The lecture will begin at 08:15.

Please use the chat to ask questions, disable your camera and mute your microphones ☺

DD2380

Artificial Intelligence

Module: Taming uncertainty

André Pereira

DD2380

Artificial Intelligence

Probabilistic reasoning

Credits

- Original slides from Patric Jensfelt, KTH
- Based partly on materials from
  - http://ai.berkeley.edu
  - Kevin Murphy, MIT, UBC, Google
  - Danica Kragic, KTH
  - W. Burgard, C. Stachniss, M. Benewitz and K. Arras, when at Albert-Ludwigs-Universität Freiburg

Reading instructions

- Chapters 13-15, Russel & Norvig
- Additional material on Canvas for the HMM part

Motivation: Why use probabilities

- An artificial system (or human) does not have perfect knowledge of its environment and of the results of its actions, and it needs to deal with uncertainty at many levels

- Uncertainty plays an important role in: sensor interpretation, sensor fusion, map making, path planning, self-localization, control, etc.

Motivational examples

- Diagnose diseases
  - Doctor knows
    - How common a certain disease is
    - Connection with factors such as age, sex, habits, …
    - Connection with measures, e.g. temperature
  - Observe
  - Diagnose

Motivational examples

- Autonomous car: Cross intersection safely?
  - From manufacturer or learned by car
    - Sensor models
    - Statistic from different roads
    - Weather models
    - …
  - Observations from own car and others?
  - Q:
    - Can I cross if I want to be 99.99999% safe?
    - Can I cross if I want to be 99% safe?

Recap of probability theory notation

- Probability of event X: **p(X)**
- Joint probability of X and Y: **p(X,Y)**, i.e. X AND Y
- Conditional probability of X given Y: **p(X|Y)**

More notation

- For boolean variables we will use
  - p(X) to mean the probability that X is true and
  - p($\neg$X) the probability that X is false.

Rules of probability

- $p(X) \in [0,1]$ (i.e. $0 \leq p(X) \leq 1$)
- $p(X) = 1 - p(\neg X)$
- Prob sum to one: $1 = \sum_{\text{all } X} p(X)$
- Product rule: $p(X,Y) = p(Y|X)p(X)$
- Sum rule (marginalization): $p(X) = \sum_{Y} p(X,Y)$

Sum rule (marginalization)

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$p(T) = \sum_W p(T,W)$$

$$p(W) = \sum_T p(T,W)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

In general: $p(X) = \sum_Y p(X,Y)$

Conditioning

- A variant of the sum rule (marginalization) involving conditional probabilities instead of joint probabilities

- Combining
  - $p(X)=\sum_{Y} p(X,Y)$ (sum rule)
  - $p(X,Y)=p(X|Y)p(Y)$ (product rule)

  gives

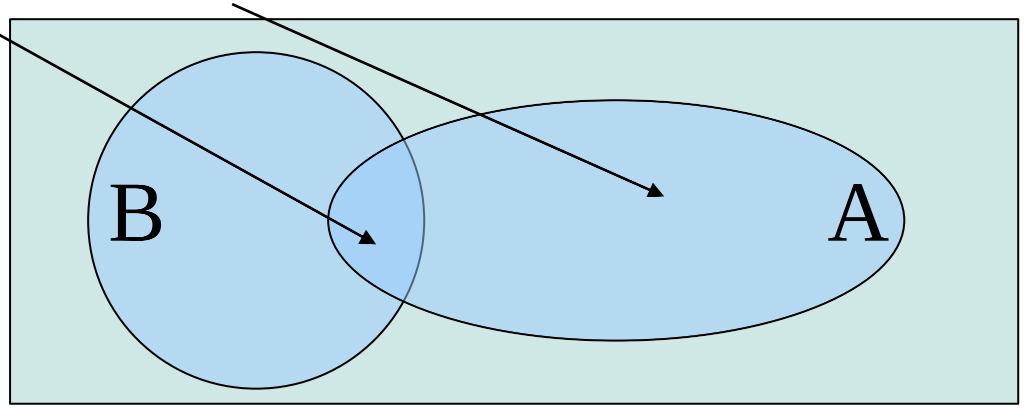
  - $\mathbf{p(X)=\sum_{Y} p(X|Y)p(Y)}$

Conditional probability

- Consider two dependent events A and B

		A	
		True	False
B	True	0.1	0.3
	False	0.4	0.2

- Probability of B given that A is true
  p(B|A) = p(A,B) / P(A)



Conditional probability (weather example)

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s | T = c) = ???$$

Conditional probability (weather example)

$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

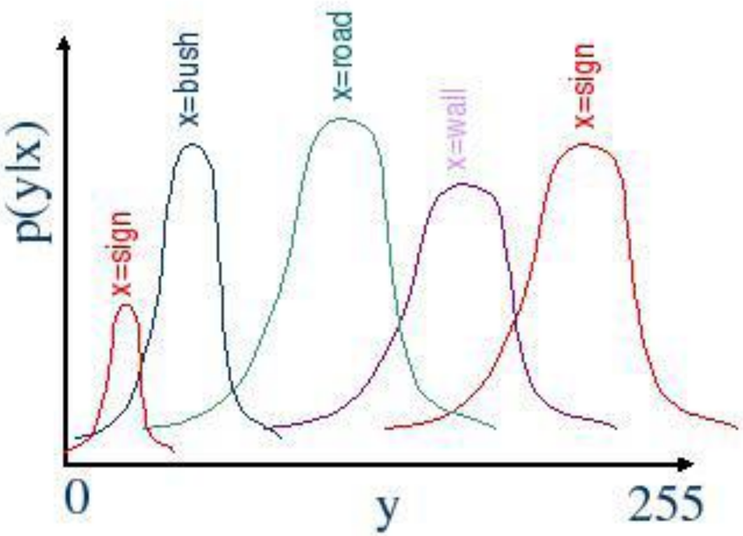
$$P(W=s|T=c)= ???$$

$$P(W=s|T=c)=\frac{P(W=s,T=c)}{P(T=c)}=\frac{0.2}{0.5}=0.4$$

$$=P(W=s,T=c)+P(W=r,T=c)$$
$$=0.2+0.3\;=0.5$$

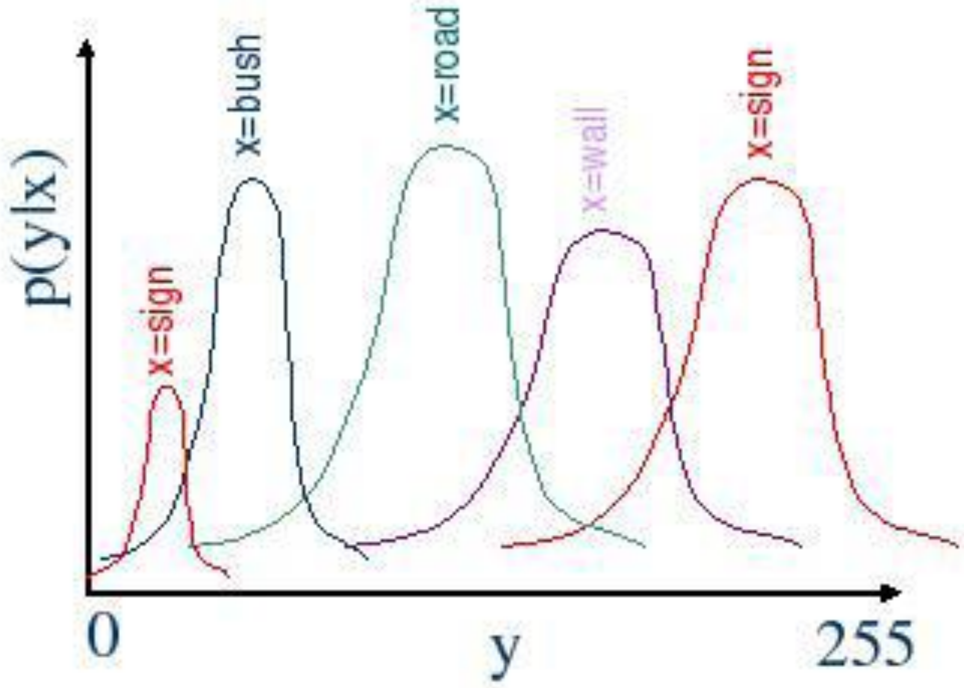
Example conditional dependence

- Knowing what we look at gives a much better idea of what to expect to measure.
- Ex: If we know that we look for a sign we expect either bright (near 255) for the white background or dark (near 0) for pixel values for the text



Simplistic probabilistic reasoning

- If we measure 115, what is the most likely category?



Probabilistic inference

Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

We generally compute conditional probabilities

P(on time | no reported accidents) = 0.90

These represent the agent's *beliefs* given the evidence

Probabilities change with new evidence:

P(on time | no accidents, 5 a.m.) = 0.95

P(on time | no accidents, 5 a.m., raining) = 0.80

Observing new evidence causes *beliefs to be updated*

Bayes rule

- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later (e.g. speech recognition)
- In the running for most important AI equation!

Bayes rule

- $p(Y|X) = p(X|Y)\, p(Y) / p(X)$

- What is the probability of Y given some evidence X can be expressed in **factors that are sometimes easier to determine**.

$$posterior = \frac{likelihood \times prior}{probability\ of\ evidence}$$

Bayes rule cont'd

- Rewrite $p(Y|X) = p(X|Y)\, p(Y) / p(X)$
- using the sum rule $p(X) = \sum_{Y} p(X|Y)p(Y)$

$$p(Y|X) = p(X|Y)\, p(Y) / \sum_{Y} p(X|Y)p(Y)$$

- $p(X) = \sum_{Y} p(X|Y)p(Y)$ is a normalization factor $1/\eta$ $\rightarrow$

$$p(Y|X) = \eta\, p(X|Y)\, p(Y)$$

So what?

