Donders Institute
for Brain, Cognition and Behaviour

BKI259: Artificial Intelligence: Principles and Techniques

Bayesian Networks (part 1/3)

Radboud University Nijmegen

Lecture/block outline

- We start with the block on **Bayesian networks**
- Topics
  - Conditional Independence
  - Formal representation of BNs
  - Factors
  - Variable elimination algorithm
  - Complexity (incl. tree-width)
  - Approximate inference
  - Learning from data (incl. missing data)
  - Elicitation of domain knowledge
  - Dynamical systems, Hidden Markov models
  - Most probable explanation & MAP

Today
2nd week
3rd week
4th week

Literature

- Required reading:
  P&MackW Chapter 6 and Section 11.2 (learning)
- Additional reading:
  Russel & Norvig, Chapter 13 and 14

- Other background material: *Textbooks*: Koller & Friedman'09, Pearl'88, Jensen and Nielson'07
- *AISpace*: http://aispace.org/bayes/
- Video lectures Daphne Koller:
  https://www.youtube.com/playlist?list=PL50E6E80E8525B59C
- And our own knowledge clip (see Weblectures)

Bayesian networks

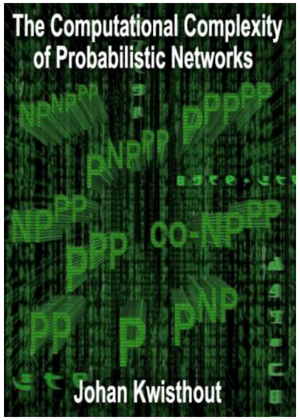
- Bayesian networks are seen by some as the most significant contribution in AI in the last decades of the 20st century – Turing award in 2011 for Judea Pearl

- Applications: spam filtering, speech recognition, robotics, forensics, decision support systems, …
- *Also*: computational cognitive models (Bayesian turn in cognitive science 2000-2010)
- *Also*: computational level theories of information processing in the brain (e.g., "Bayesian Brain")

- Ongoing research topic at AI / DCC / CS

Terminology

- Probabilistic / Bayesian / Belief network



- Math-oriented: **probabilistic**; focus on mathematical formalism and its properties
- AI-oriented: **belief**; focus on application as modeling expert beliefs in domain where one needs to reason under uncertainty
- Nowadays **Bayesian** is the more common general name, also in cognitive (neuro-)science

Background assumed

- Understand basic (discrete) probability theory
  - Look at the recap lecture if in doubt!

- Joint probability **distribution** $P(A, B, C)$
- Joint probability **value** $P(A=a, B=b, C=c)$
- Marginal probability $P(A) = \sum_{B,C} P(A, B, C)$
- Conditional probability $P(A \mid b) = P(A, b) / P(b)$

- Notation:
  - upper case = stochastic variable A
  - lower case = value of a variable a
  - Bold: sets of variables / values **A** and **a**
  - Binary variables: a and $\neg$a

Shorthand:
P(b) for P(B =b)

Product rule in probability theory

- Product rule: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1 \ldots, X_{i-1})$
- $X_1 \ldots X_n$ is *arbitrary* order of variables!
- P(A,B,C) = P(A) × P(B | A) × P(C | A, B)
  = P(B) × P(C | B) × P(A | C, B) (etc.)
- P(A) × P(B | A) × P(C | A, B) =
  P(A) × P(A, B) / P(A) × P(A, B, C) / P (A,B)
- P(B) × P(C | B) × P(A | C, B) =
  P(B) × P(B, C) / P(B) × P(A, B, C) / P (B,C)

P(A, B) = P(B, A)

Problem!

- Lots of entries in the table to fill!
- For *k* Boolean random variables, you need a table of size 2^k and have to specify $2^k - 1$ numbers
- How do we use fewer numbers? For this we need the concept of **independence**

A	B	C	P(A, B,C)
a	b	c	0.1
a	b	¬c	0.2
a	¬b	c	0.05
a	¬b	¬c	0.05
¬a	b	c	0.3
¬a	b	¬c	0.1
¬a	¬b	c	0.05
¬a	¬b	¬c	0.15

Adds to 1

Independence

- Variables A and B are **independent** in a prob. distr. P (notation $A \perp\!\!\!\perp_P B$) if any of the following holds
  - $P(A,B) = P(A) \times P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
- Knowing the outcome of B does not give you any information on the outcome of A
- Examples:
  - 1^{st} dice throw is independent of 2^{nd} throw
  - Rain in Uganda is independent of whether NEC have won, lost, or drawn last football game

Independence

- How is independence useful?
- Suppose you have *n* coin flips and you want to calculate the joint distribution $P(C_1, \ldots, C_n)$
- If the coin flips are not independent, you need to specify 2^n -1 values in the table
- If the coin flips are independent, then $P(C_1, \ldots, C_n) = \Pi_i P(C_i)$
- So, you will need only *n* values (one for each coin, or just a single one if all coin flips are equally likely)

Conditional Independence

- Independence is often **too crude** an assumption

- Variables A and B are **conditionally independent** in a probability distribution P (notation $A \perp\!\!\!\perp_P B \mid C$) if any of the following holds
  - $P(A, B \mid C) = P(A \mid C) \times P(B \mid C)$
  - $P(A \mid B, C) = P(A \mid C)$
  - $P(B \mid A, C) = P(B \mid C)$

- Knowing C already tells me everything about A; information about B is not relevant anymore for A

Independence relations

- Every probability distribution P over a set of variables V has an **independence relation** I_P describing its independences

- We call $(X,Z,Y) \in I_P$ an **independence statement** stating that X and Y are conditionally independent given Z $(X \perp\!\!\!\perp_P Y \mid Z)$

- There are many axioms that can be used to reason about whether independence relations hold, such as: $(X,Z,Y) \in I_P \Leftrightarrow (Y,Z,X) \in I_P$

- Independences between variables can also be described using a **graphical model**

Global Markov Property (undirected graphs)

- Let G be an undirected graph and let **X**, **Y**, **Z** be subsets of vertices of G

- The set **Z separates X** and **Y** in G ($X \perp\!\!\!\perp_G Y \mid Z$) if every path from a vertex X in **X** to a vertex Y in **Y** contains at least one variable Z in **Z**

