Inferring structure from data

Tom Griffiths

Department of Psychology

Program in Cognitive Science

University of California, Berkeley

Human learning	**Machine learning**
Categorization	Density estimation
Causal learning	Graphical models
Function learning	Regression
Representations	Nonparametric Bayes
Language	Probabilistic grammars
Experiment design	Inference algorithms
…	…

How much structure exists?

- How many categories of objects?
- How many features does an object have?
- How many words (or rules) are in a language?

Learning the things people learn requires using rich, unbounded hypothesis spaces

Nonparametric Bayesian statistics

- Assume the world contains infinite complexity, of which we only observe a part
- Use stochastic processes to define priors on infinite hypothesis spaces
  - Dirichlet process/Chinese restaurant process (Ferguson, 1973; Pitman, 1996)
  - Beta process/Indian buffet process (Griffiths & Ghahramani, 2006; Thibaux & Jordan, 2007)

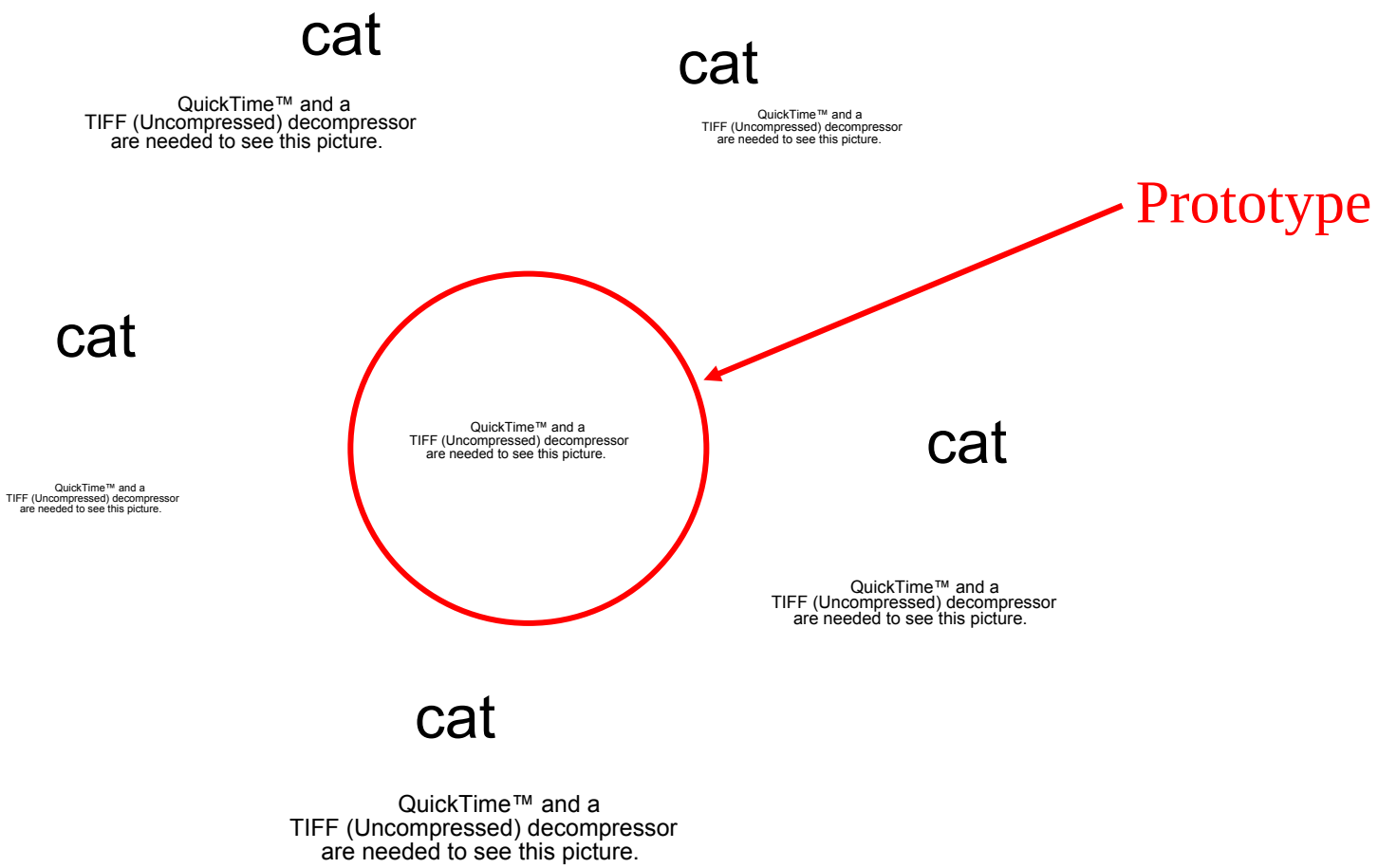
Categorization

cat

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

How do people represent categories?

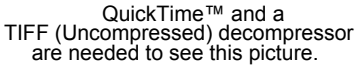
Prototypes



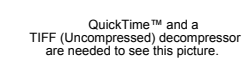
(Posner & Keele, 1968; Reed, 1972)

