

Inteligencia Artificial

Redes bayesianas



[Transparencias adaptadas de Dan Klein and Pieter Abbeel: CS188 Intro to AI, UC Berkeley (ai.berkeley.edu)]

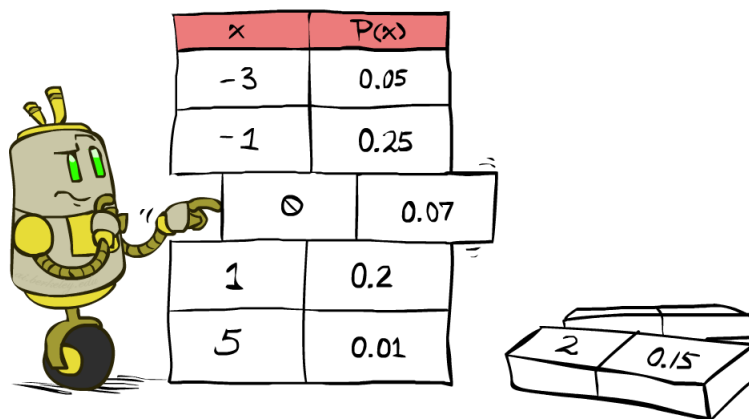
Resumiendo: Probabilidad elemental

- Leyes básicas: $0 \leq P(\omega) \leq 1$ $\sum_{\omega \in \Omega} P(\omega) = 1$
- Sucesos: subconjuntos de Ω : $P(A) = \sum_{\omega \in A} P(\omega)$
- La variable aleatoria $X(\omega)$ tiene un valor en cada ω
 - La distribución $P(X)$ expresa la probabilidad para cada valor posible x
 - La distribución conjunta $P(X,Y)$ expresa la probabilidad total por cada combinación x,y
- Marginalización: $P(X=x) = \sum_y P(X=x,Y=y)$
- Probabilidad condicional: $P(X|Y) = P(X,Y)/P(Y)$
- Regla del producto: $P(X|Y)P(Y) = P(X,Y) = P(Y|X)P(X)$
 - Generalización a la regla de la cadena: $P(X_1, \dots, X_n) = \prod_i P(X_i | X_1, \dots, X_{i-1})$

Resumendo: Inferencia por enumeración

- Dado un modelo de probabilidad $P(X_1, \dots, X_n)$
- Particiona las variables X_1, \dots, X_n en conjuntos como sigue:
 - Variables de evidencia: $E = e$
 - Variables de consulta: Q
 - Variables ocultas: H
 - Queremos: $P(Q \mid e)$

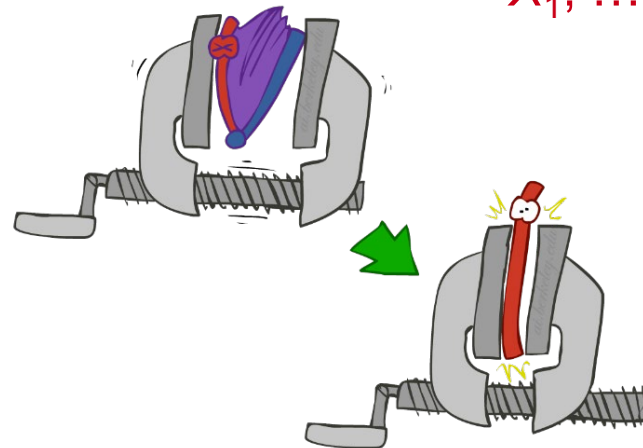
- Paso 1: Seleccionar las entradas coherentes con las evidencias



x	P(x)
-3	0.05
-1	0.25
0	0.07
1	0.2
5	0.01

- Paso 2: Marginalizar H del modelo para obtener la unión de la consulta y la evidencia

$$P(Q, e) = \sum_h \underbrace{P(Q, h, e)}_{X_1, \dots, X_n}$$



- Paso 3: Normalizar

$$P(Q \mid e) = \alpha P(Q, e)$$

Resumendo: Inferencia por enumeración

- ¿ $P(W)$?

Consulta

Evidencia

- ¿ $P(W \mid E=\text{invierno})$?

E (Estación)	Temp	W (Tiempo)	P
verano	calor	sol	0,30
verano	calor	lluvia	0,04
verano	calor	niebla	0,01
verano	calor	meteor.	0,00
verano	frío	sol	0,10
verano	frío	lluvia	0,02
verano	frío	niebla	0,03
verano	frío	meteor.	0,00
invierno	calor	sol	0,10
invierno	calor	lluvia	0,02
invierno	calor	niebla	0,03
invierno	calor	meteor.	0,00
invierno	frío	sol	0,15
invierno	frío	lluvia	0,12
invierno	frío	niebla	0,08
invierno	frío	meteor.	0,00

