Part III

Learning structured representations
Hierarchical Bayesian models

Universal Grammar

↓

Grammar

↓

Phrase structure

↓

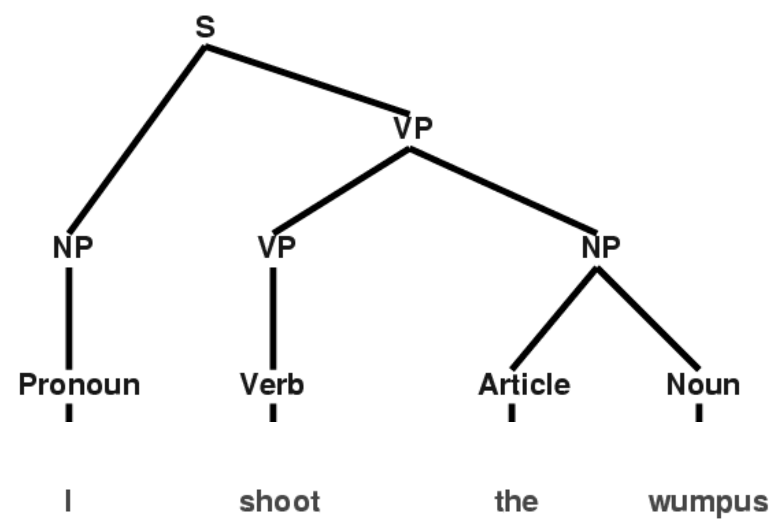
Utterance

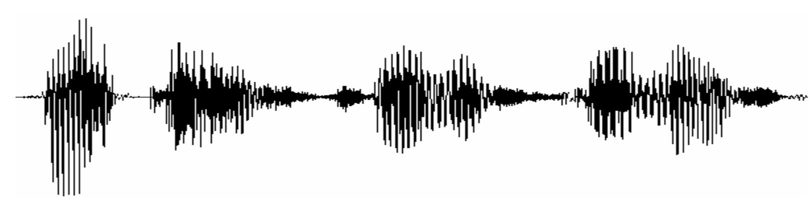
↓

Speech signal

Hierarchical phrase structure grammars (e.g., CFG, HPSG, TAG)

$$S \rightarrow NP\ VP$$
$$NP \rightarrow Det\ [Adj]\ Noun\ [RelClause]$$
$$RelClause \rightarrow [Rel]\ NP\ V$$
$$VP \rightarrow VP\ NP$$
$$VP \rightarrow Verb$$





Outline

- Learning structured representations
  - grammars
  - logical theories
- Learning at multiple levels of abstraction

A historical divide

Structured Representations

Innate knowledge

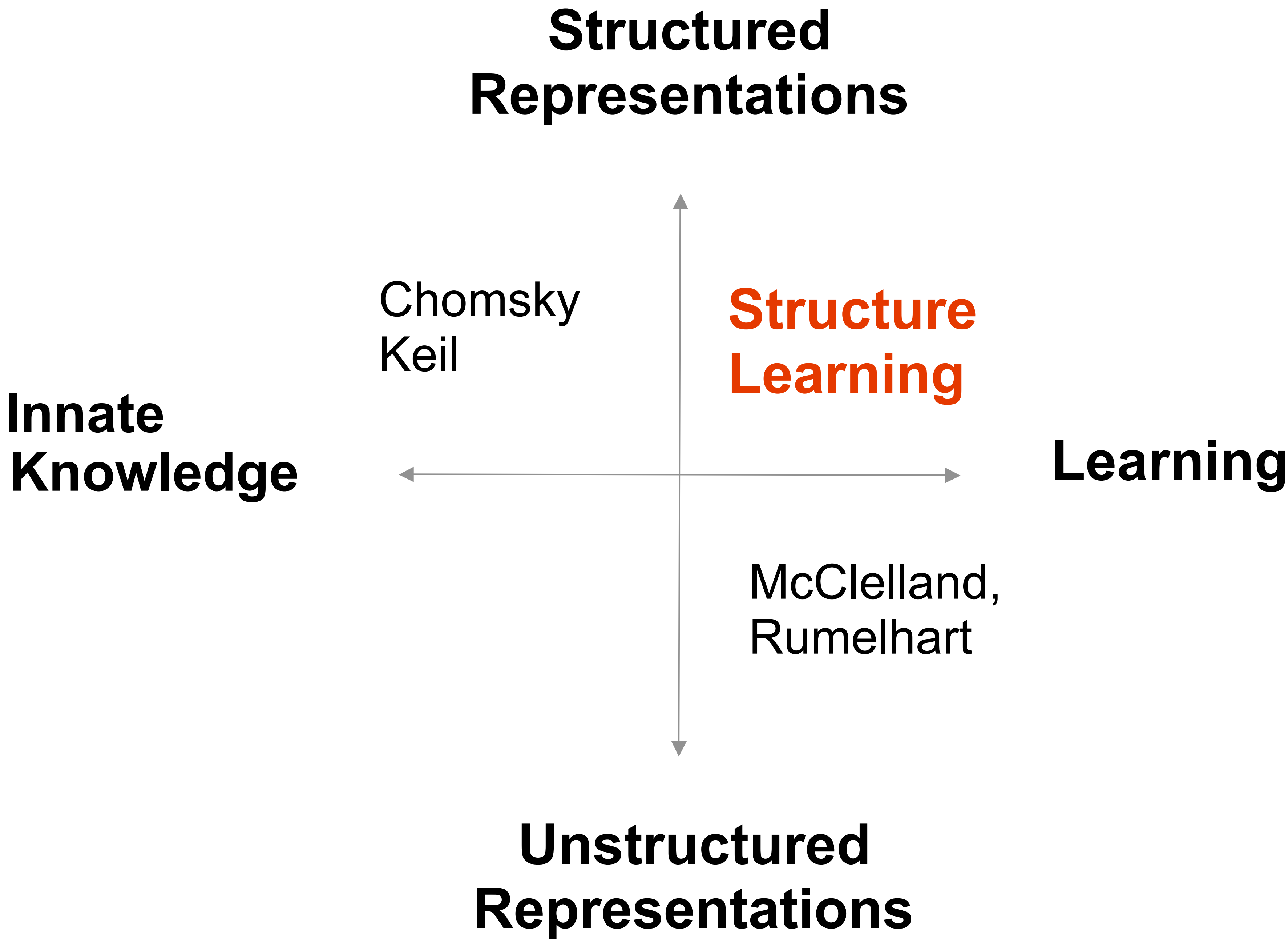
(Chomsky,
Pinker,
Keil, ...)

vs

Unstructured Representations

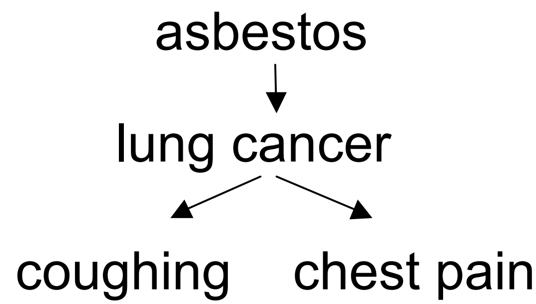
Learning

(McClelland,
Rumelhart, ...)


Structured Representations
Chomsky
Keil
Structure Learning
Innate Knowledge
Learning
McClelland, Rumelhart
Unstructured Representations

Representations

Causal networks



Grammars

$$S \rightarrow NP\ VP$$
$$NP \rightarrow Det\ [Adj]\ Noun\ [RelClause]$$
$$RelClause \rightarrow [Rel]\ NP\ V$$
$$VP \rightarrow VP\ NP$$
$$VP \rightarrow Verb$$

Logical theories

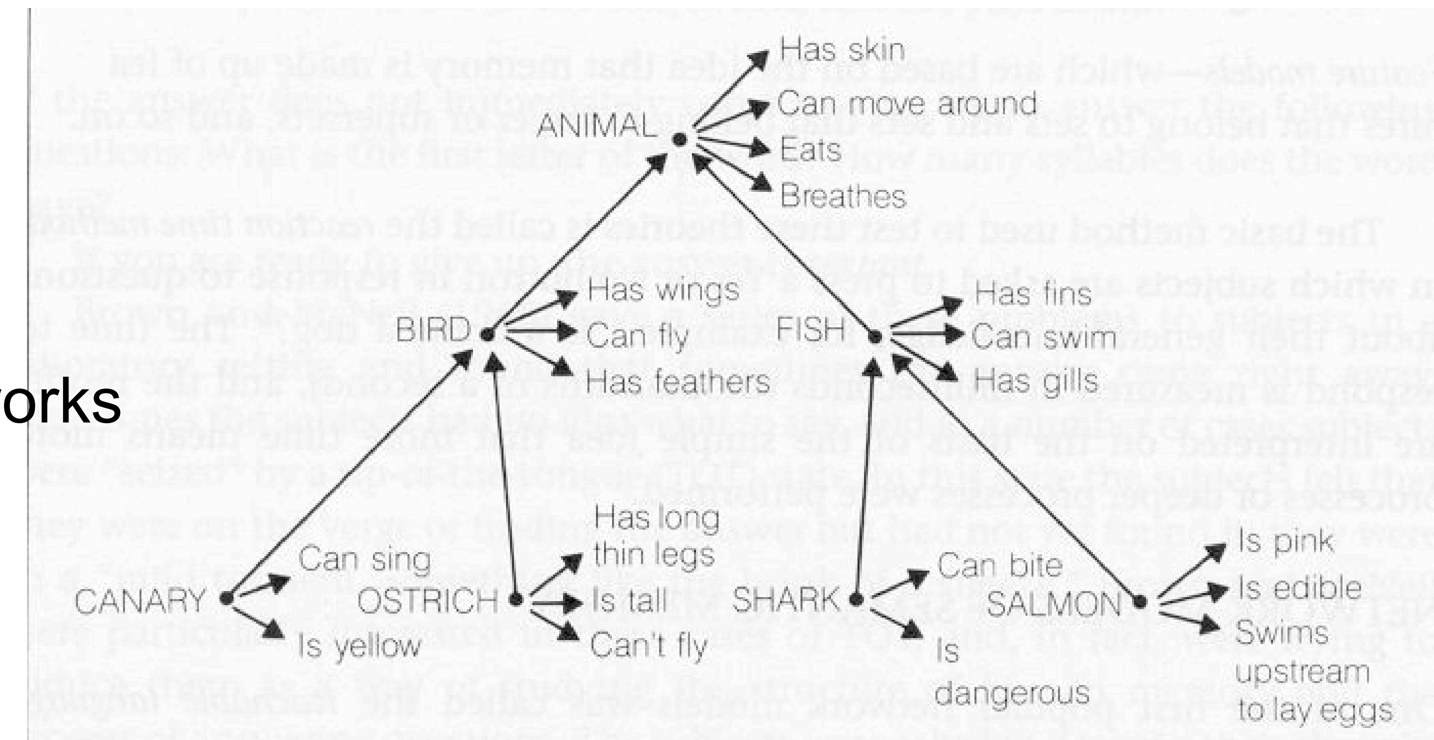
$$\forall x\ y\ \mathsf{Sibling}(x, y) \leftarrow \mathsf{Sibling}(y, x)$$
$$\forall x\ y\ \mathsf{Ancestor}(x, y) \leftarrow \mathsf{Parent}(x, y)$$

Representations

Phonological rules

$$\begin{bmatrix} +\text{syllabic} \\ -\text{consonantal} \end{bmatrix} \rightarrow [\,+\text{back}\,] \,/\, \begin{bmatrix} +\text{back} \\ +\text{syllabic} \\ -\text{consonantal} \end{bmatrix} [\,+\text{consonantal}\,]^{*} \;\_\_$$

Semantic networks



How to learn a R

- Search for R that maximizes

$$P(R|\text{Data}) \propto P(\text{Data}|R)P(R)$$

- Prerequisites
  - Put a prior over a hypothesis space of Rs.
  - Decide how observable data are generated from an underlying R.

How to learn ~~a R~~ anything

- Search for R that maximizes

$$P(R|\text{Data}) \propto P(\text{Data}|R)P(R)$$

- Prerequisites
  – Put a prior over a hypothesis space of Rs.
  – Decide how observable data are generated from an underlying R.

Context free grammar

S → N VP

VP → V

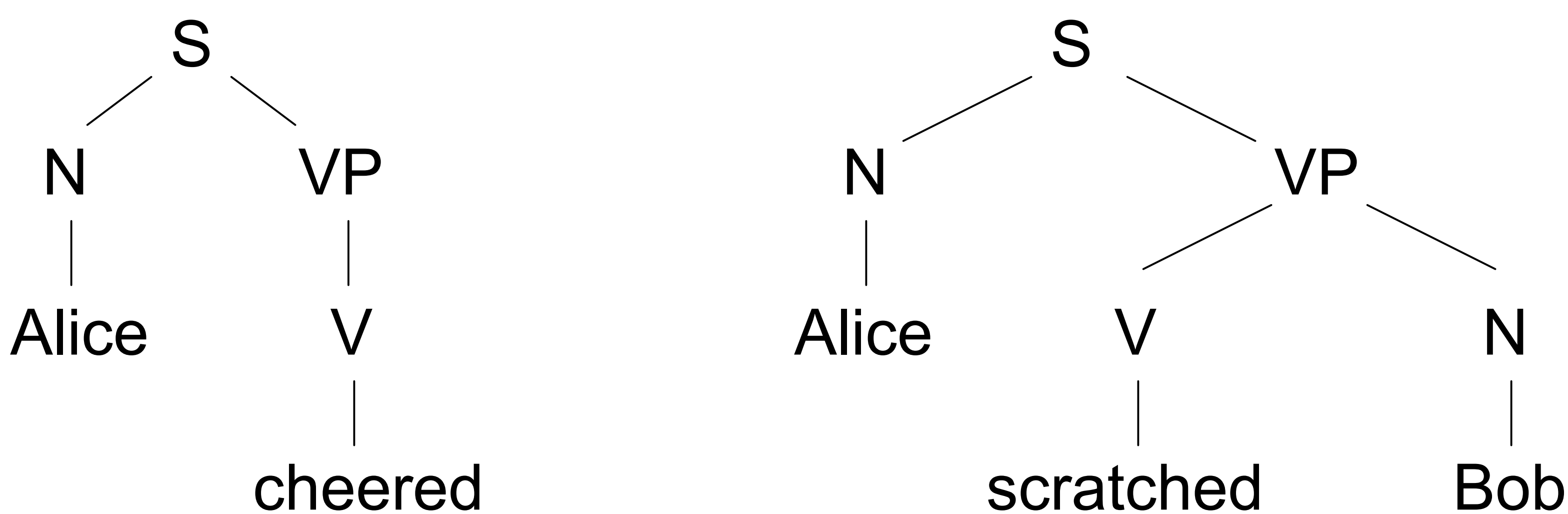
VP → V N

N → "Alice"

N → "Bob"

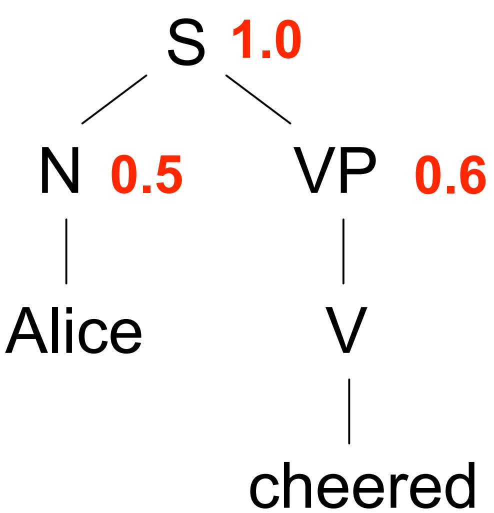
V → "scratched"

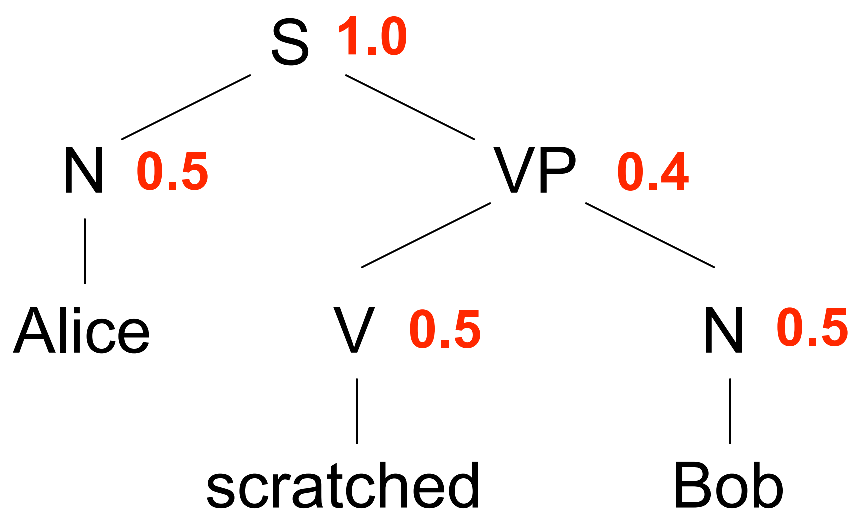
V → "cheered"



Probabilistic context free grammar

1.0 S → N VP	0.6 VP → V	0.5 N → “Alice”	0.5 V → “scratched”
	0.4 VP → V N	0.5 N → “Bob”	0.5 V → “cheered”





probability = 1.0 * 0.5 * 0.6
= 0.3

probability = 1.0*0.5*0.4*0.5*0.5
= 0.05

The learning problem

Grammar G:

$S \xrightarrow{1.0} N\ VP$

$VP \xrightarrow{0.6} V$

$VP \xrightarrow{0.4} V\ N$

$N \xrightarrow{0.5}$ "Alice"

$N \xrightarrow{0.5}$ "Bob"

$V \xrightarrow{0.5}$ "scratched"

$V \xrightarrow{0.5}$ "cheered"

Data D:

Alice scratched.
Bob scratched.
Alice scratched Alice.
Alice scratched Bob.
Bob scratched Alice.
Bob scratched Bob.

Alice cheered.
Bob cheered.
Alice cheered Alice.
Alice cheered Bob.
Bob cheered Alice.
Bob cheered Bob.

Grammar learning

- Search for G that maximizes

$$P(G|\text{Data}) \propto P(\text{Data}|G)P(G)$$

- Prior: $P(G) \propto 2^{-\text{length}(G)}$

- Likelihood: $P(\text{Data}|G)$
  - assume that sentences in the data are independently generated from the grammar.

(Horning 1969; Stolcke 1994)

Experiment

```
S --> NP VP
NP --> Det N
VP --> Vt NP
   --> Vc PP
   --> Vi
PP --> P NP
Det --> a
    --> the
Vt --> touches
   --> covers
Vc --> is
Vi --> rolls
   --> bounces
N  --> circle
   --> square
   --> triangle
P  --> above
   --> below
```

- Data: 100 sentences

```
the circle covers a square
a square is above the triangle
a circle bounces
        ⋮
```

(Stolcke, 1994)

Generating grammar:

```
S --> NP VP
NP --> Det N
VP --> Vt NP
   --> Vc PP
   --> Vi
PP --> P NP
Det --> a
    --> the
Vt --> touches
   --> covers
Vc --> is
Vi --> rolls
   --> bounces
N  --> circle
   --> square
   --> triangle
P  --> above
   --> below
```

Model solution:

```
S --> NP VP
NP --> Det N
VP --> VI
   --> X NP
X --> VT
  --> VC P
Det --> a
    --> the
Vt --> touches
   --> covers
Vc --> is
Vi --> rolls
   --> bounces
N  --> circle
   --> square
   --> triangle
P  --> above
   --> below
```

Predicate logic

- A compositional language

$$\forall x\, y\, \text{Sibling}(x, y) \leftarrow \text{Sibling}(y, x)$$

For all x and y, if y is the sibling of x then x is the sibling of y

$$\forall x\, y\, z\, \text{Ancestor}(x, z) \leftarrow \text{Ancestor}(x, y) \wedge \text{Ancestor}(y, z)$$

For all x, y and z, if x is the ancestor of y and y is the ancestor of z, then x is the ancestor of z.

