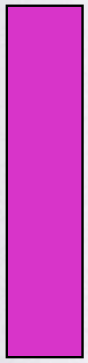
uncertainty

- uncertainty is inevitable in complex worlds
  - **logic**: uncertainty only comes in via disjunction, existential quantifier
  - **laziness**: exceptionless rules are too hard to write down or use
  - **ignorance**: lack of complete theory, insufficient data
  - probability as degree of belief summarizes uncertainty due to laziness and ignorance



a decision-theoretic agent

- computes updated probabilities for current state, given the evidence
- computes outcome probabilities for actions, given probabilities of current states
- selects action with highest expected utility, given outcome probabilities and preference/utility information

uncertainty

remove uncertainty thru assumptions

assume noise-free data

some vision programs, some ml programs, can't handle noisy data …

but noise can be informative - shadows provide useful constraints

assume relevant data

ml systems assume all data provided are instances **relevant** to task at hand;

notice how different this is from a system that has to learn 'on-line' and has to

decide which of its sensory inputs are **salient**!

circumscription: assume that only explicitly mentioned factors are relevant

frame assumption: assume that the only relevant side effects of an action are explicitly mentioned

betting on all horses

in case of uncertainty about alternatives without prospect for further evidence,

e.g. find least number of drugs to cover **all** diagnoses ...

probabilities are not certainty factors

- **$P(A \mid B) = r \quad \neq \quad B \rightarrow_r A$**

  the logical statement says: whenever B is true (**regardless** of any other info), A is true with certainty r

  the probability statement applies only when the **only** thing known is B!

  If C is known, we must refer to P(A | B, C) unless we can show that C is conditionally independent of A given B

A: it rained last night; B: my grass is wet;

given B **only**, it is reasonable to conclude A

but if B is deduced from C: the sprinkler was on last night,

then it is not reasonable to conclude A

probability theory

- ingredients

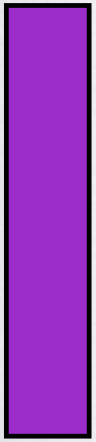
  sample / outcome space Ω represents chance experiments

  basic outcomes (e.g. 'throw a 6') ∈ ! ;

  events (e.g. 'throw even #' = {throw 2, throw 4, throw 6}) ⊆ **Ω**;

  repeated experiments, e.g. 3 throws of coin: **Ω** = {<hhh>, <hht>, …, <ttt>}

  a set A of subsets of Ω represents events [A = Pow(Ω) = {α | α ⊆ Ω}]



A is a (σ) field over Ω iff

(i) Ω ∈ A,

(ii) for each A ∈ A, ¬ A ∈ A,

(iii) if A, B ∈ A, so is A ∪ B ∈ A

(iiia) for every sequence of sets $(A_i)_{i \in N}$ where $A_i \in A$, their infinite union ∈ A

(note 1: a (σ) field is also closed under intersection)

(note 2: since field A contains Ω , it also contains its complement, i.e. ∅.)

Kolmogoroff axioms

- *P* **is a (σ) additive probability space** iff there is a **Ω**, an **A**, a P such that

  P = < **Ω**, **A**, P > and
  Ω is a non-empty set
  A is a (σ) field over **Ω**
  P is a real-valued set function with domain = **A**, and P satisfies

  A1: for all A ∈ **A**, P(A) ≥ 0
  A2: P(**Ω**) = 1
  A3: if A, B ∈ **A** and A ∩ B = ∅, then P(A ∪ B) = P(A) + P(B)
  (and analogous for infinite unions and sums)

- note: P is **uninterpreted!**

interpretations of P

- frequentist
  - probability is frequency of occurrence (in the limit)

    P(event) = 0.99 ≈ event occurs 99 times out of a 100
  - probability is an objective property (of a process, mechanism)
  - probability judgments can be empirically false

- personalist, subjectivist, Bayesian
  - probability is a subjective degree of belief
  - probability judgments can be inconsistent

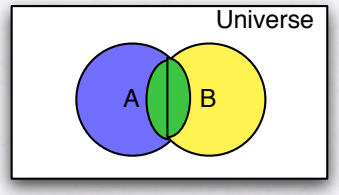
    in order to be **coherent**, probability judgments must adhere to axioms...

simple probability rules

- the axioms again .... for propositions

1. $0 \leq P(A) \leq 1$
2. $P(true) = 1; P(false) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

if you violate the axioms (and bet according to your probabilities), there is a betting strategy against you that guarantees you lose!!

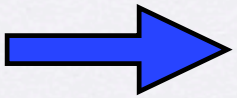


all other properties follow!!
e.g. use ¬A for B in 3:
$P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A)$, i.e. $P(true) = P(A) + P(\neg A) - P(false)$, i.e. $1 = P(A) + P(\neg A)$, i.e. $P(\neg A) = 1 - P(A)$.

- conditional probability:

$$P(A \mid B) = P(A \wedge B) / P(B), \text{ when } P(B) > 0.$$

- Rewrite: $P(A \wedge B) = P(A \mid B)\, P(B)$
  $P(A \wedge B) = P(B \mid A)\, P(A)$



product rule

subjective probabilities

- how to assess subjective probabilities (e.g. that event e occurs)
  choose between 2 gambles:

  1) if event e occurs, I get $100.

  2) I can draw a ball from an urn (with 100 balls) with n red balls and 100 - n white balls. If I draw a red ball, I get $100.

  If all balls are red, I prefer 2); if all balls are white, I prefer 1).

  There is an n for which the 2 gambles are equally attractive; for this n, my probability for e is n/100.

- subjective probabilities must satisfy the axioms or else: *Dutch book!*

random variables and joint distributions

- **random variables** (are neither ...)
  - functions: atomic events → [0, 1]
  - probabilities assigned by random variable to all values sum to 1

RV(value) = 0.2, i.e. P(RV = value) = 0.2

- **joint distribution**
  - probability assignment to all combinations of values (cross product) of *several* random variables
  - joint distribution cannot normally be computed from **marginal** (individual) distributions

	Toothache	¬Toothache
Cavity	0.04	0.06
¬Cavity	0.01	0.89

this table exhausts all possibilities, so all entries must sum to 1.

joint distribution...

- from joint distribution, we can compute everything but ...
- with many random variables, the table soon becomes too big. So?

	Toothache	¬ Toothache
Cavity	0.04	0.06
¬ Cavity	0.01	0.89

P(Cavity) = ? P(Cavity) = 0.1; sum of Cavity row

P(Toothache) = ? P(Toothache) = 0.05 sum of Toothache col

P(A | B) = P(A ∧ B) / P(B), so P(Cavity | Toothache) = 0.04 / 0.05 = 0.8

Bayes' rule

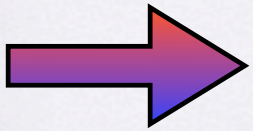
from $P(A \wedge B) = P(A \mid B)\, P(B)$

$P(A \wedge B) = P(B \mid A)\, P(A)$

the fundamental rule!

we get Bayes' rule

$$P(B \mid A) = (P(A \mid B)\, P(B)) / P(A)$$



$$P(hyp \mid evid) = \frac{P(evid \mid hyp)P(hyp)}{P(evid)}$$

this is the basis of probabilistic inference in AI systems.

Useful? We need to know 3 probabilities to compute 1...

updating beliefs with Bayes' Theorem

a common reasoning pattern:

we have a prior (subjective) probability for disease D, then we learn that somebody has a symptom S of D; now we want to update our belief that this person has D:

$$P(D \mid S) = P(S \mid D)\, P(D) / P(S)$$

important:

P(Aids | S) is different in Canada and Niger but

P(S | Aids) is the same!

