Lecture # 13
Session 2003

A Practical Introduction to Graphical Models and their use in ASR

6.345

Graphical models for ASR

- **HMMs (and most other common ASR models) have some drawbacks**
  - **Strong independence assumptions**
  - **Single state variable per time frame**

- **May want to model more complex structure**
  - **Multiple processes (audio + video, speech + noise, multiple streams of acoustic features, articulatory features)**
  - **Dependencies between these processes or between acoustic observations**

- **Graphical models provide:**
  - **General algorithms for large class of models**

    $\Rightarrow$ No need to write new code for each new model

  - **A "language" with which to talk about statistical models**

Outline

- **First half – intro to GMs**
  - **Independence & conditional independence**
  - **Bayesian networks (BNs)**
    - Definition
    - Main problems
  - **Graphical models in general**

- **Second half – dynamic Bayesian networks (DBNs) for speech recognition**
  - **Dynamic Bayesian networks -- HMMs and beyond**
  - **Implementation of ASR decoding/training using DBNs**
  - **More complex DBNs for recognition**
  - **GMTK**

(Statistical) independence

- **Definition: Given the random variables X and Y,**

$$X \perp Y \quad \Leftrightarrow \quad p(x \mid y) = p(x)$$

$$\Updownarrow \qquad\qquad \Updownarrow$$

$$p(x, y) = p(x)p(y) \quad \Leftrightarrow \quad p(y \mid x) = p(y)$$

(Statistical) conditional independence

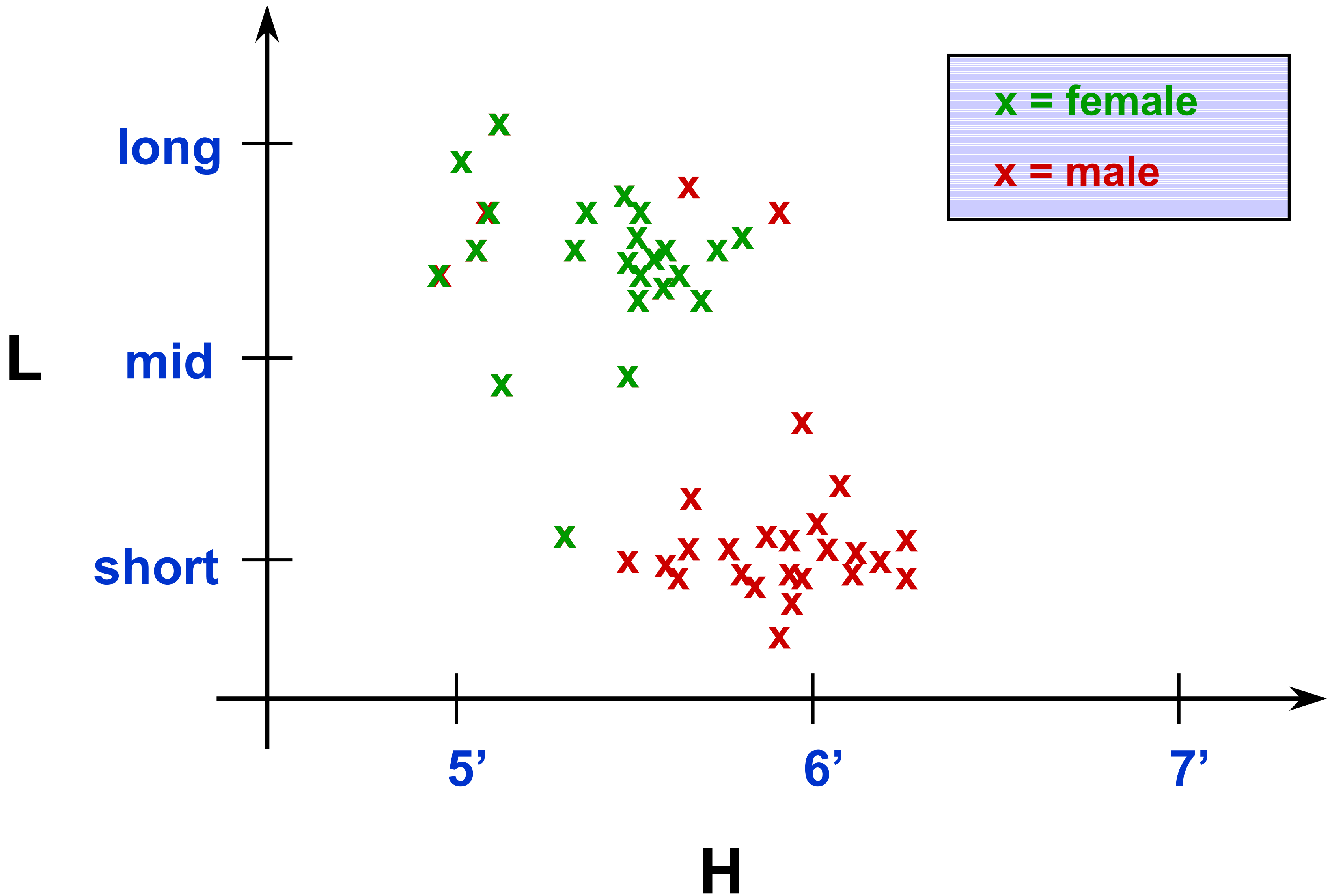
- **Definition: Given the random variables X, Y, and Z,**

$$X \perp Y \mid Z \quad \Leftrightarrow \quad p(x \mid y, z) = p(x \mid z)$$

$$\Updownarrow \qquad\qquad\qquad\qquad \Updownarrow$$

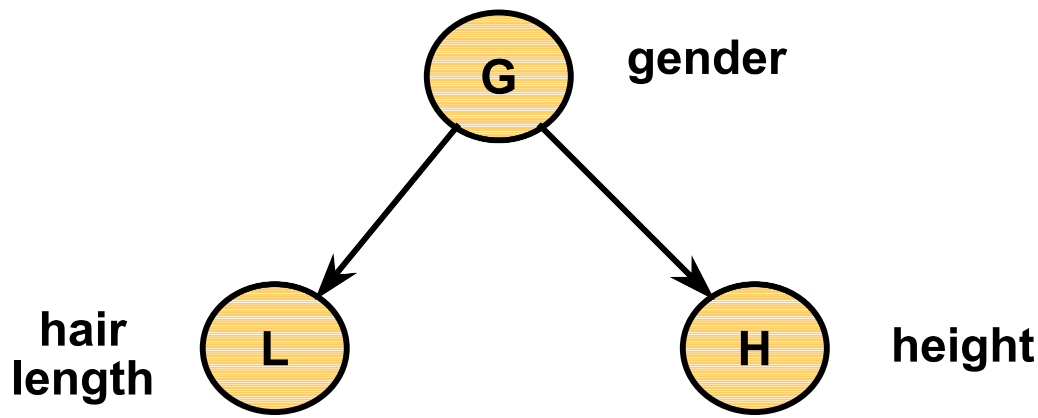
$$p(x, y \mid z) = p(x \mid z)\,p(y \mid z) \quad \Leftrightarrow \quad p(y \mid x, z) = p(y \mid z)$$

Is height independent of hair length?



Is height independent of hair length?

- Generally, no
- If gender known, yes
- This is the "common cause" scenario



$$p(h \mid l) \neq p(h) \qquad H \not\perp L$$

$$p(h \mid l, g) = p(h \mid g) \qquad H \perp L \mid G$$

Is the future independent of the past (in a Markov process)?