Learning a kinship theory

Theory T:

$$\forall x\, y\, \text{Sibling}(x, y) \leftarrow \text{Sibling}(y, x)$$

$$\forall x\, y\, z\, \text{Ancestor}(x, z) \leftarrow \text{Ancestor}(x, y) \wedge \text{Ancestor}(y, z)$$

$$\forall x\, y\, \text{Ancestor}(x, y) \leftarrow \text{Parent}(x, y)$$

$$\forall x\, y\, z\, \text{Uncle}(x, z) \leftarrow \text{Brother}(x, y) \wedge \text{Parent}(y, z)$$

Data D:

Sibling(victoria, arthur), Sibling(arthur,victoria),
Ancestor(chris,victoria), Ancestor(chris,colin),
Parent(chris,victoria), Parent(victoria,colin),
Uncle(arthur,colin), Brother(arthur,victoria) ...

(Hinton, Quinlan, …)

Learning logical theories

- Search for T that maximizes

$$P(T|\text{Data}) \propto P(\text{Data}|T)P(T)$$

- Prior: $P(T) \propto 2^{-\text{length}(T)}$

- Likelihood: $P(\text{Data}|T)$
  - assume that the data include all facts that are true according to T

(Conklin and Witten; Kemp et al 08; Katz et al 08)

Theory-learning in the lab

R(k,c) R(c,b)

R(f,c) R(c,l)

R(f,k)

R(k,l) R(l,b)

R(f,l)

R(l,h)

R(f,b) R(k,b) R(f,h)

R(b,h)

R(c,h)

R(k,h)

(cf Krueger 1979)

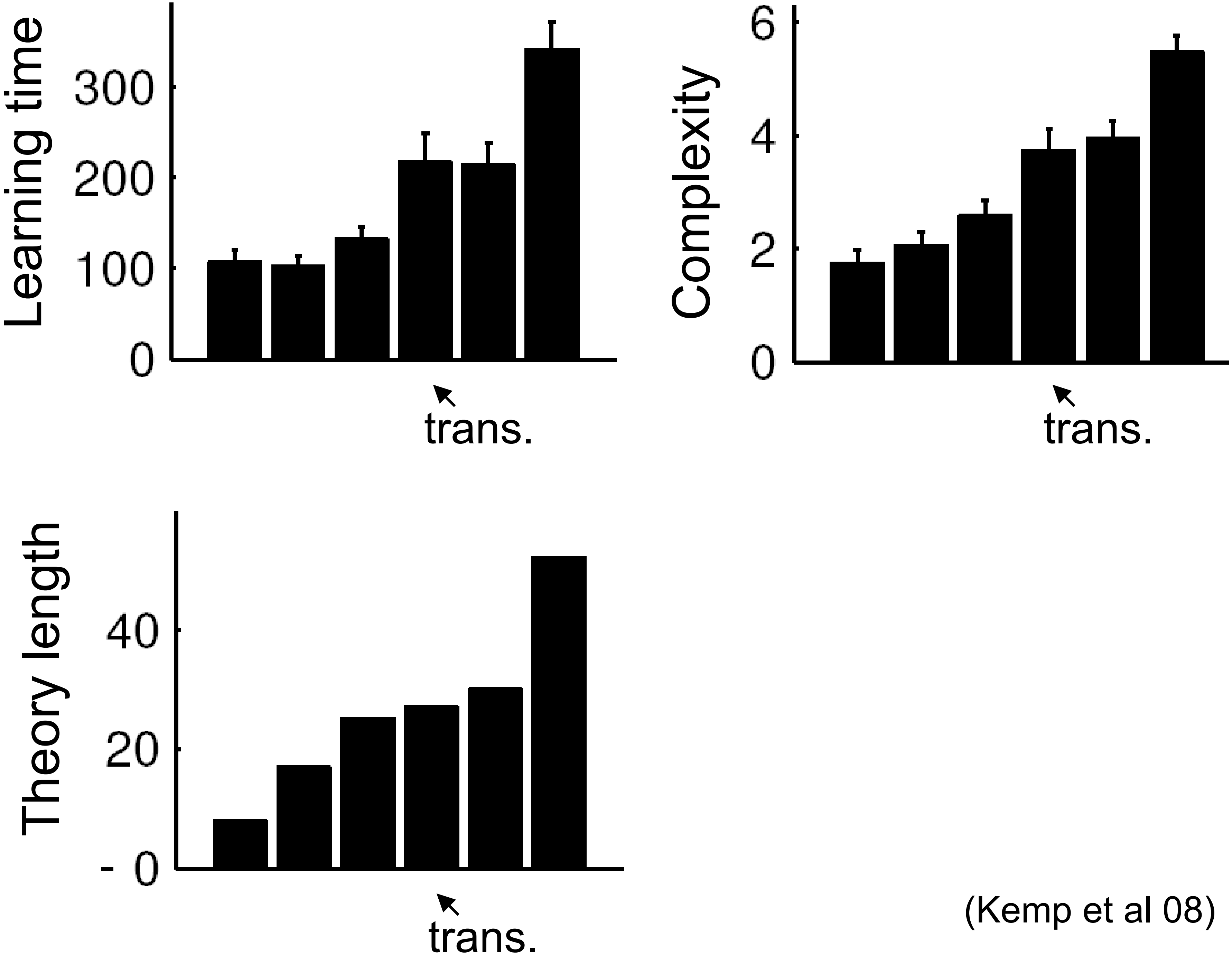
Theory-learning in the lab

Transitive: R(f,k). R(k,c). R(c,l). R(l,b). R(b,h).

$$R(X,Z) \leftarrow R(X,Y), R(Y,Z).$$

---

f,k	f,c	f,l	f,b	f,h
	k,c	k,l	k,b	k,h
		c,l	c,b	c,h
			l,b	l,h
				b,h


Learning time
300
200
100
0
trans.
Complexity
6
4
2
0
trans.
Theory length
40
20
- 0
trans.

(Kemp et al 08)

Conclusion: Part 1

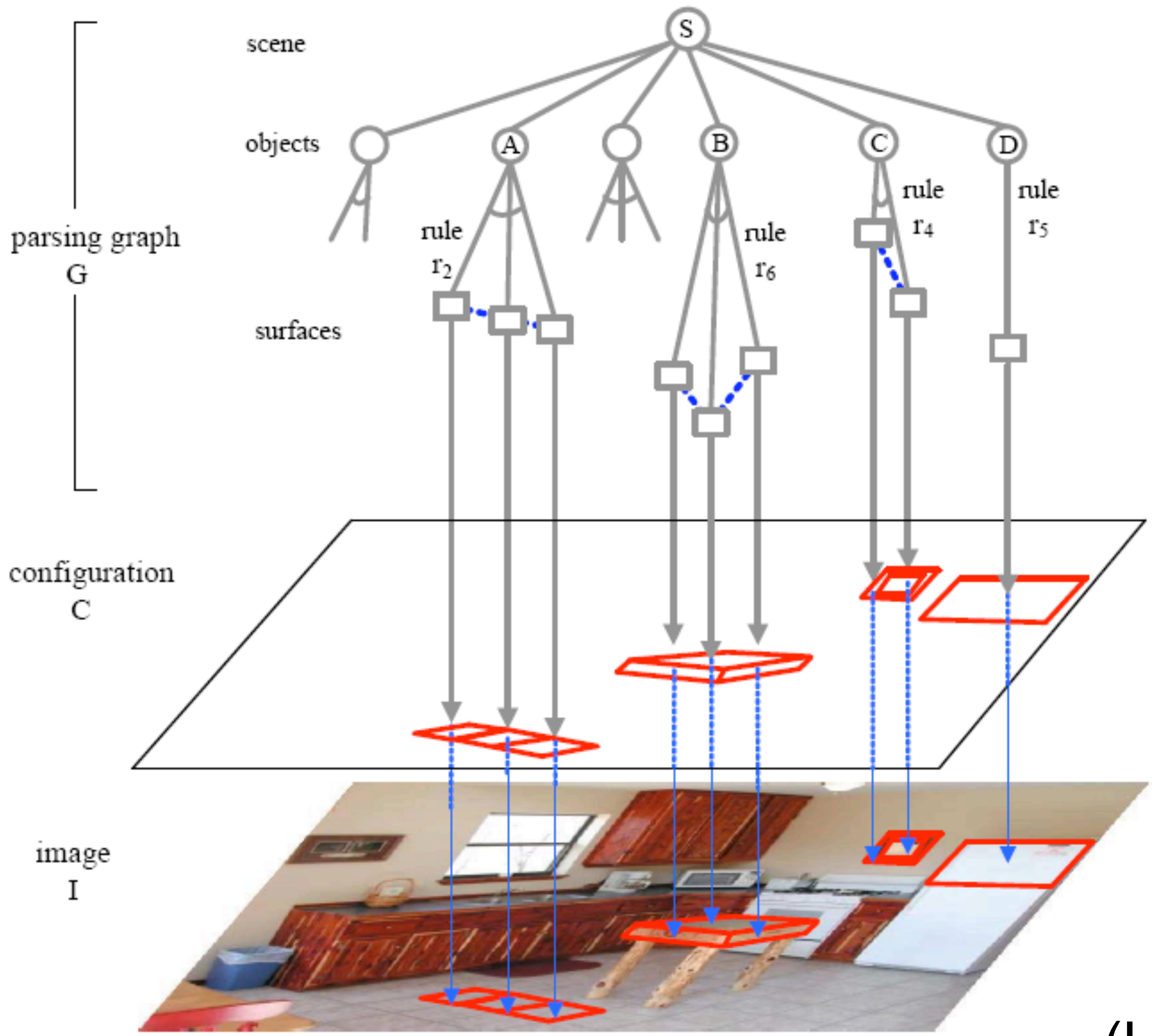
- Bayesian models can combine structured representations with statistical inference.

Outline

- Learning structured representations
  – grammars
  – logical theories

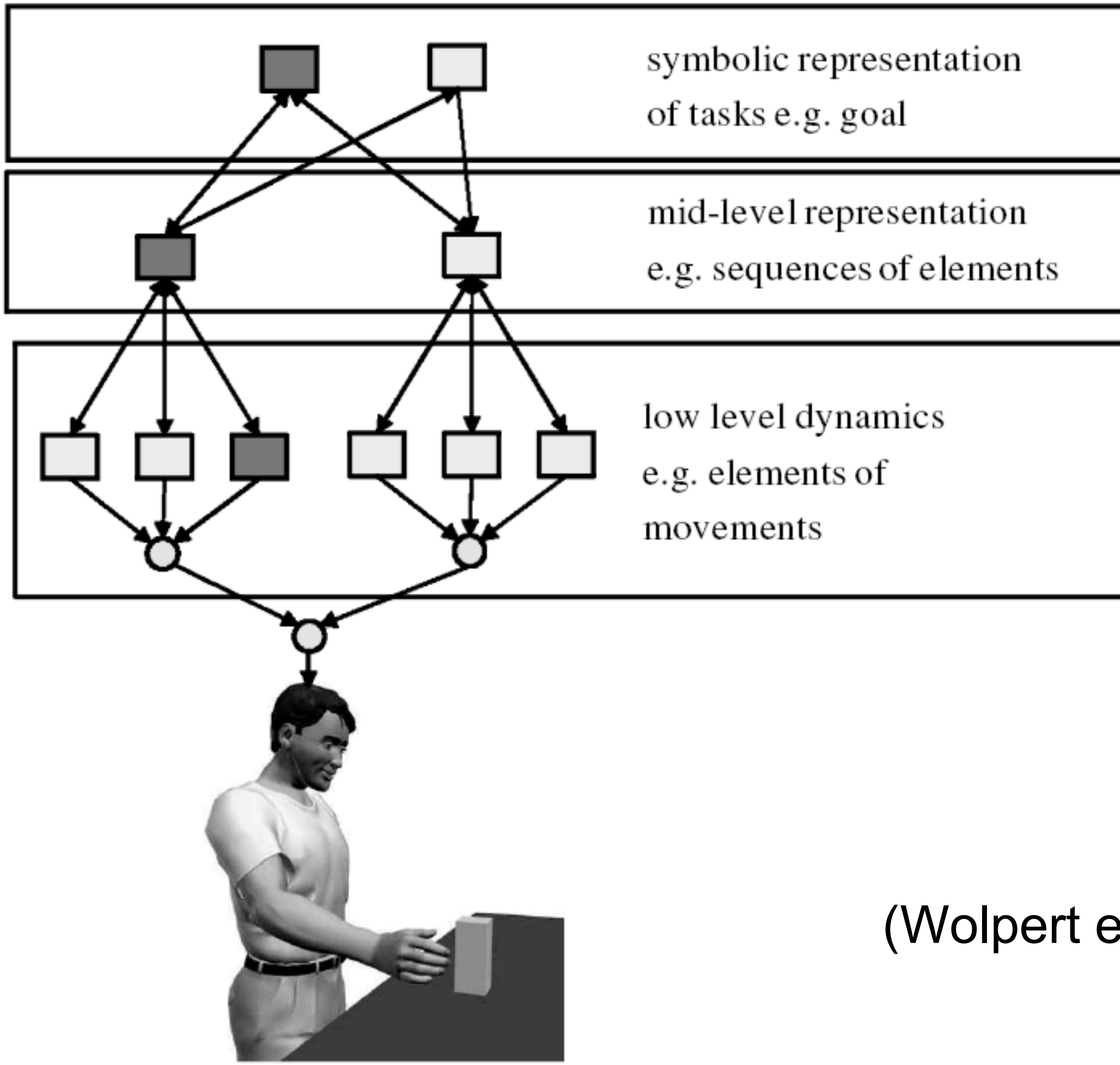
- Learning at multiple levels of abstraction

Vision



(Han and Zhu, 2006)

Motor Control



(Wolpert et al., 2003)

Causal learning

Schema

chemicals → diseases → symptoms

Causal models

asbestos → lung cancer → coughing, chest pain

mercury → minamata disease → muscle wasting

Contingency Data

Patient 1: asbestos exposure, coughing, chest pain

Patient 2: mercury exposure, muscle wasting

(Kelley; Cheng; Waldmann)

Universal Grammar

↓ P(grammar | UG)

Grammar

↓ P(phrase structure | grammar)

Phrase structure

↓ P(utterance | phrase structure)

Utterance

↓ P(speech | utterance)

Speech signal

Hierarchical phrase structure grammars (e.g., CFG, HPSG, TAG)

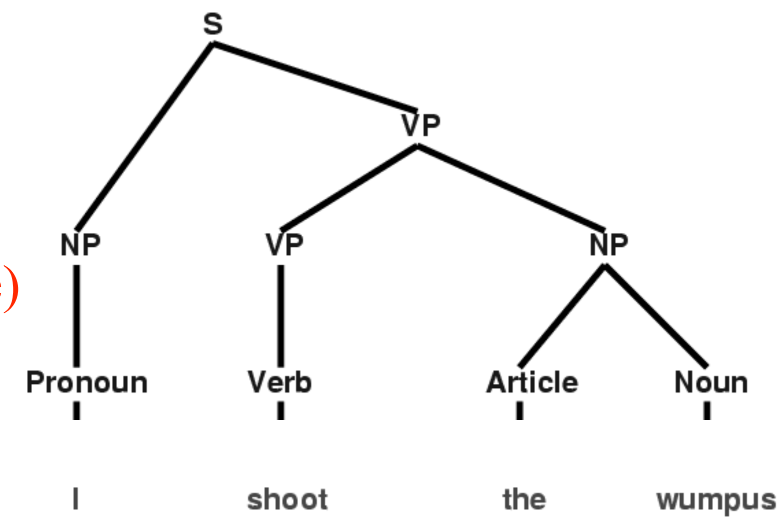
$$S \rightarrow NP\ VP$$

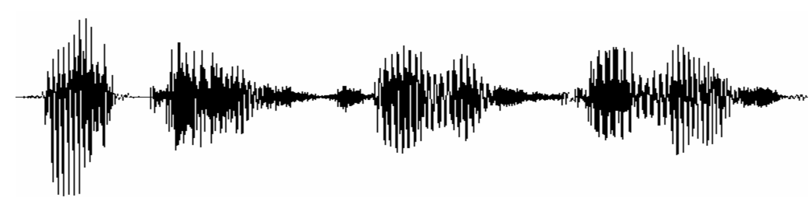
$$NP \rightarrow Det\ [Adj]\ Noun\ [RelClause]$$

$$RelClause \rightarrow [Rel]\ NP\ V$$

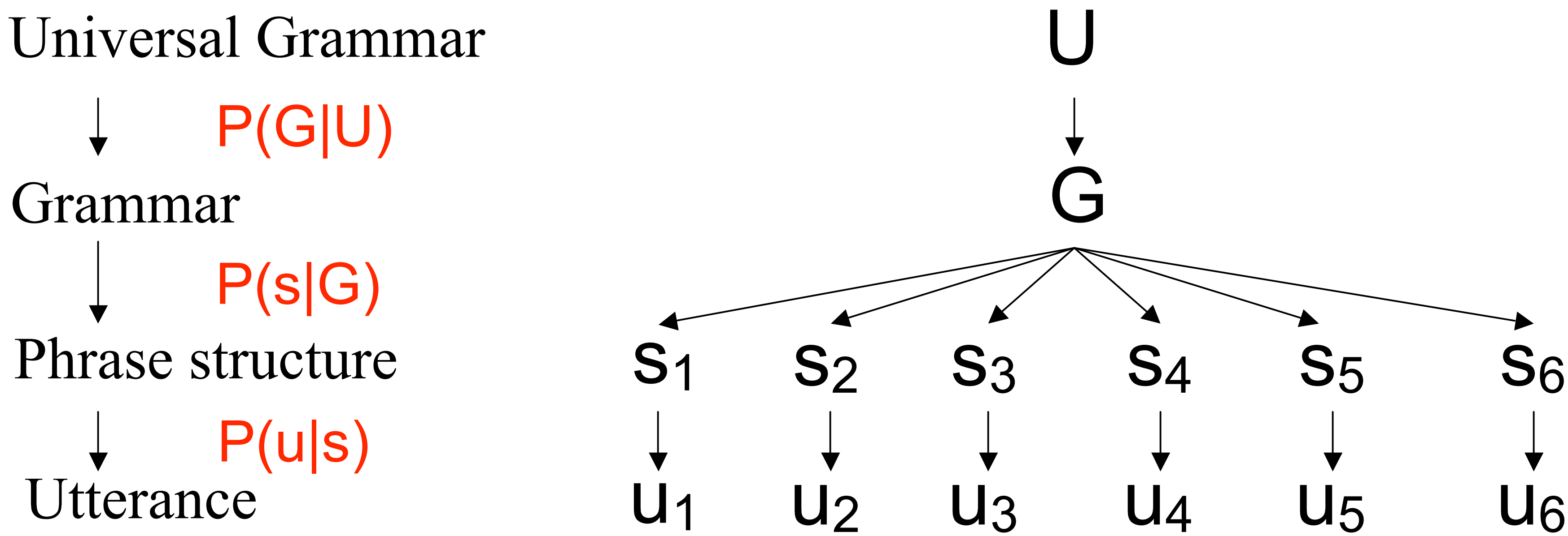
$$VP \rightarrow VP\ NP$$

$$VP \rightarrow Verb$$





Hierarchical Bayesian model

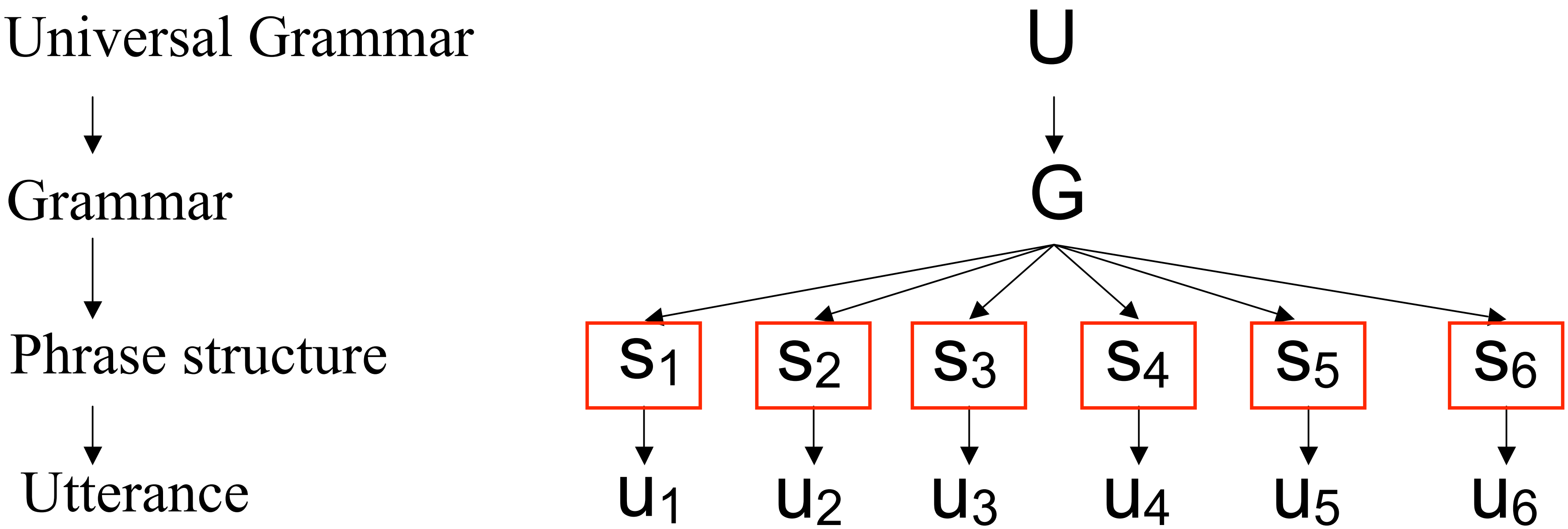


A hierarchical Bayesian model specifies a joint distribution over all variables in the hierarchy:

$$P(\{u_i\}, \{s_i\}, G \mid U)$$

$$= P(\{u_i\} \mid \{s_i\})\, P(\{s_i\} \mid G)\, P(G|U)$$

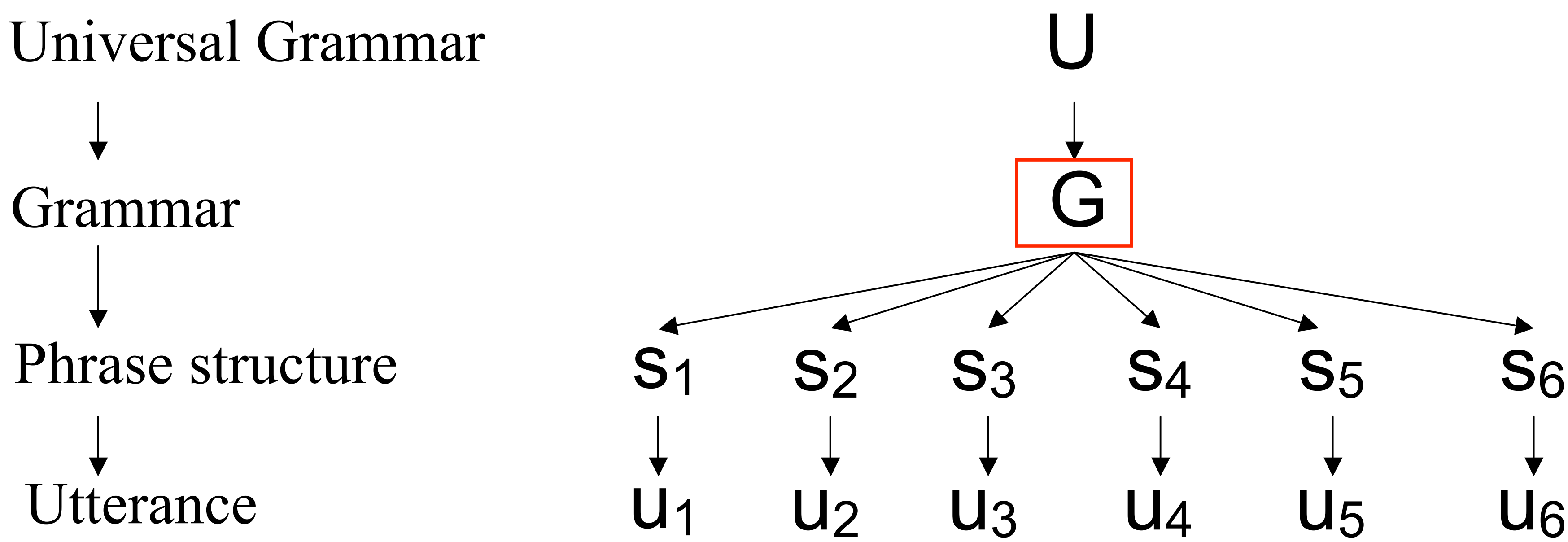
Top-down inferences



Infer $\{s_i\}$ given $\{u_i\}$, G:

$$P(\{s_i\} \mid \{u_i\}, G) \propto P(\{u_i\} \mid \{s_i\})\, P(\{s_i\} \mid G)$$

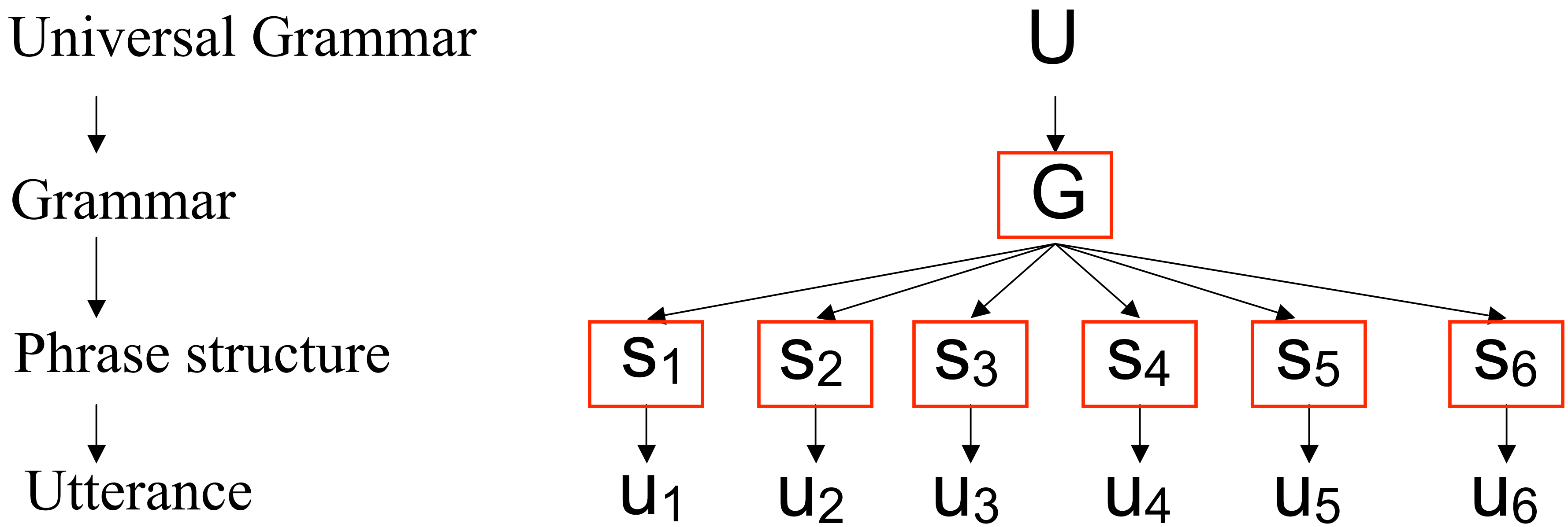
Bottom-up inferences



Infer G given $\{s_i\}$ and U:

$$P(G|\{s_i\}, U) \propto P(\{s_i\}|G)\,P(G|U)$$

Simultaneous learning at multiple levels



Infer G and $\{s_i\}$ given $\{u_i\}$ and U:

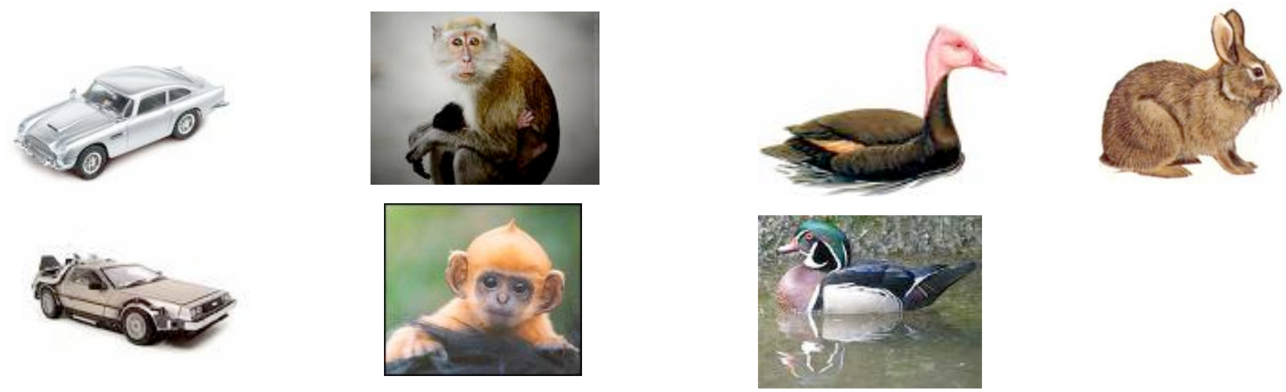
$$P(G, \{s_i\} \mid \{u_i\}, U) \propto P(\{u_i\} \mid \{s_i\})P(\{s_i\} \mid G)P(G \mid U)$$

Word learning

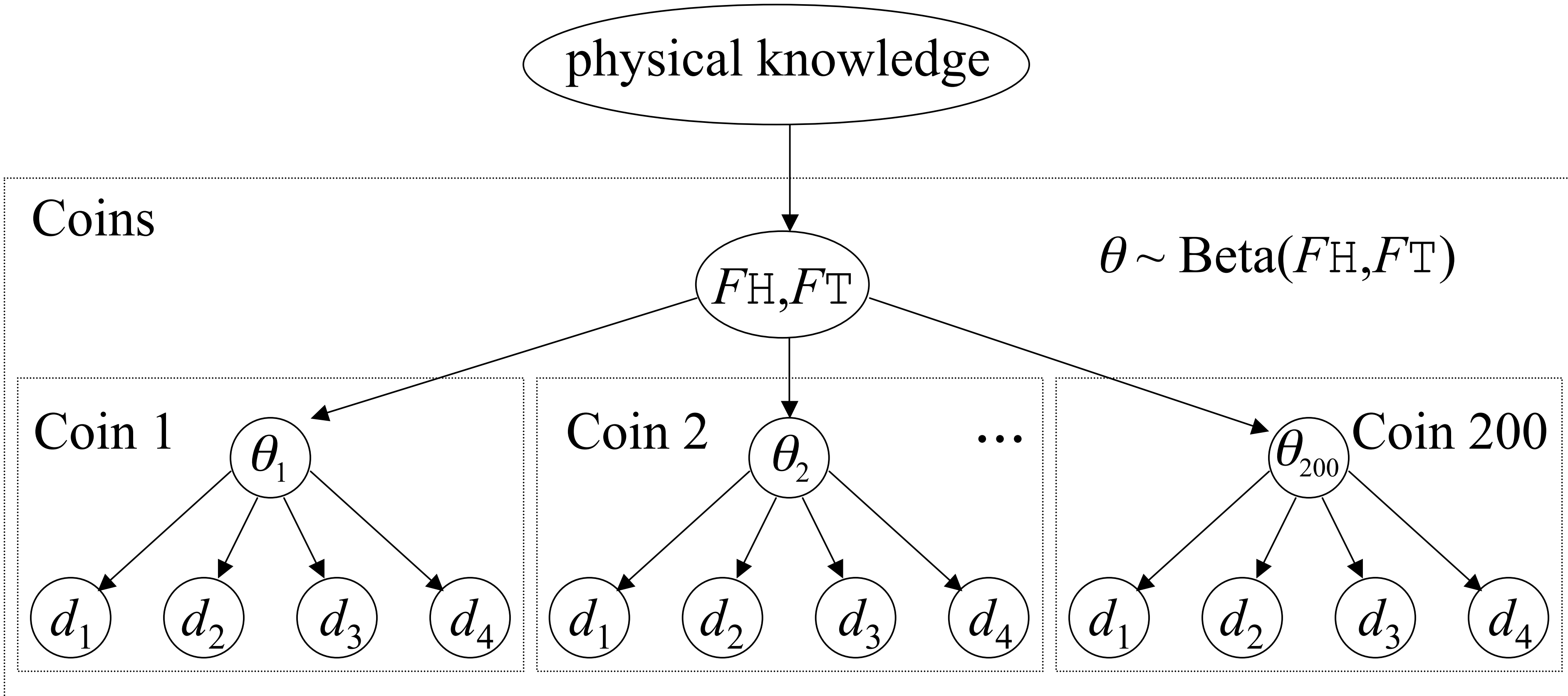
Words in general ↓ Individual words ↓ Data

Whole-object bias
Shape bias

car | monkey | duck | gavagai



A hierarchical Bayesian model



- Qualitative physical knowledge (symmetry) can influence estimates of continuous parameters (F_H, F_T).

- Explains why 10 flips of 200 coins are better than 2000 flips of a single coin: more informative about

Word Learning

"This is a dax."

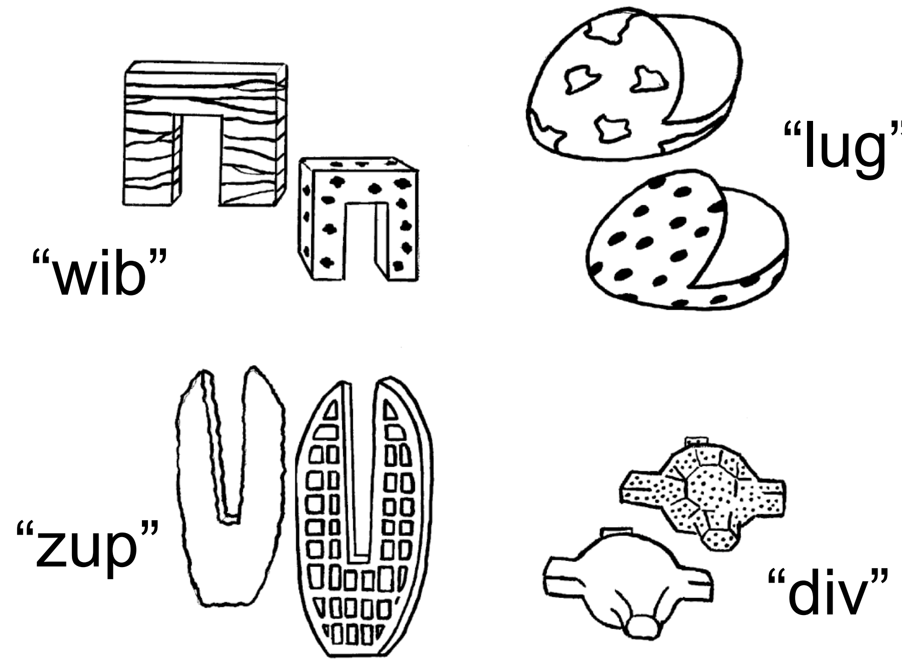
"Show me the dax."

- 24 month olds show a shape bias
- 20 month olds do not

(Landau, Smith & Gleitman)

Is the shape bias learned?

- Smith et al (2002) trained 17-month-olds on labels for 4 artificial categories:



- After 8 weeks of training 19-month-olds show the shape bias:

"This is a dax."

"Show me the dax."