Resumendo: Independencia condicional

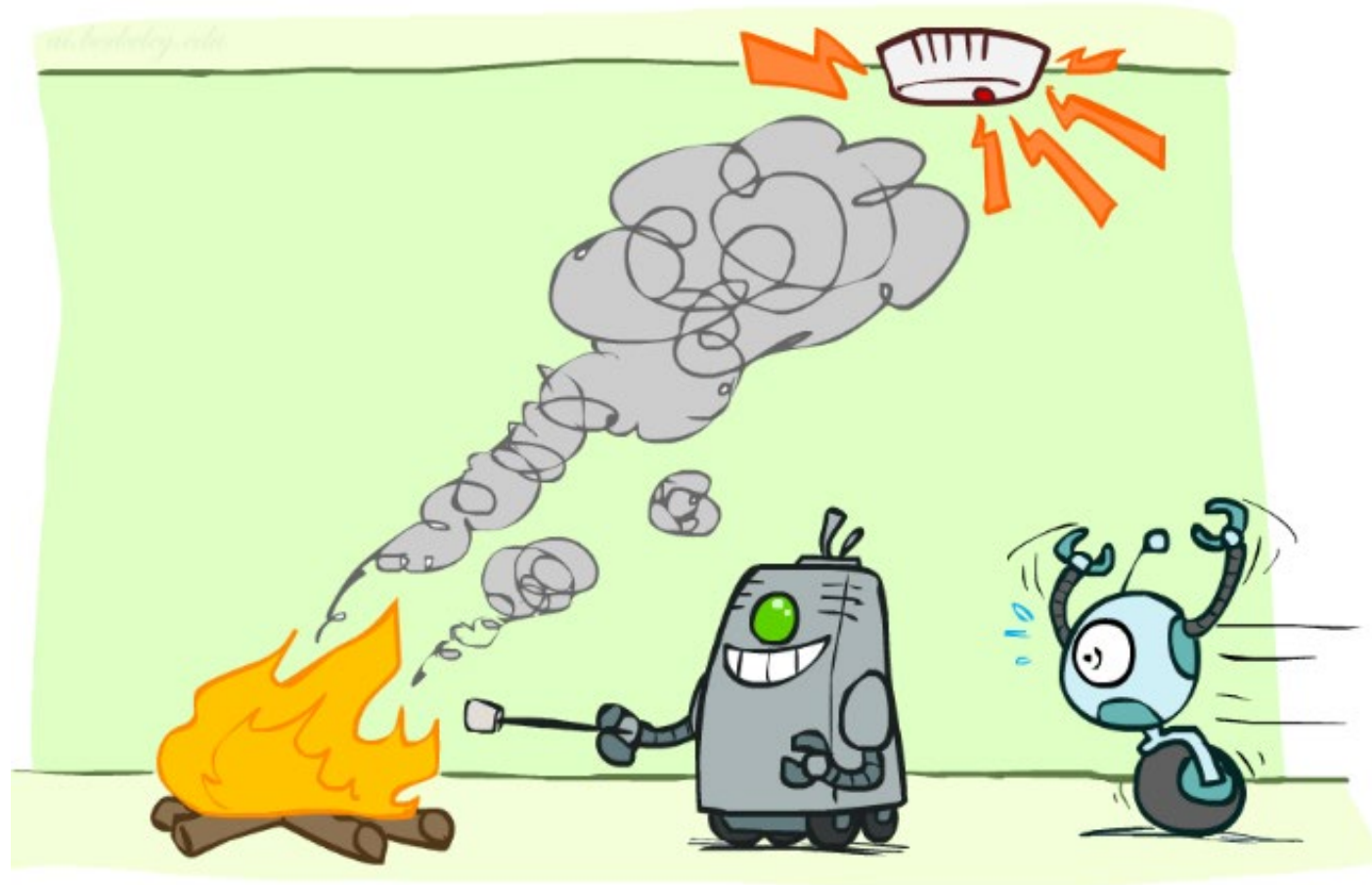
- Aunque la inferencia por enumeración puede calcular probabilidades para cualquier consulta que deseemos, representar una distribución conjunta completa en la memoria de un ordenador es poco práctico para los problemas reales (tabla con d^n números para n variables con rango d)
- La **independencia condicional** permite simplificar esto
- X es condicionalmente independiente de Y dado Z si y solo si:

$$\forall x, y, z \quad P(x \mid y, z) = P(x \mid z)$$

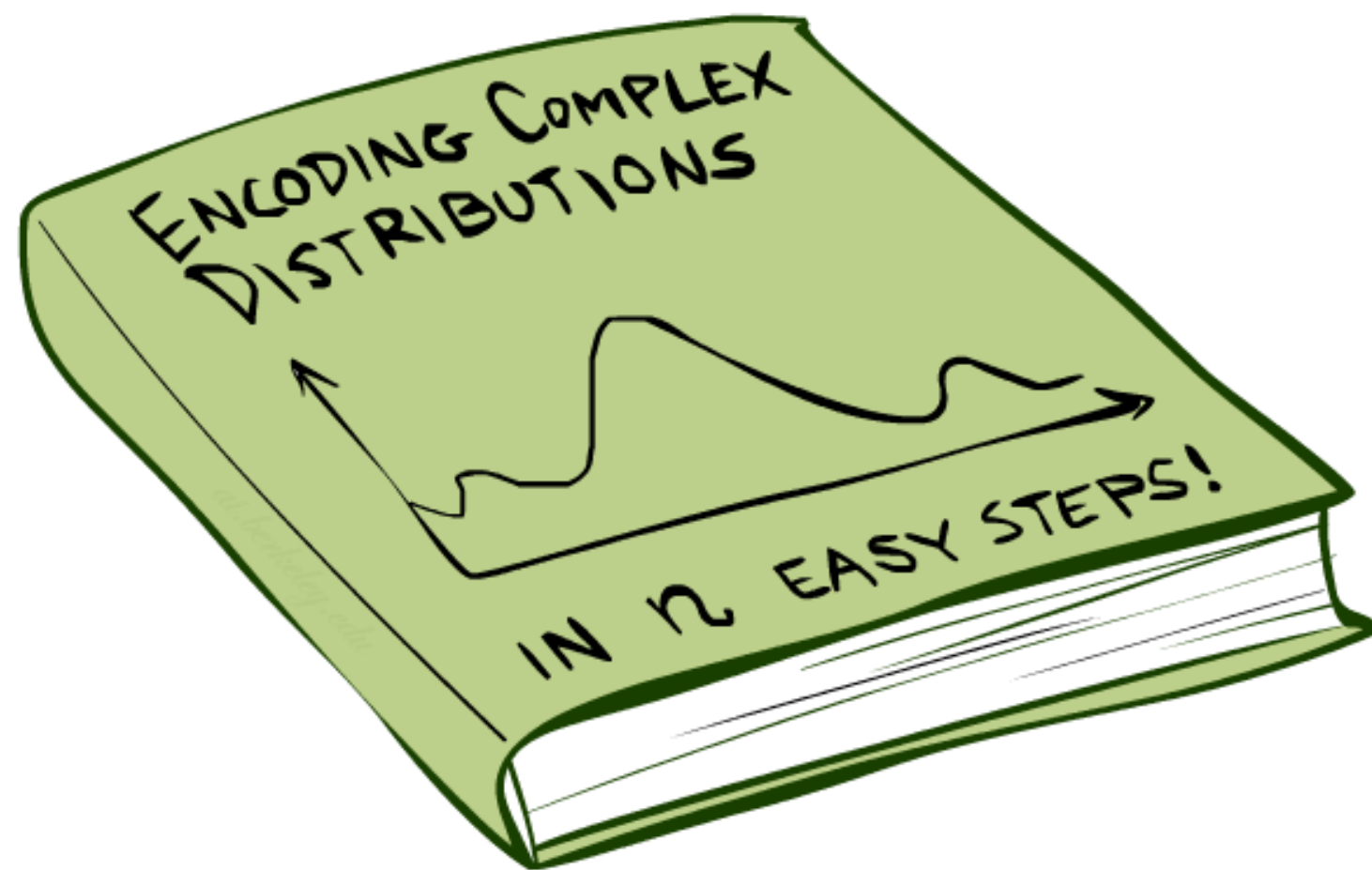
o, equivalentemente, si y solo si

$$\forall x, y, z \quad P(x, y \mid z) = P(x \mid z) P(y \mid z)$$

Recapitulando: Independencia condicional

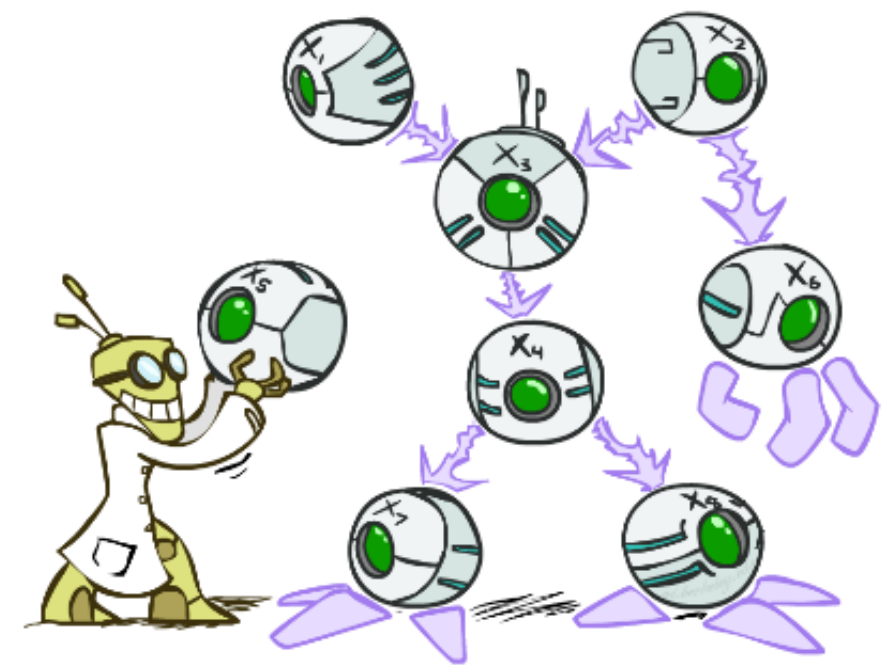


Redes bayesianas: introducción



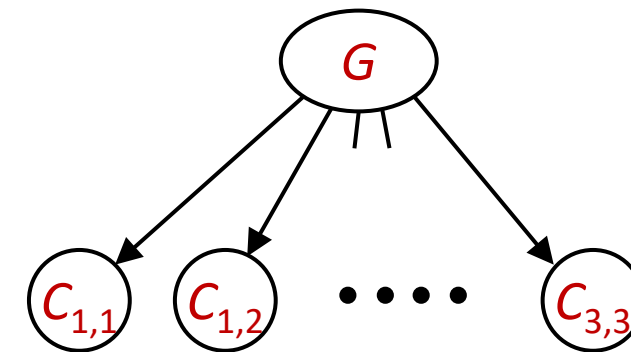
Redes bayesianas: introducción


- **Redes bayesianas**: técnica para describir **distribuciones (modelos) conjuntas complejas** usando **distribuciones condicionales simples**.
 - Un subconjunto de la clase general de **modelos gráficos**
- Usan la causalidad local/independencia condicional:
 - el mundo se compone de muchas variables,
 - cada una interactuando localmente con otras pocas
- Contenido de este bloque
 - Representación
 - Inferencia



