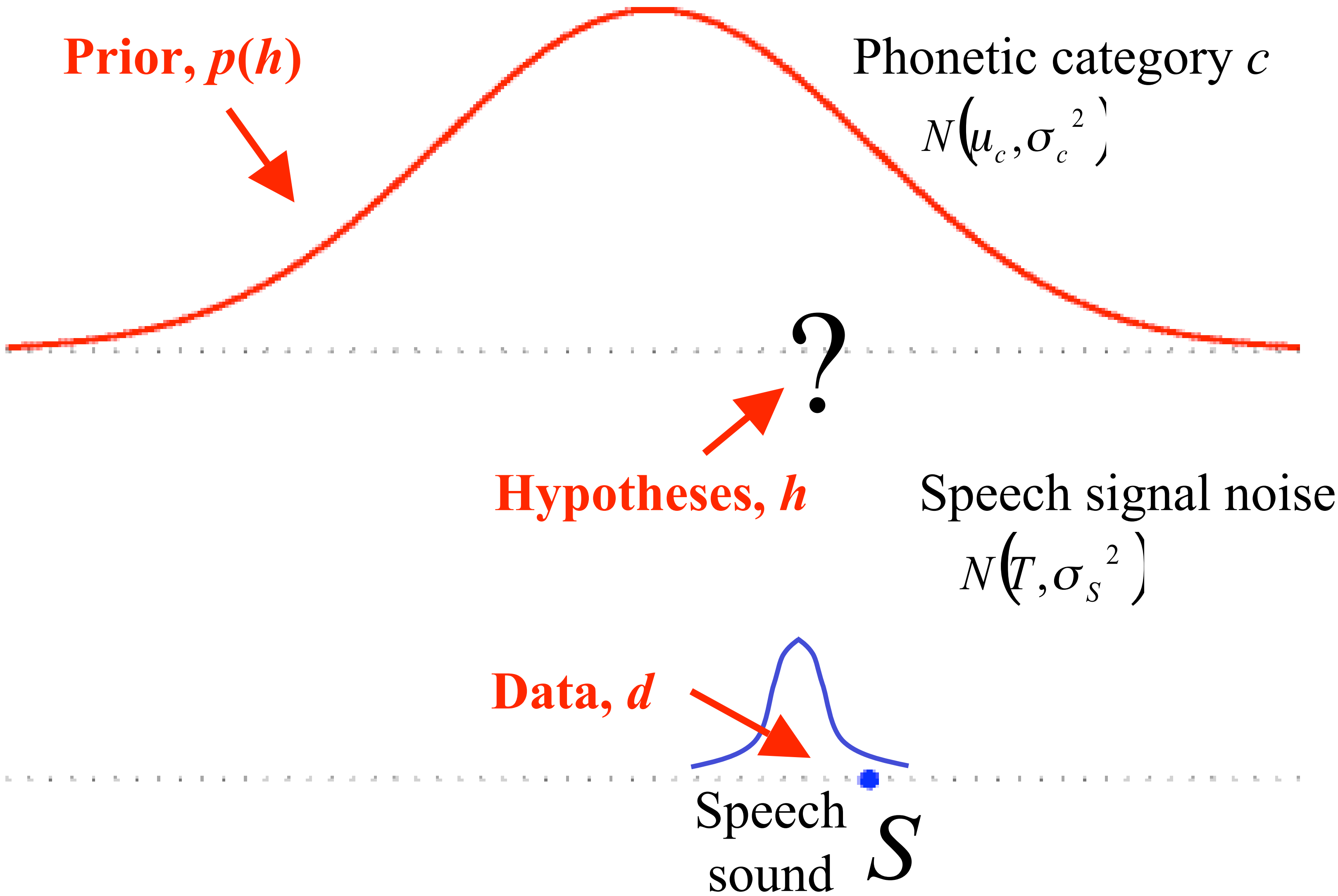
Bayes for speech perception



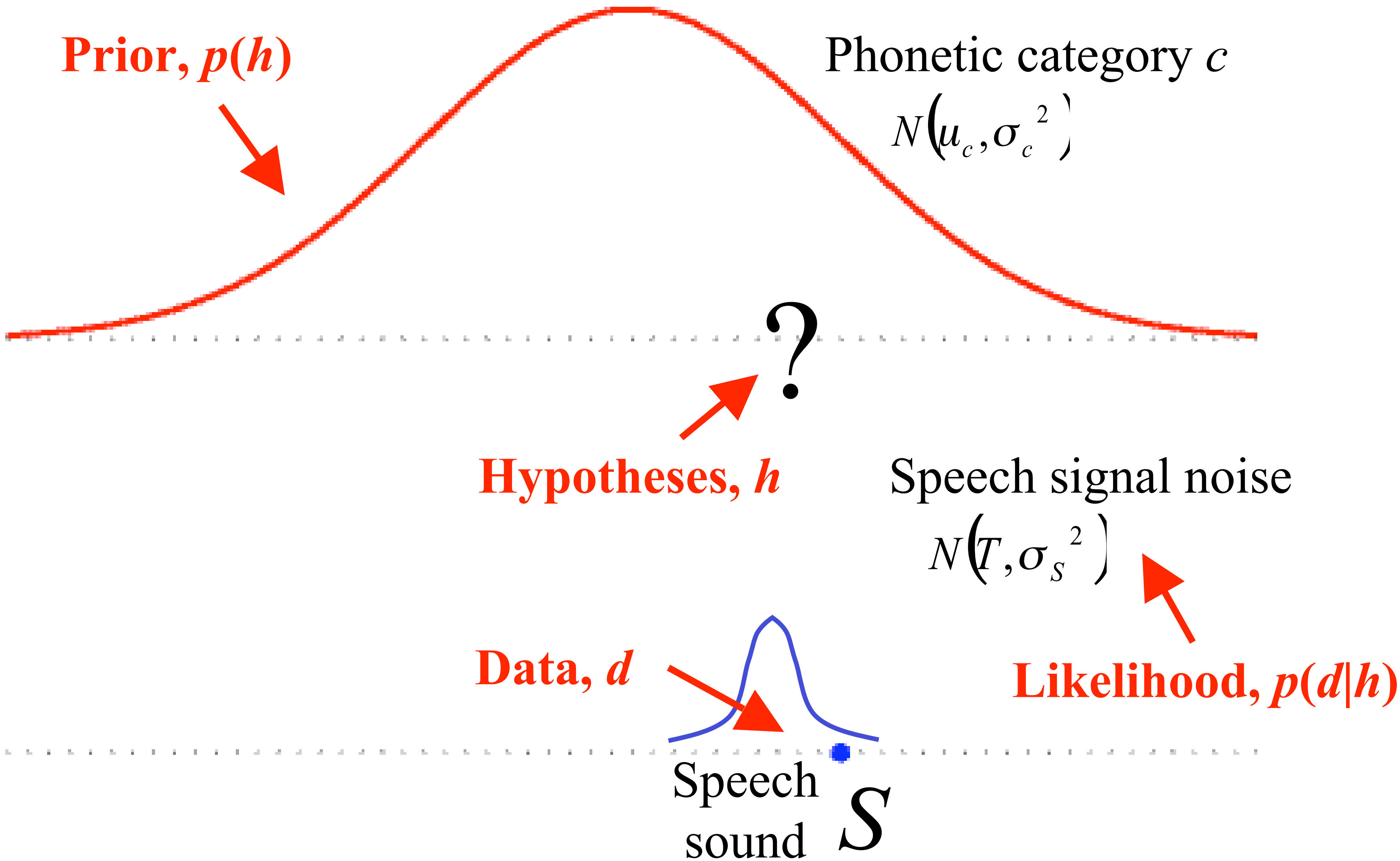
Bayes for speech perception



Bayes for speech perception



Bayes for speech perception



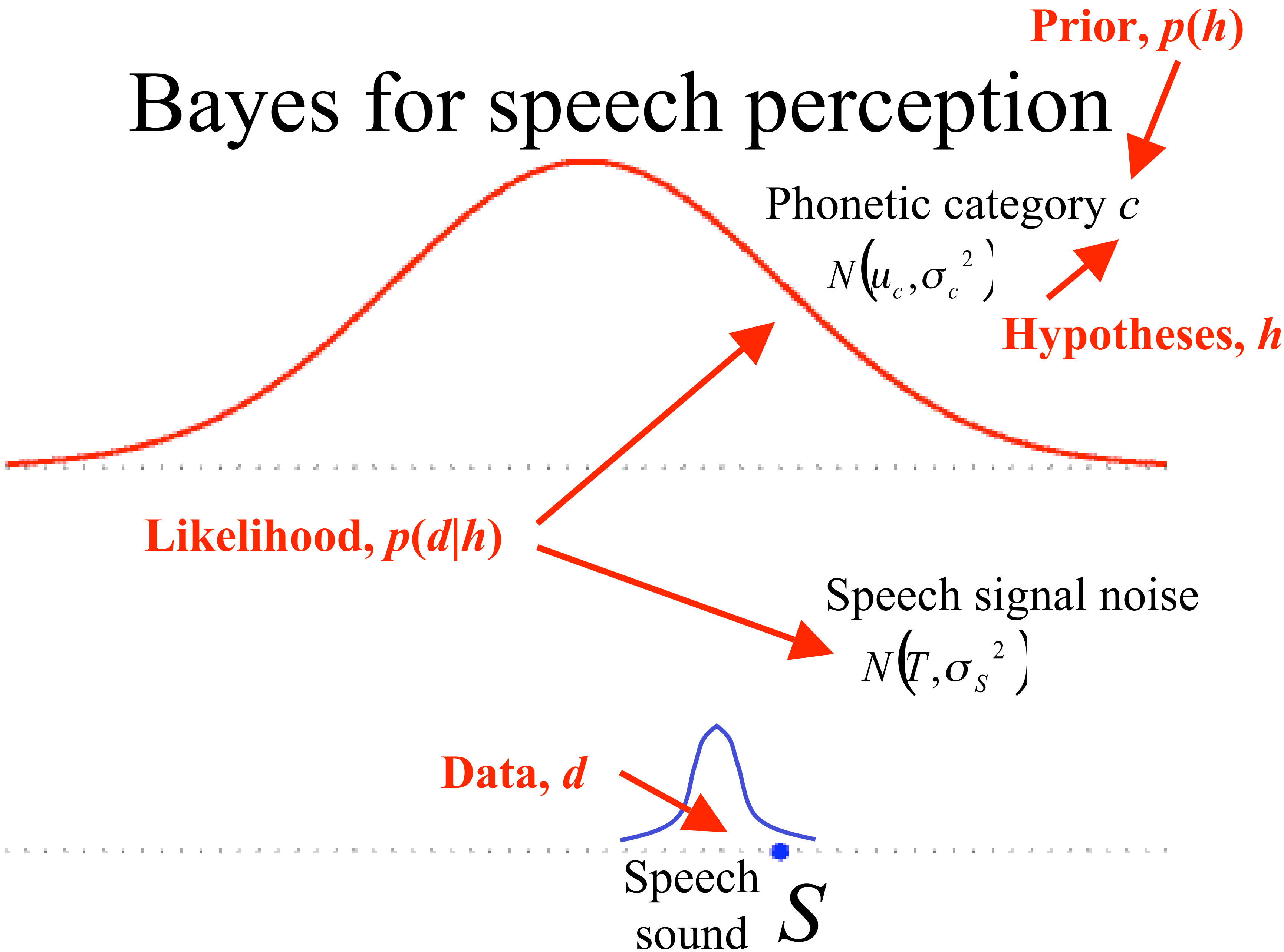
Bayes for speech perception

Listeners must invert the process that generated the sound they heard…

- data (d): speech sound S
- hypotheses (h): target productions T
- prior ($p(h)$): phonetic category structure $p(T|c)$
- likelihood ($p(d|h)$): speech signal noise $p(S|T)$