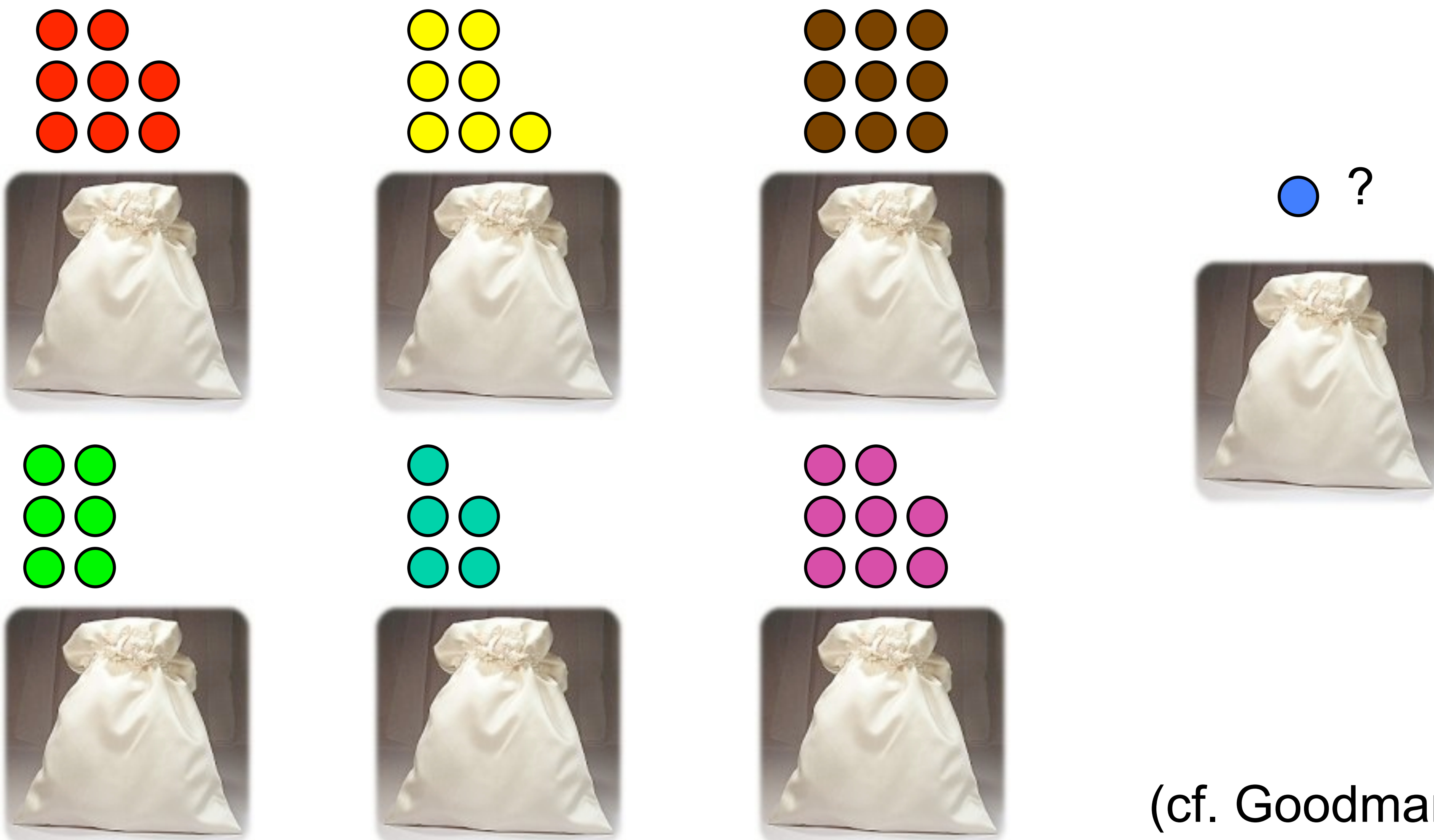
Learning about feature variability





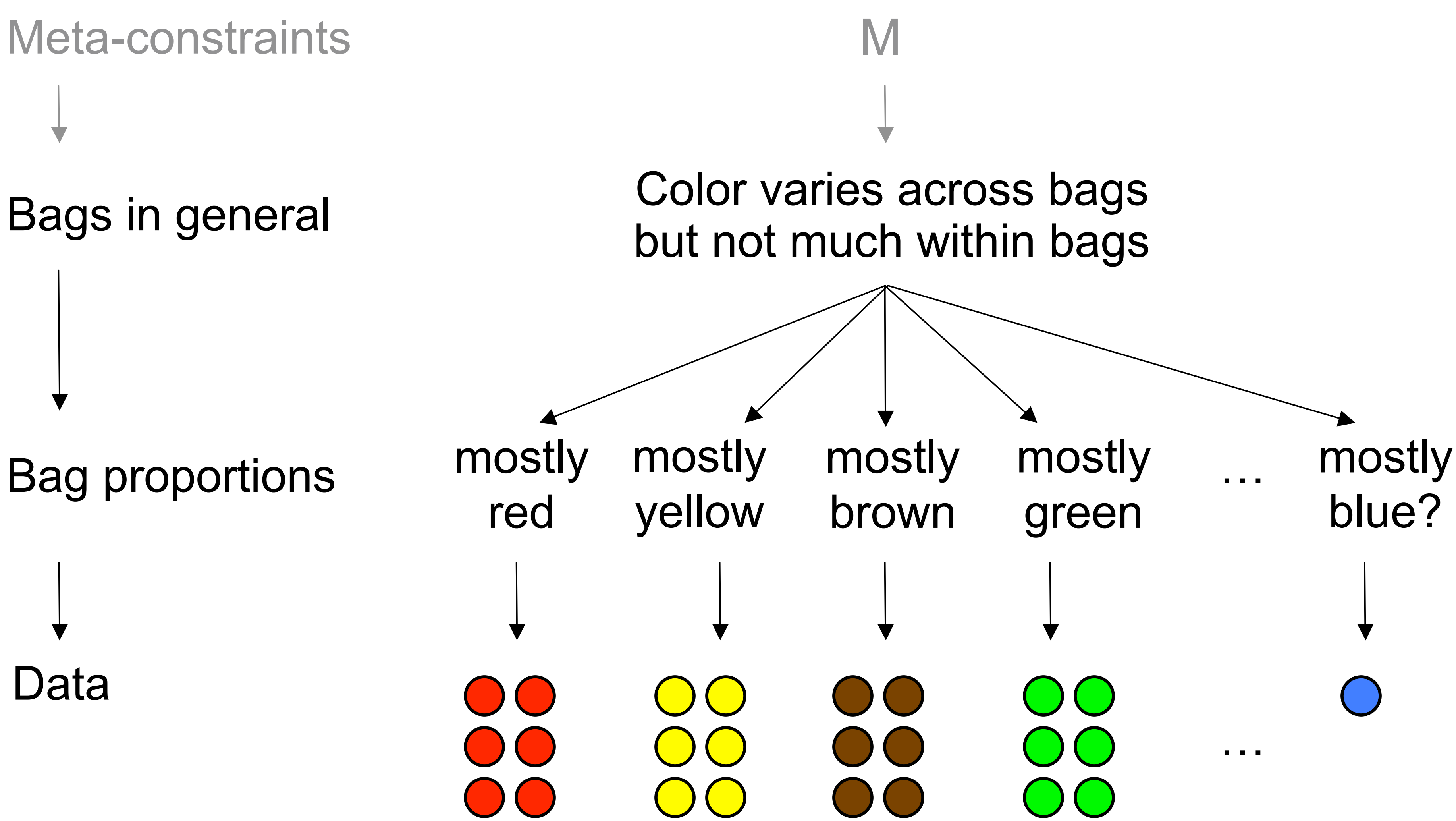

(cf. Goodman)

Learning about feature variability

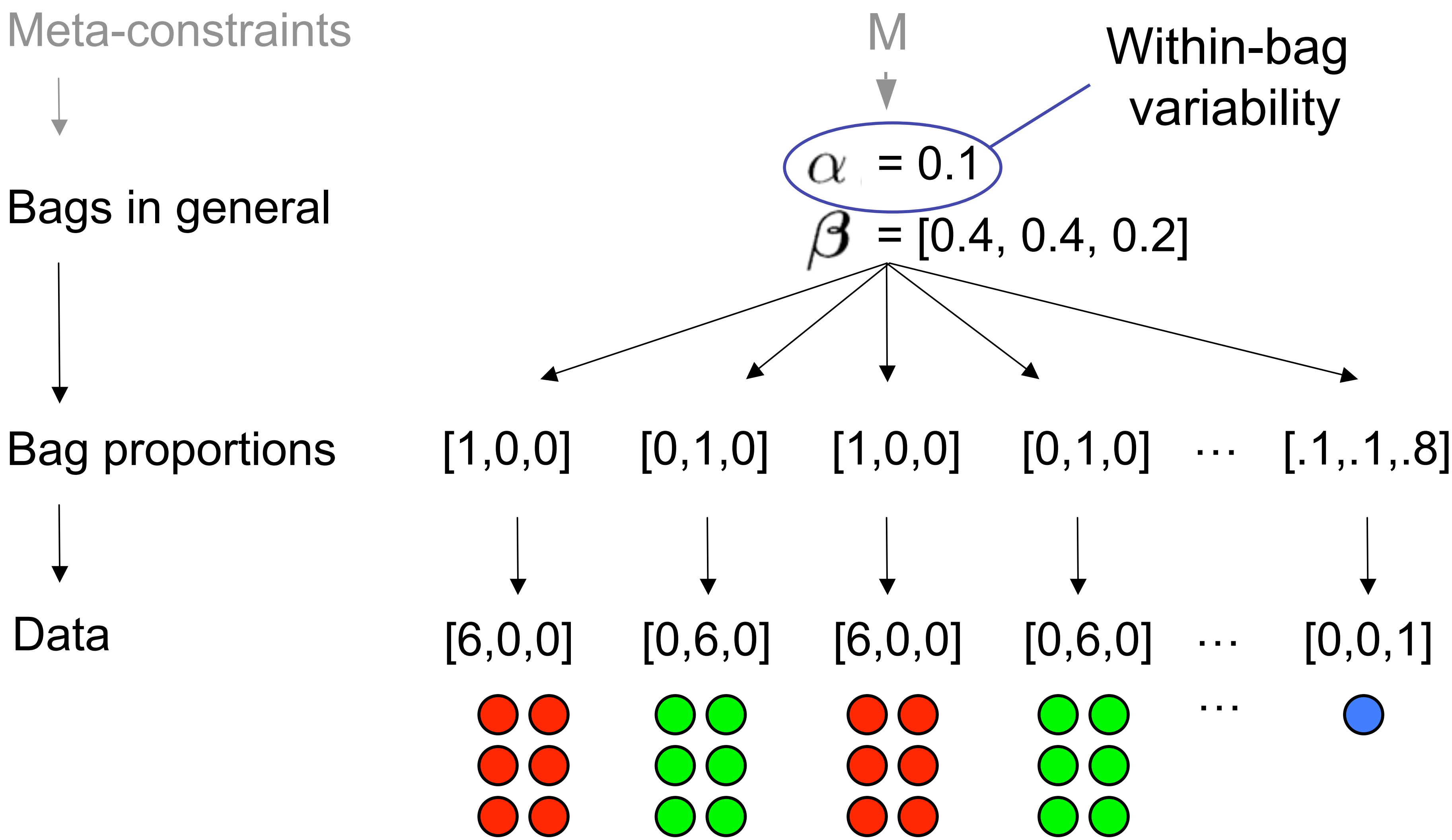


(cf. Goodman)