- If variables are dependent we can get information about one variable by measuring another!! (Often one conditional is tricky but the other one is simple)
- Foundation for most of the probabilistic reasoning and statistical machine learning

Bayes rule example

- Vision system for detecting zebras Z
- Prior: p(Z)=0.02 (zebra in 2% of images)
- Detector for “stripey areas” gives observations, O
- Detector gives yes/no
- Detector performance
  - $p(O|Z)=0.8$ (true pos)
  - $p(O|\neg Z)=0.1$ (false pos, e.g. gate)
- Task: Calculate $p(Z|O)$
  1. What does it represent?
  2. The result of that probability?

Bayes rules example solution

- p(Z|O) represents probability that there is a zebra if our detector says there is one.

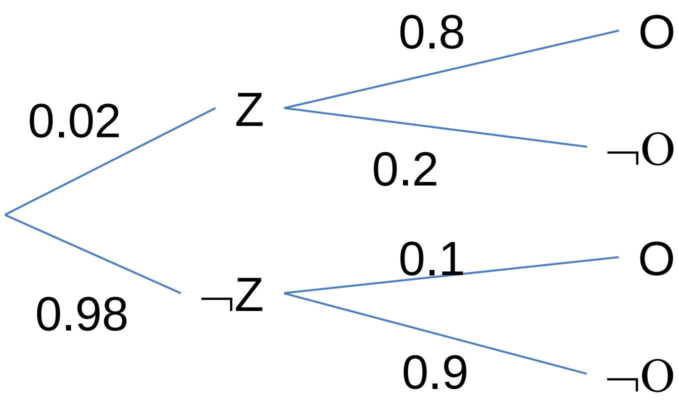
- How big? Use Bayes rule!

$$p(Z|O) = \frac{p(O|Z)p(Z)}{p(O)}$$

Bayes rules example solution

- p(Z)=0.02 (zebra in 2% of images)
- p(O|Z)=0.8 (true pos)
- p(O|¬Z)=0.1 (false pos, e.g. gate)

$$p(Z|O) = \frac{p(O|Z)p(Z)}{p(O)}$$

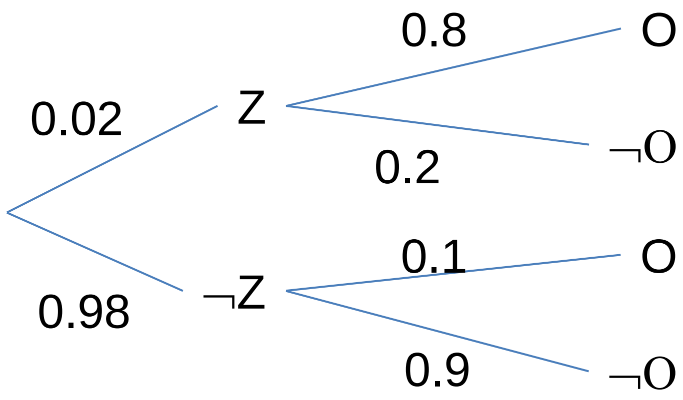


Bayes rules example solution

- p(Z|O) represents probability that there is a zebra if our detector says there is one

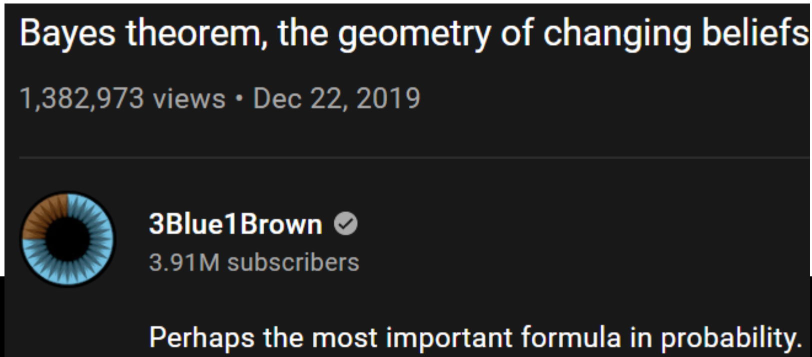
$$
\begin{aligned}
p(Z|O) &= \frac{p(O|Z)p(Z)}{p(O)} \\
&= \frac{p(O|Z)p(Z)}{p(O|Z)p(Z) + p(O|\neg Z)p(\neg Z)} \\
&= \frac{0.8 \cdot 0.02}{0.8 \cdot 0.02 + 0.1 \cdot 0.98} \\
&= \frac{0.016}{0.016 + 0.098} \\
&\approx 0.1404
\end{aligned}
$$

Conditioning



Bayes rule example discussion

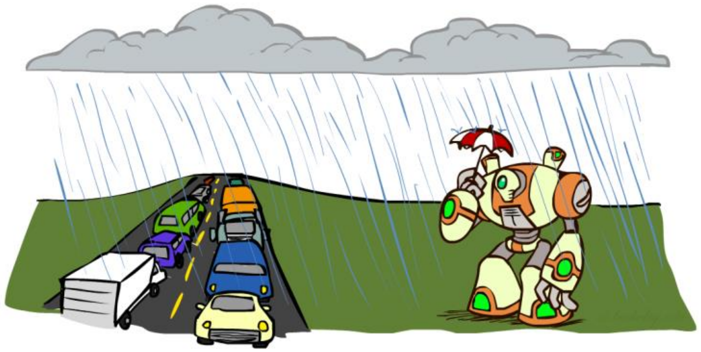
- p(Z|O) = 0.1404
- Intuition tells most people that the detector is much better than this, i.e. we would expect to see a much higher p(Z|O) since the detector is correct in 80% of the cases
- However, only 1 out of 50 images have a zebra
  → 49 out of 50 do not contain a zebra
  + detector not perfect
- Failing to account for negative evidence properly is a typical failing of human intuitive reasoning



Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.

Example:
P (Traffic)
P (Umbrella)
P (Rain)



Rain causes traffic AND people wearing umbrellas.

T is conditionally independent of U given R

Conditional independence

- If X is conditionally independent of Y given Z
- $p(X|Y,Z) = P(X|Z)$

- Which also means

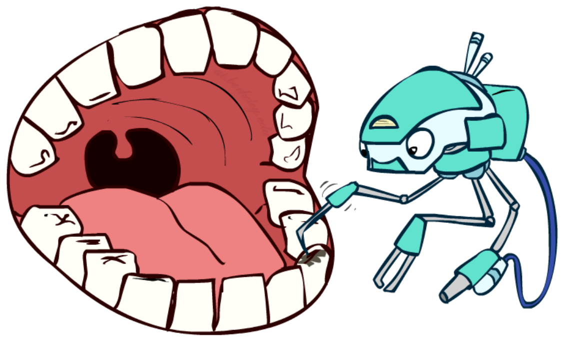
- $p(X,Y|Z)$ = {product rules} =
    - $p(X|Y,Z)p(Y|Z)$ = {conditional independence}
    - $p(X|Z)p(Y|Z)$

- NOTE: Not the same as $p(X,Y) = p(X)p(Y)$

Conditional Independence

- P(Toothache, Cavity, Catch)

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache



- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)

- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)

Probability Recap

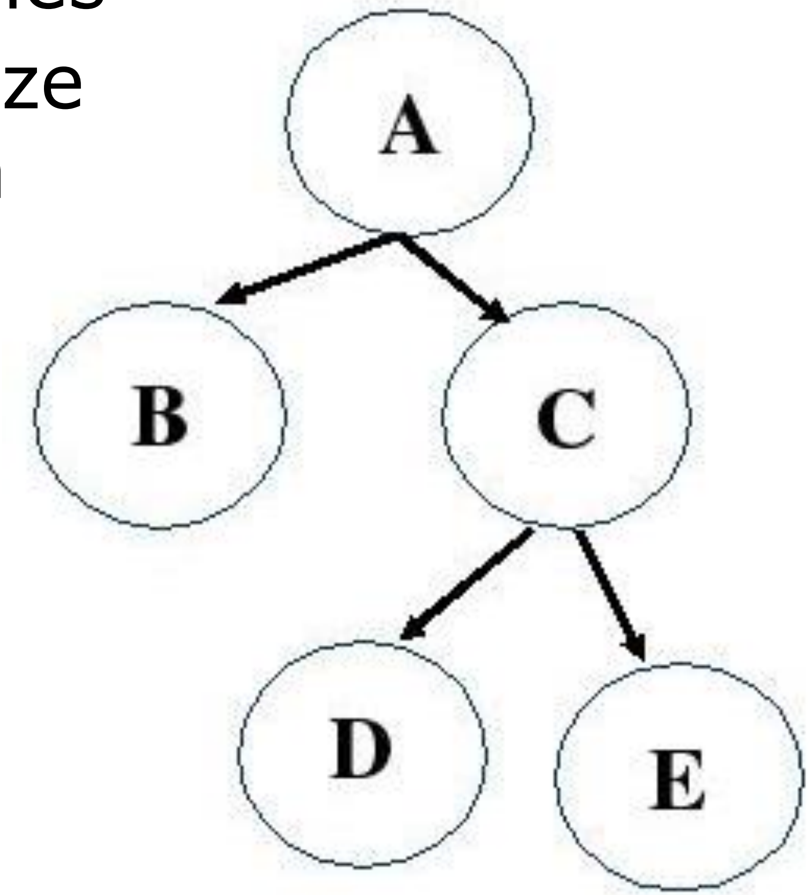
- Conditional probability $$P(x|y) = \frac{P(x,y)}{P(y)}$$
- Product rule $$P(x,y) = P(x|y)P(y)$$
- Chain rule $$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1}) \end{aligned}$$
- X, Y independent if and only if: $\forall x, y : P(x,y) = P(x)P(y)$
- X and Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y | Z$

$$\forall x, y, z : P(x,y|z) = P(x|z)P(y|z)$$

Shifting gear, hang on!