Exemplars

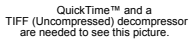
cat



cat

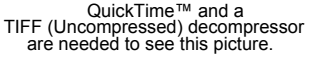


cat

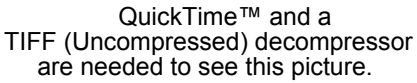


Store every instance (exemplar) in memory

cat

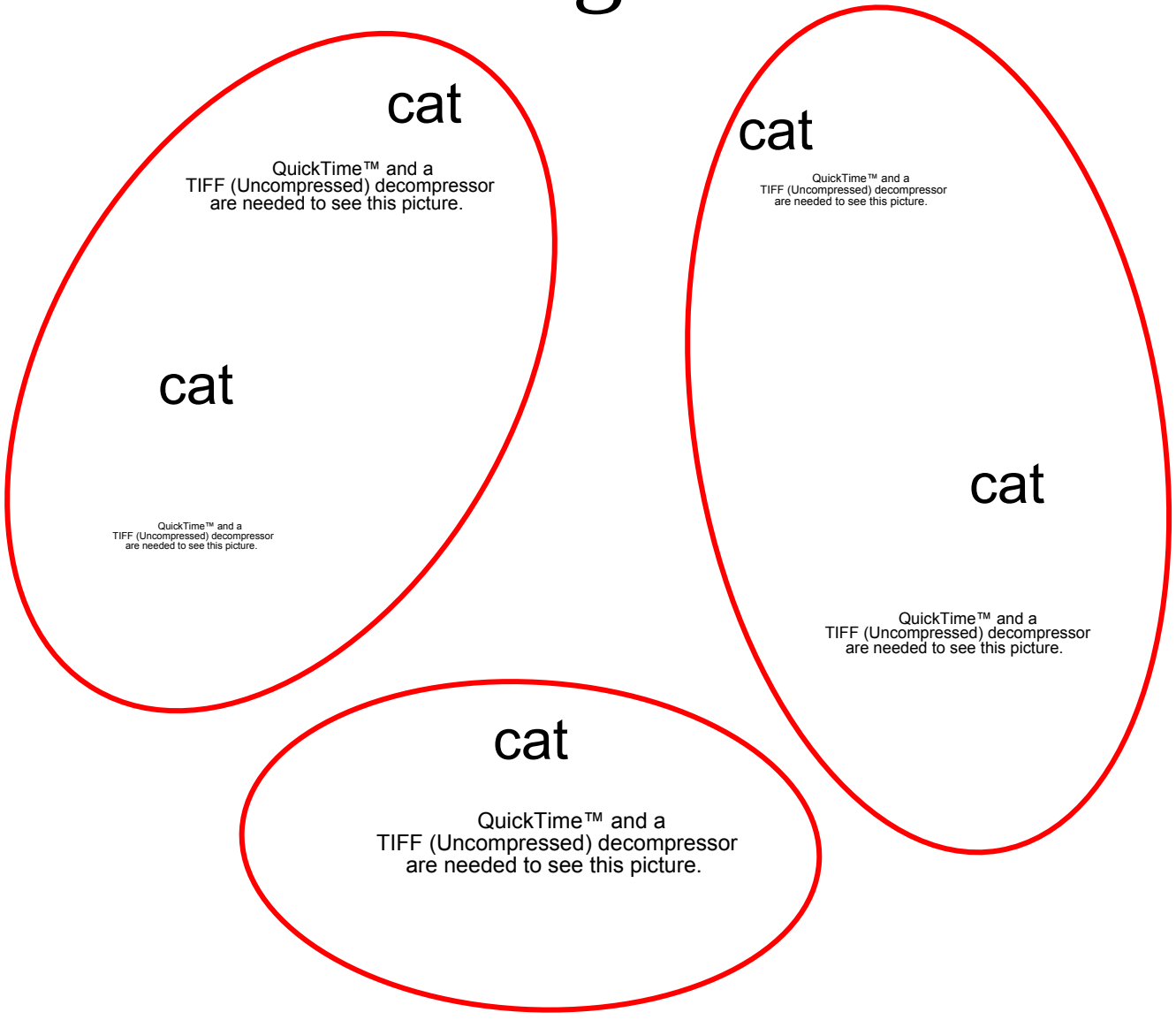


cat



(Medin & Schaffer, 1978; Nosofsky, 1986)

Something in between



(Love et al., 2004; Vanpaemel et al., 2005)

A computational problem

- Categorization is a classic inductive problem
  - data: stimulus x
  - hypotheses: category c
- We can apply Bayes' rule:

$$P(c \mid x) = \frac{p(x \mid c)P(c)}{\sum_{c} p(x \mid c)P(c)}$$

and choose c such that $P(c|x)$ is maximized

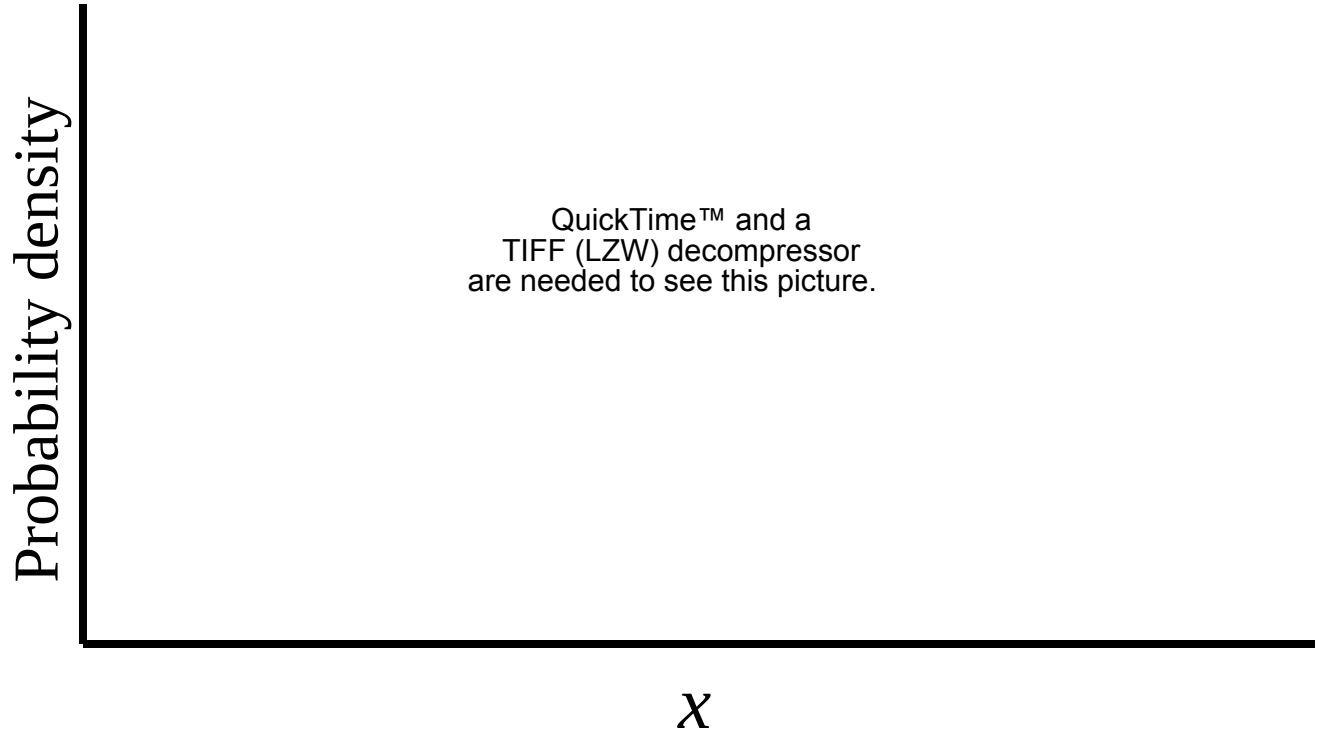
Density estimation

- We need to estimate some probability distributions
  - what is $P(c)$?
  - what is $p(x|c)$?
- Two approaches:
  - parametric
  - nonparametric
- These approaches correspond to prototype and exemplar models respectively

(Ashby & Alfonso-Reese, 1995)

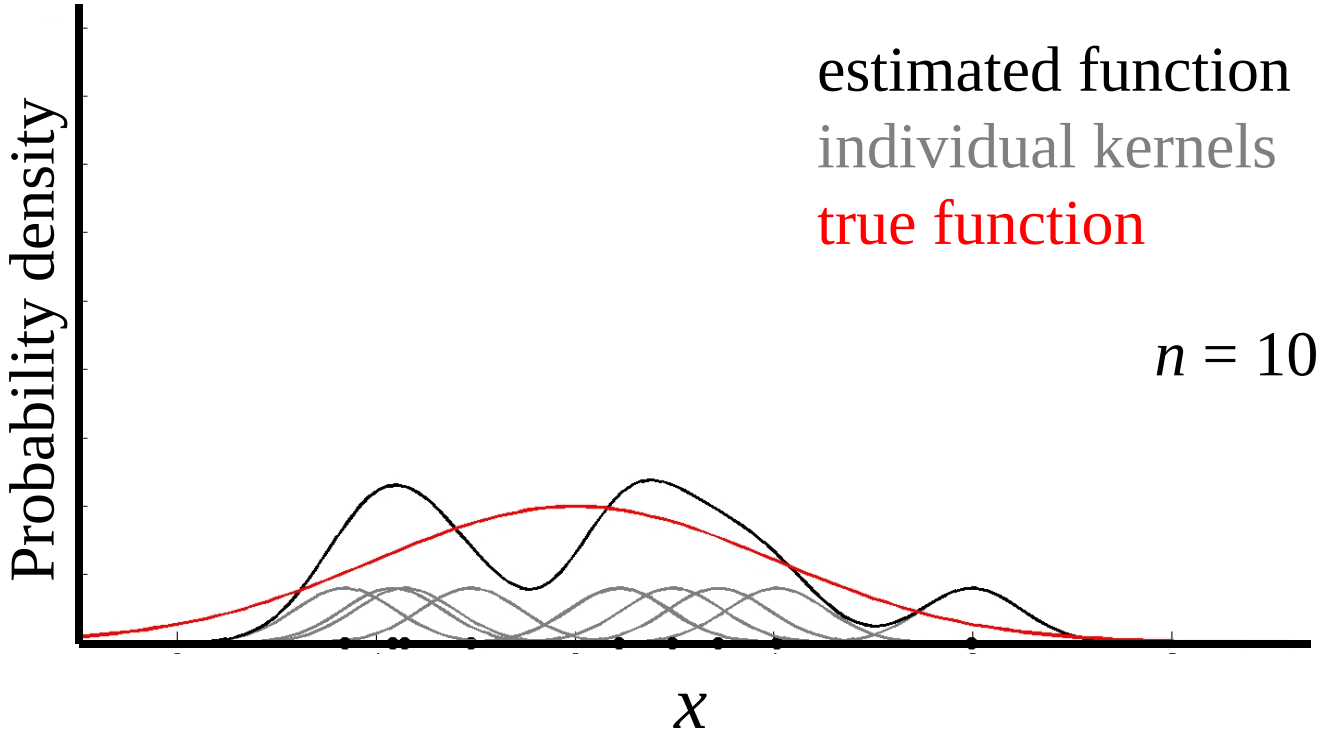
Parametric density estimation

Assume that $p(x|c)$ has a simple form, characterized by parameters θ (indicating the prototype)



