Probabilistic Graphical Models and Bayesian Networks

Machine Learning
CS5824/ECE5424
Bert Huang
Virginia Tech

Independence

independent & identically distributed (i.i.d.)

full joint distributions

amount of dependence

cheap, easy, embarrassingly parallel

super expensive

Outline

- Probabilistic graphical models
- Bayesian networks
- Naive Bayes and Logistic Regression as Bayes nets
- Time Series Bayes Nets

Probabilistic Graphical Models

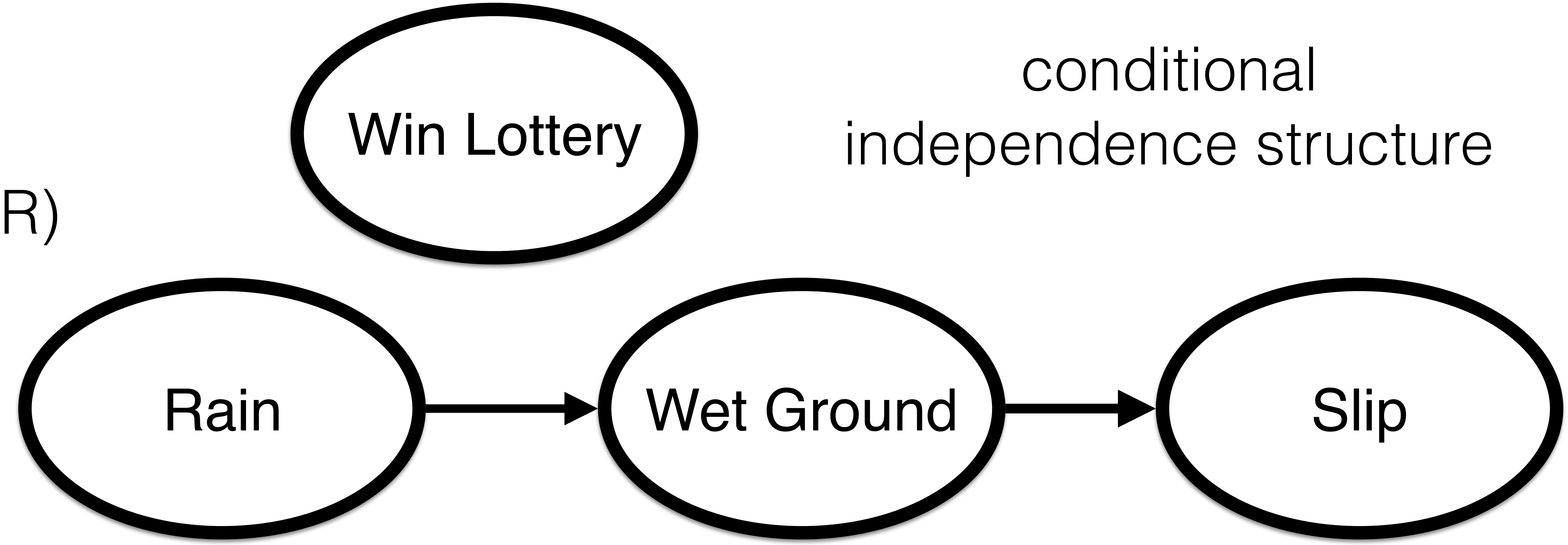
- PGMs represent probability distributions
- They encode conditional independence structure with graphs
- They enable graph algorithms for inference and learning

Bayesian Networks

$P(L, R, W)$

$= P(L)\, P(R)\, P(W \mid R)$

conditional independence structure

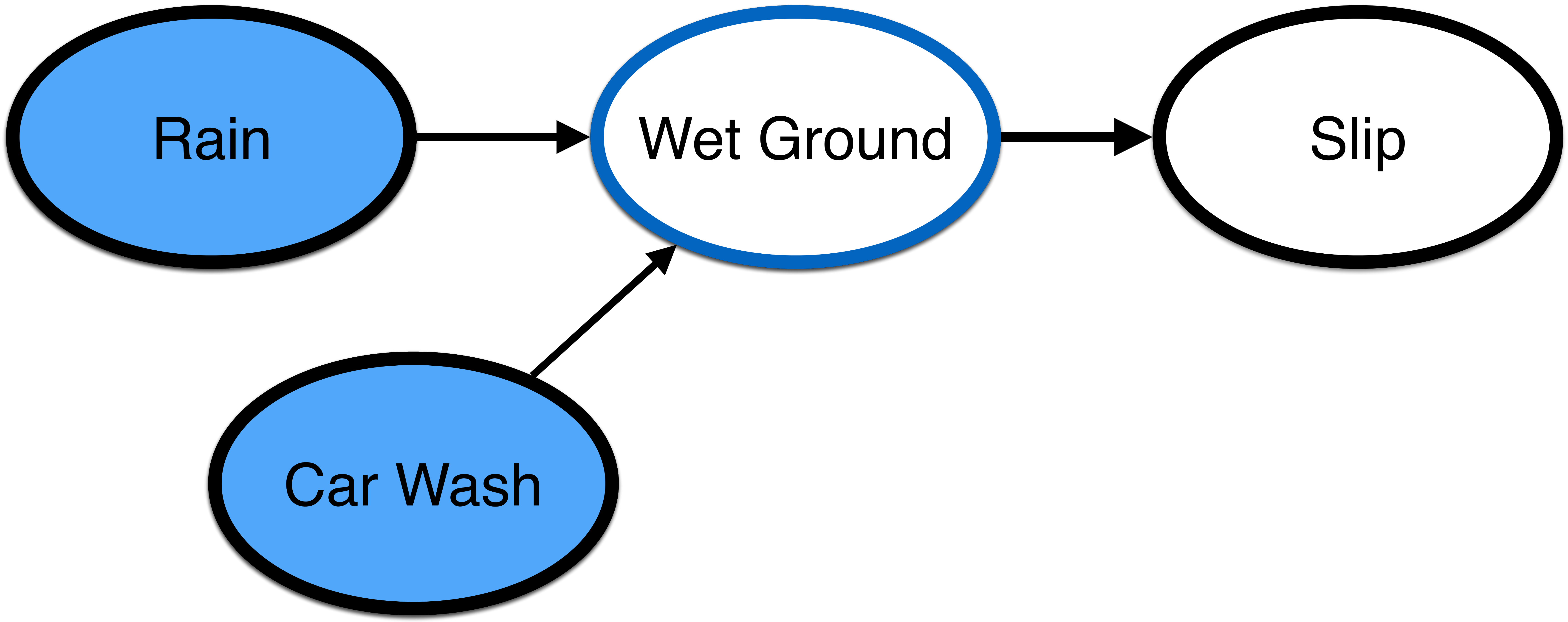


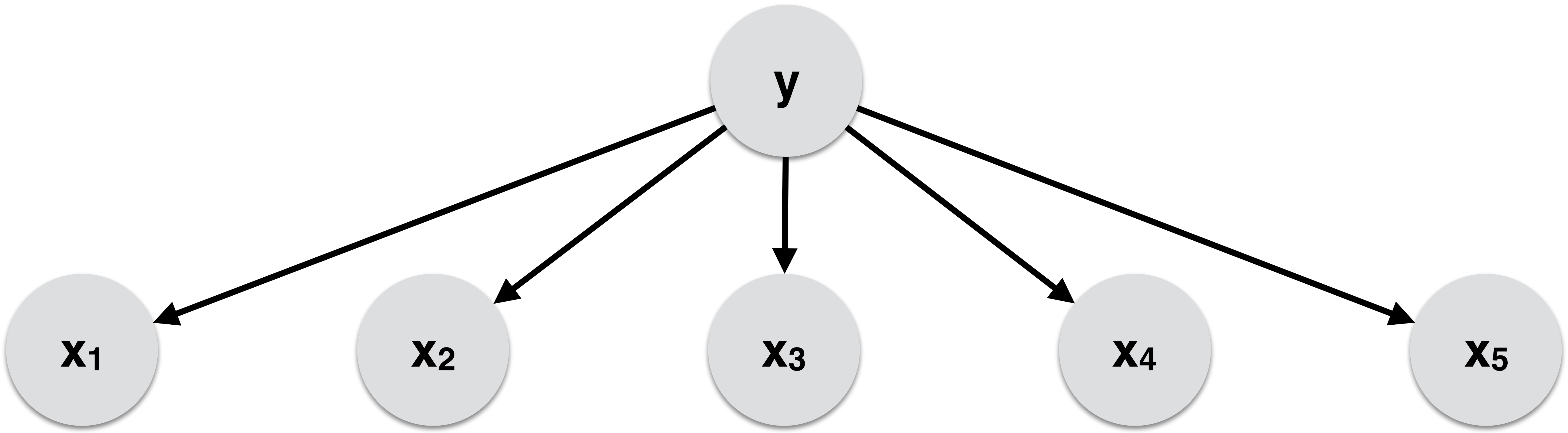
$P(L, R, W, S) = P(L)\, P(R)\, P(W \mid R)\, P(S \mid W)$