Probabilistic Graphical Models

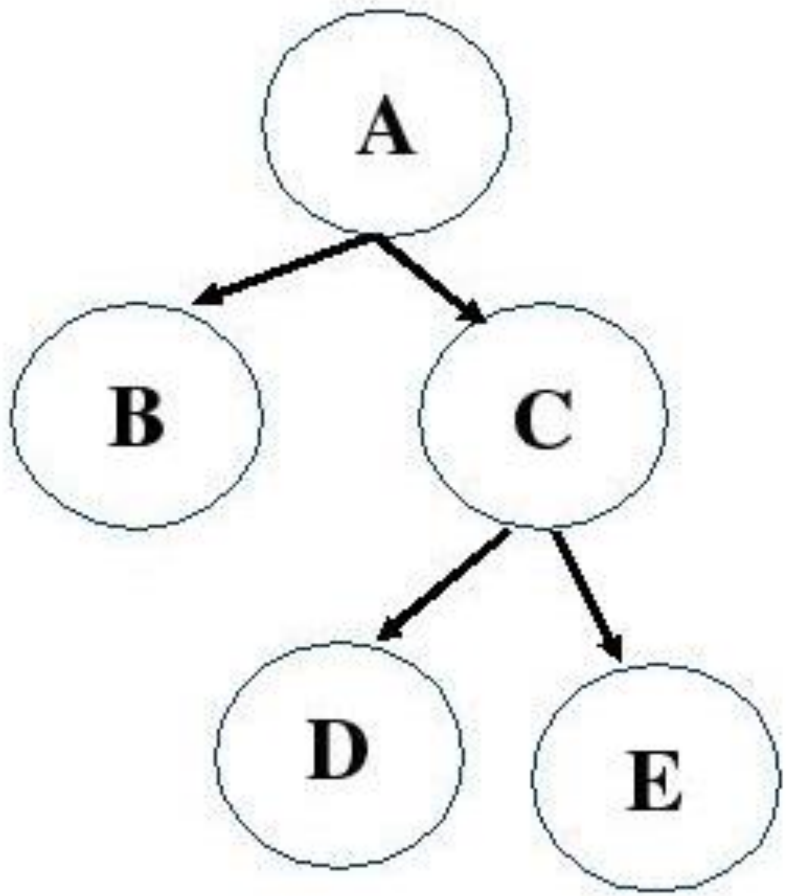
- Compact repr. of the joint distribution over a set of variables
- Graphical repr. that helps analyze and structure prob. information
- Each variable is encoded as a node
- Conditional independence assumptions coded as arcs
- Here: a Bayesian network (directed acyclic graph (DAG))



Bayesian network

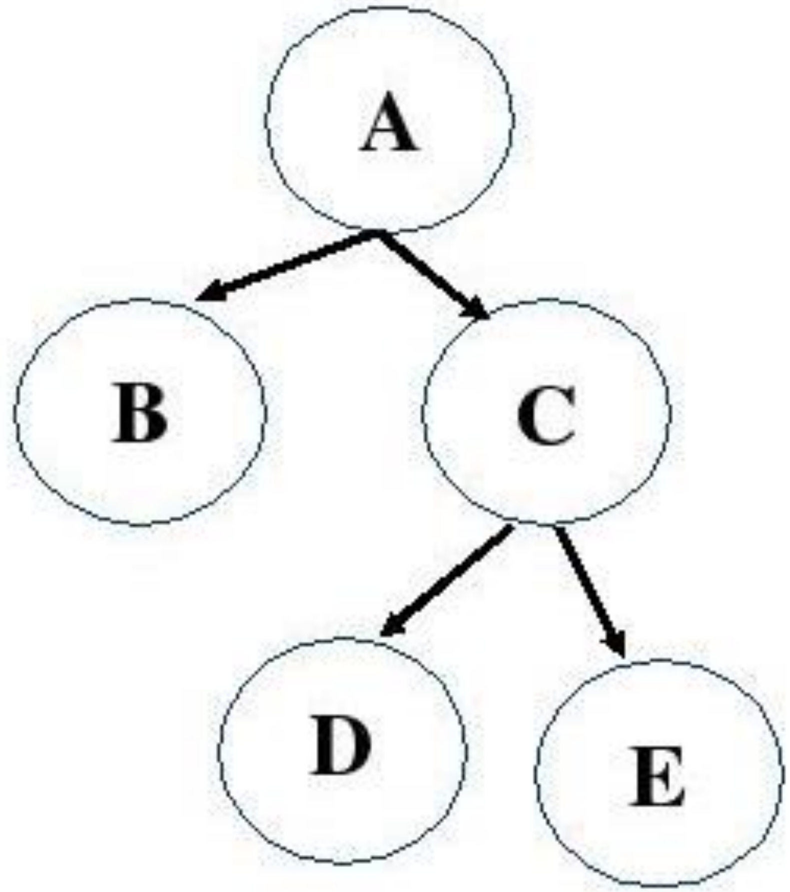
- A root node
- B, D, E leaf nodes
- A parent to B and C
- B and C children of A
- A ancestor of D and E

