Bayesian Generative Modeling

Jason Eisner

Summer School on Machine Learning
Lisbon, Portugal – July 2011

Bayesian Generative Modeling

what's a model?

Jason Eisner

Summer School on Machine Learning
Lisbon, Portugal – July 2011

Bayesian Generative Modeling

what's a generative model?

Jason Eisner

Summer School on Machine Learning
Lisbon, Portugal – July 2011

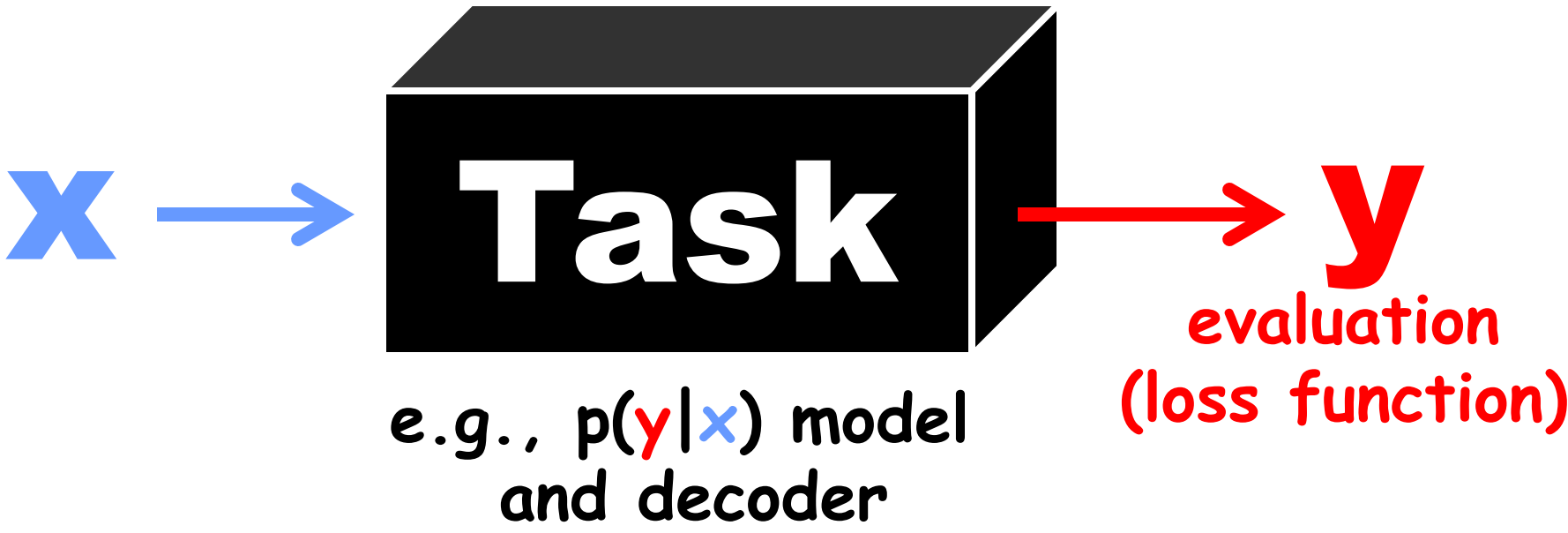
Bayesian Generative Modeling

what's Bayesian?

Jason Eisner

Summer School on Machine Learning
Lisbon, Portugal – July 2011

Task-centric view of the world



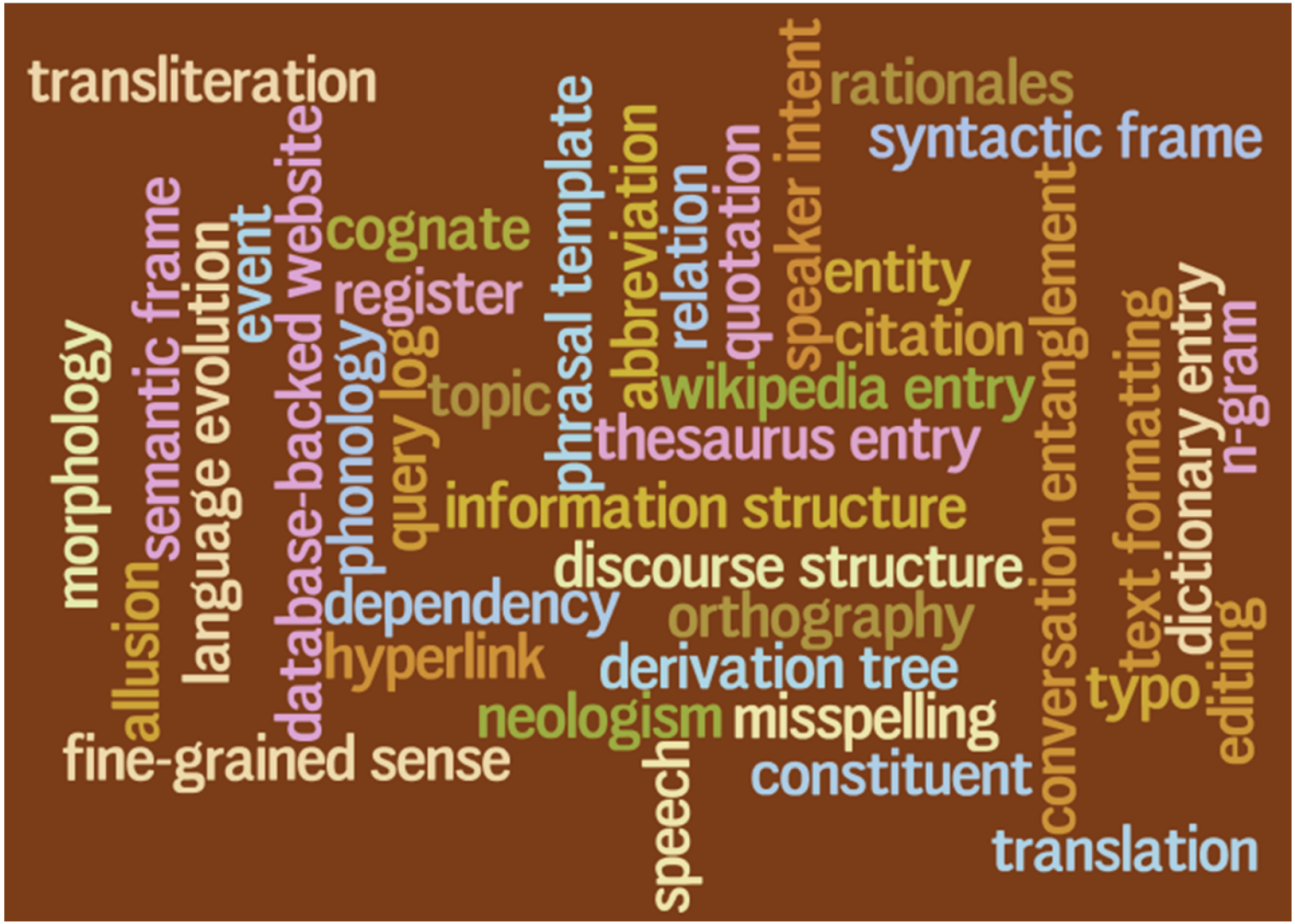
Task-centric view of the world

- ☺ Great way to track progress & compare systems
  - ☹ But may fracture us into subcommunities
    (our systems are incomparable & my semantics != your semantics)
- ☺ Room for all of AI when solving any NLP task

  Spelling correction could get some benefit from deep semantics, unsupervised grammar induction, active learning, discourse, etc.
  - ☹ But in practice, focus on raising a single performance number
  - ☹ Within strict, fixed assumptions about the type of available data

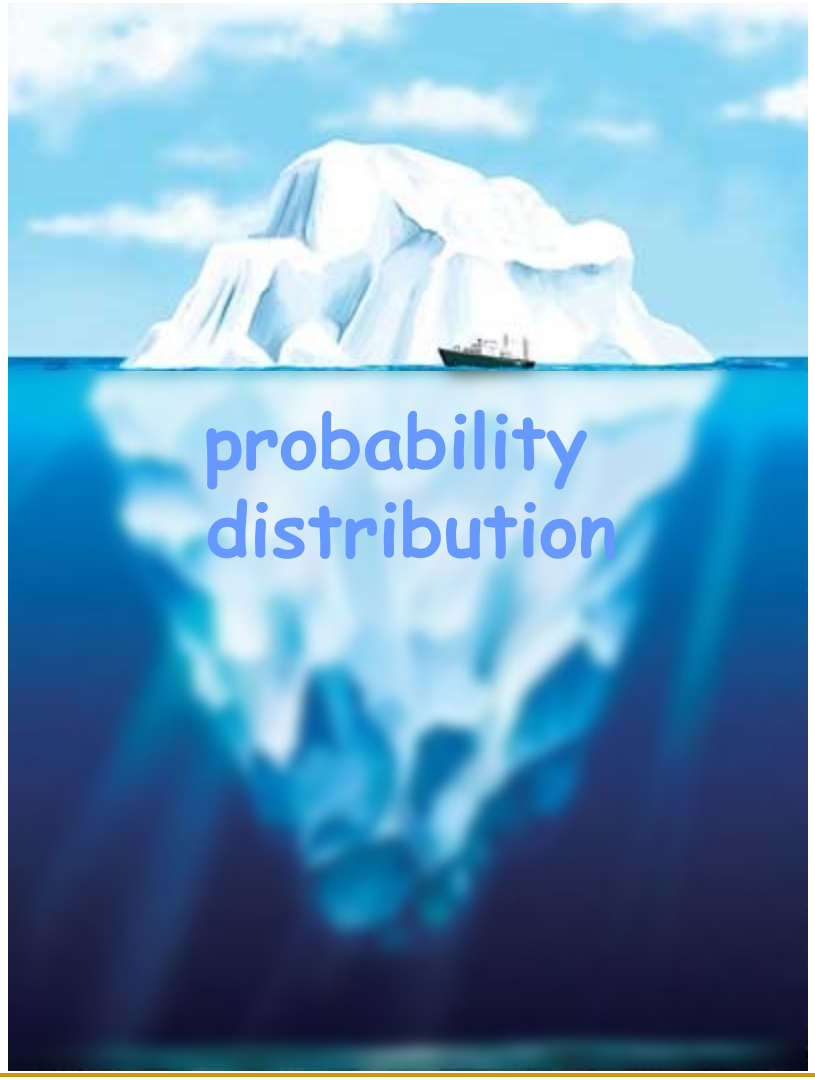
Do we want to build models & algs that are good for just one task?

Variable-centric view of the world



When we deeply understand language, what representations (type and token) does that understanding comprise?

Bayesian View of the World



observed data

hidden data

Different tasks merely change which variables are observed and which ones you care about inferring

	comprehension	production	learning
sentence	✓	?	✓
syntax tree	(?)	latent	(✓)
semantics	(?)	(✓)	(✓)
facts about speaker/world	(?)	(✓)	(✓)
facts about the language	✓	✓	?

Different tasks merely change which variables are observed and which ones you care about inferring

	comprehension	production	learning
surface form of word	✓	?	✓
surface ⇔ underlying alignment	(?)	latent	(✓)
underlying form of word	(?)	latent	(✓)
abstract morphemes in word	(?)	✓	(✓)
underlying form of morphemes (lexicon)	✓	✓	(?)
constraint ranking (grammar)	✓	✓	(?)

Different tasks merely change which variables are observed and which ones you care about inferring

	MT decoding	MT training	cross-lingual projection
Chinese sentence	✓	✓	✓
Chinese parse	latent	latent	?
English parse	latent	latent	✓
English sentence	?	✓	✓
translation & language models	✓	?	?

All you need is "p"

- Science = a descriptive theory of the world
- Write down a formula for

$$p(\text{everything})$$

- everything = observed ∪ needed ∪ latent

- Given observed, what might needed be?
- Most probable settings of needed are those that give comparatively large values of

$$\sum_{\text{latent}} p(\text{observed}, \text{needed}, \text{latent})$$

- Formally, we want p(needed | observed)
  = p(observed, needed) / p(observed)

  Since observed is constant, the conditional probability of needed varies with p(observed, needed), which is given above
- (What do we do then?)

All you need is "p"

- Science = a descriptive theory of the world
- Write down a formula for

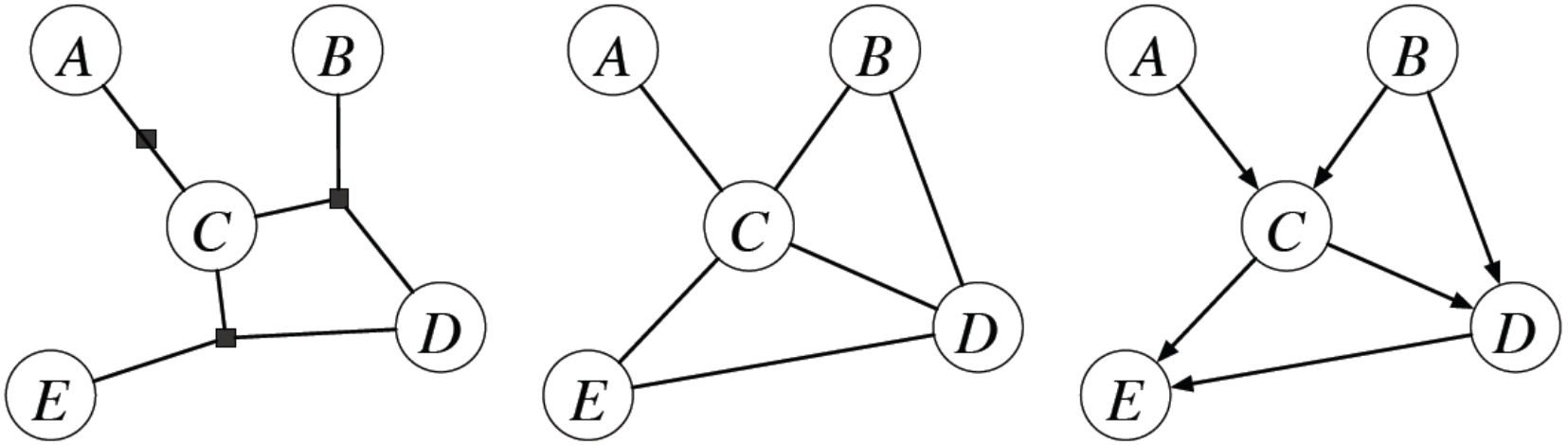
$$p(\text{everything})$$

- everything = observed $\cup$ needed $\cup$ latent

- p can be any non-negative function you care to design
  - (as long as it sums to 1)
  - (or another finite positive number: just rescale)

- But it's often convenient to use a **graphical model**
  - Flexible modeling technique
  - Well understood
  - We know how to (approximately) compute with them

Graphical model notation

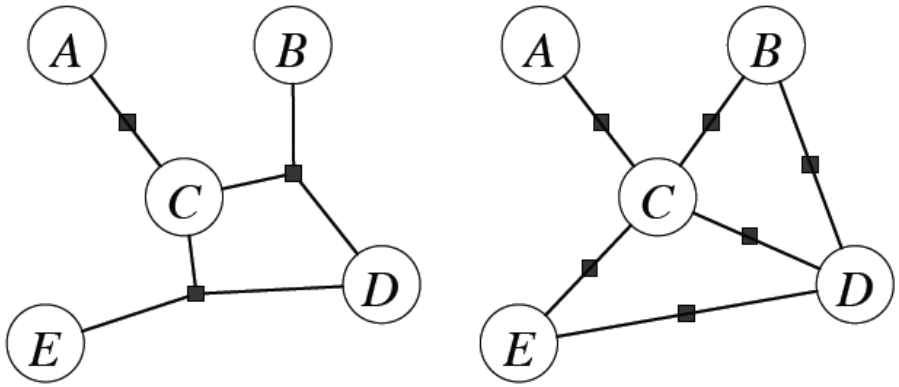


factor graph | undirected graph | directed graph

- Nodes correspond to random variables
- Edges represent statistical dependencies between the variables

slide thanks to Zoubin Ghahramani

Factor graphs



(a) (b)

Two types of nodes:

- The circles in a factor graph represent random variables (e.g. A).
- The filled dots represent factors in the joint distribution (e.g. $g_1(\cdot)$).

