Recitation: Bayes Nets and Friends

What's your favorite thing to do to unwind?

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Lyle Ungar

Heavily adapted from slides by Mitch Marcus
With contributions from Tony Liu

Recitation Plan

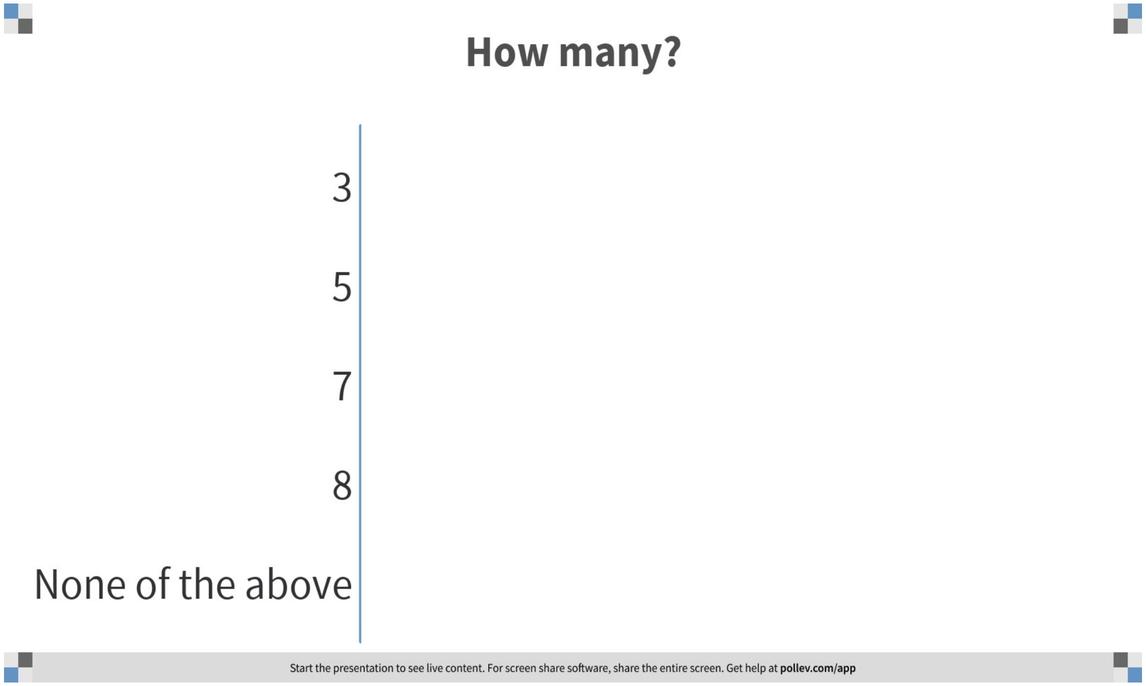
- Naïve Bayes Exercise
- LDA Example
- Bayes Net Exercises
- HMM Example

Recall: Naïve Bayes

- What’s the model?
- How do you estimate the parameters
- How is NLP Naïve Bayes different?

Naïve Bayes Exercise

Consider binary classification where ***x*** has 2 binary features.
How many parameters are there in a Naıve Bayes classifier?



How many?