- A "causes" B and C
  → Value of A influences value of B and C



Bayesian network

- Arrow → "direct influence over"
- A has direct influence over B

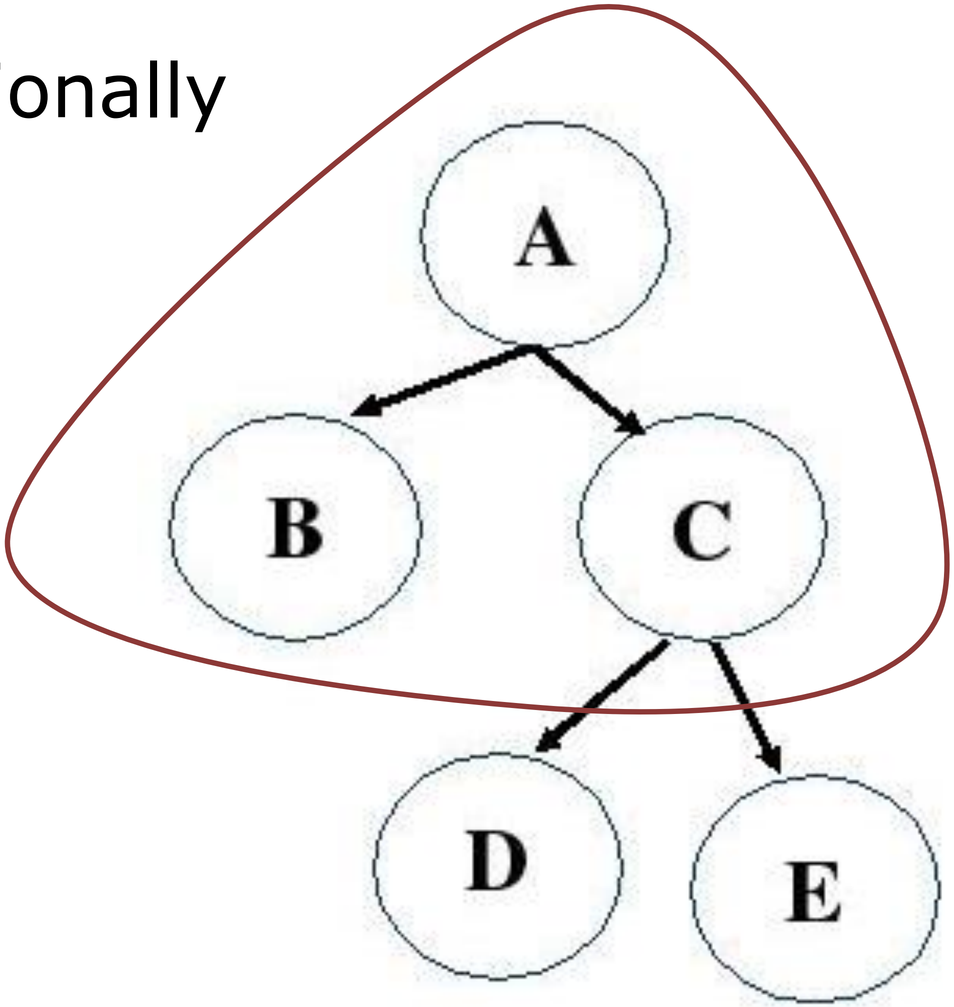


Bayesian network

- B and C are dependent
- HOWEVER, they are conditionally independent given A

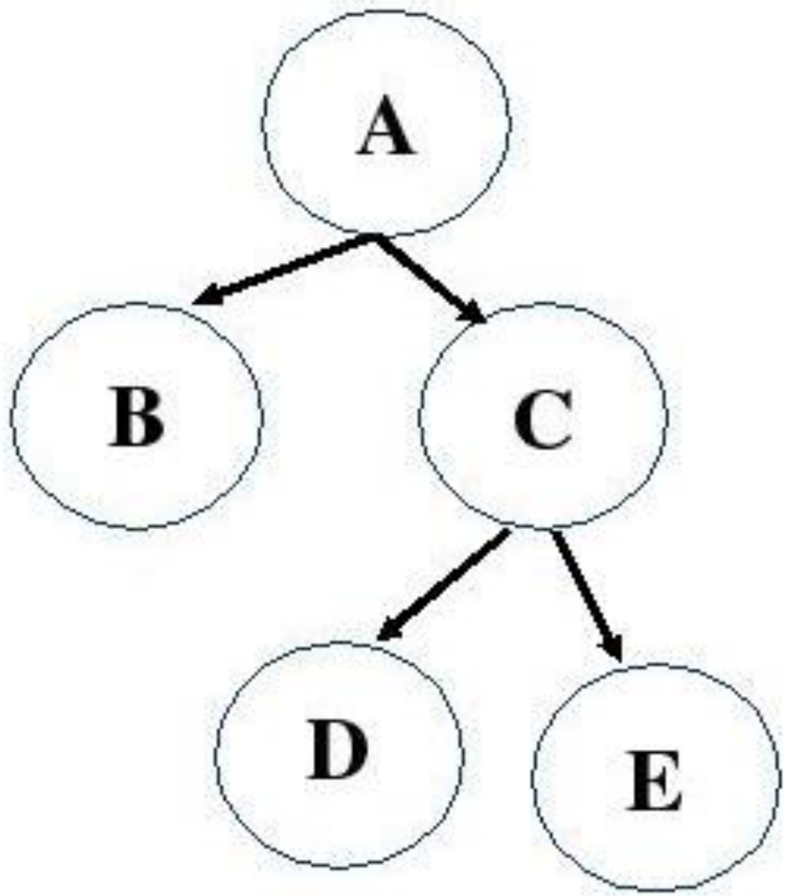
- Q: Write down the formula that relates A, B and C?

P(B,C|A)= ??



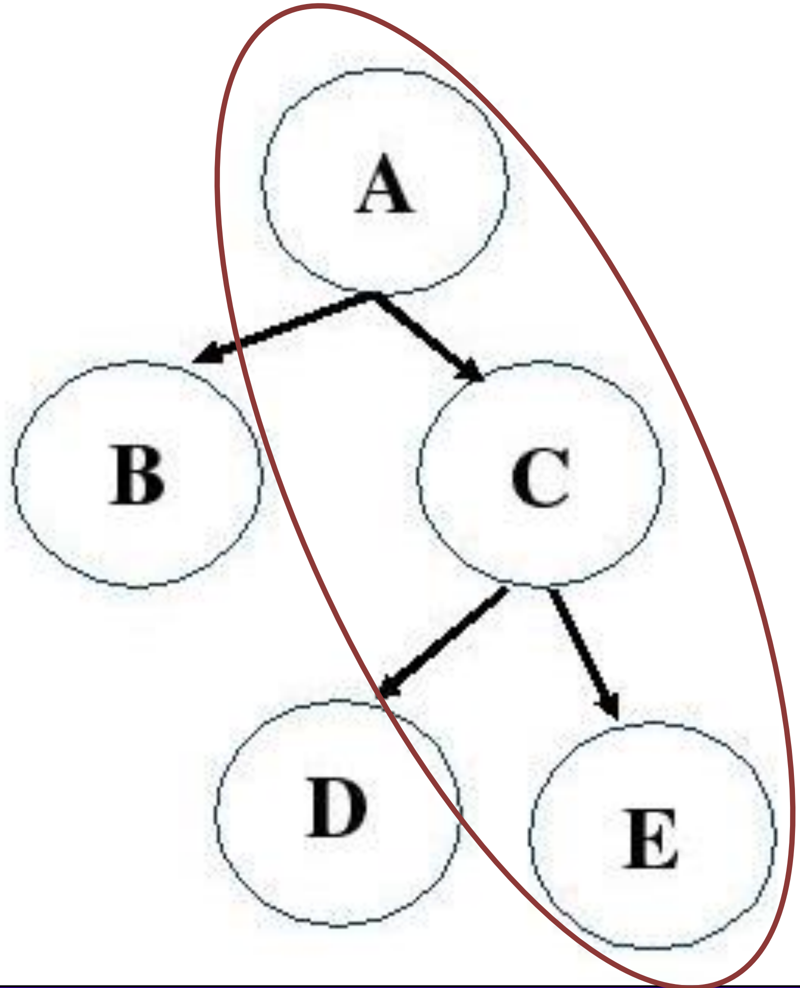
B and C independent given A

- $P(B,C|A)=P(B|A)P(C|A)$
- If we do not know A, knowing something about B will tell us something about C (tells us about A which tells us about C)
- but if we know A then knowing B does not tell us anything more about C than we already knew because of A.



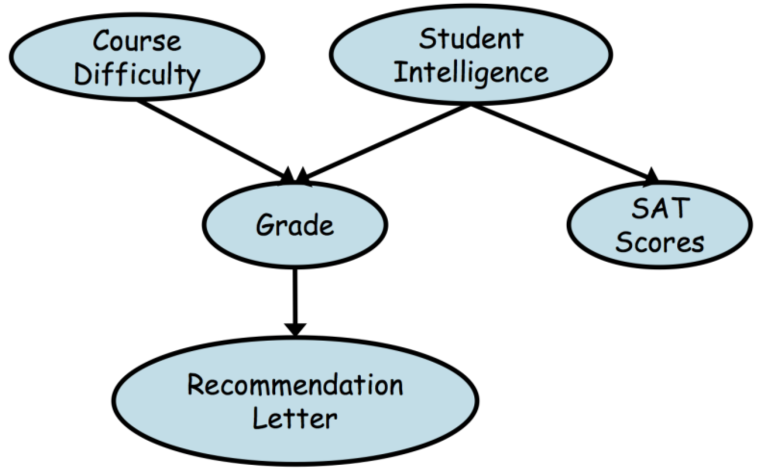
Bayesian network

- C depends on A
- E depends on A and C. **HOWEVER**, E is conditionally independent of A given C.
- That is, C captures all the information in A relevant to determine E.

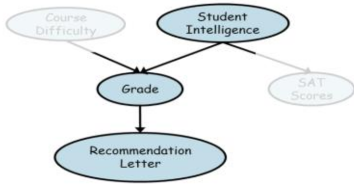


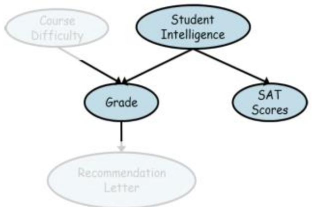
Flow of probabilistic influence

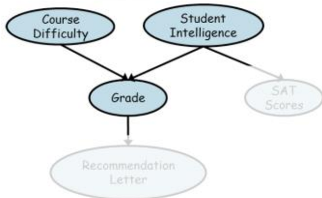
- When can X influence Y?
  - Case1: No evidence about Z
  - Case2: Evidence about Z



Graphical Structure	NO evidence about Z	YES evidence about Z
$X \rightarrow Z \rightarrow Y$	✓	✗
$X \leftarrow Z \rightarrow Y$	✓	✗
$X \rightarrow Z \leftarrow Y$	✗	✓







- Adapted from Daphne Koller: https://www.coursera.org/lecture/probabilistic-graphical-models/flow-of-probabilistic-influence-1eCp1

Joint distribution

- A probabilistic model of a domain must represent the Joint Probability Distribution (JPD), i.e. the probability of every possible event as defined by the combination of the values of all the variables.

- There are exponentially many such events, but Bayesian networks achieve compactness by factoring the JPD into local, **conditional distributions for each variable given its parents**.

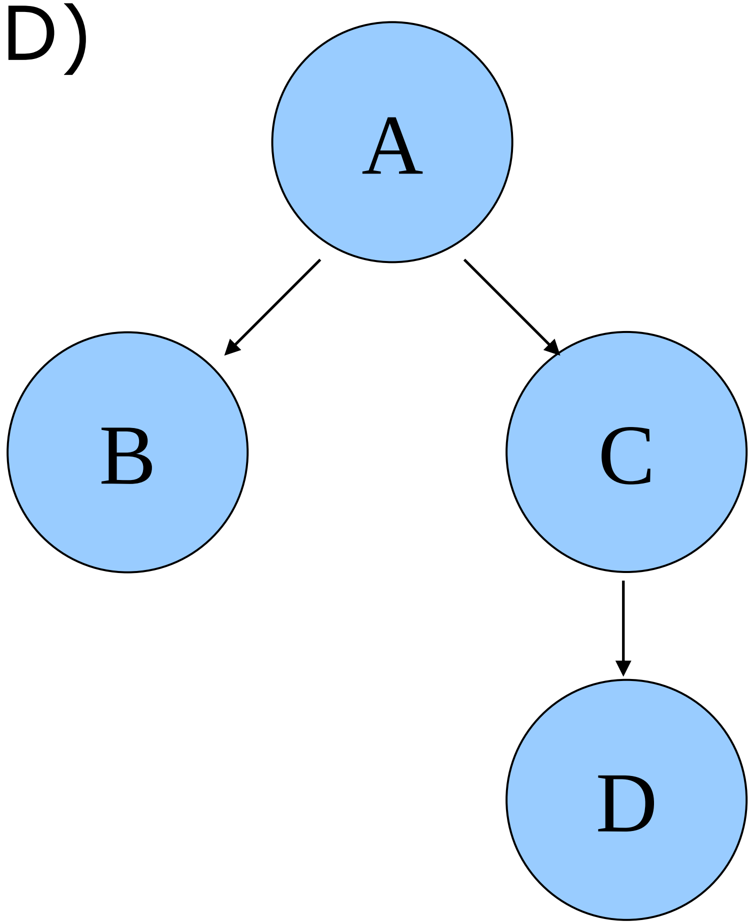
Joint distribution example

- Exercise2: Factorize p(A,B,C,D)

