Week 21: Bayesian Networks

Martha Lewis
(based on slides from Raul Santos Rodriguez)

Have a look at ...

... Russell and Norvig (Ch. 14)

... David Barber's Bayesian reasoning and Machine Learning:
`http://www.cs.ucl.ac.uk/staff/d.barber/brml/`

Outline

This week we look at probabilistic graphical models. These are very powerful representations that enable us to scale probabilistic models to large real problems. The objective is to learn the following topics:

- Bayes's theorem
- Definition of a DAG (or Bayesian network)
- Conditional independence in DAGs
- Dynamic Bayesian networks
- Hidden Markov Models

Bayes's theorem: The Horrible Disease?

You are about to be tested for a rare disease. How worried should you be if the test result is positive? The doctor has bad news and good news...

Bad You tested positive for a serious disease, and that the test is 99% accurate *(i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease)*.

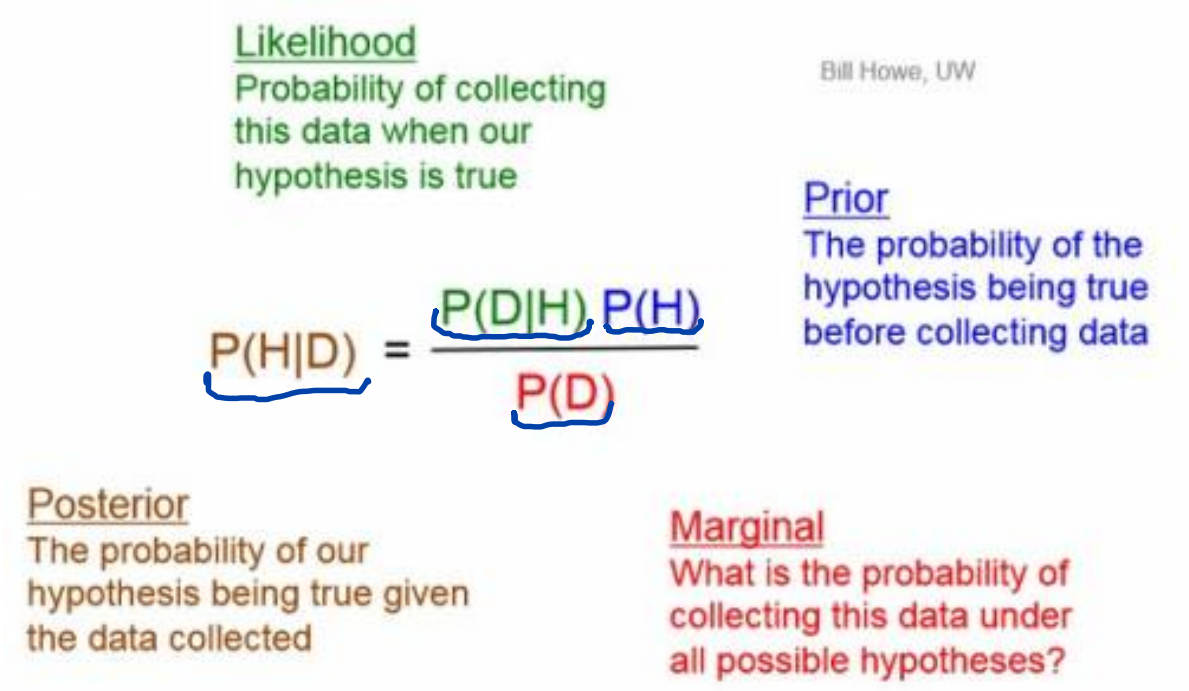
Good This is a rare disease, striking only 1 in 10,000 people.

What are the chances that you actually have the disease?

Bayes's theorem

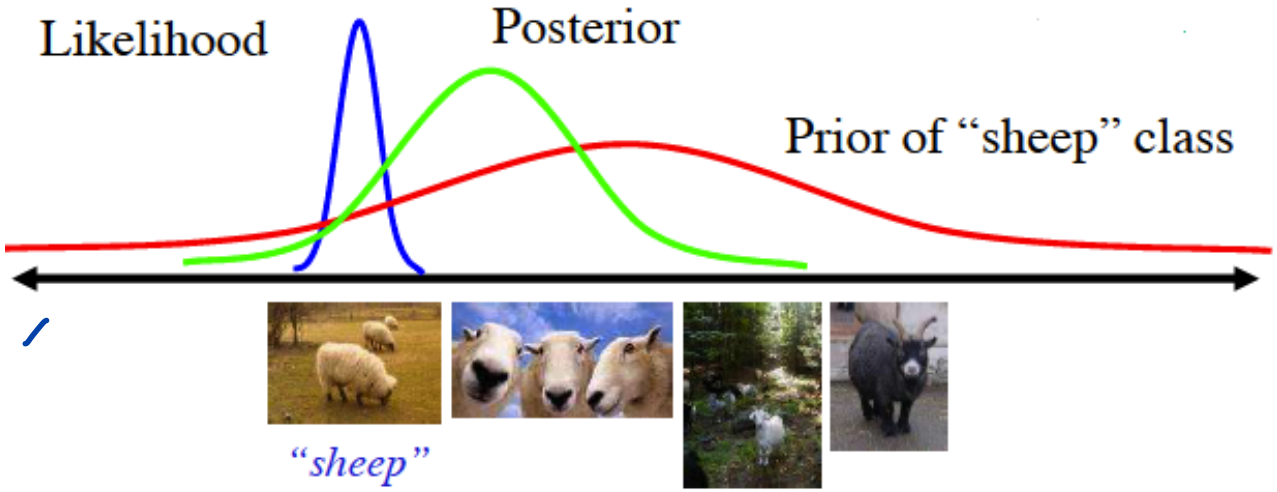
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

Bayes's theorem



Bayesian inference

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$



[De Freitas, 2012]

Bayes's theorem: example

The test is 99% accurate:

$$P(T = 1|D = 1) = 0.99$$

$$P(T = 0|D = 0) = 0.99$$

where T denotes test and D denotes disease.

The disease affects 1 in 10000:

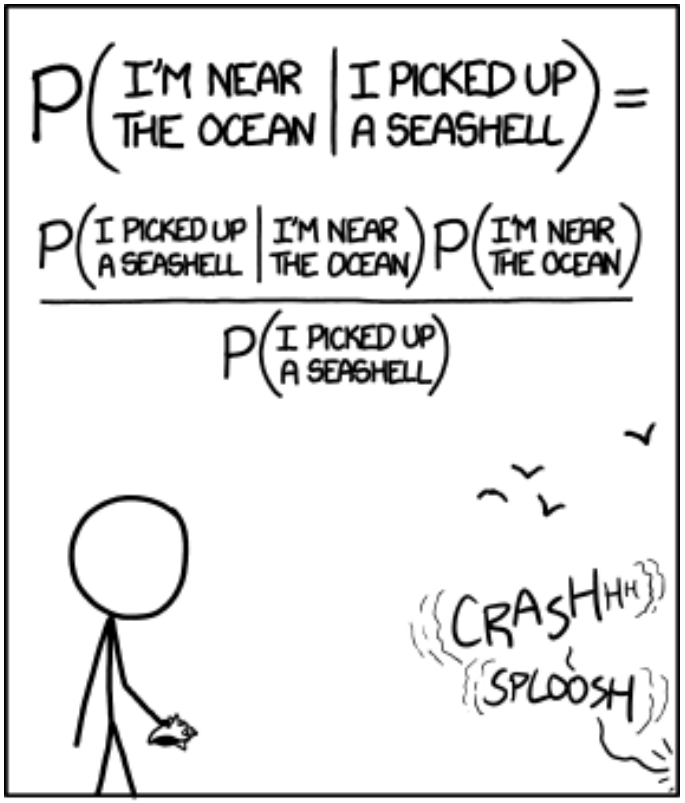
$$P(D = 1) = 0.0001$$

$$P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1|D = 0)P(D = 0) + P(T = 1|D = 1)P(D = 1)}$$

$$P(D = 1|T = 1) = 0.0098$$

[https://www.intelligentinvestor.com.au/the-theory-that-cracked-the-enigma-code-1811436]

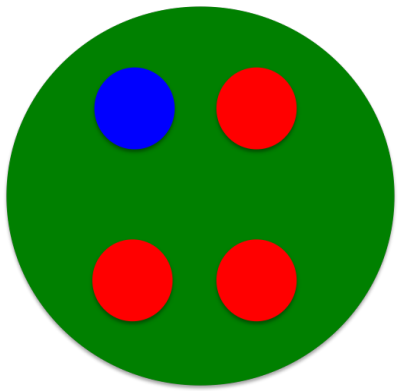
Bayes's theorem



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

A simple graphical model: Example

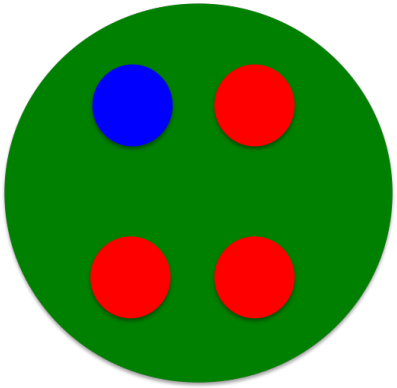
Problem: drawing balls from the set {r,r,r,b}



What are $P(B_1)$, $P(B_2|B_1)$ and $P(B_1, B_2)$?