(a) $p(A,B,C,D,E) = \frac{1}{Z} g_1(A,C) g_2(B,C,D) g_3(C,D,E)$

(b) $p(A,B,C,D,E) = \frac{1}{Z} g_1(A,C) g_2(B,C) g_3(C,D) g_4(B,D) g_5(C,E) g_6(D,E)$

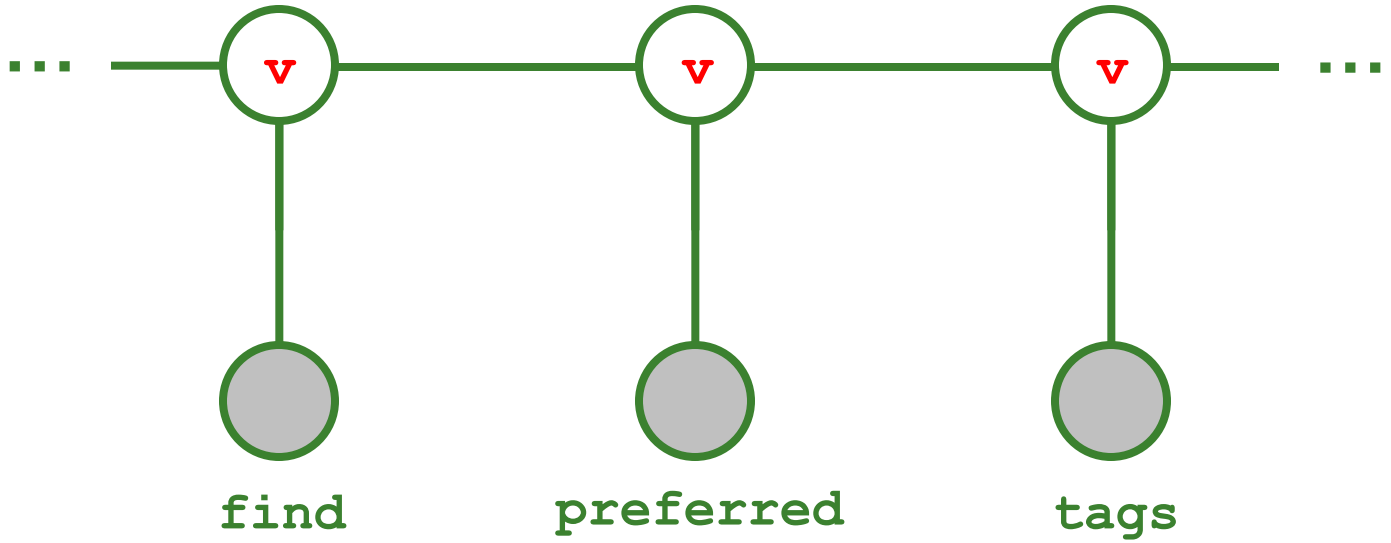
The g_i are non-negative functions of their arguments, and Z is a normalization constant. E.g. in (a), if all variables are discrete and take values in $\mathcal{A} \times \mathcal{B} \times \mathcal{C} \times \mathcal{D} \times \mathcal{E}$:

$$Z = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}} \sum_{e \in \mathcal{E}} g_1(A=a, C=c) g_2(B=b, C=c, D=d) g_3(C=c, D=d, E=e)$$

Rather basic NLP example

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

Possible tagging (i.e., assignment to remaining variables)

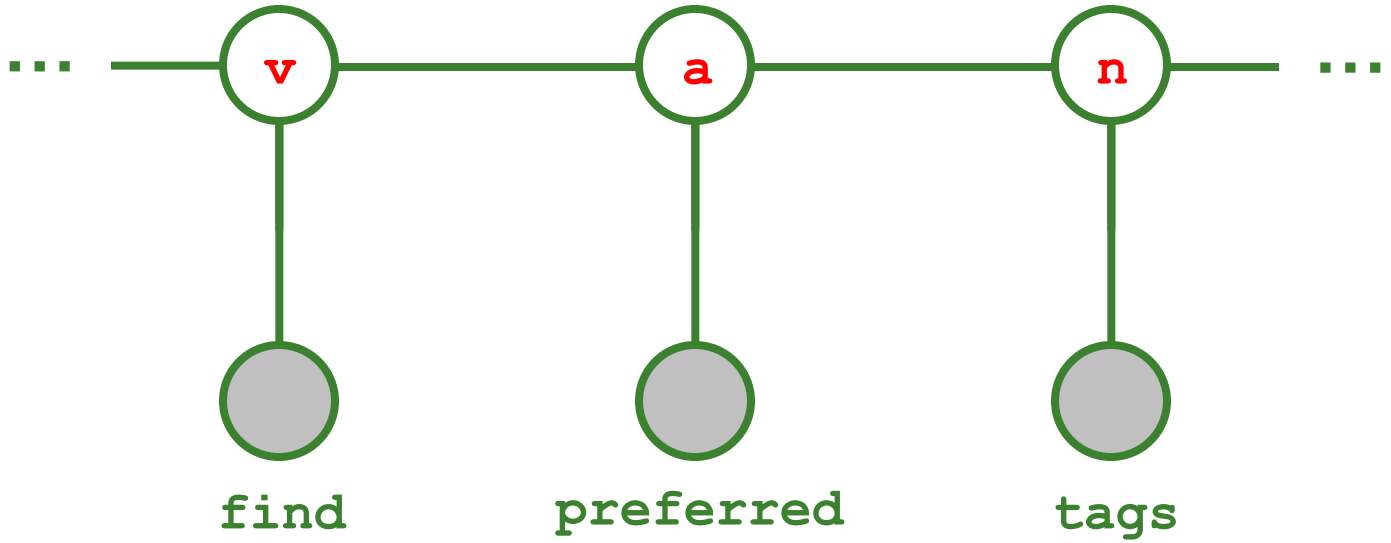


Observed input sentence (shaded)

Rather basic NLP example

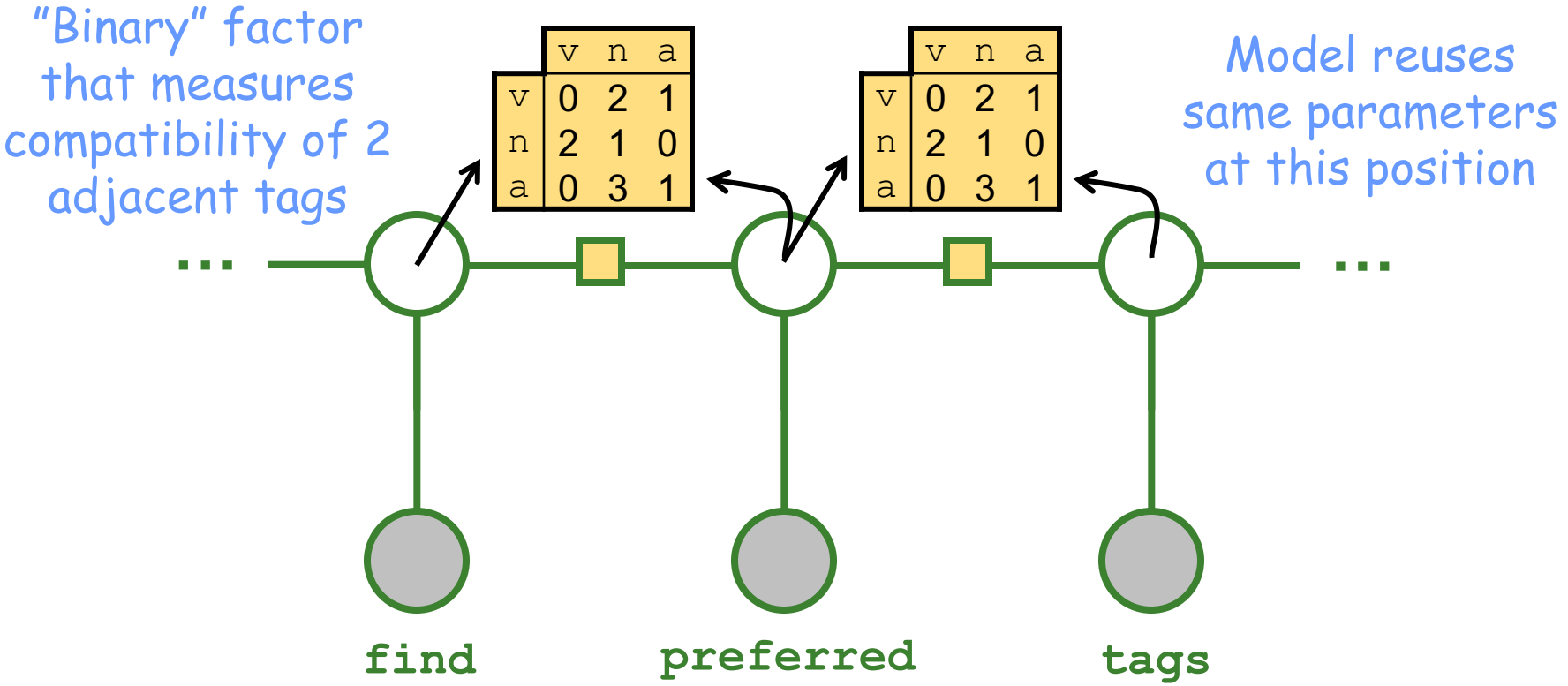
- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

Possible tagging (i.e., assignment to remaining variables)
Another possible tagging

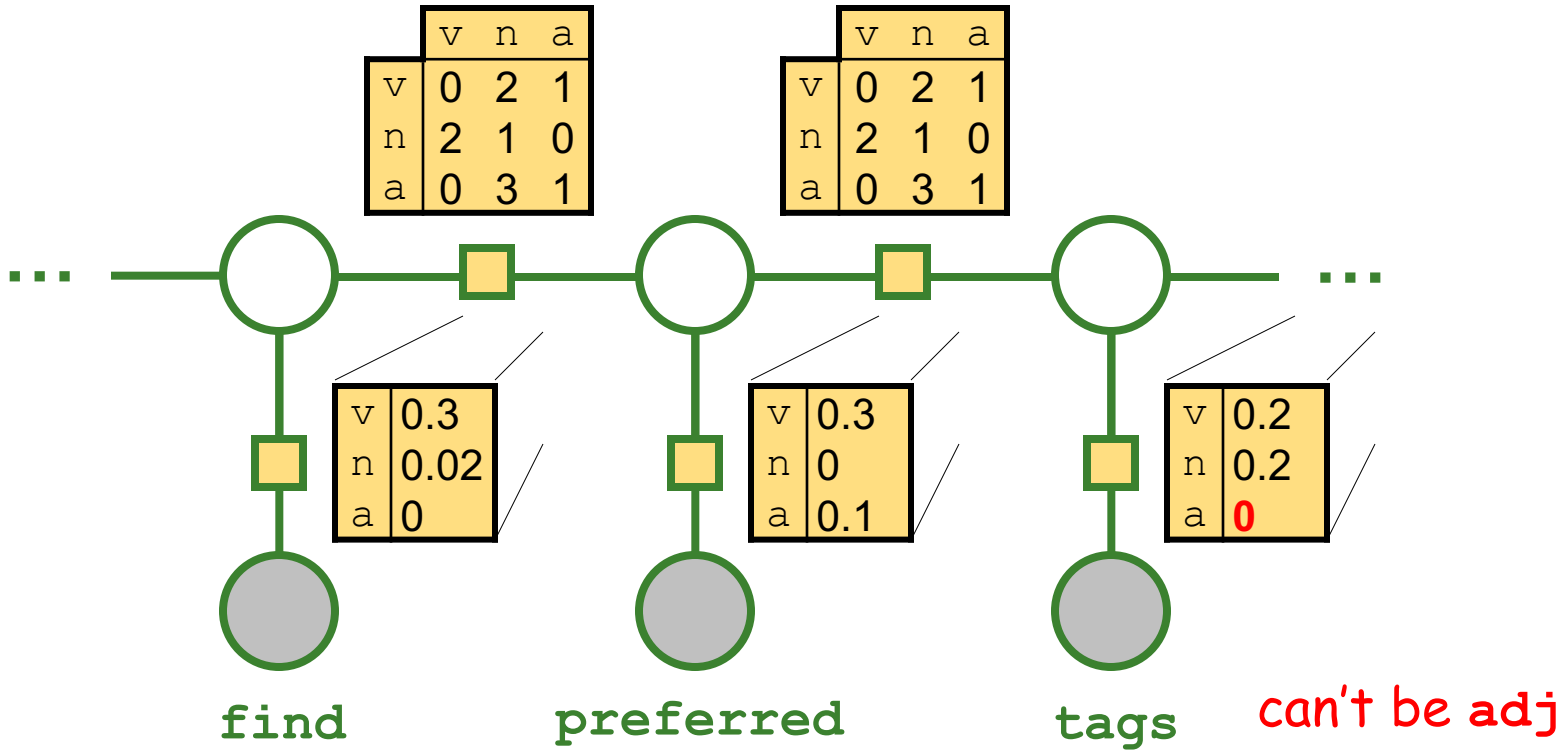


Observed input sentence (shaded)