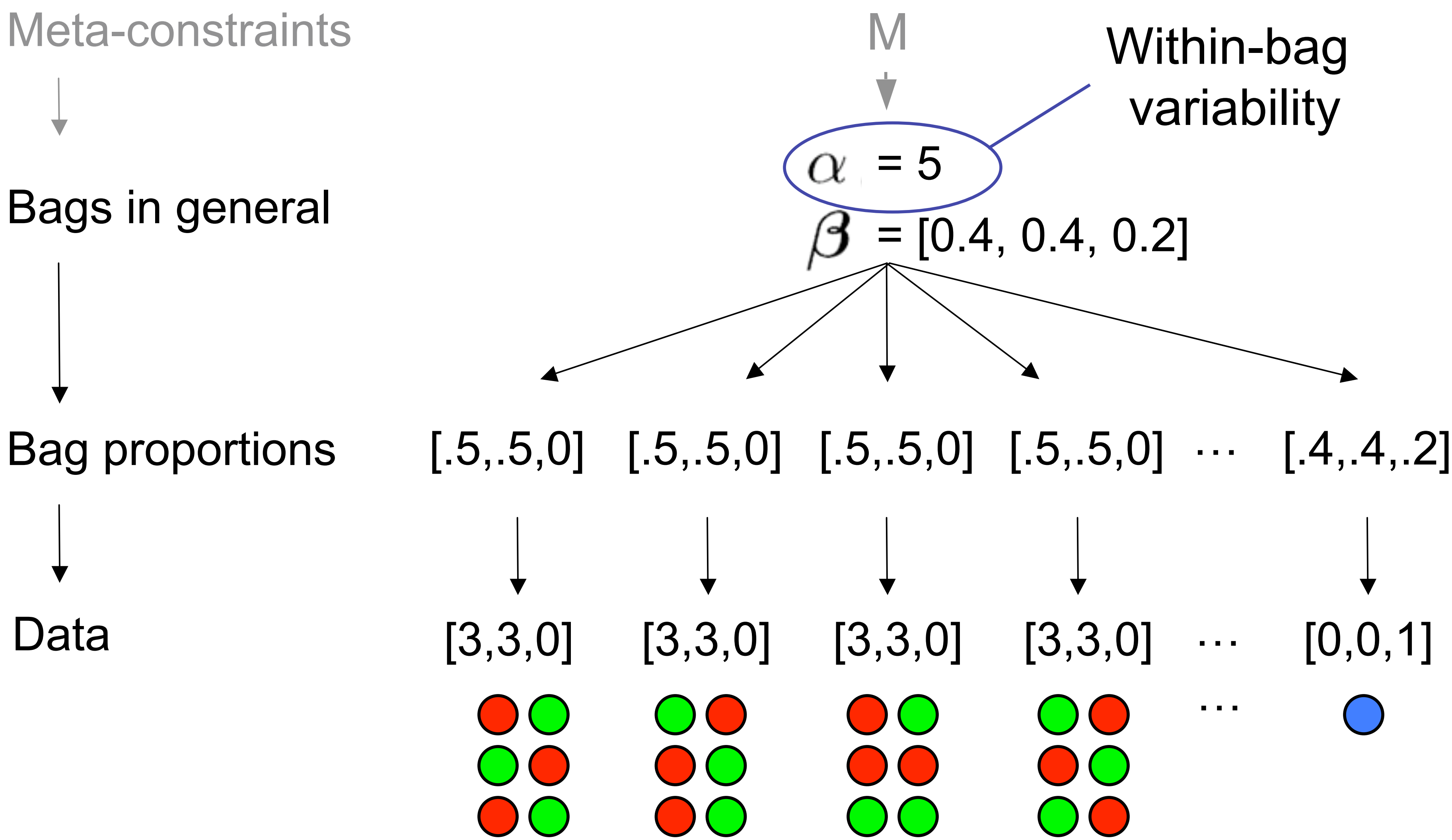
A hierarchical model



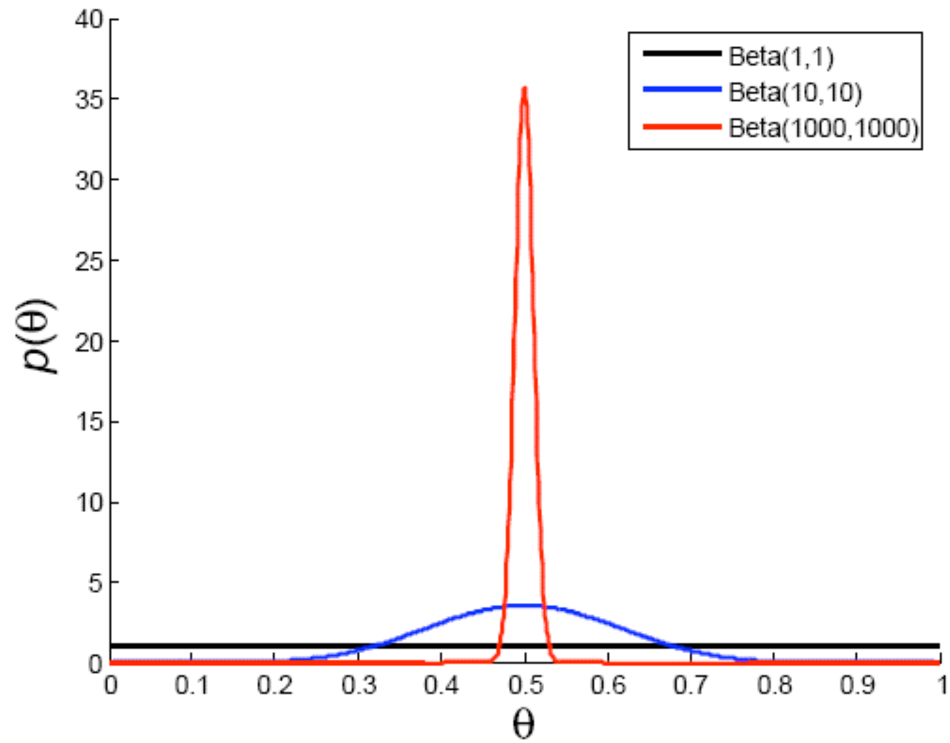
A hierarchical Bayesian model

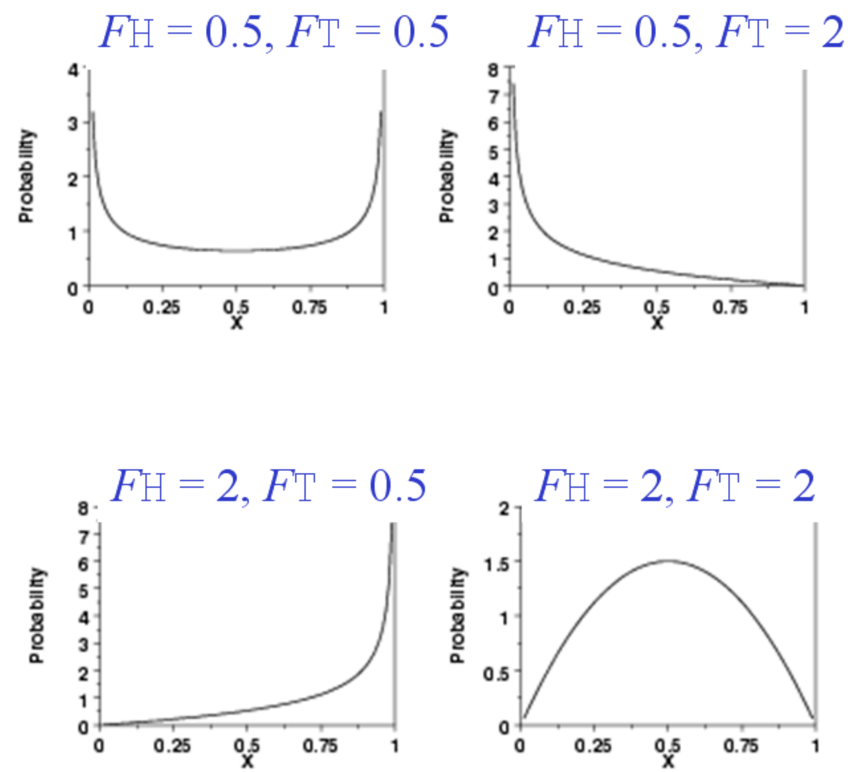


A hierarchical Bayesian model



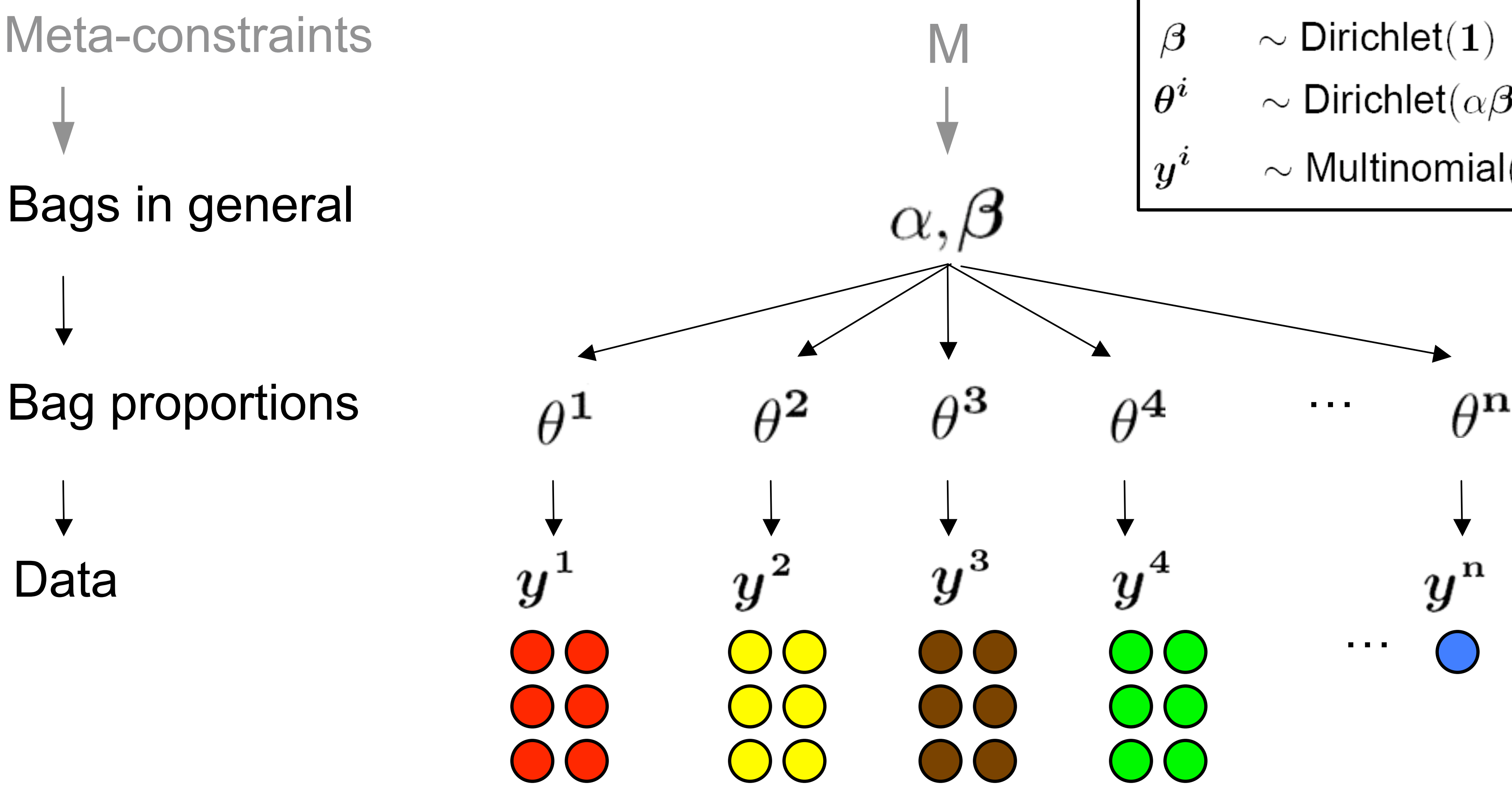
Shape of the Beta prior





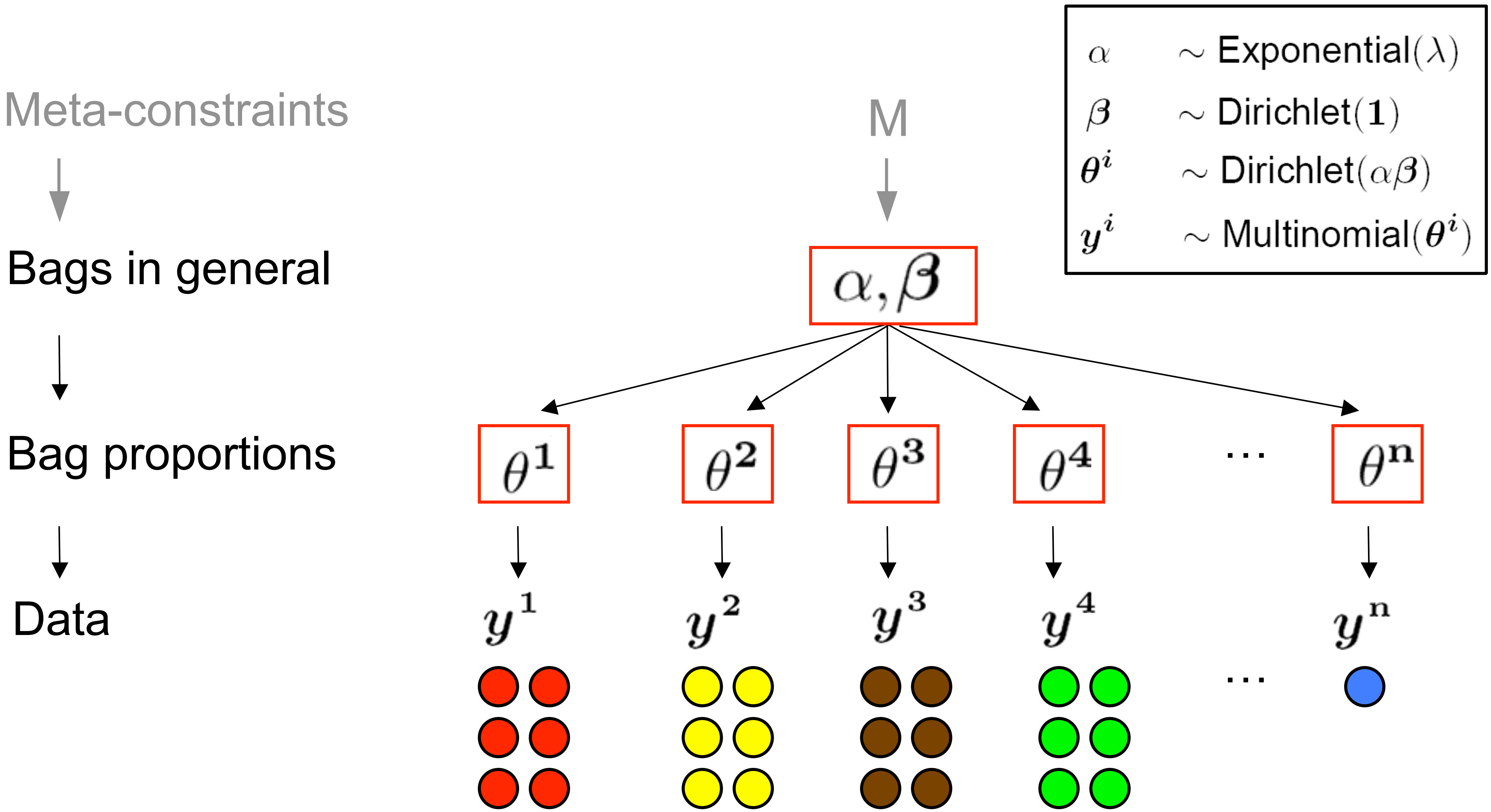
A hierarchical Bayesian model

$$
\begin{aligned}
\alpha &\sim \text{Exponential}(\lambda)\\
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1})\\
\boldsymbol{\theta}^i &\sim \text{Dirichlet}(\alpha\boldsymbol{\beta})\\
\boldsymbol{y}^i &\sim \text{Multinomial}(\boldsymbol{\theta}^i)
\end{aligned}
$$



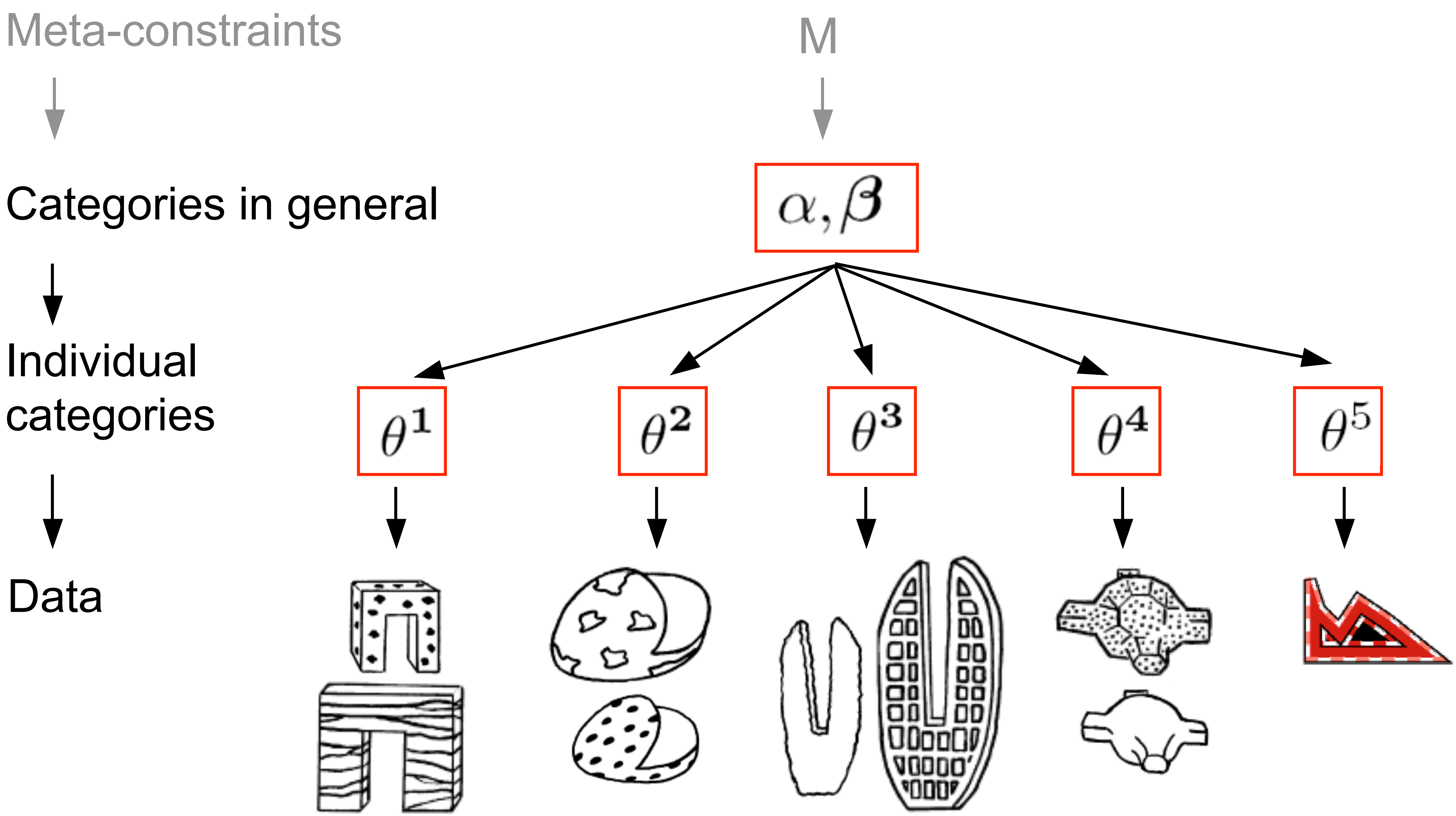
$$p(\{y^i\}, \{\theta^i\}, \alpha, \boldsymbol{\beta} | \lambda)$$

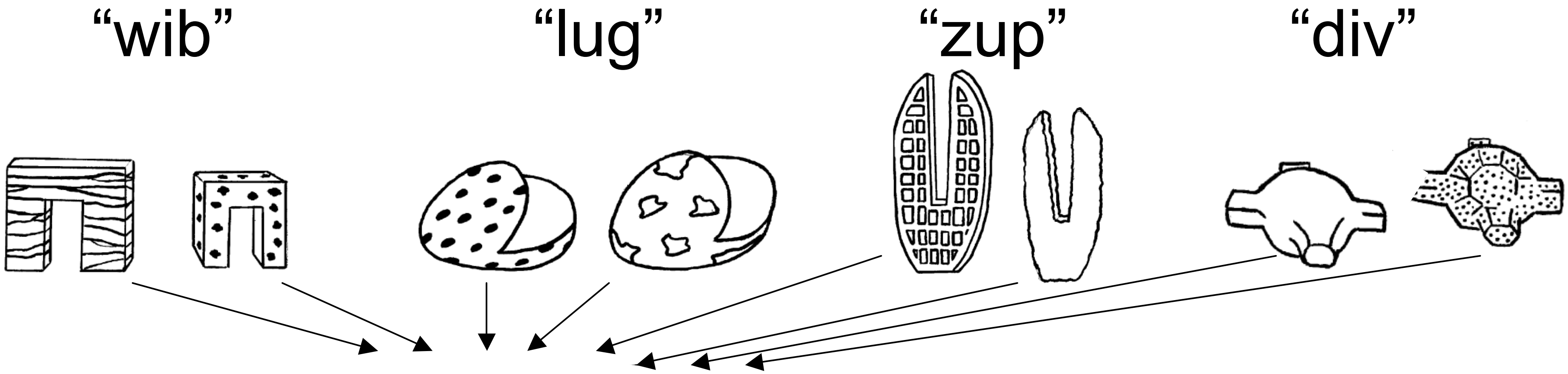
A hierarchical Bayesian model



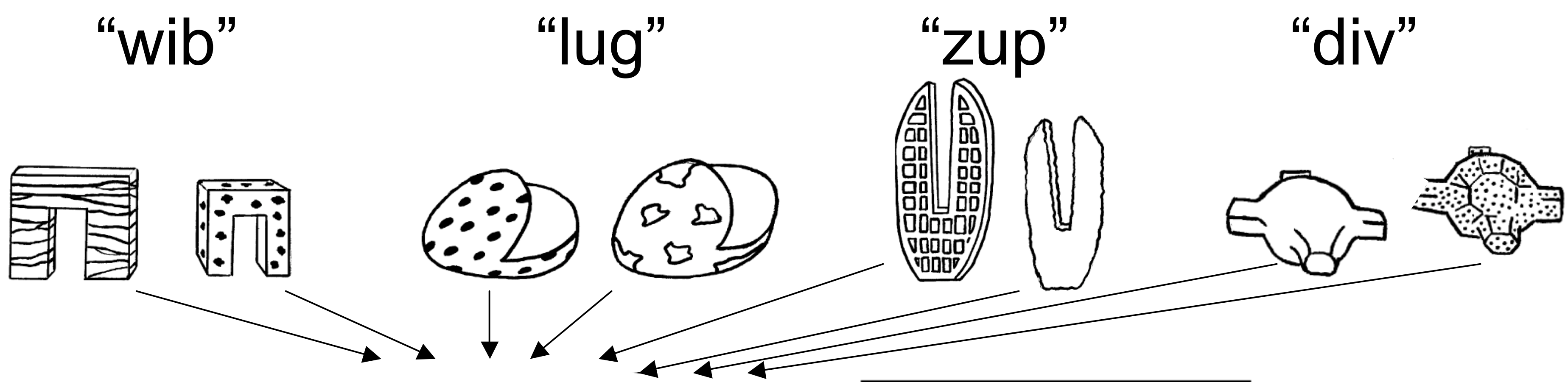
$$p(\{\theta^i\}, \alpha, \boldsymbol{\beta} | \{y^i\}, \lambda)$$

Learning about feature variability



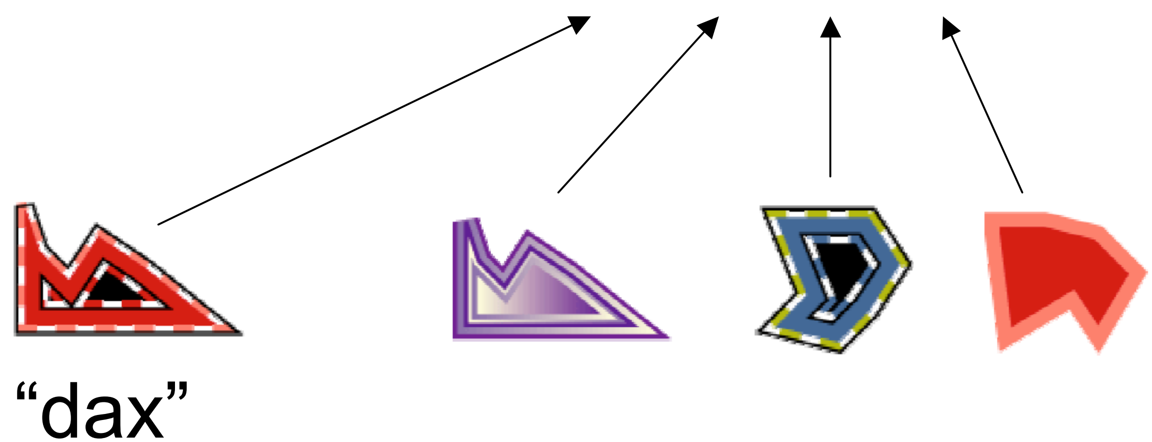


Category	1	1	2	2	3	3	4	4
Shape	1	1	2	2	3	3	4	4
Texture	1	2	3	4	5	6	7	8
Color	1	2	3	4	5	6	7	8
Size	1	2	1	2	1	2	1	2



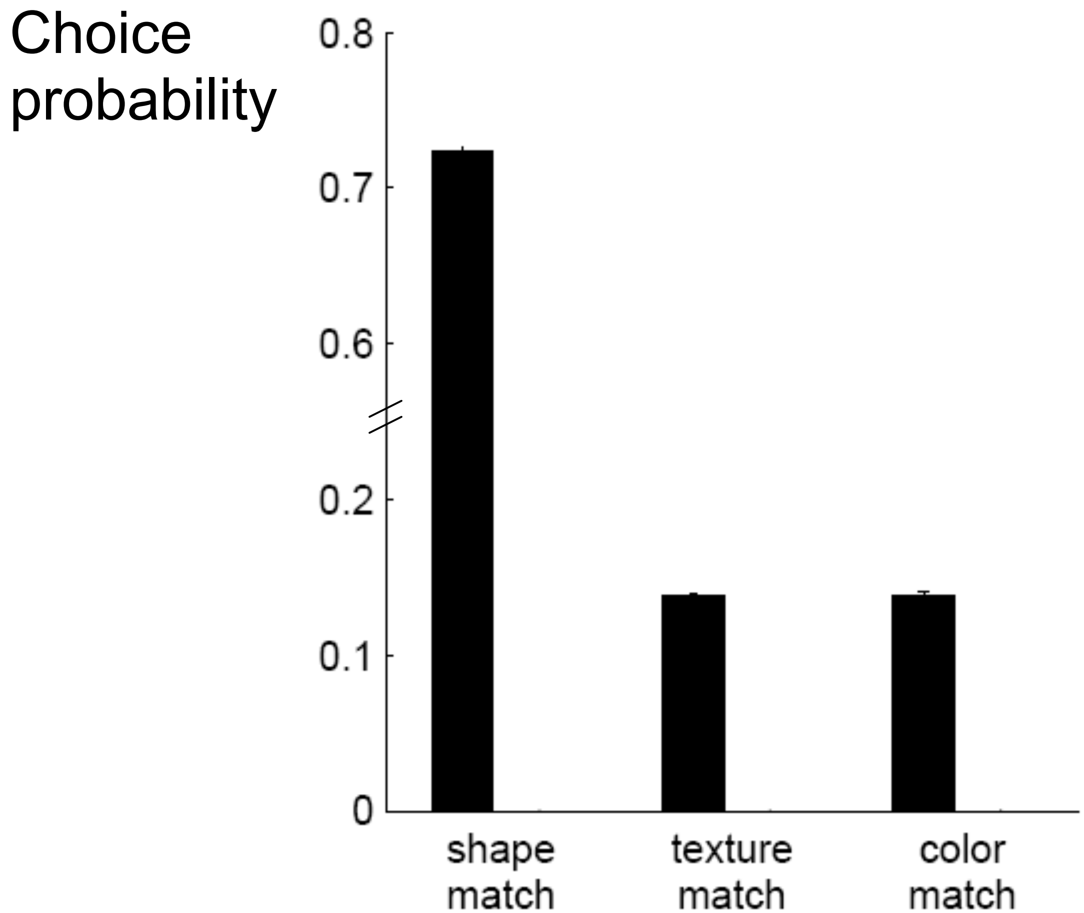
Category	1	1	2	2	3	3	4	4
Shape	1	1	2	2	3	3	4	4
Texture	1	2	3	4	5	6	7	8
Color	1	2	3	4	5	6	7	8
Size	1	2	1	2	1	2	1	2

5	?	?	?
5	5	6	6
9	10	9	10
9	10	10	9
1	1	1	1



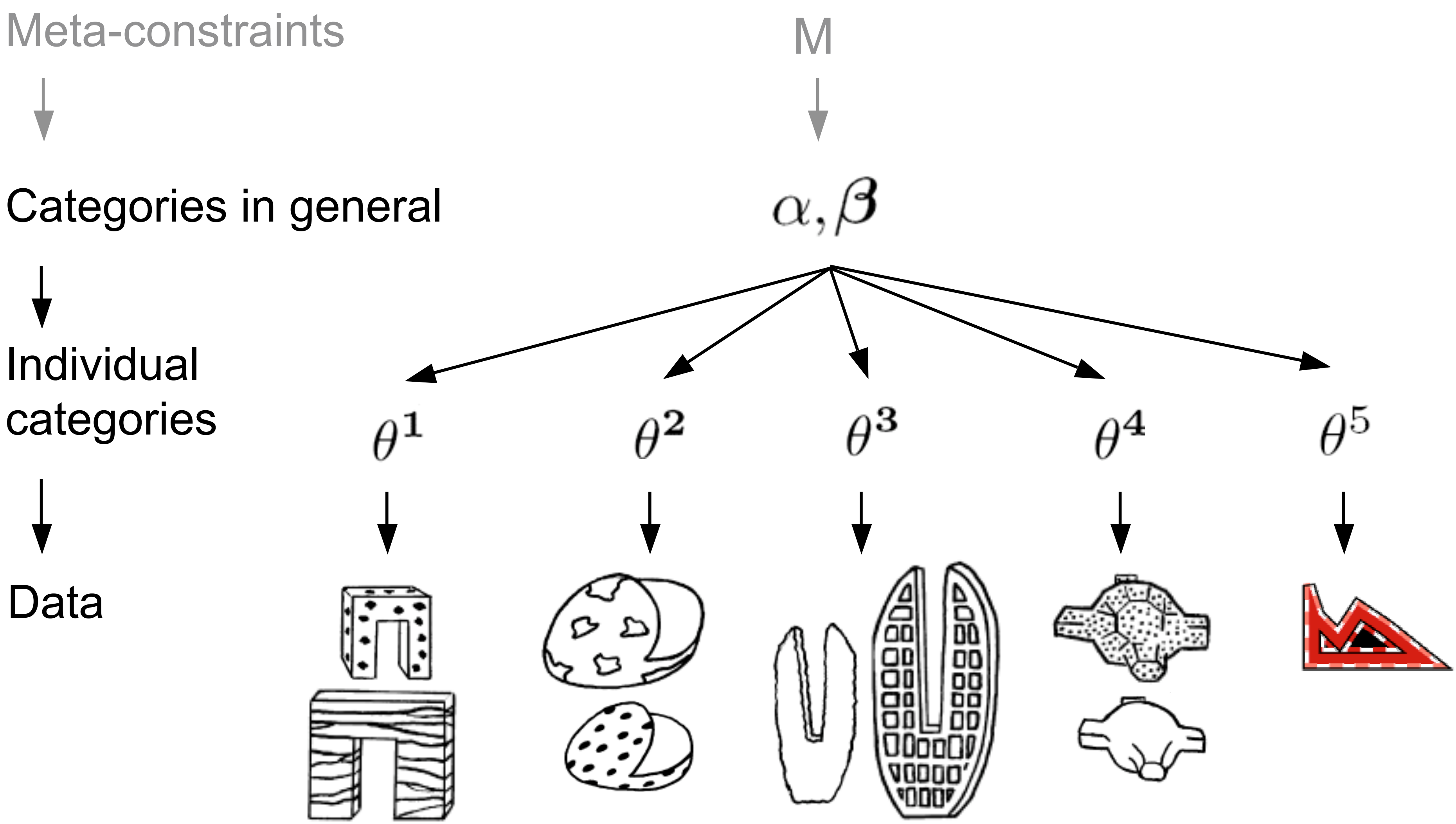
Model predictions

“dax”



“Show me the dax:”

Where do priors come from?