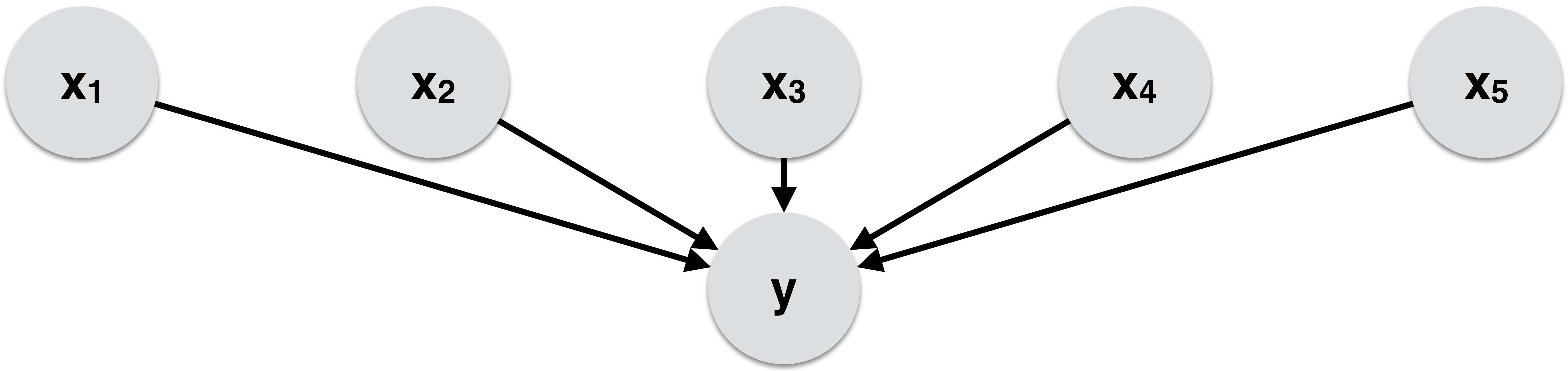
$\cancel{P(S \mid W, R)}$

Bayesian Networks

$P(R, W, S, C) = P(R)\,P(C)\,\boxed{P(W \mid C, R)}\,P(S \mid W)$ $\quad P(X \mid \text{Parents}(X))$