Conditional Random Field (CRF)

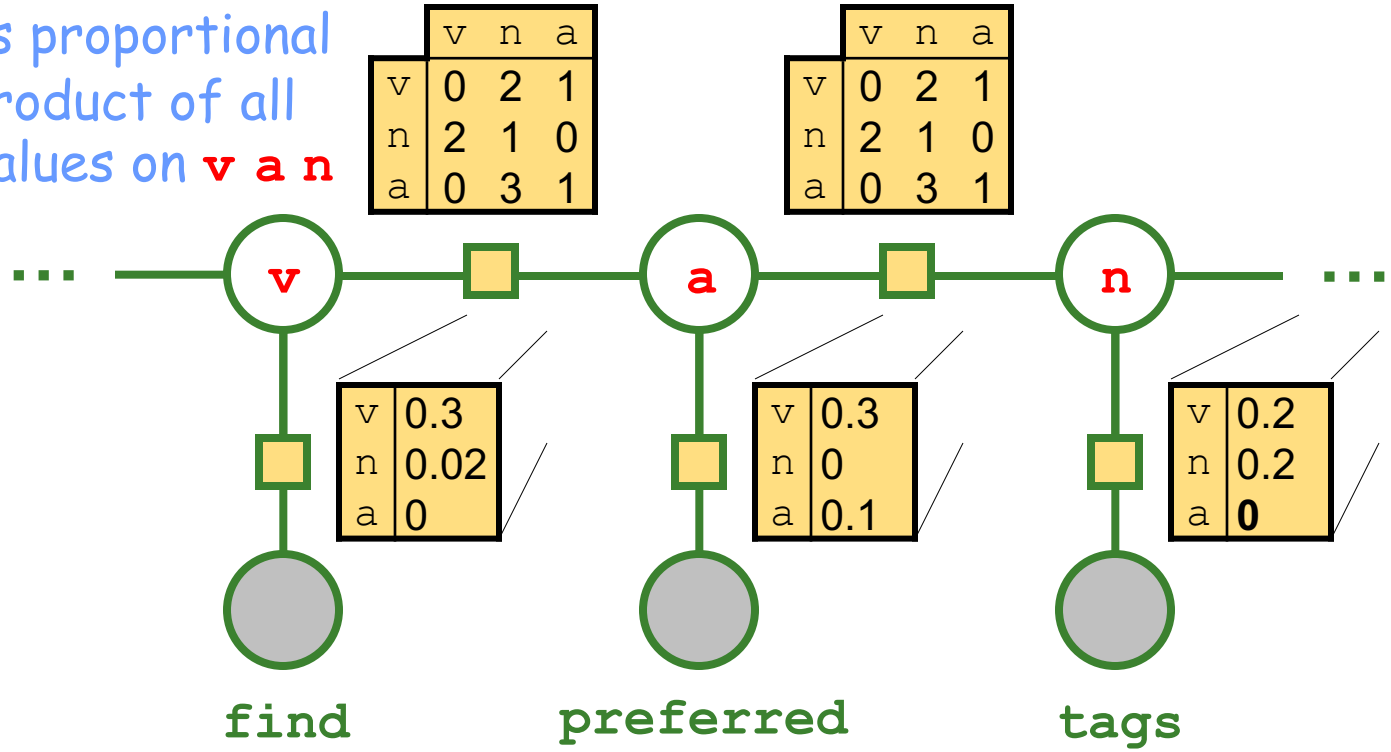


Conditional Random Field (CRF)

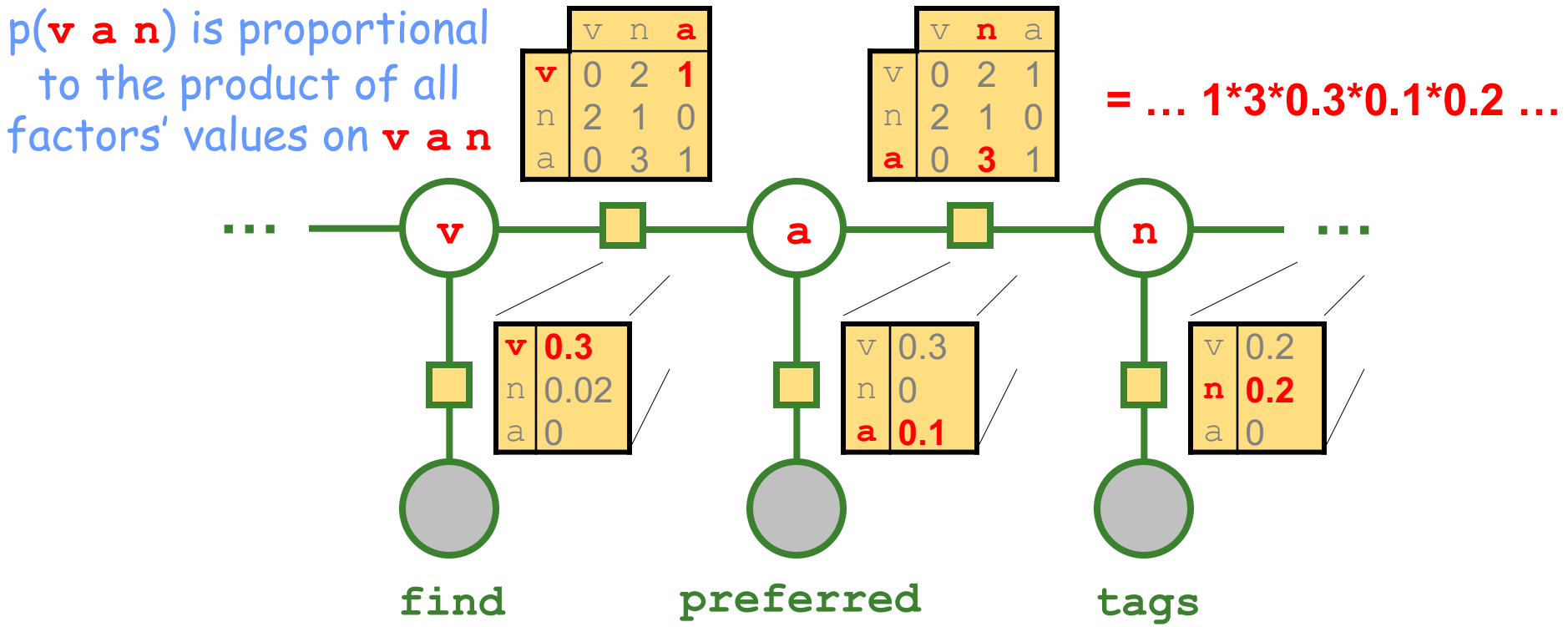


Conditional Random Field (CRF)

p(**v a n**) is proportional to the product of all factors' values on **v a n**

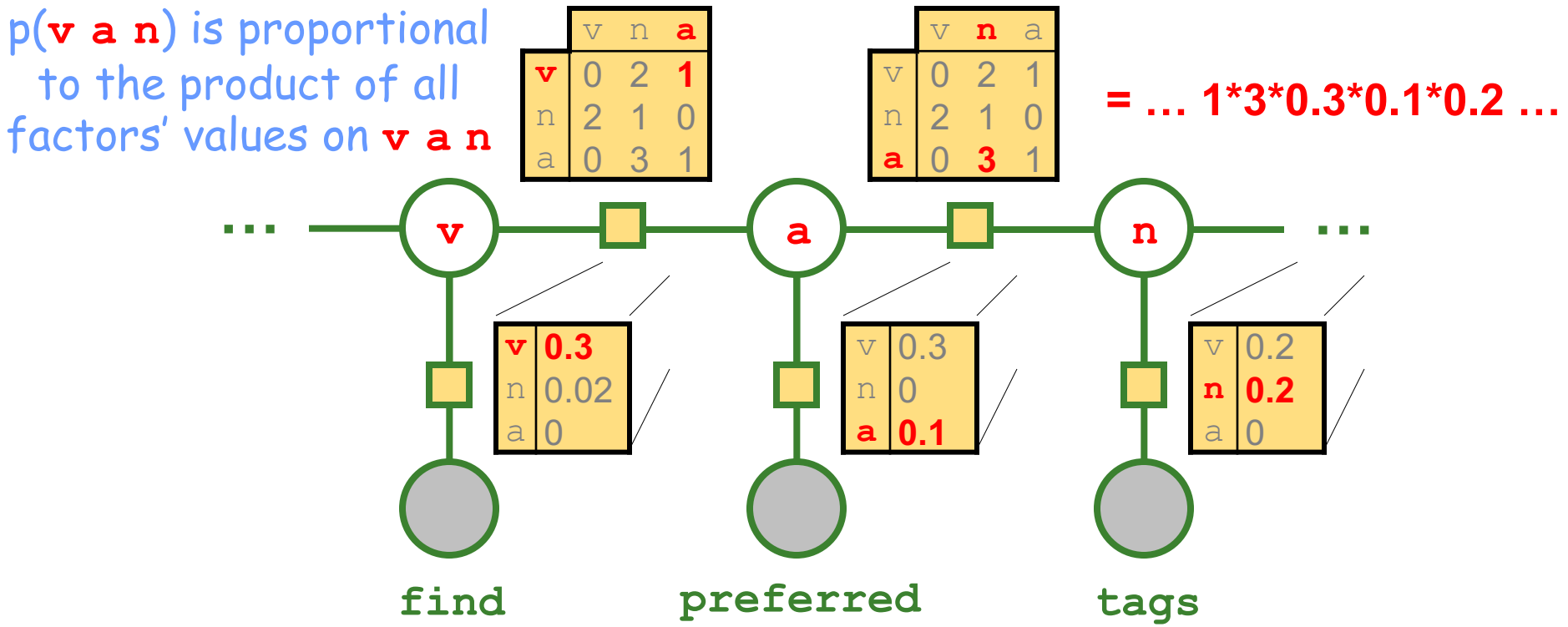


Conditional Random Field (CRF)



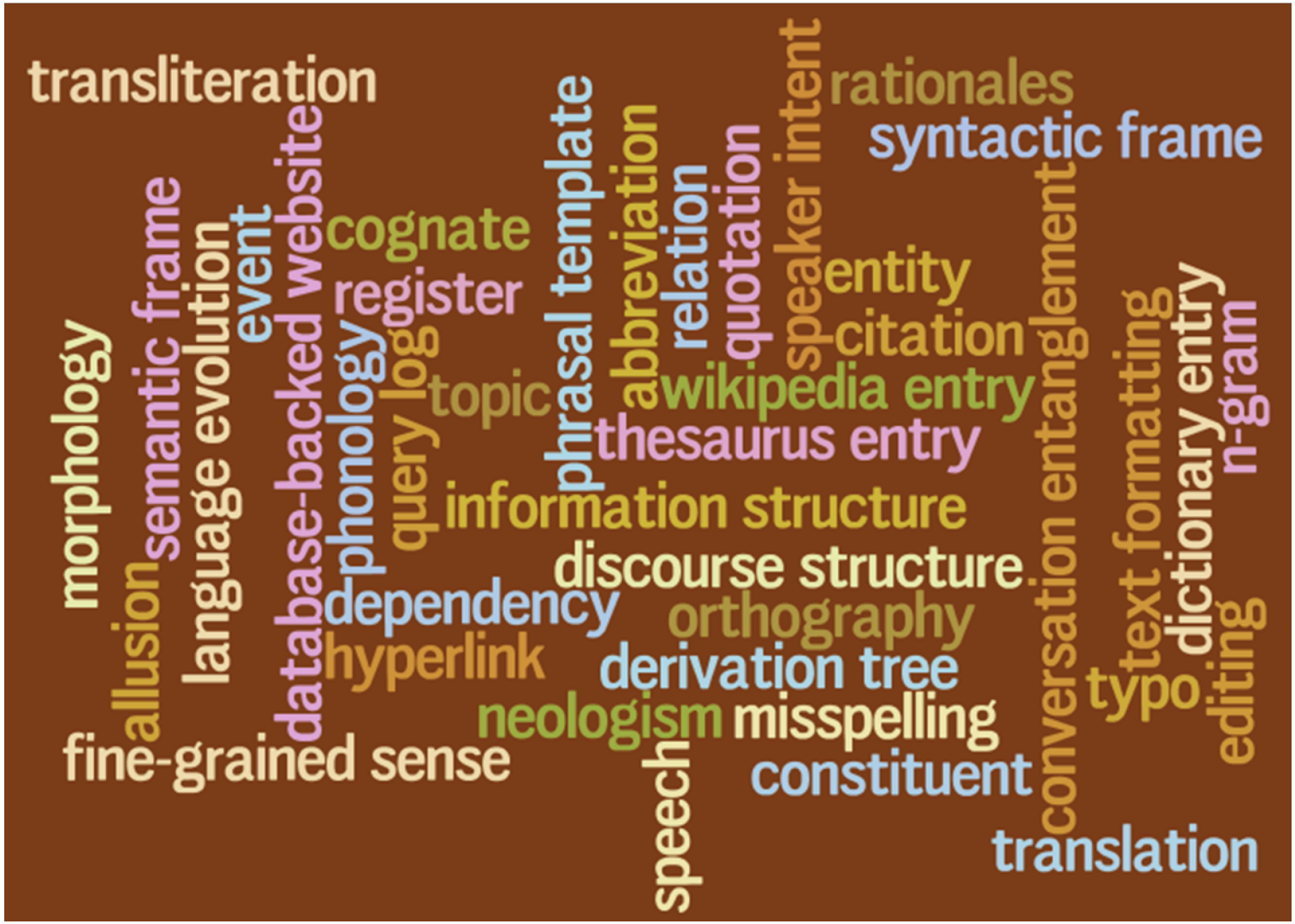
MRF vs. CRF?

Inference: What do you know how to compute with this model?