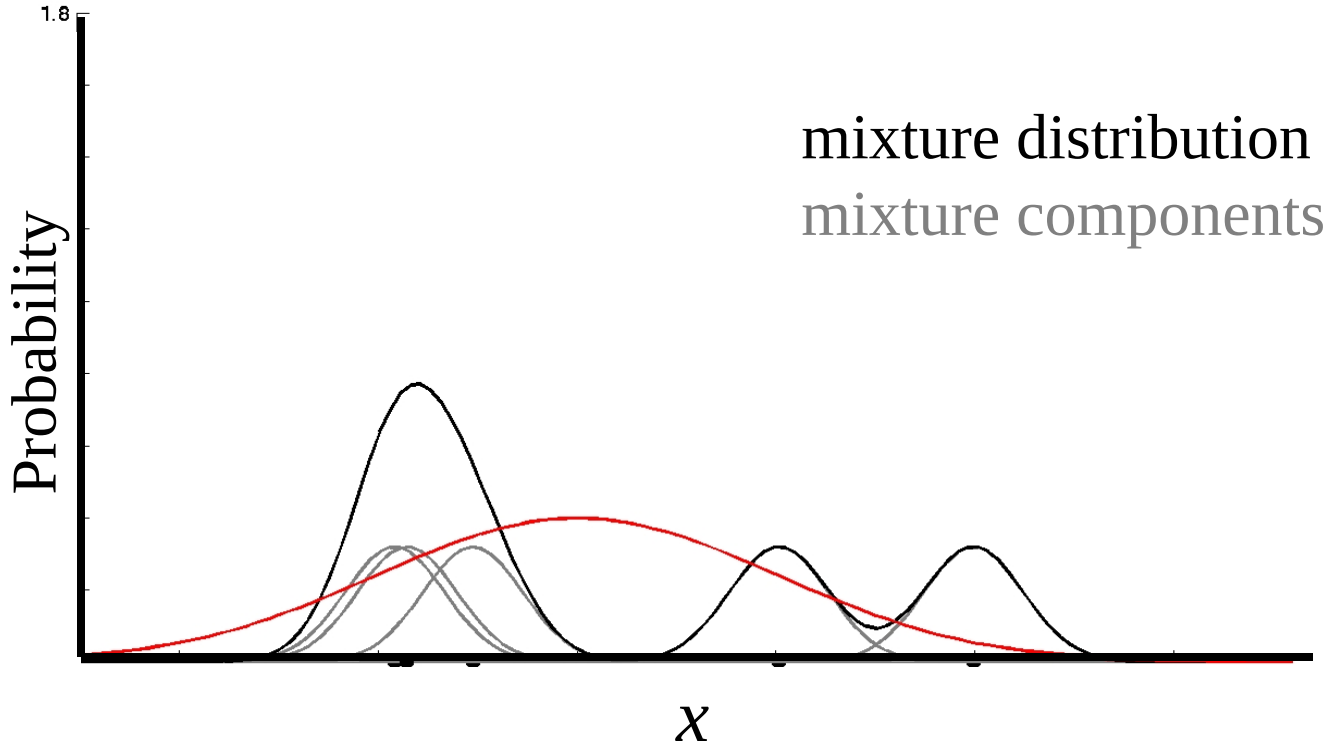
Nonparametric density estimation

Approximate a probability distribution as a sum of many “kernels” (one per data point)



Something in between

Use a “mixture” distribution, with more than one component per data point



(Rosseel, 2002)

Anderson's rational model

(Anderson, 1990, 1991)

- Treat category labels like any other feature
- Define a joint distribution $p(x,c)$ on features using a mixture model, breaking objects into clusters
- Allow the number of clusters to vary…

$$P(\text{cluster } j) \propto \begin{cases} n_j & j \text{ is old} \\ \alpha & j \text{ is new} \end{cases}$$

a Dirichlet process mixture model
(Neal, 1998; Sanborn et al., 2006)

The Chinese restaurant process

- *n* customers walk into a restaurant, choose tables z_i with probability

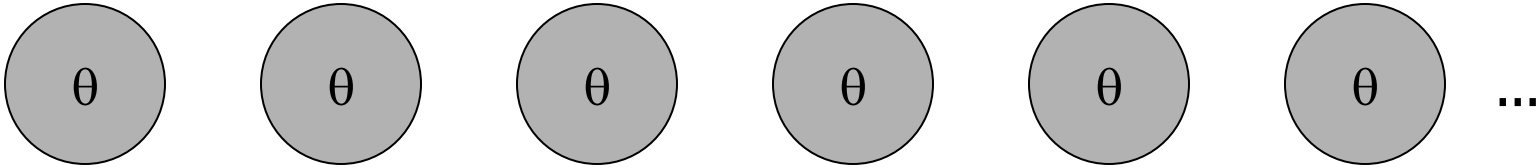
$$P(z_i = j \mid z_1, \ldots, z_{i-1}) = \begin{cases} \dfrac{n_j}{i - 1 + \alpha} & \text{existing table } j \\ \dfrac{\alpha}{i - 1 + \alpha} & \text{next unoccupied table} \end{cases}$$

- Defines an exchangeable distribution over seating arrangements (inc. counts on tables)

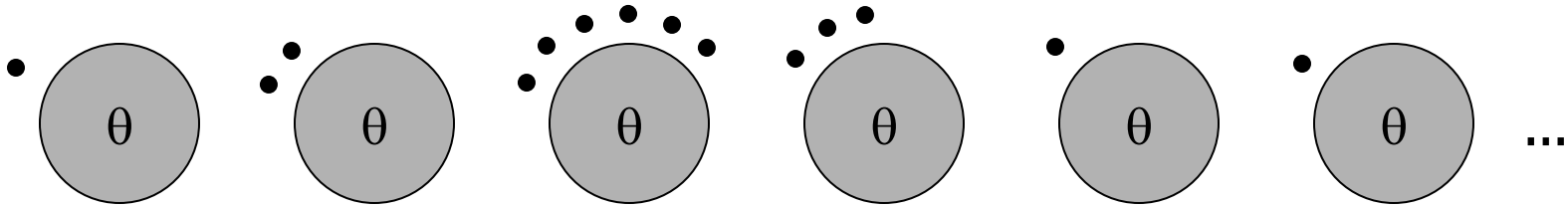
(Aldous, 1985; Pitman, 1996)

Dirichlet process mixture model

1. Sample parameters for each component



2. Assign datapoints to components via CRP



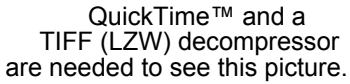
A unifying rational model

- Density estimation is a unifying *framework*
  - a way of viewing models of categorization
- We can go beyond this to define a unifying *model*
  - one model, of which all others are special cases
- Learners can adopt different representations by adaptively selecting between these cases

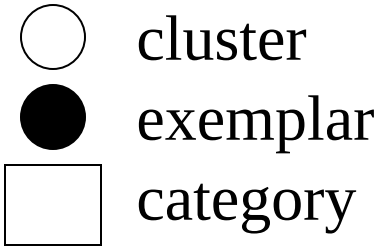
- Basic tool: two interacting levels of clusters
  - results from the hierarchical Dirichlet process

(Teh, Jordan, Beal, & Blei, 2004)

The hierarchical Dirichlet process



A unifying rational model



	$\gamma \in (0,\infty)$	$\gamma \to \infty$ prototype
$\alpha \to 0$	QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.	QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
$\alpha \in (0,\infty)$	QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.	QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
$\alpha \to \infty$	QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.	QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

Anderson

exemplar

$HDP_{+,\infty}$ and Smith & Minda (1998)