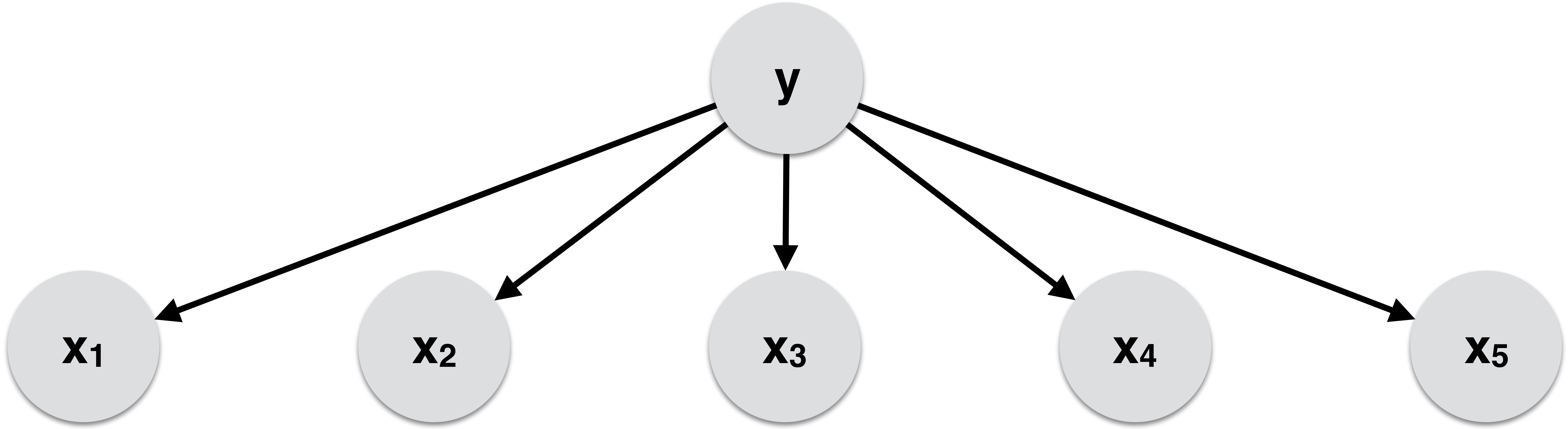


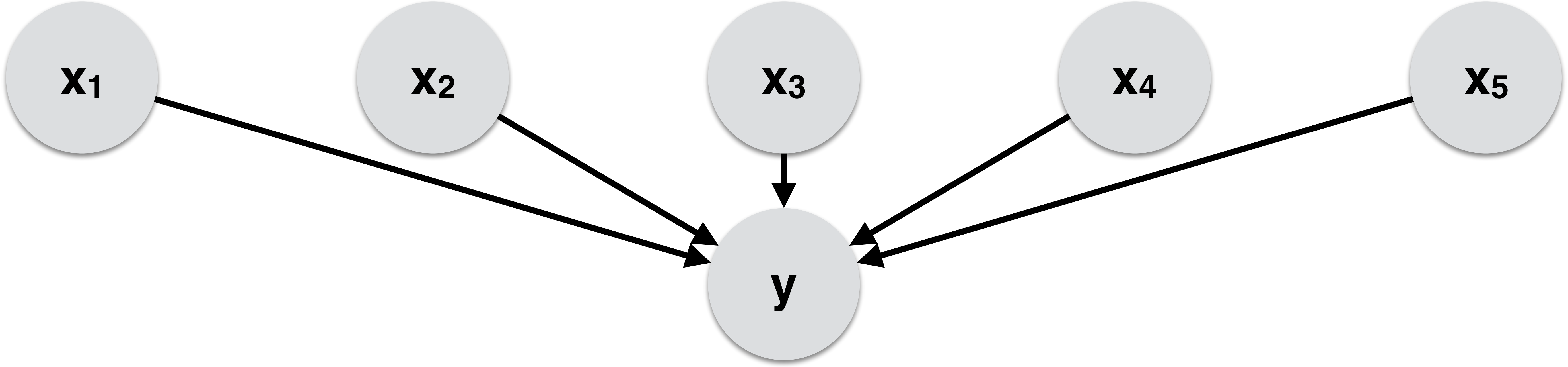

y
x1
x2
x3
x4
x5


x1
x2
x3
x4
x5
y

naive Bayes

$$p(y) \prod_{i=1}^{5} p(x_i|y)$$

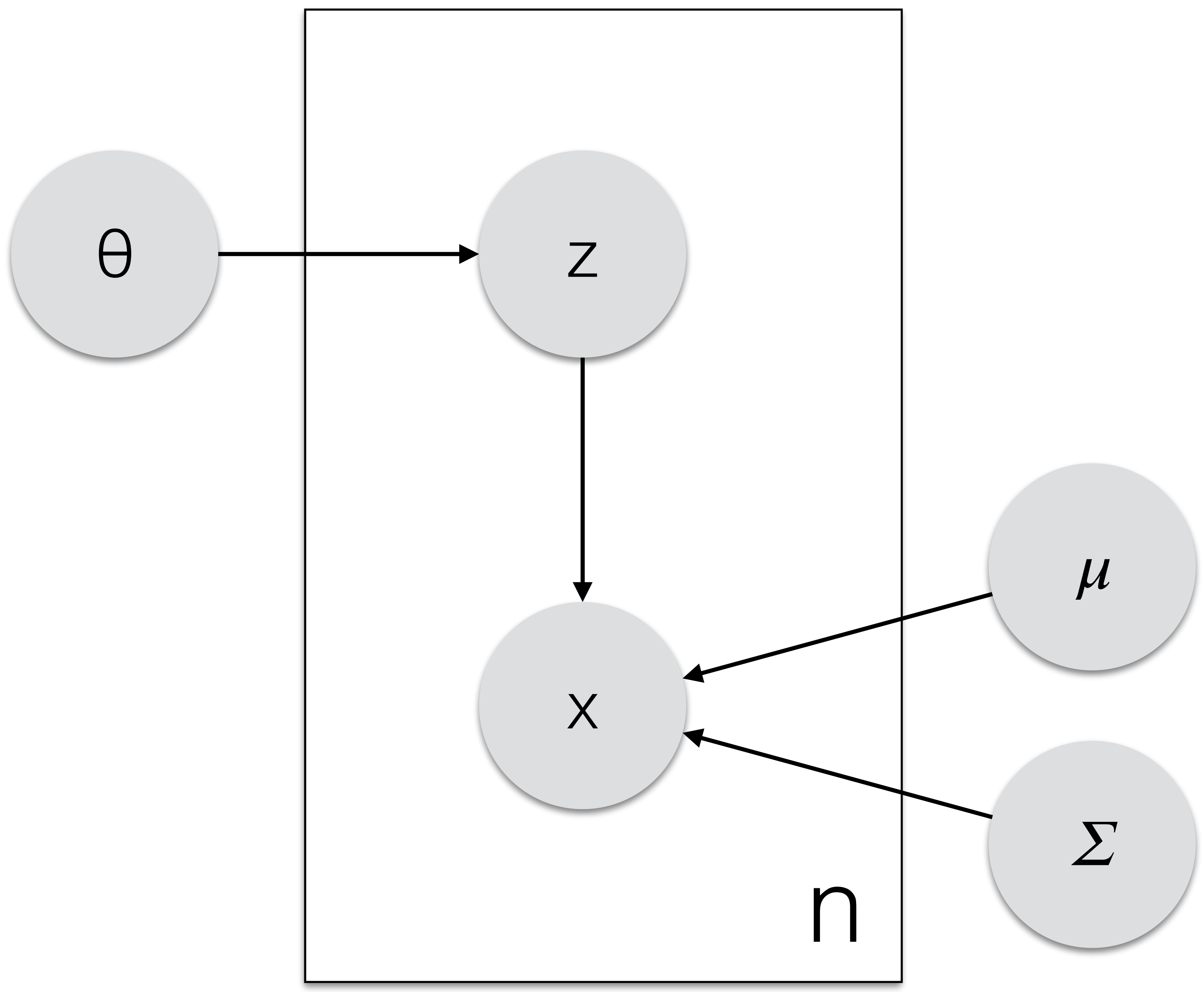




$$p(y|x_1, x_2, x_3, x_4, x_5) \prod_{i=1}^{5} p(x_i)$$

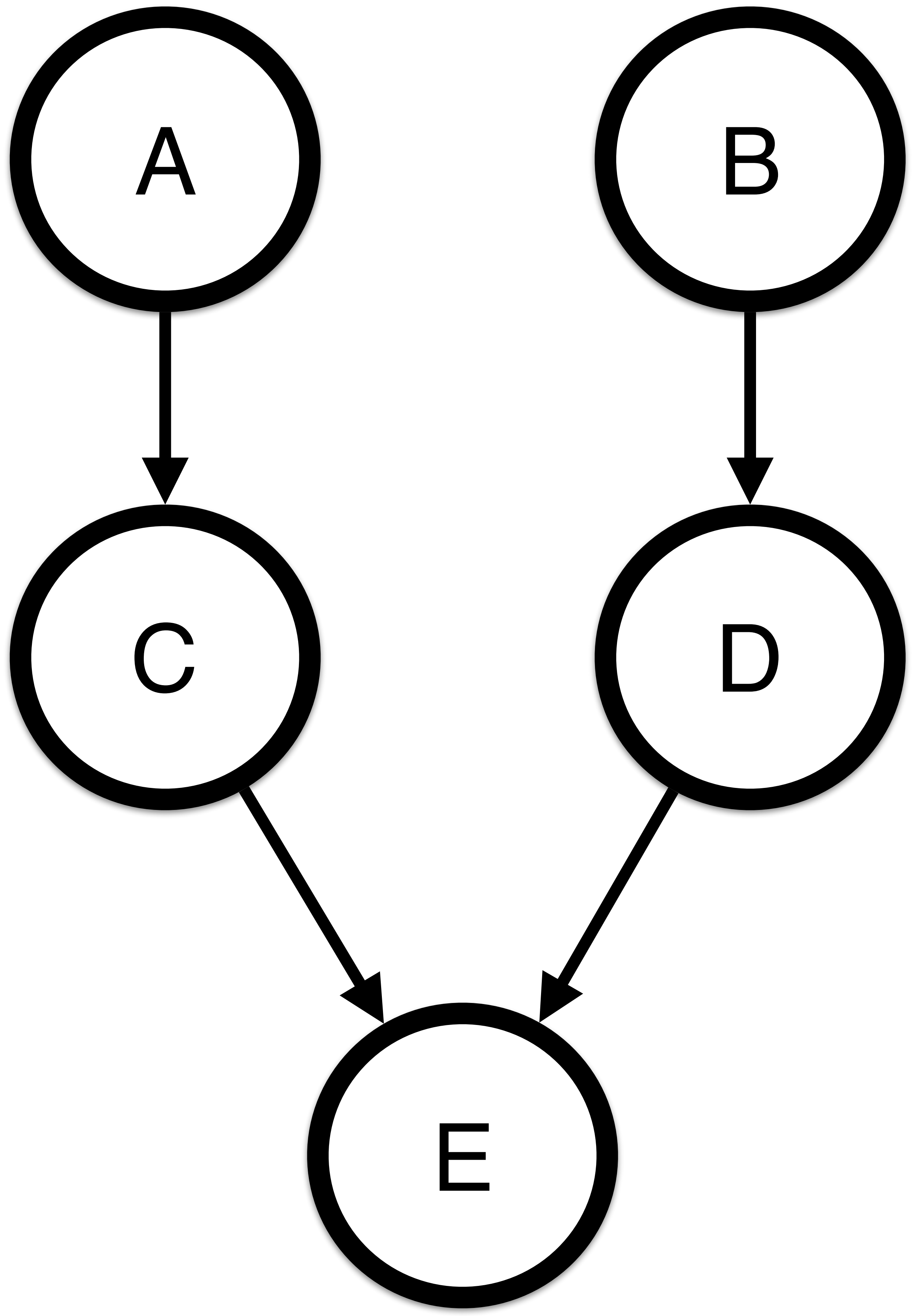
logistic regression (with input likelihood)