P(S | D) is a widely applicable **causal** relationship about mechanisms, more useful to learn than P(D | S).

but what about P(S)???

total probability and Bayes'

conditioning (total probability):

$P(A) = P(A \mid B)\, P(B) + P(A \mid \neg B)\, P(\neg B) = P(A \wedge B) + P(A \wedge \neg B)$

'from cause to effect'

$P(\text{effect}) = \sum_j P(\text{effect} \mid \text{cause}_j)\, P(\text{cause}_j)$

(fever could be caused by flu, measles, aids, etc… P(fever) = ∑ …)

Bayes': 'from effect to cause'

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j) \cdot P(A_j)}$$

$P(A_i \mid B)$: P(flu | fever);

$P(B \mid A_i)$: **causal knowledge** [that A_i causes symptom B] is easier to learn and more useful

base rate neglect: everything except likelihood is ignored...

1 in 100 people have disease D.
Test for D has a false positive rate of 0.2, false negative rate of 0.1, so

P(pos | D) = 0.9

You just got a positive test result. **What is the probability you have the disease ? 90%, 80%, 70%???**

P(D | pos) = [P(pos | D) P(D)] / [P(pos | D) P(D) + P(pos | ¬D) P(¬D)]

= [0.9 * 0.01] / [0.9 * 0.01 + 0.2 * 0.99] =
0.009/0.009+0.198 = 0.043 ≈ **4%!!!!**

Gigerenzer's experiments with doctors...

It is recommended that women screen for breast cancer with mammograms.

Probability that a woman between 40 and 50 has breast cancer = 0.8%.
If a woman has breast cancer then it is 90% probable that her mammogram is positive.
If she has no breast cancer then it is 7% probable that she still has a positive mammogram.
What is the probability that she actually has breast cancer?

48 doctors were asked: Estimates range from 1% to 90%, with mean of 70%!!! (1/3 say 90%, 1/3 say 50-80%, 1/3 10% or less. Some of the ones with the right value gave wrong reasons).

This is **serious**!

Gigerenzer got much better results when data were described as **natural frequencies** rather than as **conditional probabilities**.

better reasoning with natural frequencies

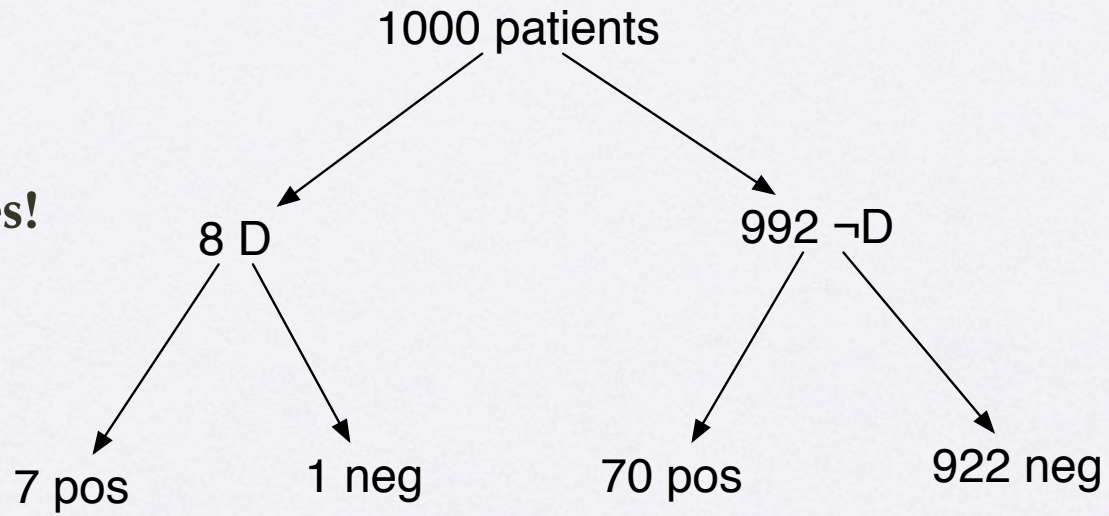
Bayes version: P(D) = 0.008, P(pos | D) = 0.9, P(pos | ¬D) = 0.07, so

$$P(D \mid pos) = \frac{0.008 \bullet 0.9}{0.008 \bullet 0.9 + 0.992 \bullet 0.07}$$

natural frequencies version:
Of 1000 women 8 have breast cancer. Of these 8, 7 have positive mammograms. Of the 992 without cancer, about 70 will also have positive mammograms. Of the 77 women with positive mammograms, how many have really cancer?
Only 7 out of 77, or 1 out of 11, or ca. 9%!!!

better reasoning with natural frequencies

frequency tree instead of Bayes!



P(D | pos) = 7 / (7 + 70). Easy! 😃

good representations make problem solving much easier!

a simple example

you tested positive on test t for **serious** disease d; t is 98% accurate (+/-);
1 person in 1000 has d.
Are you in great danger of having disease d?

normalization (with 1/P(t)): given

1) $P(d|t)=P(t|d)P(d)/\mathbf{P(t)}$ and 2) $P(\neg d|t)=P(t|\neg d)P(\neg d)/\mathbf{P(t)}$; add 1), 2);

since $P(d|t) + P(\neg d|t) = 1$, we get (or just use **conditioning**)

$P(t) = P(t|d)P(d) + P(t|\neg d)P(\neg d)$; substitute into 1):

$$\mathbf{P(d \mid t) = P(t \mid d)\, P(d) \,/\, (P(t \mid d)\, P(d) + P(t \mid \neg d)\, P(\neg d))};$$

[= P(t)]

no need to know P(t), i.e. priors of symptoms!

$$(0.98 * 0.001) / (0.98 * 0.001 + 0.02 * 0.999) = 0.047$$

so: don't worry!

3 doors example

assume: 3 doors, car is behind one of them; wlg, you pick a door, e.g. door1;
moderator opens another door, e.g. door3 (he **knows** car is not behind door3).

Question: if given a choice now , should you switch to door2?

$P(C_i)$ = prob that car is behind door$_i$; $P(C_i \mid M_j)$ = prob that car is behind door$_i$ given that moderator opens door$_j$

$$\mathbf{P(C_2 \mid M_3) = P(M_3 \mid C_2)\, P(C_2) \,/\, \{P(M_3 \mid C_2)\, P(C_2) + P(M_3 \mid C_1)\, P(C_1) + P(M_3 \mid C_3)\, P(C_3)\}}$$

where $P(C_1) = P(C_2) = P(C_3) = 1/3$;

$P(M_3 \mid C_3) = 0$ because that would end the game;

$P(M_3 \mid C_1) = 1/2$ because choice between door2 and door3 is arbitrary;

$P(M_3 \mid C_2) = 1$ because that's the only thing he can do (since you picked door1); thus

$\mathbf{P(C_2 \mid M_3)} = 1 * 1/3 \,/\, \{1/3 + 1/6\} = 1/3 * 2 = \mathbf{2/3}$ and

$\mathbf{P(C_1 \mid M_3)} = 1/6 \,/\, \{1/6 + 1/3\} = 1/6 * 2 = \mathbf{1/3}$.