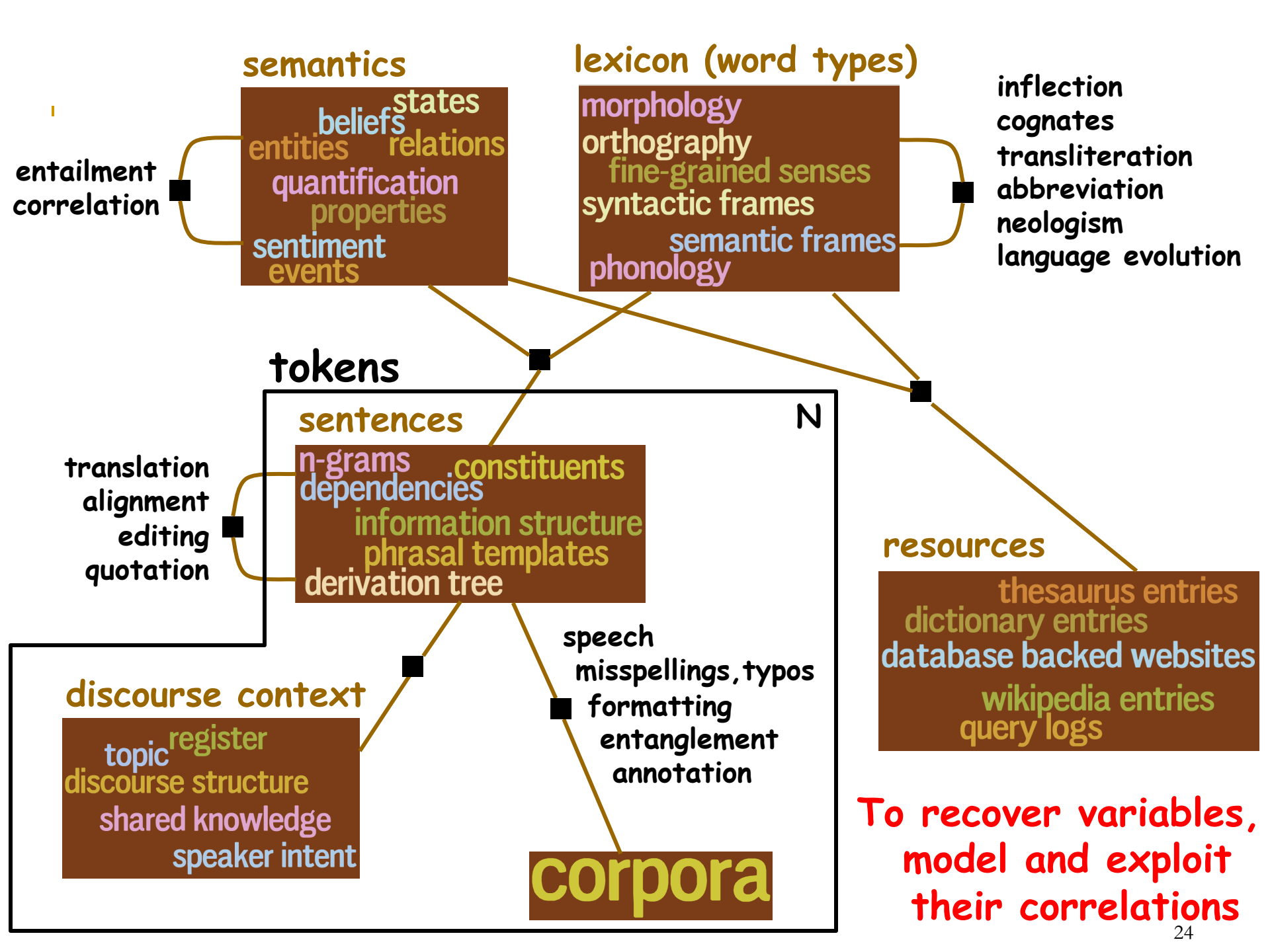


Maximize, sample, sum …

Variable-centric view of the world



When we deeply understand language, what representations (type and token) does that understanding comprise?

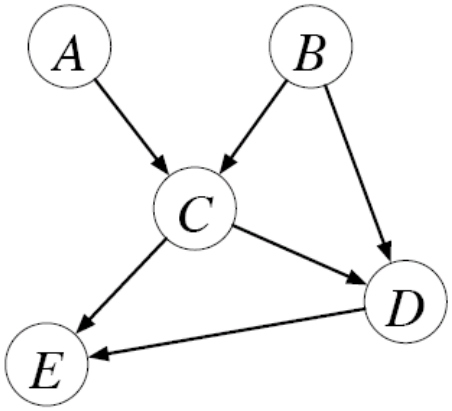

semantics
states
beliefs
entities
relations
quantification
properties
sentiment
events
entailment
correlation
lexicon (word types)
morphology
orthography
fine-grained senses
syntactic frames
semantic frames
phonology
inflection
cognates
transliteration
abbreviation
neologism
language evolution
tokens
N
sentences
n-grams
constituents
dependencies
information structure
phrasal templates
derivation tree
translation
alignment
editing
quotation
resources
thesaurus entries
dictionary entries
database backed websites
wikipedia entries
query logs
speech
misspellings, typos
formatting
entanglement
annotation
discourse context
topic
register
discourse structure
shared knowledge
speaker intent
corpora
To recover variables, model and exploit their correlations

How do you design the factors?

- It's easy to connect "English sentence" to "Portuguese sentence" …
  - … but you have to design a specific function that measures how compatible a pair of sentences is.

- Often, you can think of a generative story in which the individual factors are themselves probabilities.
  - May require some latent variables.

Directed graphical models (Bayes nets)



A DAG Model / Bayesian network[1] corresponds to a factorization of the joint probability distribution:

Under any model: $p(A, B, C, D, E) = p(A)p(B|A)p(C|A,B)p(D|A,B,C)p(E|A,B,C,D)$

Model above says: $p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|B, C)p(E|C, D)$

In general:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | X_{\mathsf{pa}(i)})$$

where $\mathsf{pa}(i)$ are the parents of node i.

Unigram model for generating text

$p(w_1) \cdot p(w_2) \cdot p(w_3) \ldots$

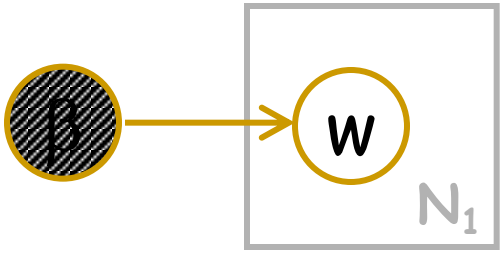
Explicitly show model's parameters β

"β is a vector that says which unigrams are likely"

$p(\beta) \cdot p(w_1 \mid \beta) \cdot p(w_2 \mid \beta) \cdot p(w_3 \mid \beta) \ldots$

"Plate notation" simplifies diagram

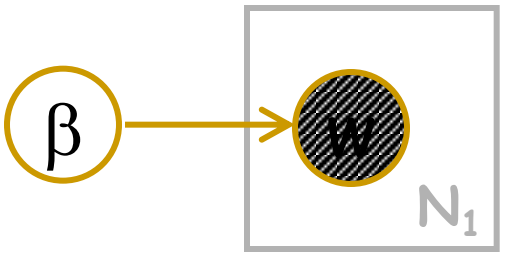
"β is a vector that says which unigrams are likely"



$p(\beta) \cdot p(w_1 \mid \beta) \cdot p(w_2 \mid \beta) \cdot p(w_3 \mid \beta) \ldots$

Learn β from observed words

(rather than vice-versa)

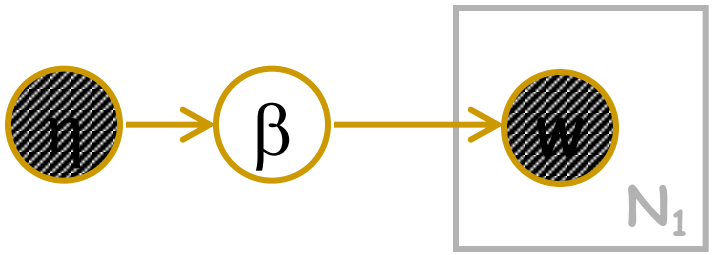


$$p(\beta) \cdot p(w_1 \mid \beta) \cdot p(w_2 \mid \beta) \cdot p(w_3 \mid \beta) \ldots$$

Explicitly show prior over β *(e.g., Dirichlet)*

η given
$\beta \sim \text{Dirichlet}(\eta)$
$w_i \sim \beta$

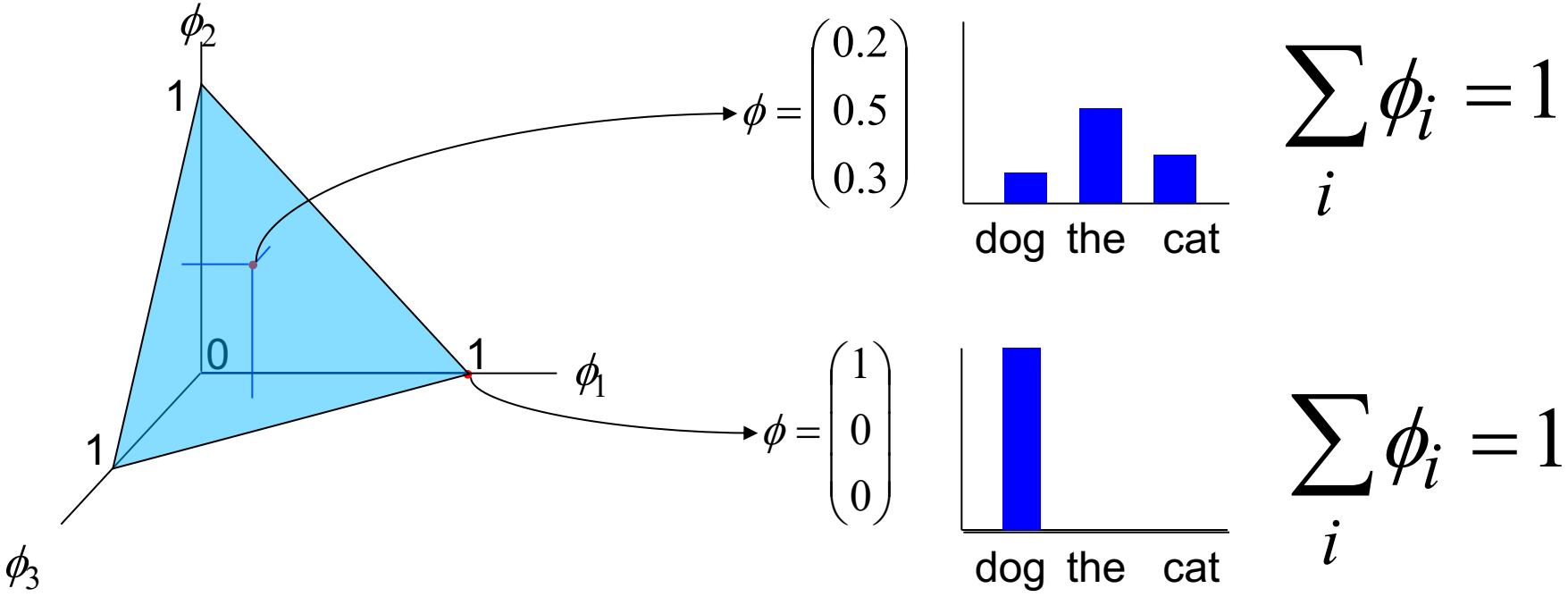
"Even if we didn't observe word 5, the prior says that $\beta_5 = 0$ is a terrible guess"



$p(\eta) \cdot p(\beta \mid \eta) \cdot p(w_1 \mid \beta) \cdot p(w_2 \mid \beta) \cdot p(w_3 \mid \beta) \ldots$

Dirichlet Distribution

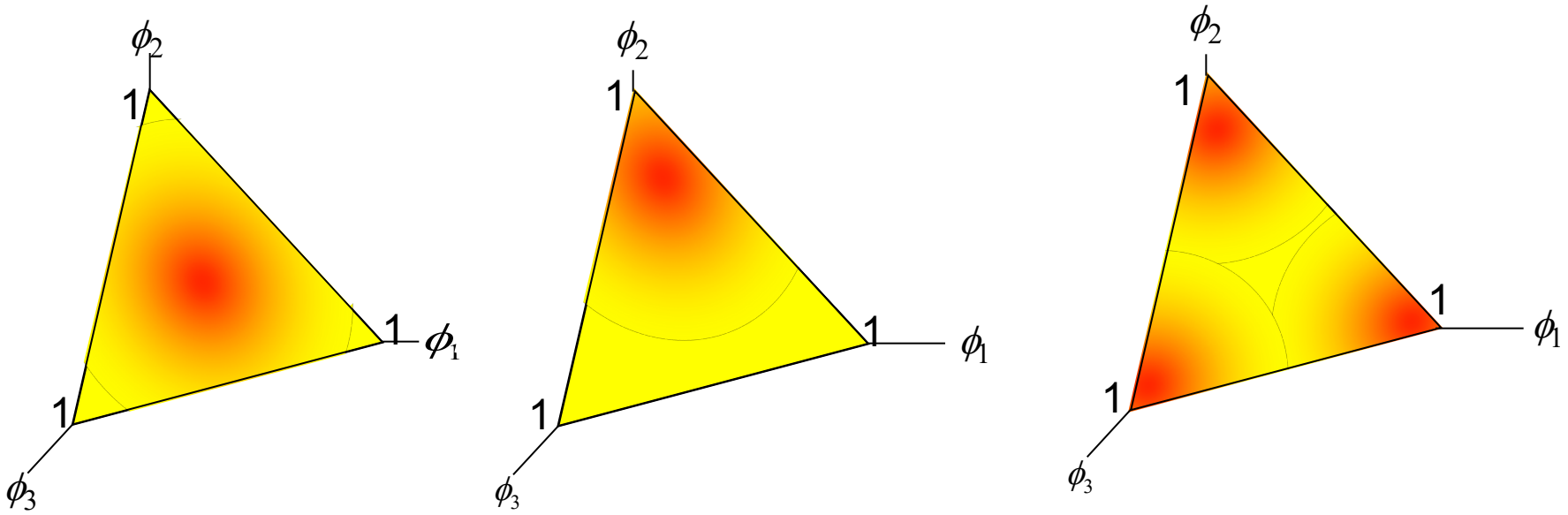
Each point on a k dimensional simplex is a multinomial probability distribution:



slide thanks to Nigel Crook

Dirichlet Distribution

A Dirichlet Distribution is a distribution over multinomial distributions ϕ *in the simplex.*