- Generally, no
- If present state is known, then yes

$$p(q_{i+1} \mid q_{i-1}) \neq p(q_{i+1})$$

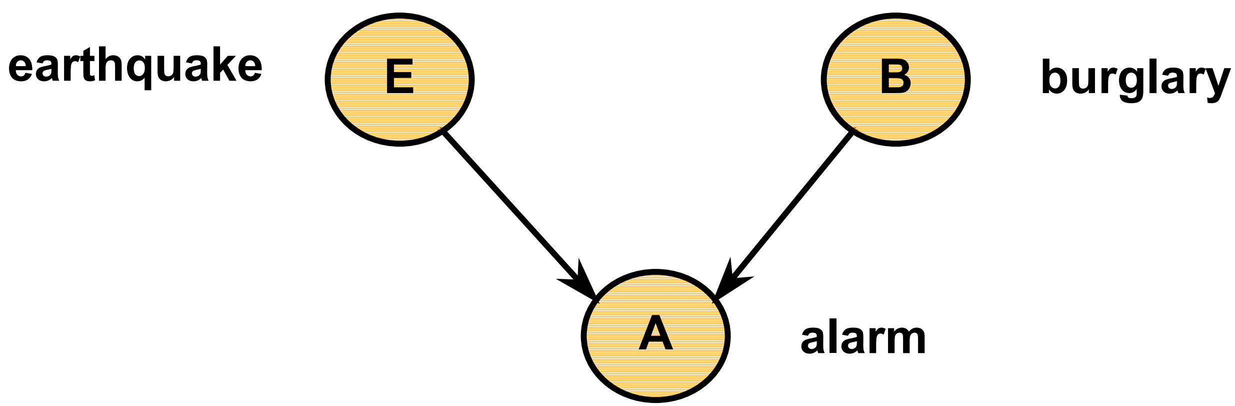
$$p(q_{i+1} \mid q_{i-1}, q_i) = p(q_{i+1} \mid q_i)$$

$$Q_{i+1} \not\perp Q_{i-1}$$

$$Q_{i+1} \perp Q_{i-1} \mid Q_i$$

Are burglaries independent of earthquakes?

- Generally, yes
- If alarm state known, no
- Explaining-away effect: the earthquake “explains away” the burglary

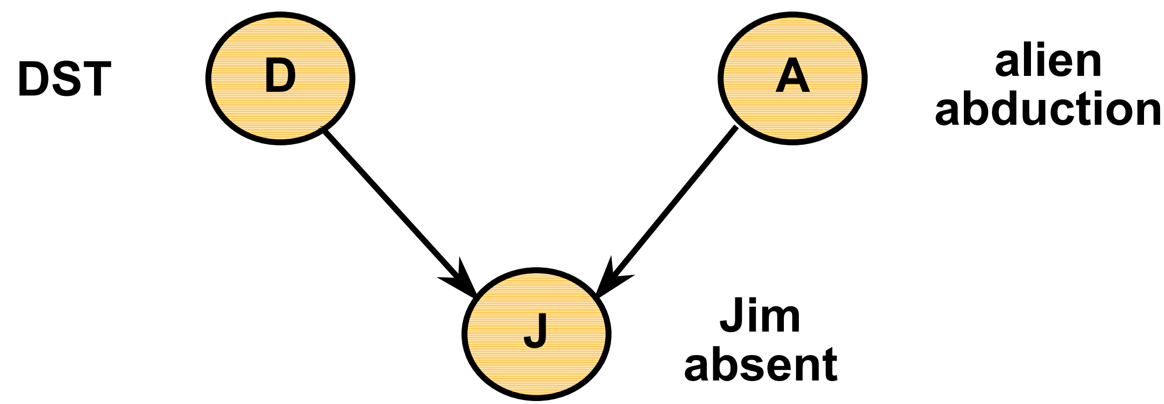


$$p(b \mid e) = p(b) \qquad E \perp B$$

$$p(b \mid e, a) \neq p(b \mid a) \qquad E \not\perp B \mid A$$

Are alien abductions independent of daylight savings time?

- Generally, yes
- If Jim doesn't show up for lecture, no
- Again, explaining-away effect



$$p(a \mid d) = p(a)$$

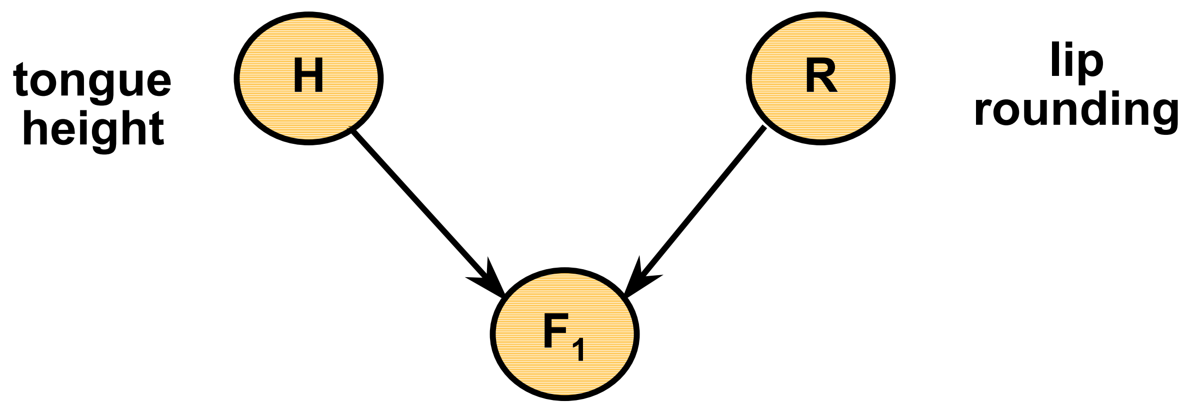
$$p(a \mid d, j) \neq p(a \mid j)$$

$$D \perp A$$

$$D \not\perp A \mid J$$

Is tongue height independent of lip rounding?

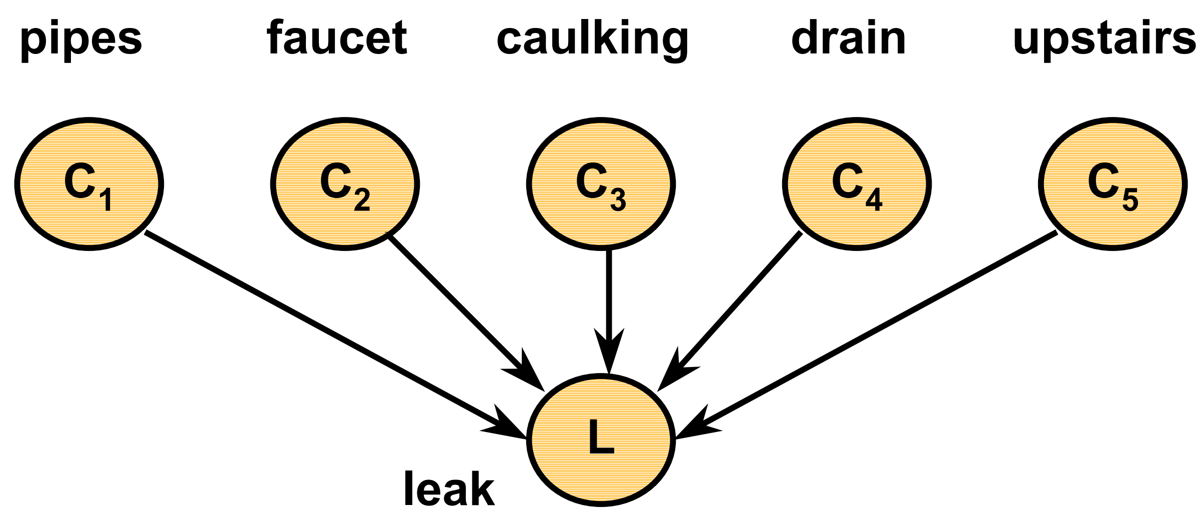
- Generally, yes
- If F_1 is known, no
- Yet again, explaining-away effect...



$$p(h \mid r) = p(h) \qquad H \perp R$$

$$p(h \mid r, f_1) \neq p(h \mid f_1) \qquad H \not\perp R \mid F_1$$

More explaining away...