$$p(h \mid d) \propto p(d \mid h)\,p(h)$$

Bayes for speech perception



Bayes for speech perception

Listeners must invert the process that generated the sound they heard…

- data (d): speech sound S
- hypotheses (h): phonetic category c
- prior ($p(h)$): probability of category $p(c)$
- likelihood ($p(d|h)$): combination of category variability $p(T|c)$ and speech signal noise $p(S|T)$

$$p(S \mid c) = \int p(S \mid T)\, p(T \mid c)\, dT$$

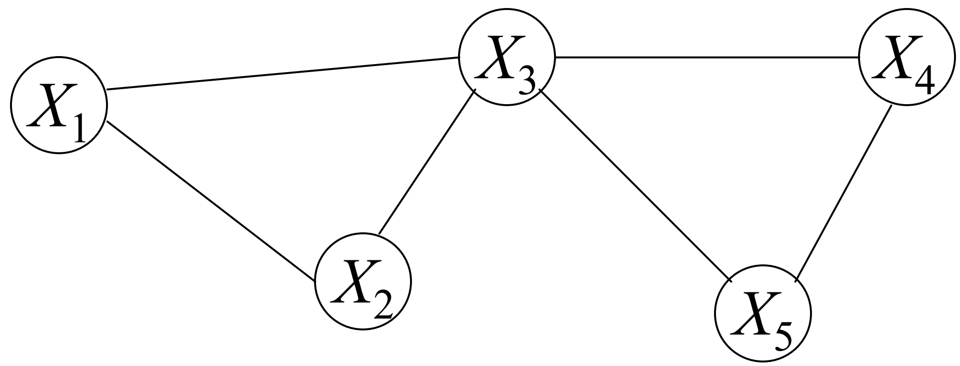
Challenges of generative models

- Specifying well-defined probabilistic models involving many variables is hard
- Representing probability distributions over those variables is hard, since distributions need to describe all possible states of the variables
- Performing Bayesian inference using those distributions is hard

Graphical models

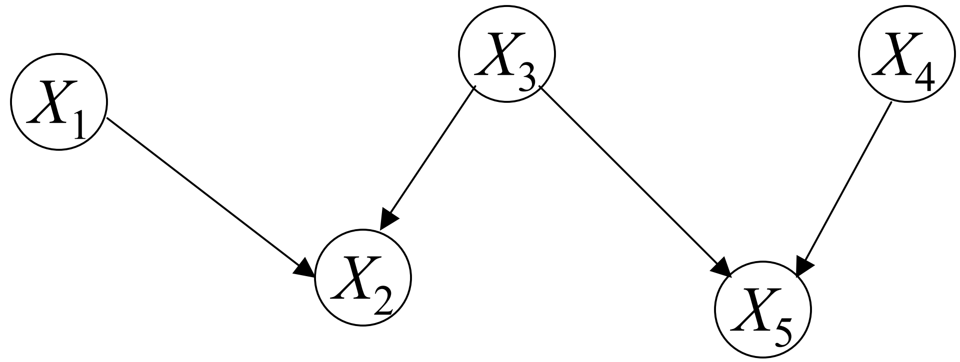
- Express the probabilistic dependency structure among a set of variables (Pearl, 1988)
- Consist of
  - a set of nodes, corresponding to variables
  - a set of edges, indicating dependency
  - a set of functions defined on the graph that specify a probability distribution

Undirected graphical models



- Consist of
  - a set of nodes
  - a set of edges
  - a *potential* for each *clique*, multiplied together to yield the distribution over variables
- Examples
  - statistical physics: Ising model, spinglasses
  - early neural networks (e.g. Boltzmann machines)

Directed graphical models



- Consist of
  - a set of nodes
  - a set of edges
  - a *conditional probability distribution* for each node, conditioned on its parents, multiplied together to yield the distribution over variables
- Constrained to directed acyclic graphs (DAGs)
- Called Bayesian networks or Bayes nets

Statistical independence

- Two random variables X_1 and X_2 are *independent* if $P(x_1|x_2) = P(x_1)$
  - e.g. coinflips: $P(x_1=\mathtt{H}|x_2=\mathtt{H}) = P(x_1=\mathtt{H}) = 0.5$
- Independence makes it easier to represent and work with probability distributions
- We can exploit the product rule:

$$P(x_1,x_2,x_3,x_4) = P(x_1 | x_2,x_3,x_4)P(x_2 | x_3,x_4)P(x_3 | x_4)P(x_4)$$

If x_1, x_2, x_3, and x_4 are all independent…

$$P(x_1,x_2,x_3,x_4) = P(x_1)P(x_2)P(x_3)P(x_4)$$

The Markov assumption

Every node is conditionally independent of its non-descendants, given its parents

$$P(x_i \mid x_{i+1},\ldots,x_k) = P(x_i \mid \mathbf{Pa}(X_i))$$

where $\mathbf{Pa}(X_i)$ is the set of parents of X_i

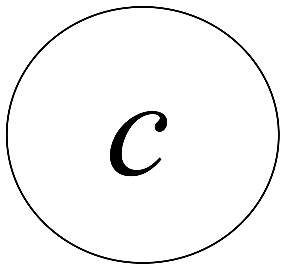
$$P(x_1,\ldots,x_k) = \prod_{i=1}^{k} P(x_i \mid \mathbf{Pa}(X_i))$$

(via the product rule)

Representing generative models

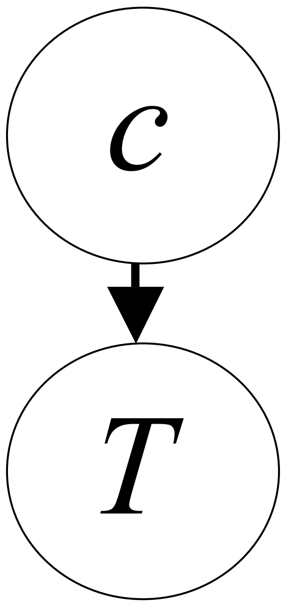
- Graphical models provide solutions to many of the challenges of probabilistic models
  - defining structured distributions
  - representing distributions on many variables
  - efficiently computing probabilities
- Graphical models also provide an intuitive way to define generative processes…

Graphical model for speech



Choose a category c with probability $p(c)$

Graphical model for speech

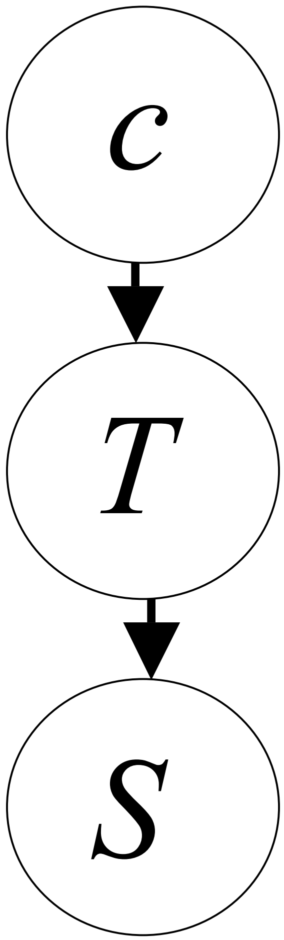


Choose a category *c* with probability *p*(*c*)

Articulate a target production *T* with probability *p*(*T*|*c*)

$$p(T \mid c) = N(\mu_c, \sigma_c^{\,2})$$

Graphical model for speech



Choose a category c with probability $p(c)$

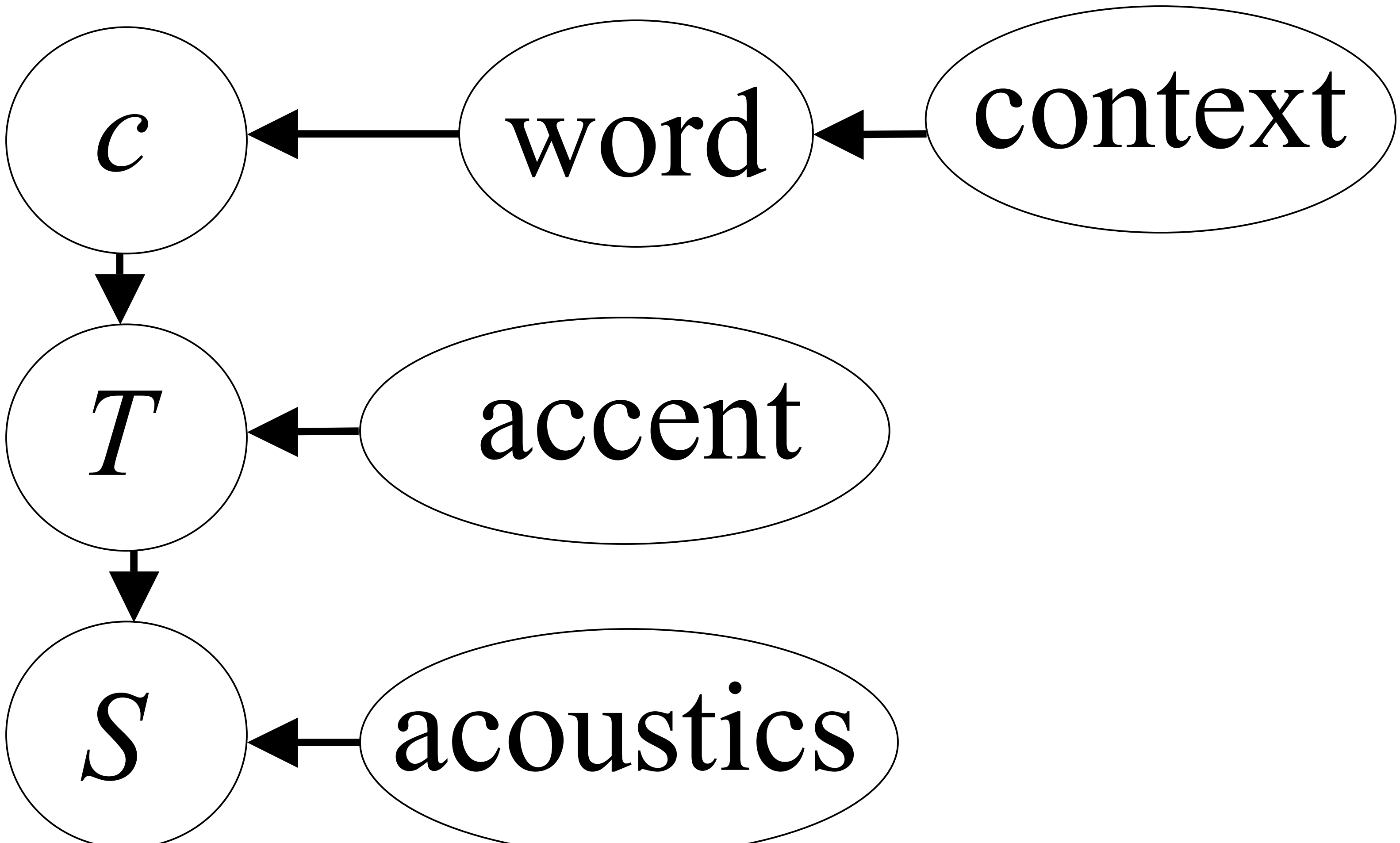
Articulate a target production T with probability $p(T|c)$

$$p(T \mid c) = N(\mu_c, \sigma_c^2)$$

Listener hears speech sound S with probability $p(S|T)$

$$p(S \mid T) = N(T, \sigma_S^2)$$

Graphical model for speech



Performing Bayesian calculations

- Having defined a generative process you are ready to invert that process using Bayes' rule