Gaussian Mixture Model



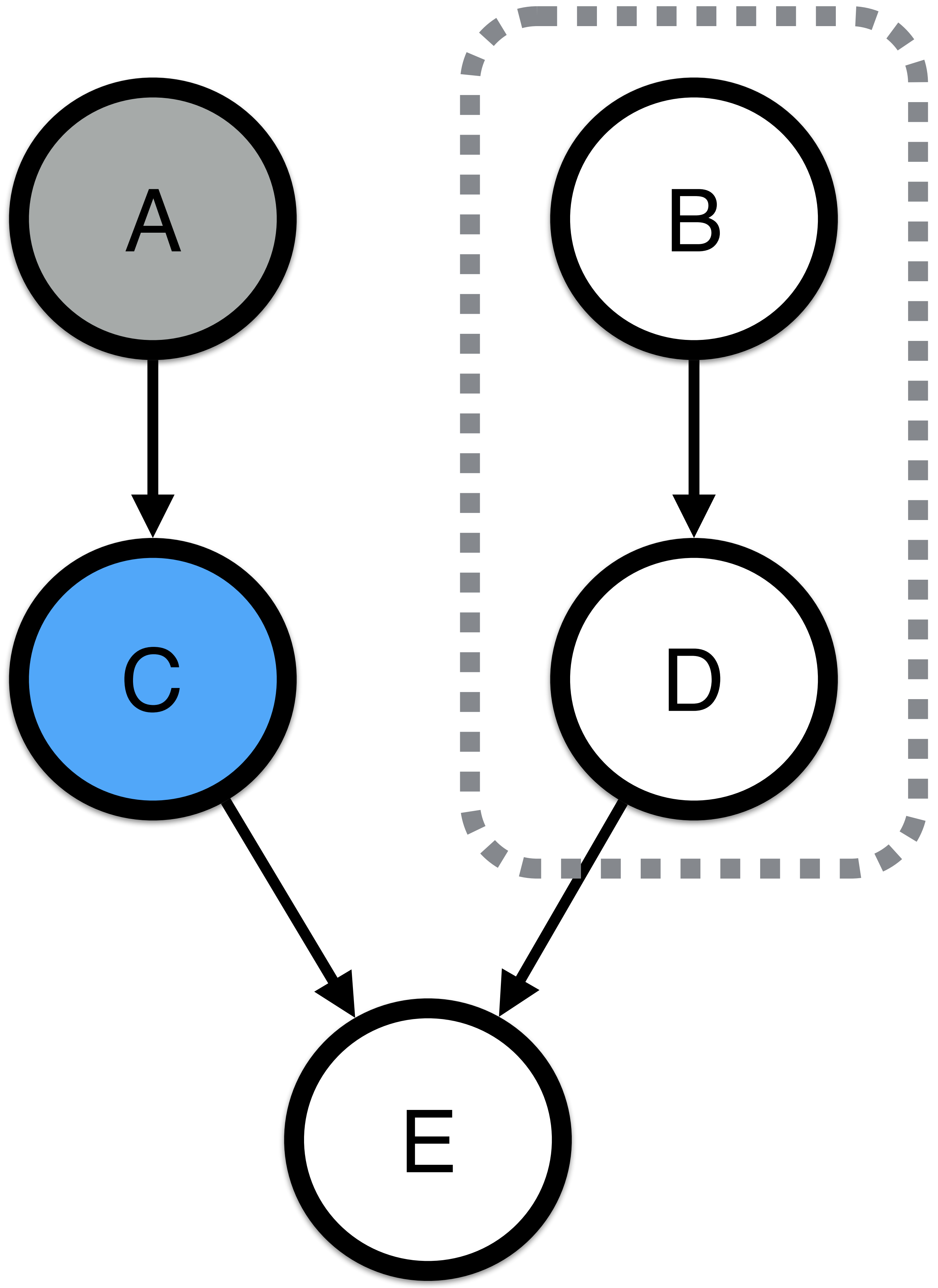
Independence in Bayes Nets

- Each variable is conditionally independent of its **non-descendents** given its **parents**



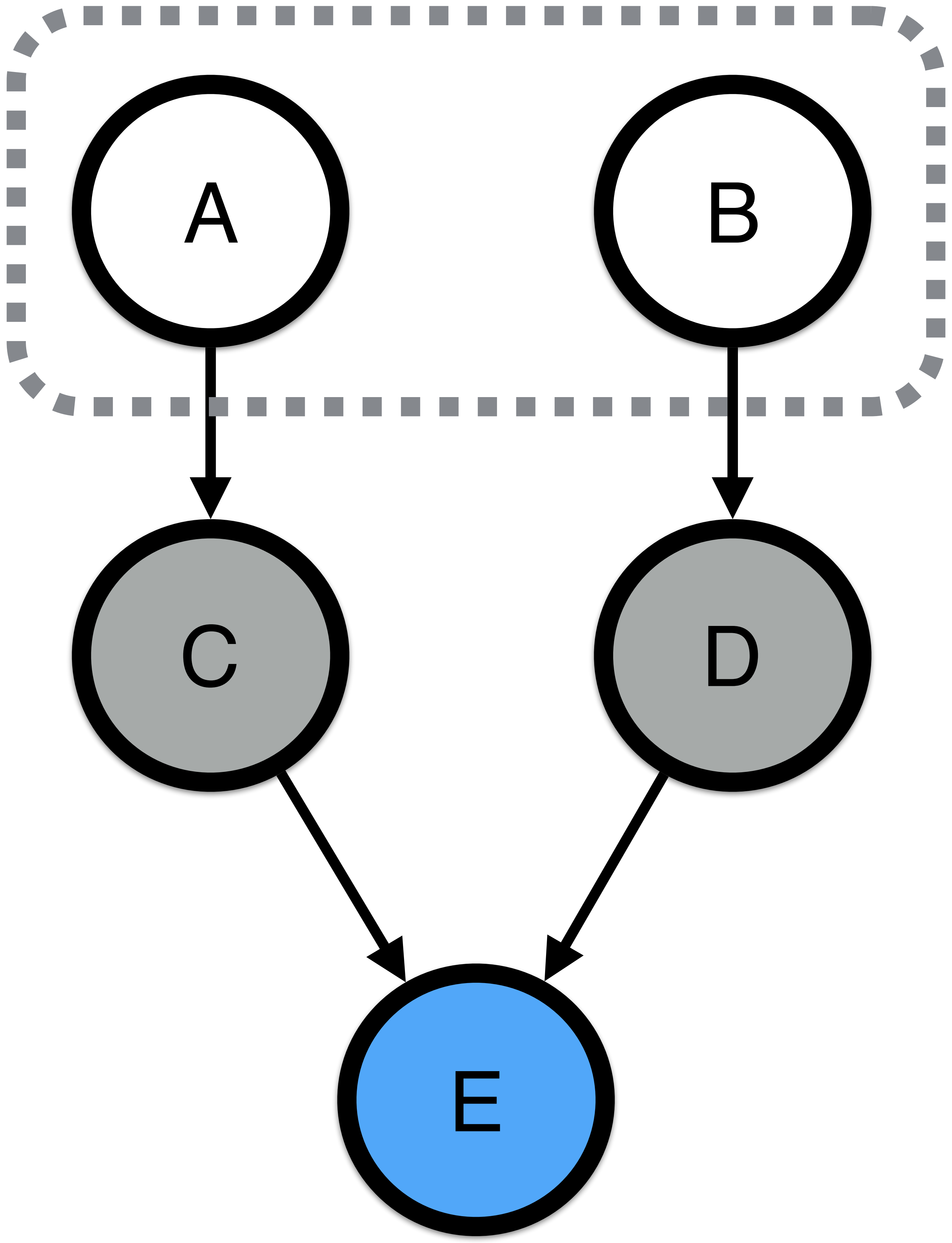
Independence in Bayes Nets

- Each variable is conditionally independent of its **non-descendents** given its **parents**



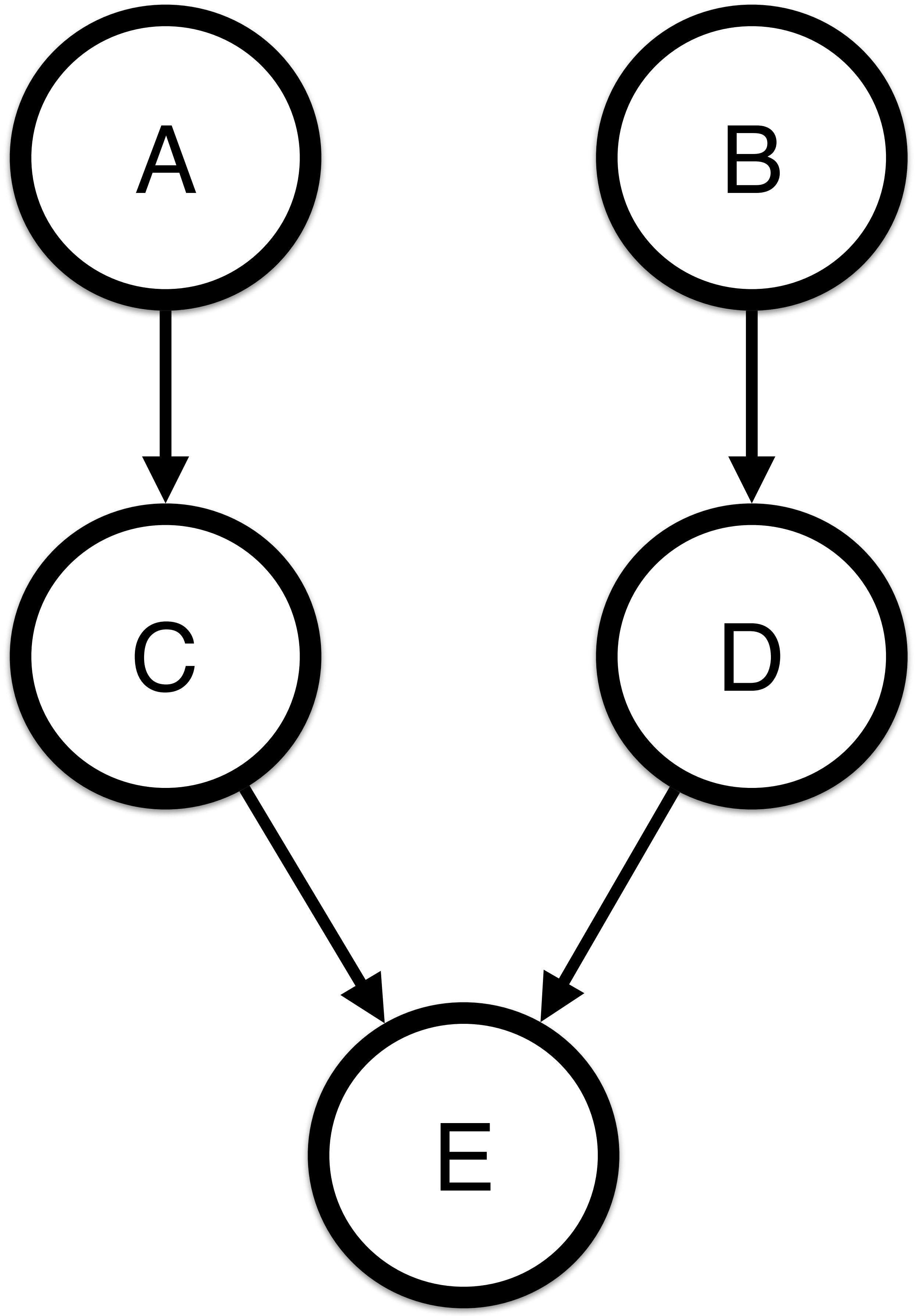
Independence in Bayes Nets

- Each variable is conditionally independent of its **non-descendents** given its **parents**