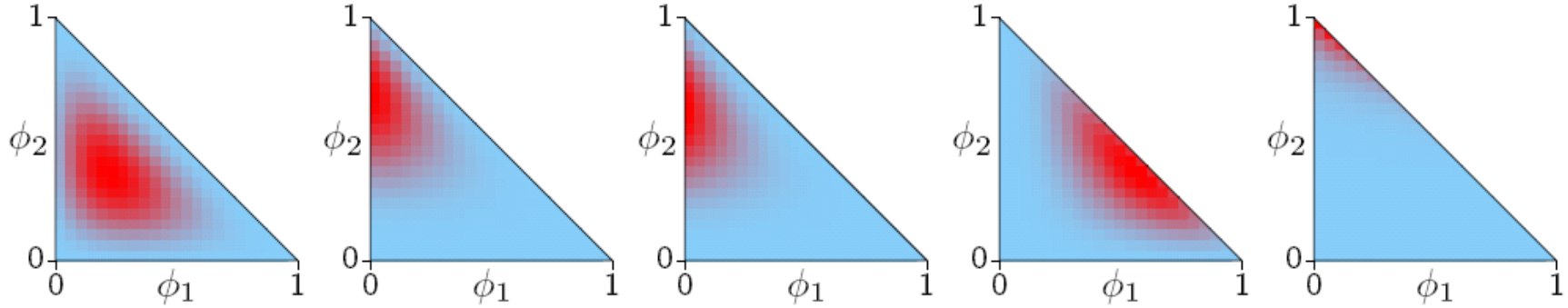
slide thanks to Nigel Crook

Distributions over multinomial parameters

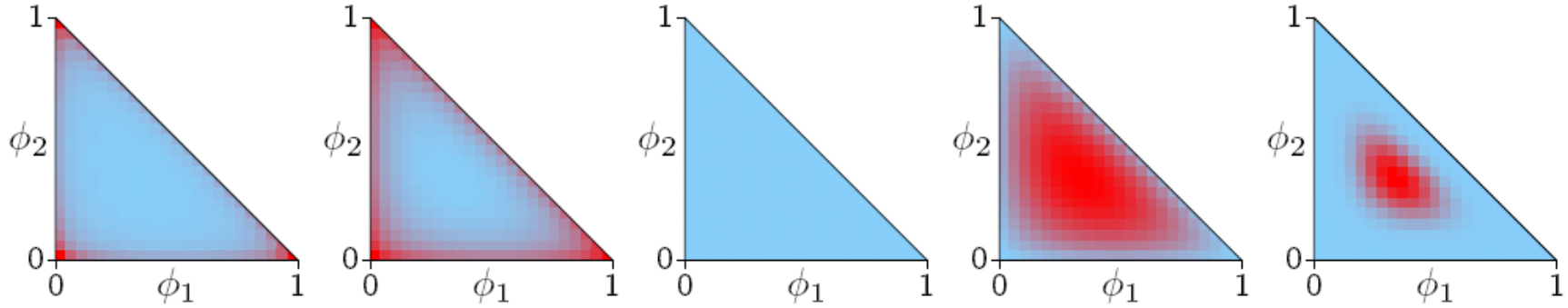
A Dirichlet distribution is a distribution over multinomial parameters ϕ in the simplex.

Like a Gaussian, there's a notion of mean and variance.

Different means:



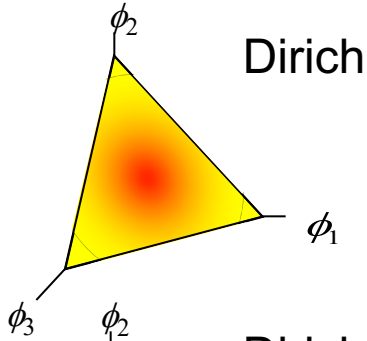
Different variances:



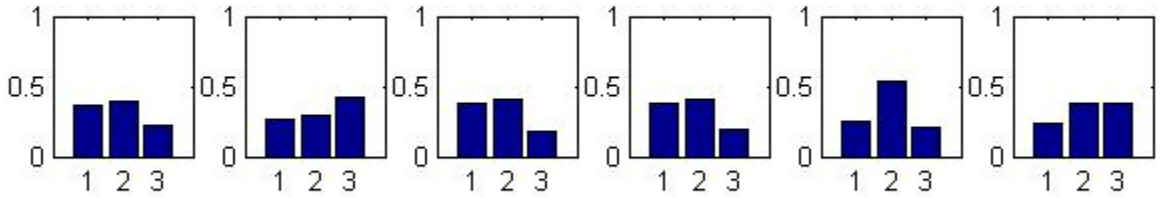
slide thanks to Percy Liang and Dan Klein

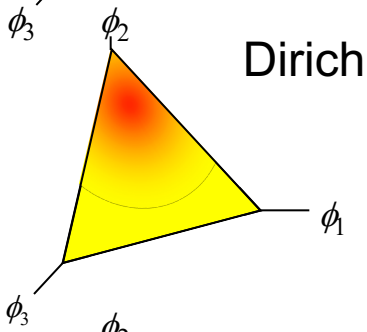
Dirichlet Distribution

Example draws from a Dirichlet Distribution over the 3-simplex:

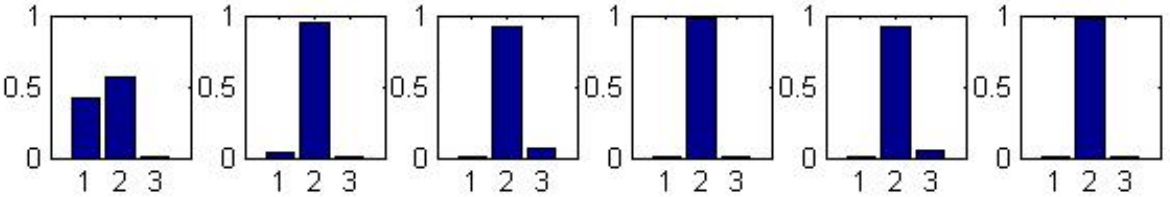


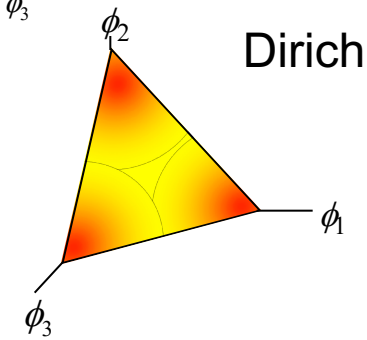
Dirichlet(5,5,5)



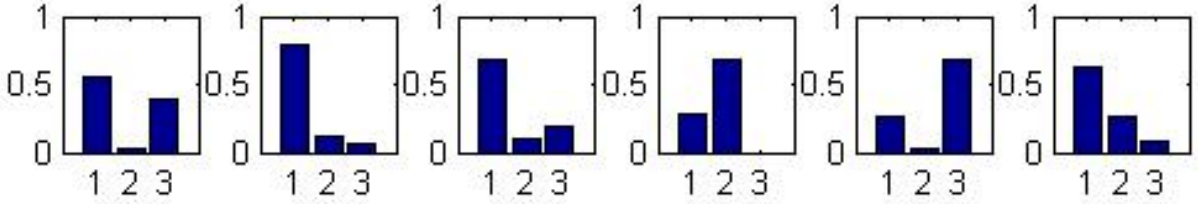


Dirichlet(0.2, 5, 0.2)





Dirichlet(0.5,0.5,0.5)



slide thanks to Nigel Crook

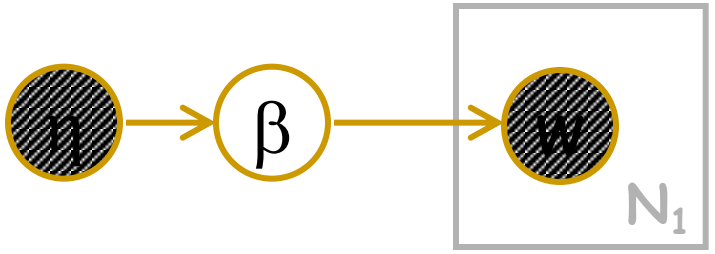
Explicitly show prior over β *(e.g., Dirichlet)*

Posterior distribution
$p(\beta \mid \eta, \mathbf{w})$
is also a Dirichlet
just like the prior $p(\beta \mid \eta)$.

"Even if we didn't observe word 5, the prior says that $\beta_5 = 0$ is a terrible guess"

prior = Dirichlet(η) ➔ posterior = Dirichlet(η+counts(**w**))

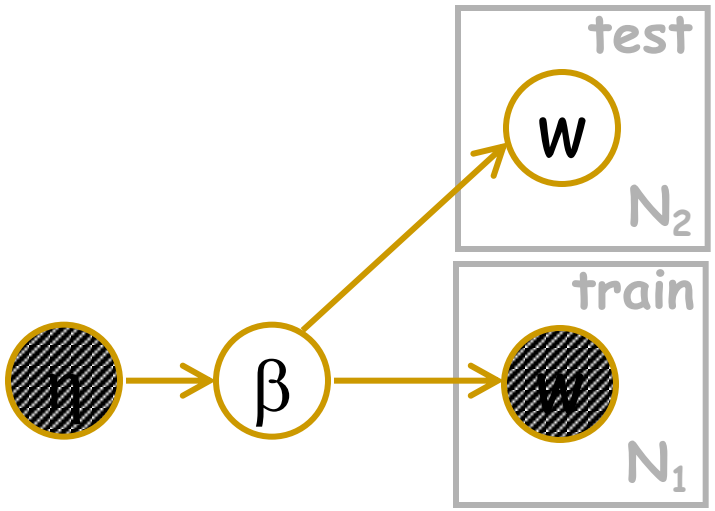
Mean of posterior is like the max-likelihood estimate of β, but smooth the corpus counts by adding "pseudocounts" η. (But better to use whole posterior, not just the mean.)



$p(\eta) \cdot p(\beta \mid \eta) \cdot p(w_1 \mid \beta) \cdot p(w_2 \mid \beta) \cdot p(w_3 \mid \beta) \ldots$

Training and Test Documents

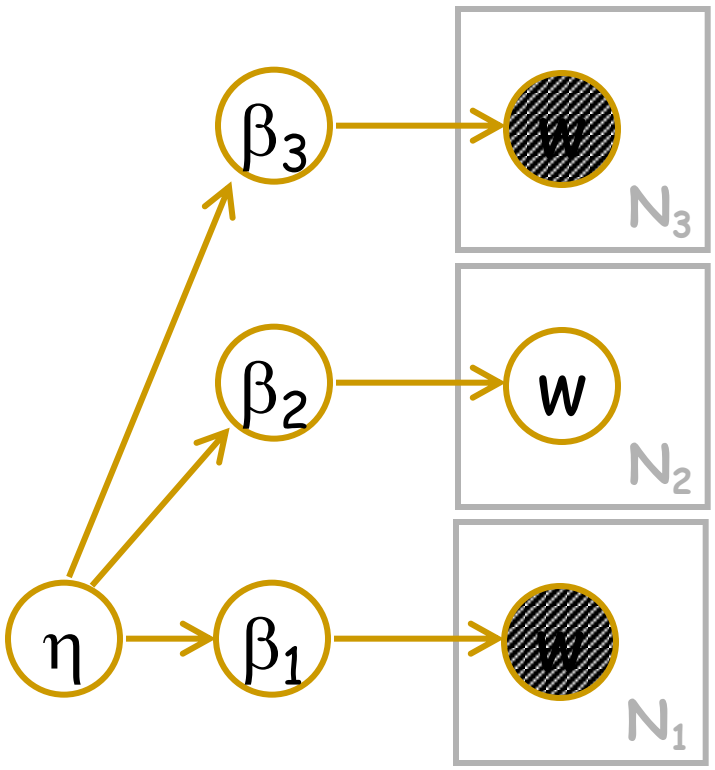
"Learn β from document 1, use it to predict document 2"



What do good configurations look like if N_1 is large?

What if N_1 is small?

Many Documents



"Each document has its own unigram model"

Now does observing docs 1 and 3 help still predict doc 2?

Only if η learns that all the β's are similar (low variance).

And in that case, why even have separate β's?

Many Documents

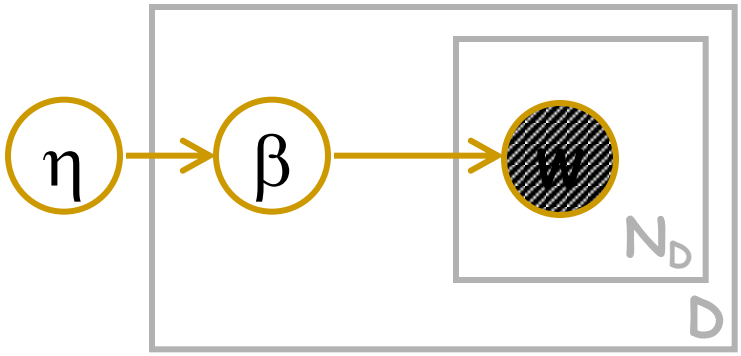
"Each document has its own unigram model"

or tuned to maximize training or dev set likelihood

η given

$\beta_d \sim \text{Dirichlet}(\eta)$

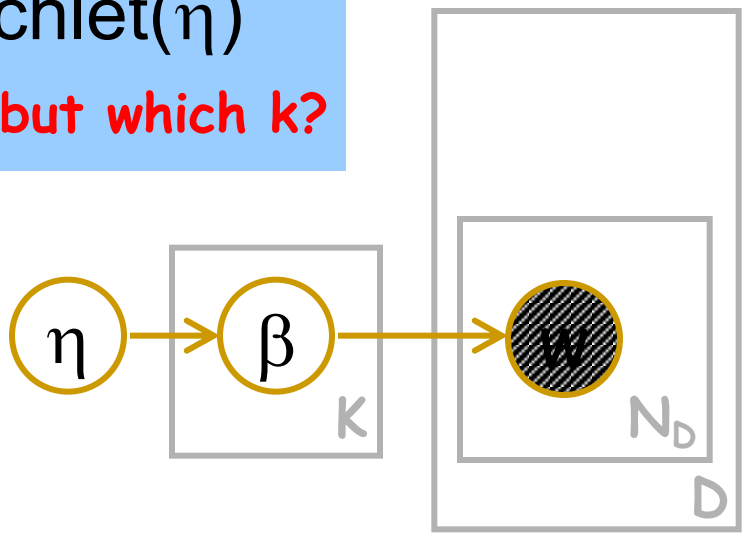
$w_{di} \sim \beta_d$



Bayesian Text Categorization

"Each document chooses one of only K topics (unigram models)"