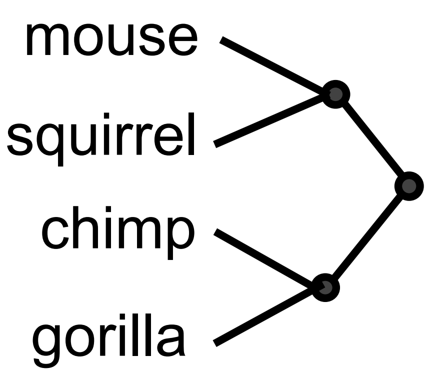
Knowledge representation

Mendeleev's Periodic Table of 1869[1]

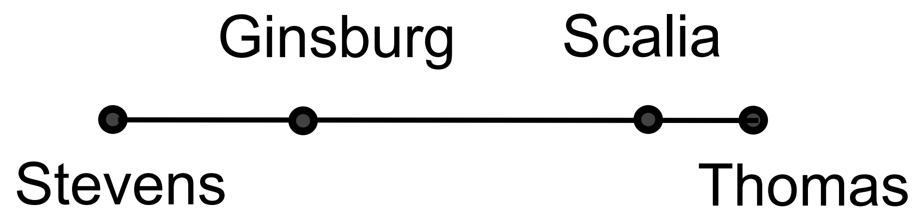
			Ti 50	Zr 90	? 100
			V 51	Nb 94	Ta 182
			Cr 52	Mo 96	W 186
			Mn 55	Rh 104.4	Pt 197.4
			Fe 56	Ru 104.4	Ir 198
			Ni, Co 59	Pd 106.6	Os 199
H 1			Cu 63.4	Ag 108	Hg 200
	Be 9.4	Mg 24	Zn 65.2	Cd 112	
	B 11	Al 27.4	? 68	U 116	Au 197?
	C 12	Si 28	? 70	Sn 118	
	N 14	P 31	As 75	Sb 122	Bi 210?
	O 16	S 32	Se 79.4	Te 128?	
	F 19	Cl 35.5	Br 80	I 127	
Li 7	Na 23	K 39	Rb 85.4	Cs 133	Tl 204
		Ca 40	Sr 87.6	Ba 137	Pb 207
		? 45	Ce 92		
		Er? 56	La 94		
		Yt? 60	Di 95		
		In 75.6?	Th 118?		

The discovery of structural form

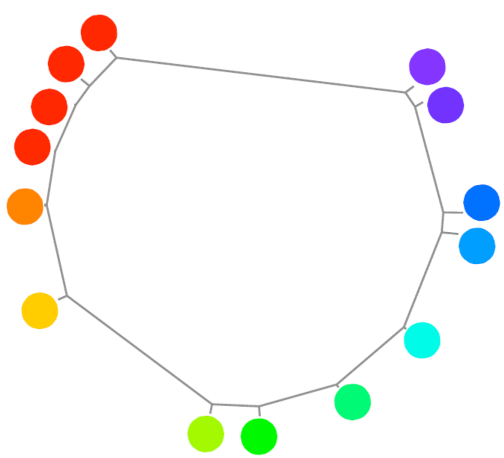
BIOLOGY



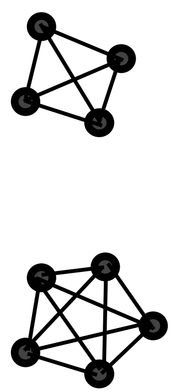
POLITICS



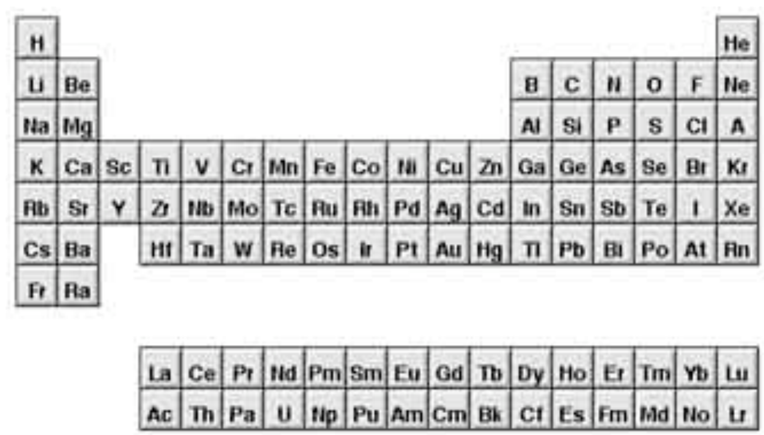
COLOR



FRIENDSHIP



CHEMISTRY



Children discover structural form

- Children may discover that
  - Social networks are often organized into cliques
  - The months form a cycle
  - “Heavier than” is transitive
  - Category labels can be organized into hierarchies

A hierarchical Bayesian model

Meta-constraints

↓

Form

↓

Structure

↓

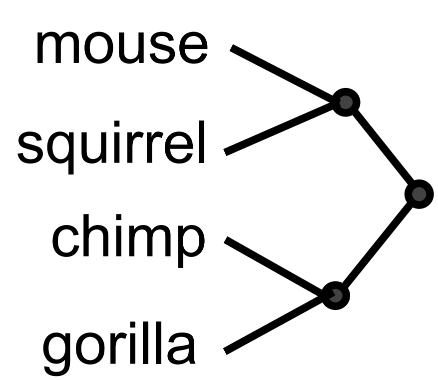
Data

M

↓

Tree

↓



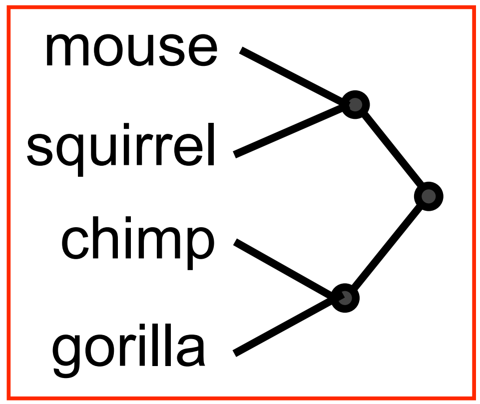
↓

	whiskers	hands	tail
mouse	●	○	●
squirrel	●	○	●
chimp	○	●	○
gorilla	○	●	○

A hierarchical Bayesian model

Meta-constraints → F: form → S: structure → D: data

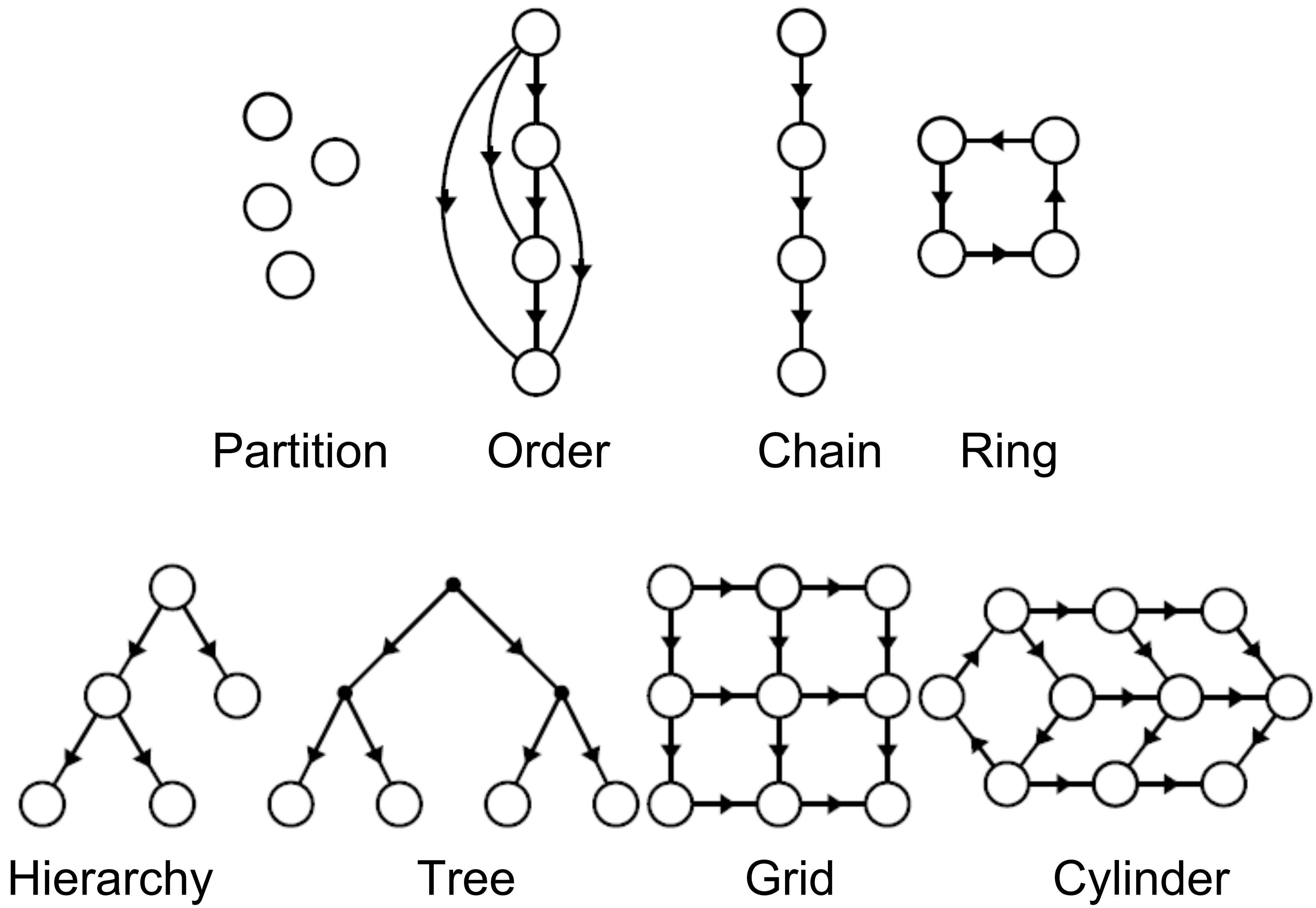
M → Tree →



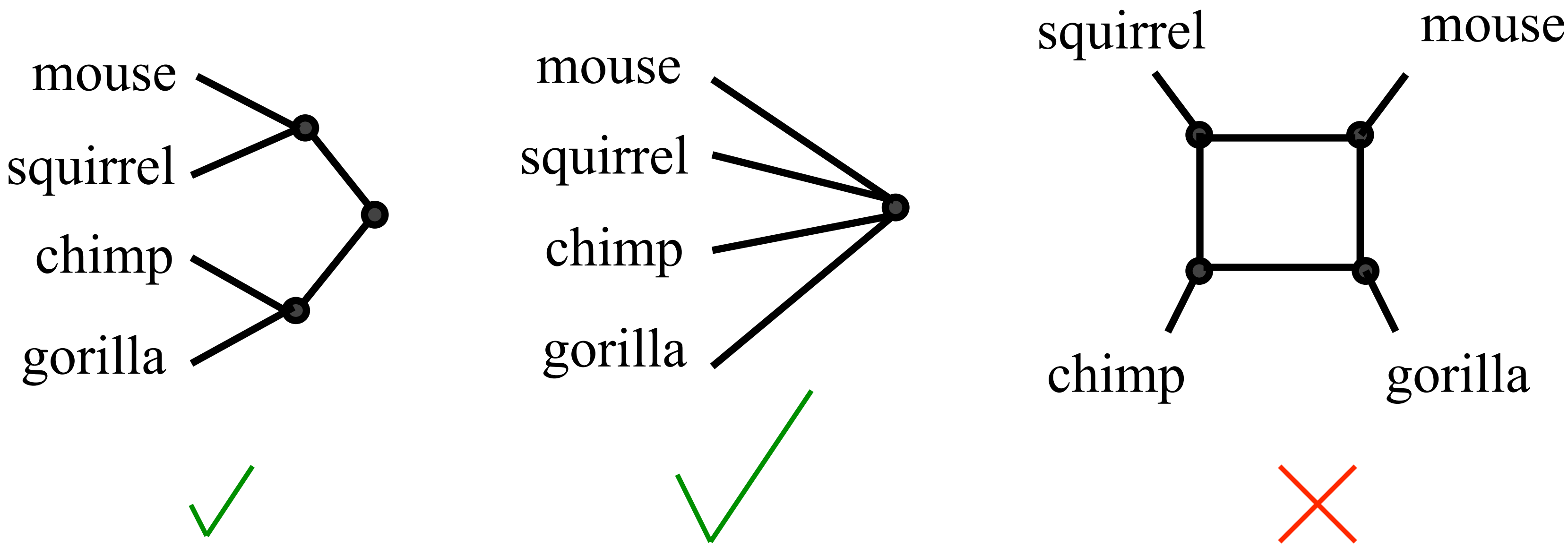
	whiskers	hands	tail
mouse	●	○	●
squirrel	●	○	●
chimp	○	●	○
gorilla	○	●	○

$$P(S, F | D, n) \propto P(D|S) P(S|F, n) P(F)$$

Structural forms



P(S|F,n): Generating structures



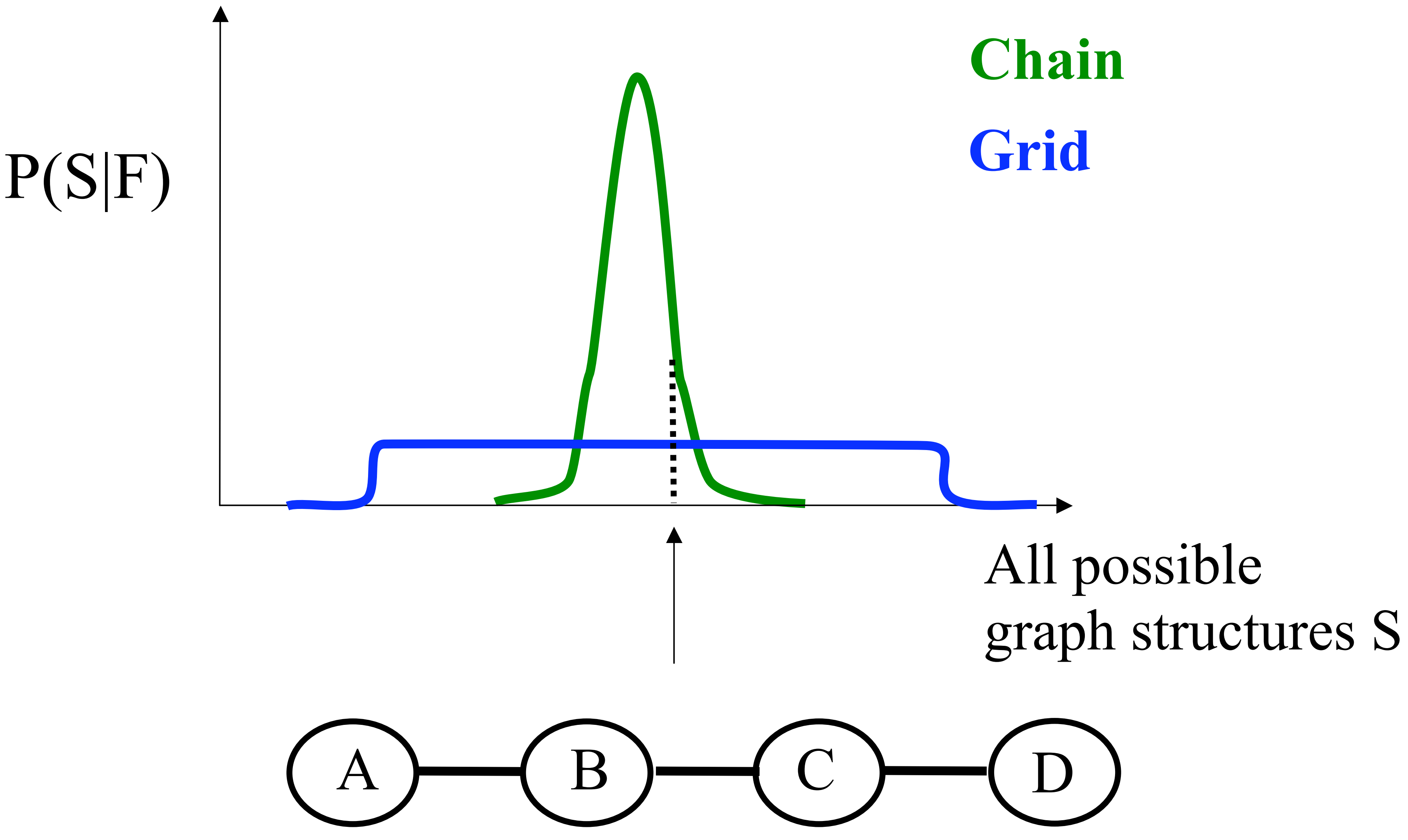
- Each structure is weighted by the number of nodes it contains:

$$P(S|F) \propto \begin{cases} 0 & \text{if S inconsistent with F} \\ \theta(1-\theta)^{|S|} & \text{otherwise} \end{cases}$$