A simple graphical model: Example

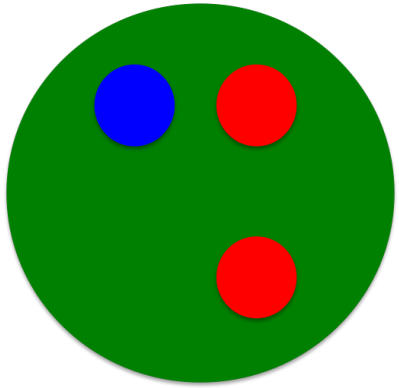
Problem: drawing balls from the set {r,r,r,b}



$$P(B_1) = \begin{array}{|c|c|} \hline \bullet\ B_1 & ? \\ \hline \bullet\ B_1 & ? \\ \hline \end{array}$$

A simple graphical model: Example

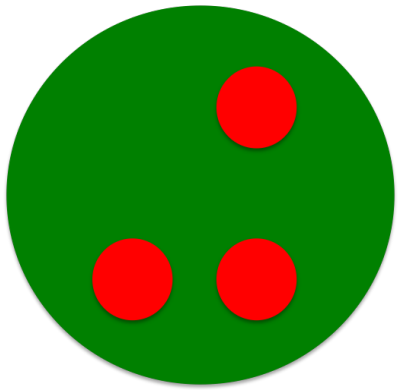
Problem: drawing balls from the set {r,r,r,b}



$$P(B_1) = \begin{array}{|c|c|} \hline B_1 & 3/4 \\ \hline B_1 & ? \\ \hline \end{array}$$

A simple graphical model: Example

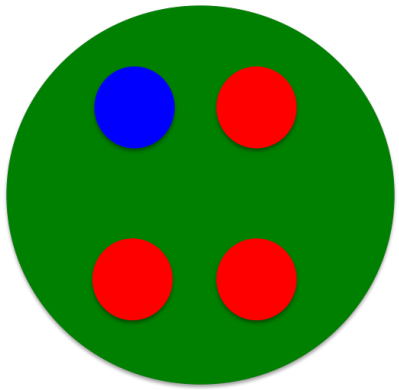
Problem: drawing balls from the set {r,r,r,b}



$$P(B_1) = \begin{array}{|c|c|} \hline B_1 & 3/4 \\ \hline B_1 & 1/4 \\ \hline \end{array}$$

A simple graphical model: Example

Problem: drawing balls from the set {r,r,r,b}

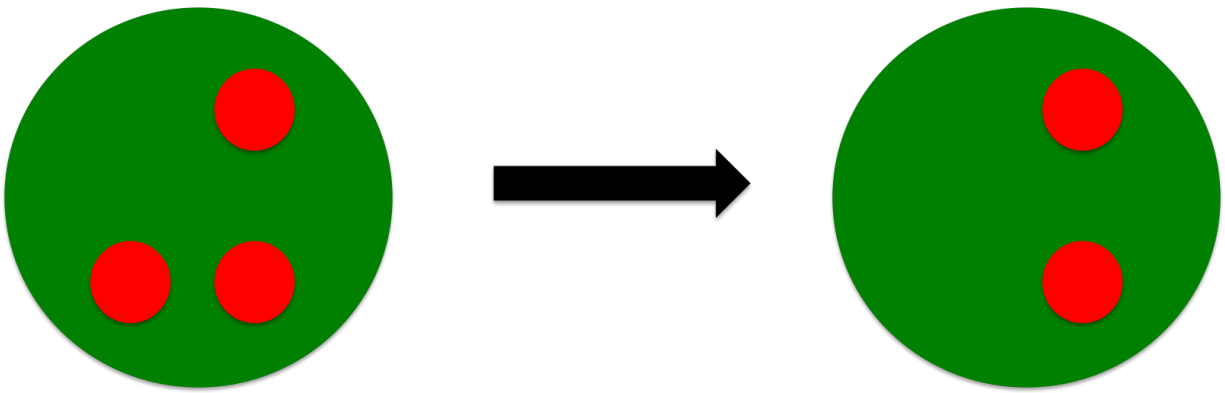


$P(B_2|B_1) =$

	B_2	B_2
B_1	?	?
B_1	?	?

A simple graphical model: Example

Problem: drawing balls from the set {r,r,r,b}

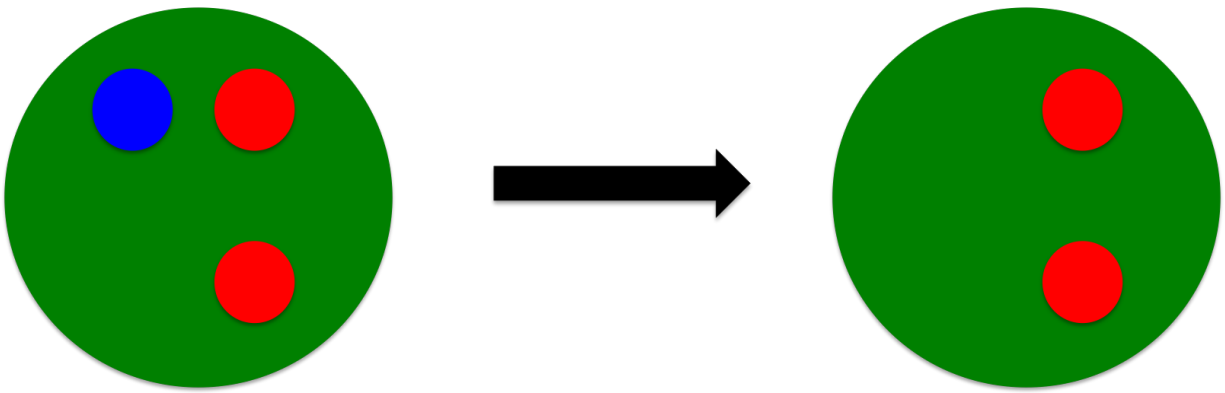


$P(B_2|B_1) =$

	B_2	B_2
B_1	?	?
B_1	1	0

A simple graphical model: Example

Problem: drawing balls from the set {r,r,r,b}

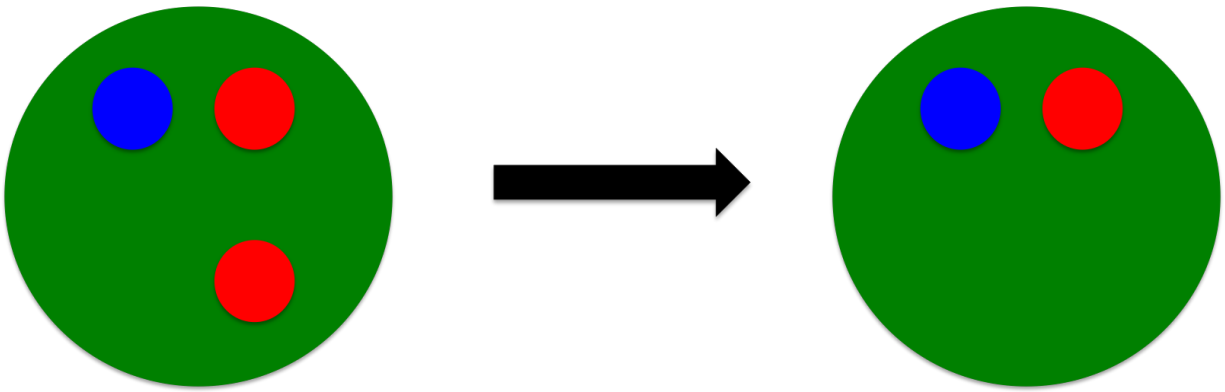


$P(B_2|B_1) =$

	B_2	B_2
B_1	?	1/3
B_1	1	0

A simple graphical model: Example

Problem: drawing balls from the set {r,r,r,b}



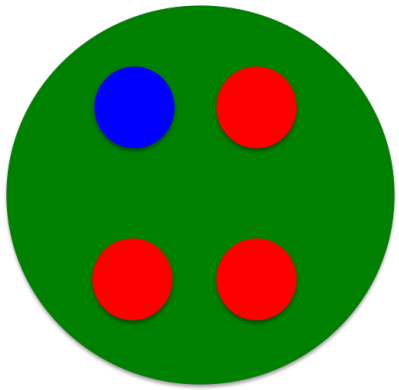
$P(B_2|B_1) =$

	B_2	B_2
B_1	2/3	1/3
B_1	1	0

A simple graphical model: Example

Problem: drawing balls from the set {r,r,r,b}

$P(B_1, B_2) := P(B_1 \cap B_2)$



$P(B_1, B_2) =$

	B_2	B_2
B_1	1/2	1/4
B_1	1/4	0

A simple graphical model: Example

Problem: drawing balls from the set {r,r,r,b}

$P(B_1) =$

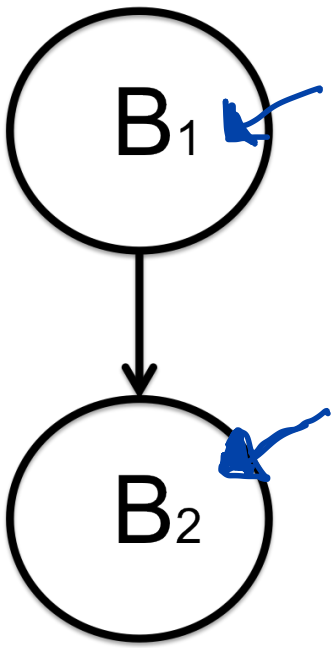
B_1	3/4
B_1	1/4

$P(B_2|B_1) =$

	B_2	B_2
B_1	2/3	1/3
B_1	1	0

$P(B_1, B_2) =$

	B_2	B_2
B_1	1/2	1/4
B_1	1/4	0



$P(B_1 \cap B_2) = P(B_2|B_1)P(B_1)$

Summary

- Bayes's theorem allows us to calculate the probability of a hypothesis being true given the data
- We looked at some worked examples

Directed probabilistic graphical models

A Bayesian Network is a directed graph in which each node is annotated with quantitative probability information.

Recall that we have $P(x_1, ..., x_n) = \Pi_i P(x_i | x_{i+1}, ..., x_n)$. Calculating this becomes intractable.

Bayesian Network

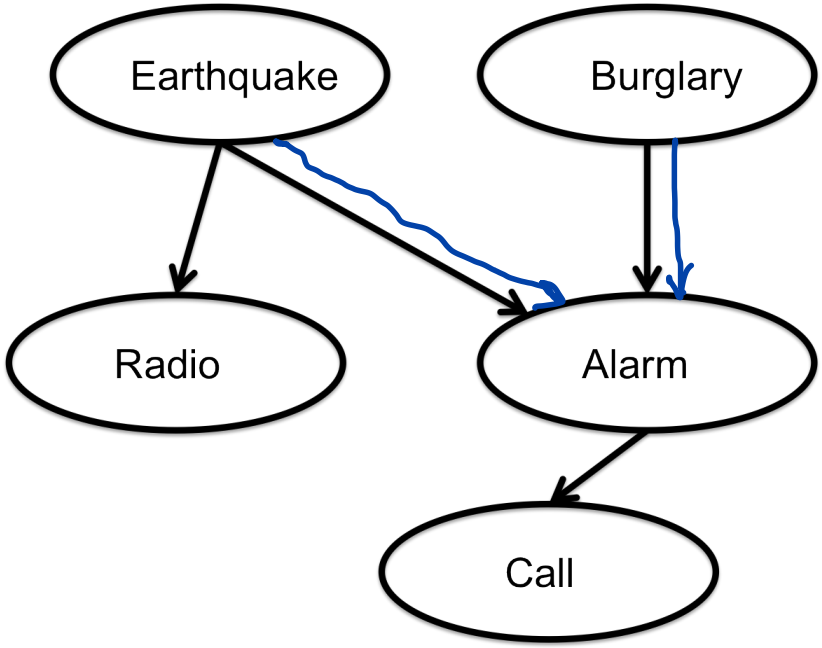
- Each node corresponds to a random variable (discrete or continuous).
- Each edge represents influence.
- The graph has no directed cycles: directed acyclic graph (DAG).
- Each node x_i has a conditional probability distribution $P(x_i | Parents(x_i))$ that quantifies the effect of the parents on the node.