Thus: you should always switch!!!!

joint probability distributions

for variable A with states $a_1, \ldots, a_n$, P(A) is $(x_1,\ldots,x_n)$, $\sum_i x_i = 1$. (i.e. $P(A = a_i) = x_i$.)

suppose P(A | B):

	b1	b2	b3
a1	.4	.3	.6
a2	.6	.7	.4

suppose P(B) = (.4, .4, .2);

P(A, B) = ?

via **fundamental rule: $P(a_i \mid b_j)\, P(b_j) = P(a_i, b_j)$**

we get **joint probability table** P(A, B)
(for each j, multiply col for b_j in P(A | B) by $P(b_j)$:

	b1	b2	b3
a1	.16	.12	.12
a2	.24	.28	.08

note: sum of all entries = 1

marginalization

get P(A) from table P(A, B): A is in state a_i for m mutually exclusive events $(a_i, b_1),\ldots, (a_i, b_m)$, so (axiom 3): $\mathbf{P(a_i) = \sum_{j=1}^{m} P(a_i, b_j).}$

marginalizing B out of P(A, B):

	b1	b2	b3
a1	.16	.12	.12
a2	.24	.28	.08

$\mathbf{P(A) = \sum_B P(A, B) = P(A, b_1) + \ldots + P(A, b_m)}$

i.e sum the rows in P(A, B): P(A) = (.4, .6)

get P(B | A) from P(A, B): P(B | A) = [P(A | B) P(B)] / P(A)

	a1	a2
b1	.16	.24
b2	.12	.28
b3	.12	.08

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

now divide P(A, B) by P(A)=(.4, .6) $\Rightarrow$

	a1	a2
b1	.4	.4
b2	.3	.47
b3	.3	.13

independence & conditional independence

A and B are **independent** iff

$$P(A \mid B) = P(A)$$
$$P(B \mid A) = P(B)$$
$$P(A, B) = P(A) * P(B)$$

if everything depends on everything else, efficient probabilistic reasoning is impossible!

A and B are **conditionally independent given C** iff

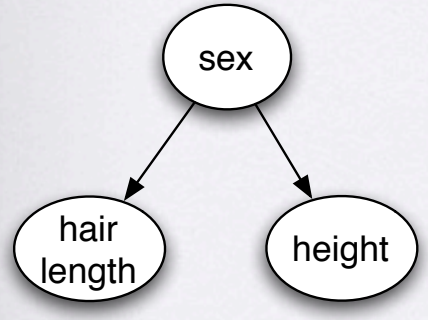
$$P(A \mid B, C) = P(A \mid C)$$
$$P(B \mid A, C) = P(B \mid C)$$
$$P(A, B \mid C) = P(A \mid C) * P(B \mid C)$$

conditional independence example

something is wrong with your car. If the radio works (R), it's likely that the starter will turn over (S). But if ¬R then maybe the battery is dead (¬B), and thus ¬S. So R, S, B are all mutually dependent.

Now you learn that the battery is ok (B). Now R says nothing about S!

R and S are conditionally independent given B.



if we don't know sex, seeing hair length tells us something about sex which tells us something about height. But if we know the person is a man, length of hair tells us nothing about height…
so: hair length and height are **conditionally independent** given sex.

conditional independence and updating

a toothache (T) is a symptom of a cavity (C), and so is a spot on an X-ray (X). But given C, T and X are conditionally independent: having pain is independent of seeing a spot, given the cavity.

Bayesian updating with several pieces of evidence:

$$P(C \mid T, X) = P(T, X \mid C)\, P(C) \;/\; P(T, X)$$

assume conditional independence of T and X given C; then

$$P(C \mid T, X) = P(T \mid C)\, P(X \mid C)\, P(C) \;/\; P(T, X),$$ i.e. conditionally independent symptoms can be multiplied in incrementally, serially, in any order. That's good.

What about P(T, X)?

Bayesian updating: dividing by P(T, X)

normalizing factor:

$$P(C \mid T, X) + P(\neg C \mid T, X) = 1$$

so

$$[P(T \mid C)\, P(X \mid C)\, P(C)]/P(T, X) \;+\; [P(T \mid \neg C)\, P(X \mid \neg C)\, P(\neg C)]/P(T, X) = 1$$

now multiply with P(T, X):

$$P(T \mid C)\, P(X \mid C)\, P(C) \;+\; P(T \mid \neg C)\, P(X \mid \neg C)\, P(\neg C) = P(T, X)$$

this is denominator in Bayes' for 2 pieces of evidence. etc...

$$P(C \mid T,X) = \frac{P(T \mid C)P(X \mid C)P(C)}{P(T \mid C)P(X \mid C)P(C) + P(T \mid \neg C)P(X \mid \neg C)P(\neg C)}$$

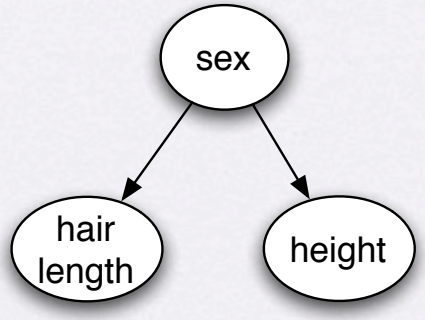
Naive Bayes, joint probability distribution, Bayesian Networks (BN)

Naive Bayes: all variables assumed independent ---> unrealistic

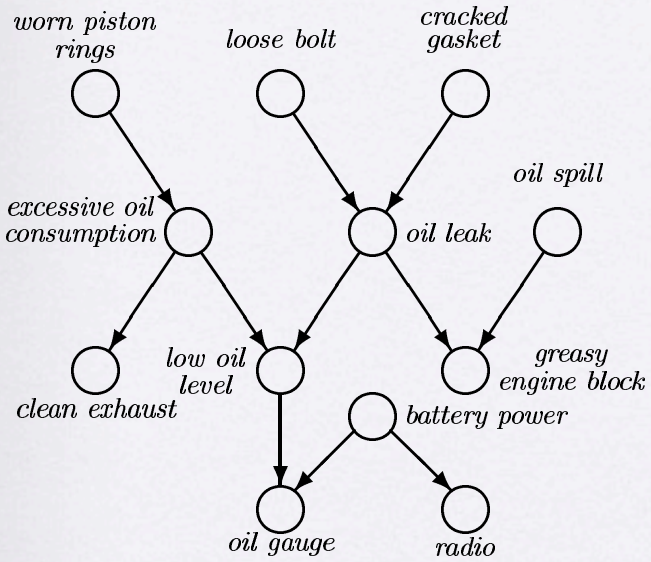
joint probability distribution: usually unavailable ---> unrealistic

Bayesian nets exploit independencies, local structure, in a domain

BN describe **causal dependencies** graphically and thus give a compact representation of the joint probability table



Bayes Belief Nets store joint probability distribution **compactly**



12 propositional variables:

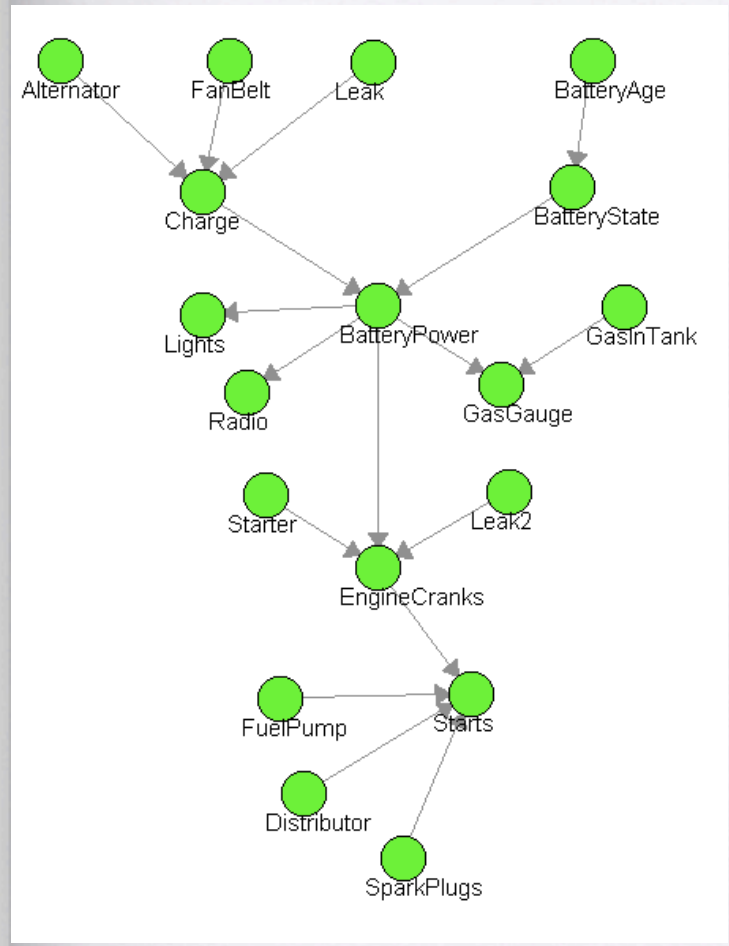
2^{12} = 4096 probabilities needed!

using independencies in the model:

54 (or 27: P(¬X)=1-P(X)) probabilities needed!

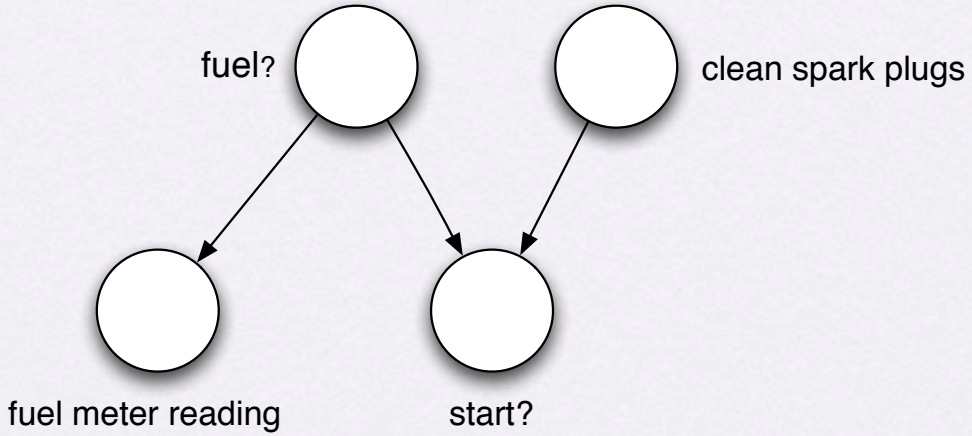
in general: instead of 2^n values for n propositional nodes, we want to exploit independencies in the domain, i.e. separabilities among variables....

causal reasoning



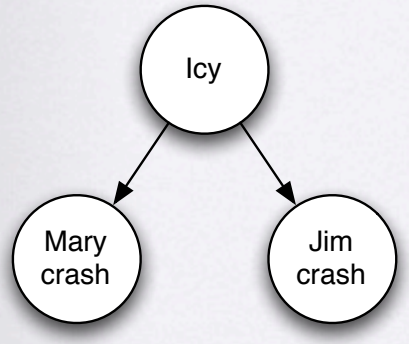
the car doesn't start. The starter turns, so there is power in the battery. Most likely causes:

no fuel or dirty spark plugs. Fuel meter shows there is fuel, so the spark plugs must be dirty… etc

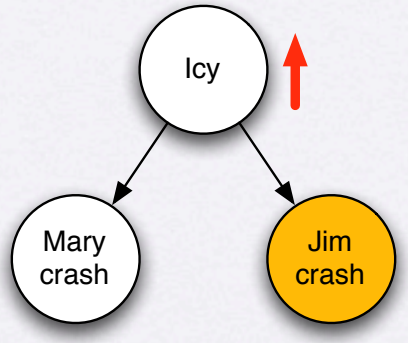


simple causal reasoning: qualitative component of Bayesian Nets

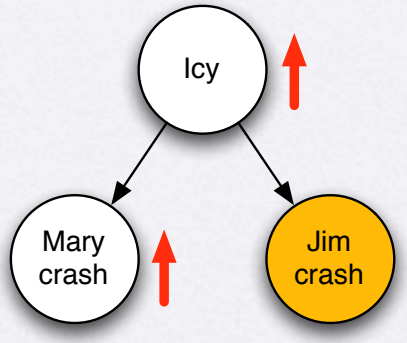
Joe waits for Mary and Jim. They are driving to him in 2 cars. It's winter and the roads might be icy. Now Joe is informed that Jim crashed. Joe thinks 'the roads must be really icy, I guess Mary will crash too'.



Icy makes Mary crash and Jim crash more likely.



Jim crash makes Icy more likely.

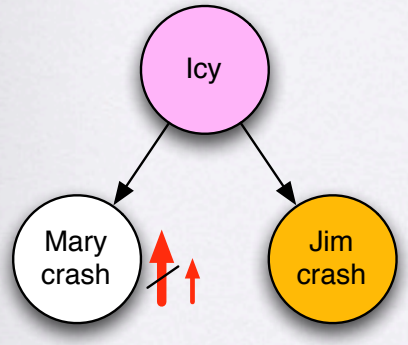


Icy more likely makes Mary crash more likely.

Note: • info flows also against arrow direction!
• Mary crash, Jim crash, Icy are all mutually **dependent**

simple causal reasoning, cont.

Joe waits for Mary and Jim. They are driving to him in 2 cars. It's winter and the roads might be icy. Now Joe is informed that Jim crashed. Joe thinks 'the roads must be really icy, I guess Mary will crash too'. Next he hears that the roads are not icy, so now he has no reason to think that Mary will also crash.

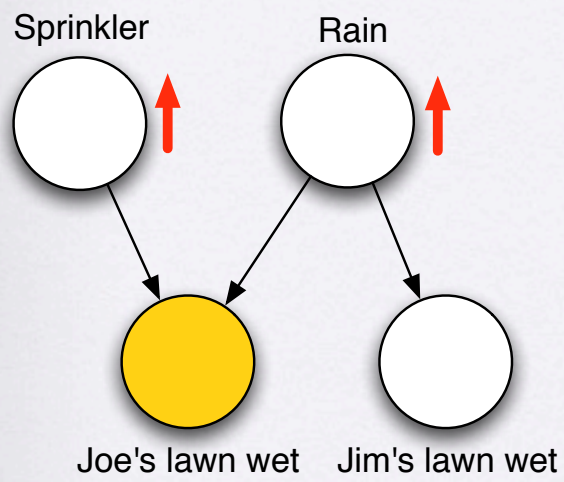


if it's not Icy, then the fact that Jim crash says nothing about whether Mary crash, so Joe's belief in Mary crash is weakened.

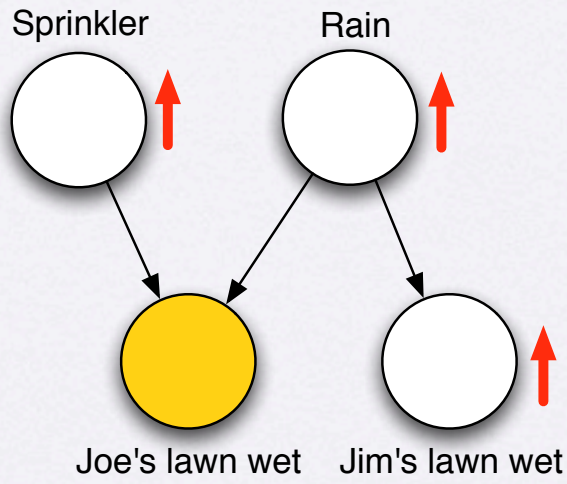
Mary crash and Jim crash are conditionally independent given (¬)Icy.

simple causal reasoning, cont.