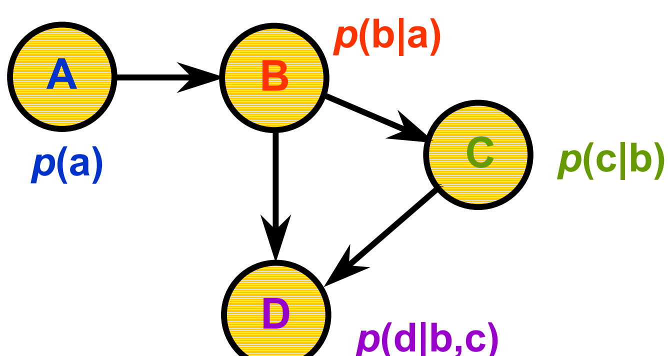
$$p(c_i \mid c_j) = p(c_i) \qquad C_i \perp C_j \;\; \forall i, j$$

$$p(c_i \mid c_j, l) \neq p(c_i \mid l) \qquad C_i \not\perp C_j \mid L \;\; \forall i, j$$

Bayesian networks

- **The preceding slides are examples of simple Bayesian networks**

- **Definition:**
  - **Directed acyclic graph (DAG) with a one-to-one correspondence between nodes (vertices) and variables $X_1, X_2, \ldots, X_N$**
  - **Each node X_i with parents $pa(X_i)$ is associated with the "local" probability function $p_{X_i|pa(X_i)}$**
  - **The joint probability of all of the variables is given by the product of the local probabilities, i.e. $p(x_i, \ldots, x_N) = \prod p(x_i|pa(x_i))$**



$\Rightarrow$ $p(a,b,c,d) = p(a)\, p(b|a)\, p(c|b)\, p(d|b,c)$

- **A given BN represents a *family* of probability distributions**

Bayesian networks, cont'd

- **Missing edges in the graph correspond to independence assumptions**

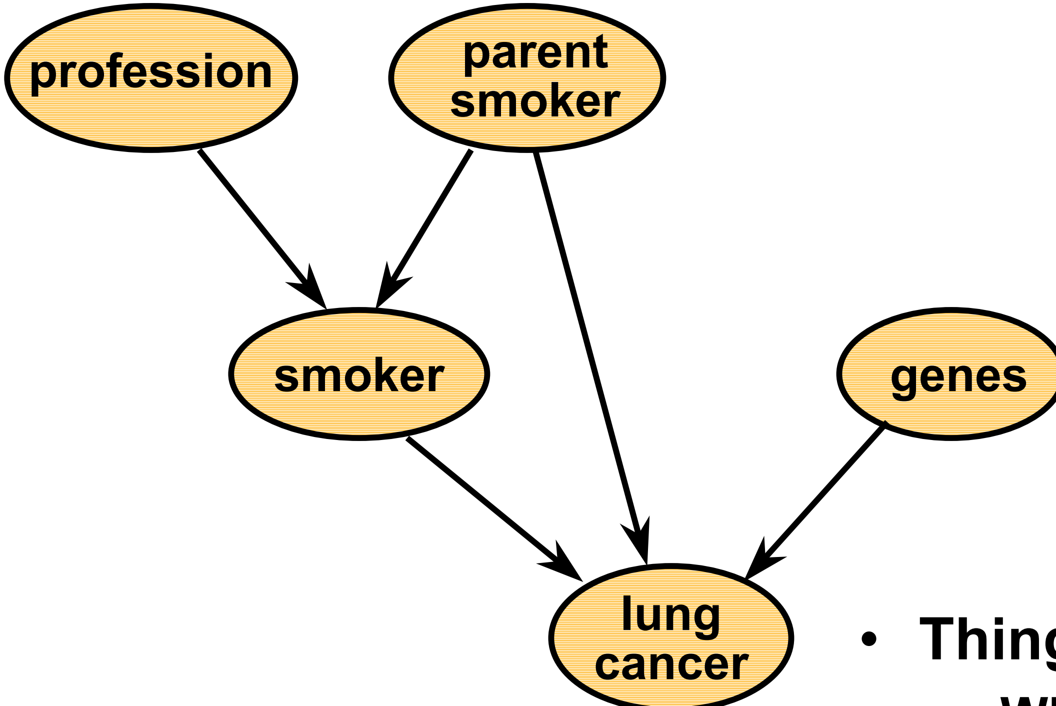
- **Joint probability can always be factored according to the chain rule:**

$$p(a,b,c,d) = p(a)\, p(b|a)\, p(c|a,b)\, p(d|a,b,c)$$

- **But by making some independence assumptions, we get a *sparse* factorization, i.e. one with fewer parameters**

$$p(a,b,c,d) = p(a)\, p(b|a)\, p(c|b)\, p(d|b,c)$$

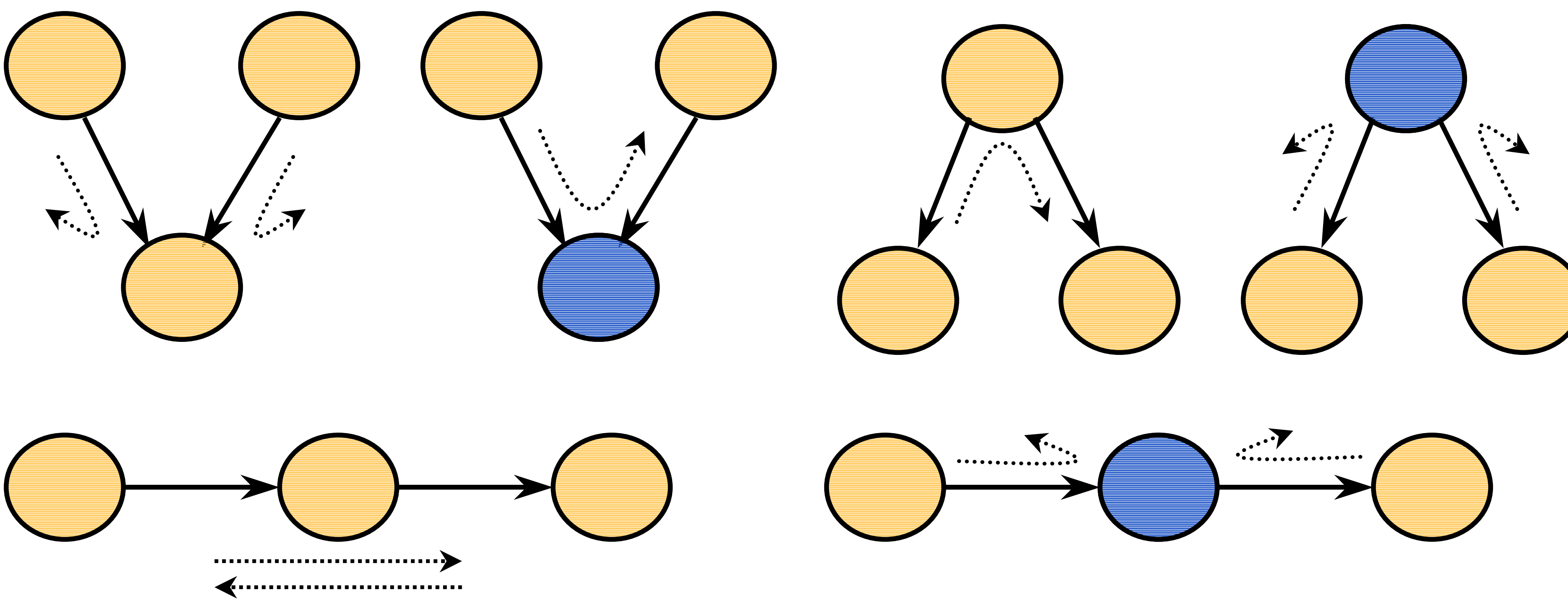
Medical example



- **Things we may want to know:**
  - **What independence assumptions does this model encode?**
  - **What is *p*(*lung cancer* | *profession*) ? *p*(*smoker* | *parent smoker, genes*) ?**
  - **Given some of the variables, what are the most likely values of others?**
  - **How do we estimate the local probabilities from data?**