The DAG tells us that if we have n variables x_i, the joint distribution of these variables factorises as follows:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(x_i))$$

Directed probabilistic graphical models: example

A Bayesian Network is a directed graph in which each node is annotated with quantitative probability information.



Russell+Norvig

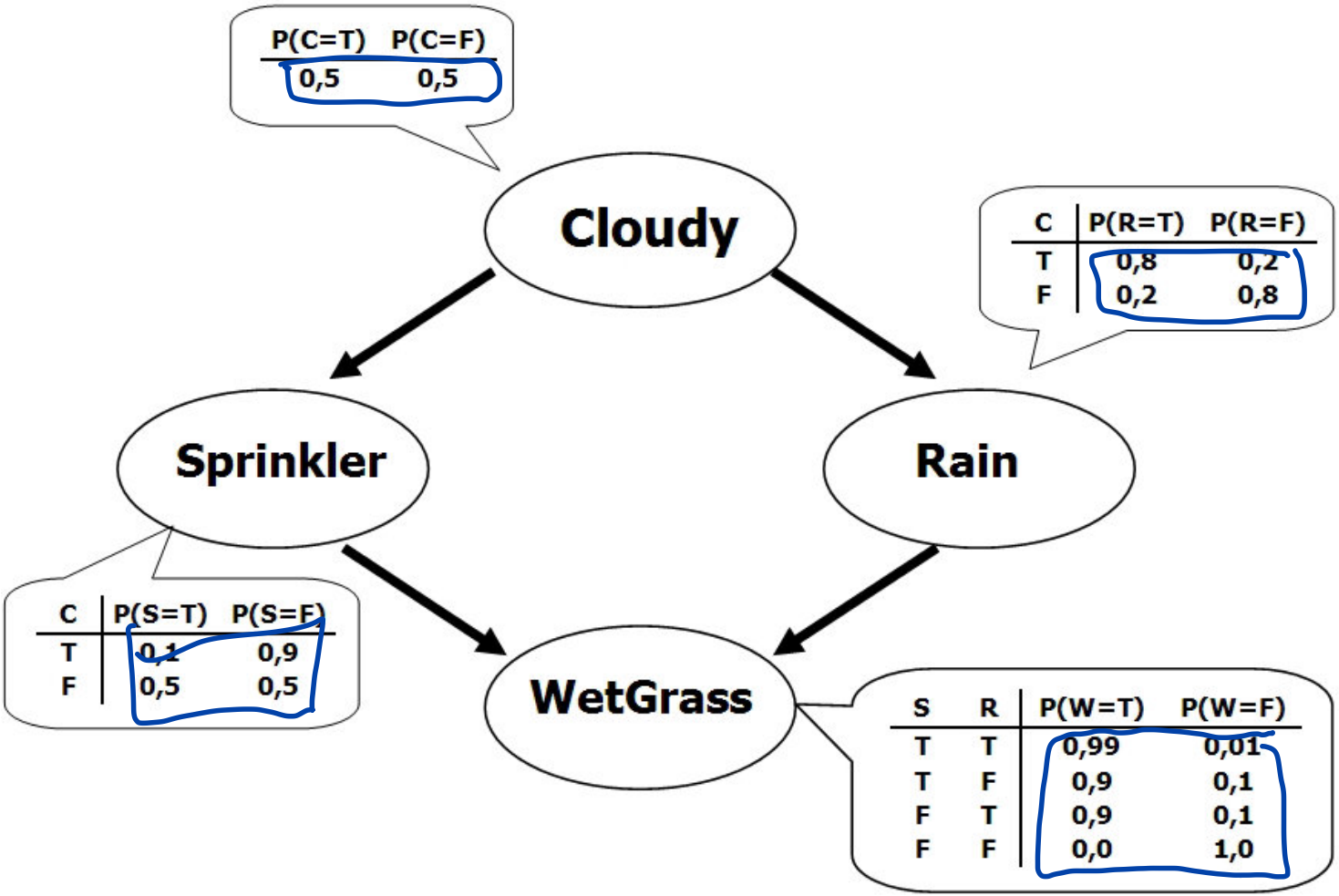
$$P(A, R, E, B) = P(A|R, E, B)P(R|E, B)P(E|B)P(B)$$

Can be reduced to:

$$P(A, R, E, B) = P(A|E, B)P(R|E)P(E)P(B)$$

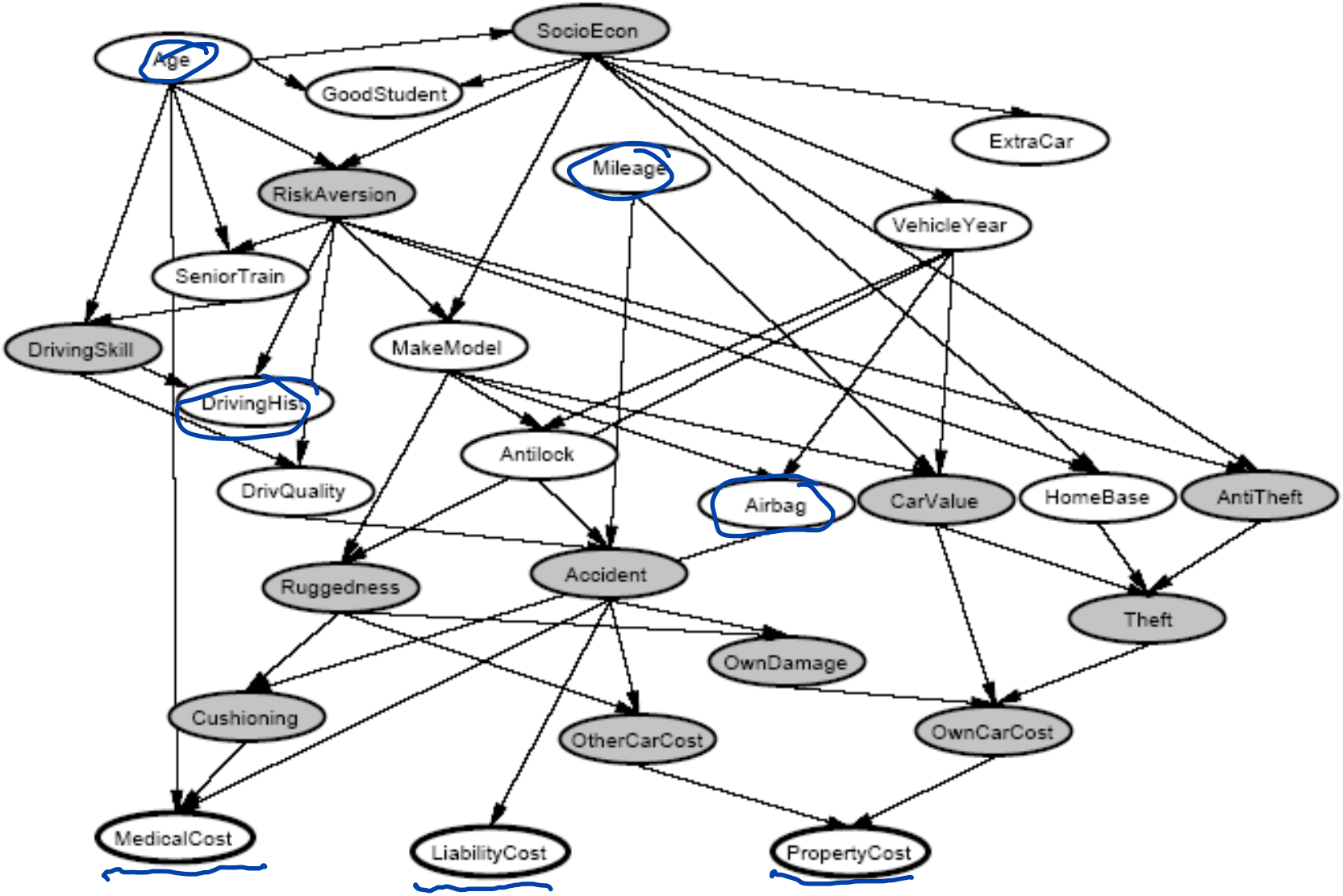
Joint vs Factorised joint distributions

$$p(C, S, R, W) = p(C)p(S|C)p(R|C)p(W|S, R)$$

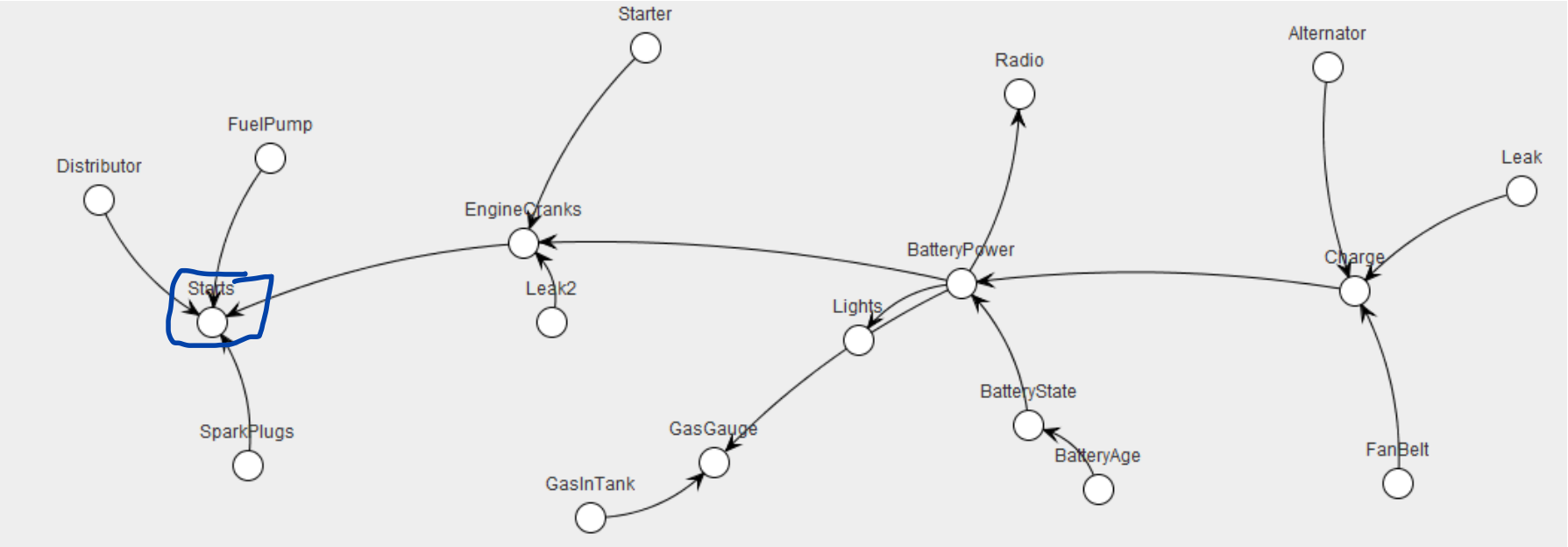


Joint distribution $\rightarrow$ 15 values vs Factorised joint distribution $\rightarrow$ 9 values