Joe sees his lawn is wet. Did it rain or was the sprinkler on? His neighbor Jim's lawn is also wet. So it probably rained.



wet lawn makes both Rain and Sprinkler more likely.

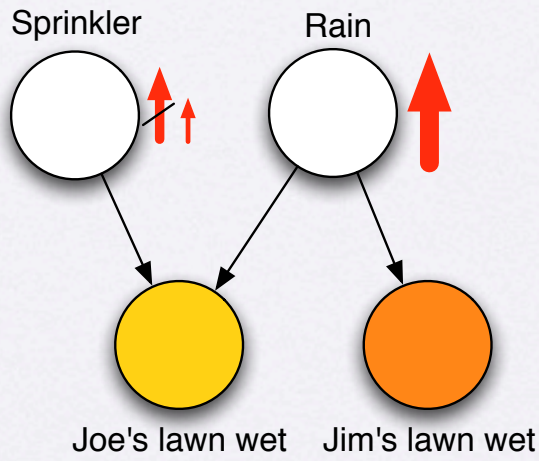


even before seeing Jim's lawn, greater likelihood of Rain makes Jim's lawn wet more likely.

simple causal reasoning, cont.

Joe sees his lawn is wet. Did it rain or was the sprinkler on? His neighbor Jim's lawn is also wet. So it probably rained.

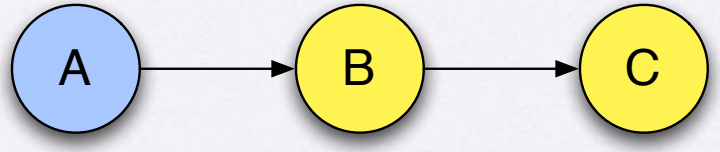
now Joe sees Jim's lawn is wet too,
so **probability of Rain grows (2 pieces of evidence)** and **probability of Sprinkler drops!**



Explaining away: several possible causes; each gets more probable when a symptom is observed; but once a particular cause wins out, the probability of the others drops!

causal networks

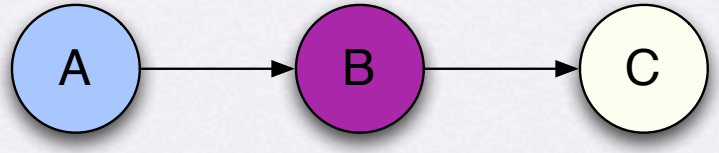
serial connections:



evidence on A influences *certainty* of B (whose truth value is unknown), which then influences *certainty* of C; evidence on C can influence A via B.

if battery dead (A) then car won't start (B), so car won't move (C). Finding that battery is dead tells us something about whether car will or will not move...

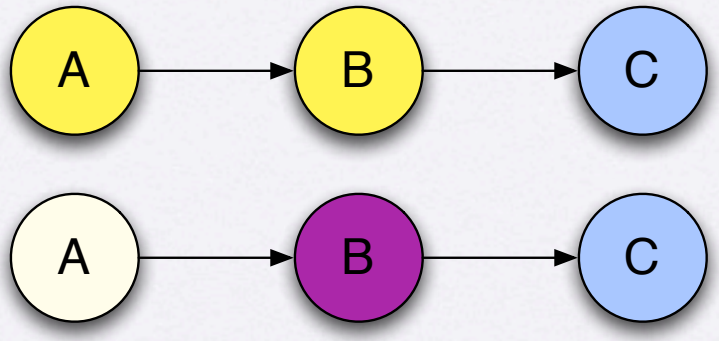
If B is known (**instantiated**), the channel is blocked:
A and C are **d-separated** given B (independent given B).



if we know B (car won't start), knowing about A (battery dead) tells us nothing about C....

causal networks, cont.

backward serial connections:

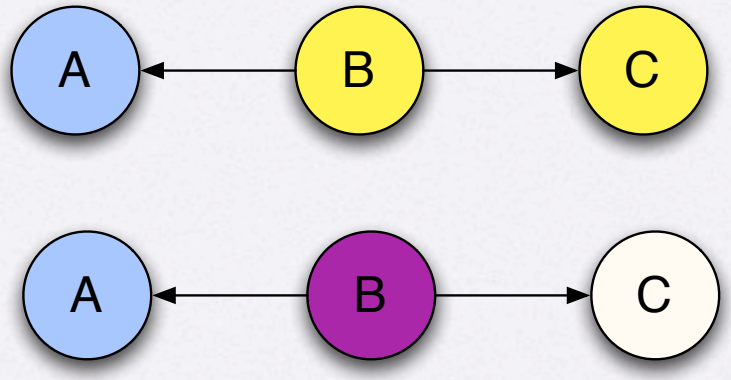


knowing about C tells us something about A but if we know B, knowing about C does not propagate back to A.

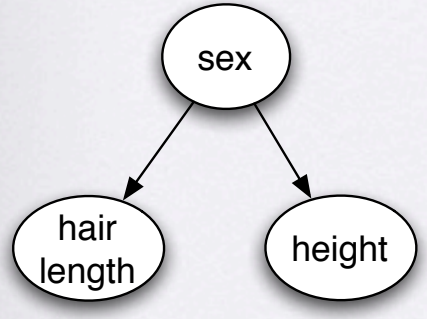
If you know that car won't start (B), then finding that it won't move (C), tells you nothing about whether the battery is dead (A)

causal networks

diverging connections:



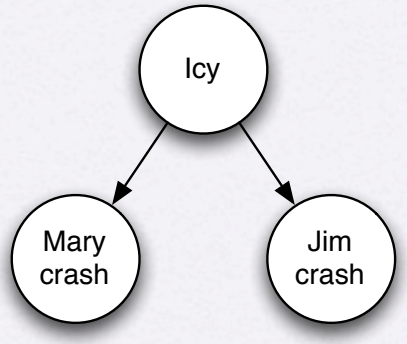
knowing about A propagates through B to C;
knowing about C propagates through B to A; **but**
if B is known, propagation in both directions is blocked



if we don't know sex, seeing hair length tells us something about sex which tells us something about height. But if we know the person is a man, length of hair tells us nothing about height…
so: hair length and height are **conditionally independent** given sex.

causal networks

diverging connections:



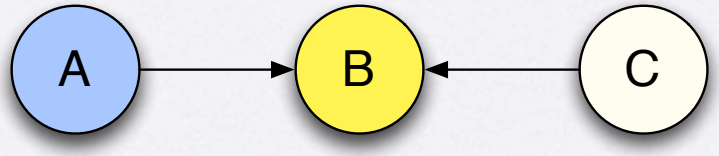
knowing about either child tells us something about the other but if we know the root (Icy), the information flow is stopped.

The children are d-separated given the root.