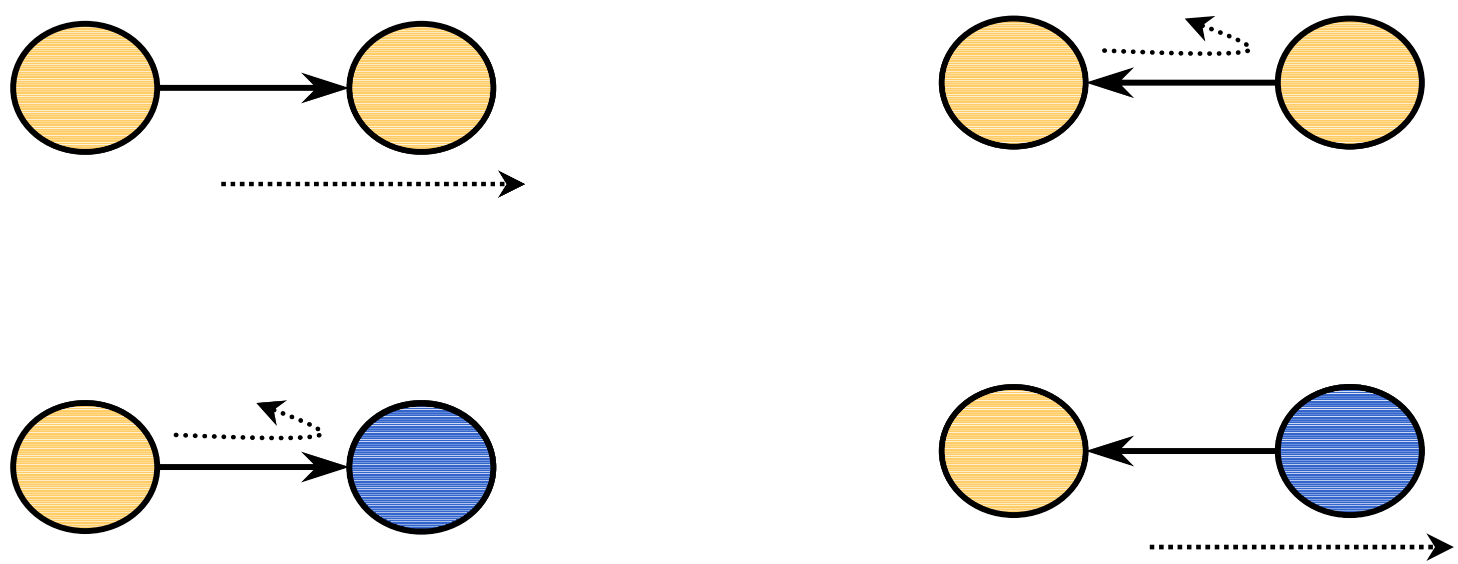
Determining independencies from a graph

- There are several ways...
- Bayes-ball algorithm ("Bayes-Ball: The Rational Pastime", Schachter 1998)
  - Ball bouncing around graph according to a set of rules
  - Two nodes are independent given a set of observed nodes if a ball can't get from one to the other

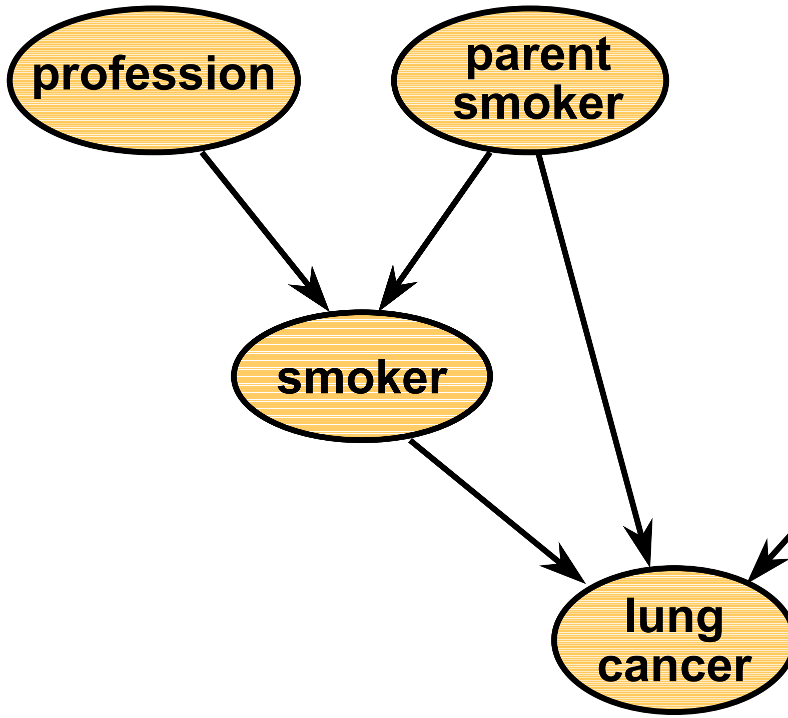


Bayes-ball, cont'd

- **Boundary conditions:**



Bayes-ball in medical example



- **According to this model:**
  - Are a person's genes independent of whether they have a parent who smokes? What about if we know the person has lung cancer?
  - Is lung cancer independent of profession given that the person is a smoker?
  - (Do the answers make sense?)

Inference

- **Definition:**
  - **Computation of the probability of one subset of the variables given another subset**

- **Inference is a subroutine of:**
  - **Viterbi decoding**

$$q^* = \text{argmax}_q \, p(q|obs)$$

  - **Maximum-likelihood estimation of the parameters of the local probabilities**

$$\lambda^* = \text{argmax}_{\lambda} \, p(obs| \lambda)$$

Graphical models (GMs)

- **In general, GMs represent families of probability distributions via graphs**
  - **directed, e.g. Bayesian networks**
  - **undirected, e.g. Markov random fields**
  - **combination, e.g. chain graphs**

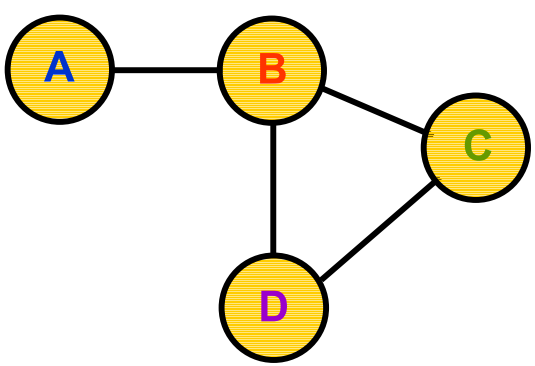
- **To describe a *particular* distribution with a GM, we need to specify:**
  - **Semantics: Bayesian network, Markov random field, ...**
  - **Structure: the graph itself**
  - **Implementation: the form of the local functions (Gaussian, table, ...)**
  - **Parameters of local functions (means, covariances, table entries...)**

- **Not all types of GMs can represent all sets of independence properties!**

Example of undirected graphical models: Markov random fields

- **Definition:**
  - Undirected graph
  - Local function ("potential") defined on each maximal clique
  - Joint probability given by normalized product of potentials