3

5

7

8

None of the above

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Naïve Bayes Models

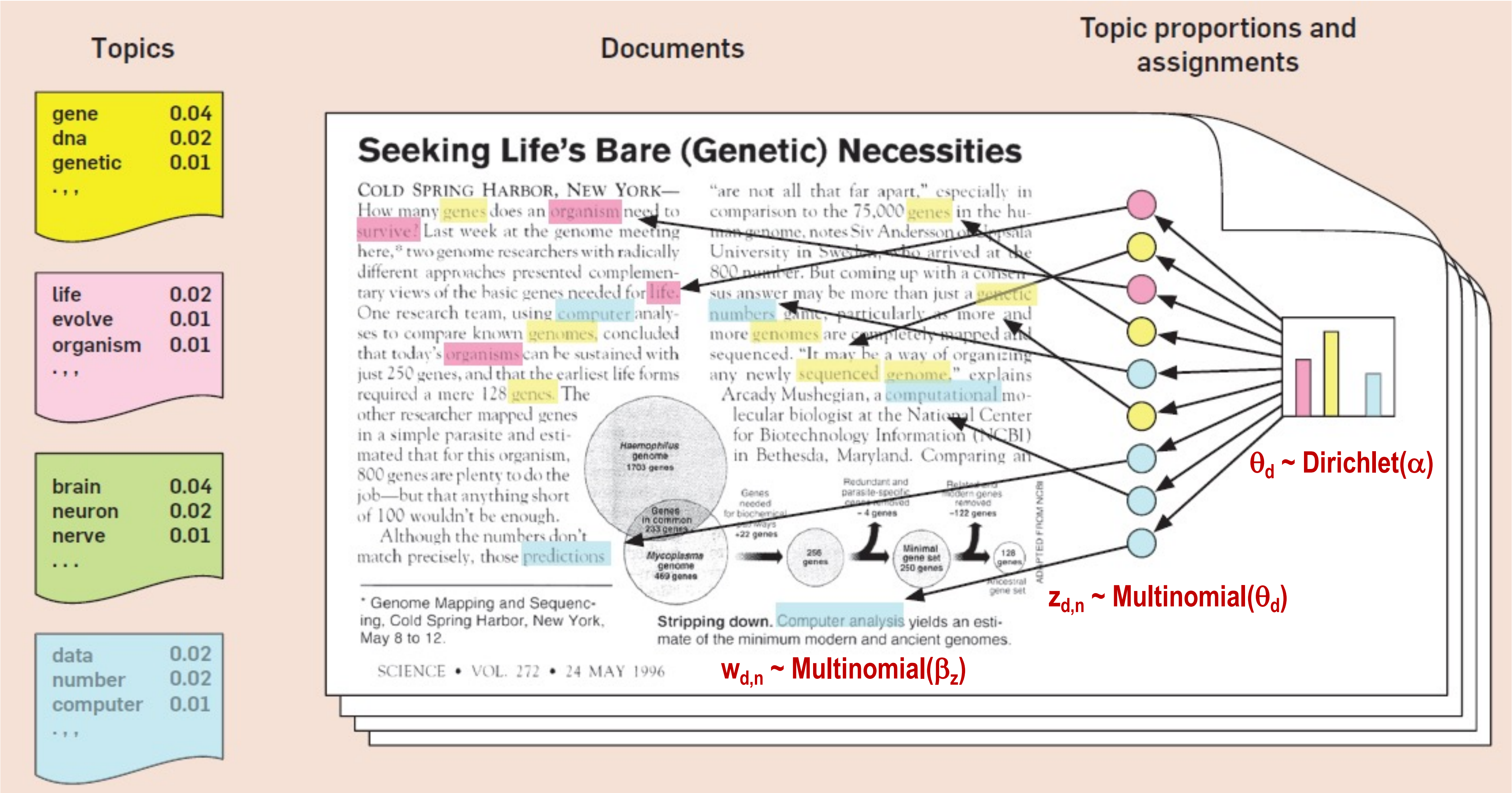
- **Different models**
  - $p(y|x) \sim p(x_1|y)\ (x_2|y)\ ..\ p(x_p|y)\ \ p(y)$
  - $p(x_j|y)$ can be Bernoulli or Gaussian or …

Naïve Bayes: Parameter Estimation

- **MAP – why not MLE?**
  - P(“apple”|class) = (#docs in class with “apple”) / (#docs in class)
- **Uninformed prior (Laplace smoothing)**
  - Add a document with each word to each class
  - (#docs in class with “apple” + 1) / (#docs in class + v)
- **Informed prior (Empirical Bayes)**
  - Add prior counts of each word proportionally to their frequency
  - (#docs in class with “apple” + m p(“apple")) / (#docs in class + m)

Naïve Bayes for NLP

- What additional assumption is made in Naïve Bayes for NLP?