- $HDP_{+,\infty}$ will automatically infer a representation using exemplars, prototypes, or something in between (with α being learned from the data)
- Test on Smith & Minda (1998, Experiment 2)

Category A:	Category B:
000000	111111
100000	011111
010000	101111
001000	110111
000010	111011
000001	111110
111101	000100
	exceptions

$HDP_{+,\infty}$ and Smith & Minda (1998)

Probability of A

Log-likelihood

HDP

exemplar

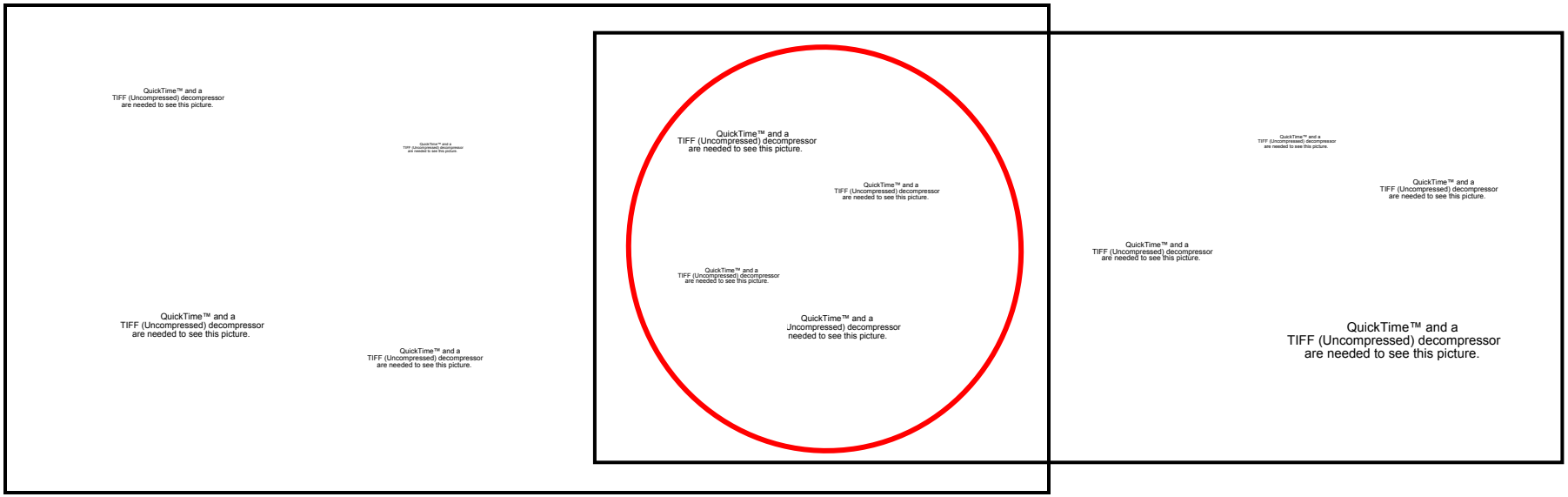
prototype

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

The promise of $HDP_{+,+}$

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

- In $HDP_{+,+}$, clusters are shared between categories
  - a property of hierarchical Bayesian models
- Learning one category has a direct effect on the prior on probability densities for the next category



Other uses of Dirichlet processes

- Nonparametric Block Model from causality lecture
  - extends DPMM to relations
- Models of language, where number of words, syntactic classes, or grammar rules is unknown
- Any multinomial can be replaced by a CRP…
- Extensions:
  - hierarchical, nested, dependent, two-parameter, distance-dependent, …

Learning the features of objects

- Most models of human cognition assume objects are represented in terms of abstract features
- What are the features of this object?

- What determines what features we identify?

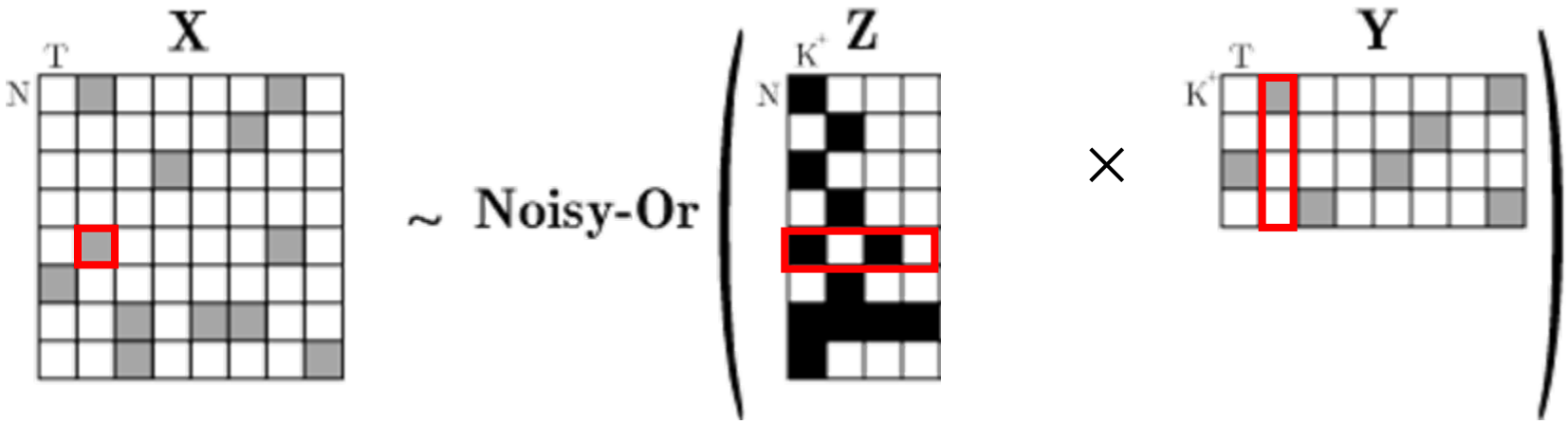
(Austerweil & Griffiths, 2009)

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

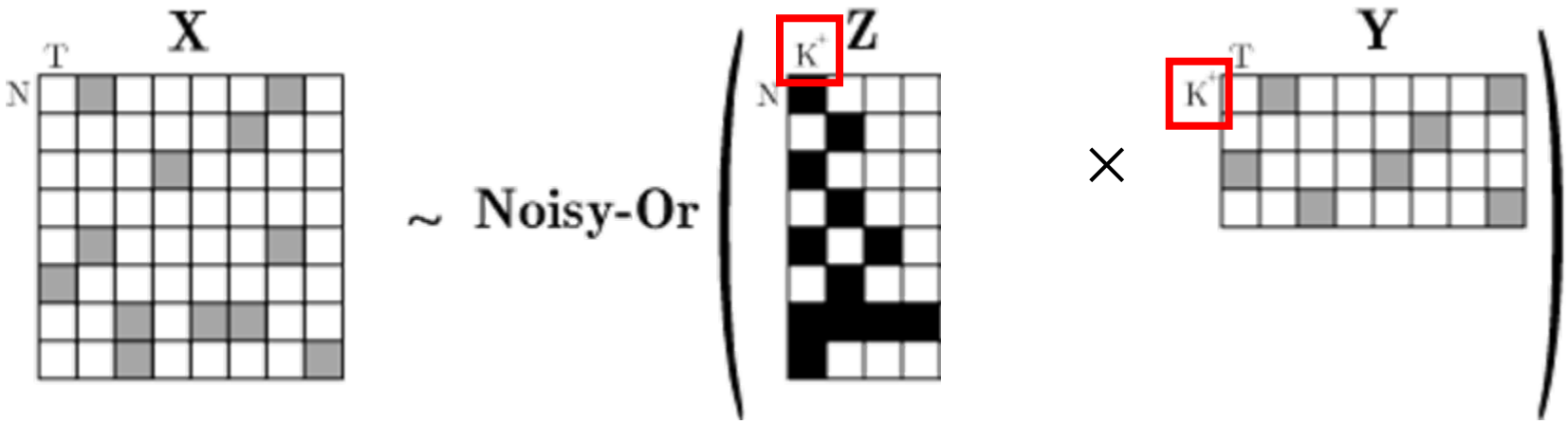
Binary matrix factorization

$$P(x_{i,t} = 1 | \mathbf{Z}, \mathbf{Y}) = 1 - (1 - \lambda)^{<\mathbf{z}_{i,:}, \mathbf{y}_{:,t}>} (1 - \epsilon)$$