where $|S|$ is the number of nodes in *S*

P(S|F, n): Generating structures from forms

- Simpler forms are preferred



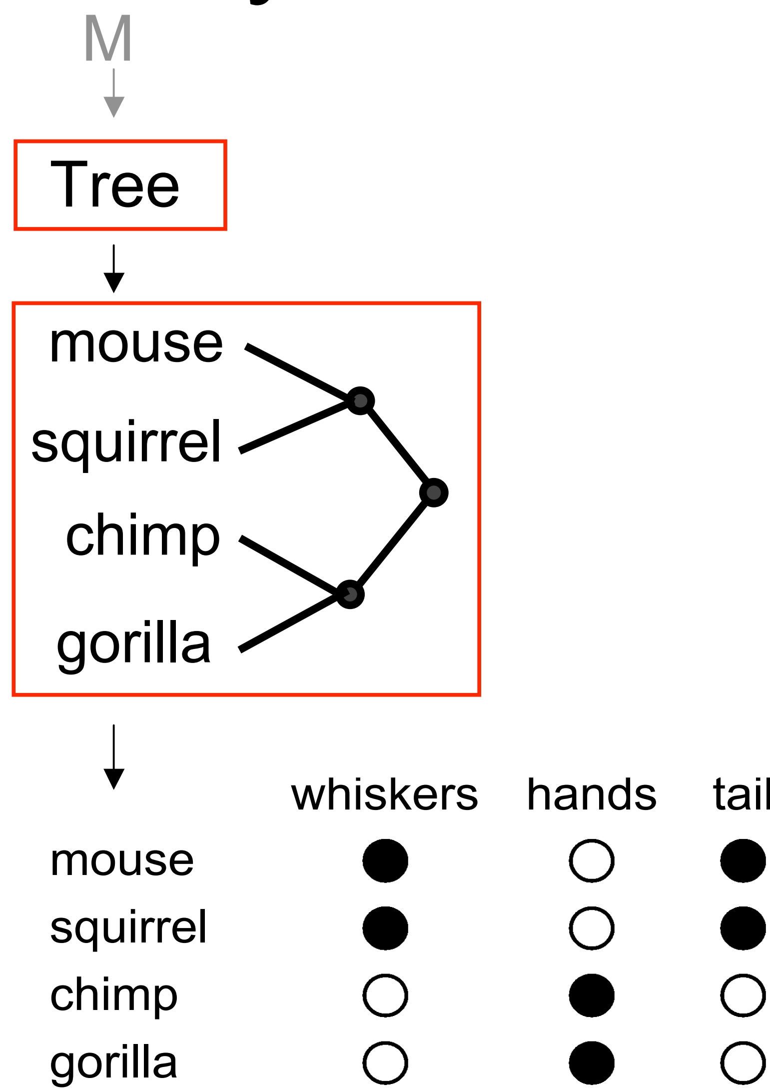
A hierarchical Bayesian model

Meta-constraints

F: form

S: structure

D: data

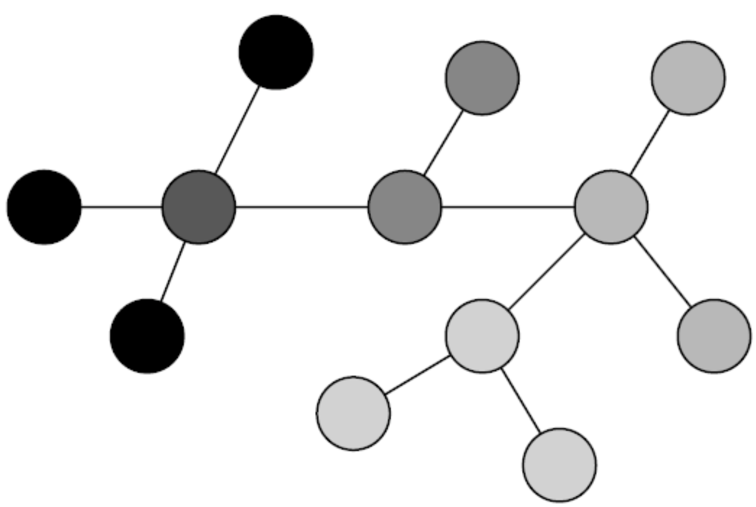


$$P(S, F|D, n) \propto P(D|S)P(S|F, n)P(F)$$

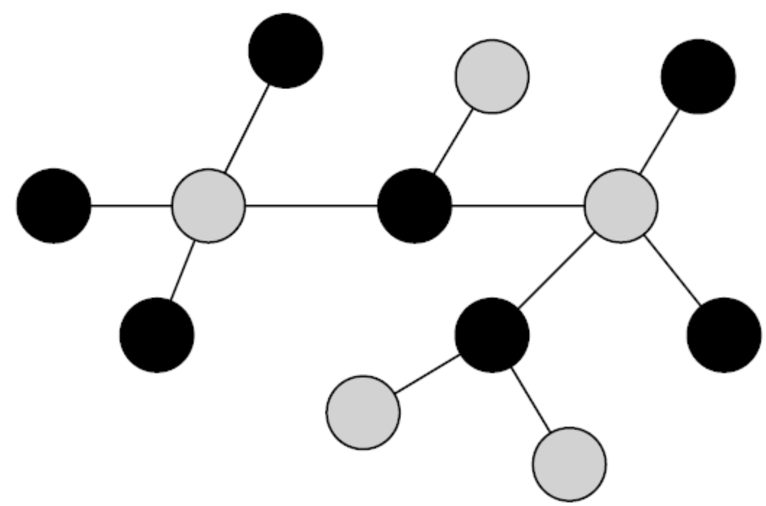
p(D|S): Generating feature data

- Intuition: features should be smooth over graph S

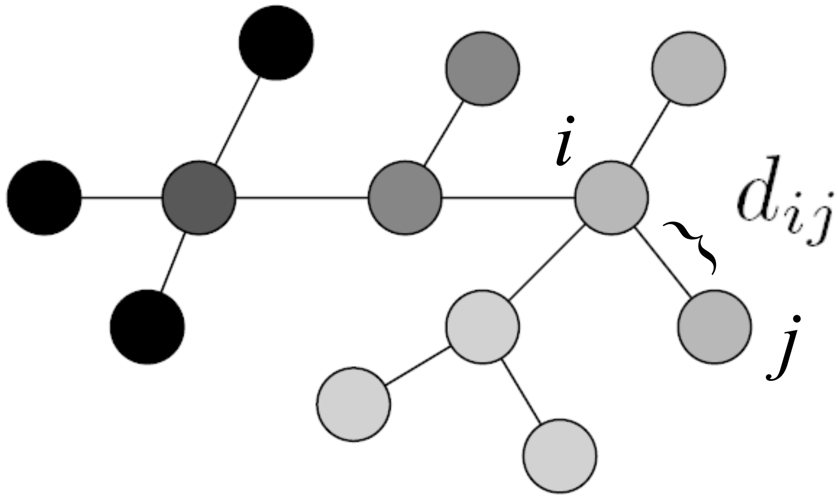
Relatively smooth



Not smooth



p(D|S): Generating feature data



Let f_i be the feature value at node *i*

$$p(f) \propto \exp\left(-\frac{1}{4}\sum_{i,j}\frac{(f_i - f_j)^2}{d_{ij}} - \frac{1}{2\sigma}f^\top f\right)$$

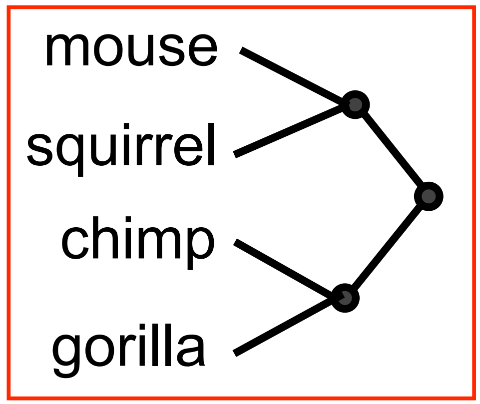
(Zhu, Lafferty & Ghahramani)

A hierarchical Bayesian model

Meta-constraints → M

F: form → Tree

S: structure

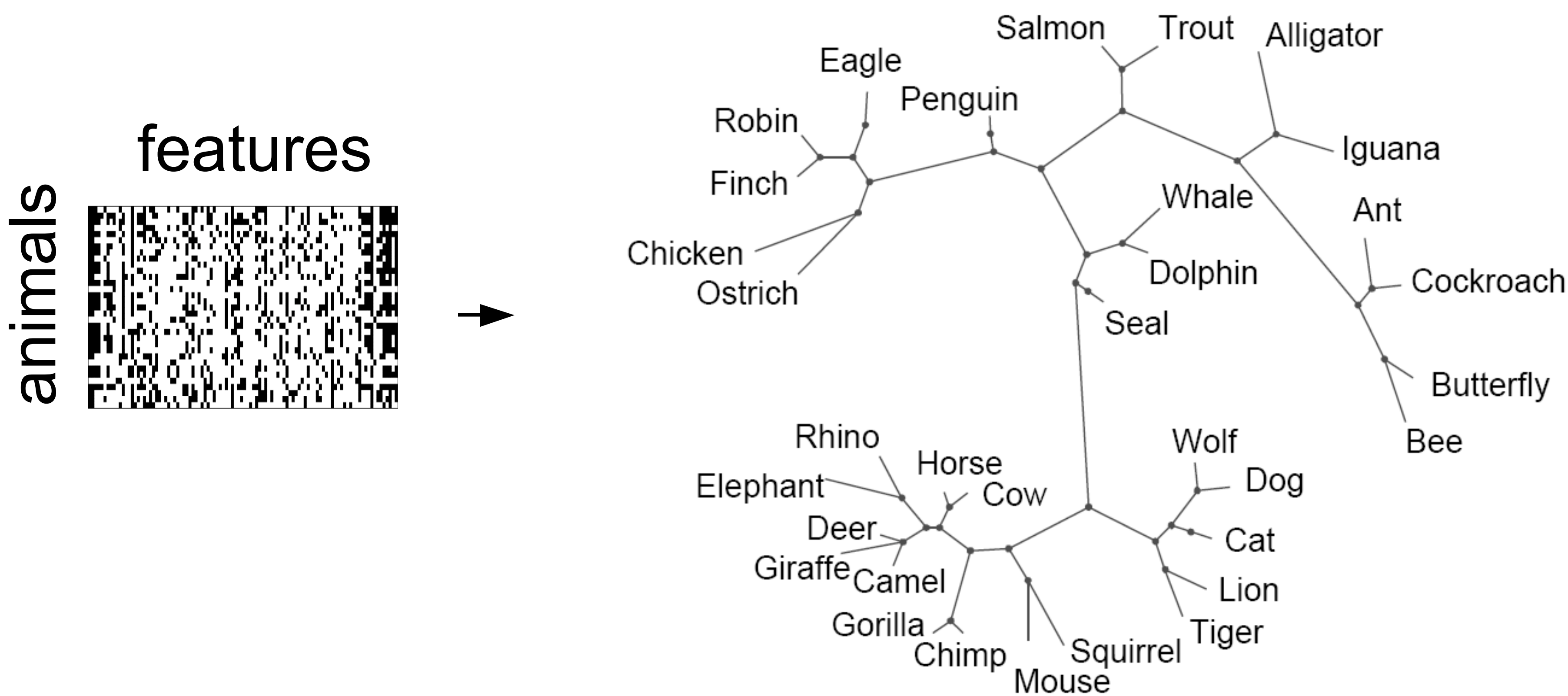


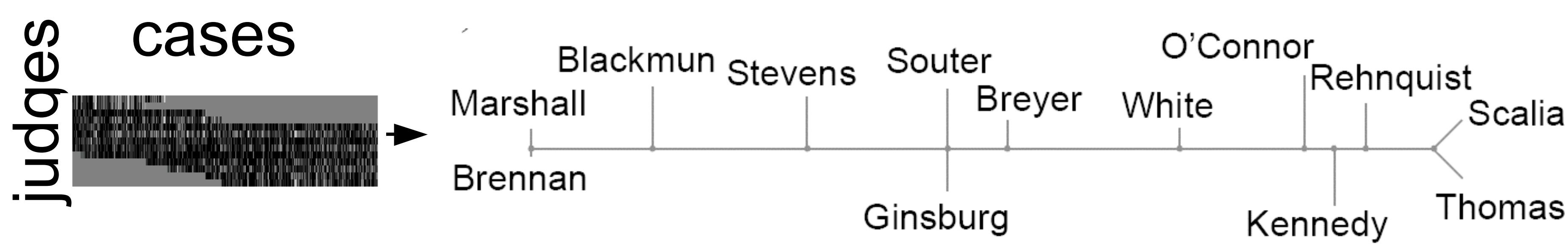
D: data

	whiskers	hands	tail
mouse	●	○	●
squirrel	●	○	●
chimp	○	●	○
gorilla	○	●	○

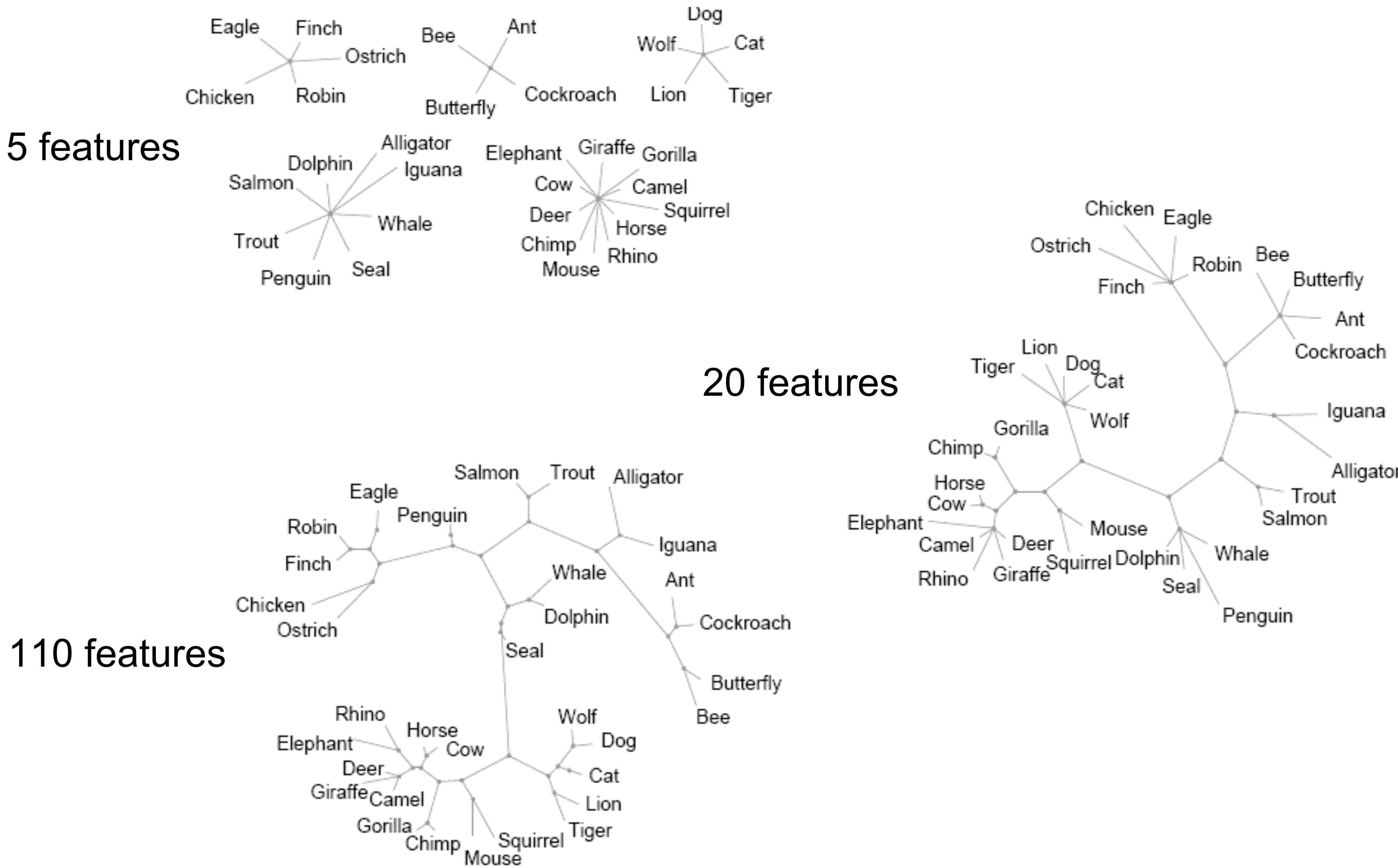
$$P(S, F|D, n) \propto P(D|S)P(S|F, n)P(F)$$