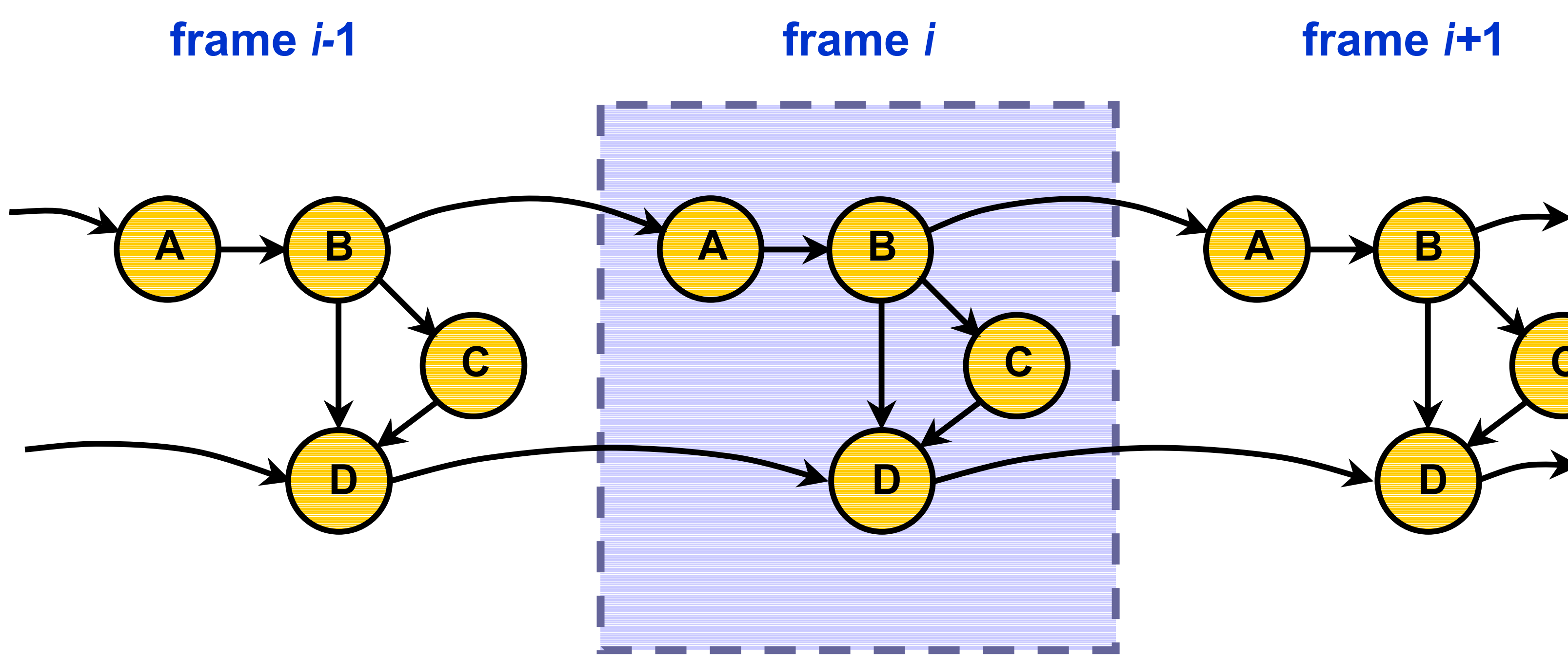
- **Independence properties can be deduced via simple graph separation**



$$p(a,b,c,d) \propto \psi_{A,B}(a,b)\psi_{B,C,D}(b,c,d)$$

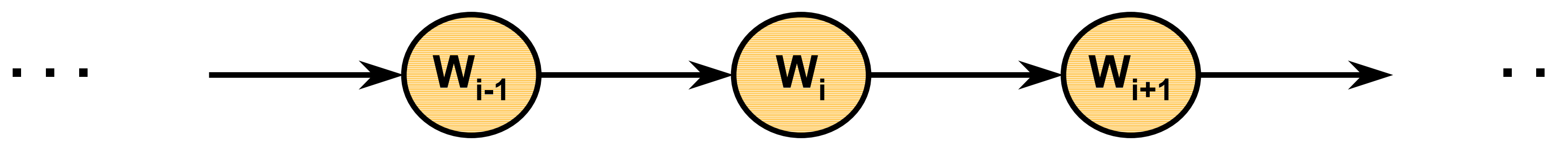
Dynamic Bayesian networks (DBNs)

- **BNs consisting of a structure that** repeats **an indefinite (or dynamic) number of times**
  - **Useful for modeling time series (e.g. speech)**