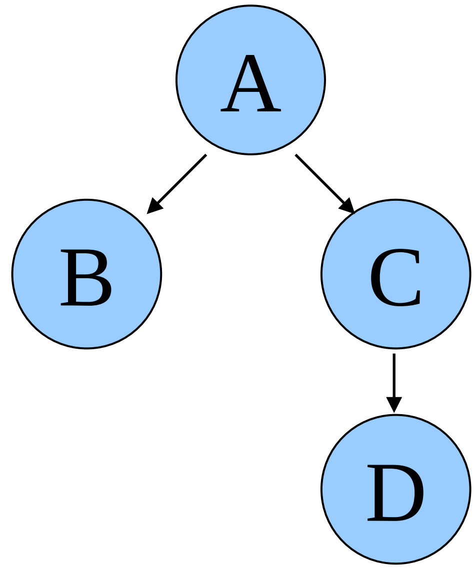
- Remember the chain rule

$$P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$$
$$= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$$

- Tip: Work from the top and factor out A, then B, C and D



Joint distribution example



$$P(A, B, C, D) = P(A)\, P(B|A)\, \underbrace{P(C|A,B)}_{\text{C cond. indep. of B given A}}\, \underbrace{P(D|A,B,C)}_{\text{D cond. indep. of A,B given C}}$$

$$= P(A)\, P(B|A)\, P(C|A)\, P(D|C)$$

- In general

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{i=n} p(X_i|Parents(X_i))$$

Note

- We could have used the chain or product rule in any order
- Looking at the graph we can make use of the conditional independencies
- Other factorization possible, but not as compact

Zebra example

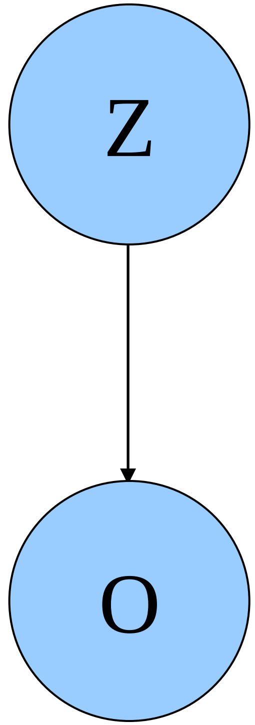
- Vision system for detecting zebras Z
- Prior: p(Z)=0.02 (zebra in 2% of images)
- Detector for "stripey areas" giving observations, O
- Detector gives true/false
- Detector performance
  - $p(O|Z)=0.8$ (true pos)
  - $p(O|\neg Z)=0.1$ (false pos, e.g. gate)

- Q: Draw as Bayesian network

Zebra example

- Vision system for detecting zebras Z
- Prior: p(Z)=0.02 (zebra in 2% of images)
- Detector for "stripey areas" giving observations, O
- Detector gives true/false
- Detector performance
  - $p(O|Z)=0.8$ (true pos)
  - $p(O|\neg Z)=0.1$ (false pos, e.g. gate)

- Q: Draw as Bayesian network

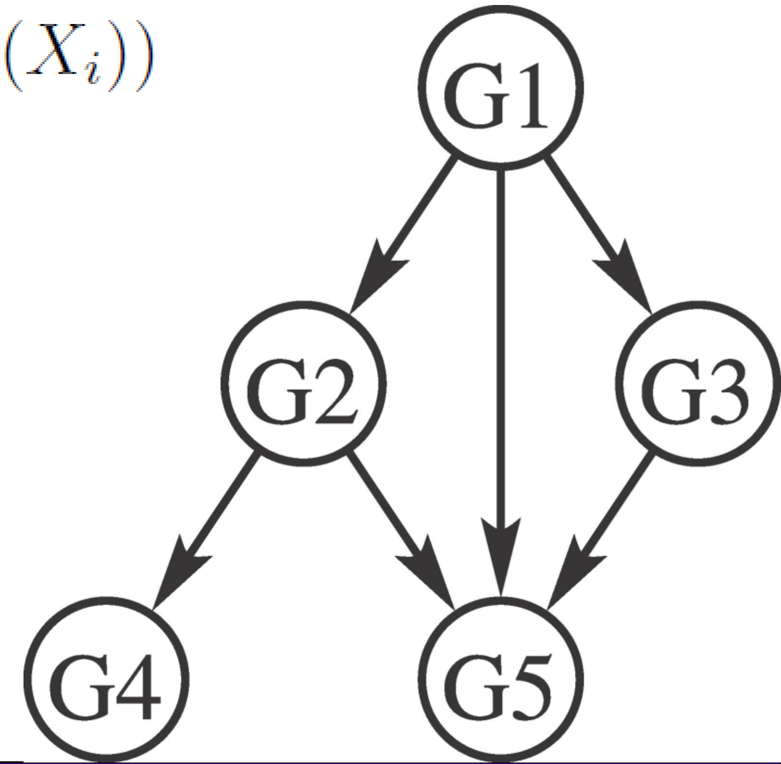


Order? : The existence of the zebra influences the observation and not vice versa

Exercise3: Factorize the graph

- p(G1,G2,G3,G4,G5) = ?
- remember

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{i=n} p(X_i | Parents(X_i))$$



Alarm example

Structure

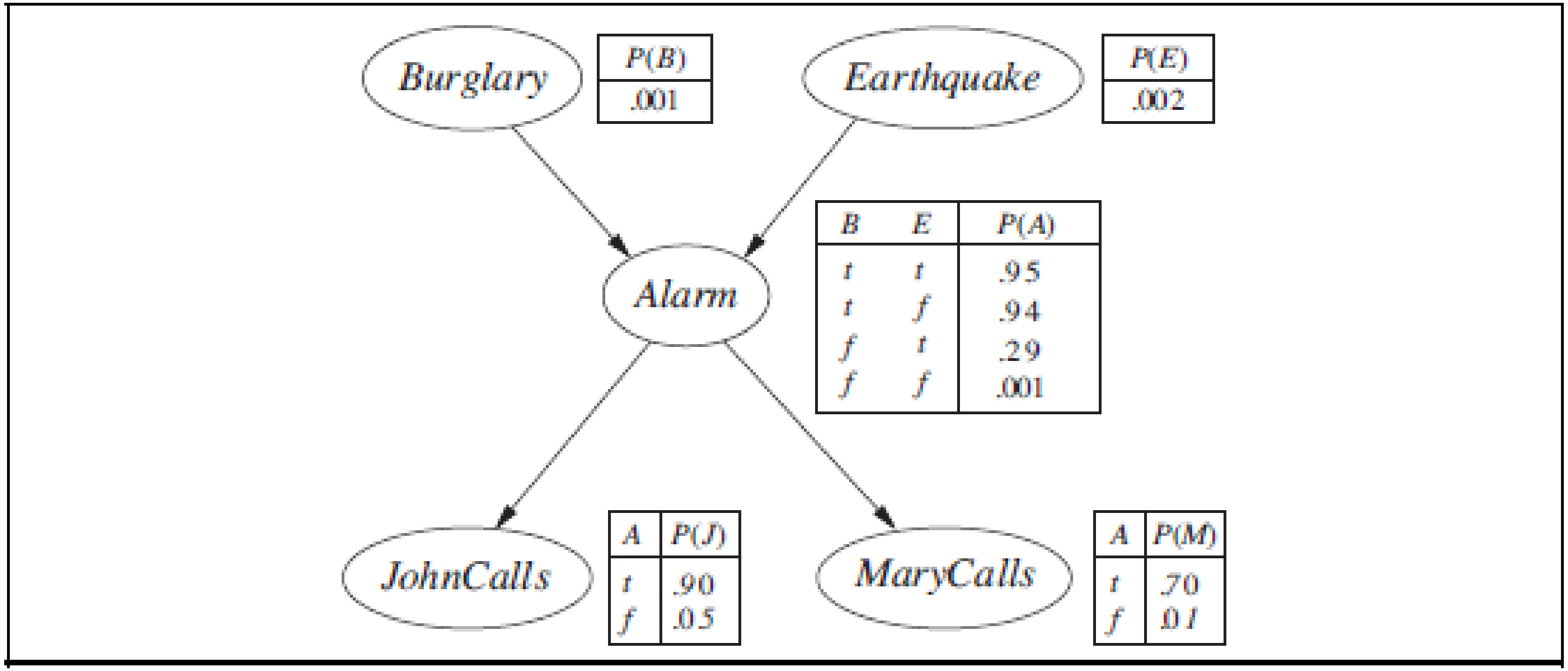
- You have an **alarm**. It reacts reliably to burglaries but is sometimes triggered also by small earthquakes.
- Two neighbors **John** and **Mary** promised to call you at work when they hear the alarm (not **earthquake** or **burglary**!)

Qualitative information about **probabilities**

- John calls almost every time there is an alarm but sometimes confuses it with a phone ringing and calls then too
- Mary plays loud music and sometimes misses it completely but rarely mix other things with it
- Draw Bayesian network! What variables?