Topics

gene	0.04
dna	0.02
genetic	0.01
. , ,	

life	0.02
evolve	0.01
organism	0.01
. , ,	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. , ,	

Documents

Topic proportions and assignments

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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

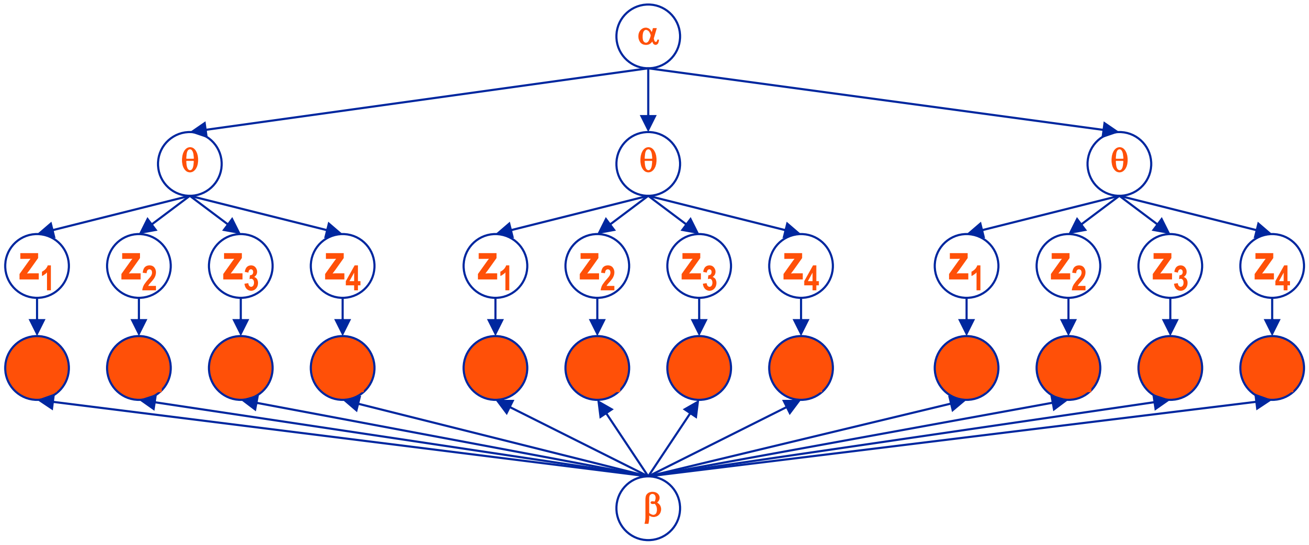
$\theta_d \sim \text{Dirichlet}(\alpha)$

$z_{d,n} \sim \text{Multinomial}(\theta_d)$

$w_{d,n} \sim \text{Multinomial}(\beta_z)$

David Blei 2012: Probabilistic Topic Models (on course website)

Recall: The LDA Model



Topic distribution

Topics of words

Words (observed)

- **For each document,**
  - Choose the topic distribution $\theta \sim \text{Dirichlet}(\alpha)$
  - For each of the N words w_n:
    - Choose a topic $z \sim \text{Multinomial}(\theta)$
    - Then choose a word $w_n \sim \text{Multinomial}(\beta_z)$
      - Where each topic has a different parameter vector β for the words

LDA Parameter Estimation

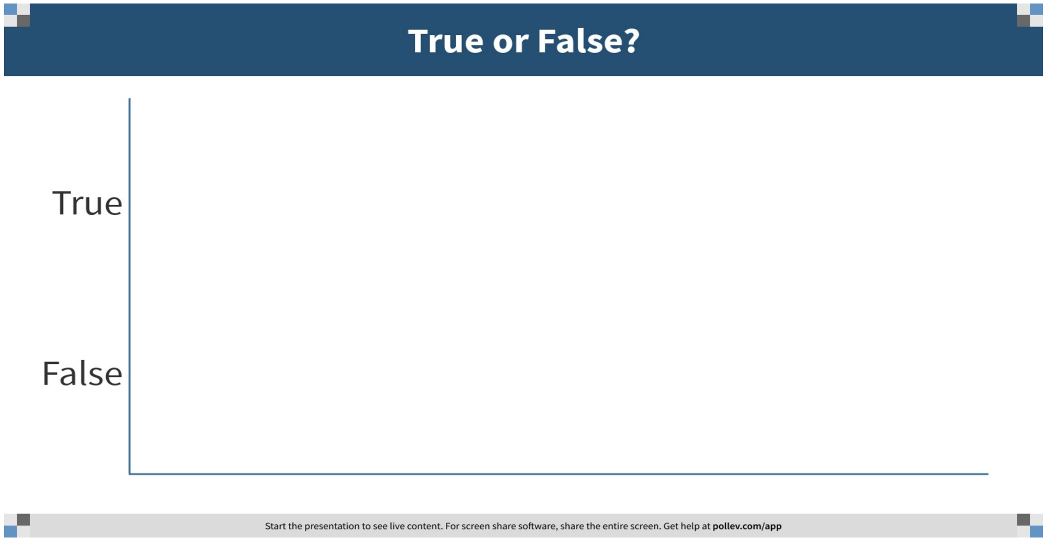
- Given a corpus of documents, find the parameters α and β which maximize the likelihood of the observed data (words in documents), marginalizing over the hidden variables θ, z
- **E-step:**
  - Compute $p(\boldsymbol{\theta},\mathbf{z}|\mathbf{w},\alpha,\beta)$, the posterior of the hidden variables $(\theta,\mathbf{z})$ given each document **w**, and parameters α and β.
- **M-step**
  - Estimate parameters α and β given the current hidden variable distribution estimates