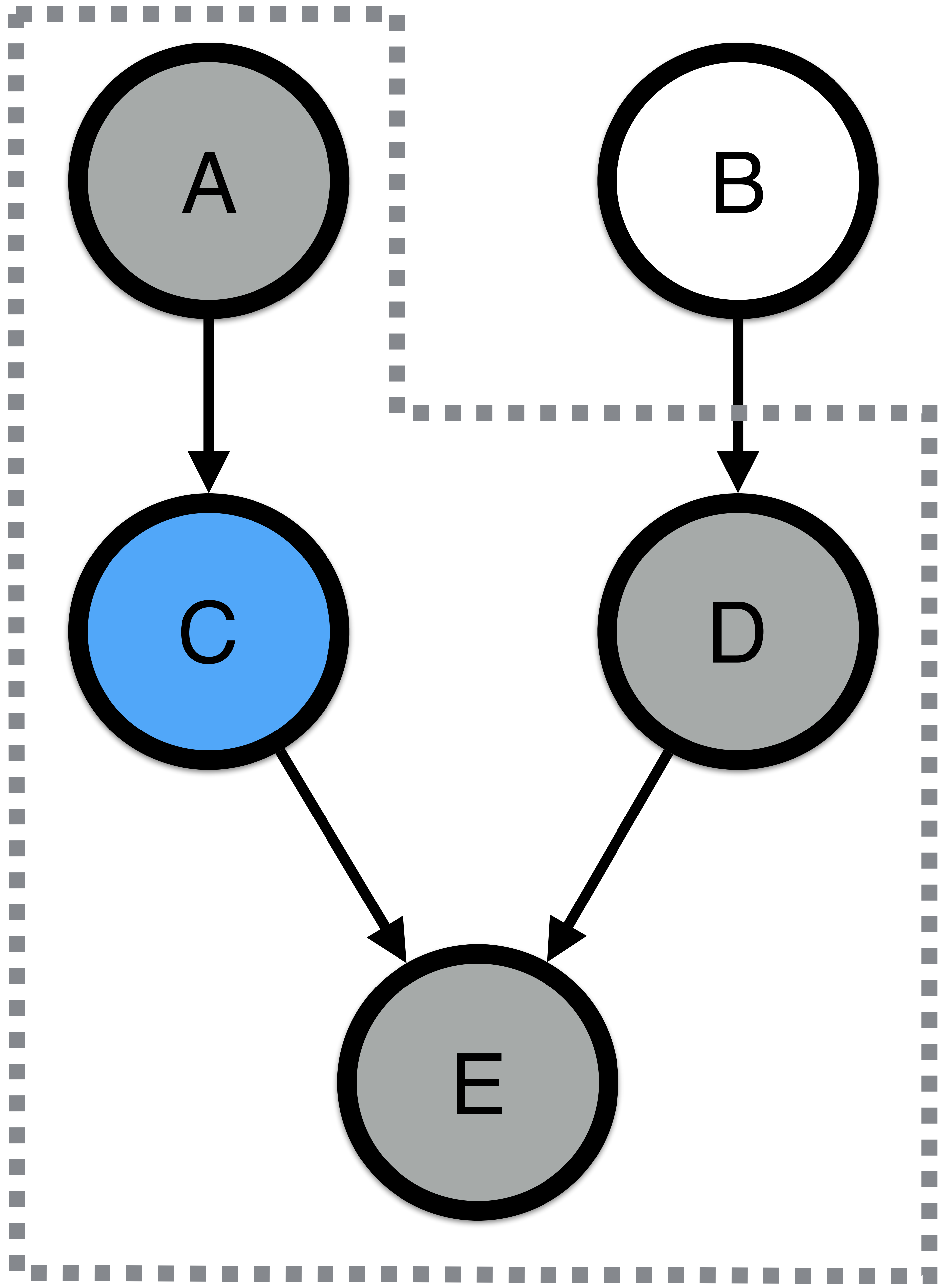
Independence in Bayes Nets

- Each variable is conditionally independent of its **non-descendents** given its **parents**
- Each variable is conditionally independent of any other variable given its **Markov blanket**
  - Parents, children, and children's parents



Independence in Bayes Nets

- Each variable is conditionally independent of its **non-descendents** given its **parents**
- Each variable is conditionally independent of any other variable given its **Markov blanket**
    - Parents, children, and children's parents



General Inference: Variable Elimination

- Every variable that is not an ancestor of a query variable or evidence variable is irrelevant to the query. Sum out irrelevant variables.
- Iterate:
  - choose variable to eliminate
  - sum terms relevant to variable, generate new factor
  - until no more variables to eliminate
- Exact inference is #P-Hard
  - in tree-structured BNs, linear time (in number of table entries)

Learning in Bayes Nets

- Super easy!
- Estimate each conditional probability
    - just like we did for naive Bayes

Bayesian Networks Summary

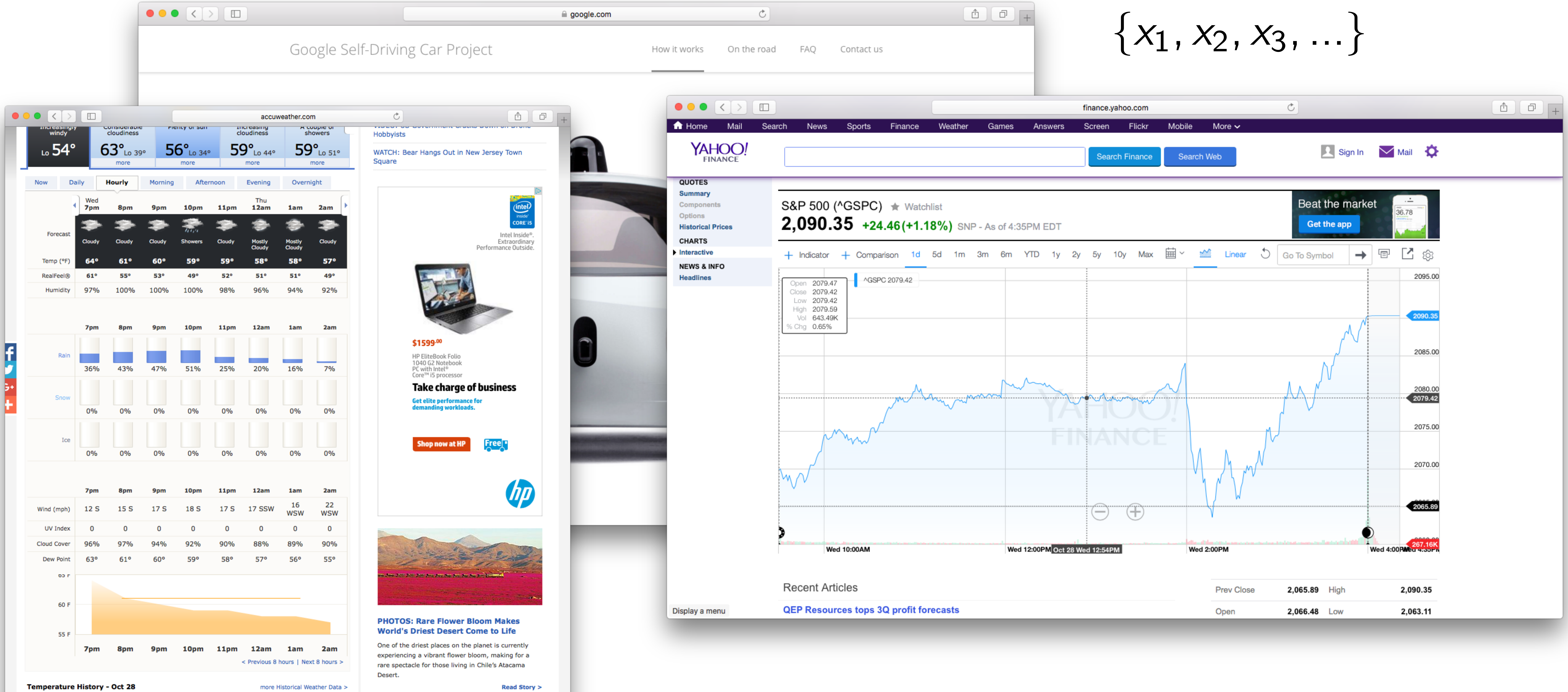
- Directed graph represents conditional dependence structure.
- Each variable conditioned on parents.
- General graph-based inference and learning algorithms