Binary matrix factorization

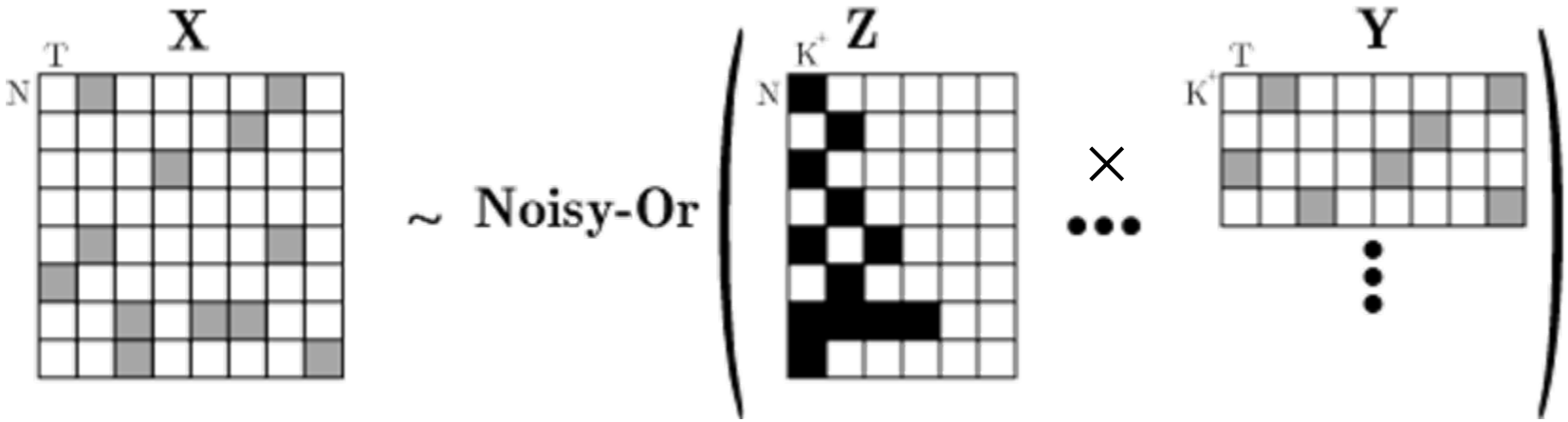
$$P(x_{i,t} = 1 | \mathbf{Z}, \mathbf{Y}) = 1 - (1 - \lambda)^{<\mathbf{z}_{i,:}, \mathbf{y}_{:,t}>}(1 - \epsilon)$$



How should we infer the number of features?

The nonparametric approach

Assume that the total number of features is unbounded, but only a finite number will be expressed in any finite dataset



Use the Indian buffet process as a prior on **Z**

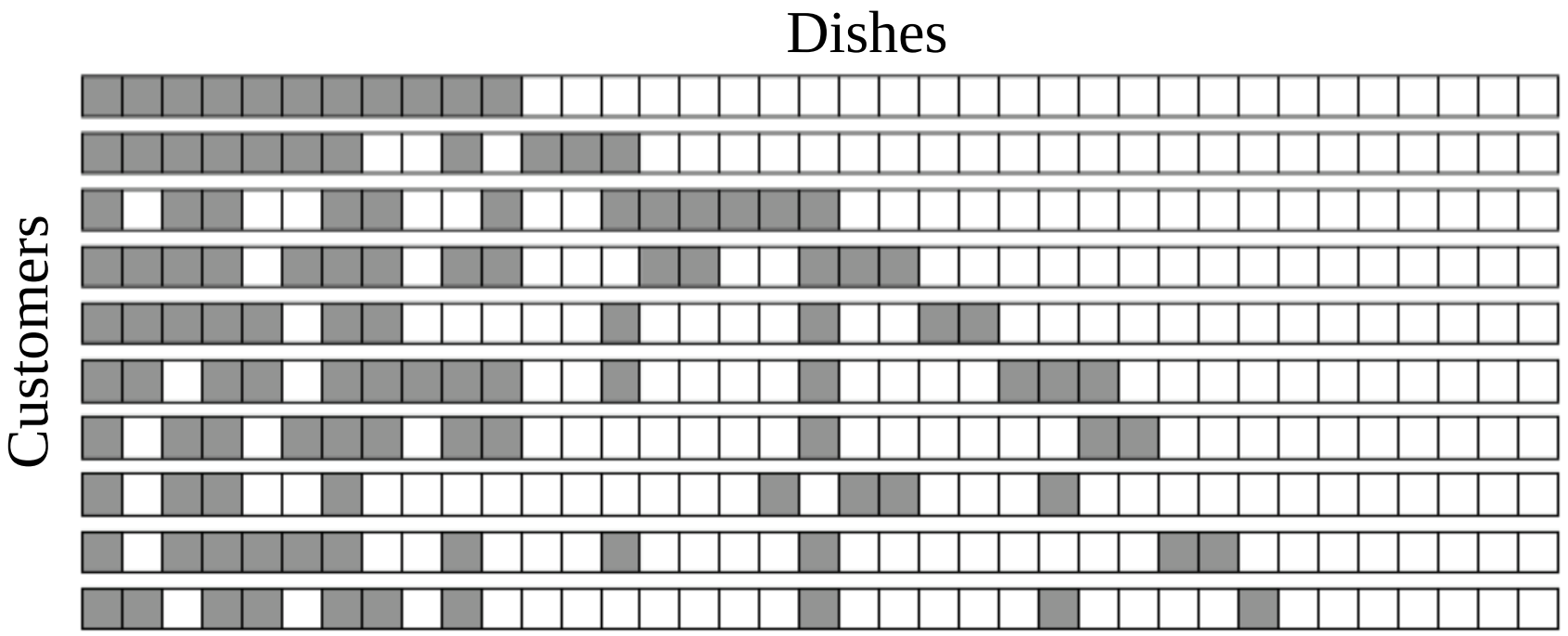
(Griffiths & Ghahramani, 2006)

The Indian buffet process

- First customer walks into Indian restaurant, and tastes Poisson (α) dishes from the buffet
- The ith customer tastes previously-tasted dish k with probability m_k/i, where m_k is the number of previous tasters, and Poisson (α/i) new dishes
- Customers are exchangeable, as in the CRP

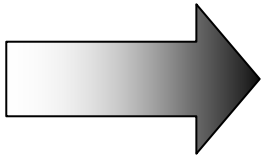
(Griffiths & Ghahramani, 2006)

The Indian buffet process



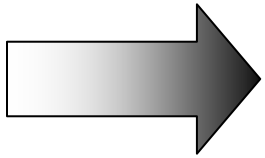
(Griffiths & Ghahramani, 2006)

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.



QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

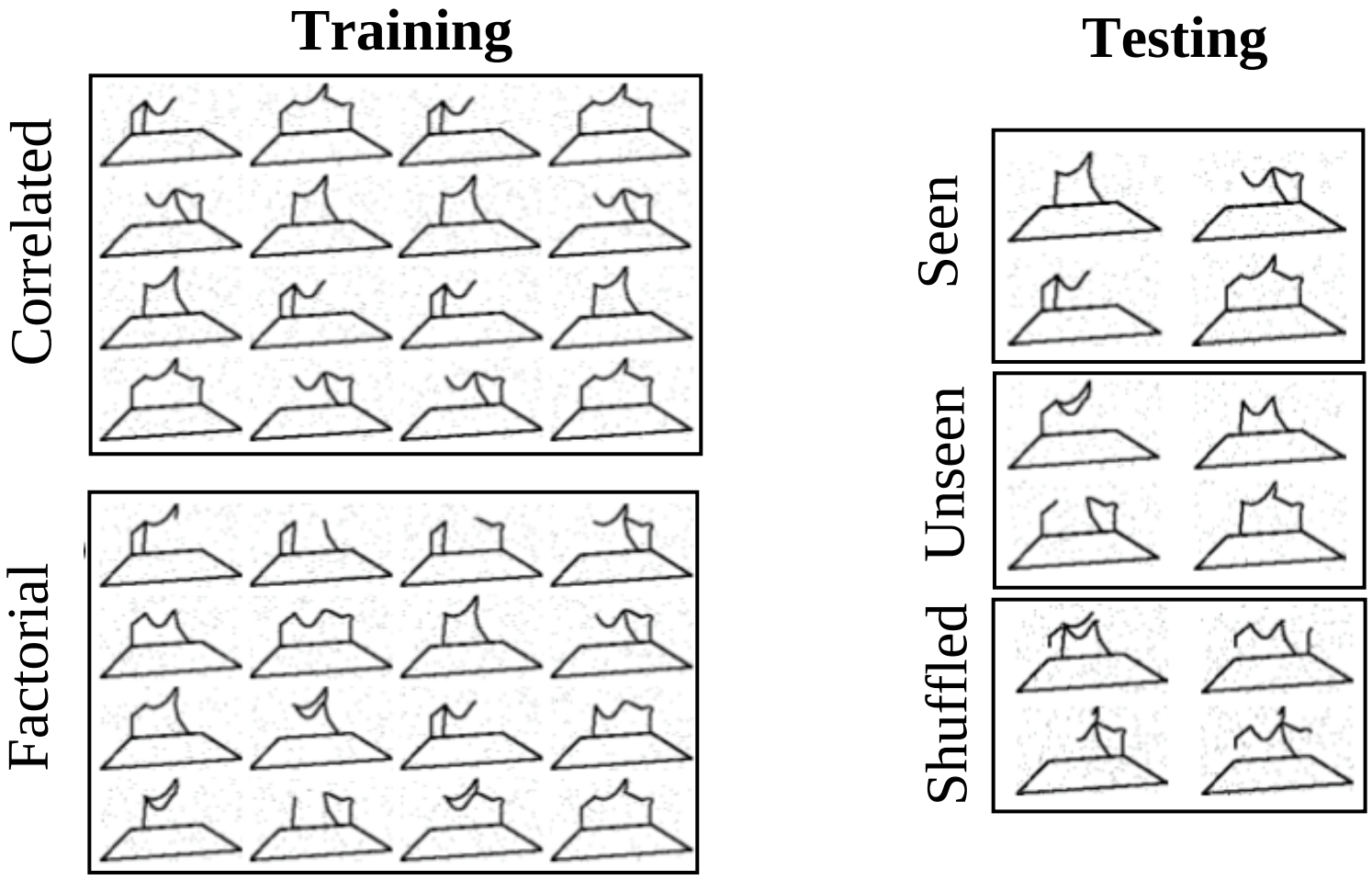
QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.



QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

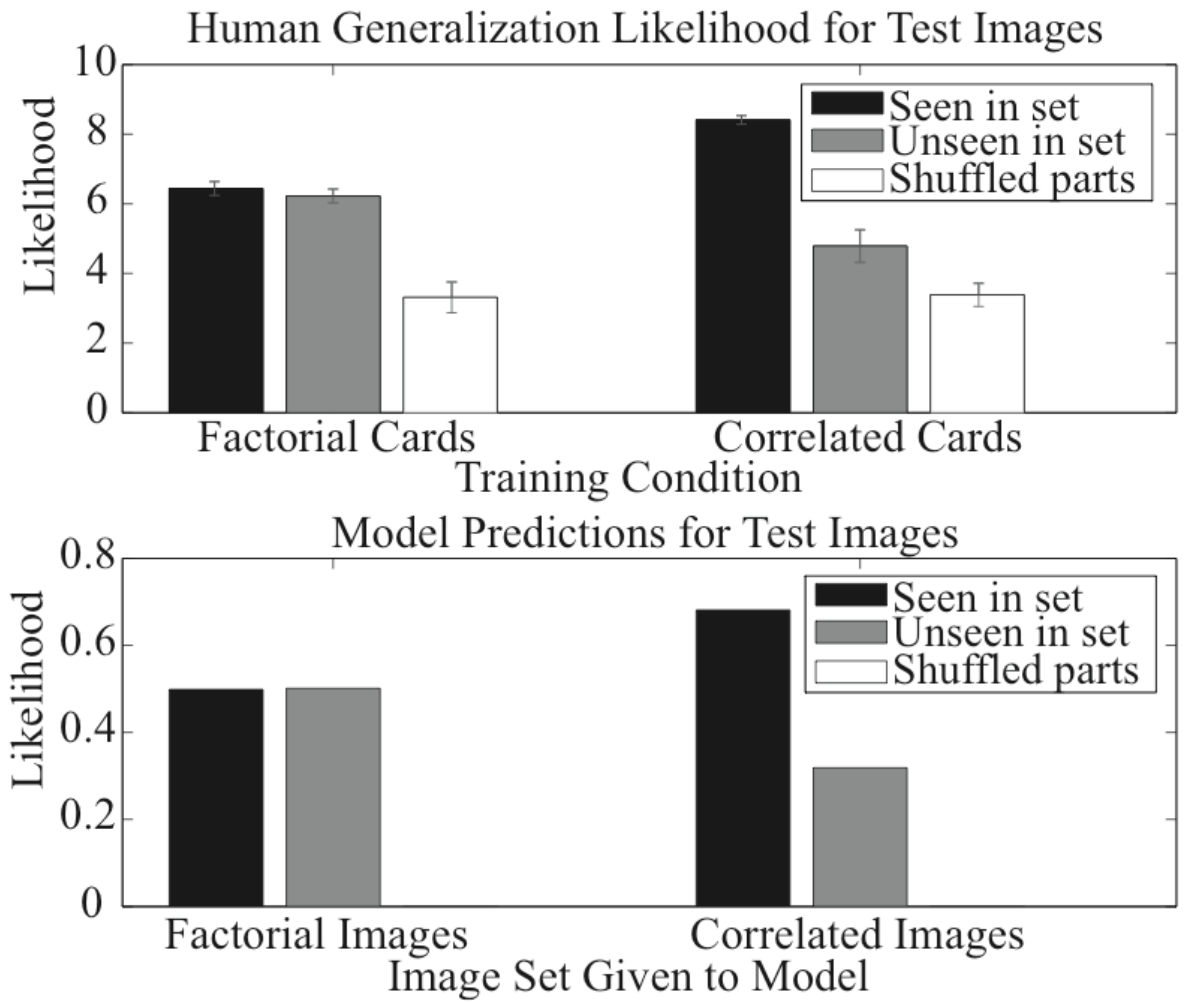
(Austerweil & Griffiths, 2009)

An experiment…



(Austerweil & Griffiths, 2009)

Results



(Austerweil & Griffiths, 2009)

Other uses of the IBP

- Prior on sparse binary matrices, used for number of dimensions in *any* sparse latent feature model
  - PCA, ICA, collaborative filtering, …
- Prior on adjacency matrix for bipartite graph with one class of nodes having unknown size
  - e.g., inferring hidden causes
- Interesting link to Beta processes (like CRP to DP)
- Extensions:
  - two parameters, three parameters, phylogenetic, …

Nonparametric Bayes and the mind

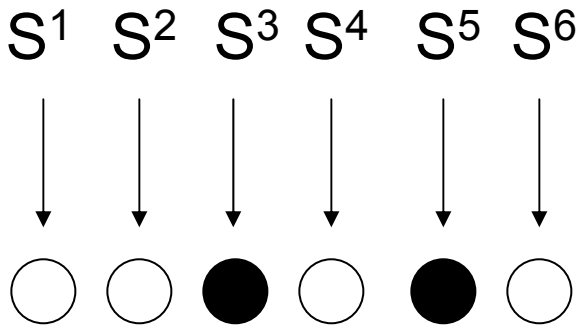
- Nonparametric Bayesian models provide a way to answer questions of how much structure to infer
- For questions like “how many clusters?” we can use the Chinese restaurant/Dirichlet process
- For questions like “how many features?” we can use the Indian buffet/Beta process