θ: topic distribution for the document,
z: topic for each word in the document

You don't need to know the details;
Only what is hidden and what is observed;
And that EM works here.

LDA: True or False?

In LDA, the words in each document are assumed to be drawn from a Dirichlet distribution.
These distributions can vary across documents.



Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Recall: Bayes Nets

- Local Markov Assumption
- Active Trails
- D Separation

Active Trails

A trail $\{X_1, X_2, \cdots, X_k\}$ in the graph (no cycles) is an **active trail** if for each consecutive triplet in the trail:

- $X_{i-1} \rightarrow X_i \rightarrow X_{i+1}$, and X_i is not observed
  $X_{i-1} \leftarrow X_i \leftarrow X_{i+1}$, and X_i is not observed
  $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$, and X_i is not observed
  $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, and X_i is observed or one of its descendants is observed

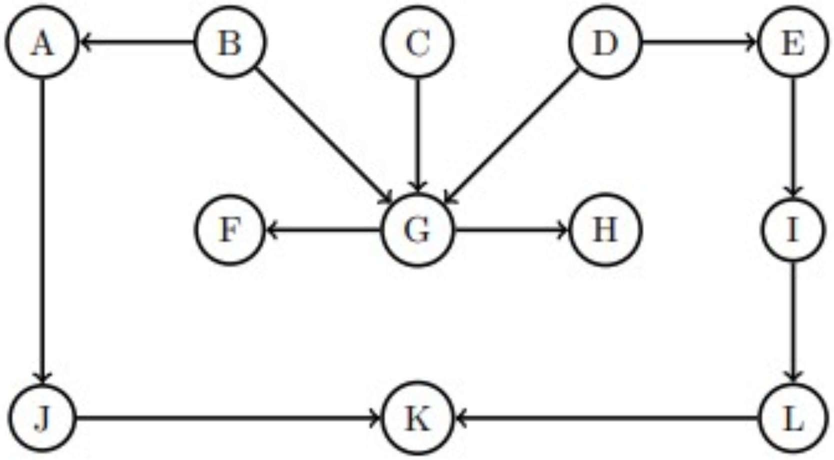
Variables connected by active trails are not conditionally independent