Alarm example cont'd

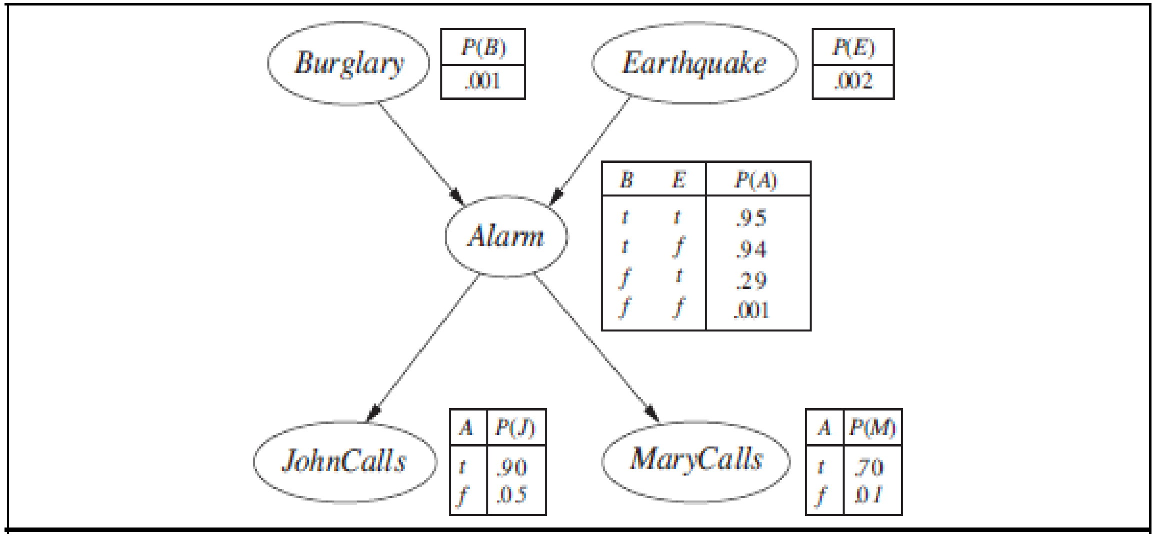
- John and Mary calling does not depend on what triggered the alarm only the alarm itself (simplification)
- CPT – conditional probability table. (Values not specified in the text before!! Made up here)



Exercise4: Alarm example calculation

- Calculate p(J,M,A,¬B,¬E) Means what??

- J : JohnCalls
- M : MaryCalls
- A : Alarm
- B : Burglary
- E : Earthquake



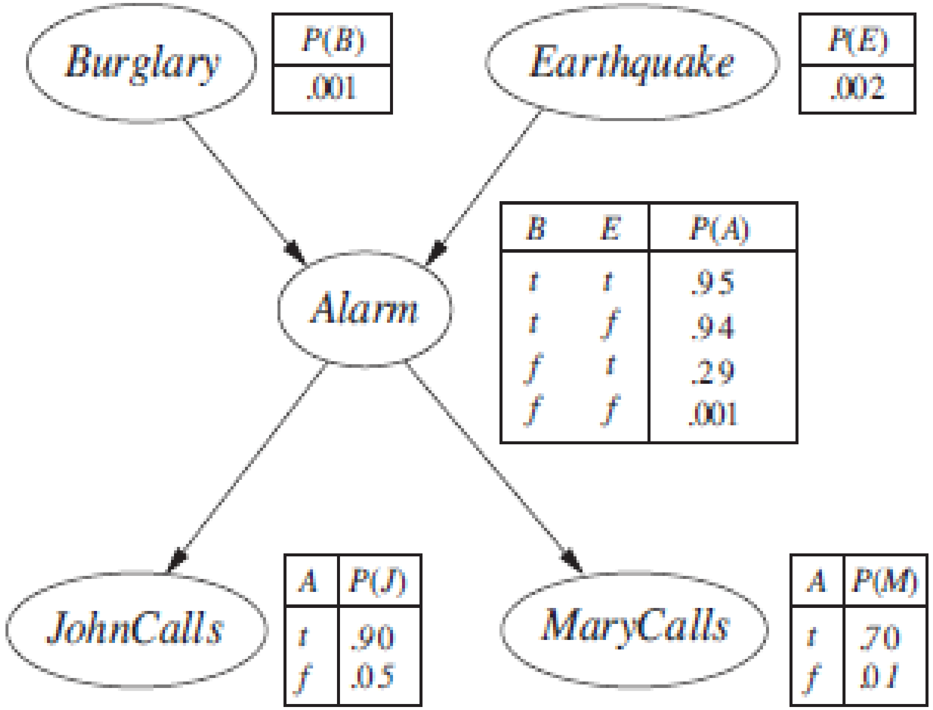
- Remember that p(¬X) = 1-p(X) and

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{i=n} p(X_i | Parents(X_i))$$

Exercise4: Alarm example calculation

- Calculate $p(J,M,A,\neg B,\neg E)$ Means what??

- J : JohnCalls
- M : MaryCalls
- A : Alarm
- B : Burglary
- E : Earthquake



$$p(J,M,A,\neg B,\neg E) = p(\neg B)p(\neg E)p(A|\neg B,\neg E)p(J|A)p(M|A)$$

Alarm example calculation

- $p(J,M,A,\neg B,\neg E) = 0.00062$
- So what does this tell us?

Very unlikely that both John and Mary call and the alarm has gone off when there is no burglary or earthquake

Two views on Bayesian networks

- Representation of the joint probability distribution
  - Helps to understand how to construct it
- Encoding of a collection of independence statements
  - Helpful in designing inference procedure

Tip

- When constructing the network try to use ordering based on **cause → symptom (causal)** rather than symptom → cause (diagnostic)
- Need to specify fewer numbers and numbers are easier to get
- Ex: Alarm → MaryCalls means we have to specify p(MaryCalls|Alarm) which is a lot easier than p(Alarm|MaryCalls) for MaryCalls → Alarm

Reasoning over Time or Space

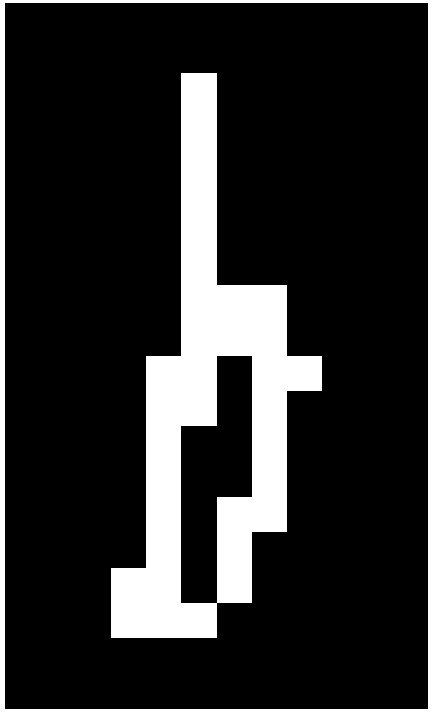
- Often, we want to reason about a sequence of observations
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring

- Need to introduce time (or space) into our models

Sequential data

- Measurement of time series
- Example: Sign recognition

- Measure: drawn path
- Want: characters



Sequential data

- Measurement of time series
- Example: Speech recognition

- Measure: audio signal
- Want: Words/sentences

- **DD2118 Speech and speaker recognition**

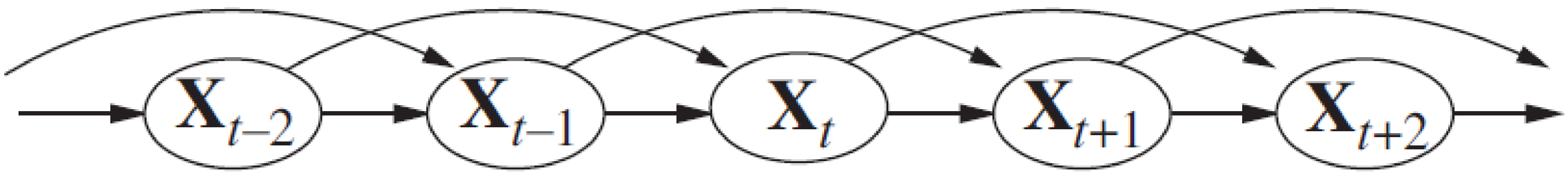
Let's get cracking with the theory again

Sequential data
Markov model

- A Markov model describes processes that evolve through time, in a sequence of steps
- The (first order) Markov assumption
  - The present (current state) can be predicted using local knowledge of the past (state at the previous step)
  - X_t is **conditionally independent** of all X_k ,k=t-2, …, given X_{t-1} ,i.e. $\mathbf{p(X_t|X_{t-1},X_{t-2},X_{t-3},\ldots) = p(X_t|X_{t-1})}$

Second-order Markov Model

- State at time k depends on the states at times k-1 and k-2



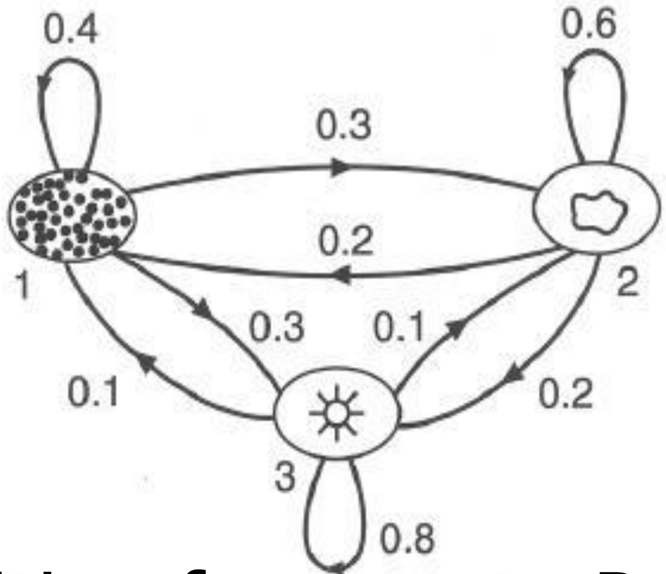
Example: Weather prediction

Once each day (e.g., at noon), the weather is observed and classified as being one of the following:

- State 1—Rain (or Snow; e.g. precipitation)
- State 2—Cloudy
- State 3—Sunny

with state transition probabilities:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

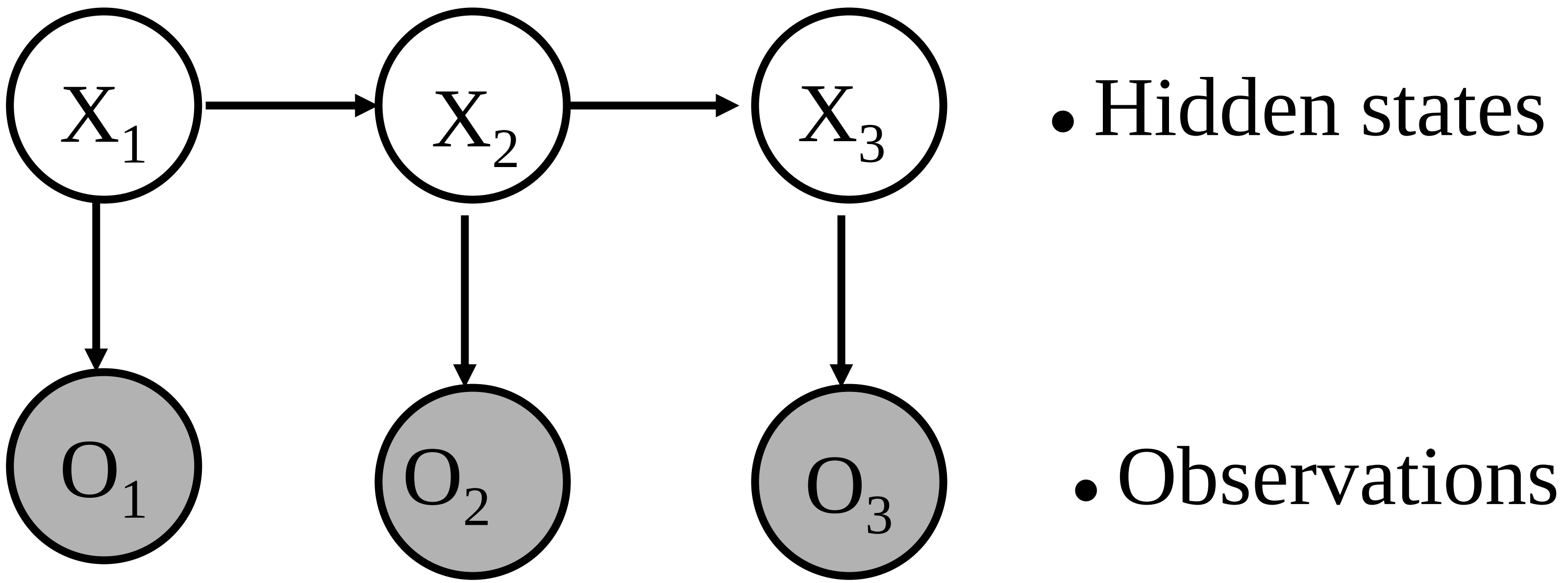


- Ex: 30% chance to transition from state Rain to state Cloudy (and Sunny).

What if we cannot observe the state?

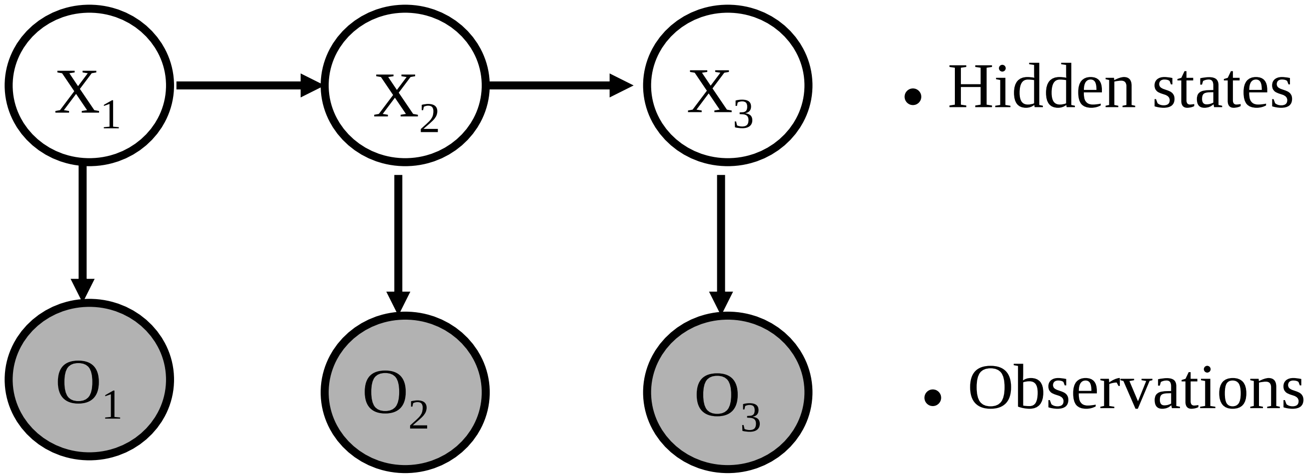
- Examples:
  - Cannot observe the weather, only the temperature
    - State=weather, observation=temperature
  - Cannot observe the words spoken, only the sound uttered
    - State=word, observation=sound
  - Cannot observe the position of the car only the laser scanner readings
    - State=position, observation=laser data

Hidden Markov Models (HMMs)



- State transition model: $p(X_t=j|X_{t-1}=i)=A(i,j)=a_{ij}$
- Observation model: $p(O_t=j|X_t=i)=b_{ij}$

Hidden Markov Models (HMMs)



- Two important (conditional) independence properties:
  - Markov hidden process: future depends on the past via the present
  - Current observation independent of all else given current state

Elements of HMM

1. Number of states N, $x \in \{1, \ldots, N\}$;

2. Number of events K, $k \in \{1, \ldots, K\}$;

3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \leq i \leq N$;

4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$;

5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.

Elements of HMM

The number of possible observation types

1. Number of states N, $x \in \{1, \ldots, N\}$;

2. Number of events K, $k \in \{1, \ldots, K\}$;

3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \leq i \leq N$;

4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$;

5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.

Elements of HMM

1. Number of states N, $x \in \{1, \ldots, N\}$;

2. Number of events K, $k \in \{1, \ldots, K\}$;

Note that we can start with the probability spread over several states

3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \leq i \leq N$;

4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$;

5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.

The model if often called λ

- λ is the model, i.e.,
  $\lambda=(A, B, \pi)$

state transition matrix

Distribution for initial state

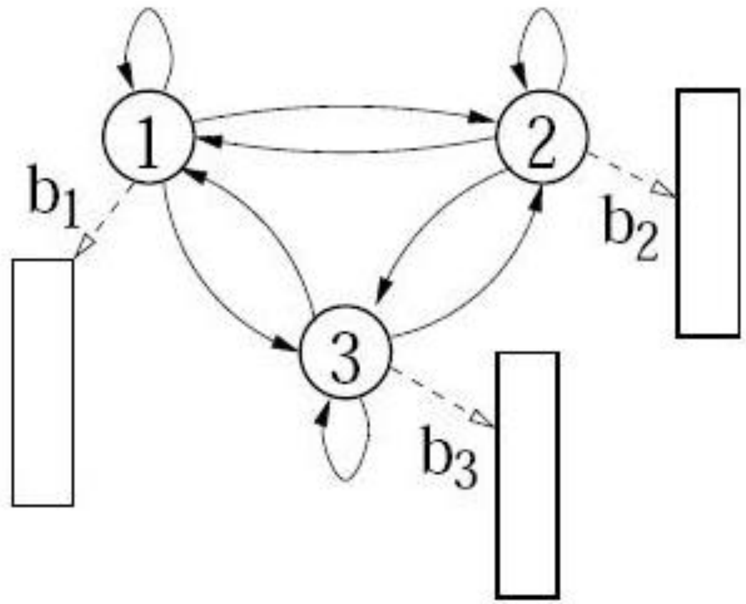
Output Matrix
Output sometimes called "emissions"

- λ is sometimes called *M*
- A, B and π are row-stochastic matrices (their rows sum to 1)

Elements of HMM, λ

- Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time t = 1 (sometimes t=0).
- Often referred to as π

Ex: π =[0.5 0.2 0.3], i.e.,
–$p(X_1=1)=0.5$
–$p(X_1=2)=0.2$
–$p(X_1=3)=0.3$



Elements of HMM, λ

- State transition matrix A: holding the probability of transitioning from one hidden state to another hidden state.
- Ex: a_{21} gives $p(X_{t+1}=1|X_t=2)$, i.e. probability to transition from state 2 to state 1

	$X_{t+1}=1$	$X_{t+1}=2$	…	$X_{t+1}=N$
$X_t=1$	a_{11}	a_{12}	…	a_{1N}
$X_t=2$	a_{21}	a_{22}	…	a_{2N}
…	…	…	…	…
$X_t=N$	a_{N1}	a_{N2}	…	a_{NN}

Elements of HMM, λ

- Output matrix B: Contains the probability of observing a particular measurement given that the hidden model is in a particular hidden state.
- $b_i(O_t=j)$ is the probability to observe j in state i