Example: insurance

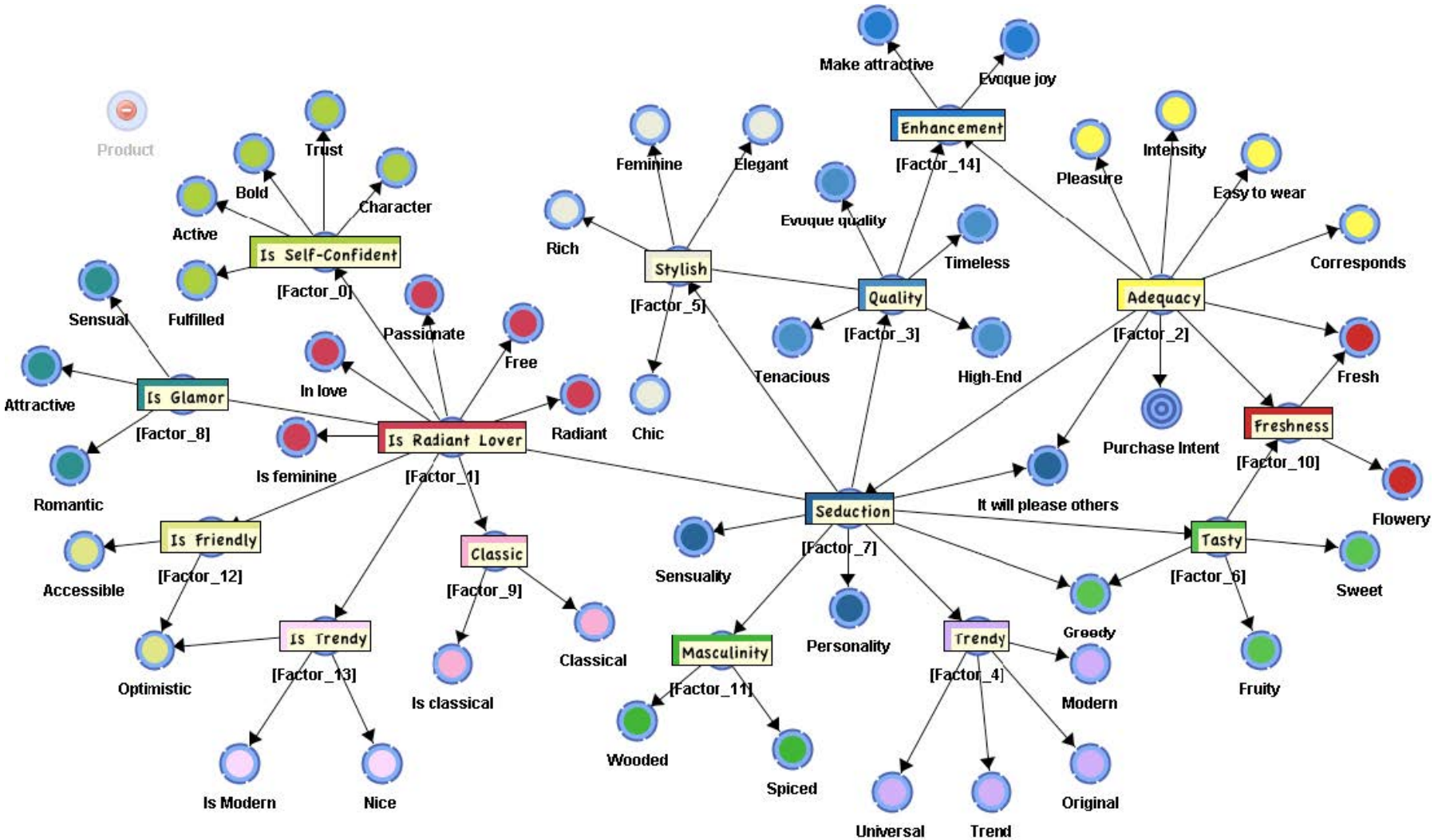


Example: car engine



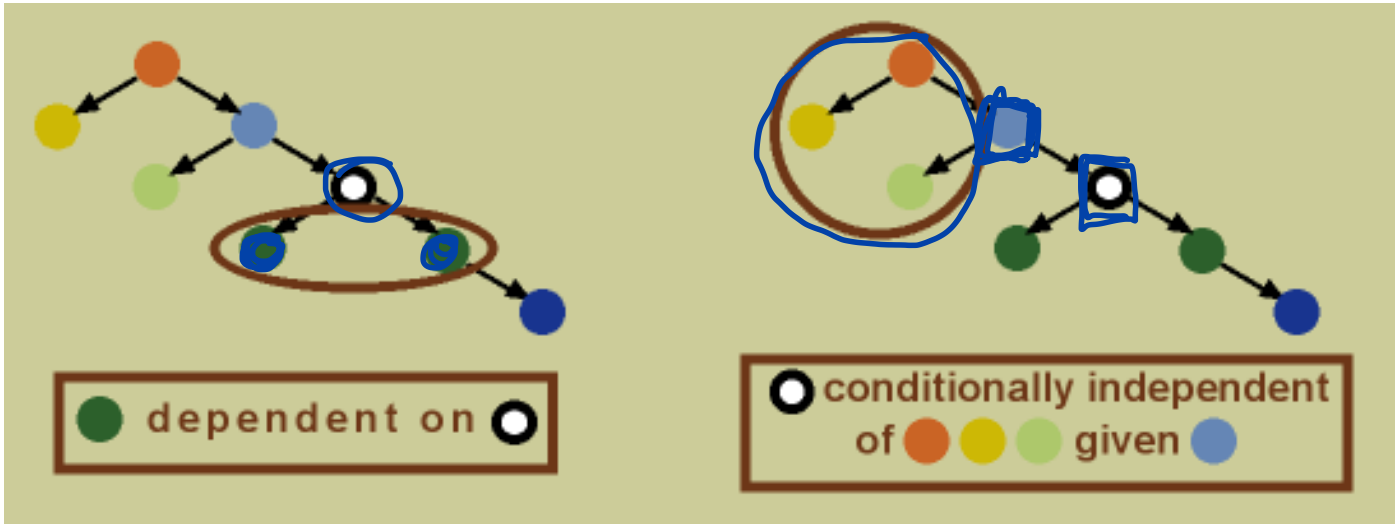
[David Heckerman and John S. Breese and Koos Rommelse]

Example: perfume market analysis



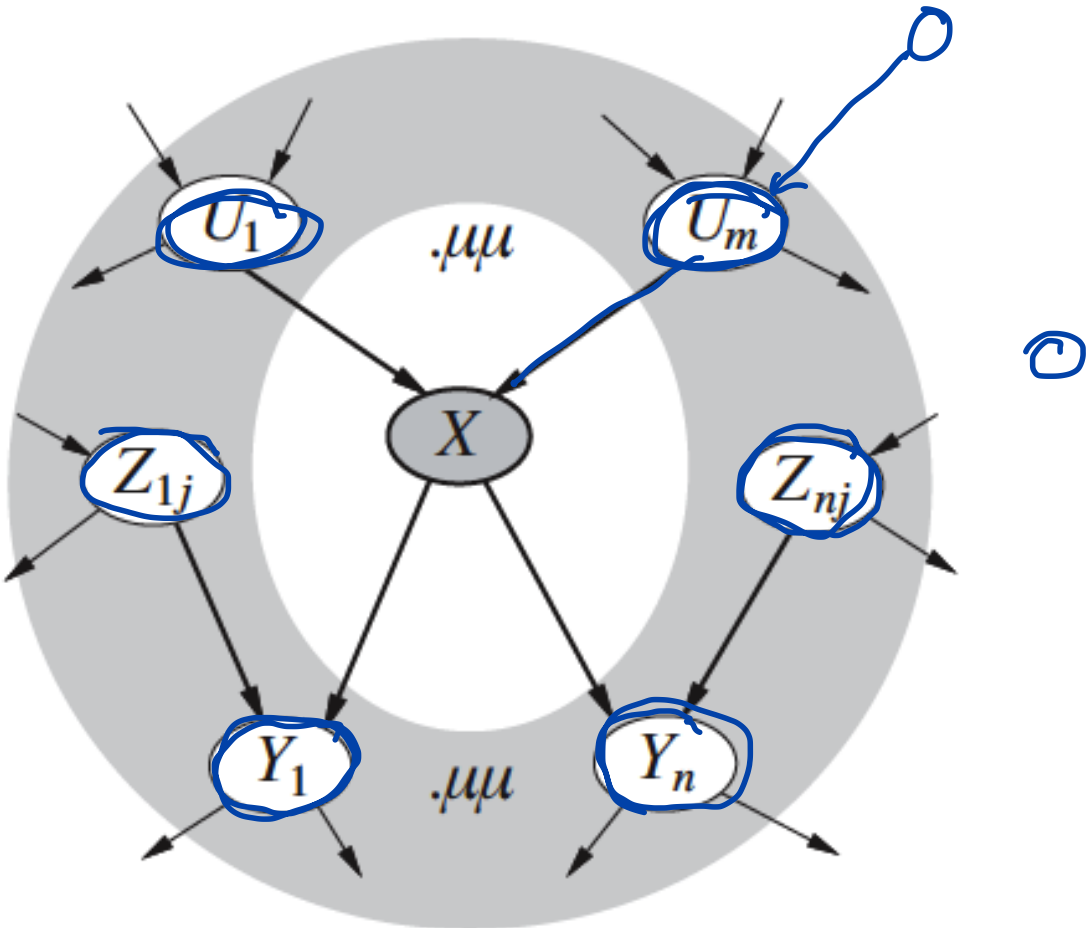
[http://www.bayesia.com/en/applications/marketing.php]

Conditional independence



Conditional independence: Markov Blanket

Given its Markov Blanket, X is independent of all other nodes



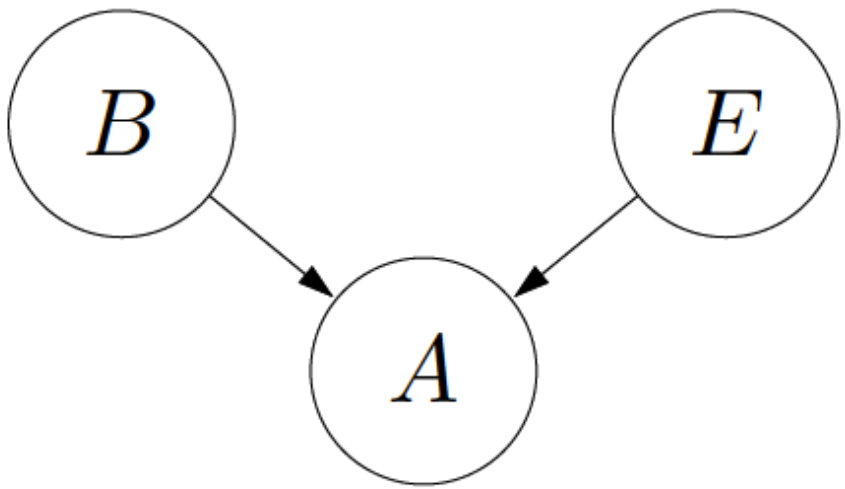
$$MB(X) = Parents(X) \cup Children(X) \cup Parents(Children(X))$$

Summary

- A directed graphical model allows us to represent the dependencies an a Bayesian network.
- By factoring out the network we can calculate probabilities more efficiently.
- In applications this problem can become huge!