- Note the similarity as well as the difference in notation: $X \perp\!\!\!\perp_P Y \mid Z$ denotes independence between variables in a probability distribution P, whereas $\mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}$ denotes that **Z** blocks the paths from **X** to **Y** in G

- We can relate the two notions using the concepts D-Maps, I-Maps, and P-Maps

D-Map, I-Map, and P-Map

- **Definition**: Let G be an undirected graph and let I_P be an independence relation defined on a probability distribution P. We call G:

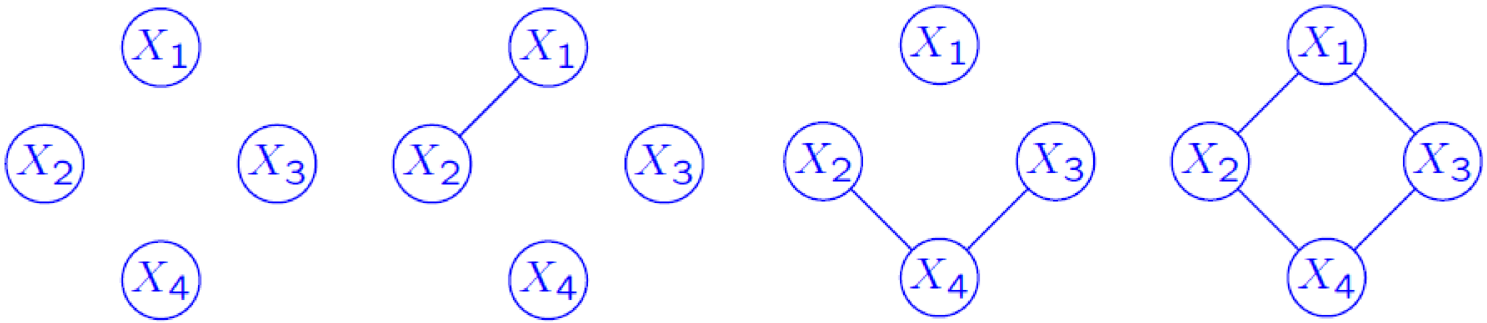
  - A D-Map of P if for all sets of variable X,Y,Z in P it holds that $(X \perp\!\!\!\perp_P Y \mid Z) \Rightarrow (X \perp\!\!\!\perp_G Y \mid Z)$

  - An I-Map of P if for all sets of variable X,Y,Z in P it holds that $(X \perp\!\!\!\perp_G Y \mid Z) \Rightarrow (X \perp\!\!\!\perp_P Y \mid Z)$

  - A P-Map of P if for all sets of variable X,Y,Z in P it holds that $(X \perp\!\!\!\perp_G Y \mid Z) \Leftrightarrow (X \perp\!\!\!\perp_P Y \mid Z)$

D-Map, I-Map, and P-Map

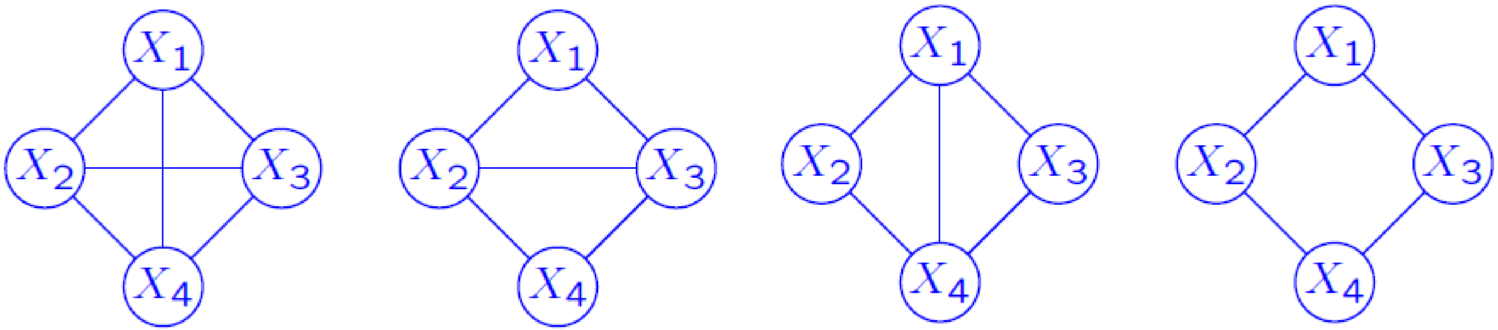
- Let $(X_1, \{X_2,X_3\}, X_4) \in I_P$ and $(X_2, \{X_1,X_4\}, X_3) \in I_P$
- (i.e., X_1 is conditionally independent of X_4 given X_2 and X_3, and X_2 is conditionally independent of X_3 given X_1 and X_4)
- These are all **D-Maps**:
  all independences in P
  are in G (but maybe more)



D-Map, I-Map, and P-Map

- Let $(X_1, \{X_2,X_3\}, X_4) \in I_P$ and $(X_2, \{X_1,X_4\}, X_3) \in I_P$
- (i.e., X_1 is conditionally independent of X_4 given X_2 and X_3, and X_2 is conditionally independent of X_3 given X_1 and X_4)

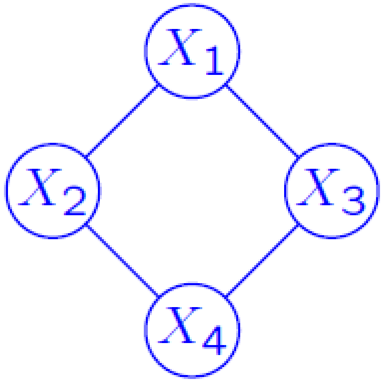
  - These are all **I-Maps**:
    all independences in G
    are in P (but maybe more)



D-Map, I-Map, and P-Map

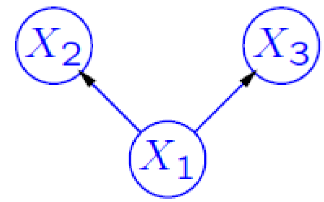
- Let $(X_1, \{X_2,X_3\}, X_4) \in I_P$ and $(X_2, \{X_1,X_4\}, X_3) \in I_P$
- (i.e. X_1 is conditionally independent of X_4 given X_2 and X_3, and X_2 is conditionally independent of X_3 given X_1 and X_4)

- This is the **P-Map**:
  the independences in P
  and in G perfectly match

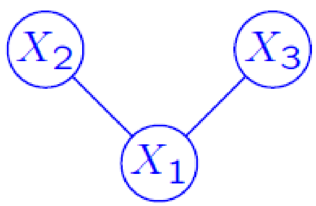


Directed graphical models

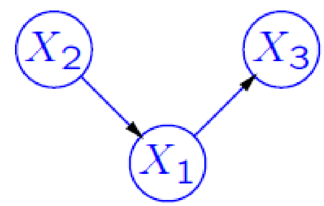
- Directed graphical models introduce an additional source of information: the direction of the arcs!