Redes bayesianas: introducción

- **Nodos:** variables (con dominios)
 - Pueden tener un valor asignado (observadas) o no (no observadas)
- **Arcos:** interacciones
 - Indican la "influencia directa" entre variables
 - Formalmente: la ausencia de arco codifica la independencia condicional (lo veremos más adelante)
- **Probabilidades condicionales**
 - $P(\text{variable} \mid \text{padres})$
 - CPT (tablas de probabilidad condicional)



0.11		0.11
0.11	0.11	0.11
0.11	0.11	0.11

Ejemplo: lanzar moneda al aire

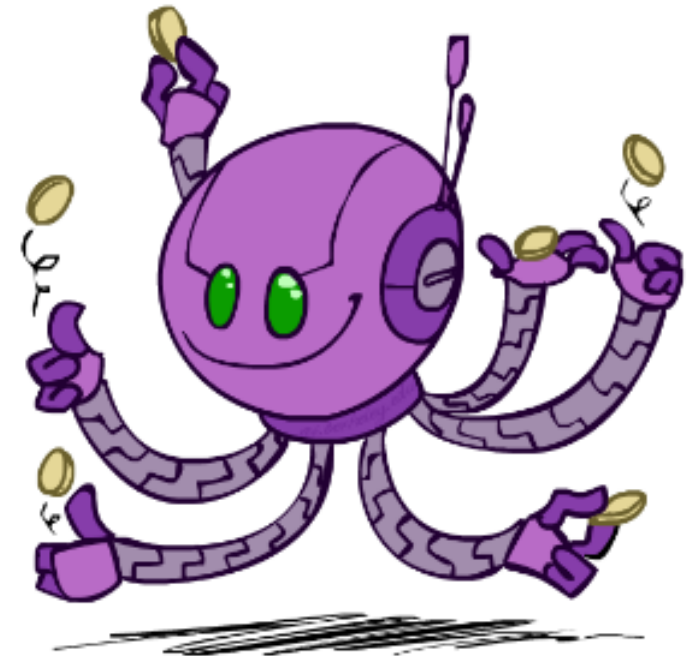
- n tiradas de moneda al aire independientes

X_1

X_2

...