- Different models and modeling goals require different methods…
  - mathematical analysis
  - special-purpose computer programs
  - general-purpose computer programs

Mathematical analysis

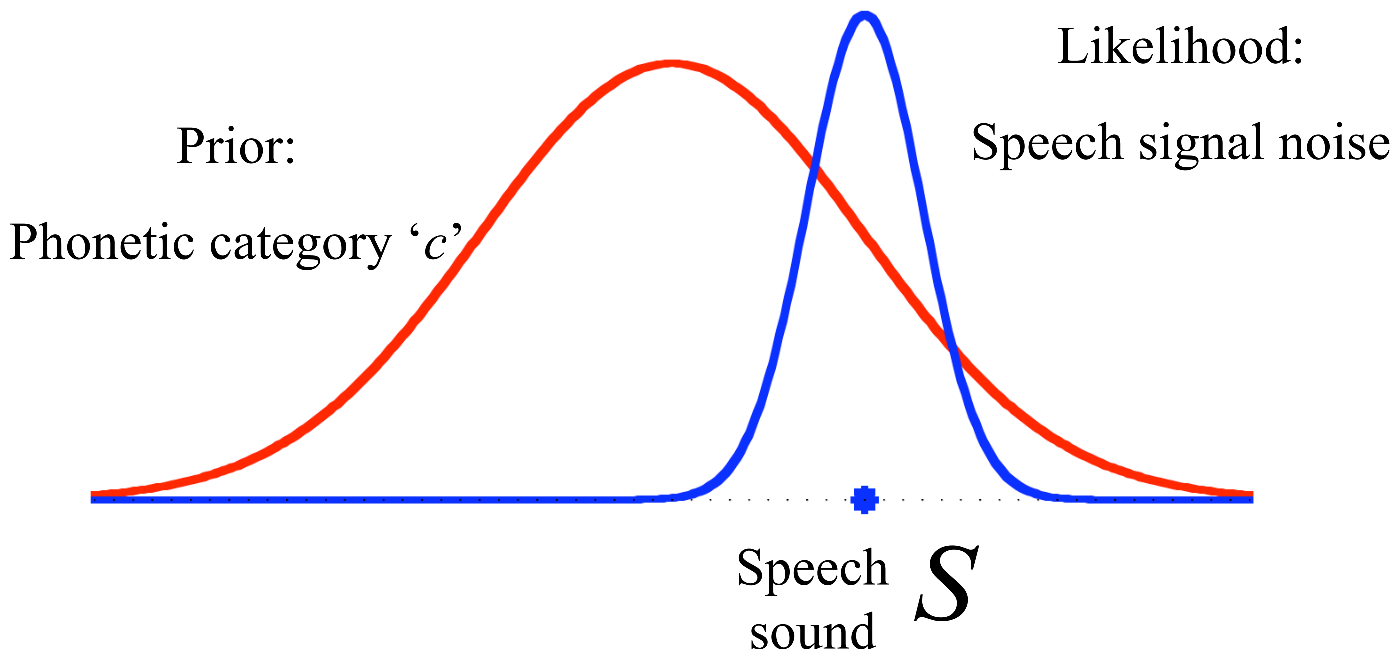
- Work through Bayes' rule by hand
  - the only option available for a long time!
- Suitable for simple models using a small number of hypotheses and/or conjugate priors

One phonetic category

Bayes' rule: $$p(T \mid S) \propto p(S \mid T)\,p(T)$$

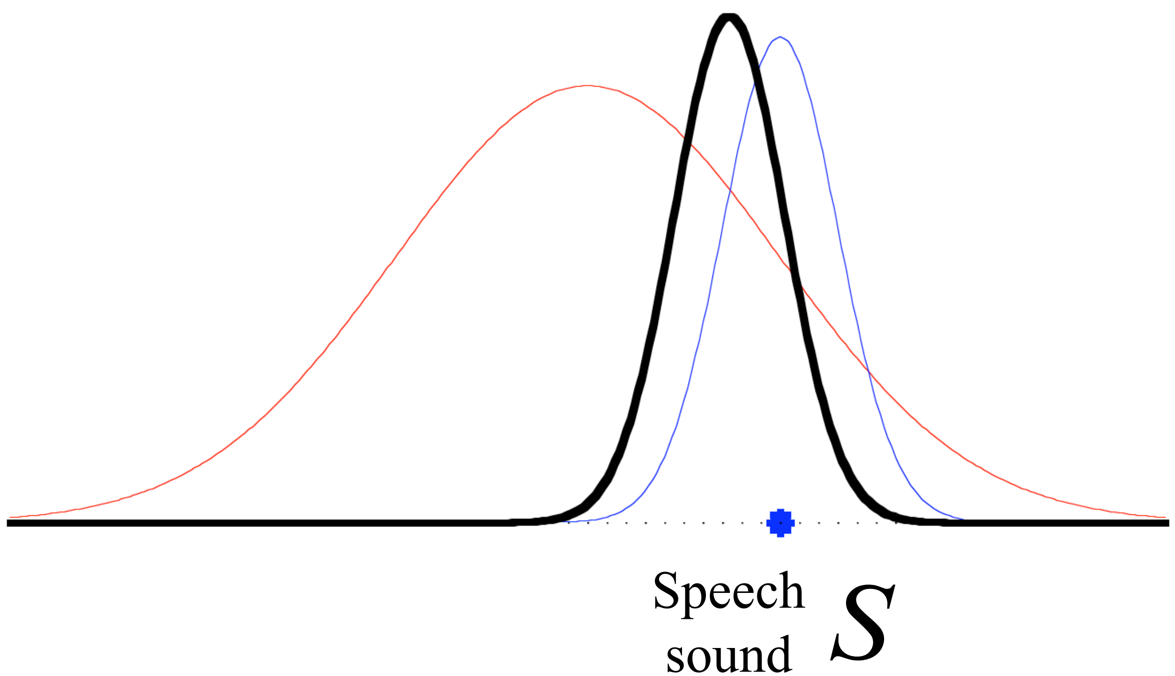
One phonetic category

Bayes' rule: $p(T \mid S) \propto p(S \mid T)p(T)$



One phonetic category

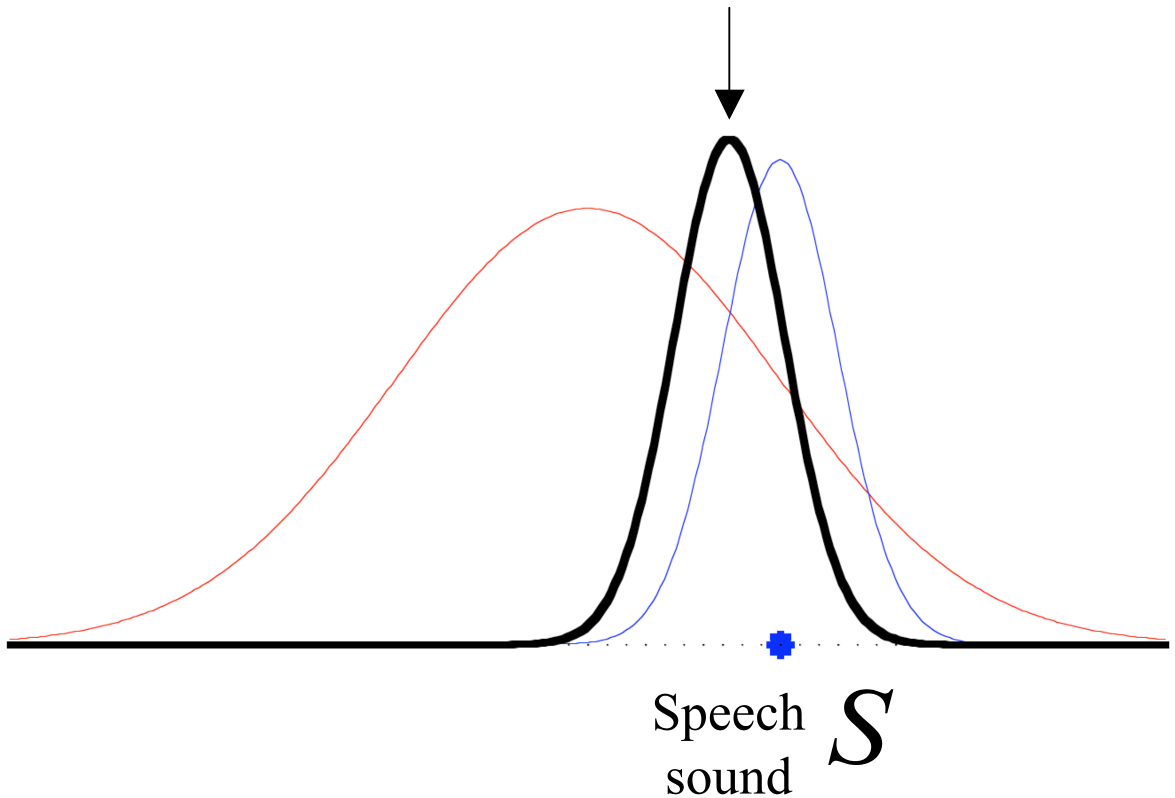
This can be simplified to a Gaussian distribution:



One phonetic category

Which has the expectation (mean):

$$E[T \mid S] = \frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2}$$

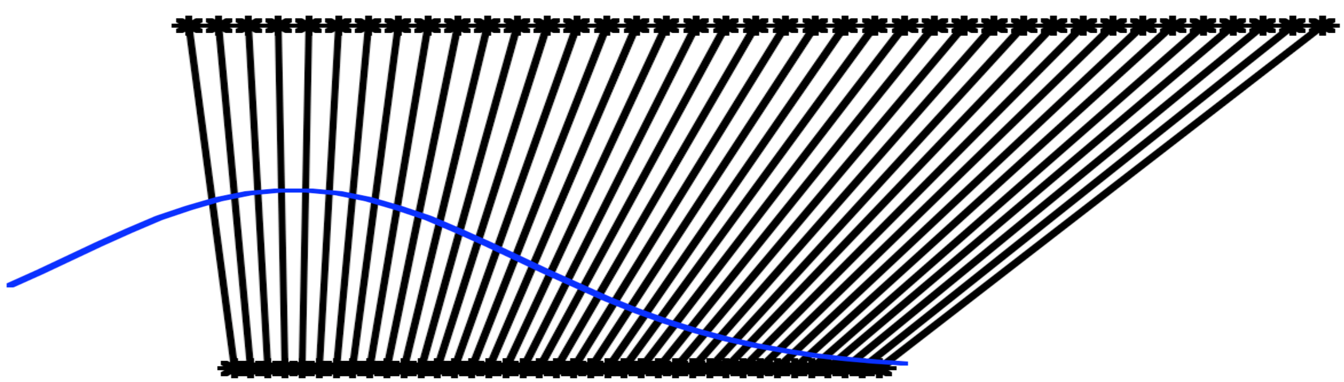


Perceptual warping

Perception of speech sounds is pulled toward the mean of the phonetic category

(shrinks perceptual space)