$(X_2, \emptyset, X_3) \notin I_P$
$(X_2, X_1, X_3) \in I_P$

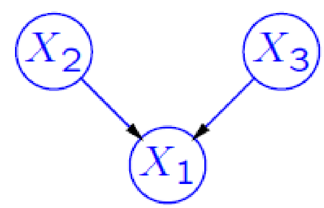


vs



$(X_2, \emptyset, X_3) \notin I_P$
$(X_2, X_1, X_3) \in I_P$

$(X_2, \emptyset, X_3) \notin I_P$
$(X_2, X_1, X_3) \in I_P$



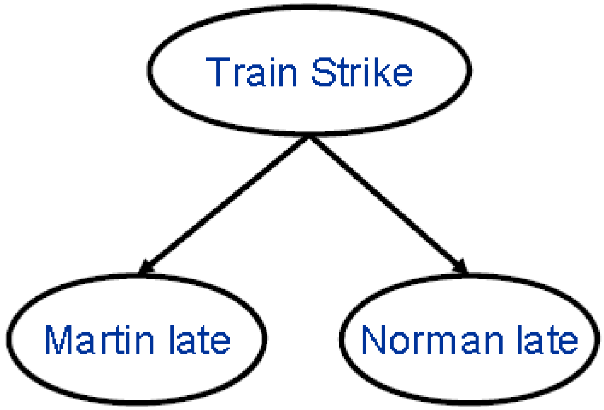
$(X_2, \emptyset, X_3) \in I_P$
$(X_2, X_1, X_3) \notin I_P$

Causal Chain

- Age and Blood Pressure are dependent, $P(B \mid A) \neq P(B)$ $\quad B \not\perp\!\!\!\perp A$

- but conditionally independent given Weight: $P(B \mid A, W) = P(B \mid W)$ $\quad B \perp\!\!\!\perp A \mid W$

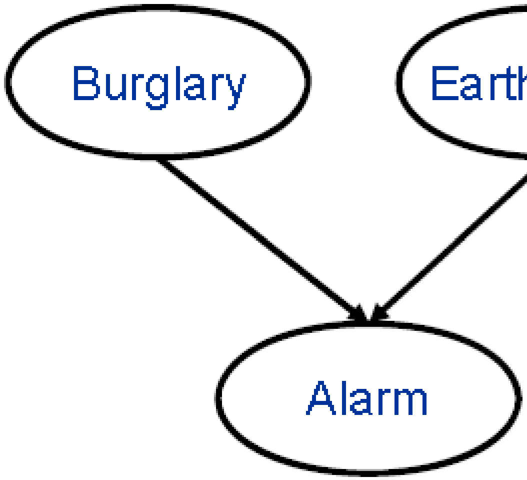
Common Cause



- Martin late and Norman late are dependent, $M \not\perp\!\!\!\perp N$
  $P(M,N) \neq P(M)\, P(N)$

- but conditionally independent given Train Strike: $M \perp\!\!\!\perp N \mid T$
  $P(M,N \mid T) = P(M|T)\, P(N|T)$