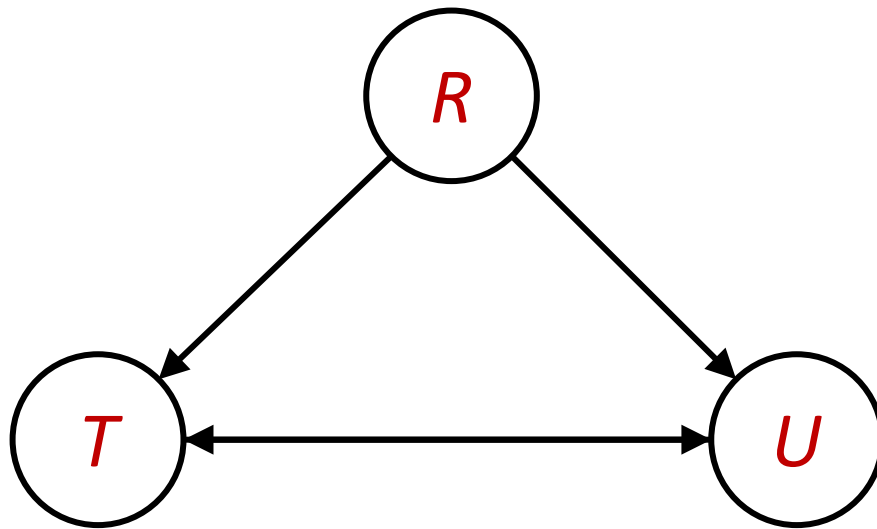
X_n



- Sin interacciones entre variables: independencia absoluta

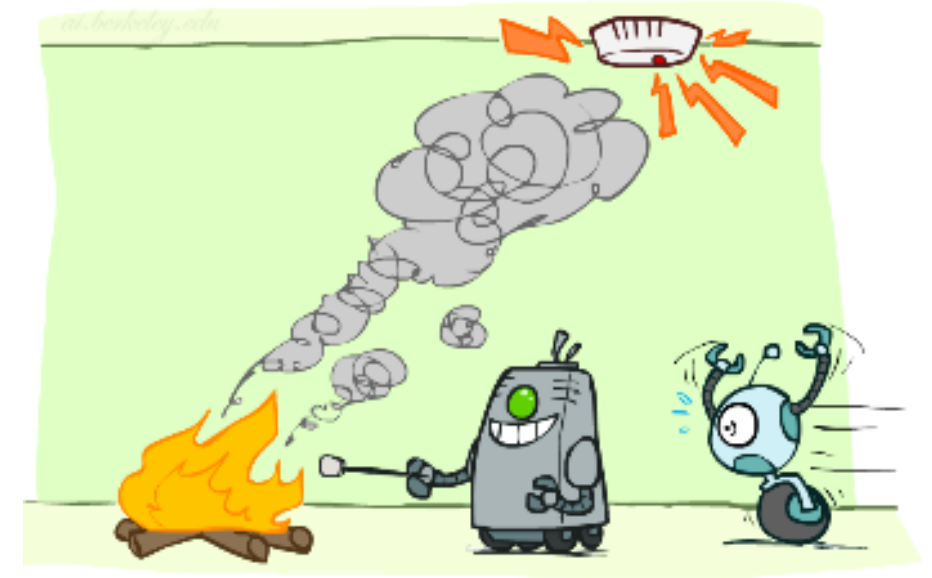
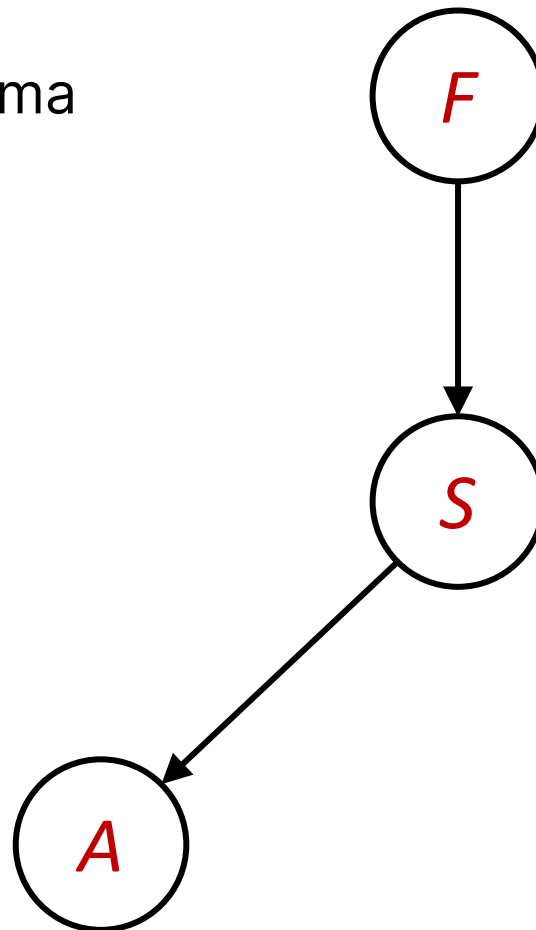
Ejemplo: tráfico