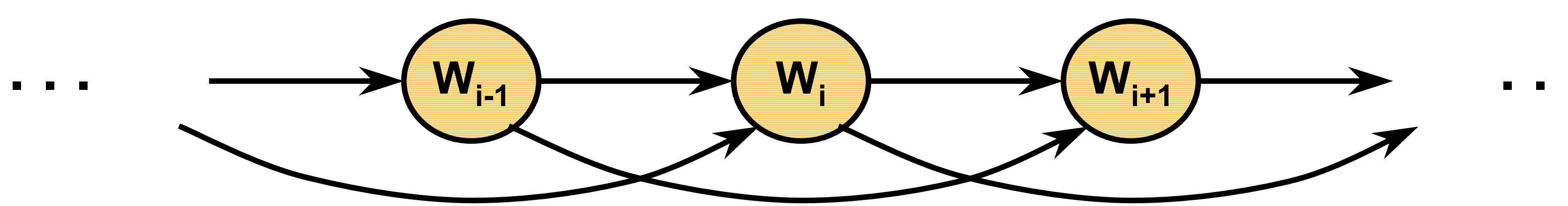


DBN representation of n-gram language models

- **Bigram:**

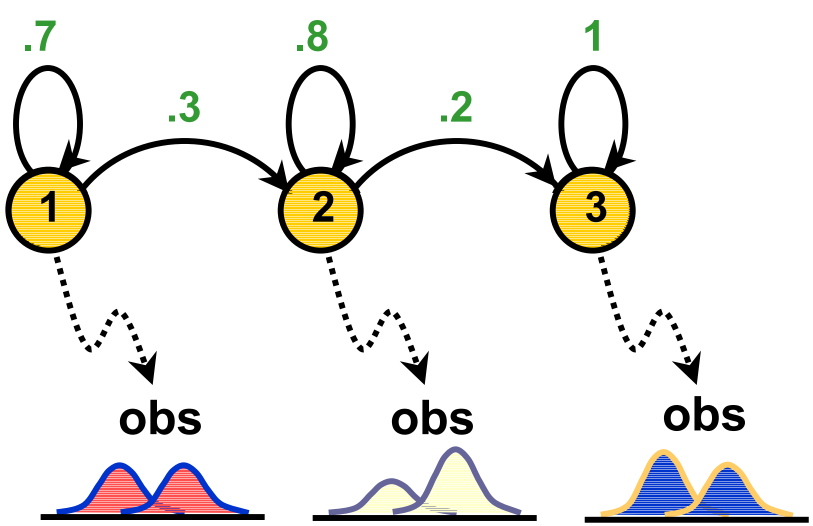


- **Trigram:**

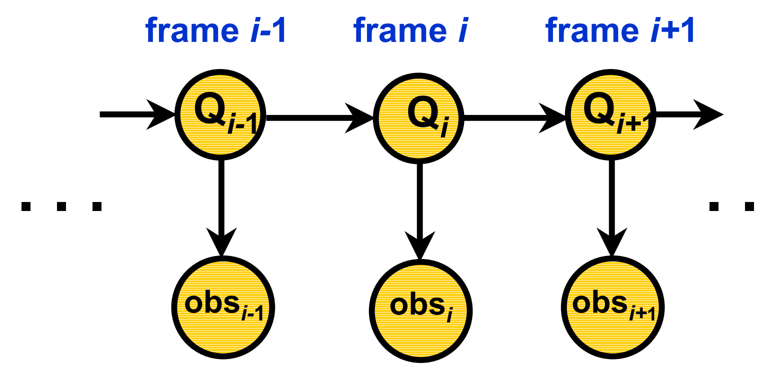


Representing an HMM as a DBN

HMM



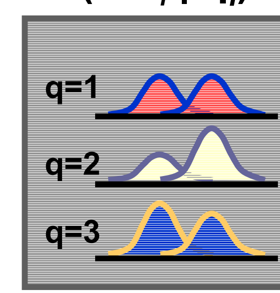
DBN



$P(q_i|q_{i-1})$

q_{i-1} \ q_i	1	2	3
1	.7	.3	0
2	0	.8	.2
3	0	0	1

$P(obs_i \mid q_i)$



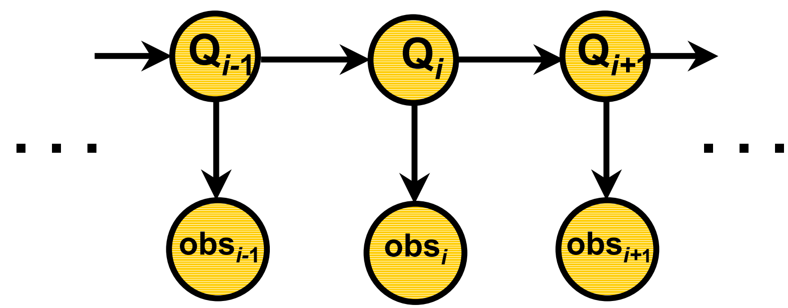
● = state

→ = allowed transition

● = variable

→ = allowed dependency

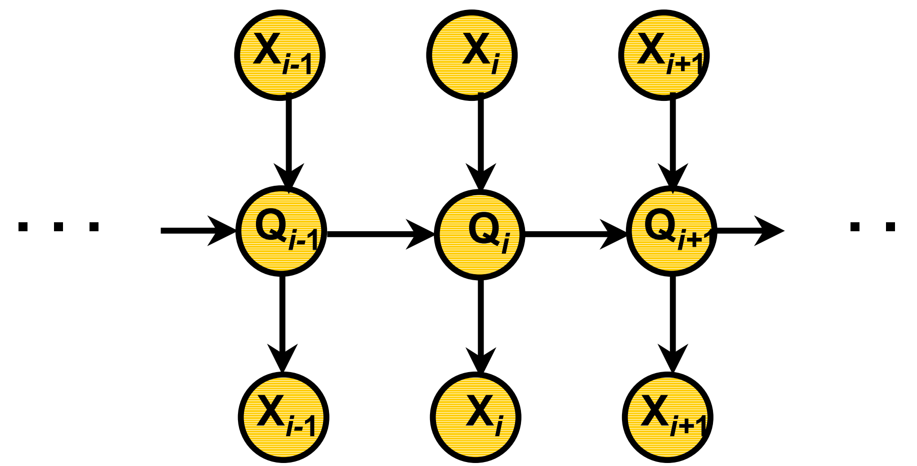
Casting HMM-based ASR as a GM problem



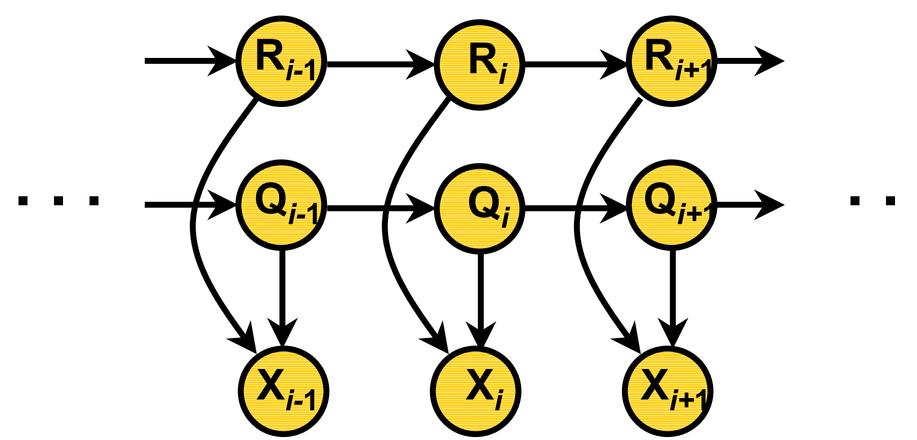
- **Viterbi decoding** ➔ **finding the most probable settings for all q_i given the acoustic observations $\{obs_i\}$**
- **Baum-Welch training** ➔ **finding the most likely settings for the parameters of $P(q_i|q_{i-1})$ and $P(obs_i \mid q_i)$**
- **Both are special cases of the standard GM algorithms for Viterbi and EM training**

Variations

- **Input-output HMMs**



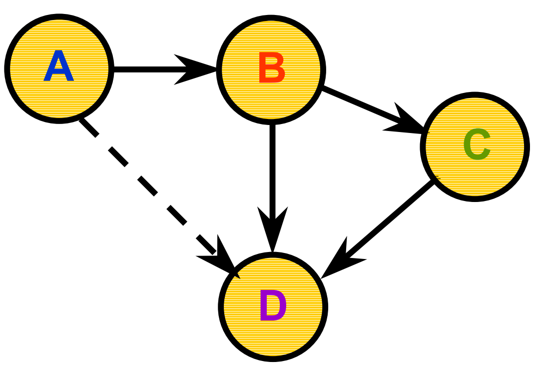
- **Factorial HMMs**



Switching parents

- **Definition:**
  - A variable *X* is a switching parent of variable *Y* if the value of *X* determines the parents and/or implementation of *Y*