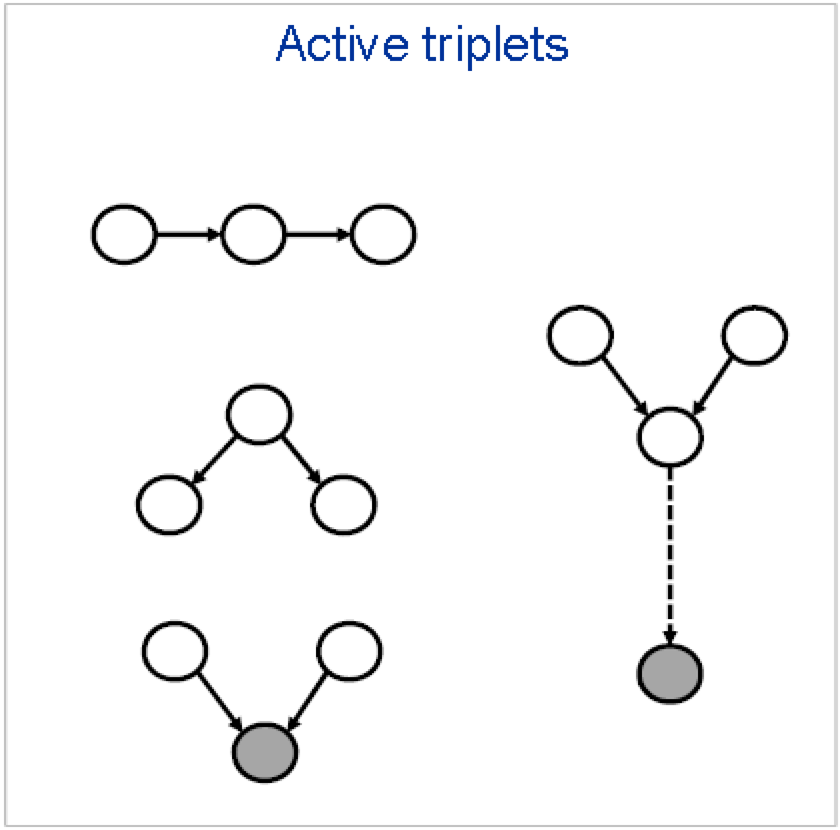
Common Effect

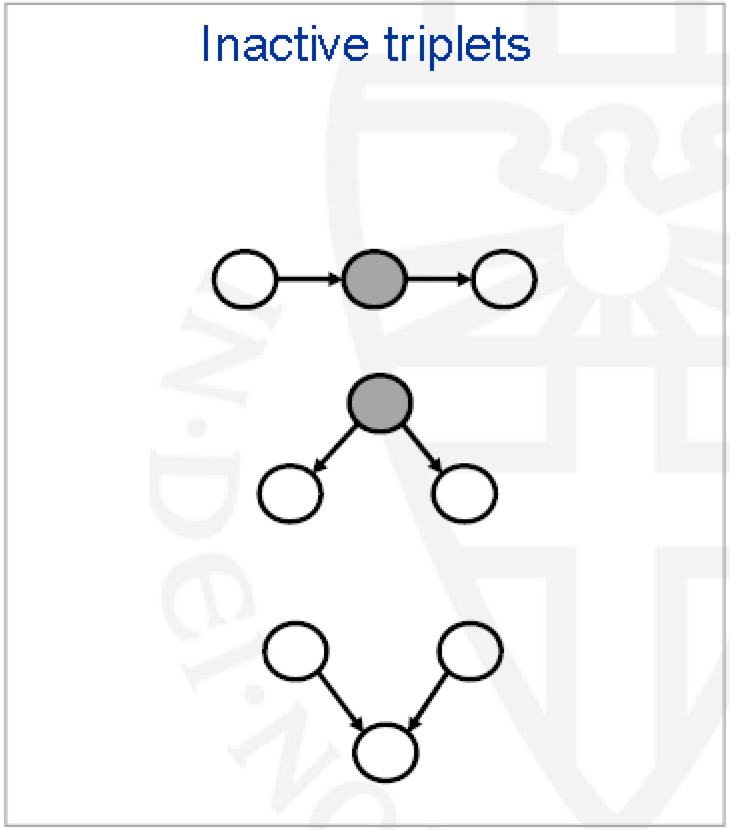


- Burglary and Earthquake are independent, $P(B,E) = P(B)\,P(E)$ $B \perp\!\!\!\perp E$
- but conditionally dependent given Alarm: $P(B,E \mid A) \neq P(B|A)\,P(E|A)$ $B \not\!\perp\!\!\!\perp E \mid A$

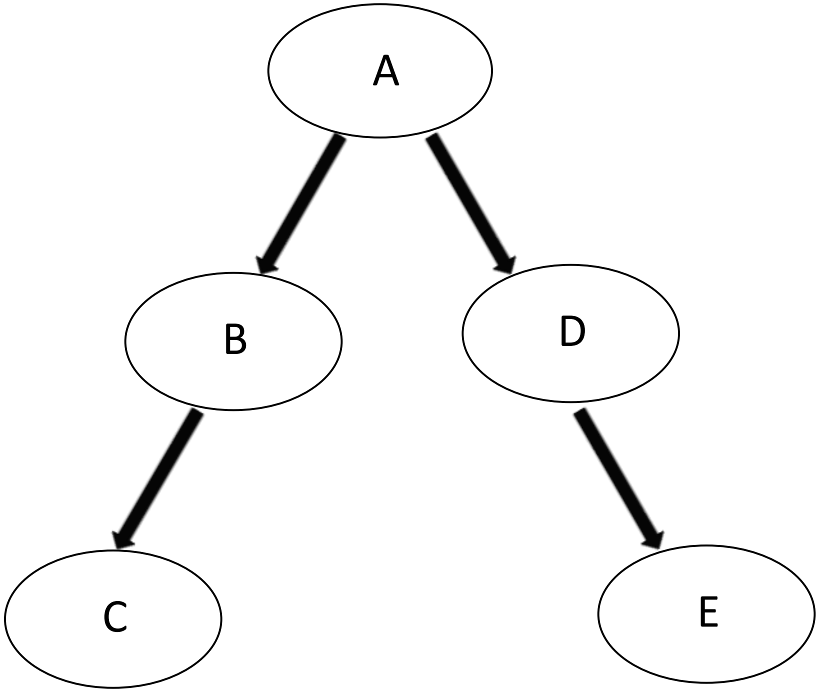
D-separation: reachability

Two nodes are D-separated if all chains connecting them are inactive



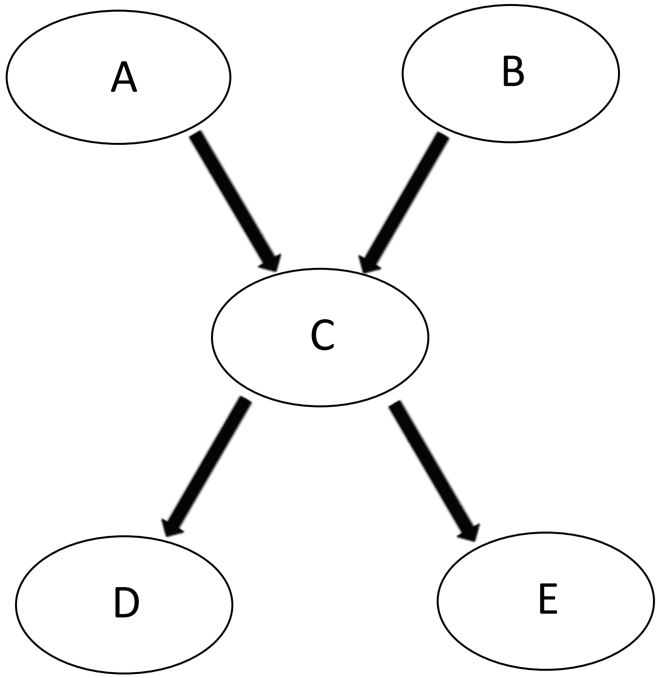


Reading off independence (example 1)