Time Series Bayes Nets

- Markov models
- Variable elimination in Markov models
- Forward message-passing inference
- Hidden Markov Models
- Forward-backward inference
- Learning

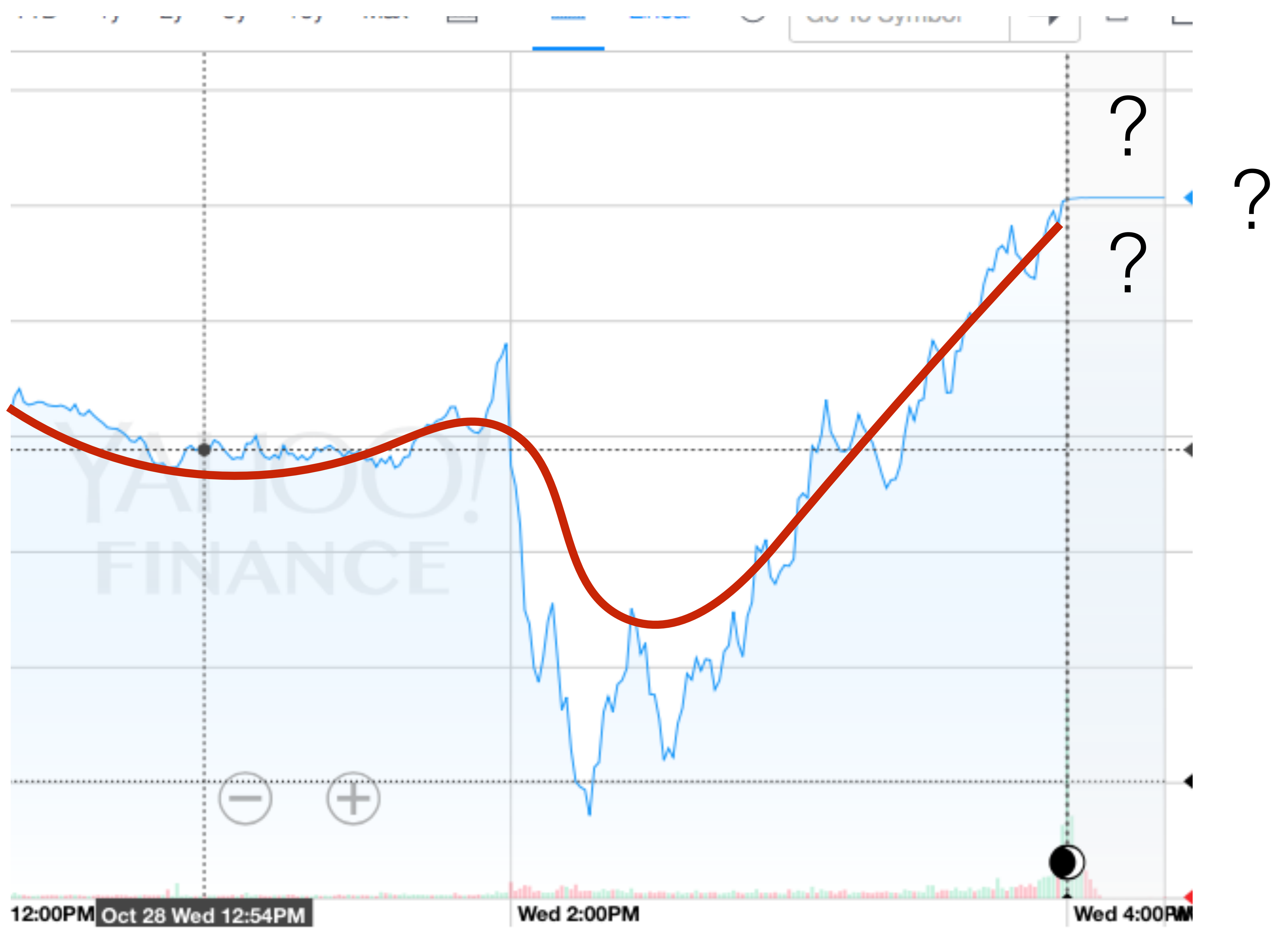
Time Series

$$\{x_1, x_2, x_3, \ldots\}$$



Time Series

- Goals:
  - Prediction
  - Filtering, smoothing



Markov Models

Markov assumption: the past is independent of the future given the present

$$p(x_i, x_k | x_j) = p(x_i | x_j) p(x_k | x_j) \qquad i < j < k$$

$$p(x_1, \ldots, x_T) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1} | x_t)$$

usually parameterized with function independent of ***t***

Variable Elimination

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)$$

$p(x_4)$?

$$p(x_4) = \sum_{x_1,x_2,x_3} p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)$$

$$p(x_2) = \alpha_2(x_2) = \sum_{x_1} p(x_1)p(x_2|x_1)$$

$$p(x_4) = \sum_{x_2,x_3} \alpha_2(x_2)p(x_3|x_2)p(x_4|x_3)$$

$$p(x_3) = \alpha_3(x_3) = \sum_{x_2} \alpha_2(x_2)p(x_3|x_2)$$

$$p(x_4) = \sum_{x_3} \alpha_3(x_3)p(x_4|x_3)$$

Forward Message Passing

$$p(X) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1}|x_t)$$

for ***t*** from 1 to (T-1):

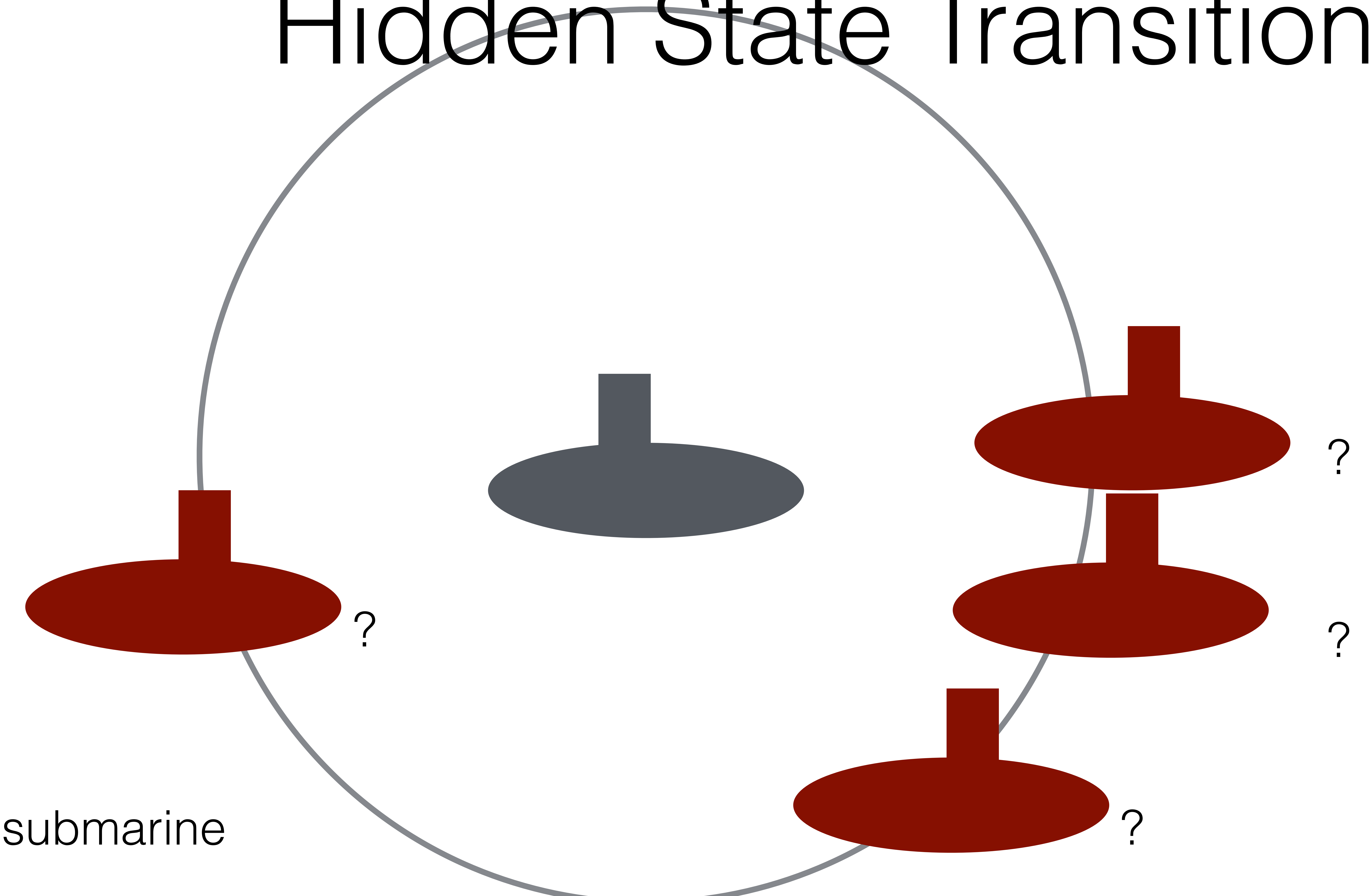
$$p(x_{t+1}) = \sum_{x_t} p(x_t) p(x_{t+1}|x_t)$$