Feature data: results

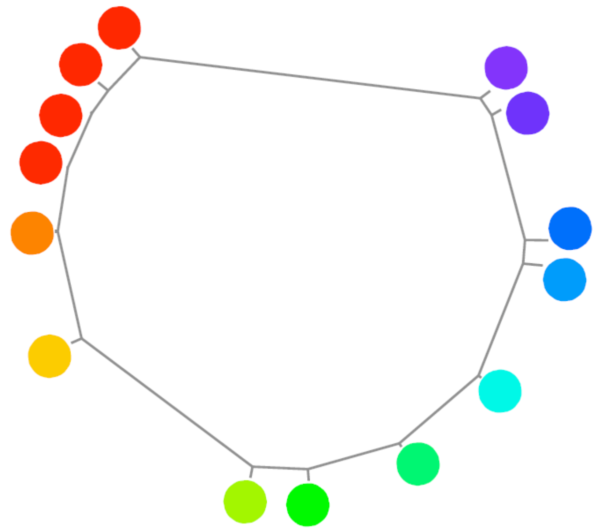


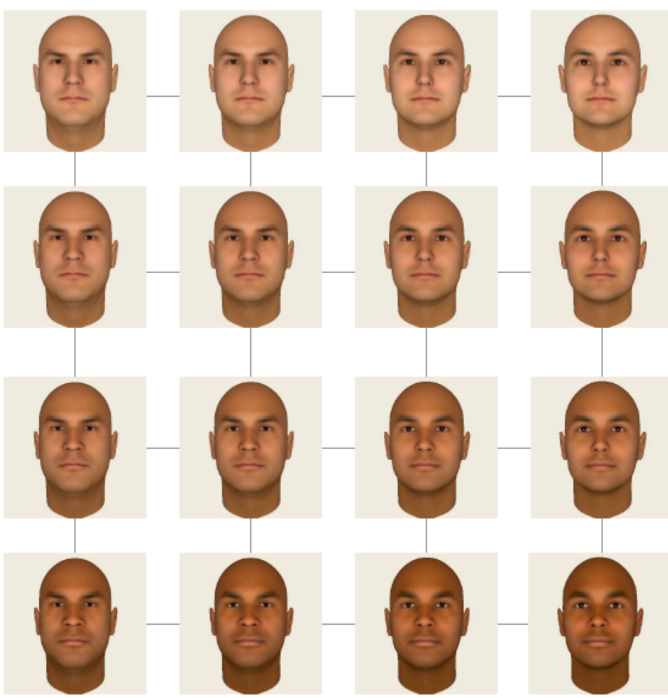


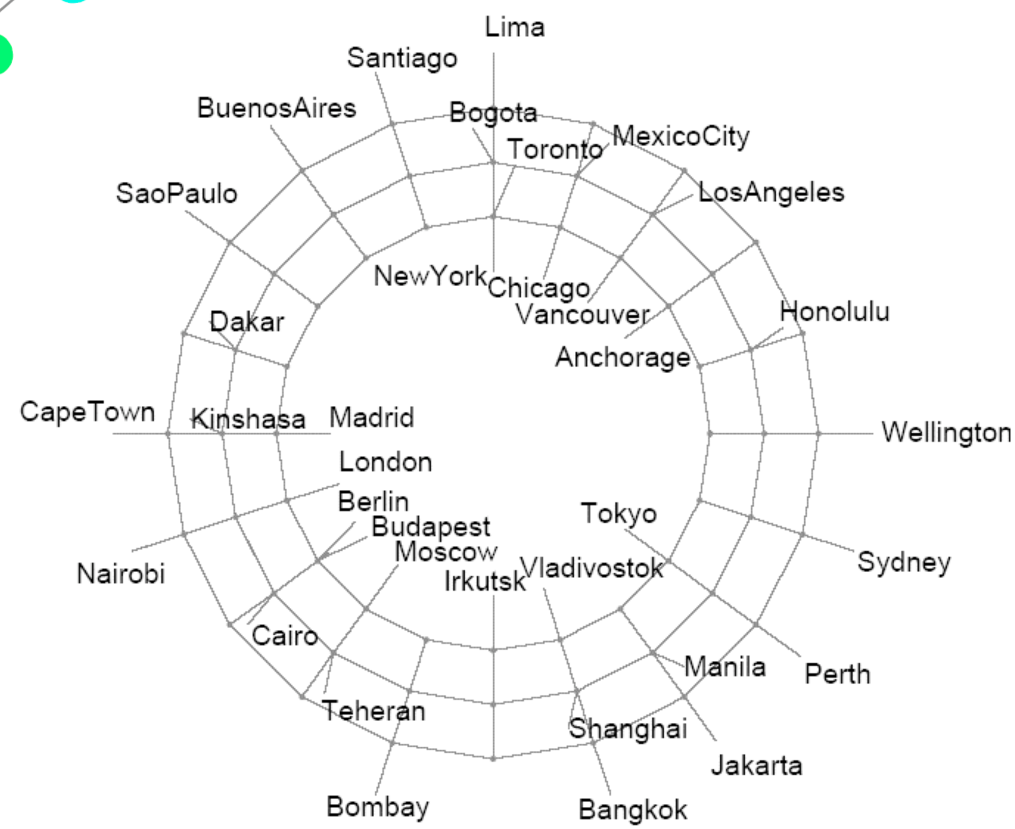
Developmental shifts



Similarity data: results





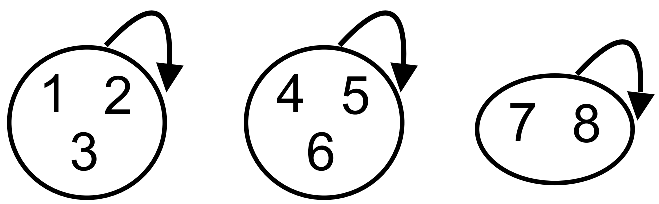


Relational data

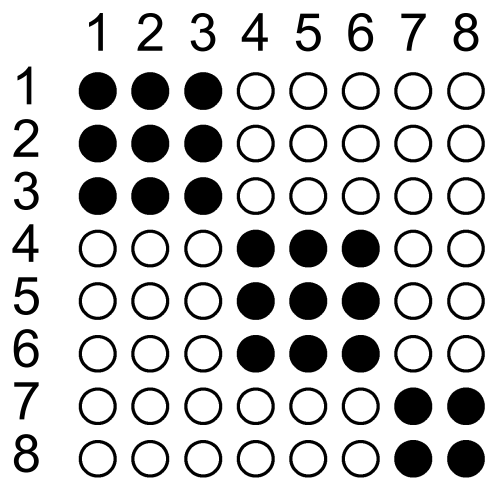
Meta-constraints — M

Form — Cliques

Structure



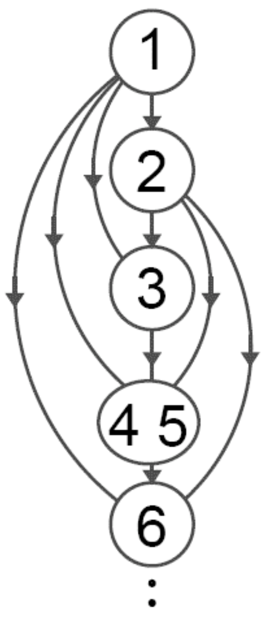
Data



Relational data: results

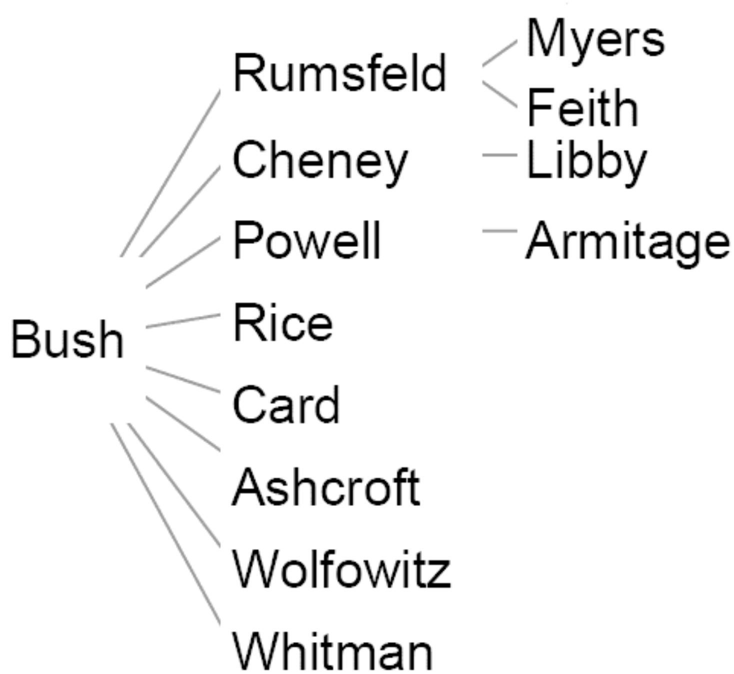
Primates

"x dominates y"



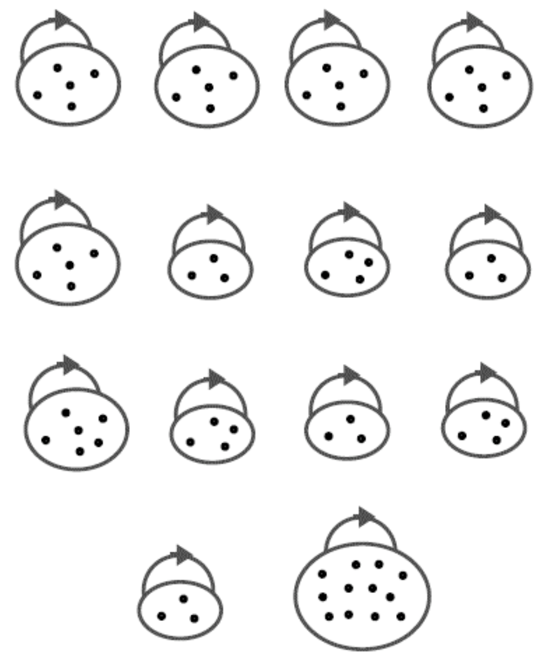
Bush cabinet

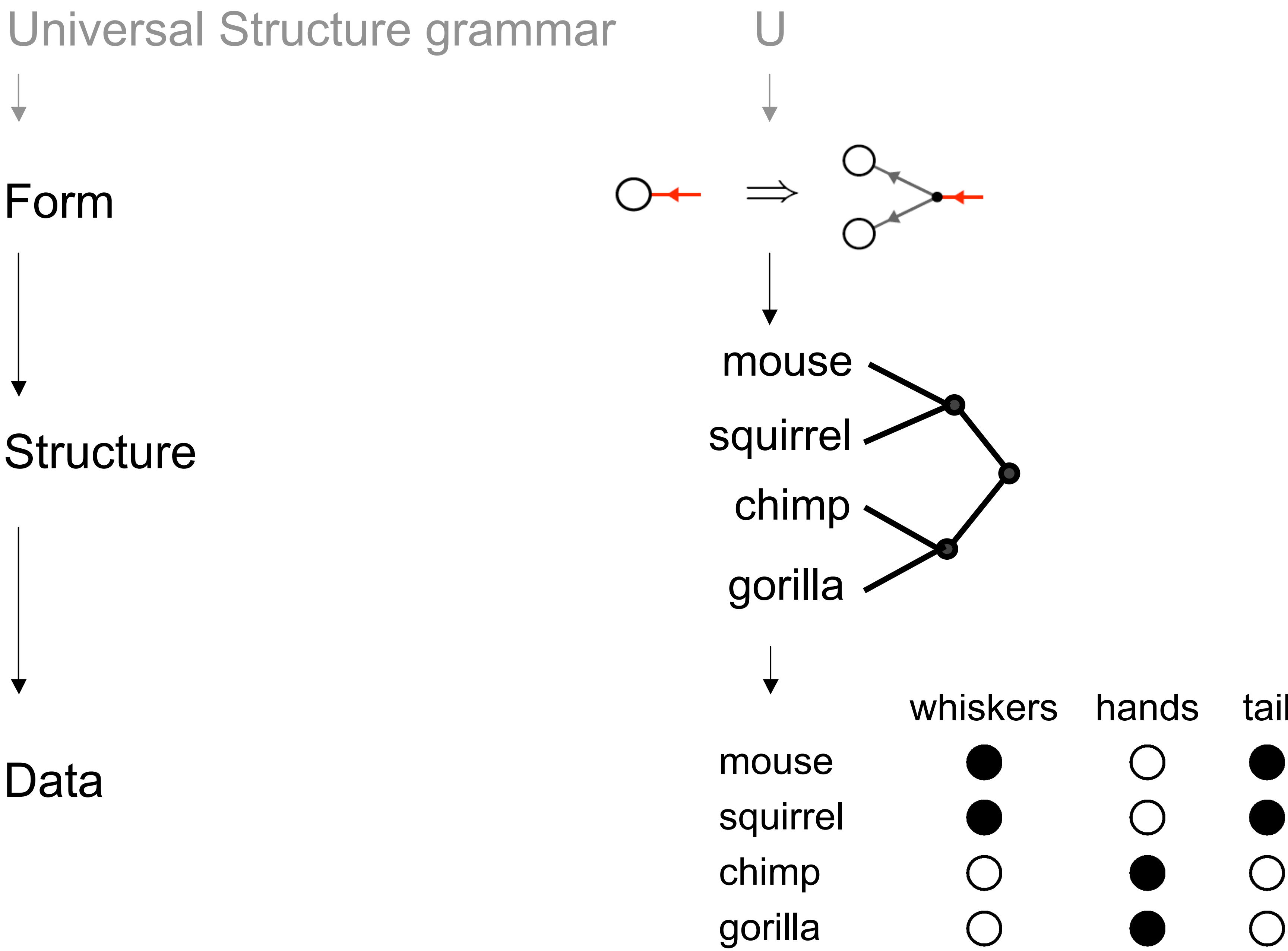
"x tells y"



Prisoners

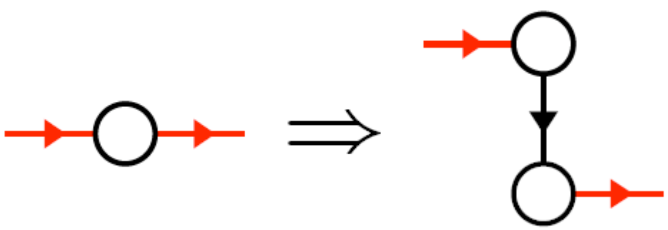
"x is friends with y"



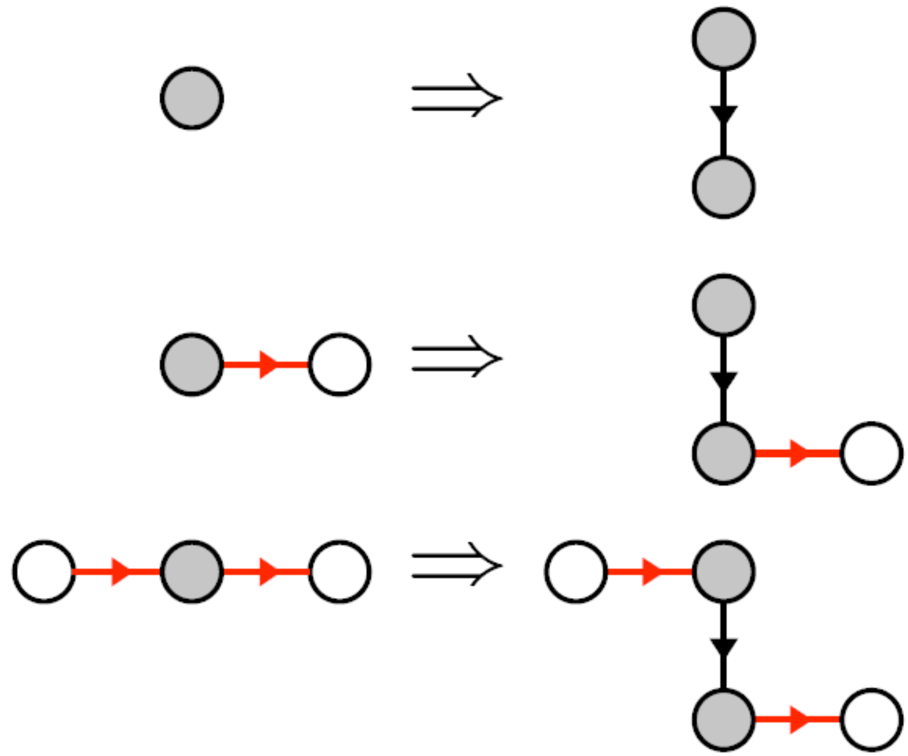

Universal Structure grammar
U
Form
Structure
Data
mouse
squirrel
chimp
gorilla
whiskers
hands
tail
mouse
squirrel
chimp
gorilla

Node-replacement graph grammars

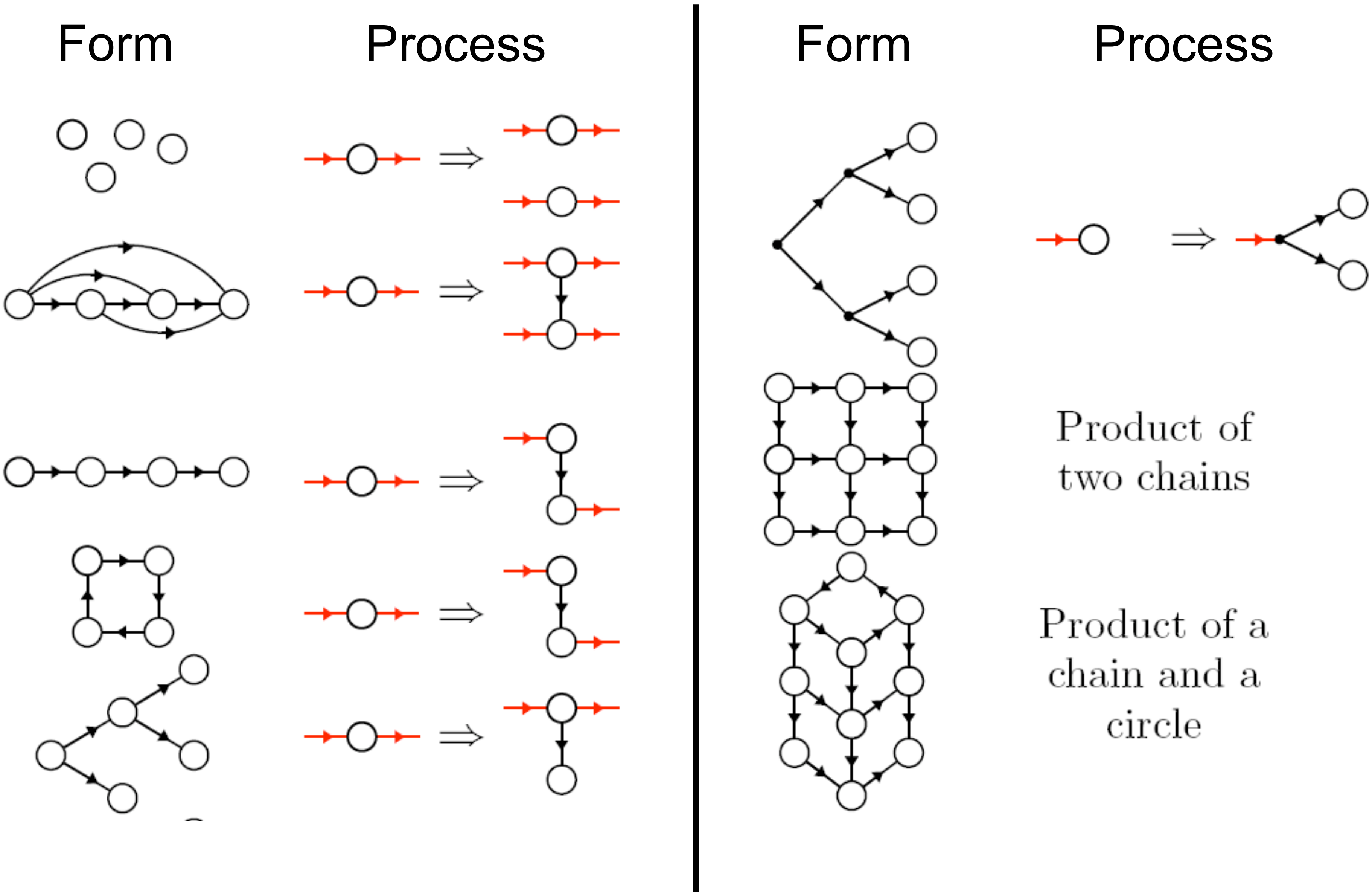
Production
(Chain)



Derivation

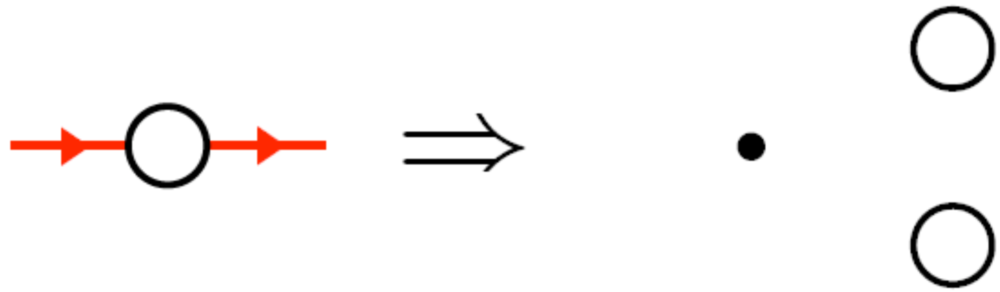


A hypothesis space of forms



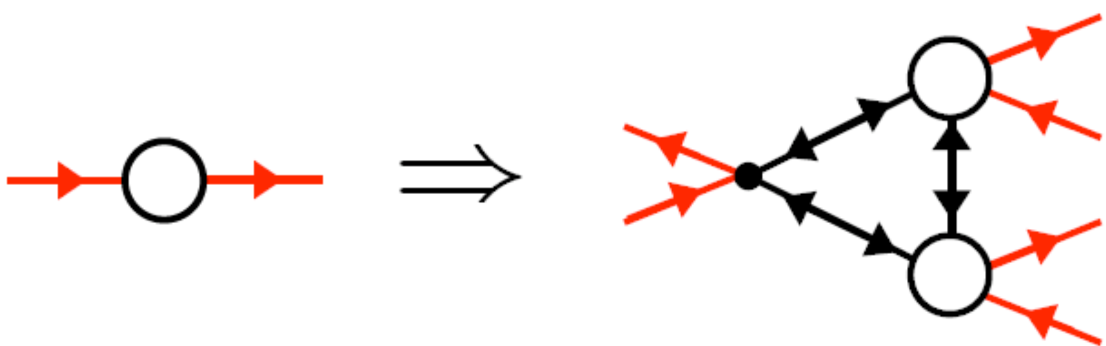
The complete space of grammars

1



⋮ ⋮

4096

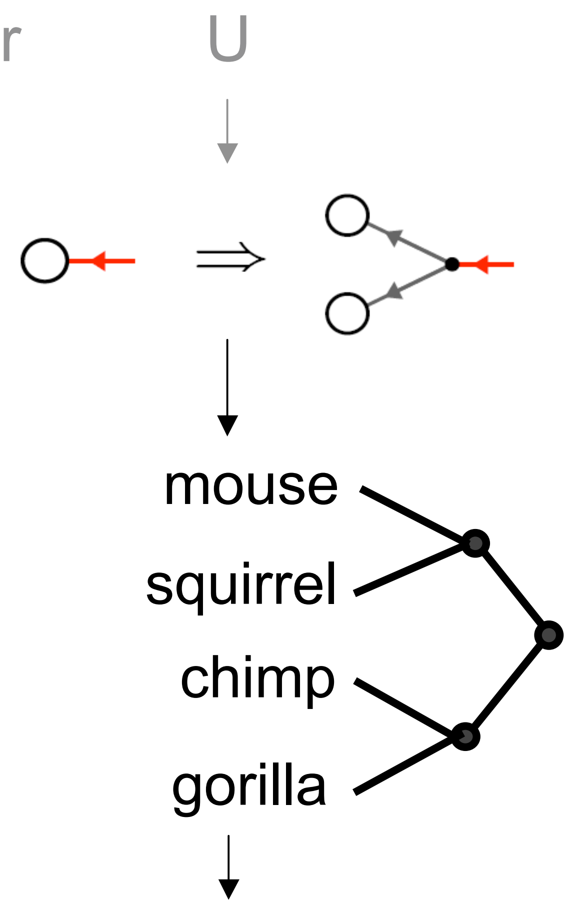


Universal Structure grammar

Form

Structure

Data



	whiskers	hands	tail	feature X
mouse	●	○	●	●
squirrel	●	○	●	?
chimp	○	●	○	?
gorilla	○	●	○	?

Conclusions: Part 2

- Hierarchical Bayesian models provide a unified framework which helps to explain:
  - How abstract knowledge is acquired
  - How abstract knowledge is used for induction

Outline

- Learning structured representations
  - grammars
  - logical theories
- Learning at multiple levels of abstraction

Handbook of Mathematical Psychology, 1963