Actual stimulus



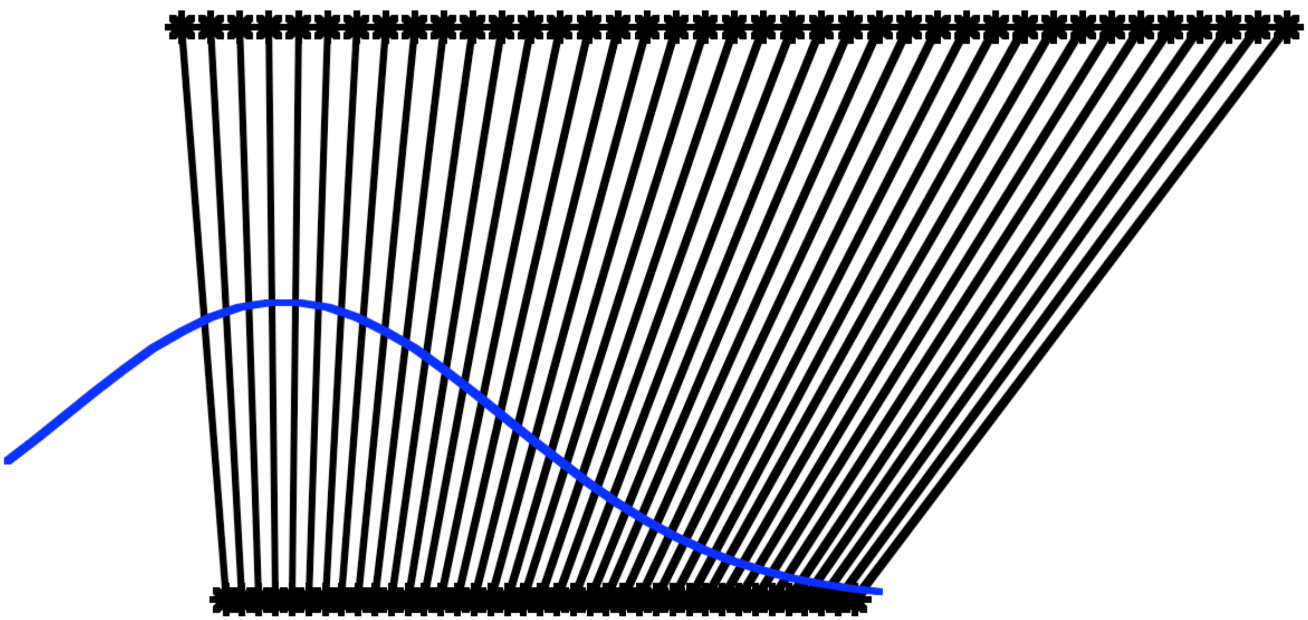
Perceived stimulus

Mathematical analysis

- Work through Bayes' rule by hand
  - the only option available for a long time!
- Suitable for simple models using a small number of hypotheses and/or conjugate priors
- Can provide conditions on conclusions or determine the effects of assumptions
  - e.g. perceptual magnet effect

Perceptual warping

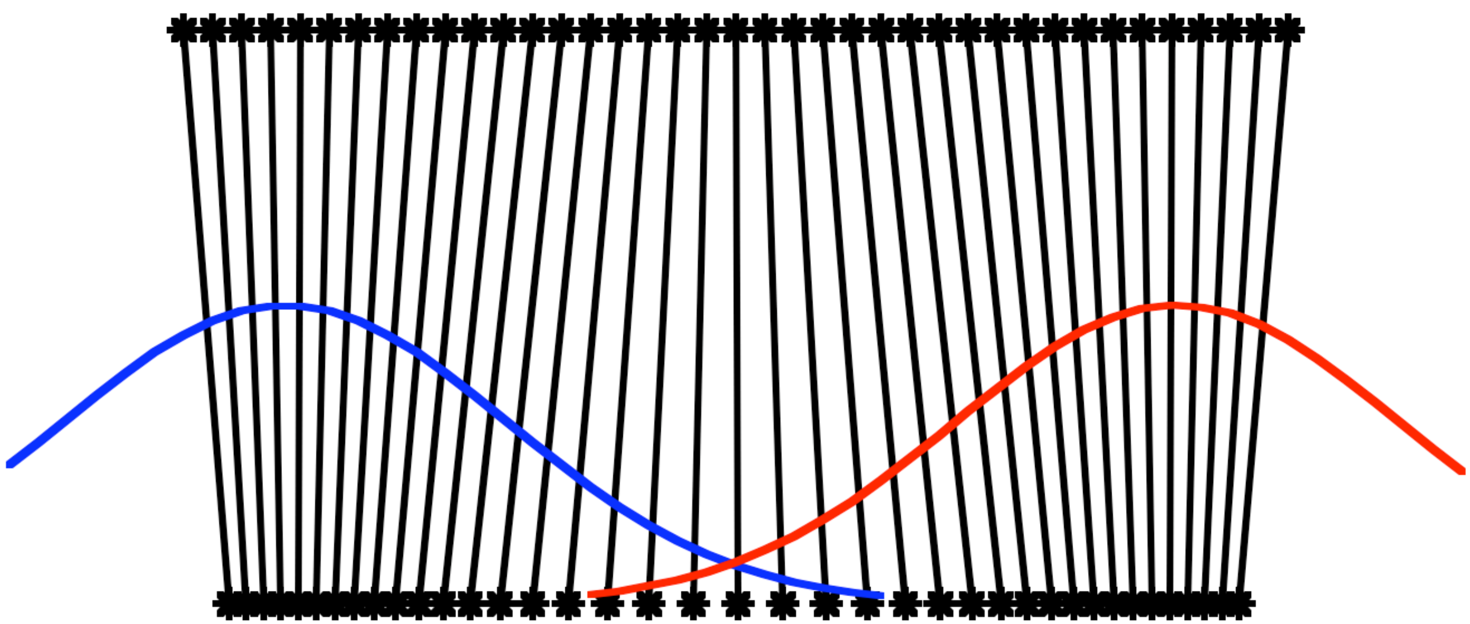
Actual stimulus



Perceived stimulus

Perceptual warping

Actual stimulus



Perceived stimulus

Characterizing perceptual warping

$$\frac{d}{dS}E[T \mid S] = \frac{d}{dS}p(c = 1 \mid S)\frac{\sigma_S^{\ 2}(\mu_1 - \mu_2)}{\sigma_c^{\ 2} + \sigma_S^{\ 2}} + \frac{\sigma_c^{\ 2}}{\sigma_c^{\ 2} + \sigma_S^{\ 2}}$$

Mathematical analysis

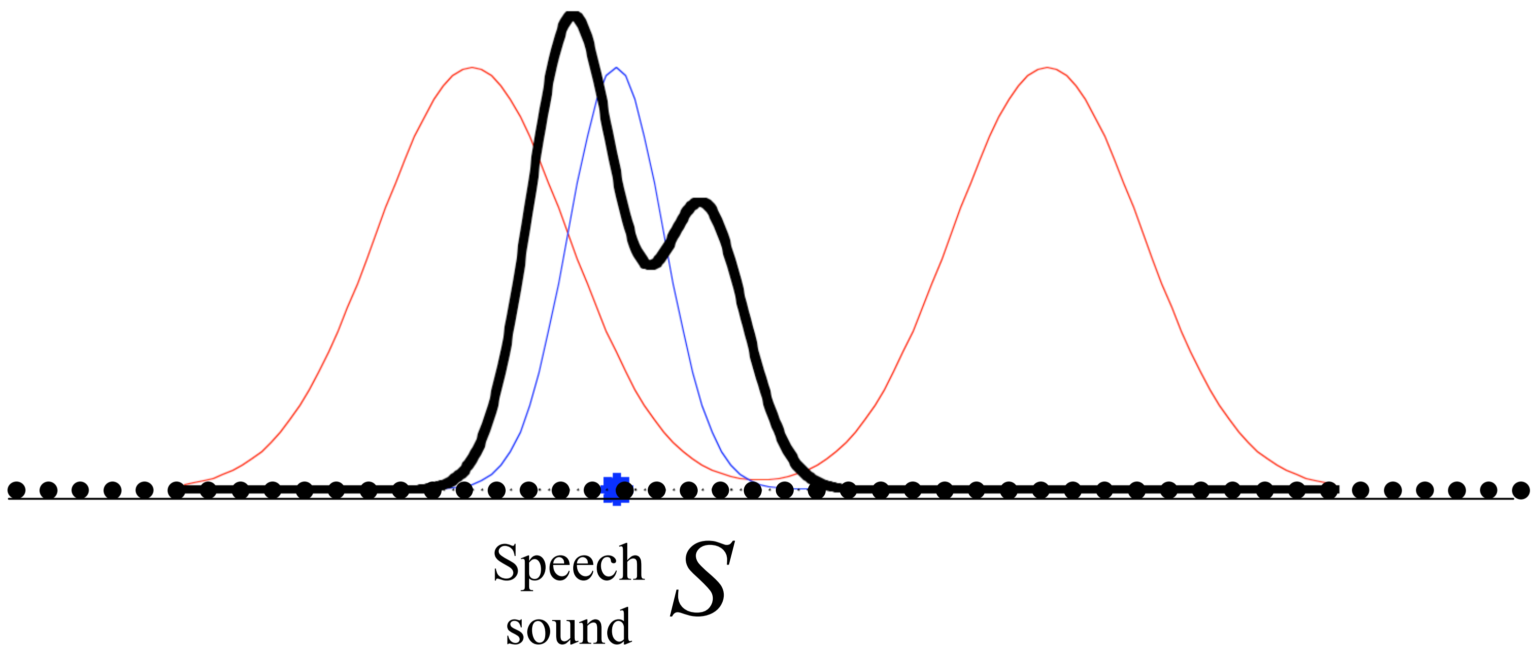
- Work through Bayes' rule by hand
  - the only option available for a long time!
- Suitable for simple models using a small number of hypotheses and/or conjugate priors
- Can provide conditions on conclusions or determine the effects of assumptions
  - e.g. perceptual magnet effect
- Lots of useful math: calculus, linear algebra, stochastic processes, …

Special-purpose computer programs

- Some models are best analyzed by implementing tailored numerical algorithms
- Bayesian inference for low-dimensional continuous hypothesis spaces (e.g.the perceptual magnet effect) can be approximated discretely

multiply $p(d|h)$ and $p(h)$ at each site
normalize over vector

Multiple phonetic categories

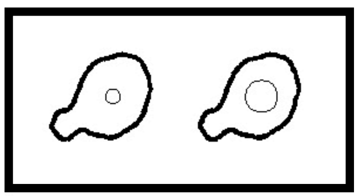


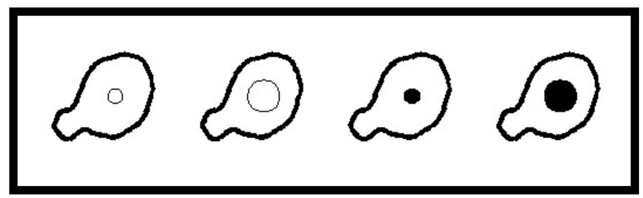
Special-purpose computer programs

- Some models are best analyzed by implementing tailored numerical algorithms
- Bayesian inference for large discrete hypothesis spaces (e.g. concept learning) can be implemented efficiently using matrices