Outline

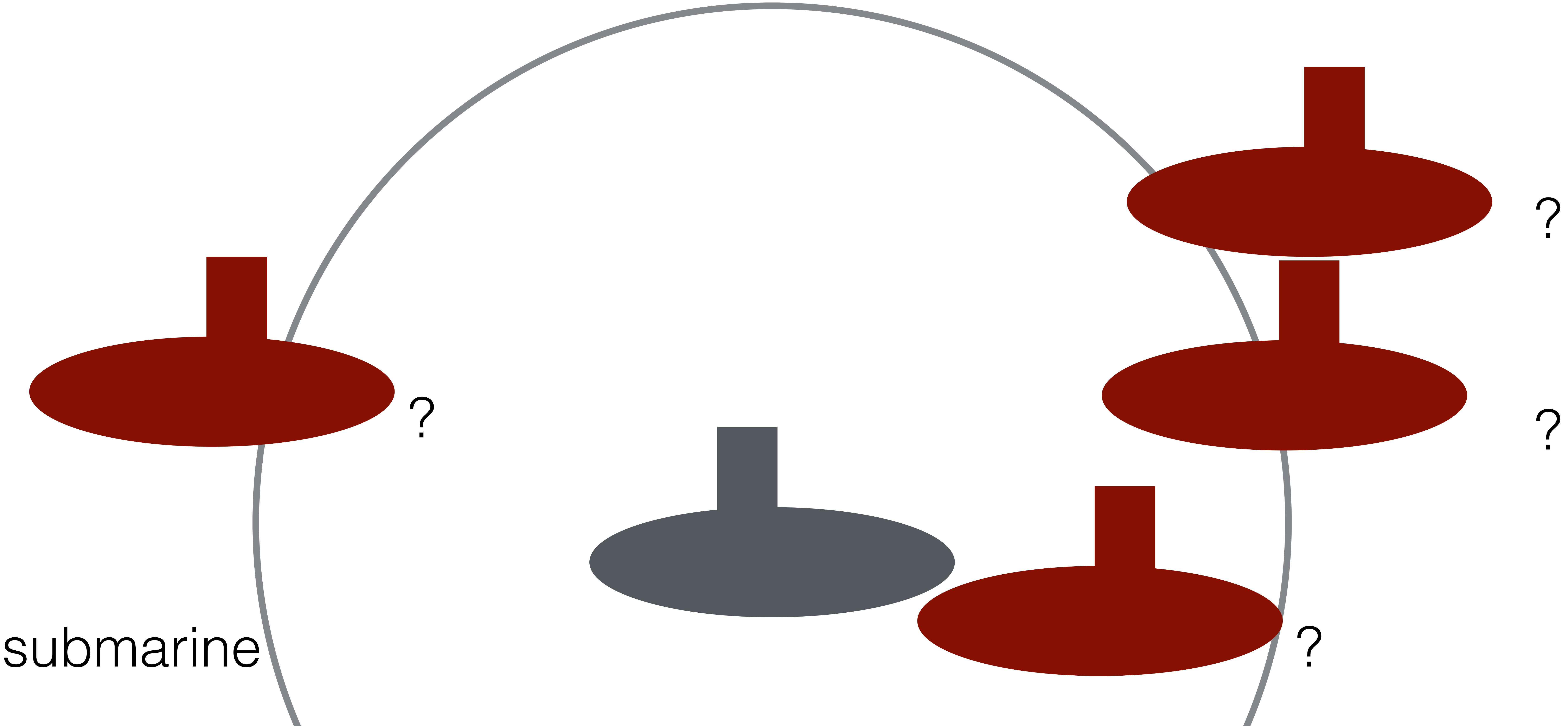
- Markov models
- Variable elimination in Markov models
- Forward message-passing inference