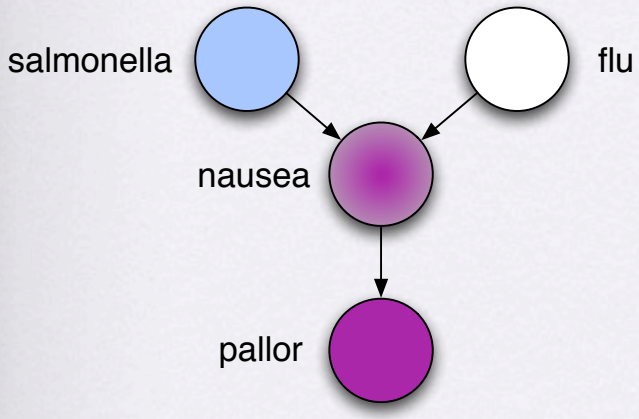
serial and **diverging** connections behave the same wrt evidence propagation: arrows can be turned around in a BN as long as no converging connections are made or destroyed!

causal networks

converging connections:



knowing of one possible cause of an event tells us nothing about other possible causes; if all we know about B is inferred from knowledge of its parents, then the parents are independent: evidence on one of them has no influence on certainty of others. But: suppose B **has** occurred **(soft** evidence), and A and C may cause it; if we learn that A has occurred**,** certainty of C decreases…
explaining away...
Evidence flows ('negatively') from A to C only if B or a child of B is instantiated.



if we know nothing of nausea or pallor, then info about salmonella tells us nothing about flu. But if person is pale, then info that he/she has salmonella weakens belief in flu…

d-separation (basic to human reasoning!)

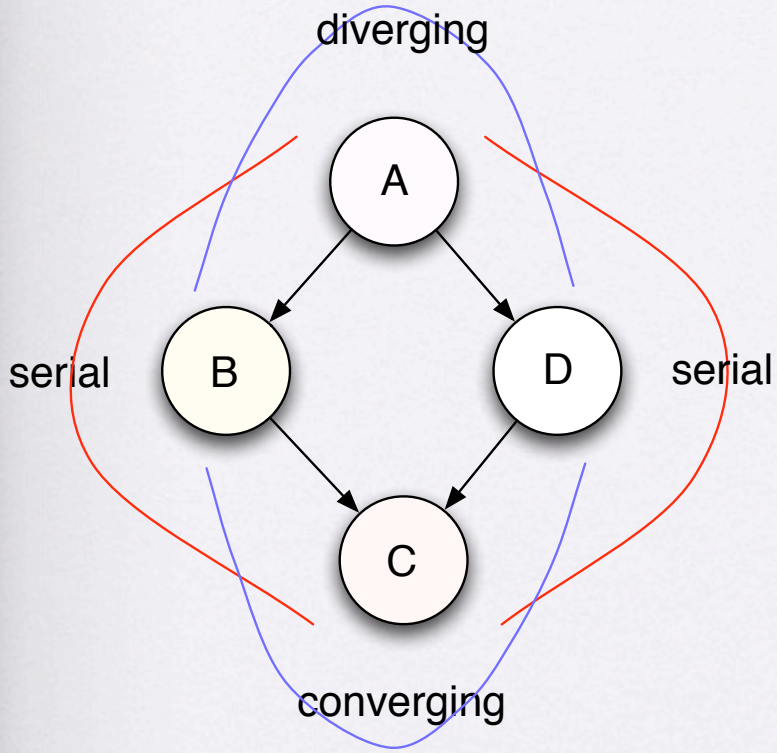
2 variables A and B are **d-separated** if, for **all** paths between A and B, there is an intermediate variable V such that either

- connection is serial or diverging and V is instantiated (fully known) or
- connection is converging and neither V nor any of its descendants have received evidence.

if A and B are d-separated, then changes in certainty of one have no impact on certainty of other.

Remember: d-separation is conditional independence....

d-separation



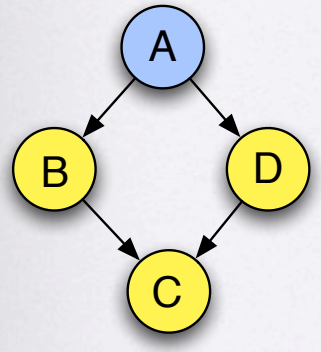
A-B-C: serial, blocked when B known

A-D-C: serial, blocked when D known

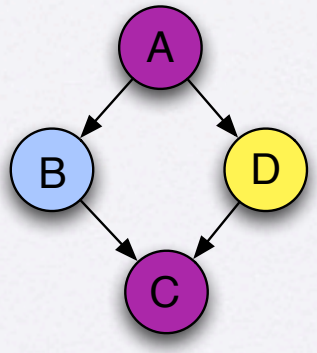
B-A-D: diverging, blocked when A known

B-C-D: converging, blocked when C **not** known

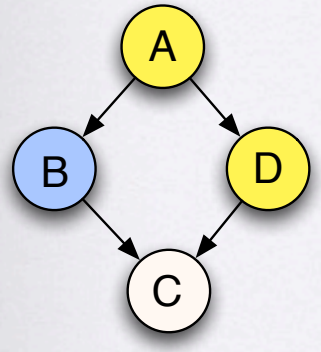
d-separation



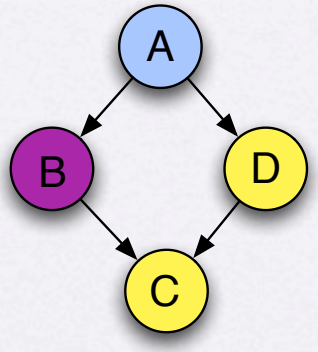
(info enters at A) A, C connected via A,B,C and A,D,C.



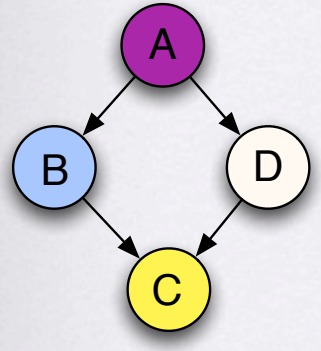
(A, C instantiated) **B, D connected**: B,A,D blocked, **B,C,D connected**.



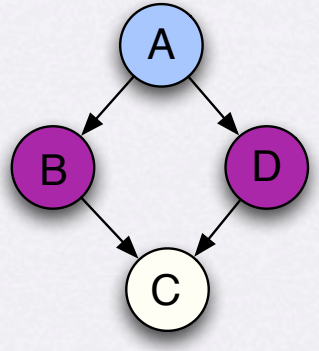
(info enters at B) A, C connected via A,B,C and A,D,C. B, D connected via B,A,D. B,C,D **blocked**.



(B instantiated) **A, C connected**: A,D,C connected, A,B,C blocked.



(A instantiated) **B, D d-sep**: B,A,D blocked, B,C,D blocked.

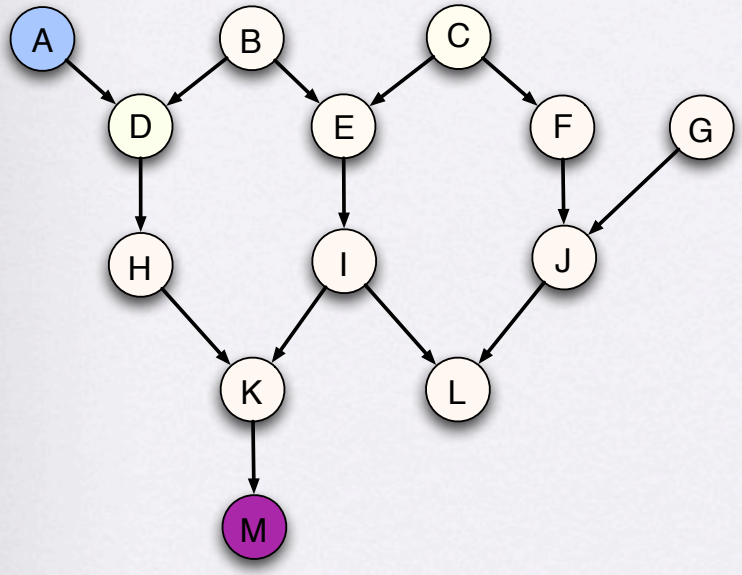


(B, D instantiated) **A, C d-sep**: A,B,C blocked, A,D,C blocked.

d-separation example

if M is fully known, is A d-separated from E?