Question

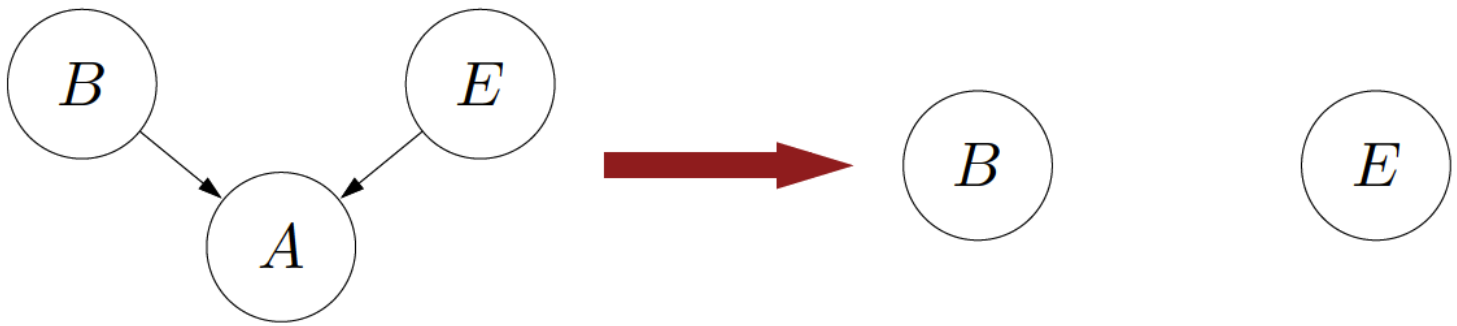
Given the following Bayesian Network,



show that B and E are independent, given A

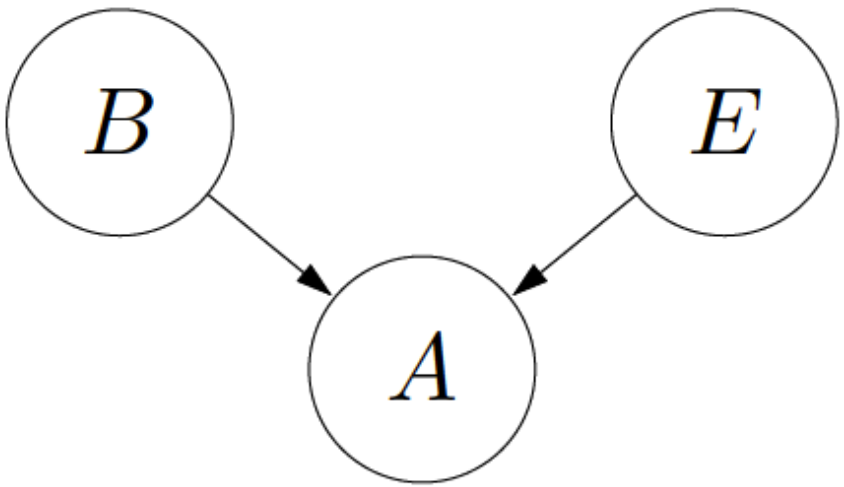
Marginalisation

Marginalisation of a leaf node yields a Bayesian network without the node.



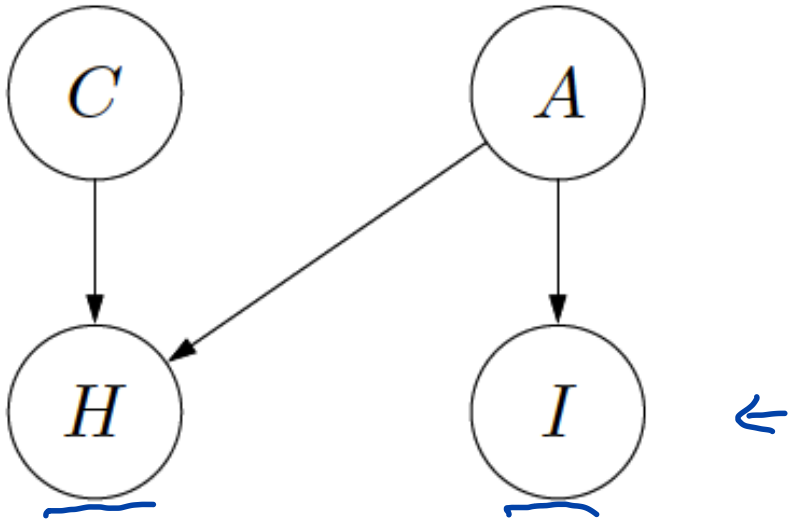
Explaining away

Suppose two causes positively influence an effect. Conditioned on the effect, conditioning on one cause reduces the probability of the other cause.



Example: cold or allergies

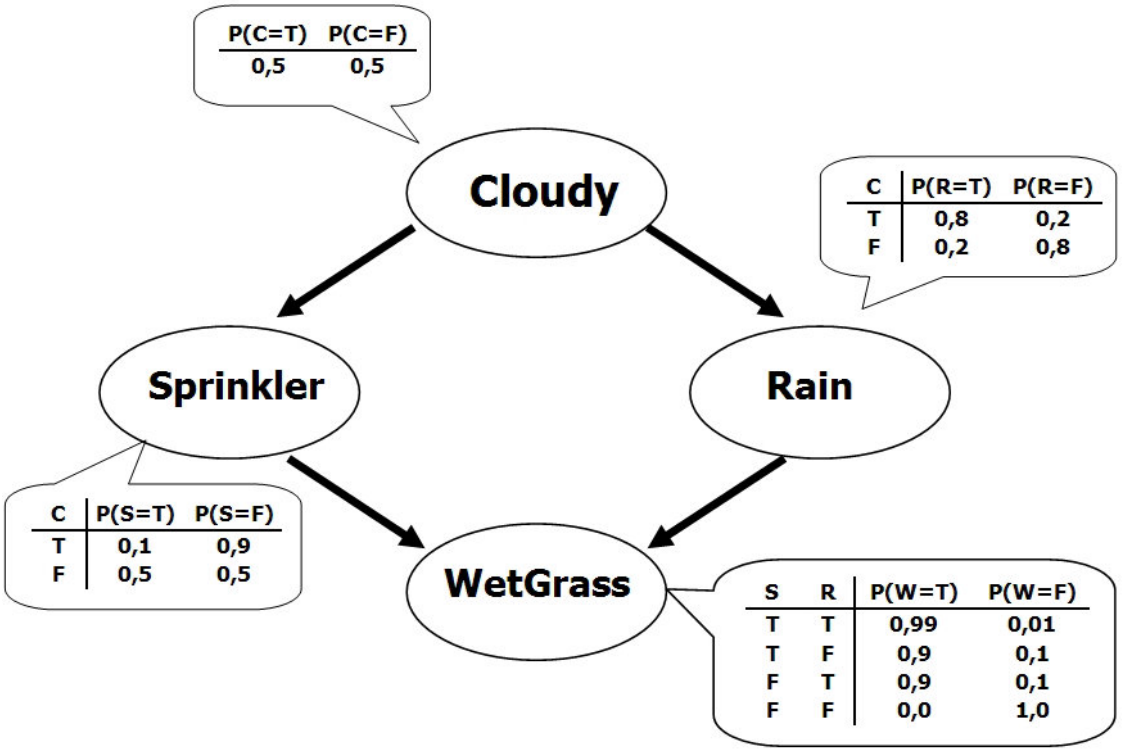
You are coughing and have itchy eyes. What do you have?



Variables: **C**old, **A**llergy, Coug**h**, **I**tchy eyes

Inference ...

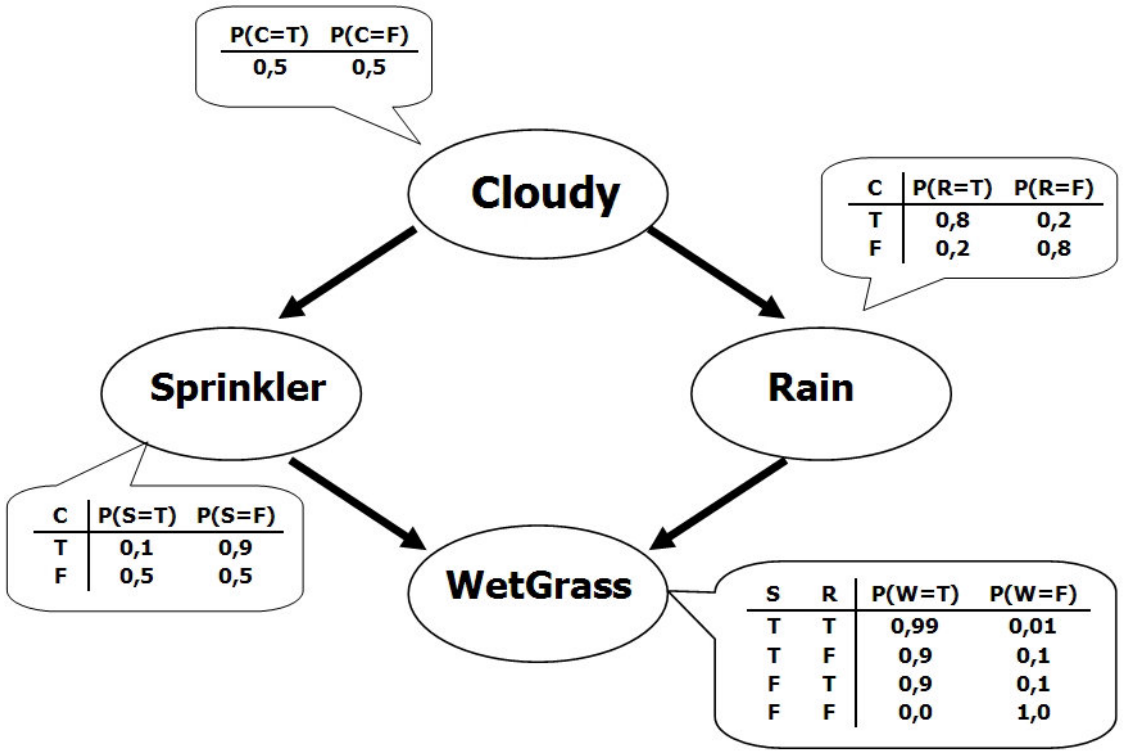
... or how can we use these models to efficiently answer probabilistic queries?