Bayesian concept learning

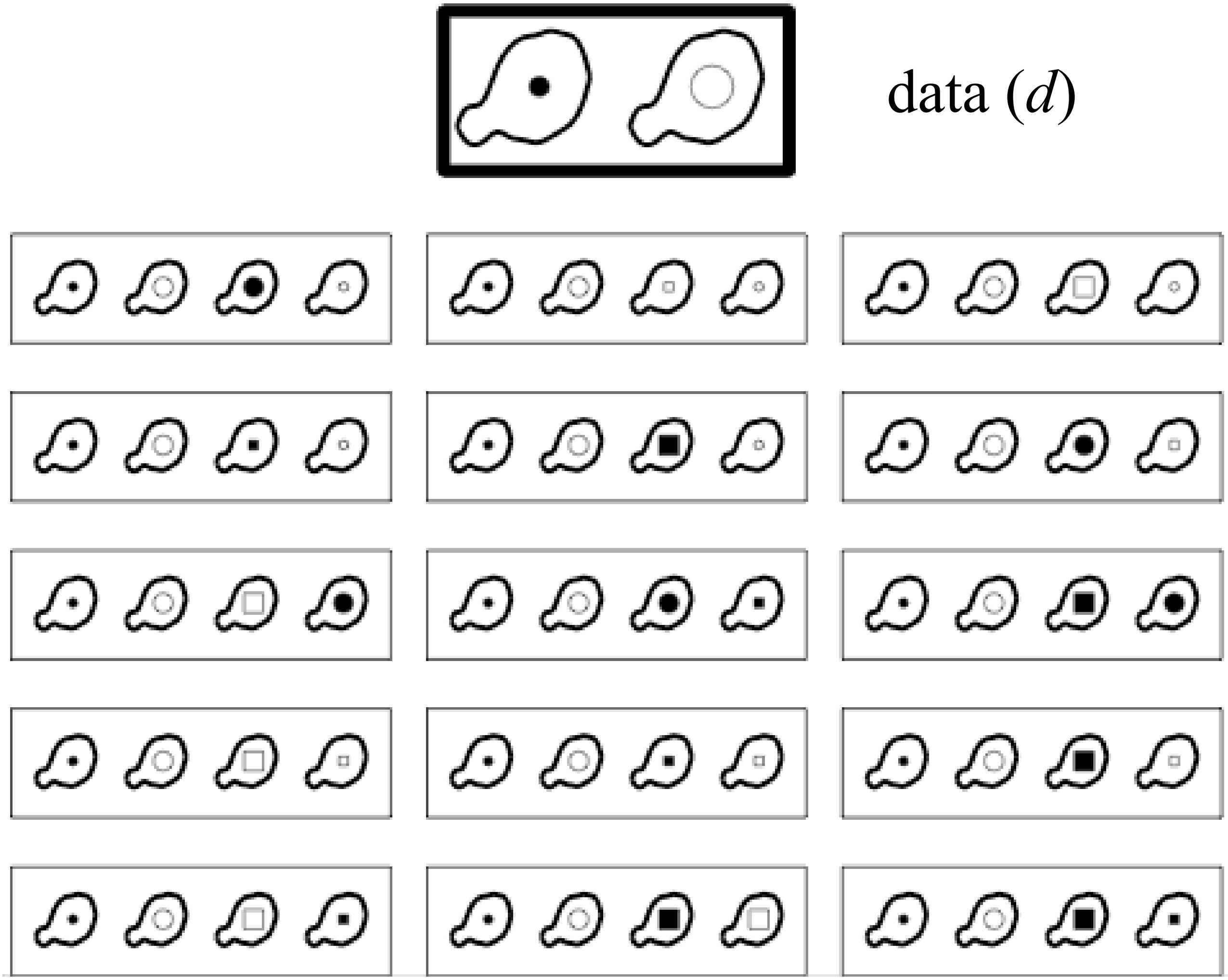
data **hypotheses**





What rule describes the species that these amoebae belong to?

Concept learning experiments



data (d)

hypotheses (h)

Bayesian model

(Tenenbaum, 1999; Tenenbaum & Griffiths, 2001)

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h' \in H} P(d \mid h')P(h')}$$

d: 2 amoebae

h: set of 4 amoebae

$$P(d \mid h) = \begin{cases} 1/|h|^m & d \in h \\ 0 & \text{otherwise} \end{cases}$$

m: # of amoebae in the set *d* (= 2)

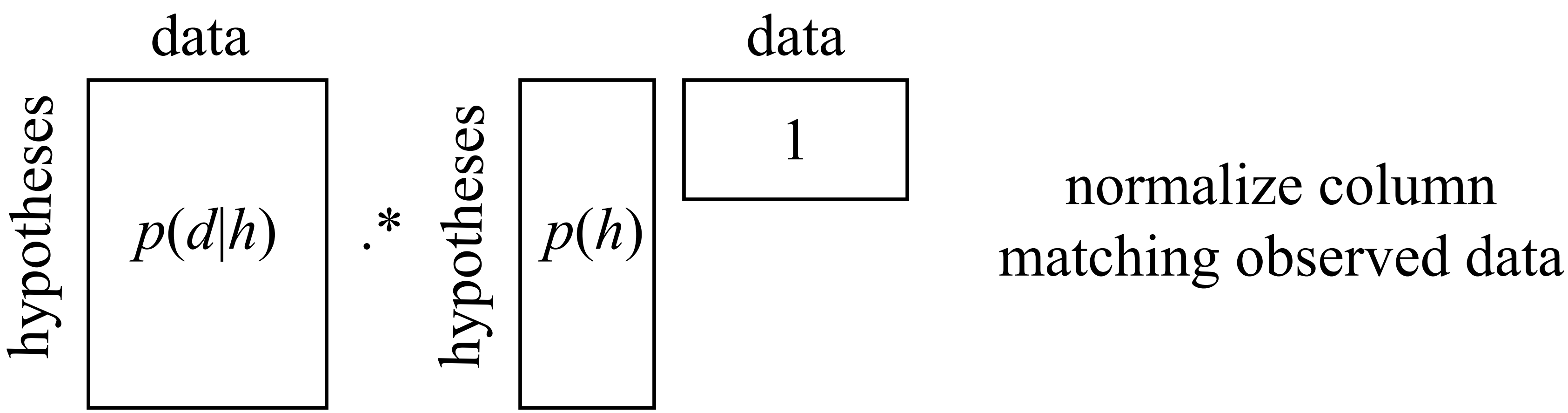
|*h*|: # of amoebae in the set *h* (= 4)

$$P(h \mid d) = \frac{P(h)}{\sum_{h' \mid d \in h'} P(h')}$$

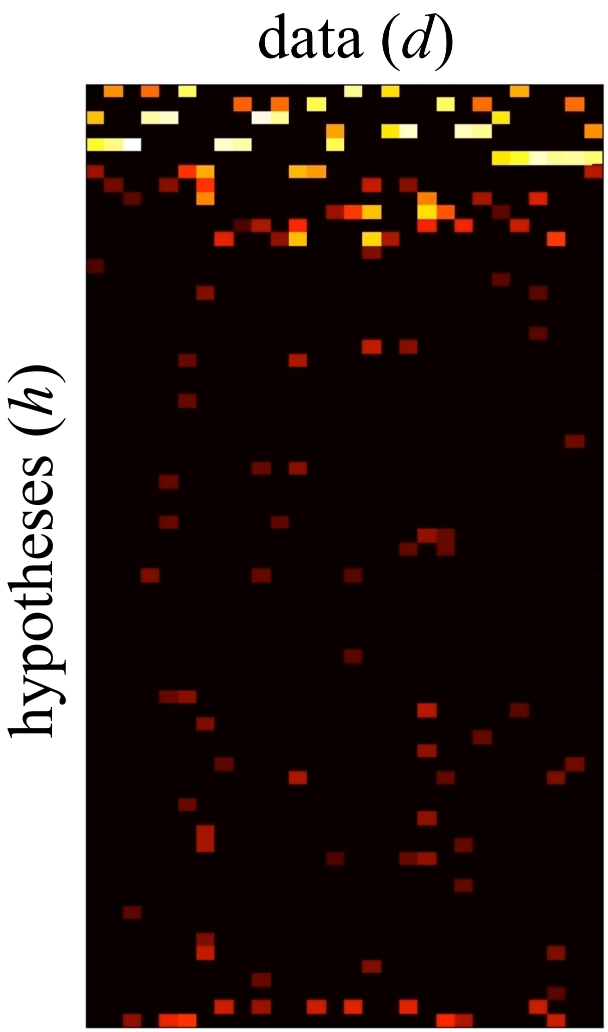
Posterior is renormalized prior

Special-purpose computer programs

- Some models are best analyzed by implementing tailored numerical algorithms
- Bayesian inference for large discrete hypothesis spaces (e.g. concept learning) can be implemented efficiently using matrices

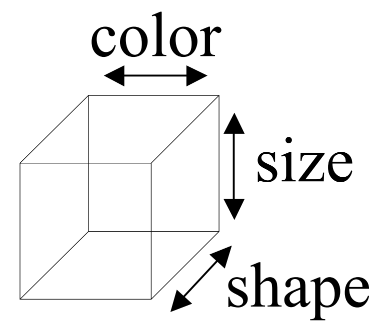


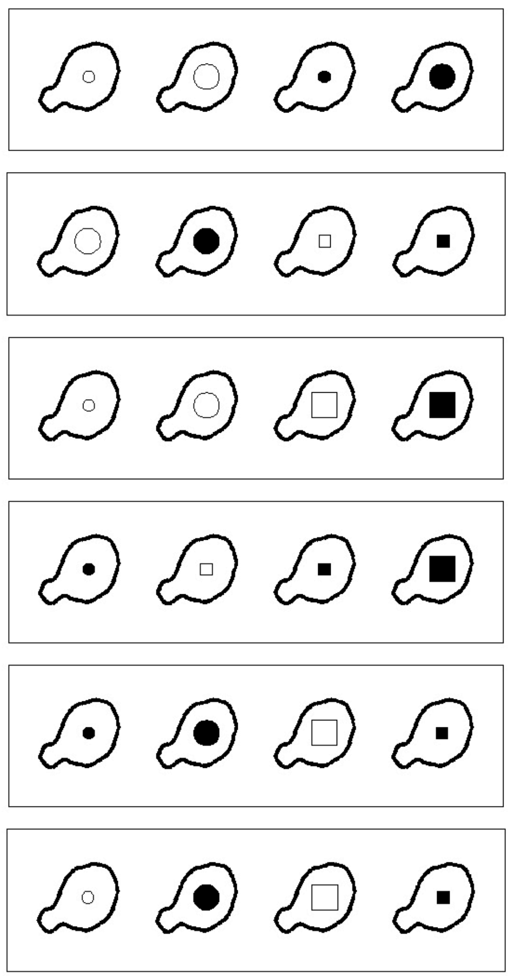
Fitting the model



Classes of concepts

(Shepard, Hovland, & Jenkins, 1961)