- Variables
 - T: Hay atasco
 - U: Estoy sujetando mi paraguas
 - R: Llueve

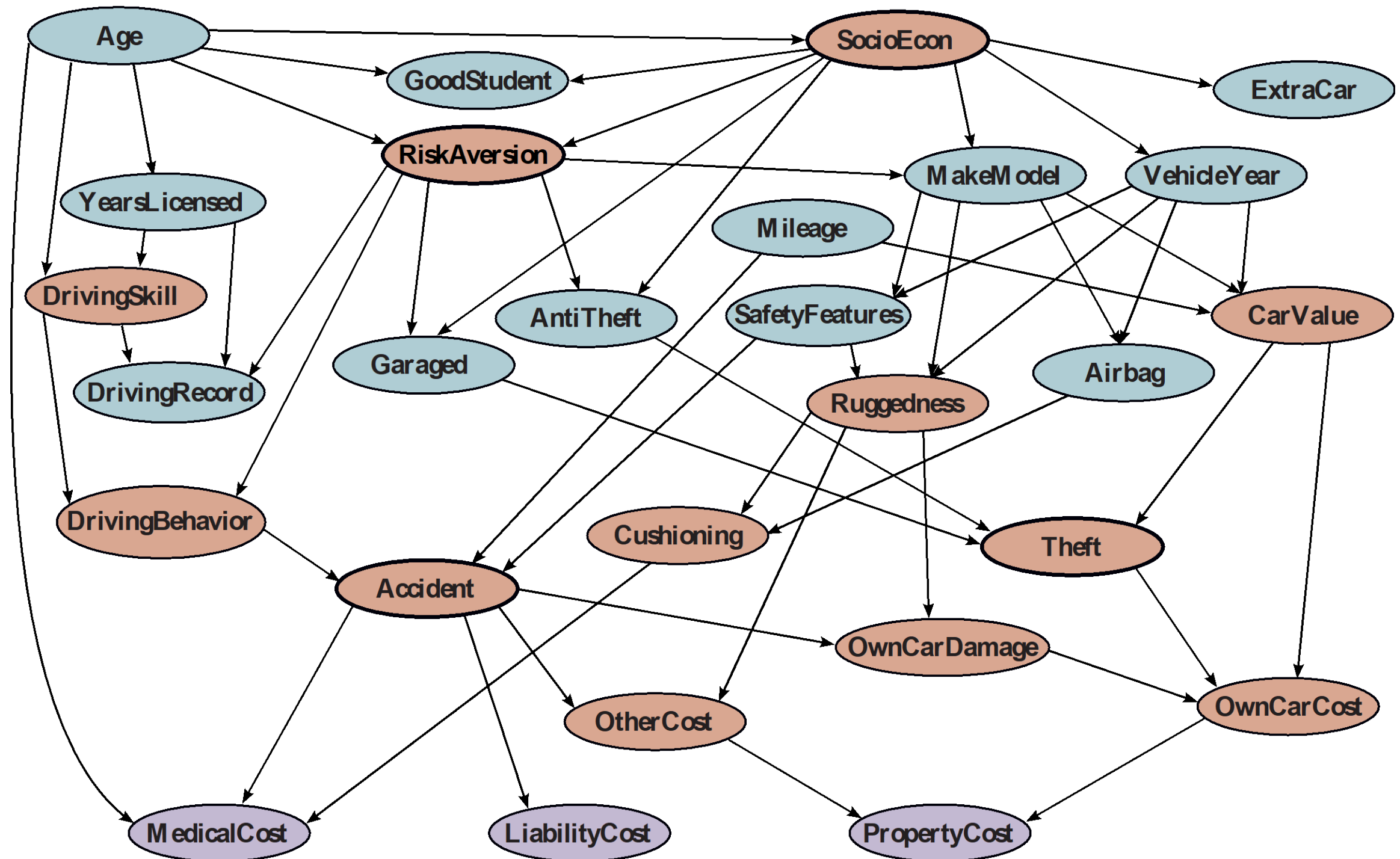


Ejemplo: alarma de incendios

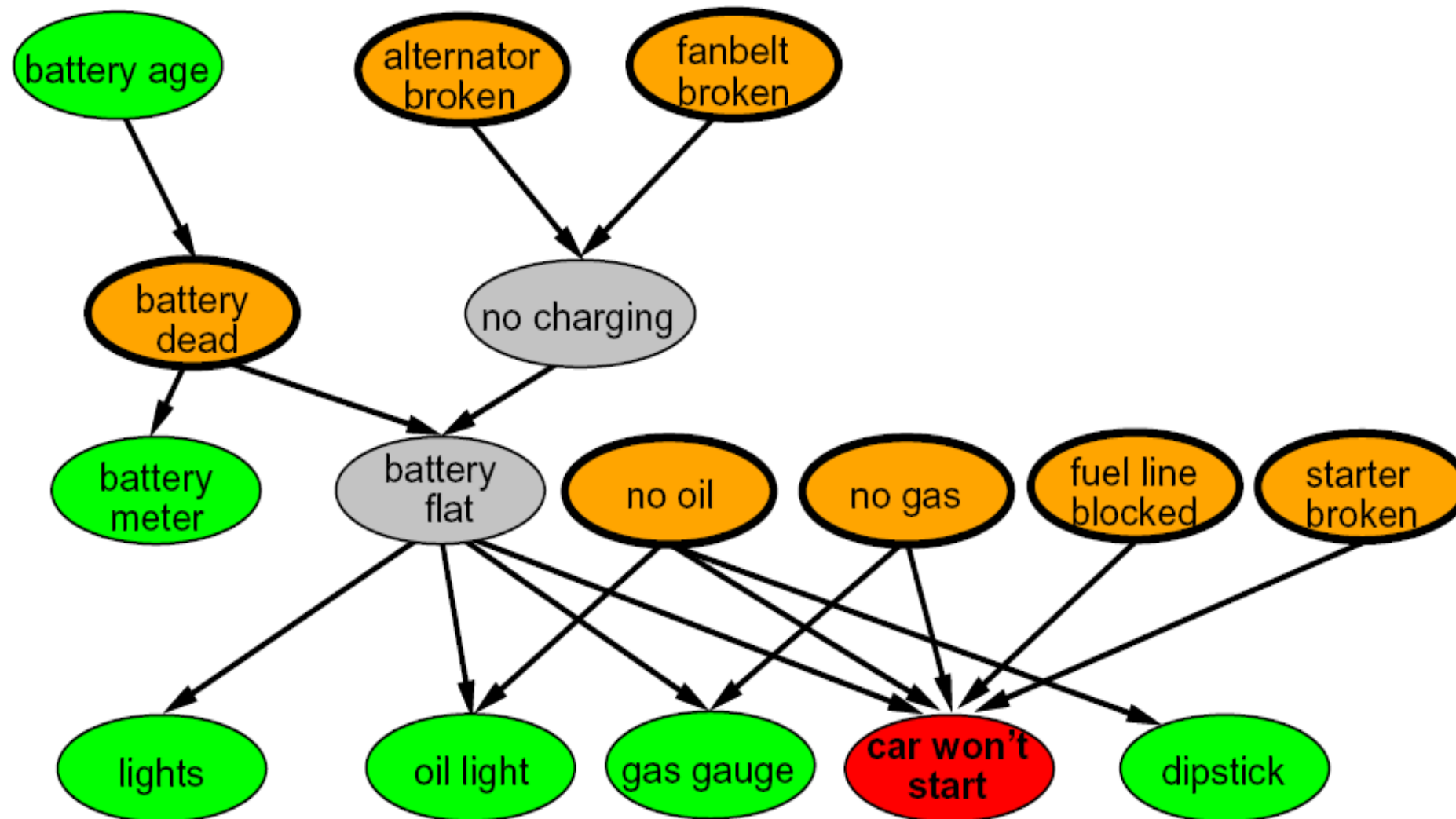
- Variables:
 - F: Hay fuego
 - S: Hay humo
 - A: Suena la alarma