Are all paths between A and E blocked?

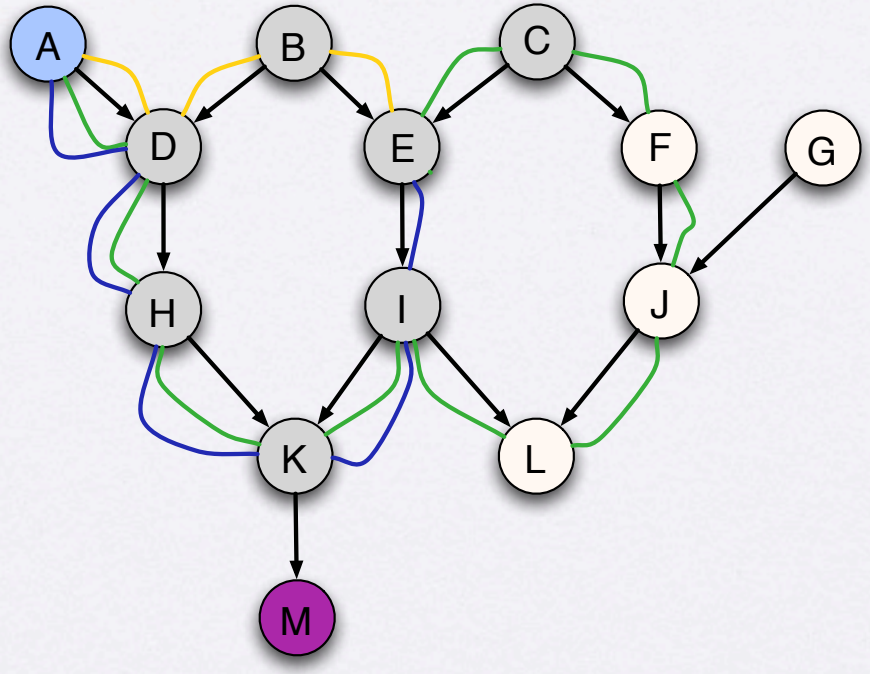


since M is instantiated, we indicate all nodes that get some evidence, i.e. that have an instantiated descendant: K,H,I,D,E,B,C.

d-separation example, cont.

there are 3 paths:
ADBE
ADHKIE
ADHKILJFCE

ADHKILJFCE not ok: ILJ is converging but L has no evidence; blocked.
ADBE ok: ADB is converging but there is evidence at D, so info flows through, and diverging DBE is ok because B is not instantiated (although there is evidence at B).
ADHKIE ok: ADHK all serial, not instantiated, so info flows; HKI converging but K has evidence, so info flows; KIE backward serial.



A and E are d-connected.
Are A and F d-separated if M is instantiated?
Are A and E d-separated if I is instantiated?
When are A and G d-separated?

Bayesian networks: compact representation of joint probability table

Bayesian Network (BN):

- set of variables and directed edges
- each variable has finite set of mutually exclusive states
- DAG: directed acyclic graph (no feedback loops!)
- each variable A with parents B1,…, Bn has an associated table P(A | B1,…, Bn) [a variable without parents has table P(A)]
- the d-separation properties implied by the structure must hold [if A and B are d-separated given evidence e, then the underlying calculus must yield P(A | e) = P(A | e, B)]
- links go (typically) in a causal direction
- the chain rule is supported in the graphical model

chain rule

$U = \{A_1, \ldots, A_n\}$. JPT $P(U) = P(A_1, \ldots, A_n)$ grows exponentially with n. Bayesian Nets store P(U) **compactly**, because

Theorem: Let BN be a Bayesian Network over $U = \{A_1, \ldots, A_n\}$. Then the joint prob. distribution P(U) is product of all potentials in BN, i.e.

$$\mathbf{P(U) = \prod_i P(A_i \mid \text{parents}(A_i))}$$

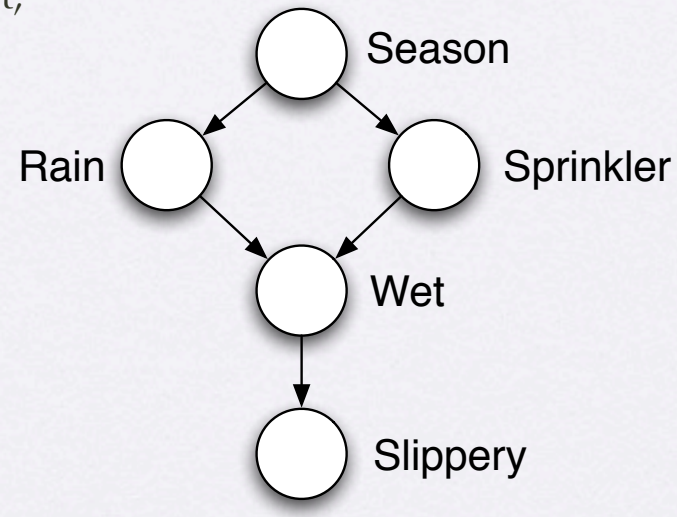
If A is **serially** connected to C via B, then P(C | A, B) = P(C | B).
Proof: via Chain rule, P(A,B,C) = P(A) P(B | A) P(C | B) = P(A, B) P(C | B);
then **P(C | B, A)** = P(A, B, C) / P(A, B) = (P(A) P(B | A) P(C | B))/(P(A) P(B | A))
= **P(C | B).**

Converging connection: if A and B are parents of C, P(A | B) = P(A),
and so P(A, B) = P(A) P(B).
Proof: P(A, B, C) = P(A) P(B) P(C | A, B). By distributivity,
$P(A, B) = \sum_C P(A) P(B) P(C \mid A, B) = P(A) P(B) \sum_C P(C \mid A, B)$. Since columns in prob table sum to 1, $\sum_C P(C \mid A, B)$ is table of 1's; thus, **P(A, B) = P(A)P(B)**.

making a simple Bayes net

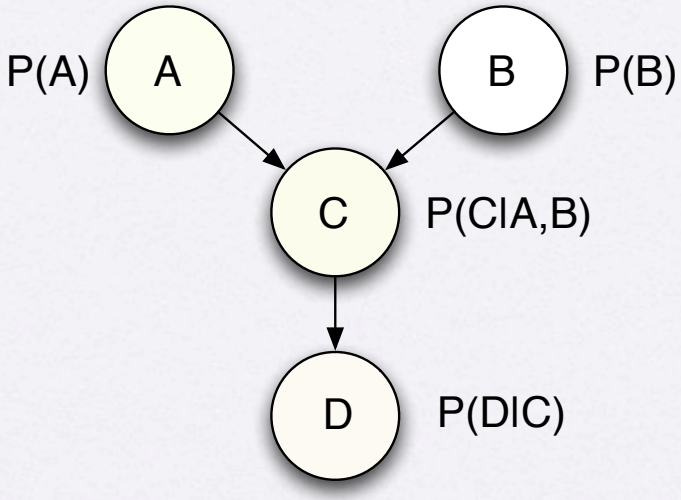
a set of variables PA_j are the **parents** of X_j if PA_j is a **minimal set of predecessors** of X_j that makes X_j **independent** of all other predecessors.