$$P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|S, R)$$

Inference: example

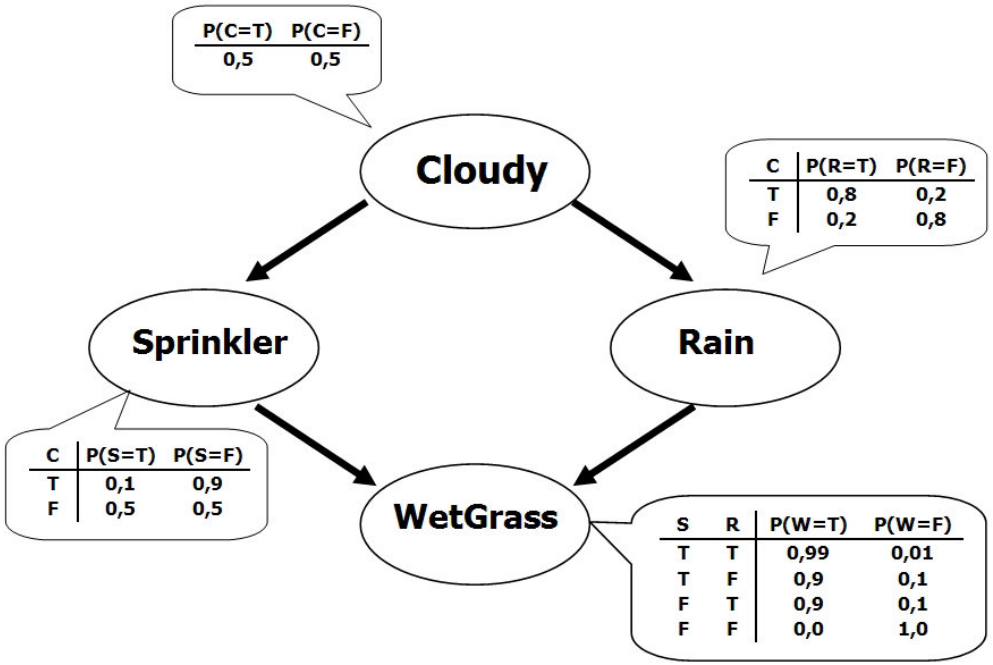
Let us use 0 to denote false and 1 to denote true.



What is the marginal probability, P(S=1), that the sprinkler is on?

Inference: example

What is the marginal probability, P(S=1), that the sprinkler is on?



$$P(S=1) = \sum_C \sum_R \sum_W P(C, R, W, S=1)$$

$$= \sum_C \sum_R \sum_W P(C)P(S=1|C)P(R|C)P(W|S=1,R)$$

Inference: brute force

What is the marginal probability, P(S=1), that the sprinkler is on?

$$P(S=1) = \sum_C \sum_R \sum_W P(W|S=1,R)P(S=1|C)P(R|C)P(C)$$

$$= P(W=0|S=1,R=0)P(S=1|C=0)P(R=0|C=0)P(C=0)$$
$$+P(W=1|S=1,R=0)P(S=1|C=0)P(R=0|C=0)P(C=0)$$
$$+P(W=0|S=1,R=1)P(S=1|C=0)P(R=1|C=0)P(C=0)$$
$$+\ldots$$

C R W
0 0 0
0 0 1
0 1 0

Inference: brute force

What is the marginal probability, P(S=1), that the sprinkler is on?

Pseudocode

Prod = 0
For C=0:1:1
  For R=0:1:1
    For W=0:1:1
      $Prod = Prod + P(C)P(R|C)P(S=1|C)P(W|S=1,R)$
    end
  end
end

Inference: variable elimination

What is the marginal probability, P(S=1), that the sprinkler is on?

$$P(S=1) = \sum_C \sum_R \sum_W P(W|S=1,R)P(S=1|C)P(R|C)P(C)$$

$$= \sum_C \sum_R P(S=1|C)P(R|C)P(C) \sum_W P(W|S=1,R)$$

$$= \sum_C \sum_R P(S=1|C)P(R|C)P(C) \cdot 1$$

$$= \sum_C P(S=1|C)P(C) \sum_R P(R|C)$$

$$= \sum_C P(S=1|C)P(C) \cdot 1$$

$$= P(S=1|C=0)P(C=0) + P(S=1|C=1)P(C=1) = 0.3$$

Inference: variable elimination

What is the marginal probability, P(S=1), that the sprinkler is on?

Pseudocode

$\Psi = 0$
$\Phi = 0$
$\Theta = 0$
For W=0:1:1
  $\Phi = \Phi + P(W|S=1,R)$
end
For R=0:1:1
  $\Psi = \Psi + P(R|C)\Phi$
end
For C=0:1:1
  $\Theta = \Theta + P(S=1|C)\Psi$
end

Inference: Dynamic programming and Approximation algorithms

If we wish to compute several marginals at the same time, we can use Dynamic Programming: **avoid redundant computation**.

Exact algorithms work well for small graphs and for graphs that are trees or close to trees (have low tree-width). For large densely connected graphs we require the use of algorithms beyond the scope of this course:

- *Variational methods*.
- Sampling (Monte Carlo) methods: *Gibbs sampling*.

Question

What is the posterior probability, $P(S = 1|W = 1)$, that the sprinkler is on given that the grass is wet?

Question

What is the posterior probability, $P(S = 1|W = 1)$, that the sprinkler is on given that the grass is wet?

$$P(S = 1|W = 1) = \frac{P(S = 1, W = 1)}{P(W = 1)}$$

$$P(W = 1) = \sum_R \sum_C \sum_S P(W = 1, S, C, R)$$

$$P(S = 1, W = 1) = \sum_R \sum_C P(W = 1, S = 1, C, R)$$

Question

What is the posterior probability, $P(S = 1|W = 1, R = 1)$, that the sprinkler is on given that the grass is wet and it is raining?

Question

What is the posterior probability, $P(S = 1|W = 1, R = 1)$, that the sprinkler is on given that the grass is wet and it is raining?

$$P(S = 1|W = 1, R = 1) = \frac{P(S = 1, W = 1, R = 1)}{P(W = 1, R = 1)}$$

$$P(W = 1, R = 1) = \sum_S \sum_C P(S, W = 1, R = 1, C)$$

$$P(S = 1, W = 1, R = 1) = \sum_C P(S = 1, W = 1, R = 1, C)$$

Summary

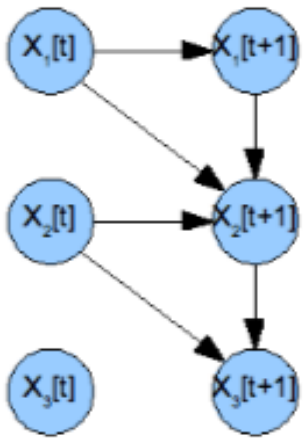
- Bayesian networks encode a joint distribution over the variables.
- We can calculate different probabilities by marginalising over the joint distribution.
- More sophisticated methods are needed for large graphs.

Temporal models

Dynamic Bayesian Networks (DBNs) are directed graphical models of stochastic processes

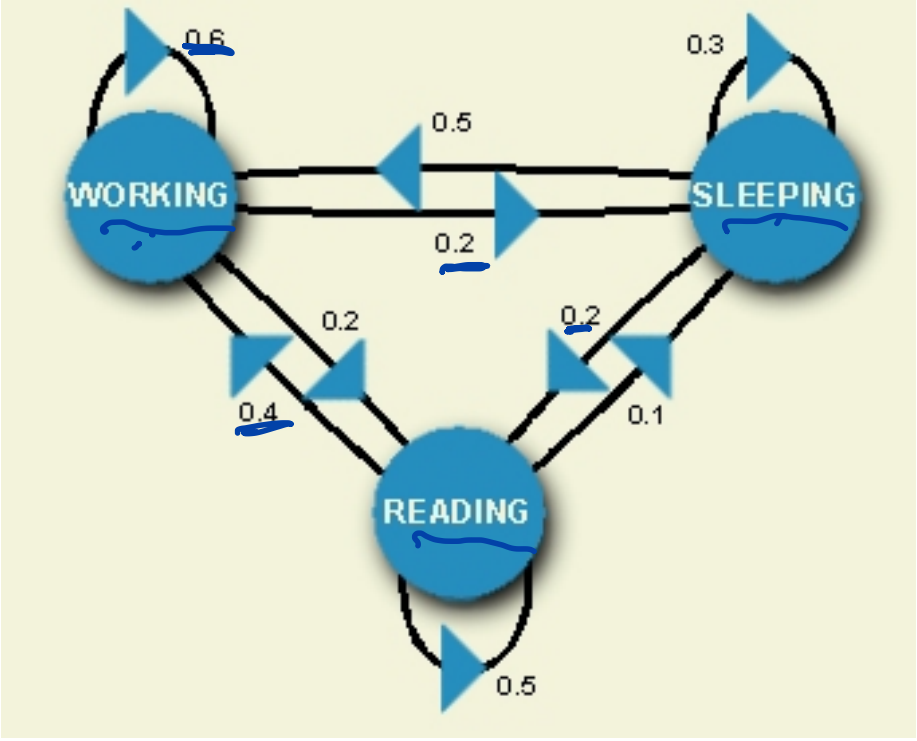
Dynamic Bayesian Networks...