- Lots of room to develop new models…

Learning language

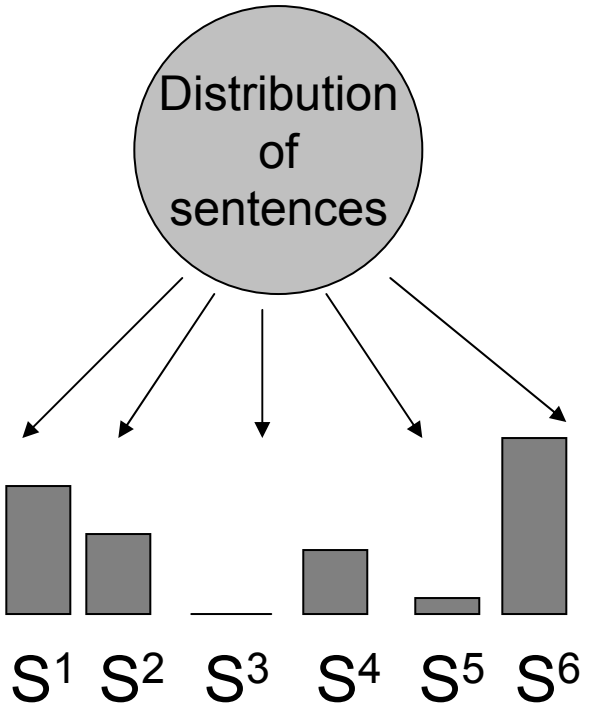
Discriminative vs. Generative

P(Grammatical|S)



Labels of grammatical or ungrammatical

P(S|Grammatical)



An artificial language:

S1) *Verb Subject Object*

S2) *Subject Verb Object*

S3) *Subject Object Verb*

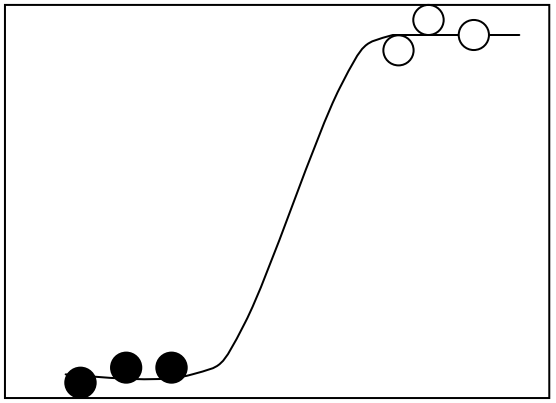
	S1	S2	S3
V1	+ (9)	+ (9)	- (6)
V2	- (3)	+(18)	- (3)
V3	+ (18)	- (3)	- (3)
V4*	+ (18)	- (0)	- (6)

Discriminative
Logistic regression

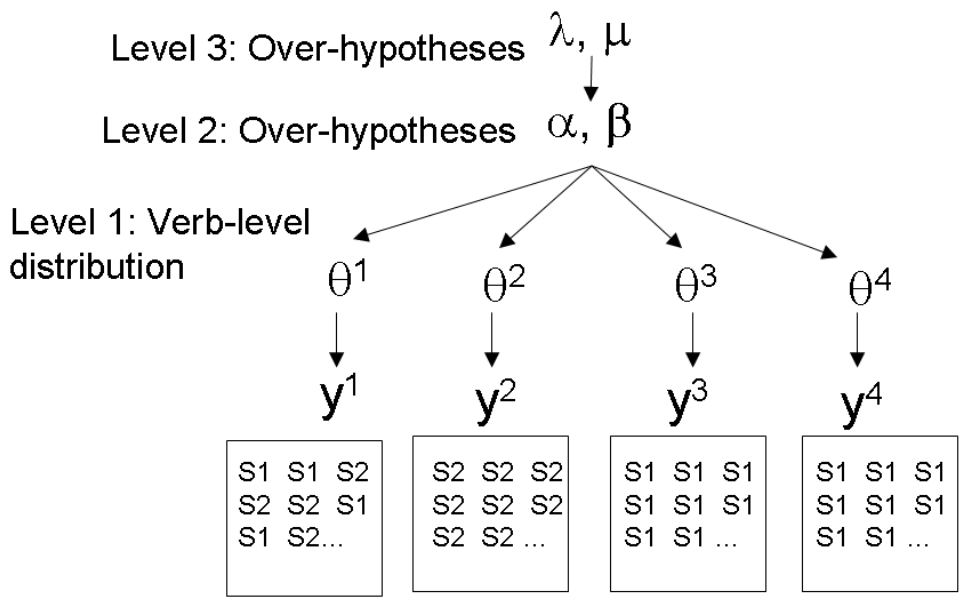
vs.