Slippery, Season (of year), Rain, Wet (pavement), Sprinkler (on).
Ok, start with **root causes** ...
given X1, X2, draw an arrow iff they are dependent;
X3: no arrow if it is independent of {X1,X2}};
if X2 screens X3 from X1,
draw arrow from X2 to X3,
if X1 screens X3 from X2,
draw arrow from X1 to X3,
if no screening,
draw arrows from X1 and X2 to X3.
At any stage,
draw arrow from each member of PA_j to X_j.



Note: if we **observe** Sprinkler = On, we infer Season = dry, maybe Rain = no, but if we **do**(Sprinkler = On), arrow from Season to Sprinkler is dropped (new mechanism!), and we make no such inferences!

chain rule example



condition only on the parents and multiply!

- P(A,B,C,D) = P(D | A,B,C) P(A,B,C)

given C, A is d-sep from D and B is d-sep from D, so P(D | A,B,C) = P(D | C):

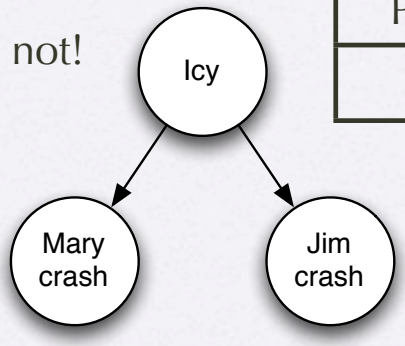
- P(A,B,C,D) = P(D | C) P(A,B,C)

P(A,B,C) = P(C | A,B) P(A,B);
P(C | A,B) is already in the table. What about P(A,B)? A and B are independent (converging!), so P(A,B) = P(A) P(B). Thus:

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

quantitative component of Bayes nets, some examples

note: rows sum to 1, cols need not!

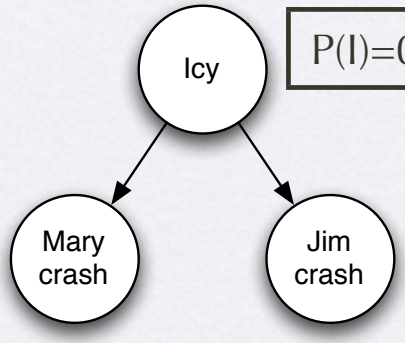


P(I=t)	P(I=f)
0.7	0.3

	P(M=t|I)	P(M=f|I)
I=t	0.8	0.2
I=f	0.1	0.9

	P(J=t|I)	P(J=f|I)
I=t	0.8	0.2
I=f	0.1	0.9

note: conditional probabilities for Mary crash and Jim crash just happen to be the same.

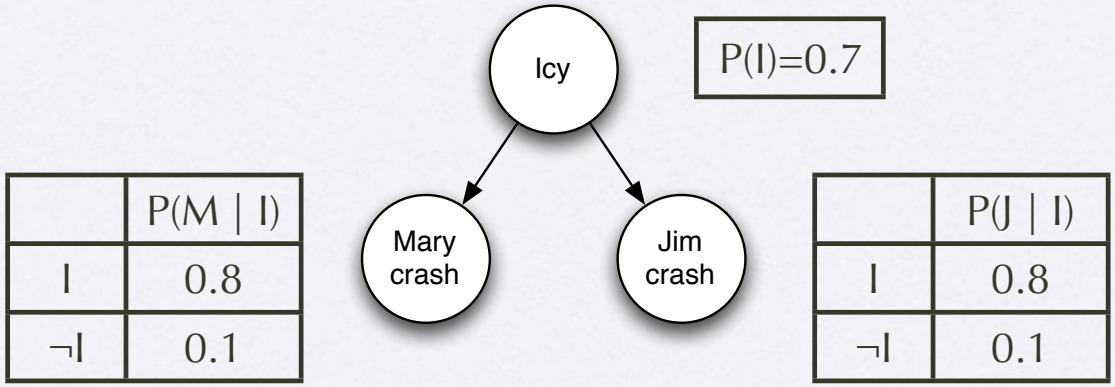


P(I)=0.7

	P(M|I)
I	0.8
¬I	0.1

	P(J|I)
I	0.8
¬I	0.1

simple example



prior P(J) = ?

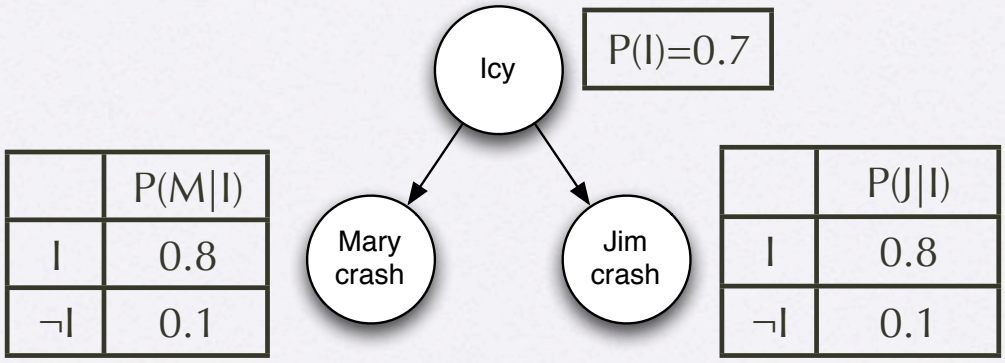
P(J) = P(J | I)P(I) + P(J | ¬I)P(¬I) = 0.8*0.7 + 0.1*0.3 = 0.59, i.e. Jim has a .6 chance of crashing no matter what -;)

now we hear that Jim crashed. How to change our belief in Icy? **P(I | J) = ?**

to go against the arrows, use Bayes theorem!

P(I | J) = P(J | I) P(I) / P(J) = 0.8*0.7 / 0.59 = 0.95;
initially we thought P(I)=0.7 but after hearing about Jim, it's gone up to 0.95..

simple example, cont.



P(M | J) = ?? =
P(M | J,I) P(I | J) + P(M | J,¬I) P(¬I | J) (**conditioning!**)

now: **M is d-sep from J given I**, so P(M | J,I) = P(M | I), P(M | J,¬I) = P(M | ¬I)!

P(M | J) = P(M | I)P(I | J) + P(M | ¬I)P(¬I | J) = 0.8*0.95+0.1*0.05=**0.765**

now we learn that it's not icy;
so P(M | ¬I, J) = P(M | ¬I) = **0.1** (M and J are d-sep given knowledge of I).

questions to answer with Bayes nets

- given some instantiated variables e, what's the probability that X has x?

P(x | e) = ? also, in general: P(X | e) = ?

- **maximum aposteriori probability (MAP):**

what value x maximizes P(X | e)?
What's the most probable explanation for some evidence?

MAP (maximum a posteriori) hypothesis

$\mathbf{h_{MAP}} = \text{argmax}_{h \in H} P(h \mid e) = \mathbf{argmax_{h \in H}\ P(e \mid h)\ P(h)}$

(denominator P(e) dropped because it's a constant independent of h)

ML (maximum likelihood) hypothesis

$\mathbf{h_{ML} = argmax_{h \in H}\ P(e \mid h)}$ (P(h) is dropped **if** all h ∈ H have the same probability).

brute force MAP or ML learning (impossible for large H)

for each h ∈ H, compute P(h | e) = P(e | h) P(h) (or just P(e | h))

output **a** hypothesis h_{MAP} or h_{ML} as one maximally probable hypothesis given all the data e