1. Extend BNs to exploit temporal redundancy
   - Markov assumption: future independent of past given present
   - Stationarity: transition model is stable over time
2. State described via random variables
3. Each variable depends only on few other



Markov chain

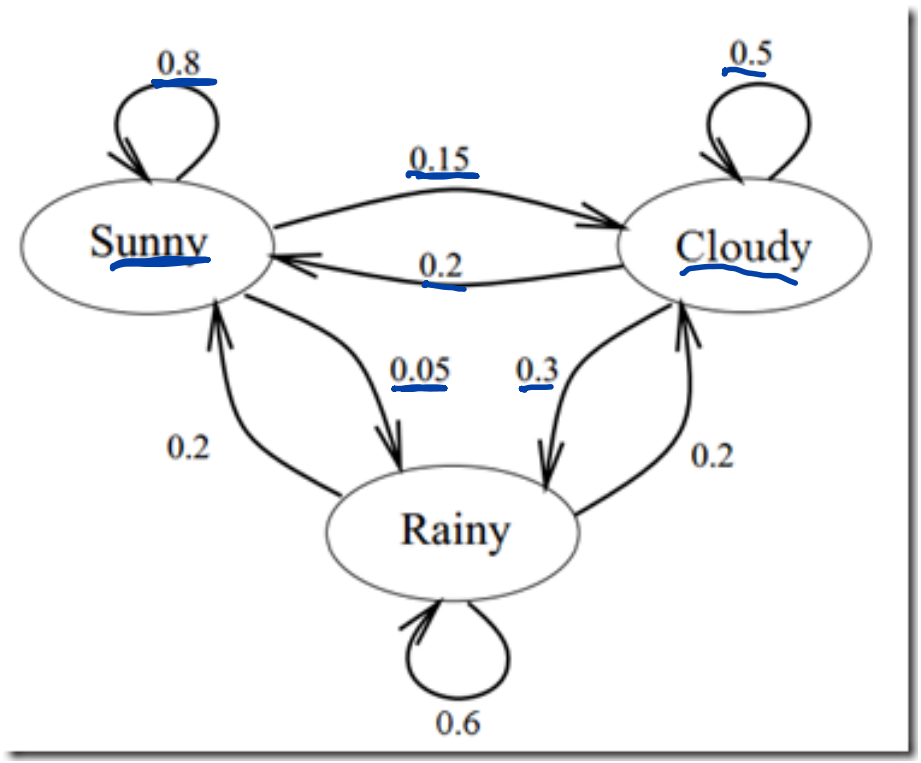
Markov chain: sequence of random variables each conditionally dependent only on the previous.



$$P(W, R, S) = P(W)P(R|W)P(S|R)$$

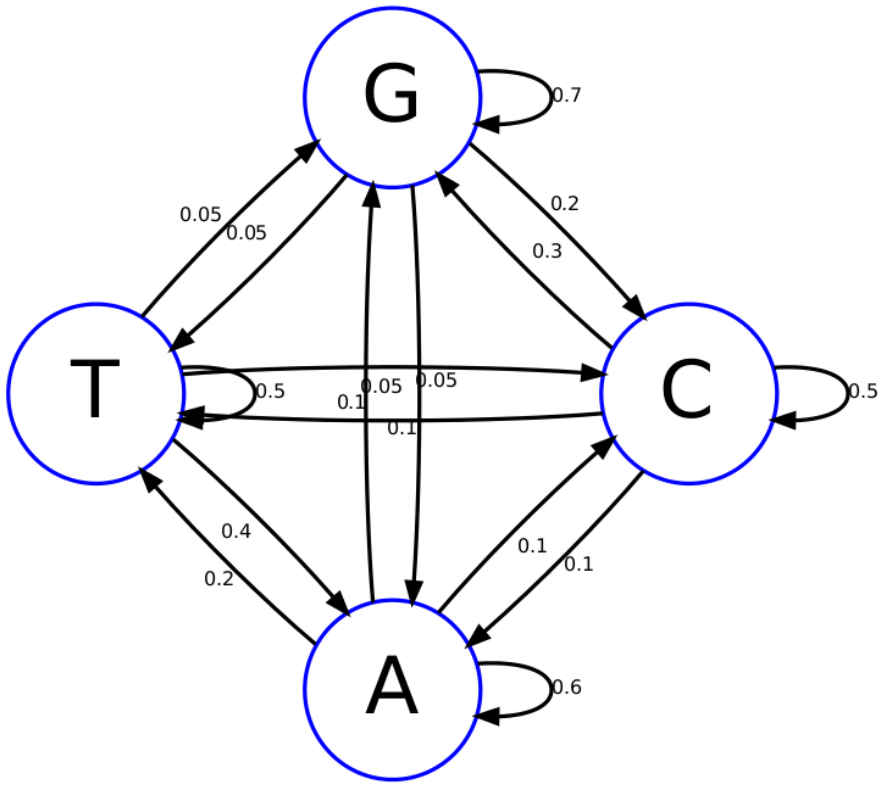
Markov chain

Markov chain: sequence of random variables each conditionally dependent only on the previous.



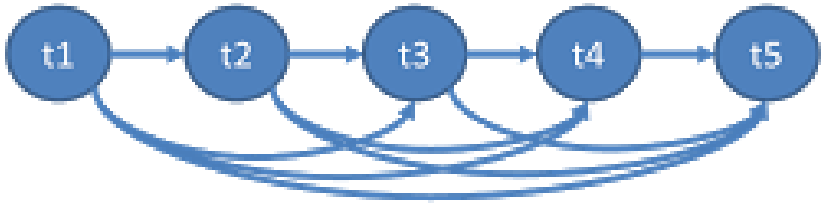
Markov chain

Markov chain: sequence of random variables each conditionally dependent only on the previous.



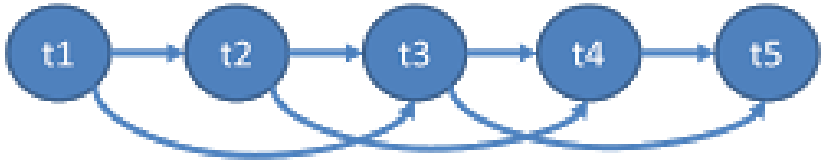
Markov chain

Fully joint distribution :
Probability distribution of a state depends on ALL preceding states



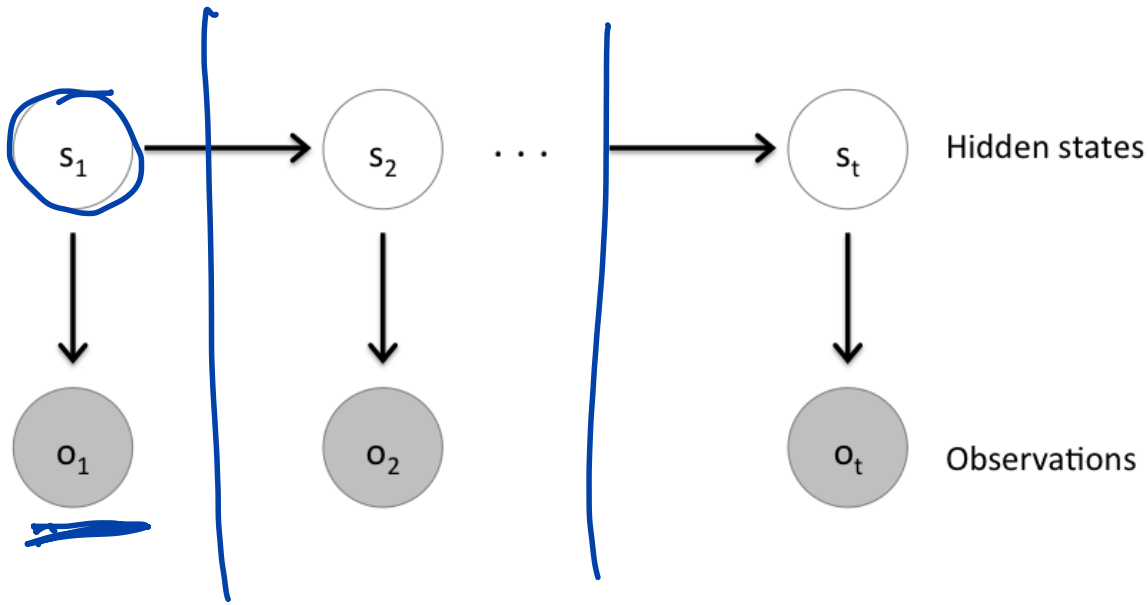
First order Markov Chain :
Probability distribution of a state depends ONLY on 1 immediate preceding state

Second order Markov Chain :
Probability distribution of a state depends ONLY on 2 immediate preceding states



Hidden Markov Models (HMMs)

The simplest kind of DBN is a Hidden Markov Model (HMM), which has one discrete hidden node and one discrete or continuous observed node per slice.



In general, we assume we have an initial distribution $P(s_1)$, a transition model $P(s_t|s_{t-1})$, and an observation model $P(o_t|s_t)$.

HMM: speech

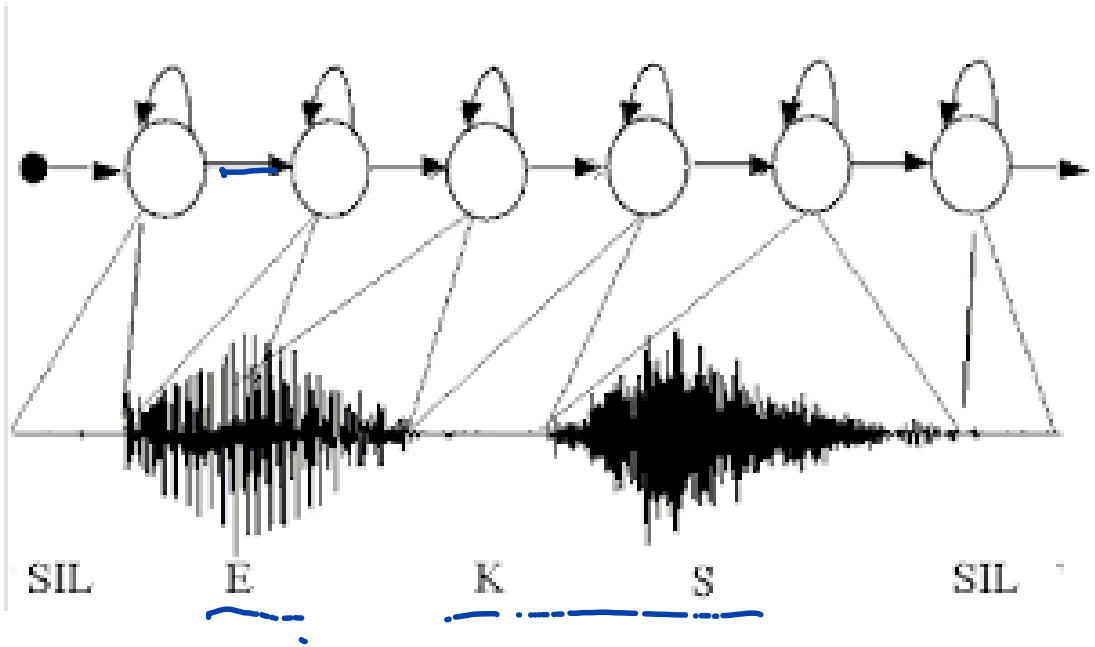
$$P(\text{words} \mid \text{sound}) \;\alpha\; P(\text{sound} \mid \text{words})\, P(\text{words})$$