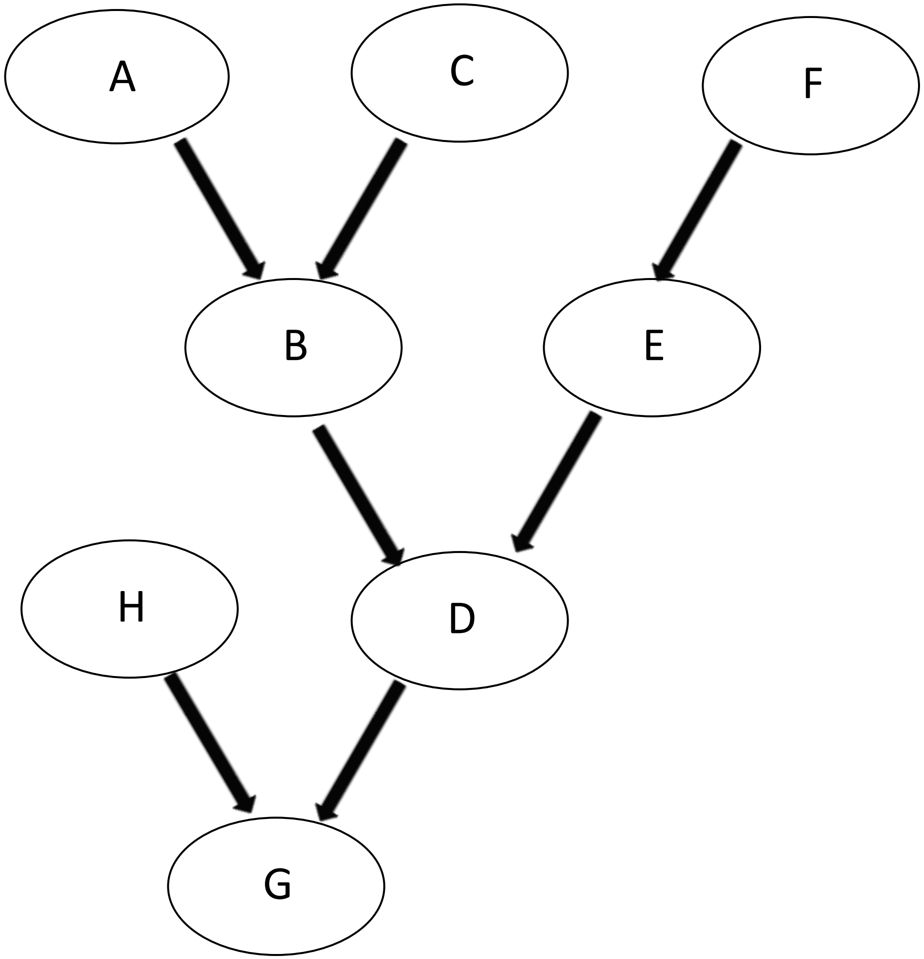
- Is C $\perp\!\!\!\perp$ A? **NO**
- Is C $\perp\!\!\!\perp$ A | B ? **YES**
- Is C $\perp\!\!\!\perp$ D? **NO**
- Is C $\perp\!\!\!\perp$ D | A ? **YES**
- Is E $\perp\!\!\!\perp$ C | D ? **YES**

Reading off independence (example 2)



- Is A $\perp\!\!\!\perp$ E? **NO**
- Is A $\perp\!\!\!\perp$ E | B ? **NO**
- Is A $\perp\!\!\!\perp$ E | C? **YES**
- Is A $\perp\!\!\!\perp$ B ? **YES**
- Is A $\perp\!\!\!\perp$ B | C ? **NO**

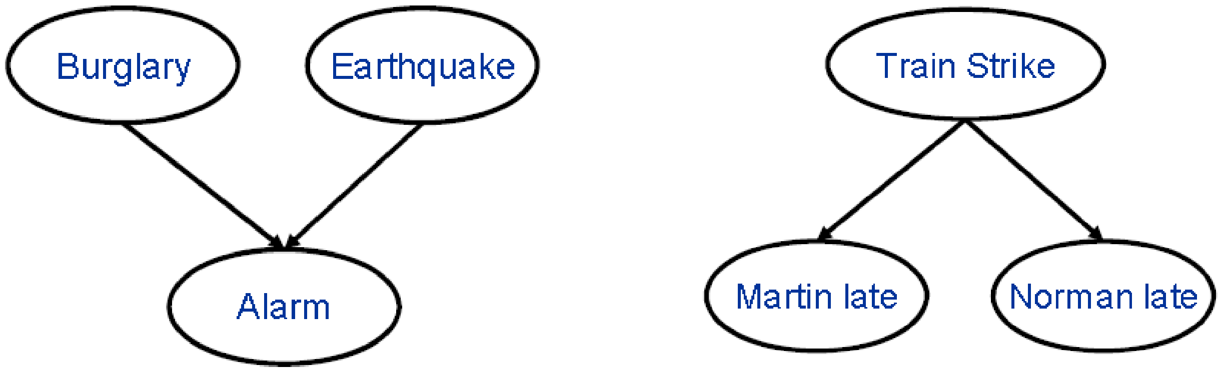
Reading off independence (example 3)



- Is $A \perp\!\!\!\perp F$? **YES**
- Is $A \perp\!\!\!\perp F \mid D$? **NO**
- Is $A \perp\!\!\!\perp F \mid G$? **NO**
- Is $A \perp\!\!\!\perp F \mid H$? **YES**

Directed vs. undirected models

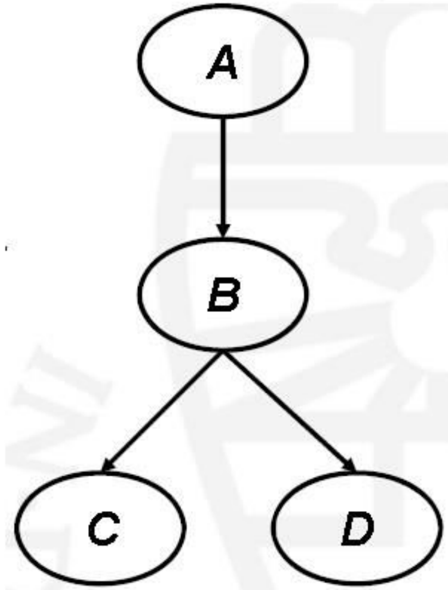
- Independences can be described using directed and undirected graphs; directed graphs have a bit more expressive power



- Directed graphs more intuitively represent statistical information
  - Easier for domain experts to formulate stochastic relations
  - Easier for humans to interpret structure and results of inferences
  - Causal interpretation (Cause → Effect)

Bayesian network

- A Bayesian network is made up of:
  - A directed acyclic graph with nodes representing random variables
  - Probability tables for each node in the graph
- The DAG describes the conditional independences in the network