- Hidden Markov Models
- Forward-backward inference
- Learning

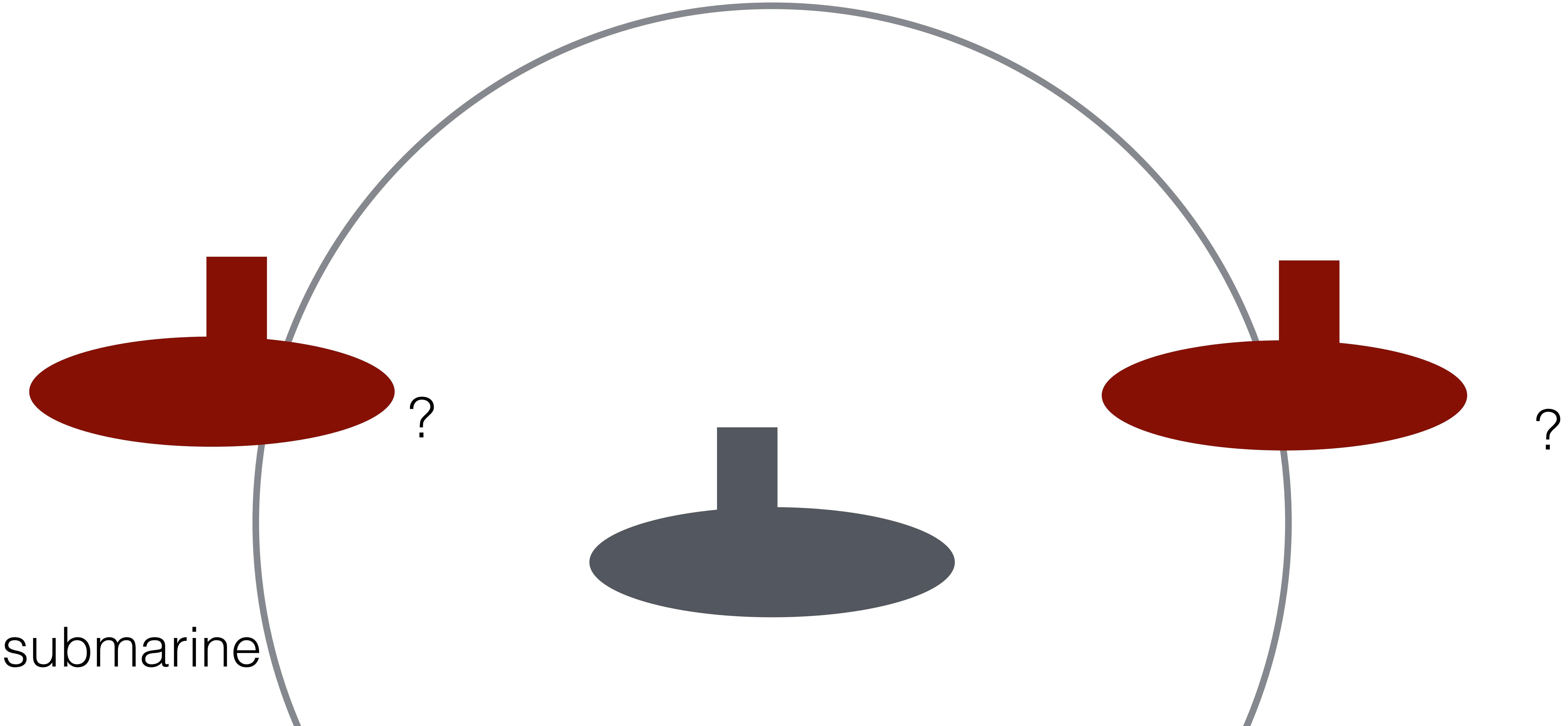
Hidden State Transitions



Hidden State Transitions



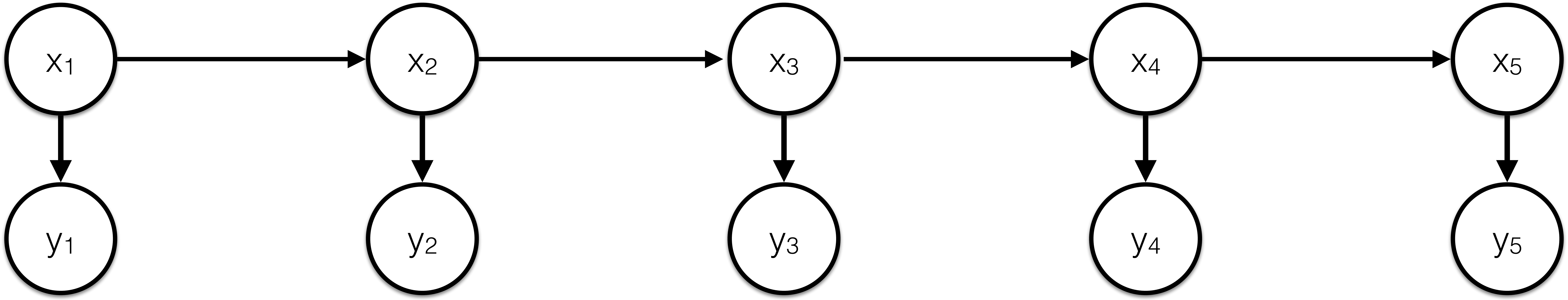
Hidden State Transitions



Hidden Markov Models

$p(y_t \| x_t)$	observation probability	SONAR noisiness
$p(x_t \| x_{t-1})$	transition probability	submarine locomotion

$$p(X, Y) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1}|x_t) \prod_{t'=1}^{T} p(y_{t'}|x_{t'})$$



Hidden State Inference

$$p(X|Y) \qquad p(x_t|Y)$$

$$\alpha_t(x_t) = p(x_t, y_1, \ldots, y_t) \qquad \beta_t(x_t) = p(y_{t+1}, \ldots, y_T|x_t)$$

$$\alpha_t(x_t)\beta_t(x_t) = p(x_t, y_1, \ldots, y_t)p(y_{t+1}, \ldots, y_T|x_t) \;\; = p(x_t, Y) \;\; \propto p(x_t|Y)$$

normalize to get conditional probability

note: not the same as $p(x_1, \ldots, x_T, Y)$

Forward Inference

$$\alpha_t(x_t) = p(x_t, y_1, \ldots, y_t)$$

$$p(x_1, y_1) = p(x_1)p(y_1|x_1) = \alpha_1(x_1)$$

$$p(x_2, y_1, y_2) = \sum_{x_1} p(x_1, y_1)p(x_2|x_1)p(y_2|x_2) = \alpha_2(x_2) = \sum_{x_1} \alpha_1(x_1)p(x_2|x_1)p(y_2|x_2)$$

$$p(x_{t+1}, y_1, \ldots, y_{t+1}) = \alpha_{t+1}(x_{t+1}) = \sum_{x_t} \alpha_t(x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})$$

Backward Inference

$$\beta_t(x_t) = p(y_{t+1}, \dots, y_T | x_t)$$

$$p(\{\}|x_T) = 1 = \beta_T(x_T)$$

$$\begin{aligned}\beta_{t-1}(x_{t-1}) = p(y_t, \dots, y_T | x_{t-1}) &= \sum_{x_t} p(x_t|x_{t-1}) p(y_t, y_{t+1}, \dots, y_T | x_t) \\ &= \sum_{x_t} p(x_t|x_{t-1}) p(y_t|x_t) p(y_{t+1}, \dots, y_T | x_t) \\ &= \sum_{x_t} p(x_t|x_{t-1}) p(y_t|x_t) \beta_t(x_t)\end{aligned}$$

Backward Inference

$$\beta_t(x_t) = p(y_{t+1}, \ldots, y_T | x_t)$$

$$p(\{\}|x_T) = 1 = \beta_T(x_T)$$

$$\beta_{t-1}(x_{t-1}) = p(y_t, \ldots, y_T | x_{t-1}) = \sum_{x_t} p(x_t|x_{t-1}) p(y_t|x_t) \beta_t(x_t)$$

Fusing the Messages

$$\alpha_t(x_t) = p(x_t, y_1, \ldots, y_t) \qquad \beta_t(x_t) = p(y_{t+1}, \ldots, y_T | x_t)$$

$$\alpha_t(x_t)\beta_t(x_t) = p(x_t, y_1, \ldots, y_t)p(y_{t+1}, \ldots, y_T | x_t) \;\; = p(x_t, Y) \;\; \propto p(x_t | Y)$$

$$\begin{aligned} p(x_t, x_{t+1} | Y) &= \frac{p(x_t, x_{t+1}, y_1, \ldots, y_t, y_{t+1}, y_{t+2}, \ldots, y_T)}{p(Y)} \\ &= \frac{p(x_t, y_1, \ldots, y_t)p(x_{t+1}|x_t)p(y_{t+2}, \ldots, y_T | x_{t+1})p(y_{t+1}|x_{t+1})}{\sum_{x_T} p(x_t, Y)} \\ &= \frac{\alpha_t(x_t)p(x_{t+1}|x_t)\beta_{t+1}(x_{t+1})p(y_{t+1}|x_{t+1})}{\sum_{x_T} \alpha_T(x_T)} \end{aligned}$$

Forward-Backward Inference

$$\alpha_1(x_1) = p(x_1)p(y_1|x_1)$$

$$\alpha_{t+1}(x_{t+1}) = \sum_{x_t} \alpha_t(x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})$$

$$\beta_T(x_T) = 1$$

$$\beta_{t-1}(x_{t-1}) = \sum_{x_t} p(x_t|x_{t-1})p(y_t|x_t)\beta_t(x_t)$$

$$p(x_t, Y) = \alpha_t(x_t)\beta_t(x_t)$$

$$p(x_t|Y) = \frac{\alpha_t(x_t)\beta_t(x_t)}{\sum_{x'_t} \alpha_t(x'_t)\beta_t(x'_t)}$$

$$p(x_t, x_{t+1}|Y) \quad = \frac{\alpha_t(x_t)p(x_{t+1}|x_t)\beta_{t+1}(x_{t+1})p(y_{t+1}|x_{t+1})}{\sum_{x_T} \alpha_T(x_T)}$$

Normalization

To avoid underflow, re-normalize at each time step

$$\tilde{\alpha}_t(x_t) = \frac{\alpha_t(x_t)}{\sum_{x_t'} \alpha_t(x_t')} \qquad \tilde{\beta}_t(x_t) = \frac{\beta_t(x_t)}{\sum_{x_t'} \beta_t(x_t')}$$

(Normalization cancels out.)

Learning

- Parameterize and learn

$p(x_{t+1}|x_t)$

conditional probability table
transition matrix

$p(y_t|x_t)$

observation model
emission model

- If fully observed, super easy!

- If $\boldsymbol{x}$ is hidden (most cases) treat as latent variable

    - E.g., expectation maximization

EM (Baum-Welch) Details

Compute $p(x_t|Y)$ and $p(x_t, x_{t+1}|Y)$ using forward-backward

Maximize weighted (expected) log-likelihood

$$p(x_1) \leftarrow \frac{1}{T}\sum_{t=1}^{T} p(x_t|Y) \quad \text{or} \quad p(x_1|Y)$$

$$p(x_{t'+1} = i | x_{t'} = j) \leftarrow \frac{\sum_{t=1}^{T-1} p(x_{t+1} = i, x_t = j|Y)}{\sum_{t=1}^{T-1} p(x_t = j|Y)}$$

e.g., Gaussian

$$\mu_x \leftarrow \frac{\sum_{t=1}^{T} p(x_t = x|Y) y_t}{\sum_{t'=1}^{T} p(x_t = x|Y)}$$

$$p(y|x) \leftarrow \frac{\sum_{t=1}^{T} p(x_t = x|Y) I(y_t = y)}{\sum_{t'=1}^{T} p(x_t = x|Y)}$$

e.g., multinomial

Time Series Bayes Net Summary

- MMs model state transitions
- HMMs represent hidden states
  - Transitions between adjacent states, observation based on states
- Forward-backward inference to incorporate all evidence
- Expectation maximization to train parameters (Baum-Welch) with latent state variables