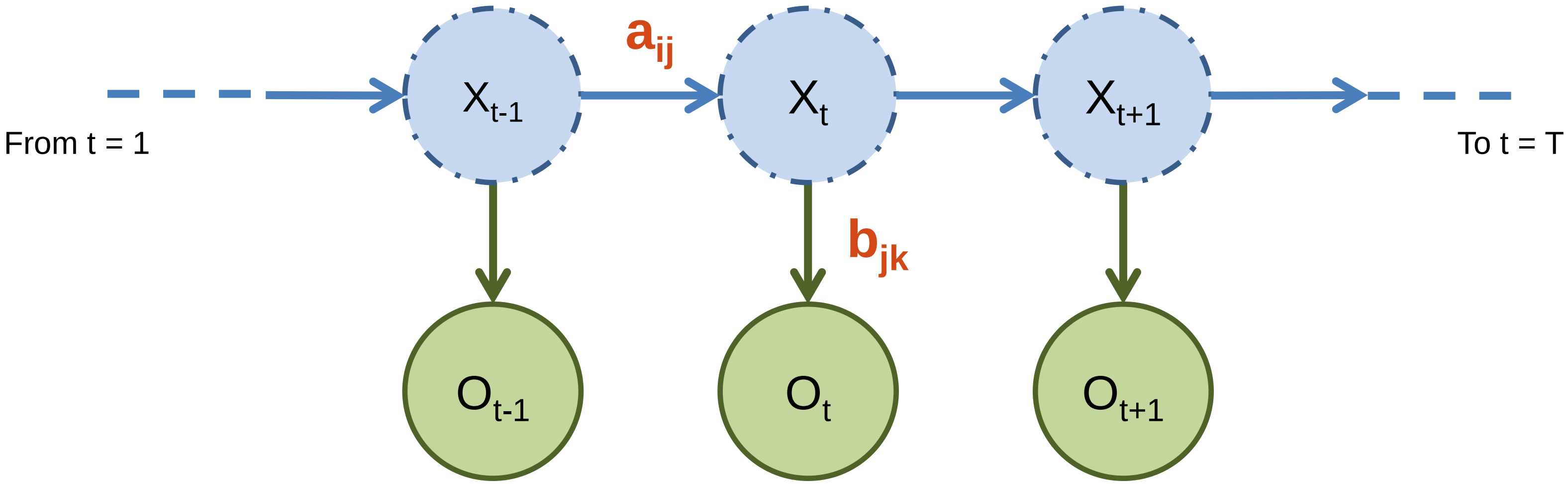
	$O_t=1$	$O_t=2$	…	$O_t=K$
$X_t=1$	$b_1(1)$	$b_1(2)$	…	$b_1(K)$
$X_t=2$	$b_2(1)$	$b_2(2)$	…	$b_2(K)$
…	…	…	…	…
$X_t=N$	$b_N(1)$	$b_N()$	…	$b_N(K)$

What is an HMM, reeeallllly?!?!?

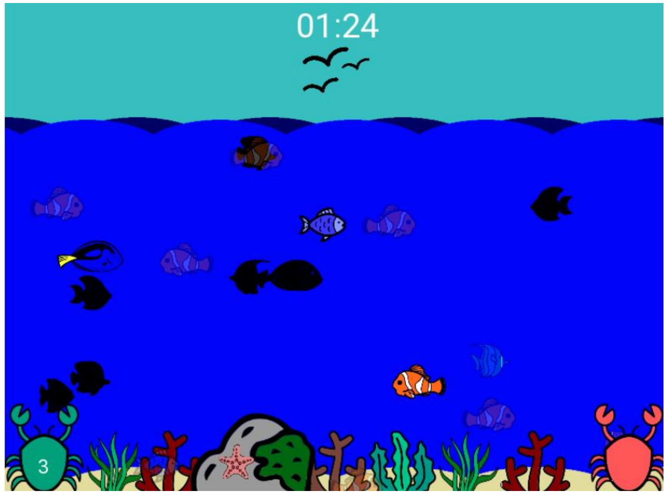
- It is a **model** (not necessarily a perfect one) for system
- It can be used to (e.g.)
  - Generate predictions about how the system will behave
  - Analyze if a sequence of measurements match a certain model, e.g., did the person say "Bayesian" (i.e. match our model for that word) or "Bearnaise" (i.e. match that model)?
  - Learn something about a system by learning the model parameters.

HMM Terminology

Time instants	t *in* {1,2 . . . T}
Hidden States / States / Emitters	X_t
Outputs / Emissions / Observations / Visible States	O_t
All possible states / states set	X_t *in* {1,2 . . . N}
All possible emissions / emissions set	O_t *in* {1,2 . . . K}
Initial state distribution / Initial state probabilities	p_i in q *or* π_i in π
Transition probabilities / State transition probabilities	a_{ij} *in* row-stochastic matrix A
Emission probabilities / Observation probabilities	b_{jk} *in* row-stochastic matrix B



HMM Assignment



- Fishing Derby!
  - Have some knowledge about how different fish swim
  - Have sparse observations of fish
  - Q: How to tell which fish belong to the same species?
  - You get to: Design a model, learn the parameters from data, make decisions, take actions

End of lecture

See exercise solutions after this...

Exercise 1: Deriving Bayes rules

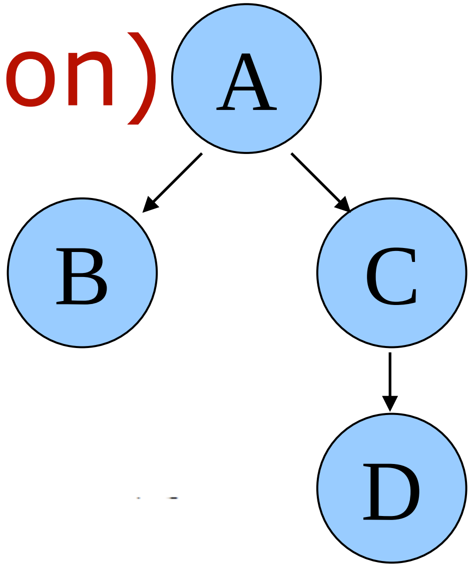
- Product rule
  - $p(X,Y)=p(Y|X)p(X)$
  - $p(Y,X)=p(X|Y)p(Y)$

- $p(X,Y)=p(Y,X)$ (symmetry)

- $p(Y|X)p(X) = p(X|Y)p(Y)$

- $\rightarrow \boxed{p(Y|X) = p(X|Y)\, p(Y) \,/\, p(X)}$

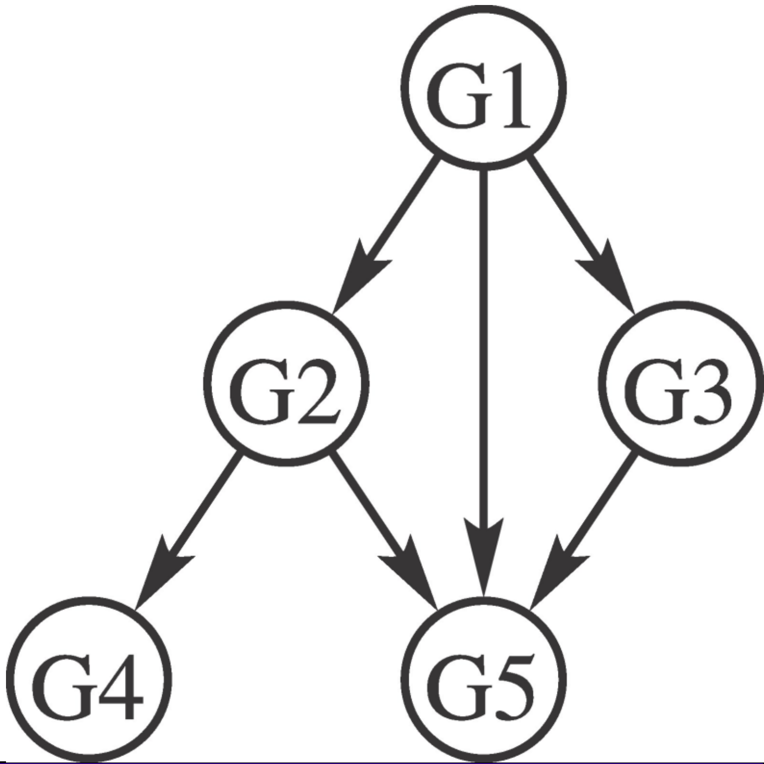
Exercise2: Factorization (alt. solution)



$$
\begin{aligned}
p(A,B,C,D) &= \{\text{product rule with X=A, Y=B,C,D}\} \\
&= p(B,C,D|A)p(A) \\
&= \{\text{product rule with X=B, Y=C,D}\} \\
&= p(C,D|A,B)p(B|A,C,D)p(A) \\
&= \{\text{B conditionally independent of C,D given A}\} \\
&= p(C,D|A,B)p(B|A)p(A) \\
&= \{\text{product rule with X=C, Y=D}\} \\
&= p(D|A,B,C)p(C|A,B)p(B|A)p(A) \\
&= \{\text{C conditionally independent of B given A}\} \\
&= p(D|A,B,C)p(C|A)p(B|A)p(A) \\
&= \{\text{D conditionally independent of A and B given C}\} \\
&= p(D|C)p(C|A)p(B|A)p(A)
\end{aligned}
$$

Exercise3: Factorize the graph

- $p(G1,G2,G3,G4,G5)$
- $= p(G1)\, p(G2|G1)\, p(G3|G1)\, p(G4|G2)\, p(G5|G1,G2,G3)$



Exercie4: Alarm example calculation

- $p(J,M,A,\neg B,\neg E)$ = {use graph!!}
- $= p(\neg B)p(\neg E)p(A|\neg B,\neg E)p(J|A)p(M|A)$
- = {read from CPT and use $p(\neg X) = 1-p(X)$}
- $= (1-0.001) * (1-0.002) * 0.001 * 0.9 * 0.7$
- $= 0.00062$