A	**P(A)**
false	0.6
true	0.4

A	**B**	**P(B|A)**
false	*false*	0.01
false	*true*	0.99
true	*false*	0.7
true	*true*	0.3

B	**C**	**P(C|B)**
false	*false*	0.4
false	*true*	0.6
true	*false*	0.9
true	*true*	0.1

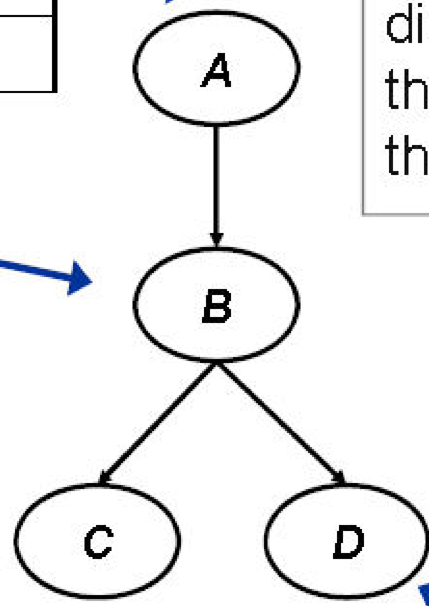
B	**D**	**P(D|B)**
false	*false*	0.02
false	*true*	0.98
true	*false*	0.05
true	*true*	0.95

Conditional Probability Tables

A	P(A)
false	0.6
true	0.4

Each node X has a conditional probability distribution $P(X \mid Parents(X))$ that quantifies the effect of the parents on the node

A	B	P(B|A)
false	*false*	0.01
false	*true*	0.99
true	*false*	0.7
true	*true*	0.3



B	C	P(C|B)
false	*false*	0.4
false	*true*	0.6
true	*false*	0.9
true	*true*	0.1

B	D	P(D|B)
false	*false*	0.02
false	*true*	0.98
true	*false*	0.05
true	*true*	0.95

Conditional Probability Tables

- For any given combination of values of the parents (eg. *B*), the entries for *P(C=true | B)* and *P(C=false | B)* must add up to 1, eg. *P(C=true | B=false)* + *P(C=false | B=false)* = 1

B	**C**	***P(C|B)***
false	*false*	0.4
false	*true*	0.6
true	*false*	0.9
true	*true*	0.1

Sums to 1

Sums to 1

- If you have a Boolean variable with *k* Boolean parents, this table has 2^{k+1} probabilities (but only 2^k need to be specified)

Bayesian network running example

$P(\text{tampering}) = 0.02$

$P(\text{fire}) = 0.01$

$P(\text{alarm} \mid \text{fire} \wedge \text{tampering}) = 0.5$

$P(\text{alarm} \mid \text{fire} \wedge \neg\text{tampering}) = 0.99$

$P(\text{alarm} \mid \neg\text{fire} \wedge \text{tampering}) = 0.85$

$P(\text{alarm} \mid \neg\text{fire} \wedge \neg\text{tampering}) = 0.0001$

$P(\text{smoke} \mid \text{fire}) = 0.9$

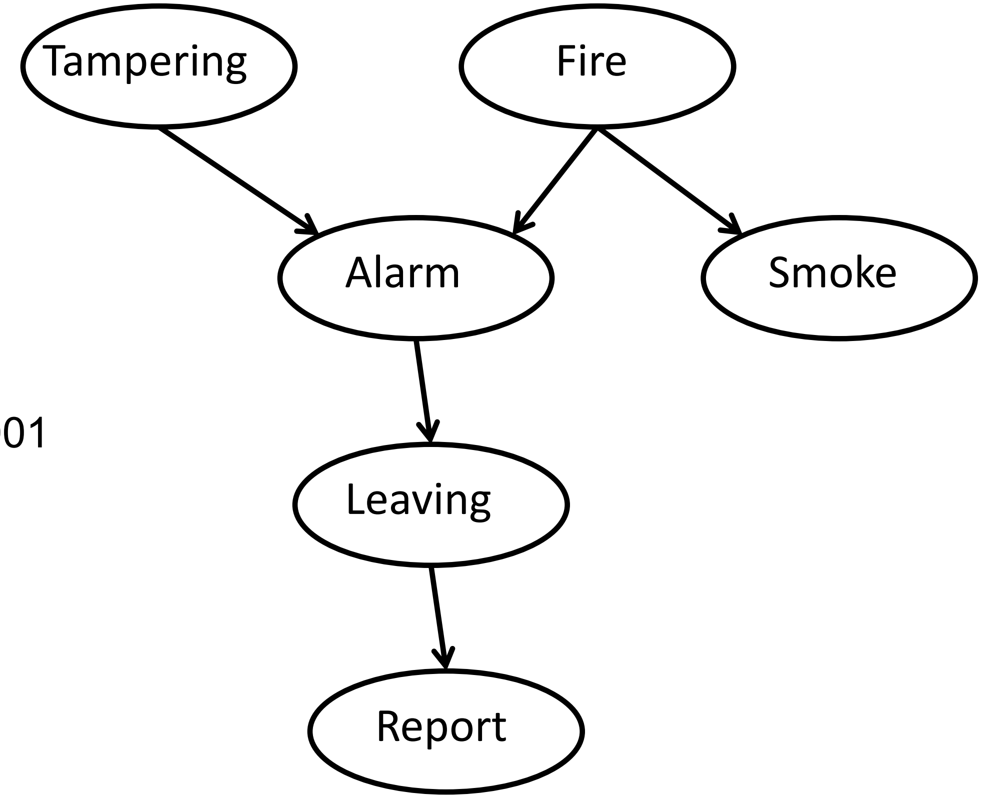
$P(\text{smoke} \mid \neg\text{fire}) = 0.01$

$P(\text{leaving} \mid \text{alarm}) = 0.88$

$P(\text{leaving} \mid \neg\text{alarm}) = 0.001$

$P(\text{report} \mid \text{leaving}) = 0.75$

$P(\text{report} \mid \neg\text{leaving}) = 0.01$



Bayesian Networks

- Some important properties of Bayesian networks
- It encodes the **conditional independence** relationships between the variables in the graph structure (as directed I-Map)
- It is a **compact representation** of the joint probability distribution over the variables
- It allows for (relatively) **efficient computations** of joint probability distributions of interest