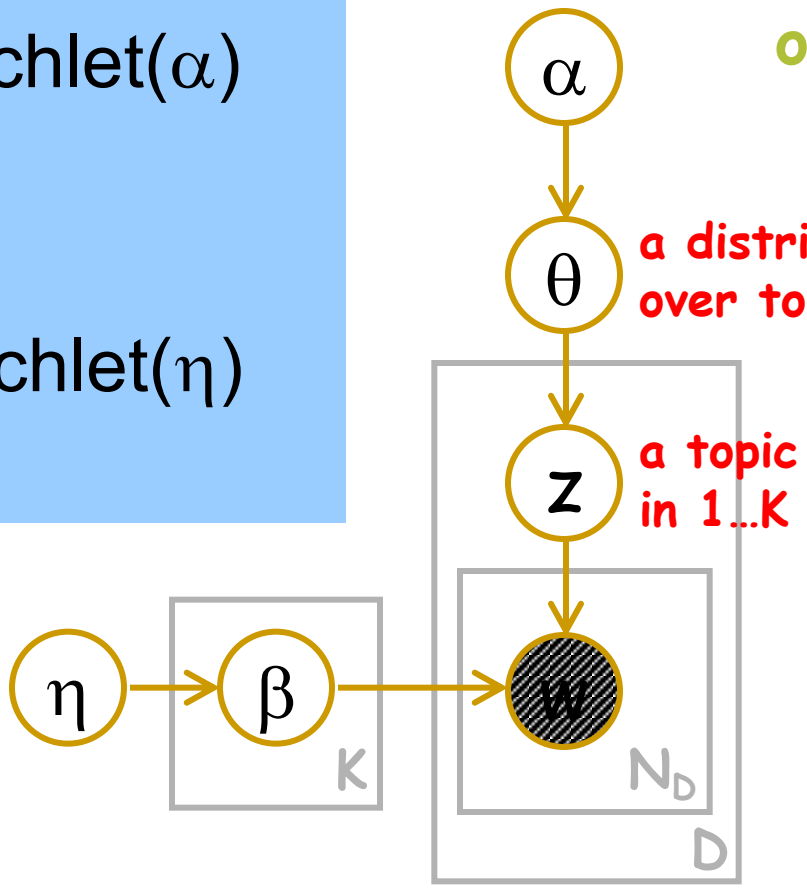
η given
$\beta_k \sim$ Dirichlet(η)
$w_{di} \sim \beta_k$ but which k?



Bayesian Text Categorization

α given
$\theta \sim$ Dirichlet(α)
$z_d \sim \theta$

η given
$\beta_k \sim$ Dirichlet(η)
$w_{di} \sim \beta_{z_d}$



"Each document chooses one of only K topics (unigram models)"

a distribution over topics 1…K

a topic in 1…K

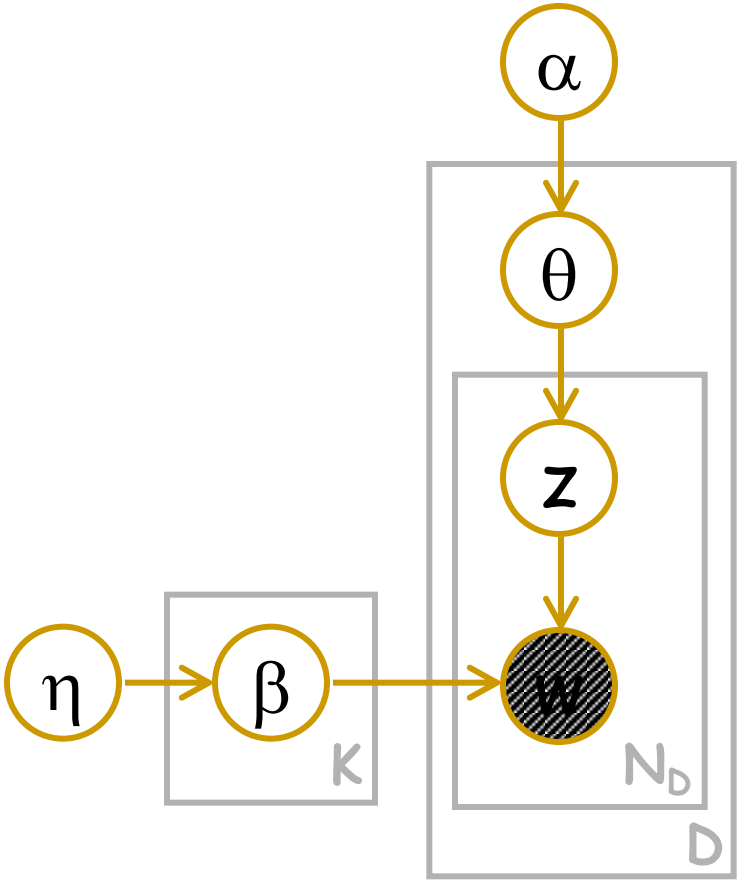
Allows documents to differ considerably while some still share β parameters.

And, we can infer the probability that two documents have the same topic z.

Might observe some topics.

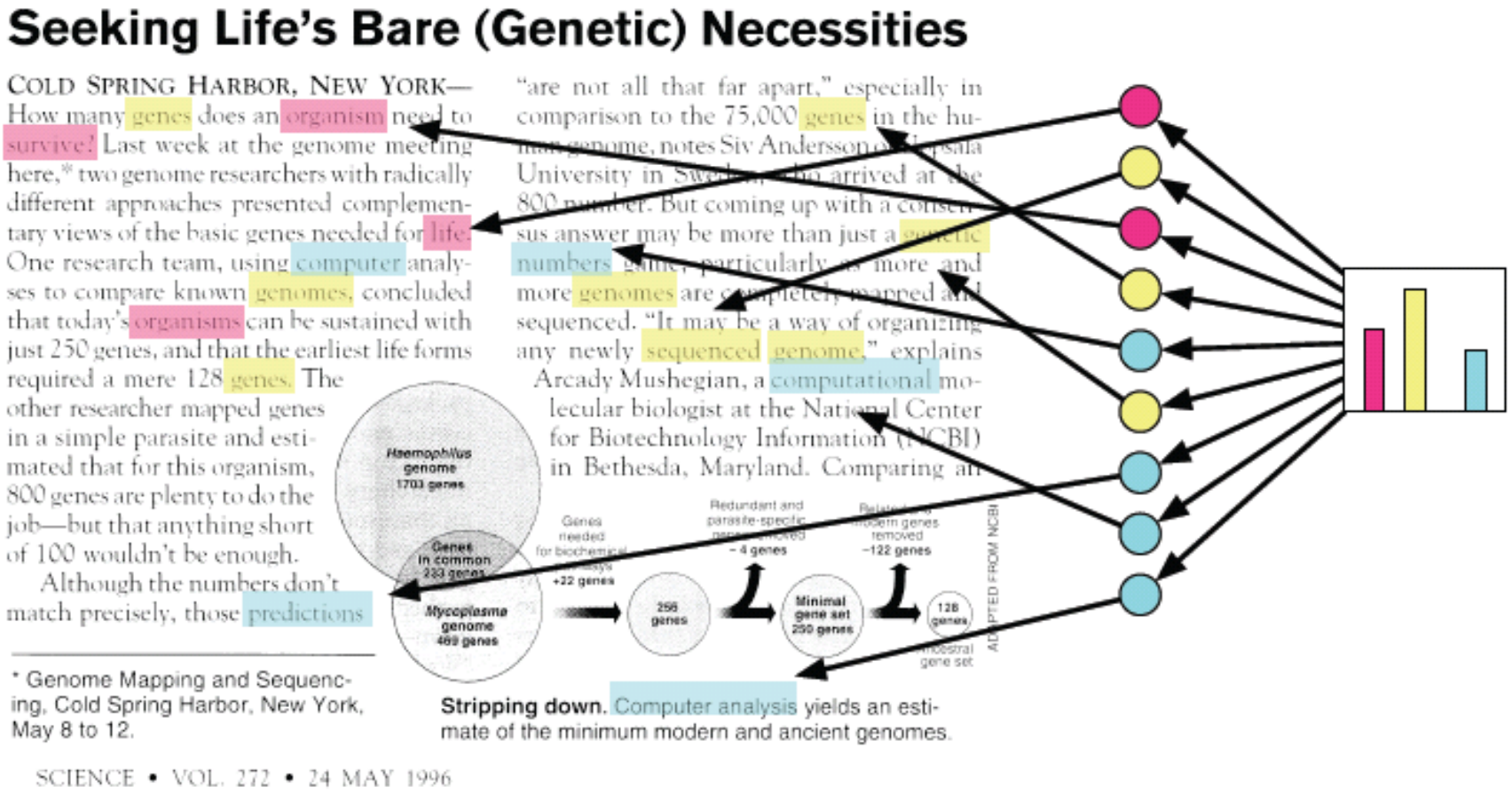
Latent Dirichlet Allocation

(Blei, Ng & Jordan 2003)



"Each document chooses a *mixture* of all K topics; each word gets its own topic"

(Part of) one assignment to LDA's variables



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

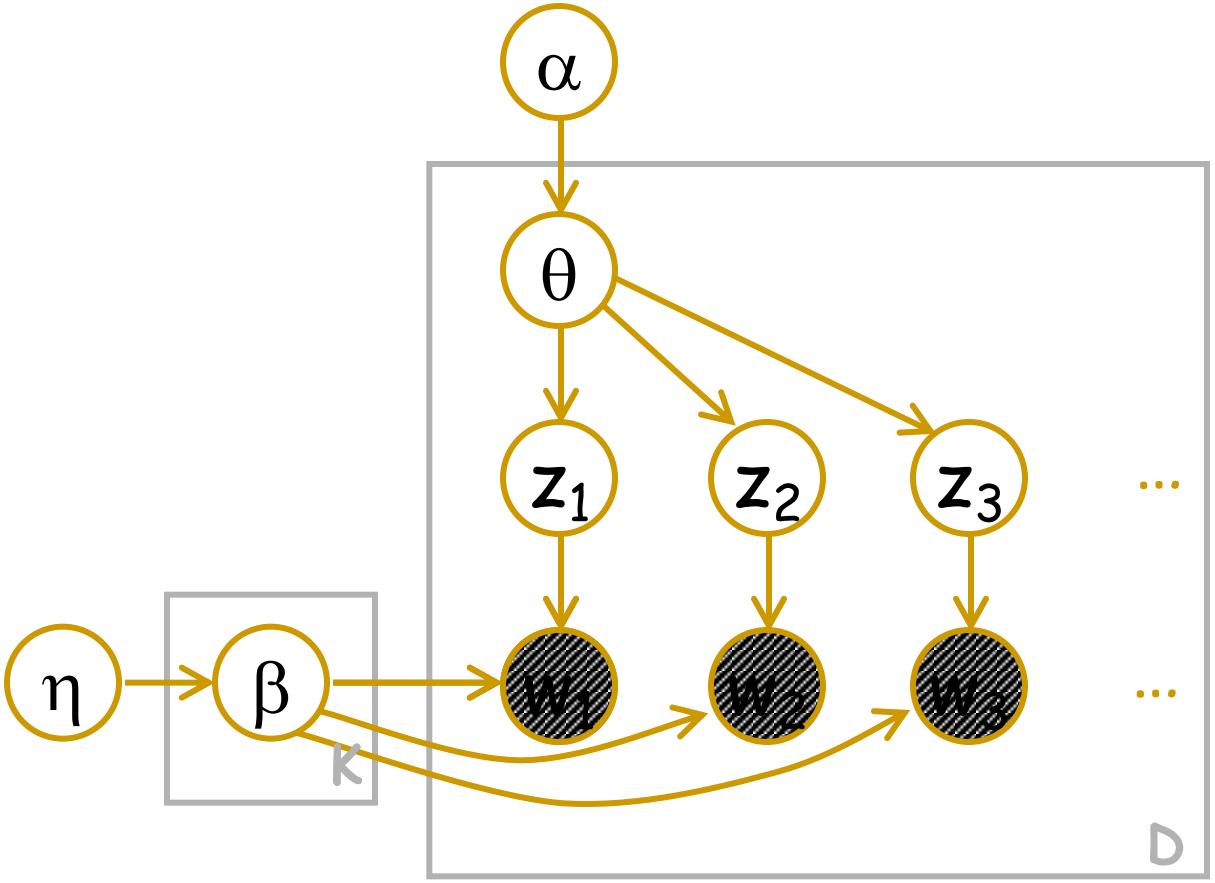
slide thanks to Dave Blei

(Part of) one assignment to LDA's variables

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

slide thanks to Dave Blei

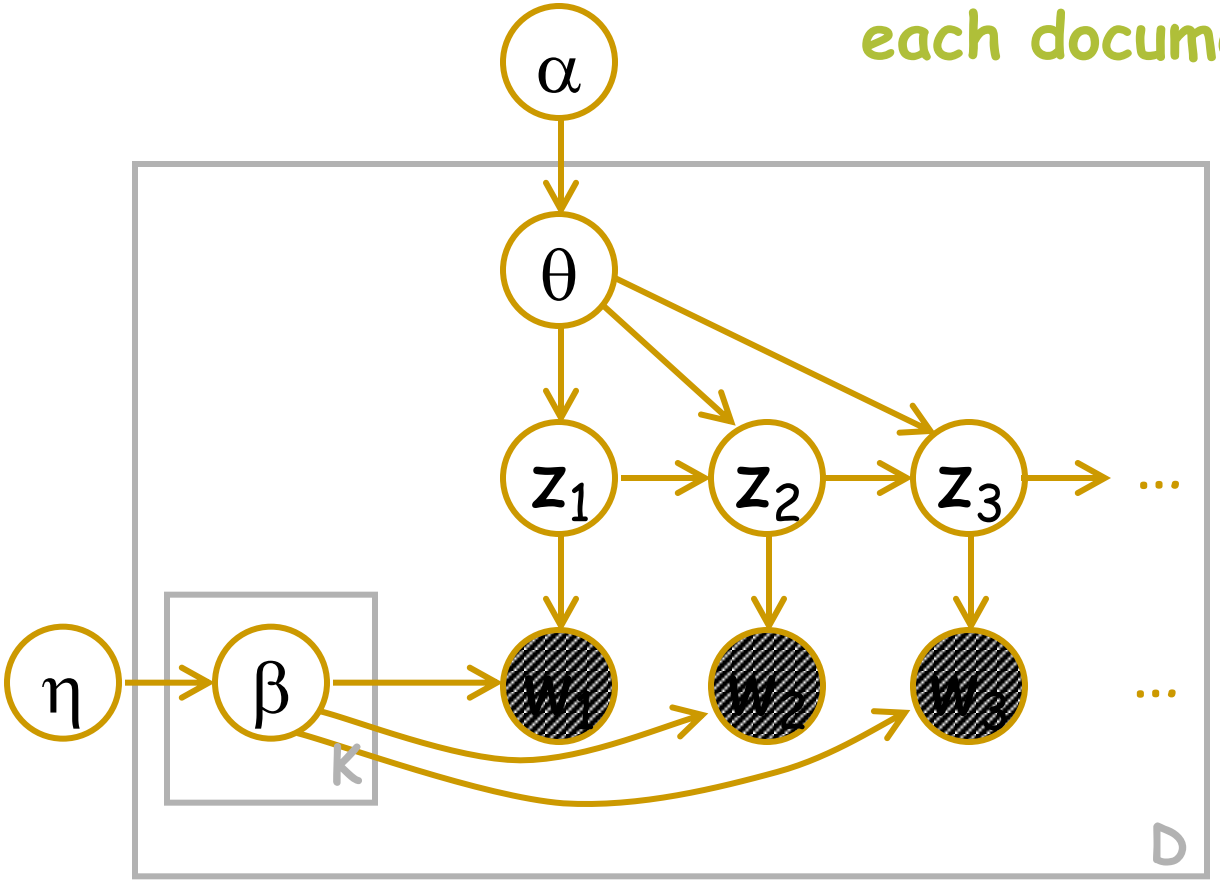
Latent Dirichlet Allocation: Inference?