Class 1

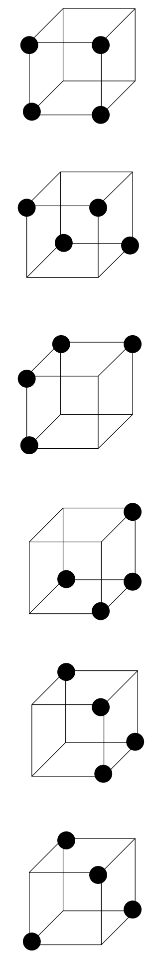
Class 2

Class 3

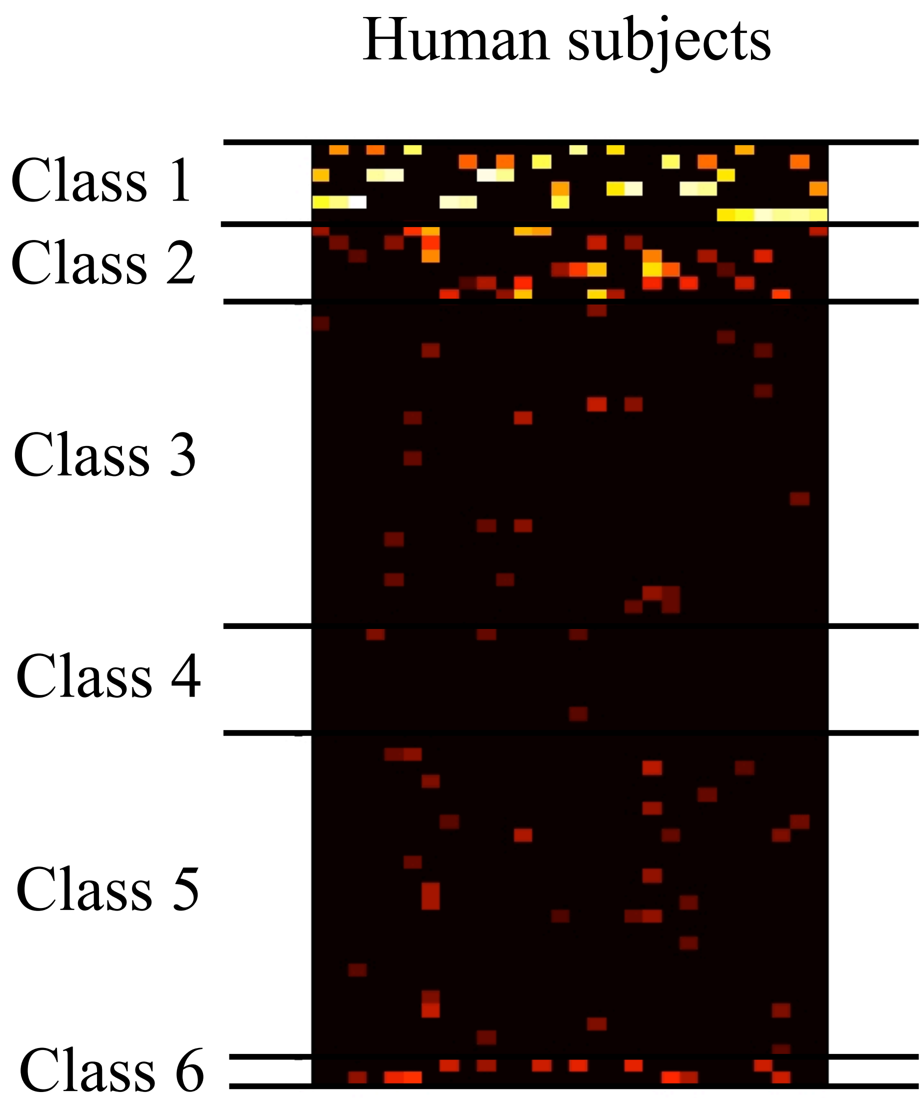
Class 4

Class 5

Class 6



Fitting the model



Special-purpose computer programs

- Some models are best analyzed by implementing tailored numerical algorithms
- Another option is Monte Carlo approximation…
- The expectation of f with respect to p can be approximated by

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

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

General-purpose computer programs

- A variety of software packages exist for performing Bayesian computations
  - Bayes Net Toolbox for Matlab
  - BUGS (Bayesian inference Using Gibbs Sampling)
  - GeNIe and SamIAm (graphical interfaces)
  - See the giant list at http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html
- Most packages require using a graphical model representation (which isn't always easy)

Six easy steps

Step 1: Find an interesting aspect of cognition

Step 2: Identify the underlying computational problem

Step 3: Identify constraints

Step 4: Work out the optimal solution to that problem, given constraints

Step 5: See how well that solution corresponds to human behavior (do some experiments!)

Step 6: Iterate Steps 2-6 until it works

(Anderson, 1990)

The perceptual magnet effect

Compare two-category model for categories /i/ and /e/ with data from Iverson and Kuhl's (1995) multidimensional scaling analysis

- compute expectation $E[T|S]$ for each stimulus
- subtract expectations for neighboring stimuli

Parameter estimation

- Assume equal prior probability for /i/ and /e/

(Tobias, 1959)

- Estimate $\mu_{/i/}$ from goodness ratings

(Iverson & Kuhl, 1995)

- Estimate $\mu_{/e/}$ and the quantity $(\sigma_c^2+\sigma_S^2)$ from identification curves

(Lotto, Kluender, & Holt, 1998)

- Find the best-fitting ratio of category variance σ_c^2 to speech signal uncertainty σ_S^2