- **Example:**

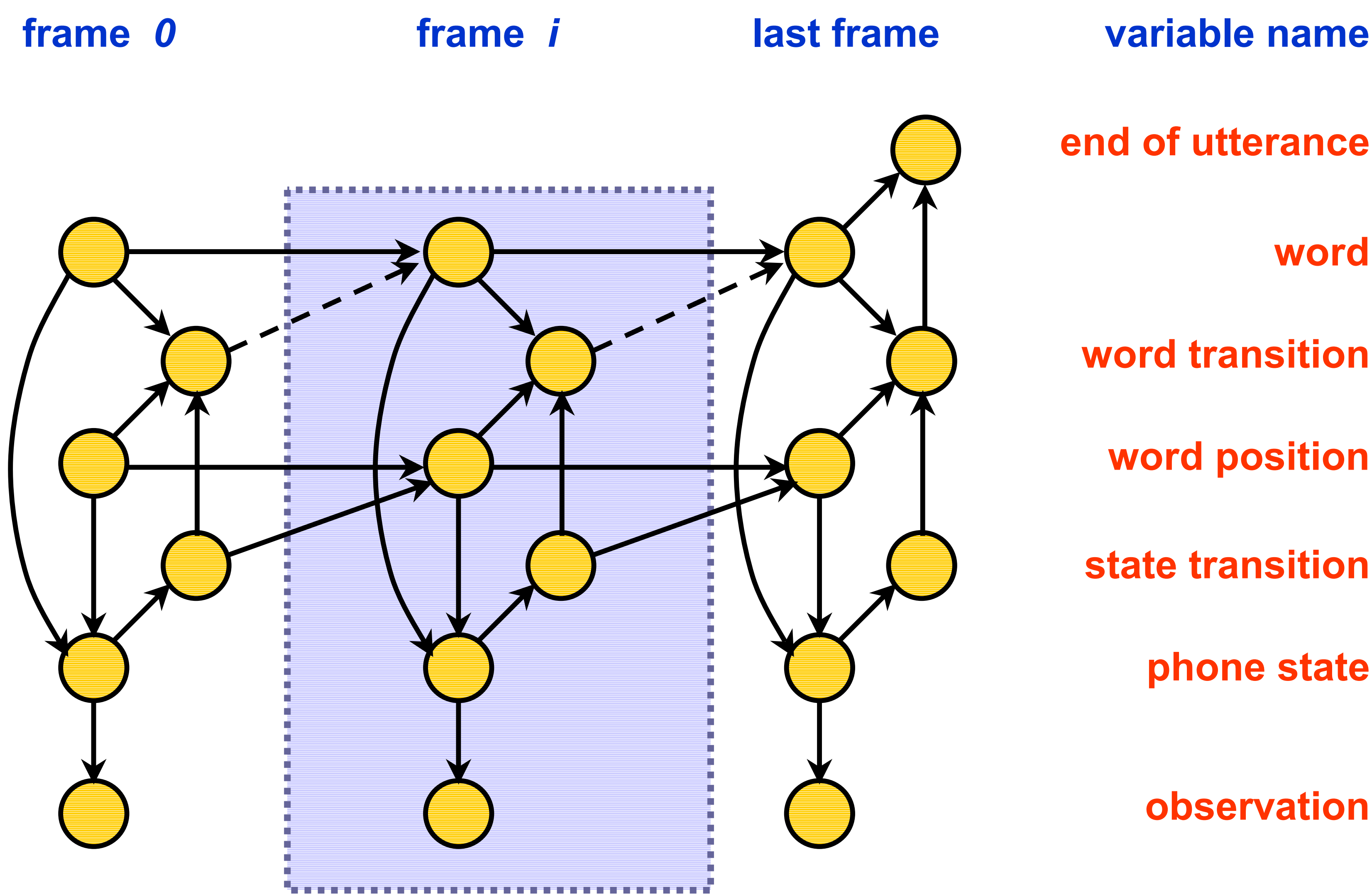


A=0 $\Rightarrow$ D has parent B with Gaussian distribution

A=1 $\Rightarrow$ D has parent C with Gaussian distribution

A=2 $\Rightarrow$ D has parent C with mixture Gaussian distribution

HMM-based recognition with a DBN



- **What language model does this GM implement?**

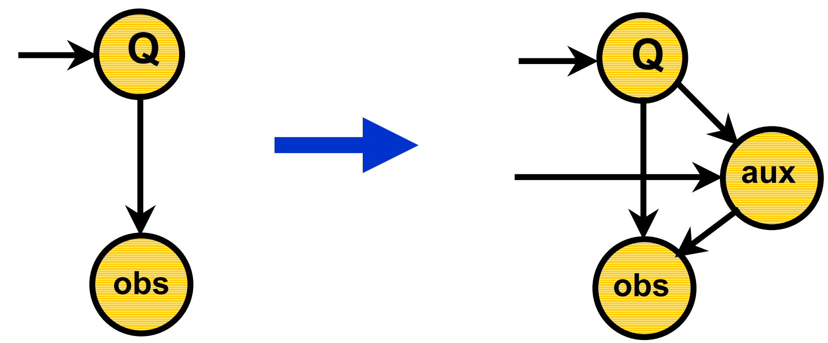
Training and testing DBNs

- **Why do we need different structures for training testing? Isn't training just the same as testing but with more of the variables observed?**

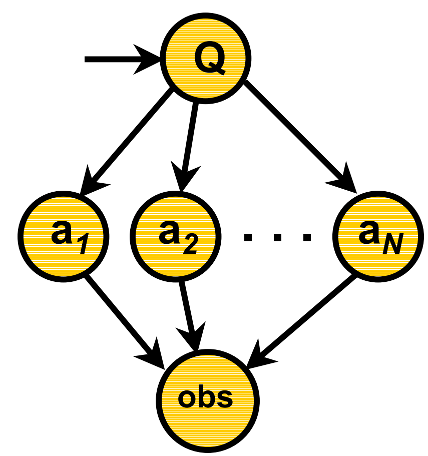
- **Not always!**
  - **Often, during training we have only *partial* information about some of the variables, e.g. the word sequence but not which frame goes with which word**

More complex GM models for recognition

- **HMM + auxiliary variables (Zweig 1998, Stephenson 2001)**
  - **Noise clustering**
  - **Speaker clustering**
  - **Dependence on pitch, speaking rate, etc.**



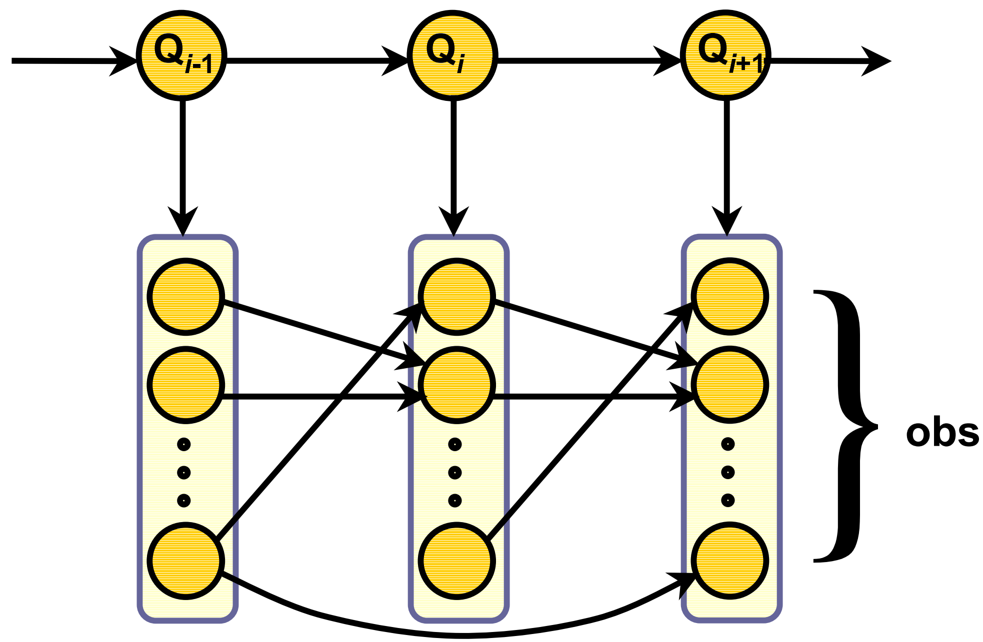
- **Articulatory/feature-based modeling**



- **Multi-rate modeling, audio-visual speech recognition (Nefian et al. 2002)**

Modeling inter-observation dependencies: Buried Markov models (Bilmes 1999)

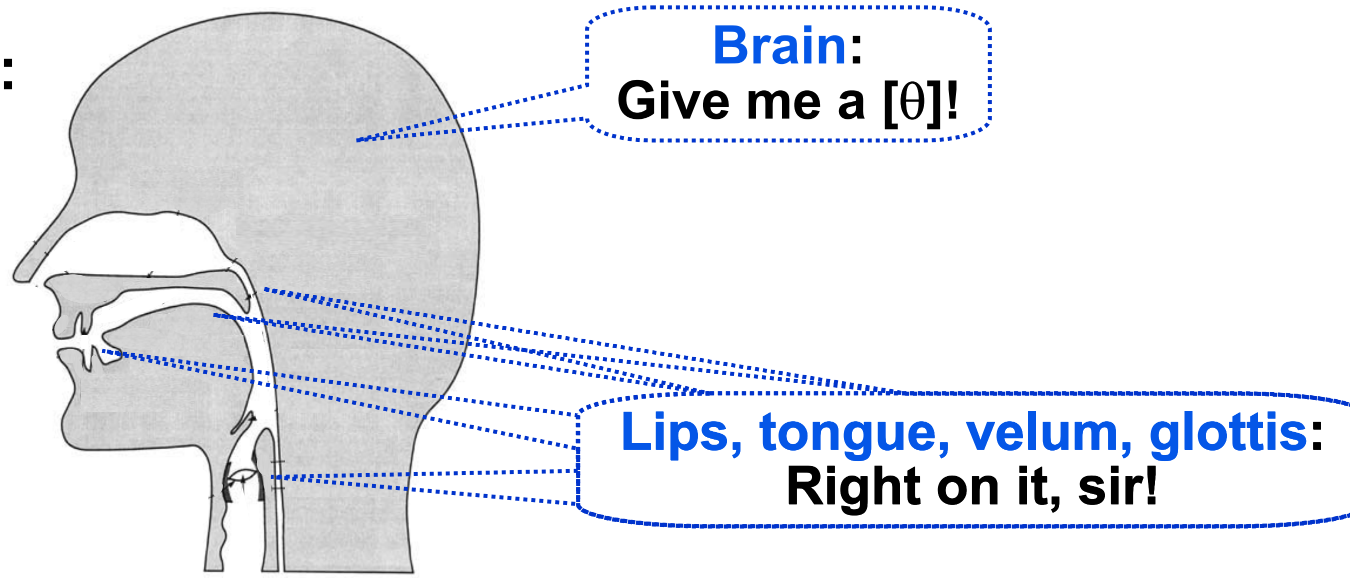
- First note that observation variable is actually a vector of acoustic observations (e.g. MFCCs)