Finite-State Dirichlet Allocation

(Cui & Eisner 2006)

"A different HMM for each document"



Variants of Latent Dirichlet Allocation

- Syntactic topic model: A word or its topic is influenced by its syntactic position.
- Correlated topic model, hierarchical topic model, …: Some topics resemble other topics.
- Polylingual topic model: All versions of the same document use the same topic mixture, even if they're in different languages. (Why useful?)
- Relational topic model: Documents on the same topic are generated separately but tend to link to one another. (Why useful?)
- Dynamic topic model: We also observe a year for each document. The k topics β used in 2011 have evolved slightly from their counterparts in 2010.

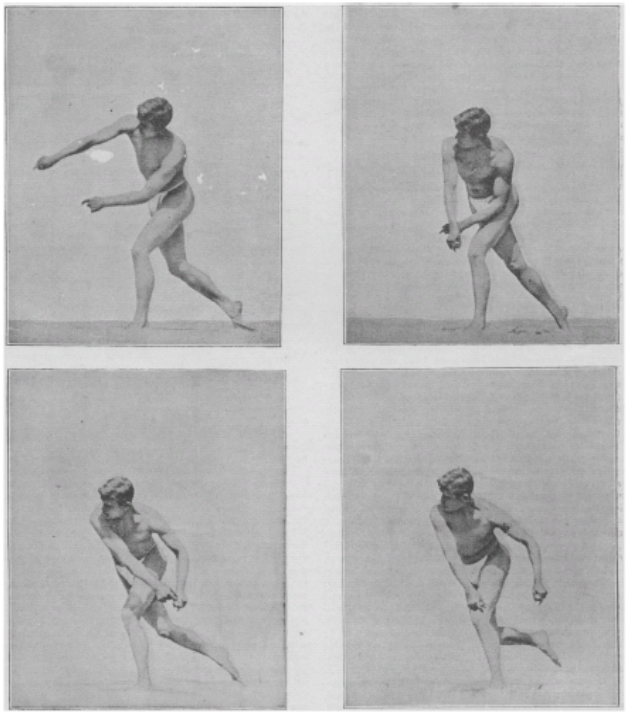
Dynamic Topic Model

"Instantaneous Photography" (1890)

Prussia, who has taken thousands of pictures of flying birds, running horses, jumping men, etc., all admirable for their perfect "technique," and for the great artistic tact and scientific skill with which the moments of exposure had been chosen. In these pictures the characteristic positions peculiar to different motions are well presented. Many of them at first appear absolutely unnatural, because the eye has never been able to observe them.

These pictures produced rich and important material for the study of motion, but Mr. Anschuetz succeeded in making his experiments more valuable by obtaining whole series of pictures giving the different phases of motion. He made it his object to get of one period of motion, for instance, of the step of a walking man, as many views as possible in equal intervals of time, and he succeeded admirably in his undertaking. He was able to observe in this manner even the fastest motion, for instance, the hurdle-jump of a racing horse, which occupies only seventy-two one-hundredths of a second, and in this short time made twenty-four pictures of the different positions in equal intervals. A dozen pictures showing the different phases of position assumed by an athlete in throwing a javelin, reproduced from instantaneous photographs taken by Mr. Anschuetz, are given on this and the preceding pages.

Mr. Anschuetz next constructed an apparatus which he called the electric tachyscope, in which he was financially assisted by the German Government. In this instrument the series o-



INSTANTANEOUS PHOTOGRAPHS OF AN ATHLETE THROWING A JAVELIN.

"Infrared Reflectance in Leaf-Sitting Neotropical Frogs" (1977)

North American frogs so examined [*Bufo debilis*, *B. boreas* (2), *B. coniferus*; *Rana pipiens* (2), *R. palmipes*, *R. catesbeiana*; *Hyla cinerea*, *H. squirella*, *H. euphorbiacea*, *H. chaneque*, and *H. cyanomma*] absorb infrared light and stand out sharply against foliage (Fig. 1).

Cott (3), using black and white infrared film, found that the Australian tree-frog *Hyla coerulea* (=*Litoria caerulea*) reflects infrared light. *Litoria caerulea*, *A. moreletii*, and *A.* (=*Pachymedusa*) *dacnicolor* all contain a newly discovered red pigment in unusual melanosomes (4). Both *fleischmanni* and *prosoblepon* groups of *Centrolenella* contain a purple pigment in their chromatophores (5). Whether these two skin pigments are identical, or play any role in infrared reflectance, has not been determined.

There are two likely functions for infrared reflectance in leaf-sitting frogs. (i) Although the near-infrared is not heat (6), photons of these wavelengths will lose energy as heat if they are absorbed by the skin. Thus, the ability to reflect infrared may play a physiological role in thermoregulation by preventing excessive heat gain. (ii) Infrared reflectance may conceal frogs from predators with infrared receptors (7). Little research has been done on near-infrared sensitivity, and supportive evidence is sparse. Both the eyes of birds and the pit organs of snakes may act as near-infrared light receptors. In pigeons and chickens, the sensitivity maxima of the eyes are shifted toward longer wavelengths than those of humans (7), and the tawny owl responds to infrared light (900 nm) (8). Visual sensitivity extending just into the near-infrared would allow birds to see most green frogs on green leaves, although centrolenids and phyllomedusines would remain camouflaged. Boid and crotaline pit organs are usually interpreted as thermal detectors, adaptations for nocturnal predation on warm-blooded prey (9). In diurnal snakes, however, these receptors may be used to detect frogs that act as infrared sinks among leaves that are reflecting light of these wavelengths. The facial pits of crotaline snakes are directionally sensitive and may allow infrared depth perception (10). Many species of birds and snakes are known to eat frogs and forage in their diurnal retreats. Predation by birds and snakes may have selected for infrared cryptic coloration in tropical leaf-sitting frogs.

PATRICIA A. SCHWALM*
PRISCILLA H. STARRETT
Department of Biological Sciences, Allan Hancock Foundation, University of Southern California, Los Angeles 90007
ROY W. MCDIARMID
Department of Biology, University of South Florida, Tampa 33620

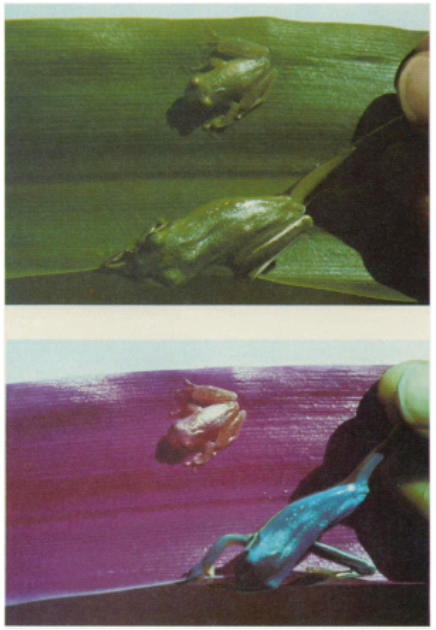


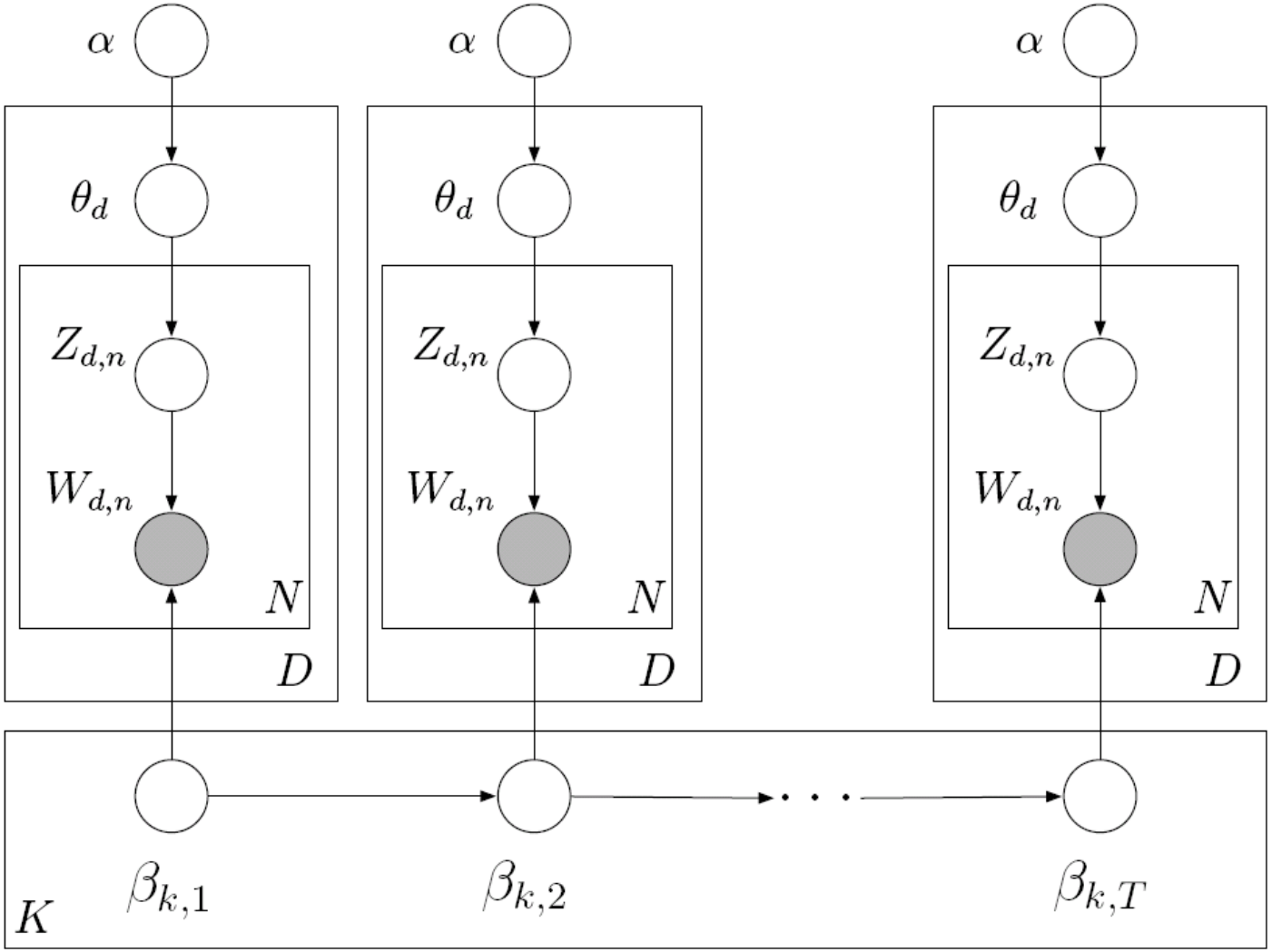
Fig. 1. A comparison of the color characteristics of a hylid and a centrolenid frog in a conventional (top) and an infrared (bottom) color photograph. Although both frogs match the green leaf in light ranges visible to man, only *Centrolenella fleischmanni* (top frog) reflects near-infrared light. This allows it to blend with foliage both in the visible and near-infrared ranges of light, unlike *Hyla cinerea* (bottom frog), which absorbs infrared and is distinguished from the leaf

References and Notes

1. Kodak Infrared Ektachrome film has a sensitivity extending to about 900 nm [*Kodak Publ. N-17* (1974)]. The three emulsion layers are sensitized to green, red, and infrared light rather than to blue, green, and red. By placing a yellow filter over the camera lens, blue light is excluded, resulting in a shift of colors in the photograph. Green objects appear blue, red objects appear green, and infrared appears red.
2. H. L. Gibson, W. R. Buckley, K. E. Whitmore, *J. Biol. Photogr. Assoc.* **33**, 1 (1965).
3. H. B. Cott, *Adaptive Coloration in Animals* (Methuen, London, 1940), pp. 9–11.
4. J. T. Bagnara and W. Ferris, *Copeia*, **3**, 592 (1975).
5. P. H. Starrett and J. M. Savage, *Bull. South. Calif. Acad. Sci.* **72**, 57 (1973).
6. Radiation in these wavelengths (700 to 900 nm) can be emitted by very hot objects, such as the sun or lamp filaments, but it is not heat per se. Similarly, the near-infrared can be reflected or emitted as luminescence by objects that are not hot themselves [*Kodak Publ. M-28* (1972)].
7. R. Granit, *Sensory Mechanisms of the Retina* (Oxford Univ. Press, London, 1947), pp. 296–296.
8. F. L. Vanderplank, *Proc. Zool. Soc. London* **1934**, 505 (1934).
9. P. H. Hartline, in *Handbook of Sensory Physiology*, vol. 3, part 3, *Electroreceptors and Other Specialized Receptors in Lower Vertebrates*, A. Fessard, Ed. (Springer-Verlag, New York,