Ejemplo de red bayesiana: Seguro de coche



Ejemplo de red bayesiana: El coche no arranca

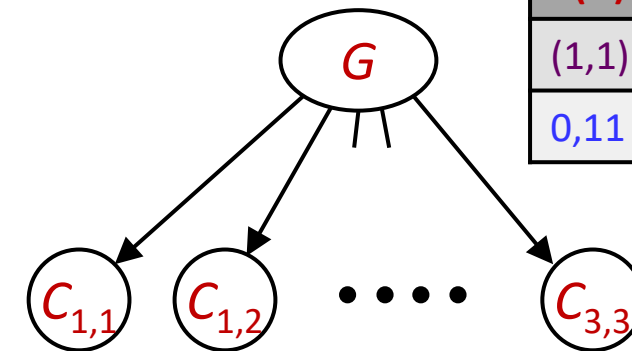


Sintaxis y semántica de las redes bayesianas



Sintaxis de las redes bayesianas

- Un conjunto de nodos, uno por variable X_i
- Un grafo acíclico dirigido
- Una distribución condicional para cada nodo dadas sus **variables padre** en el grafo
 - **CPT** (tabla de probabilidad condicional); cada fila es una distribución para el hijo dado un posible valor de sus padres

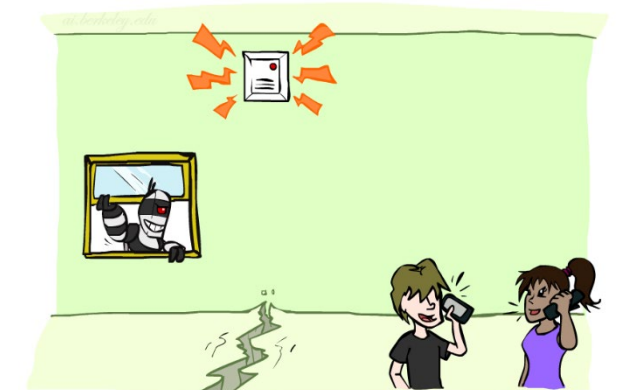
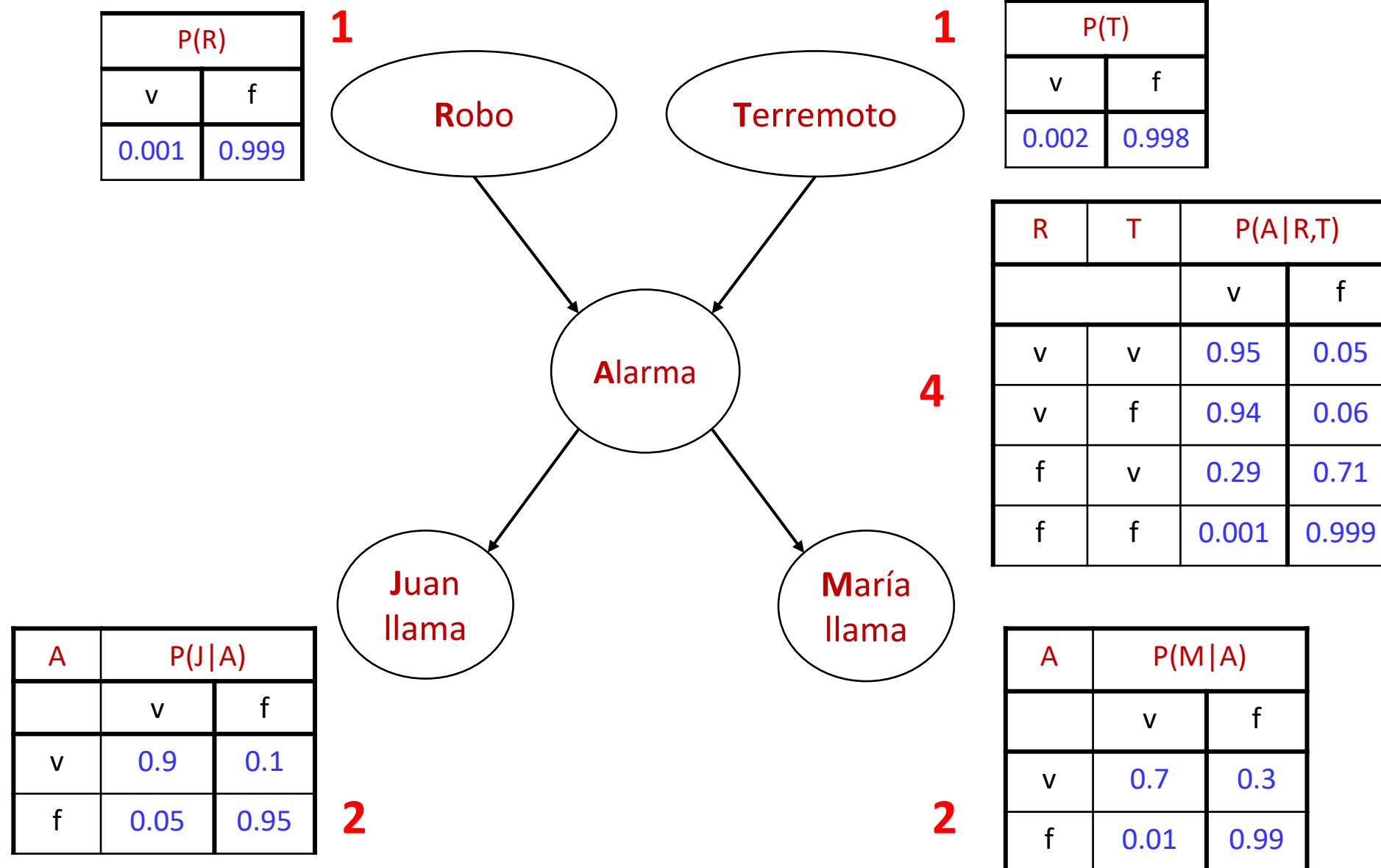


$P(G)$			
(1,1)	(1,2)	(1,3)	...
0,11	0,11	0,11	...