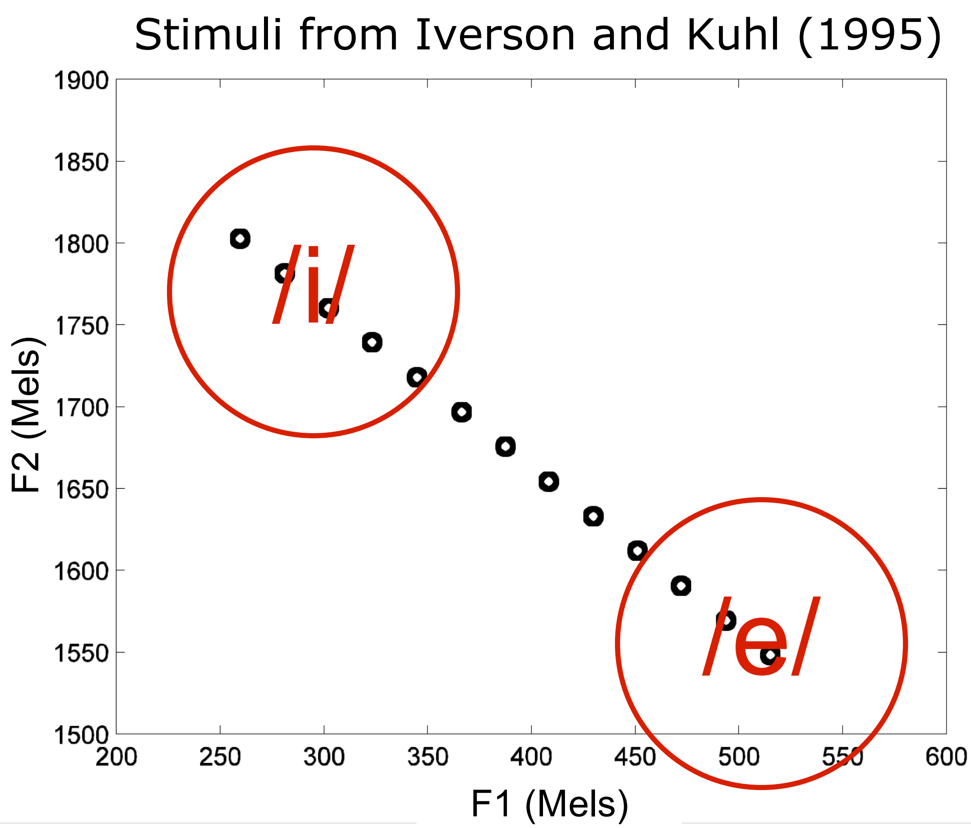
Parameter values

$\mu_{/i/}$: $F1$: 224 Hz
$F2$: 2413 Hz

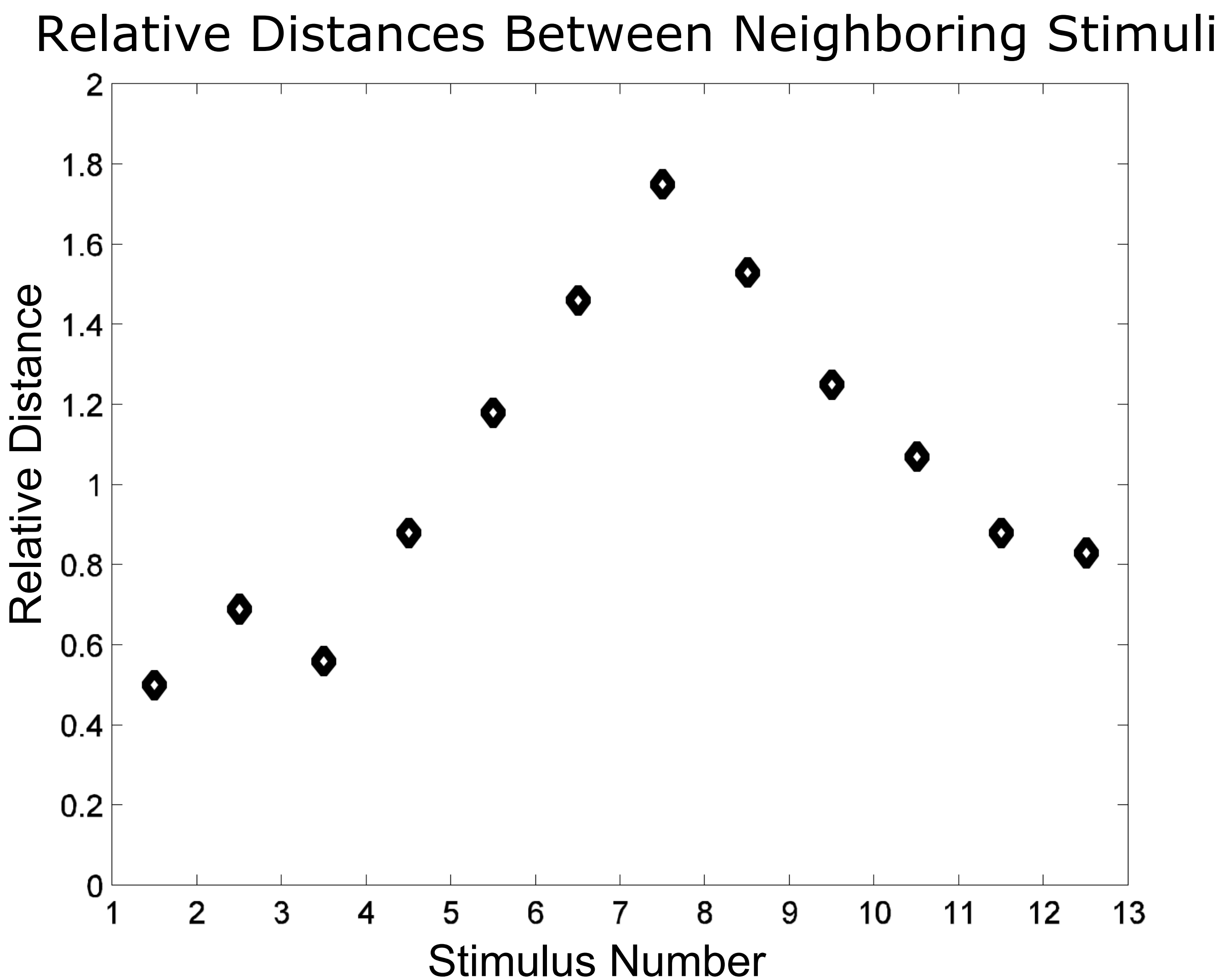
$\mu_{/e/}$: $F1$: 423 Hz
$F2$: 1936 Hz

σ_c: 77 mels

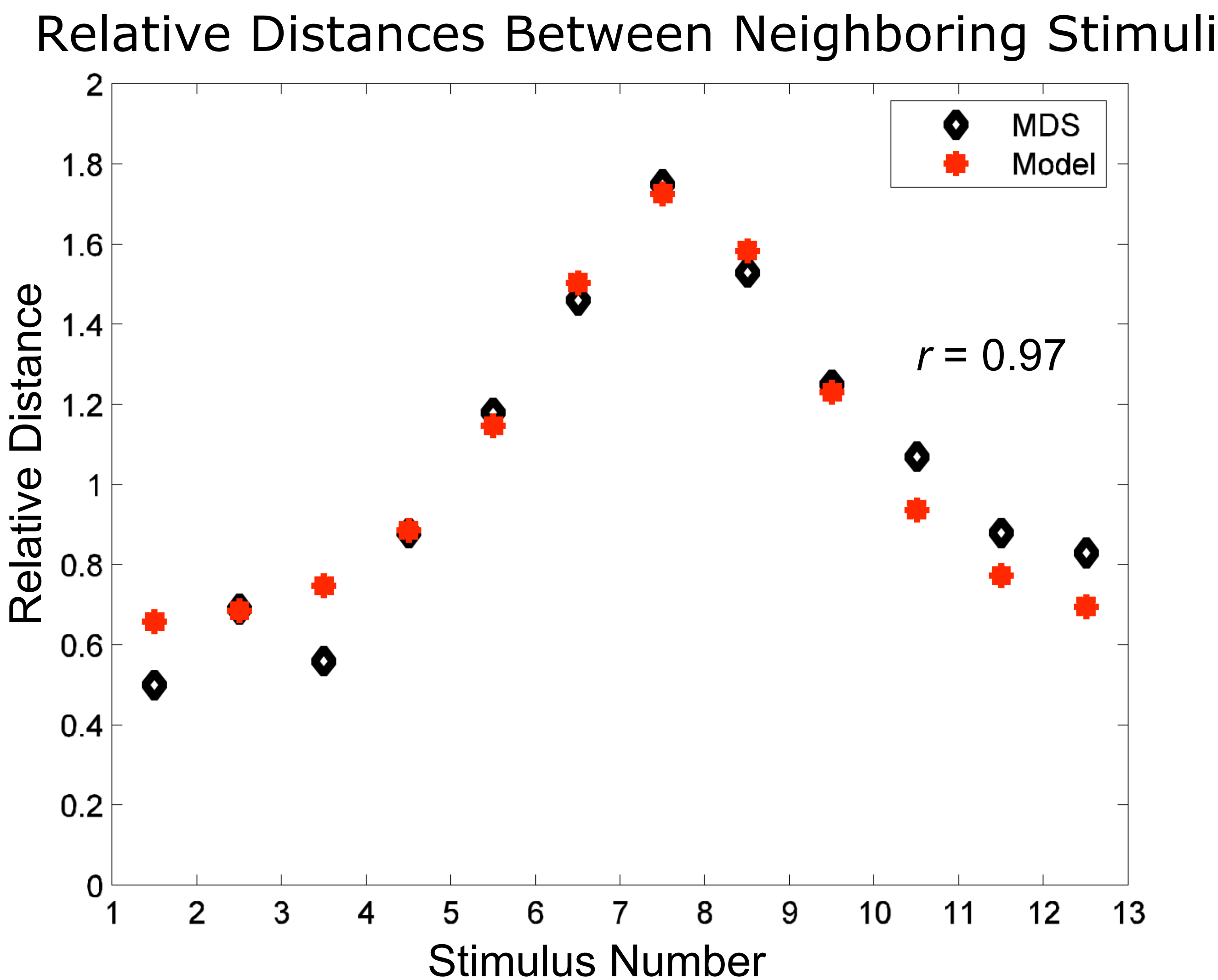
σ_S: 67 mels



Quantitative analysis

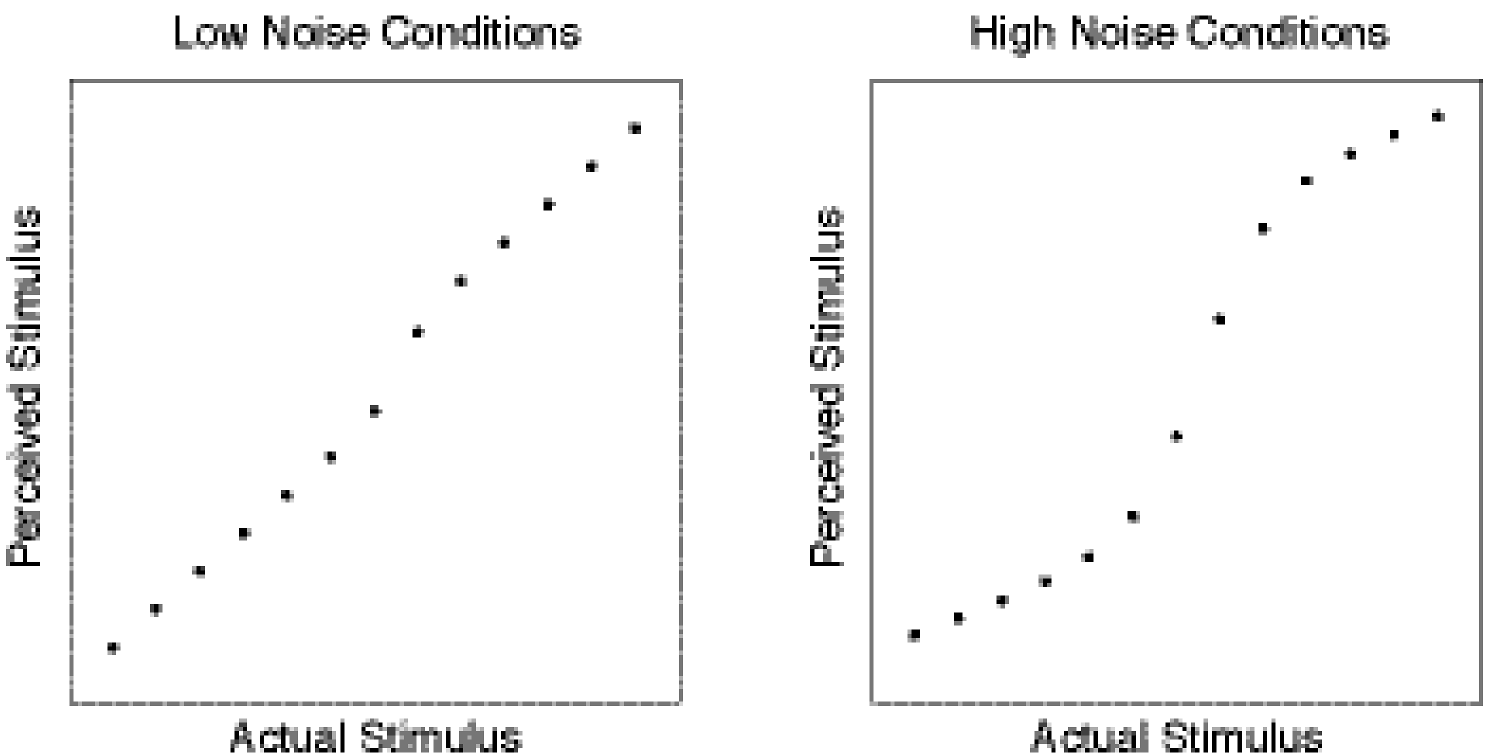


Quantitative analysis

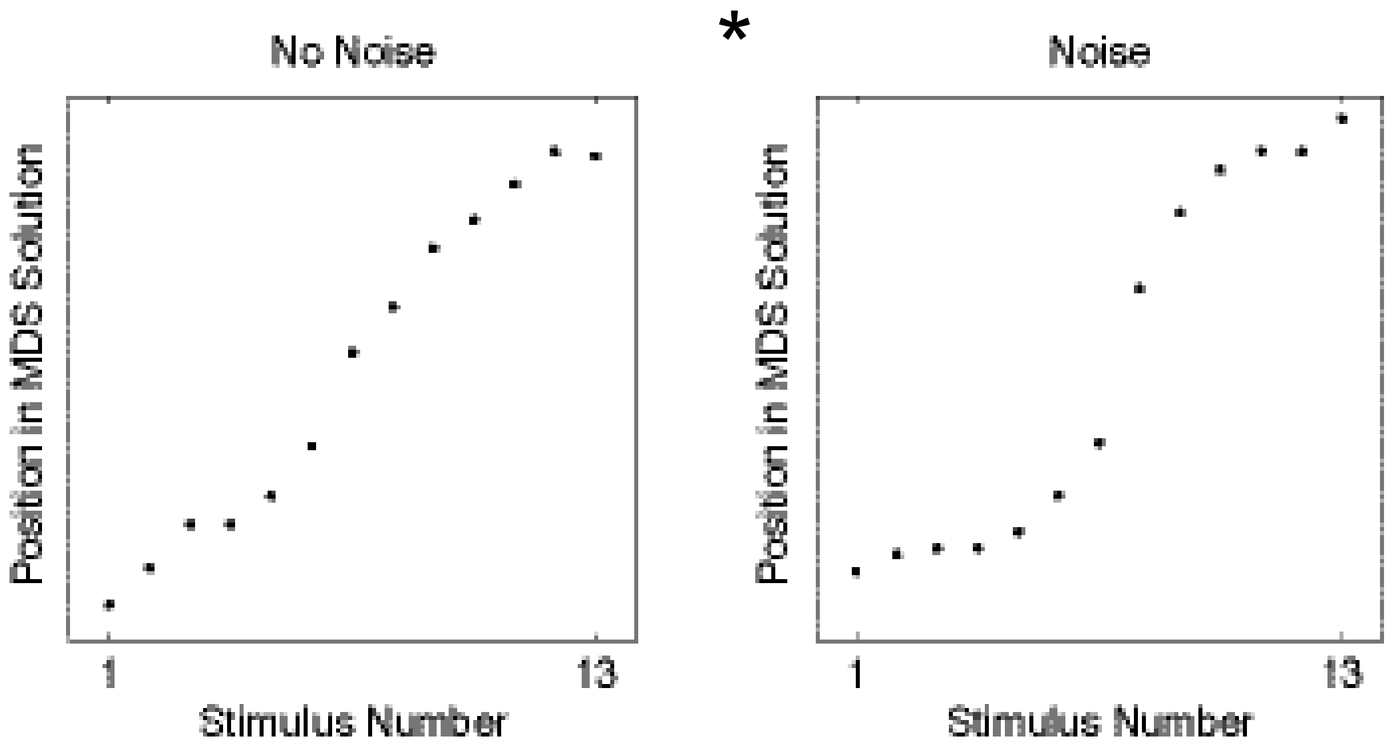


Empirical predictions

Amount of warping depends on ratio of speech signal noise to category variance:



Results

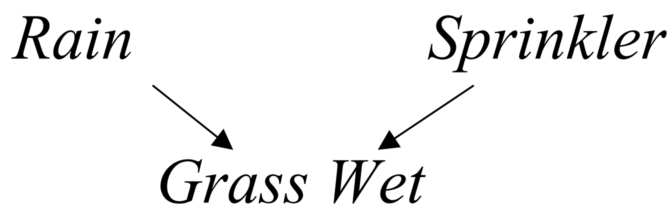


$p<0.05$ in a permutation test based on the log ratio of between/within category distances

Summary

- Bayesian models can be used to answer several questions at the computational level
- The key to defining a Bayesian model is thinking in terms of generative processes
  - graphical models illustrate these processes
  - Bayesian inference inverts these processes
- Depending on the question and the model, different tools can be useful in performing Bayesian inference (but it's usually easy for anything expressed as a graphical model)

Explaining away

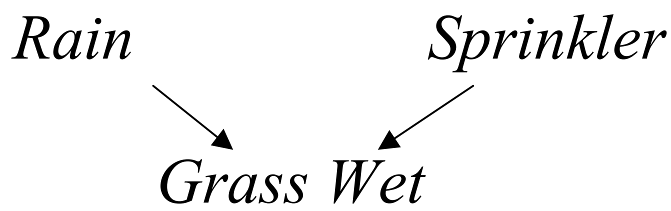


$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

Assume grass will be wet if and only if it rained last night, or if the sprinklers were left on:

$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Explaining away



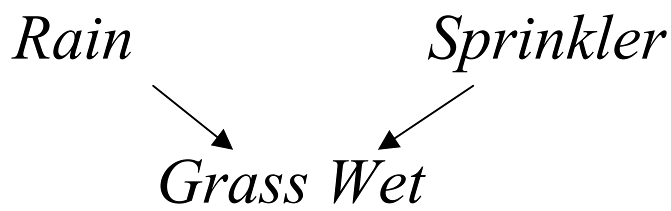
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(w \mid r)P(r)}{P(w)}$$