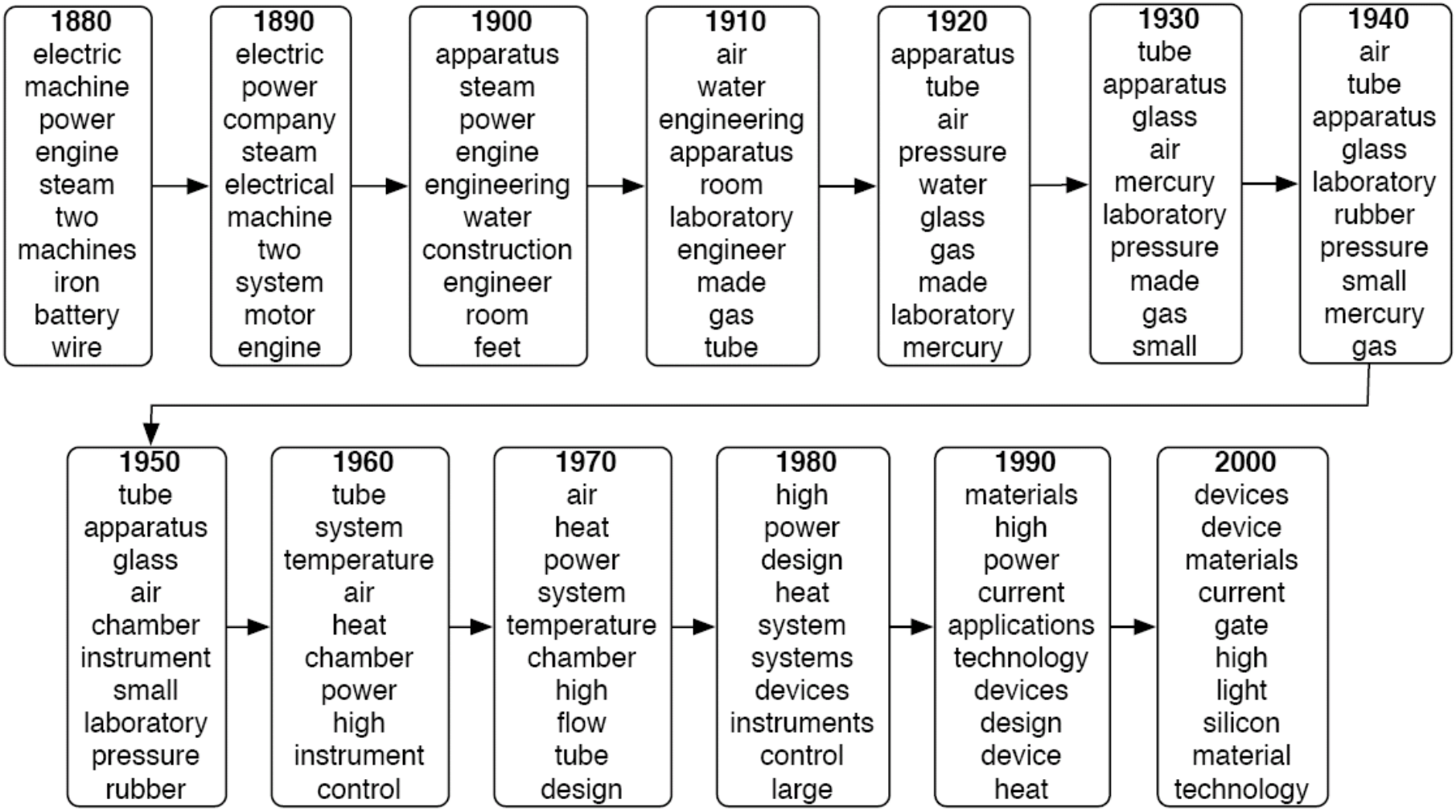
slide thanks to Dave Blei

Dynamic Topic Model



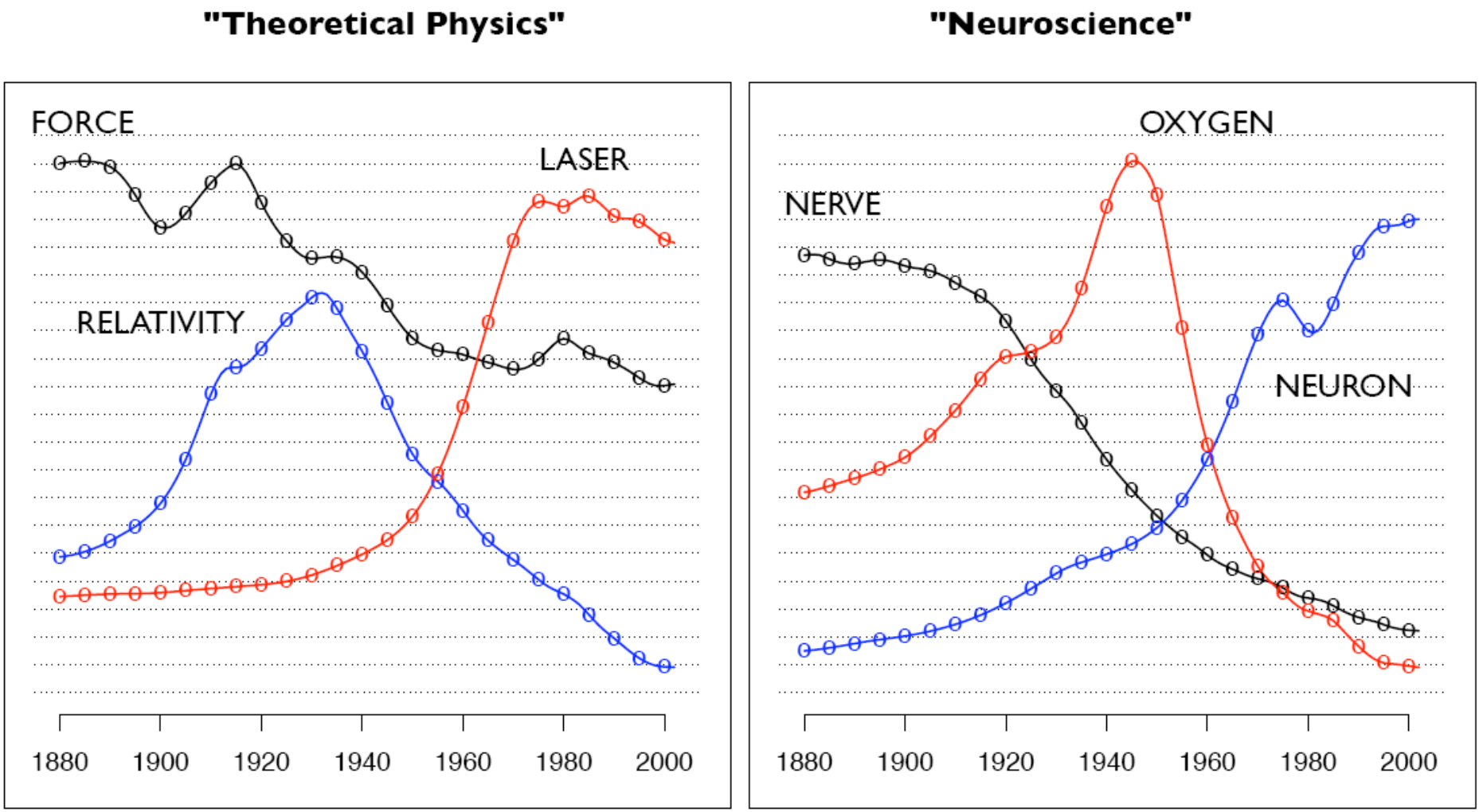
slide thanks to Dave Blei

Dynamic Topic Model



slide thanks to Dave Blei

Dynamic Topic Model

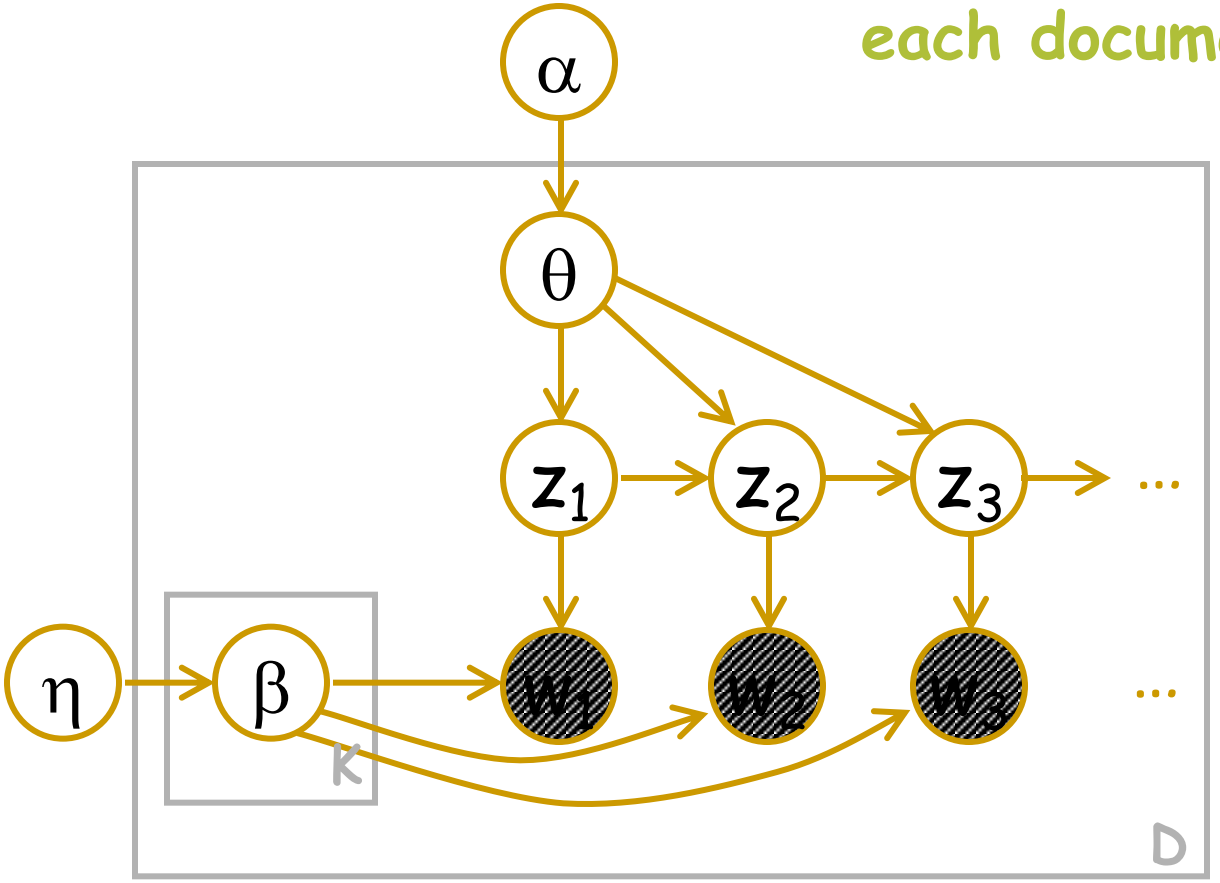


slide thanks to Dave Blei

Remember: Finite-State Dirichlet Allocation

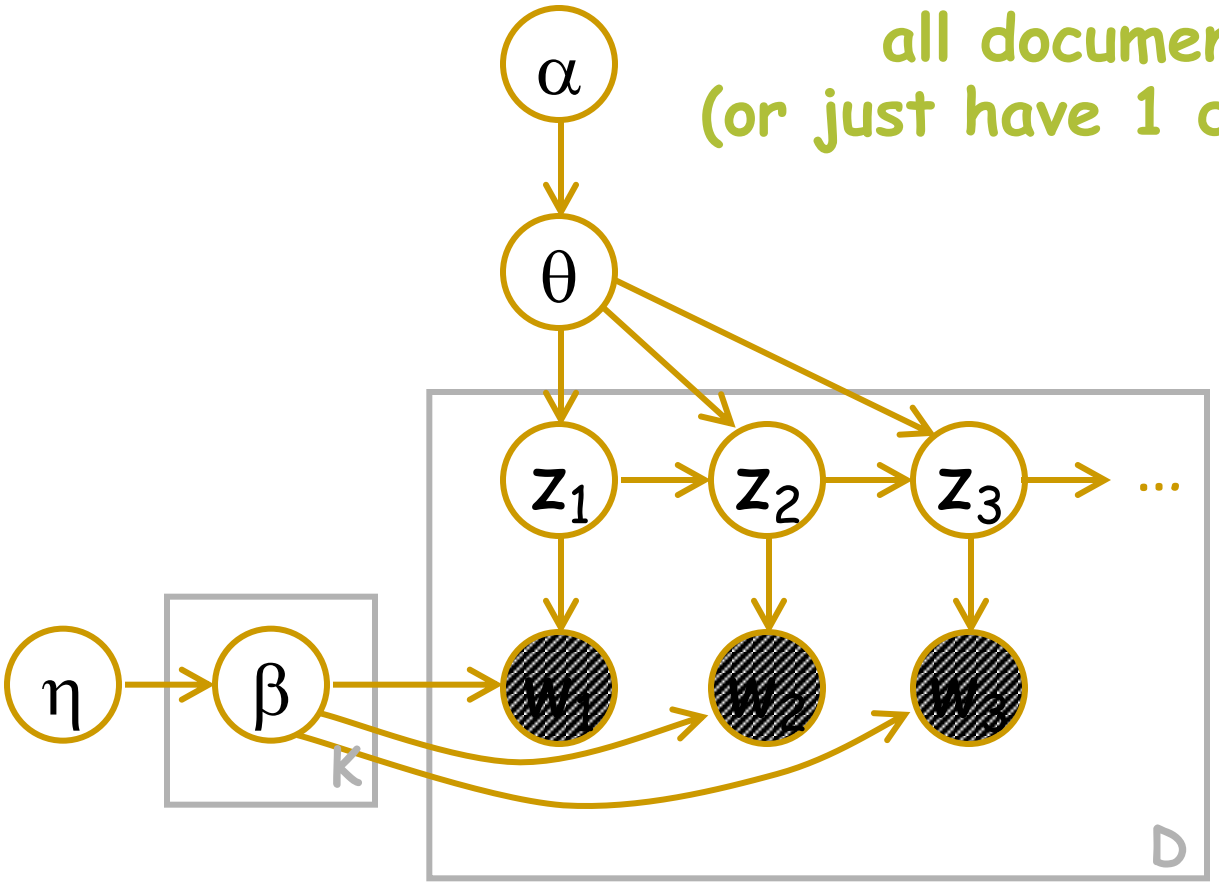
(Cui & Eisner 2006)

"A different HMM for each document"



Bayesian HMM

"Shared HMM for all documents" (or just have 1 document)



We have to estimate transition parameters θ and emission parameters β.

FIN