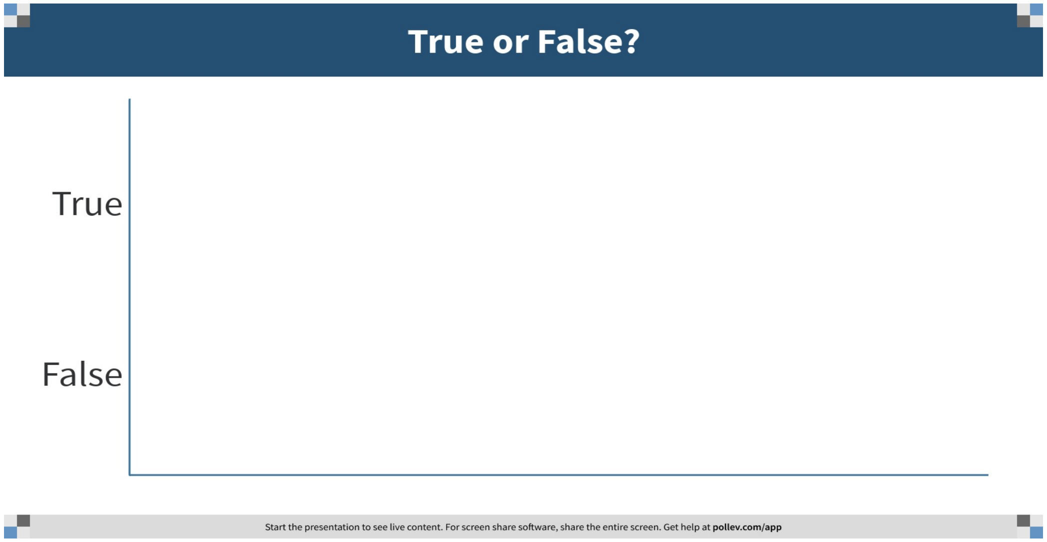
D-separation

- Variables X_i and X_j are independent if there is no *active trail* between X_i and X_j .
  - given a set of observed variables $O \subset \{X_1,\cdots,X_m\}$
  - O sometimes called a “Markov Blanket”

Bayes Net Exercises

I D-separates E and L ?

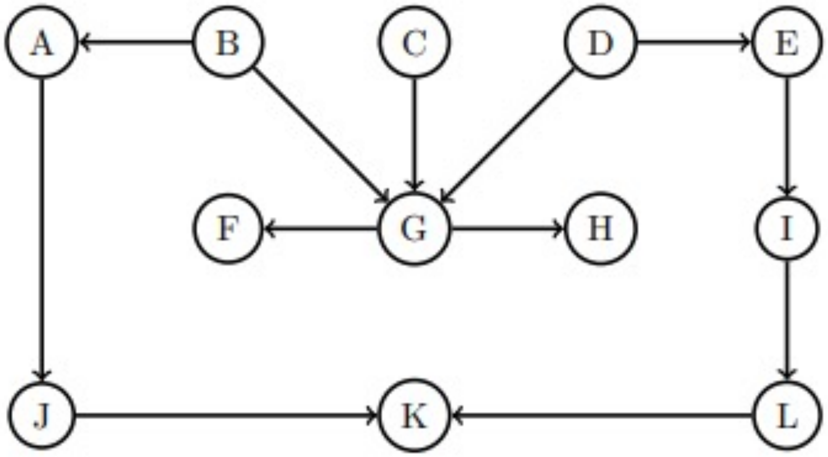


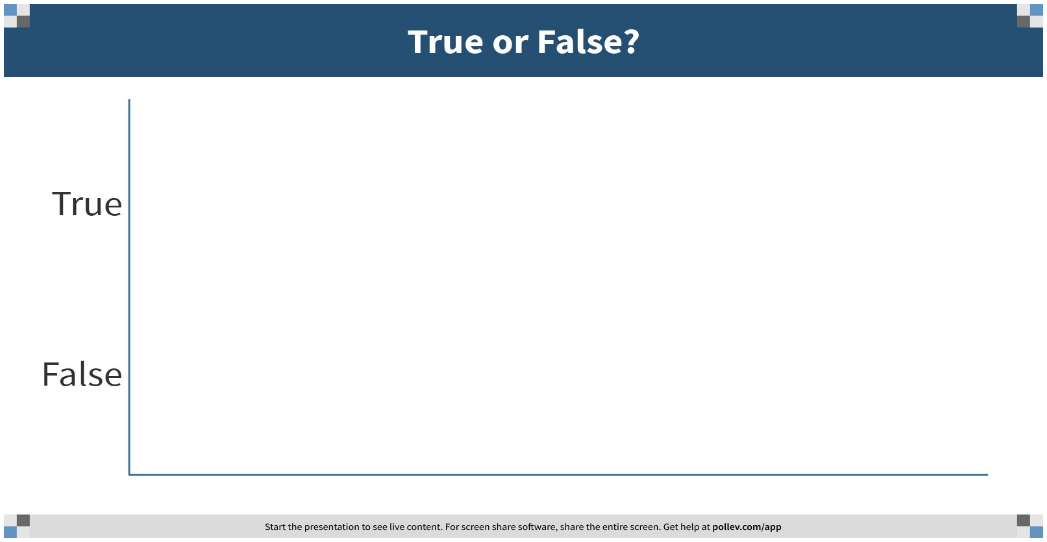


Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Bayes Net Exercises

$C \perp D \mid F$?