Explaining away



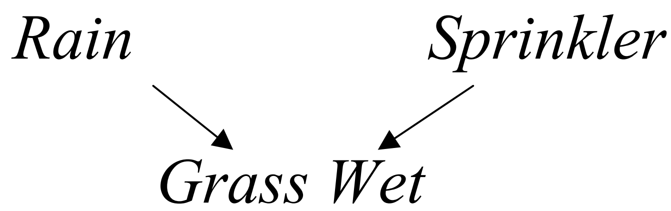
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(w \mid r)P(r)}{\sum_{r',s'} P(w \mid r', s')P(r', s')}$$

Explaining away



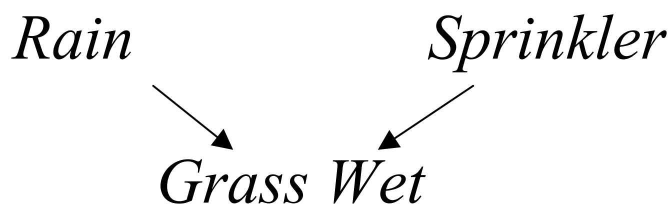
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(r)}{P(r,s) + P(r,\neg s) + P(\neg r,s)}$$

Explaining away



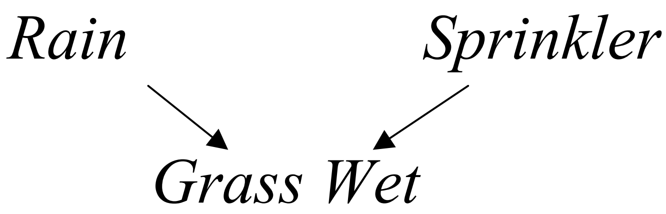
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(r)}{P(r) + P(\neg r, s)}$$

Explaining away



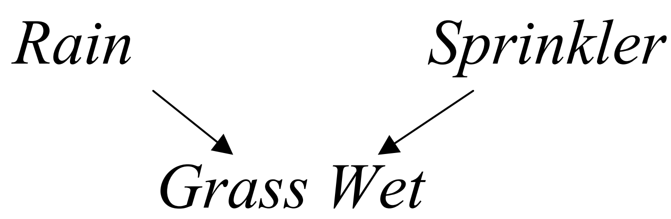
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(r)}{\underbrace{P(r) + P(\neg r)P(s)}_{\text{Between 1 and } P(s)}} > P(r)$$

Explaining away



$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

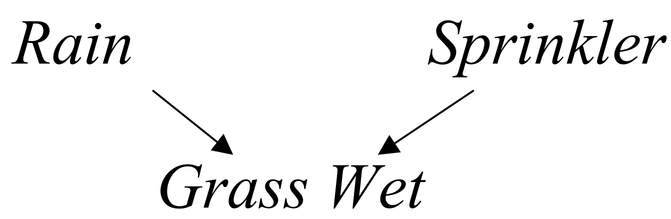
$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Compute probability it rained last night, given that the grass is wet and sprinklers were left on:

$$P(r \mid w,s) = \frac{P(w \mid r,s)P(r \mid s)}{P(w \mid s)}$$

Both terms = 1

Explaining away



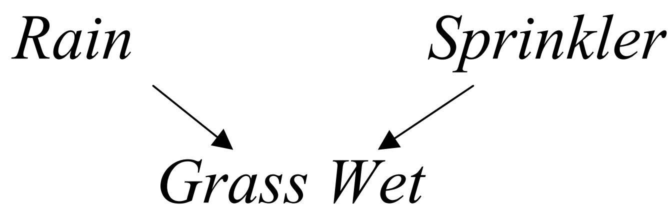
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

Compute probability it rained last night, given that the grass is wet and sprinklers were left on:

$$P(r \mid w,s) = P(r \mid s) = P(r)$$

Explaining away



$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

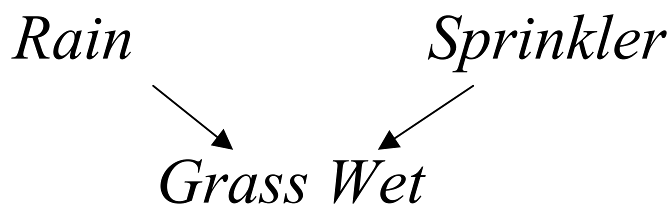
$$\begin{aligned} P(W = w \mid S,R) &= 1 \text{ if } S = s \text{ or } R = r \\ &= 0 \text{ if } R = \neg r \text{ and } S = \neg s. \end{aligned}$$

$$P(r \mid w) = \frac{P(r)}{P(r) + P(\neg r)P(s)} \quad > P(r)$$

$$P(r \mid w,s) = P(r \mid s) = P(r)$$

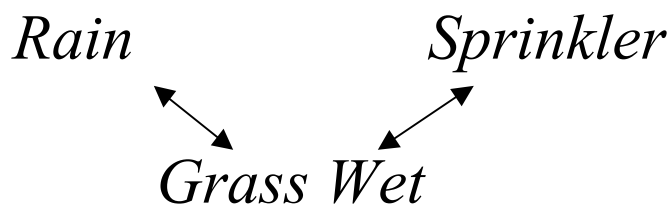
"Discounting" to prior probability.

Contrast w/ production system



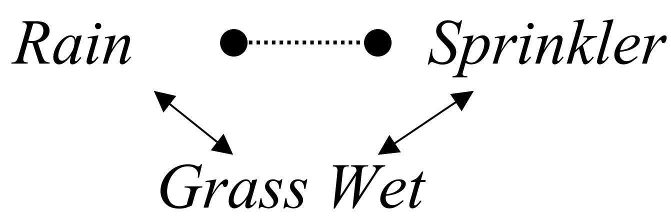
- Formulate IF-THEN rules:
  - IF *Rain* THEN *Wet*
  - ~~IF *Wet* THEN *Rain*~~ IF *Wet* AND NOT *Sprinkler* THEN *Rain*
- Rules do not distinguish directions of inference
- Requires combinatorial explosion of rules

Contrast w/ spreading activation



- Excitatory links: *Rain* $\leftrightarrow$ *Wet*, *Sprinkler* $\leftrightarrow$ *Wet*
- Observing rain, *Wet* becomes more active.
- Observing grass wet, *Rain* and *Sprinkler* become more active
- Observing grass wet and sprinkler, *Rain* cannot become less active. No explaining away!

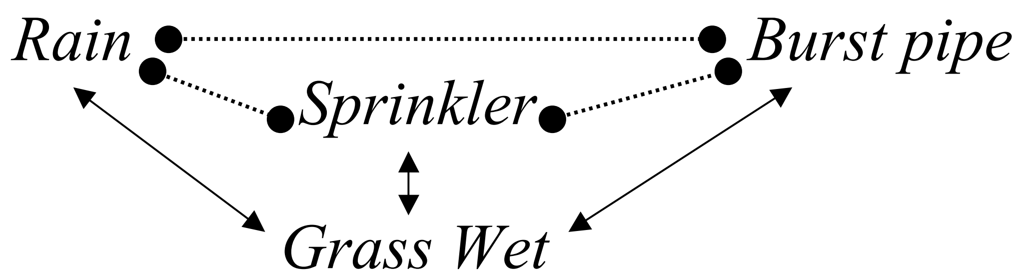
Contrast w/ spreading activation



- Excitatory links: *Rain* ↔ *Wet*, *Sprinkler* ↔ *Wet*
- Inhibitory link: *Rain* •⋯• *Sprinkler*

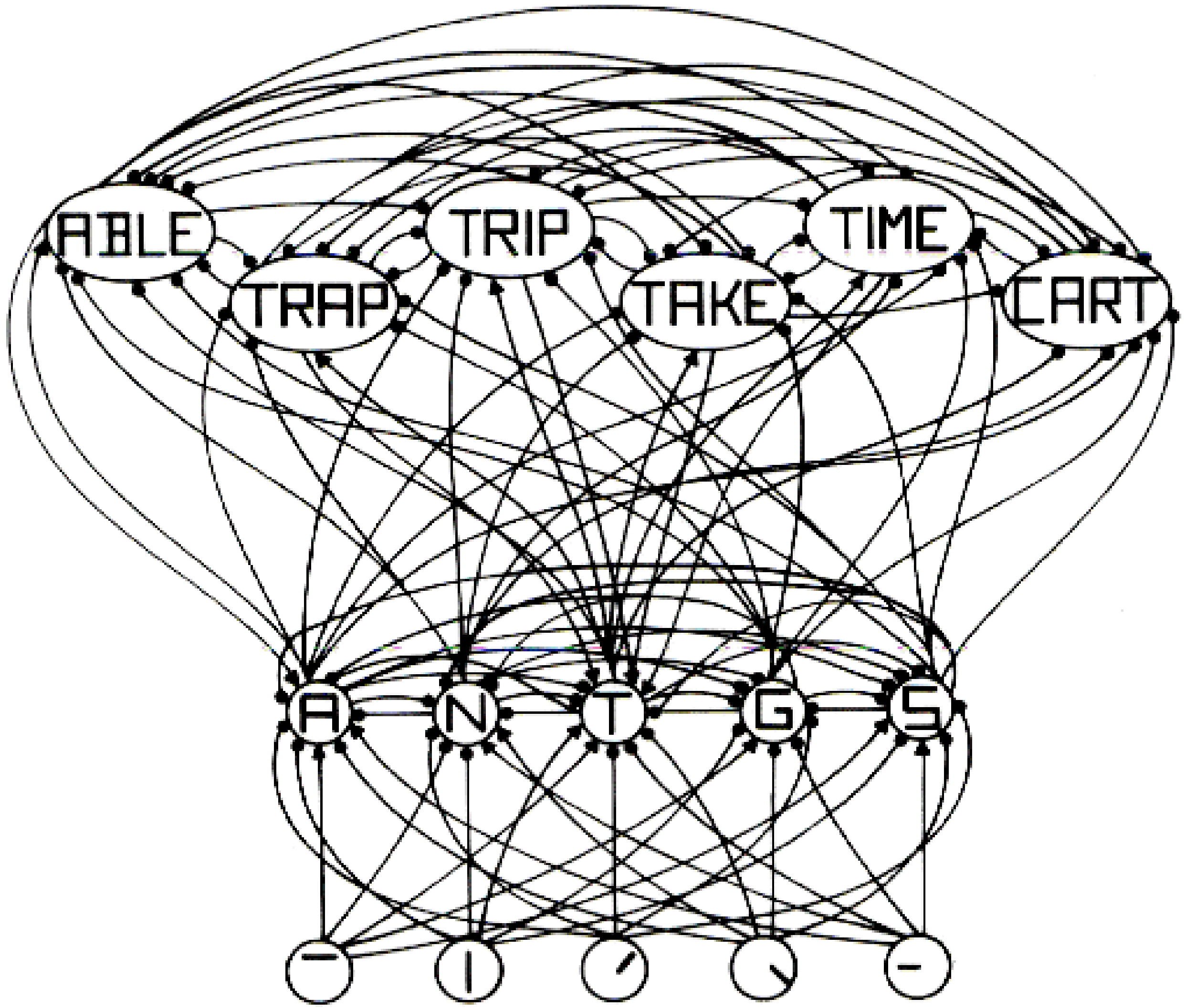
- Observing grass wet, *Rain* and *Sprinkler* become more active
- Observing grass wet and sprinkler, *Rain* becomes less active: explaining away

Contrast w/ spreading activation



- Each new variable requires more inhibitory connections
- Not modular
  - whether a connection exists depends on what others exist
  - big holism problem
  - combinatorial explosion

Contrast w/ spreading activation



(McClelland & Rumelhart, 1981)