BNs and Joint Probability Distributions

- There are (in general) many Bayesian networks that describe the same probability distribution, some more efficient than others in respecting independences

- Because of the independences in the distribution some arcs of the network can be pruned:

Chain or product rule

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

BN property

where Parents(X_i) are the parents of X_i in the graph

Joint probability distribution

- The joint probability distribution for a Bayesian network reads:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1 \dots, X_{i-1}) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$$

  where Parents(X_i) are the parents of node X_i in the graph

- Compact representation when (the ordering of the nodes is chosen such that) nodes have few parents

- Typically works best when reasoning from cause to effect

Example

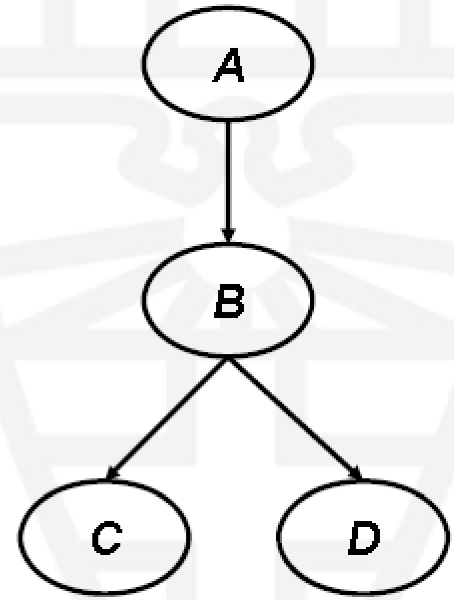
- From the previous example:

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

  $P(A)\,P(B \mid A)\,P(C \mid B)\,P(D \mid B)$

  from structure

  for any setting of the variables A, B, C, D



- Specific case:

  $P(A{=}false,\ B{=}true,\ C{=}false,\ D{=}true) =$

  $P(A{=}false)\,P(B{=}true \mid A{=}false)\,P(C{=}false \mid B{=}true)\,P(D{=}true \mid B{=}true) =$

  $\underbrace{0.6 \times 0.99 \times 0.9 \times 0.95}_{\text{from conditional probability tables}} = 0.51$

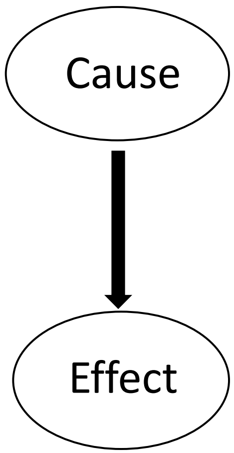
Computing with probabilities

- Relatively straightforward (and uninteresting) without any observations
- More challenging (and interesting) with observations: Bayes' rule comes into reason from observed effect to unobserved cause
- Probabilistic inference in the general case can be computationally extremely demanding (inference is an NP-hard problem)
- Approximations are available that may be of use

Axioms of probability theory

- P should obey three axioms (A. Kolmogorov):
  1. $P(A) \geq 0$ for all events A
  2. $P(\Omega) = 1$
  3. $P(A \cup B) = P(A) + P(B)$ for disjoint events A and B
- Some consequences (from set theory):
  - $P(A) = 1 - P(\Omega \setminus A)$
  - $P(\emptyset) = 0$
  - If $A \subseteq B$, then $P(A) \leq P(B)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$
- Given these axioms and a completely defined probability measure any (marginal / conditional) probability of interest can be computed!

Example: cause and effect



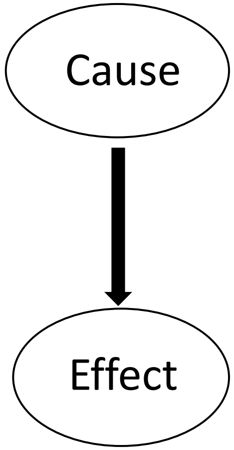
- **Cause** (e.g., disease) is often unobserved
- What we observe is the **effect**
- **Goal**: compute the probability of the cause given the effect
- $\Pr(\text{cause}) = 0.01$ $\Pr(\neg\text{cause}) = 0.99$
  $\Pr(\text{effect} \mid \text{cause}) = 0.9$ $\Pr(\text{effect} \mid \neg\text{cause}) = 0.2$
- What is $\Pr(\text{cause} \mid \text{effect})$?

Bayes' Theorem

- Definition of conditional probability:
  - Pr(E | C) = P(E,C) / P(C)
- But then, this also holds:
  - Pr(C | E) = P(E,C) / P(E)
- And thus:

$$\underbrace{\Pr(C \mid E)}_{\text{Posterior}} = \frac{\overbrace{\Pr(E \mid C)}^{\text{Likelihood}}\ \overbrace{\Pr(C)}^{\text{Prior}}}{\underbrace{\Pr(E)}_{\text{Marginal likelihood}}}$$

Cause and effect revisited



- $\Pr(\text{cause}) = 0.01$ $\quad$ $\Pr(\neg\text{cause}) = 0.99$

  $\Pr(\text{effect} \mid \text{cause}) = 0.9$ $\quad$ $\Pr(\neg\text{effect} \mid \text{cause}) = 0.1$

  $\Pr(\text{effect} \mid \neg\text{cause}) = 0.2$ $\quad$ $\Pr(\neg\text{effect} \mid \neg\text{cause}) = 0.8$

- What is $\Pr(\text{cause} \mid \text{effect})$?