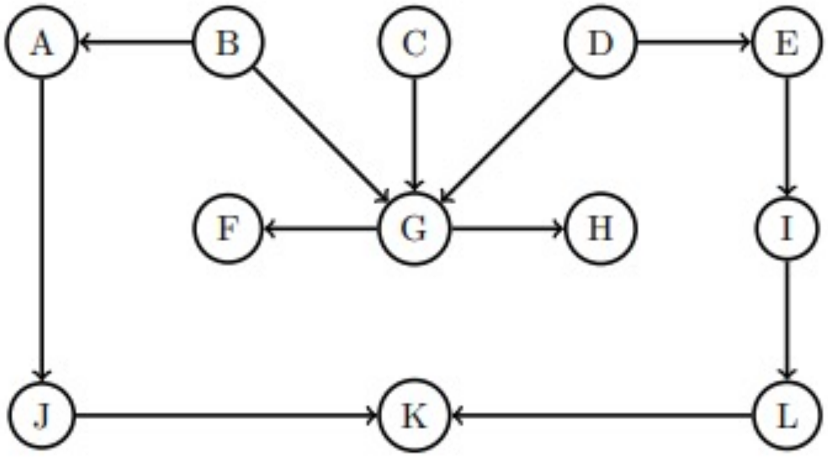


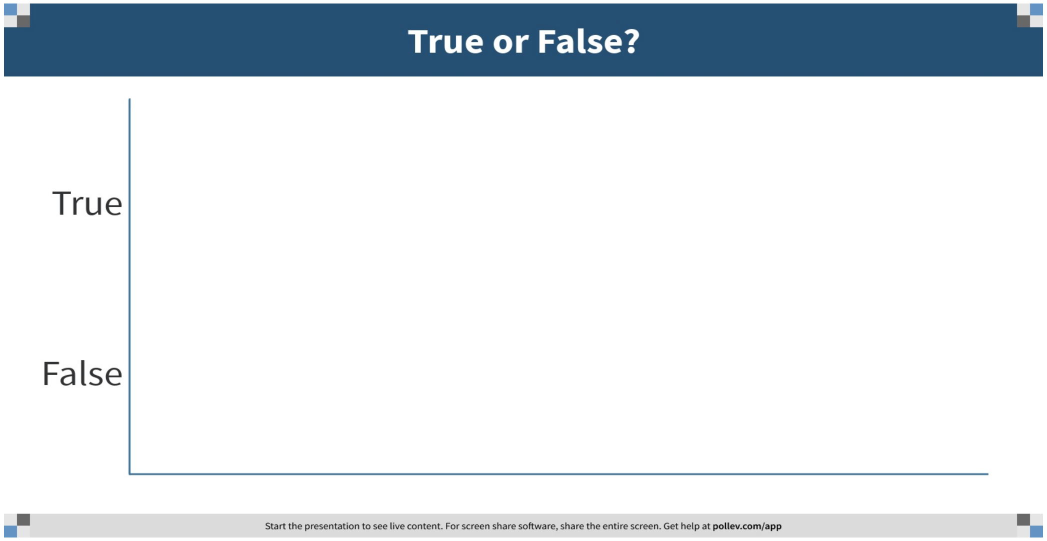


Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Bayes Net Exercises

D ⊥ I | E, F, K ?




True or False?