Final beliefs

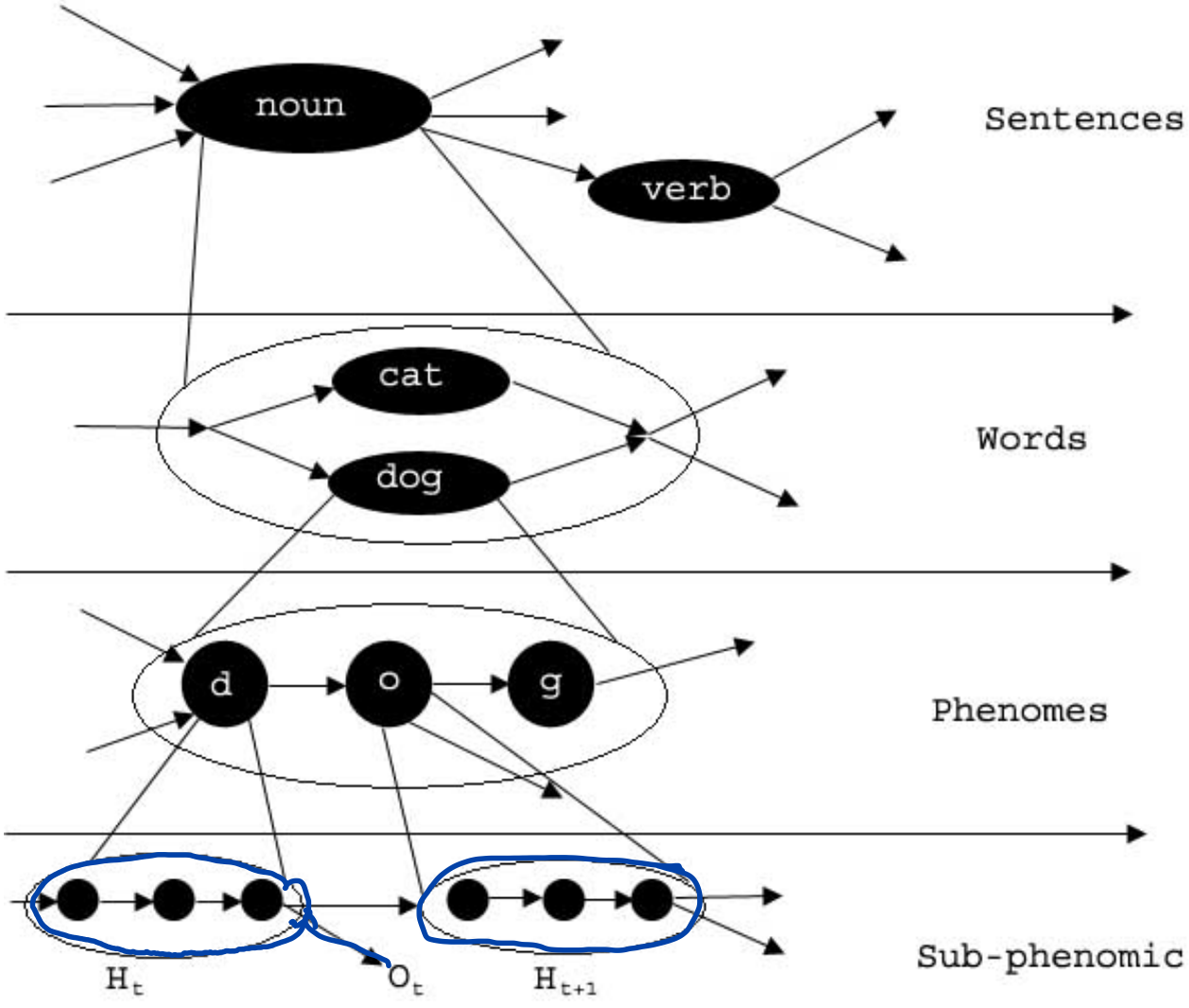
Likelihood of data
eg mixture of Gaussians

Prior language model
eg unigrams

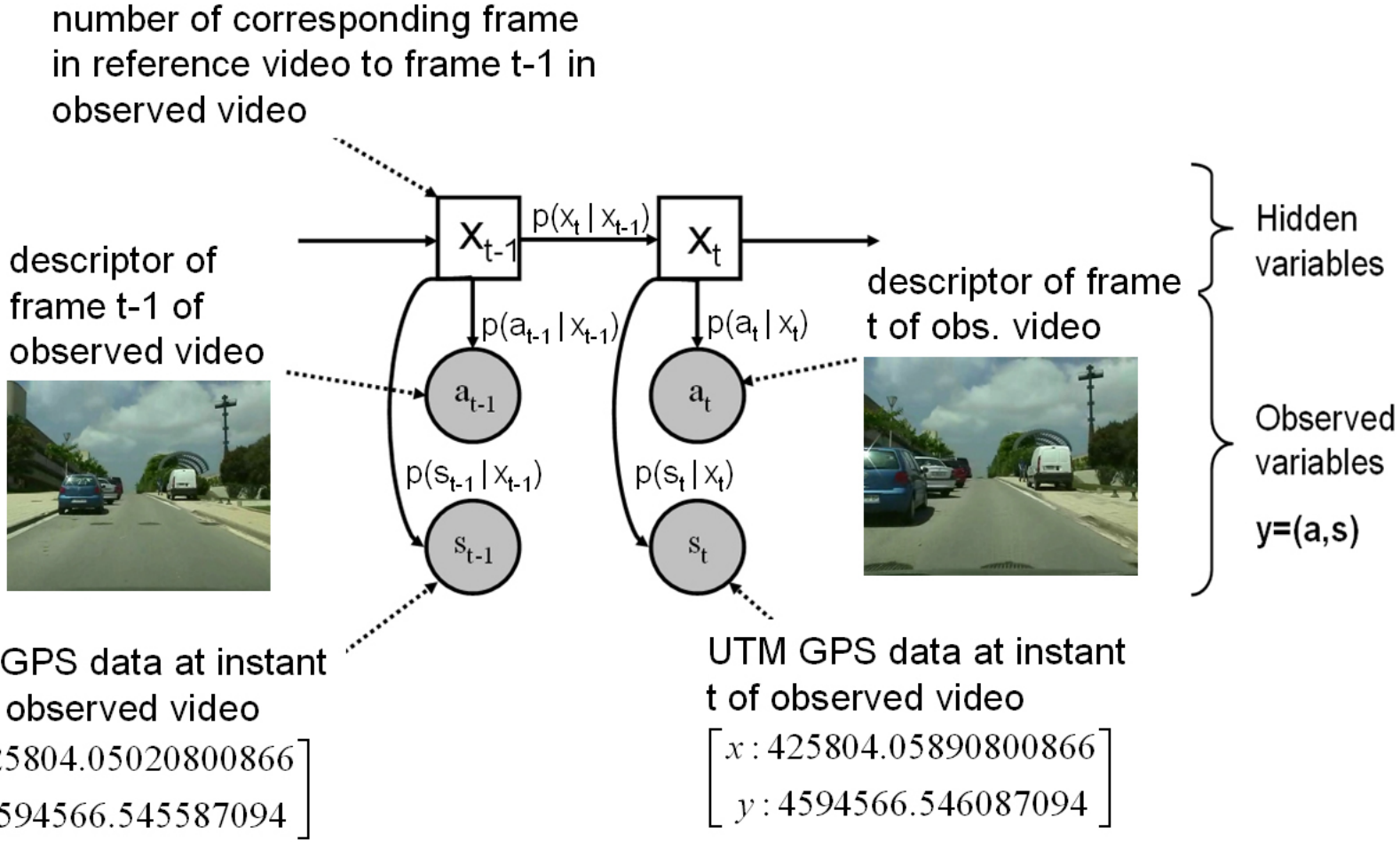
Hidden Markov Model (HMM)



HMM: language

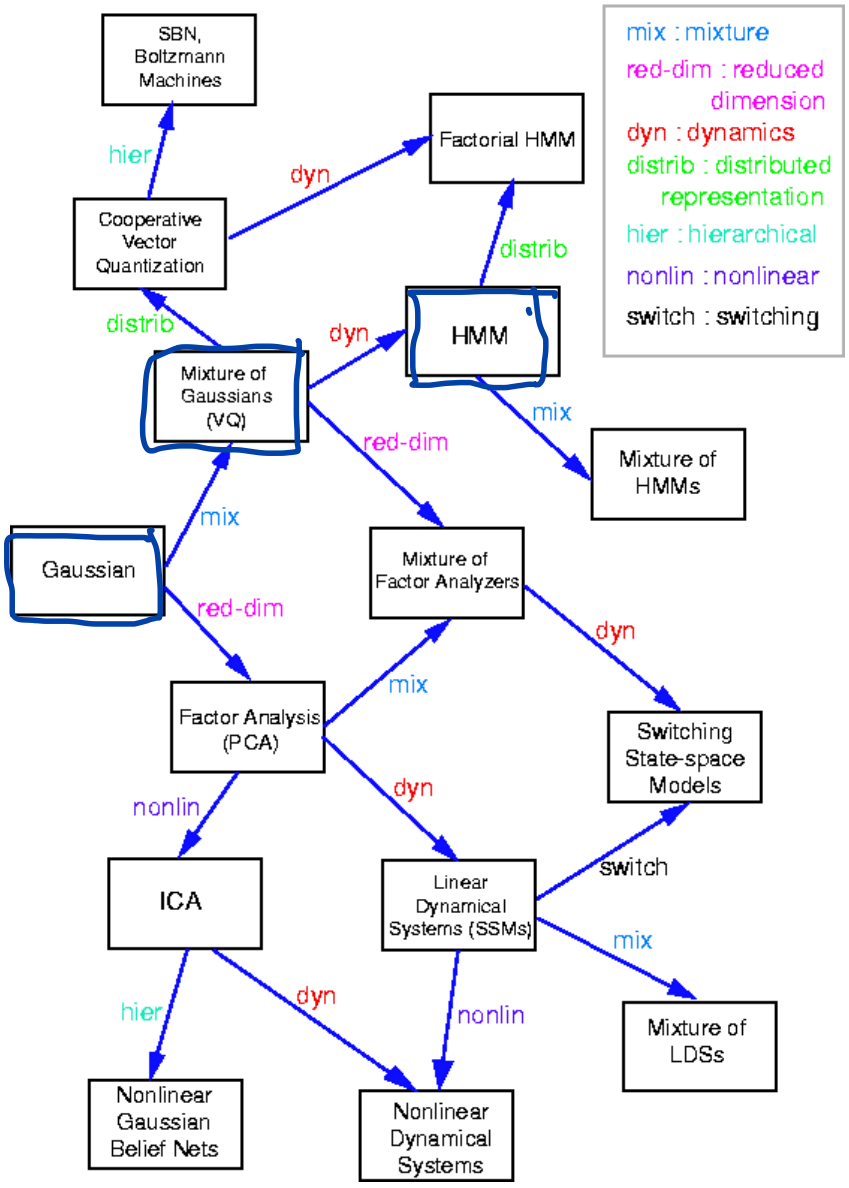


HMM: tracking



[Ferran Diego 2010]

Relationships between graphical models



[Zoubin Ghahramani and Sam Roweis]

Summary

- Bayesian networks can be modified to include temporal information
- Markov chains include temporal information up to a given point
- Hidden Markov models allow us to infer the presence of hidden variables based on the output they produce