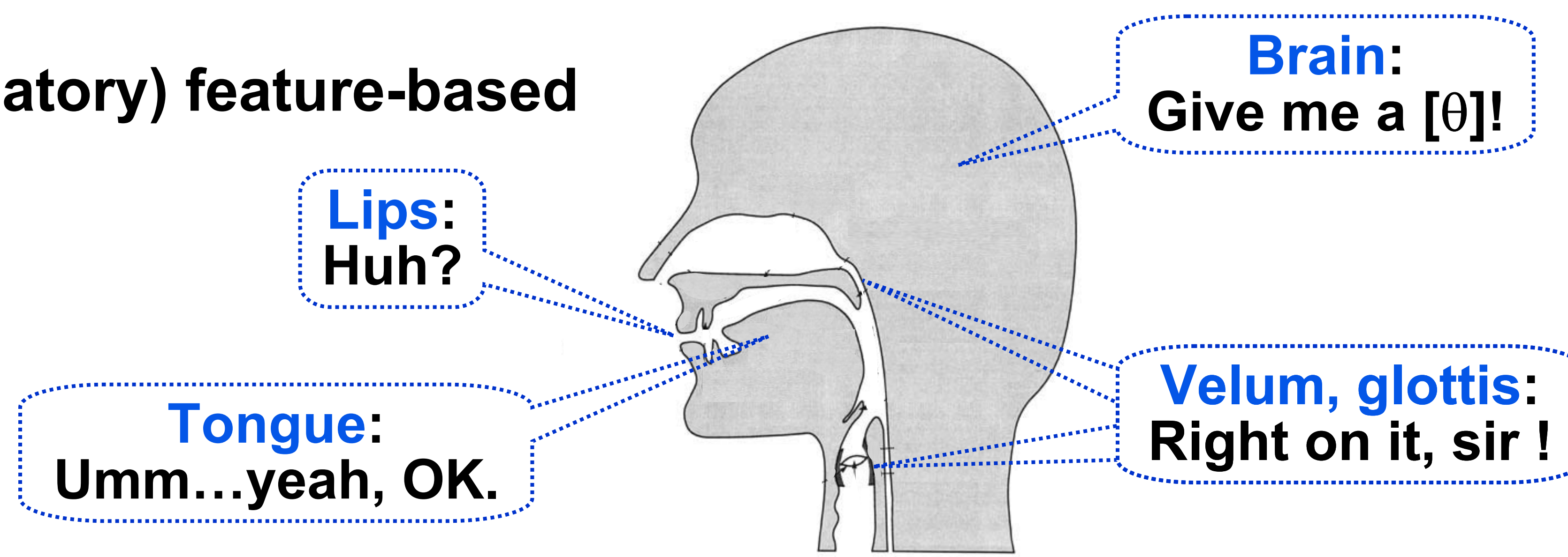
- Consider adding dependencies between observations
- Add only those that are discriminative with respect to classifying the current state/phone/word

Feature-based modeling

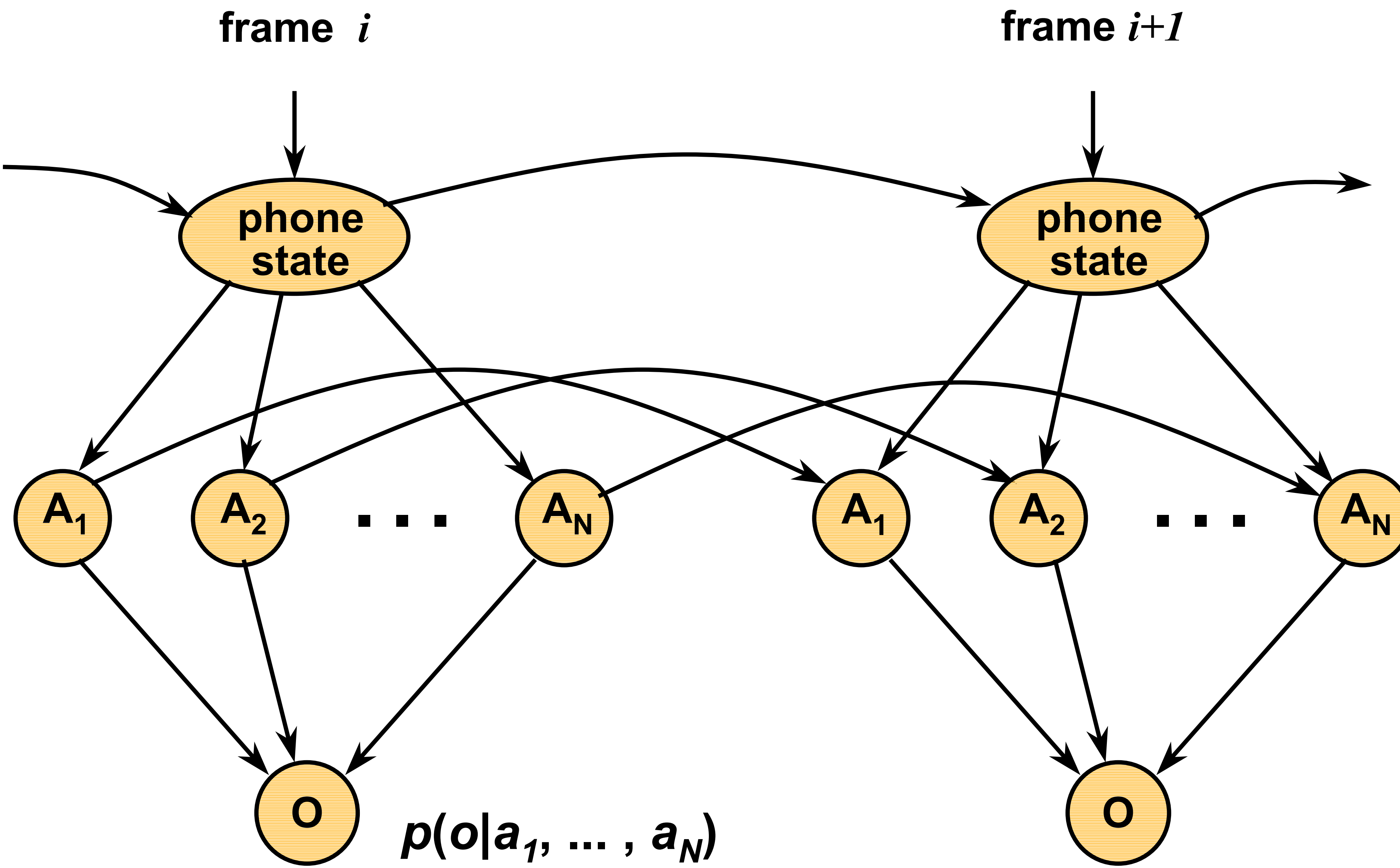
- Phone-based view:



- (Articulatory) feature-based view:



A feature-based DBN for ASR



$p(o|a_1, \ldots, a_N)$

GMTK: Graphical Modeling Toolkit (J. Bilmes and G. Zweig, ICASSP 2002)

- Toolkit for specifying and computing with dynamic Bayesian networks

- Models are specified via:
  - Structure file: defines variables, dependencies, and form of associated conditional distributions
  - Parameter files: specify parameters for each distribution in structure file

- Variable distributions can be
  - Mixture Gaussians + variants
  - Multidimensional probability tables
  - Sparse probability tables
  - Deterministic (decision trees)

- Provides programs for EM training, Viterbi decoding, and various utilities

Example portion of structure file

```
variable : phone {
    type: discrete hidden cardinality NUM_PHONES;
    switchingparents: nil;
    conditionalparents: word(0), wordPosition(0) using
         DeterministicCPT("wordWordPos2Phone");
 }


variable : obs {
    type: continuous observed OBSERVATION_RANGE;
    switchingparents: nil;
    conditionalparents: phone(0) using mixGaussian
         collection("global") mapping("phone2MixtureMapping");
 }
```

Some issues...

- For some structures, exact inference may be computationally infeasible $\Rightarrow$ approximate inference algorithms

- Structure is not always known $\Rightarrow$ structure learning algorithms

References

- J. Bilmes, "Graphical Models and Automatic Speech Recognition", in *Mathematical Foundations of Speech and Language Processing*, Institute of Mathematical Analysis Volumes in Mathematics Series, Springer-Verlag, 2003.
- G. Zweig, Speech Recognition with Dynamic Bayesian Networks, Ph.D. dissertation, UC Berkeley, 1998.
- J. Bilmes, "What HMMs Can Do", UWEETR-2002-0003, Feb. 2002.