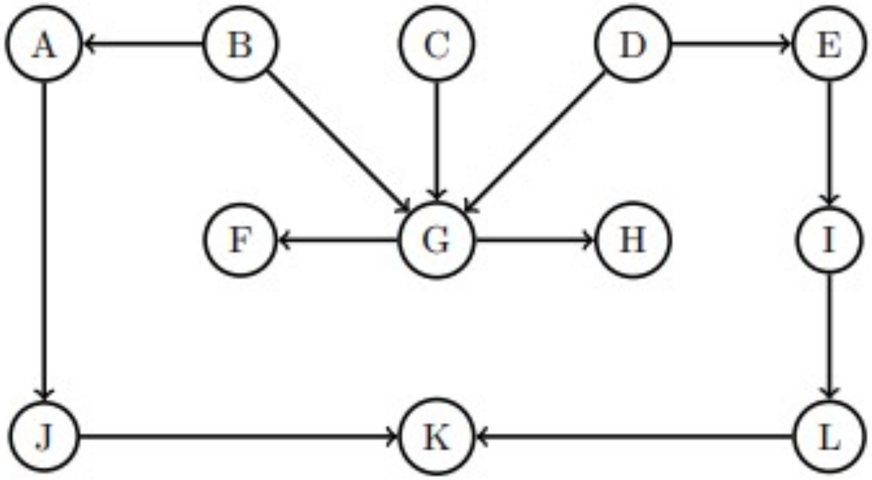
True

False

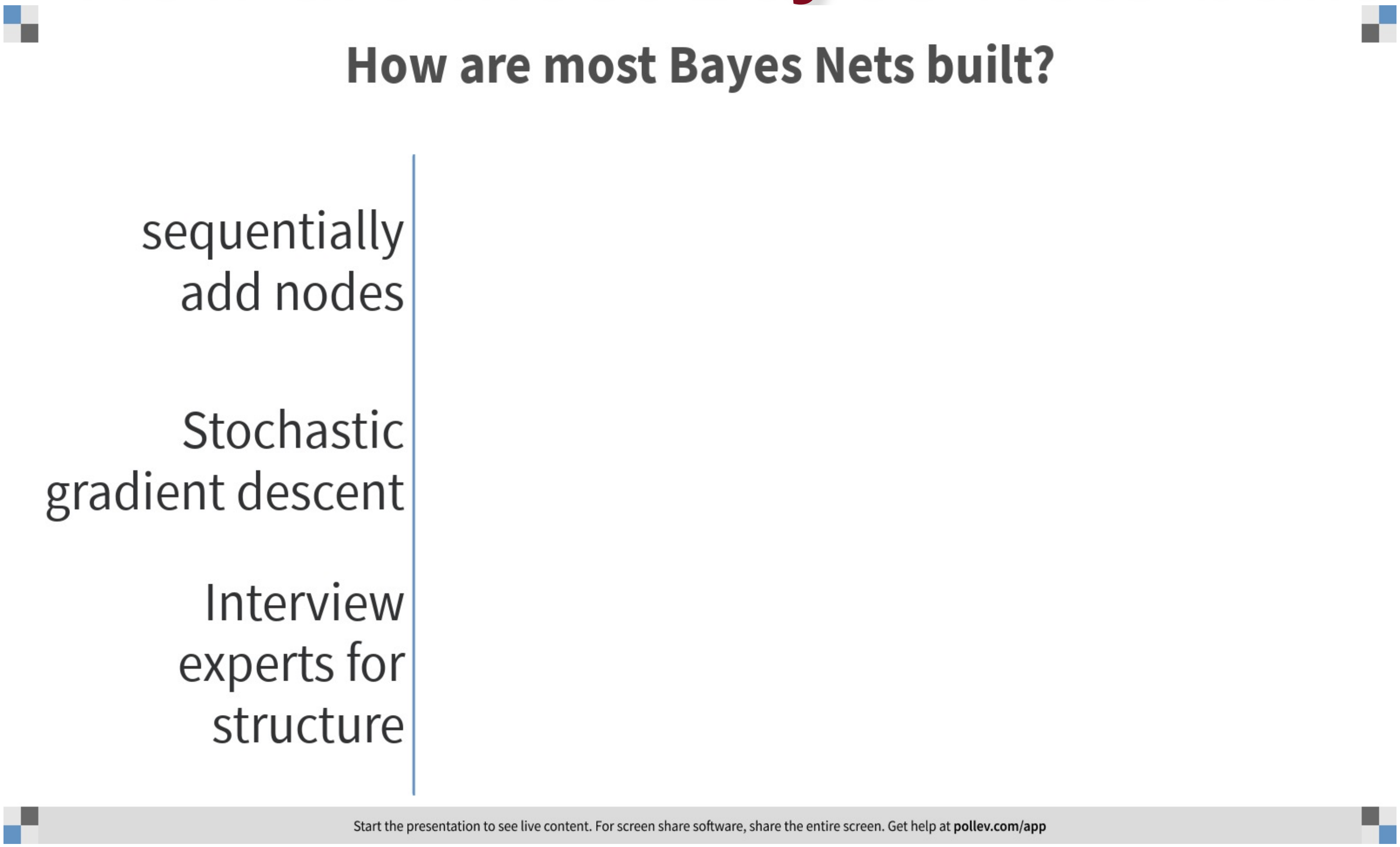
Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Bayes Net Exercises

What is the minimum number of parameters needed to represent the full joint probability P(A, B, C, D, E, F, G, H, I, J, K, L) in the above network if all the variables are binary?