- $$\Pr(c \mid e) = \Pr(e \mid c) \Pr(c) / \Pr(e) = 0.9 \times 0.01 / \Pr(e) = 0.09 / \Pr(e)$$

  $$\Pr(e) = \Pr(e \mid c) \Pr(c) + \Pr(e \mid \neg c) \Pr(\neg c) = 0.9 \times 0.01 + 0.2 \times 0.99 = 0.207$$

  $$\Pr(c \mid e) = 0.09 / 0.207 \approx 0.43$$

Useful notation: Factors

- A factor $f(X_1,\ldots,X_k)$:

$$f : X_1 \times \ldots \times X_k \rightarrow R$$

yields a real value ($r \in R$) for each concrete tuple

$$(x_1, \ldots, x_k) \in (X_1 \times \ldots \times X_k)$$

- Scope = $\{X_1,\ldots,X_k\}$ of "free variables"

From probability distributions to factors

- P(Tampering)
- P(Alarm)
- P(Report)
- P(Alarm | Tampering)
- P(Alarm | ¬tampering)
- P(¬alarm | Tampering)
- P(Smoke | Alarm)
- P(Report | Fire)
- …

Tampering	**Prob**
tampering	0.02
¬tampering	0.98

Alarm	**Prob**
alarm	0.0266
¬alarm	0.9734

Alarm	**Tamp**	**Cond Prob**
alarm	tamp	0.845
alarm	¬tamp	0.01
¬alarm	tamp	0.155
¬alarm	¬tamp	0.99

Factors

- $f_1(\text{Alarm, Tampering}) \stackrel{def}{=} P(\text{Alarm} \mid \text{Tampering})$
- $f_2(\text{Alarm}) \stackrel{def}{=} P(\text{Alarm} \mid \neg\text{tamp})$

Alarm	**Tamp**	$\mathbf{f_1}$
alarm	tamp	0.845
alarm	¬tamp	0.01
¬alarm	tamp	0.155
¬alarm	¬tamp	0.99

Alarm	**Tamp= ¬tamp**	**Cond Prob**
~~alarm~~	~~tamp~~	~~0.845~~
alarm	¬tamp	0.01
~~¬alarm~~	~~tamp~~	~~0.155~~
¬alarm	¬tamp	0.99

Alarm	**Tamp= ¬tamp**	$\mathbf{f_2}$
alarm	¬tamp	0.01
¬alarm	¬tamp	0.99

Caution

- A (conditional/joint/marginal) probability distribution can be represented by a factor

- However, a factor *does not need to represent* a particular distribution: it is nothing more than a function from a tuple to a real (or rational)

$$f : X_1 \times \ldots . \times X_k \rightarrow R$$

Factor product

- $f_1(A,B) \times f_2(B,C) = f_3(A,B,C)$
  where $f_3(a,b,c) = f_1(a,b) \times f_2(b,c)$
  for all $a \in A$, $b \in B$ and $c \in C$

f_1

a_1	b_1	0.5
a_1	b_2	0.8
a_2	b_1	0.1
a_2	b_2	0
a_3	b_1	0.3
a_3	b_2	0.9

x

f_2

b_1	c_1	0.5
b_1	c_2	0.7
b_2	c_1	0.1
b_2	c_2	0.2

=

f_3

a_1	b_1	c_1	0.5*0.5 = 0.25
a_1	b_1	c_2	0.5*0.7 = 0.35
a_1	b_2	c_1	0.8*0.1 = 0.08
a_1	b_2	c_2	0.8*0.2 = 0.16
a_2	b_1	c_1	0.1*0.5 = 0.05
a_2	b_1	c_2	0.1*0.7 = 0.07
a_2	b_2	c_1	0*0.1 = 0
a_2	b_2	c_2	0*0.2 = 0
a_3	b_1	c_1	0.3*0.5 = 0.15
a_3	b_1	c_2	0.3*0.7 = 0.21
a_3	b_2	c_1	0.9*0.1 = 0.09
a_3	b_2	c_2	0.9*0.2 = 0.18

Factor marginalization

- Summing out a factor: $\sum_{B} f_3(A,B,C) = f_4(A,C)$

f_3

a_1	b_1	c_1	0.5*0.5 = 0.25
a_1	b_1	c_2	0.5*0.7 = 0.35
a_1	b_2	c_1	0.8*0.1 = 0.08
a_1	b_2	c_2	0.8*0.2 = 0.16
a_2	b_1	c_1	0.1*0.5 = 0.05
a_2	b_1	c_2	0.1*0.7 = 0.07
a_2	b_2	c_1	0*0.1 = 0
a_2	b_2	c_2	0*0.2 = 0
a_3	b_1	c_1	0.3*0.5 = 0.15
a_3	b_1	c_2	0.3*0.7 = 0.21
a_3	b_2	c_1	0.9*0.1 = 0.09
a_3	b_2	c_2	0.9*0.2 = 0.18

f_4

a_1	c_1	0.25+0.08 = 0.33
a_1	c_2	0.35+0.16 = 0.51
a_2	c_1	0.05+0 = 0.05
a_2	c_2	0.07+0 = 0.07
a_3	c_1	0.15+0.09 = 0.24
a_3	c_2	0.21+0.18 = 0.39

Factor reduction

- $f_3(A,B,c_1) = f_5(A,B)$

f_3

a_1	b_1	c_1	0.5*0.5 = 0.25
a_1	b_1	c_2	0.5*0.7 = 0.35
a_1	b_2	c_1	0.8*0.1 = 0.08
a_1	b_2	c_2	0.8*0.2 = 0.16
a_2	b_1	c_1	0.1*0.5 = 0.05
a_2	b_1	c_2	0.1*0.7 = 0.07
a_2	b_2	c_1	0*0.1 = 0
a_2	b_2	c_2	0*0.2 = 0
a_3	b_1	c_1	0.3*0.5 = 0.15
a_3	b_1	c_2	0.3*0.7 = 0.21
a_3	b_2	c_1	0.9*0.1 = 0.09
a_3	b_2	c_2	0.9*0.2 = 0.18

$\Rightarrow$

f_5

a_1	b_1	0.25
a_1	b_2	0.08
a_2	b_1	0.05
a_2	b_2	0
a_3	b_1	0.15
a_3	b_2	0.09

Why factors?

- Fundamental building block for defining distributions in high-dimensional spaces
- Set of basic operations for manipulating these probability distributions
- We will use factors in our inference algorithm
- You will need to represent and compute with factors in the third programming assignment
- Assignment 3a: represent / compute with factors

Important highlights in this lecture

- (Conditional) Independence
  - Know what it means and how to compute it!
  - Know how a graphical model represents independences (as D-Map, I-Map, and P-Map)
  - Know and be able to use D-separation and D-connection
- Bayesian networks
  - Understand what they represent: variables, joint probability distribution, (in)dependences in the distribution
  - Know and understand the product rule property in Bayesian networks (to eliminate dependences in the product rule for computing joint distributions)
  - Go from joint probability distribution to network and v.v.