Generative
Hierarchical
Bayesian model

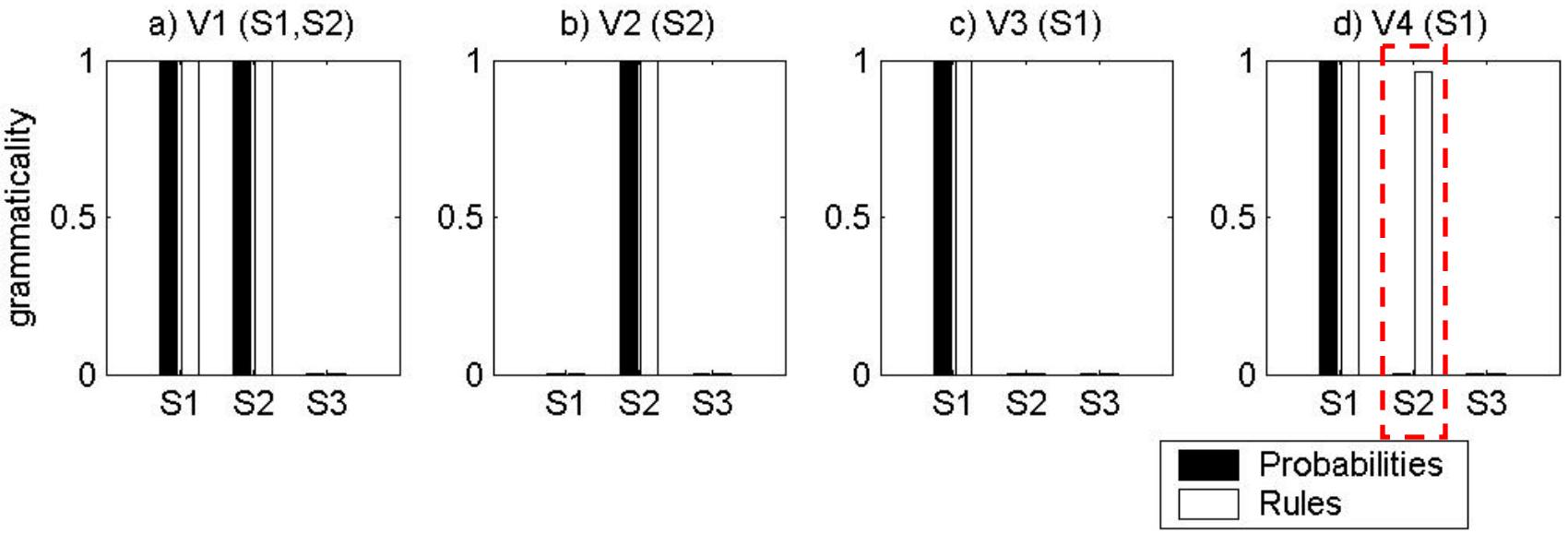
Grammatical



Ungrammatical

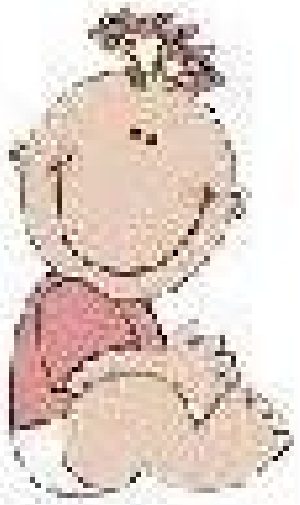


Model Predictions

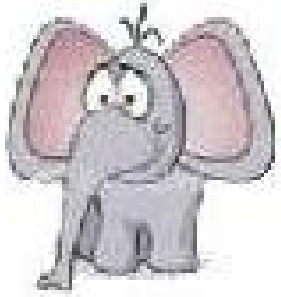


Condition 1: Generative learning

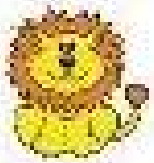
Always grammatically correct adult



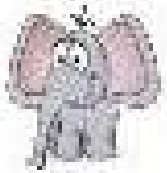
Always grammatically incorrect child



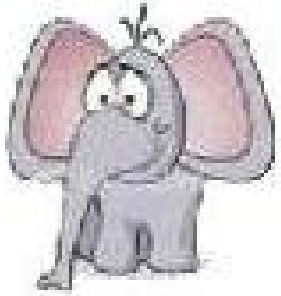
blergen



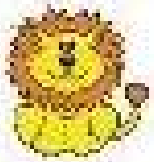
nagid



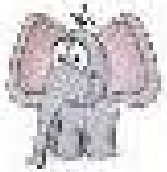
tombat



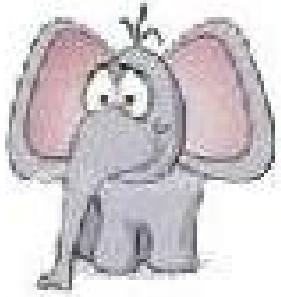
blergen



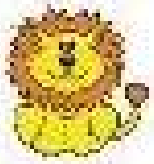
nagid



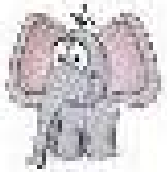
tombat



blergen

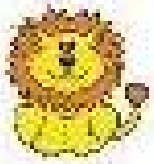


nagid

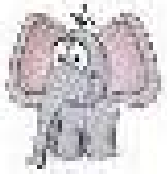


tombat

blergen

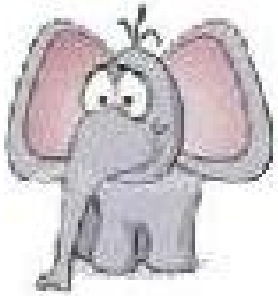


nagid

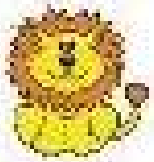


tombat

blergen norg nagid

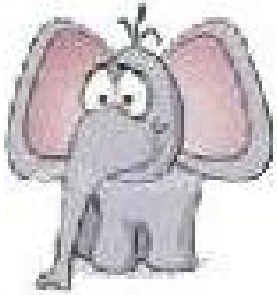


blergen

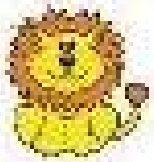


nagid

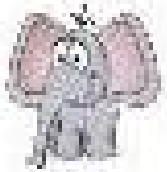
tombat



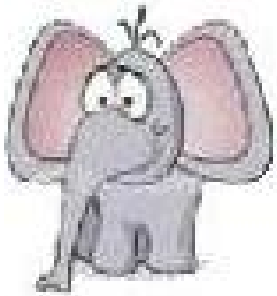
blergen



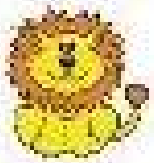
nagid



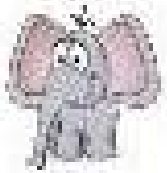
tombat



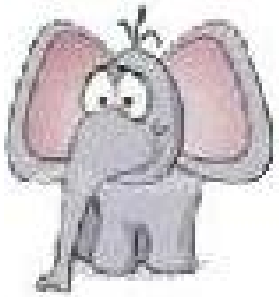
blergen



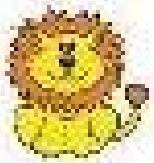
nagid



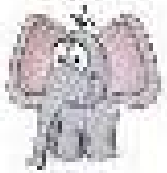
tombat



blergen

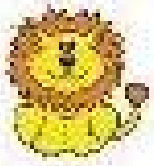


nagid

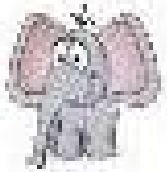


tombat

blergen

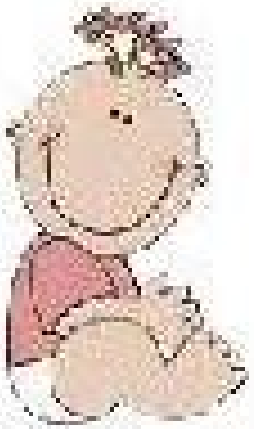


nagid



tombat

nagid blergen semz



Condition 2: Discriminative learning

scene 1/84

tombat blergen flern

Was that sentence grammatical?

Grammatical Ungrammatical

blergen

nagid

tombat

Condition 2: Discriminative learning

Condition 2: Discriminative learning

scene 4/84

blergen semz tombat

Ungrammatical

Was that sentence grammatical?

Grammatical Ungrammatical

Sorry you were wrong.

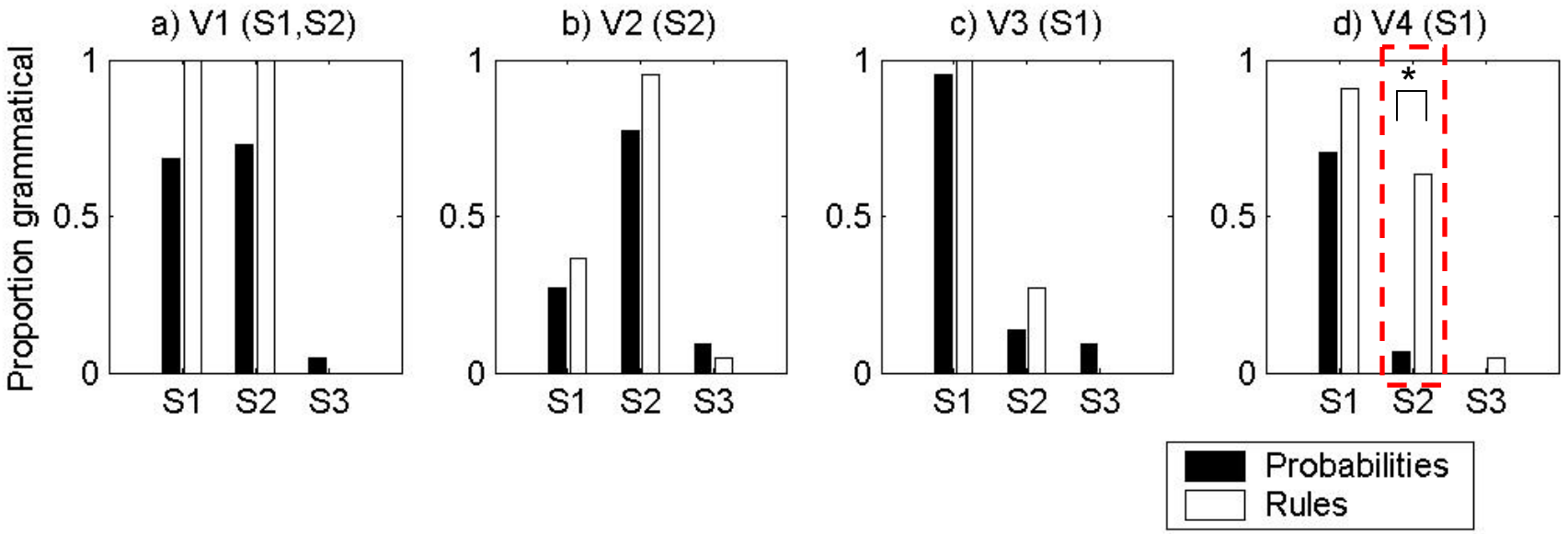
blergen

nagid

tombat

Click here for next scene

Human language learning



* $\chi 2(1) = 7.28$, $p = 0.007$