How are most Bayes Nets built?



How are most Bayes Nets built?

sequentially add nodes

Stochastic gradient descent

Interview experts for structure

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Recall: Hidden Markov Models

- Markov assumption
- Model form and parameters
- Unrolling the model

Parameters of an HMM

- ***States***: A set of states $S = s_1, \ldots, s_k$
- ***Markov transition probabilities***: $A = a_{1,1}, a_{1,2}, \ldots, a_{k,k}$ Each $a_{i,j} = p(s_j \mid s_i)$ represents the probability of transitioning from state s_i to s_j.
- ***Emission probabilities***: A set B of functions of the form $b_i(o_t) = p(o|s_i)$ giving the probability of observation o_t being emitted by s_i
- ***Initial state distribution***: the probability π_i that s_i is a start state

Markov Model Example

$S_1 = [0, 1]$

	Tomorrow’s Weather		
Today’s Weather		Sunny	Rainy
	Sunny	0.8	0.2
	Rainy	0.6	0.4

Markov Transition Matrix A

What is the expected value of s_1?

What is the expected value of $s_{1,000,000}$?

Steady state at [0.75, 0.25]
(first eigenvector, with eigenvalue of 1)

Hidden Markov Model Example

S_1 = [0.5, 0.5]

We observe:
(umbrella, no umbrella)

We can ask questions like:

- What is the joint probability of the states (rain, sun) and our observations?

	Tomorrow's Weather		
Today's Weather		Sunny	Rainy
	Sunny	0.8	0.2
	Rainy	0.6	0.4

Markov Transition Matrix A

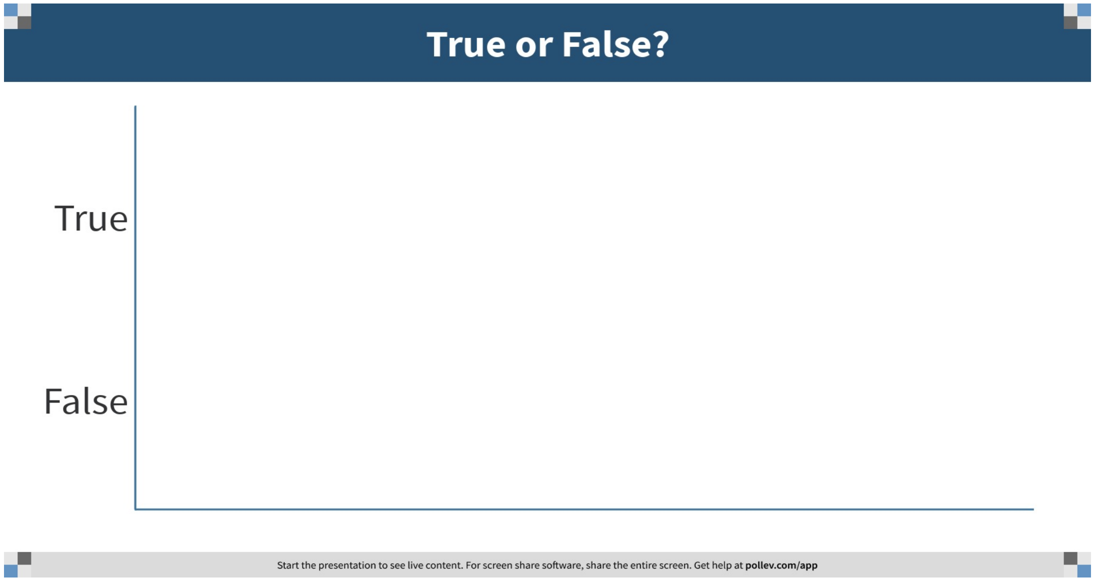
Weather		
	Sunny	Rainy
Umbrella	0.1	0.8
No Umbrella	0.9	0.2

Emission Probabilities B

HMM Exercise

True or False? The following statement about hidden Markov models holds for all 1 ≤ t ≤ T and k

$$P(O_{t+1} = o_{t+1}, ..., O_T = o_T | O_1 = o_1, ..., O_t = o_t, S_t = k)$$
$$= P(O_{t+1} = o_{t+1}, ..., O_T = o_T | S_t = k)$$



True or False?

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**