G	$P(C_{1,1} G)$			
	g	y	o	r
(1,1)	0,01	0,1	0,3	0,59
(1,2)	0,1	0,3	0,5	0,1
(1,3)	0,3	0,5	0,19	0,01
...				

Red bayesiana = Topología (gráfico) + Probabilidades condicionales locales

Ejemplo: red de la alarma



Número de **parámetros libres** en cada CPT:

Tamaños rango padres d_1, \dots, d_k

Tamaño del rango del hijo d
Cada fila debe sumar 1

$$(d-1) \prod_i d_i$$

Fórmula general para RB dispersas

- Supongamos
 - n variables
 - El tamaño máximo del rango es d
 - El máximo número de padres es k
- La distribución conjunta completa tiene tamaño $O(d^n)$
- La red bayesiana tiene tamaño $O(n \cdot d^k)$
 - Escala linealmente con n siempre y cuando la estructura causal sea local

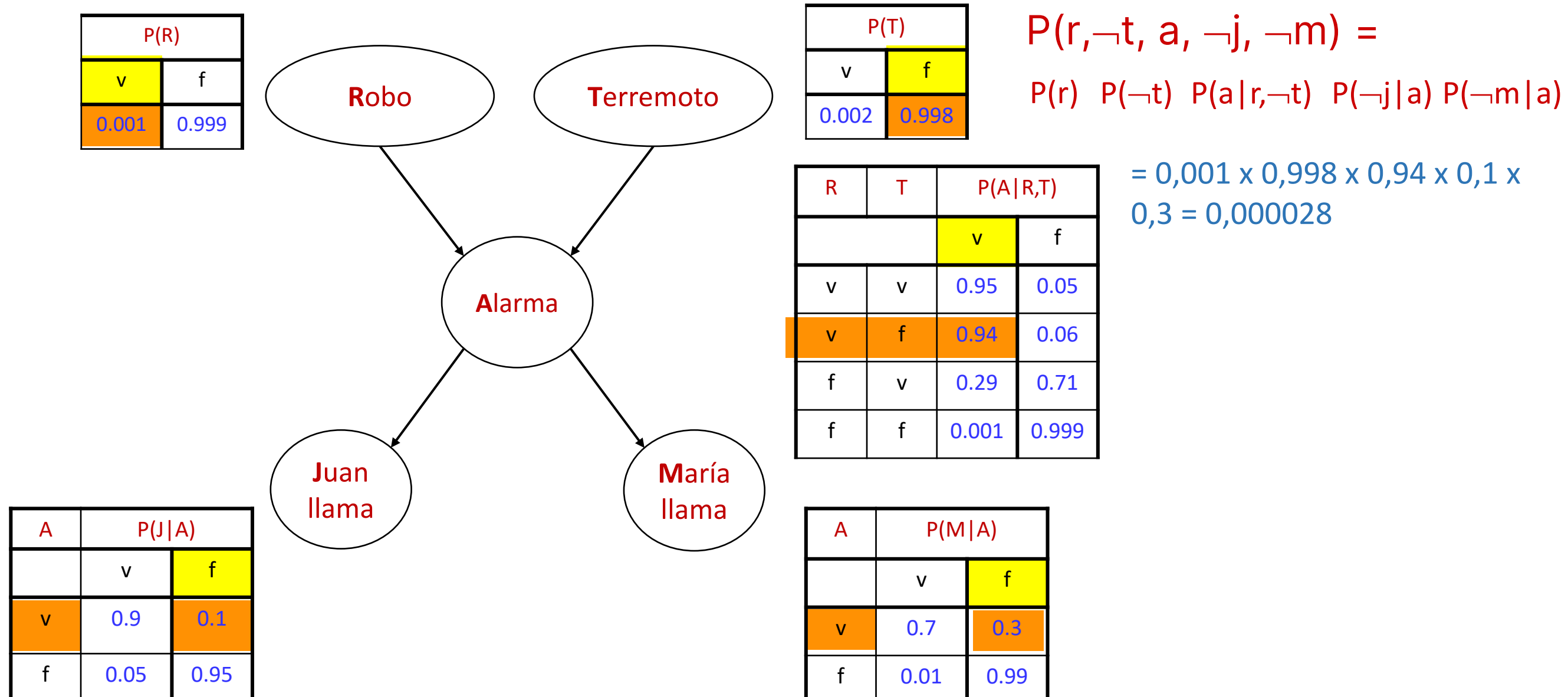
Semántica global de las redes bayesianas

- Las redes bayesianas expresan distribuciones conjuntas como el producto de distribuciones condicionales en cada variable:

$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid \text{Padres}(X_i))$$



Ejemplo



Independencia condicional en redes bayesianas

- Comparando la semántica global de las redes bayesianas

$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid \text{Padres}(X_i))$$

con la identidad de la regla de la cadena

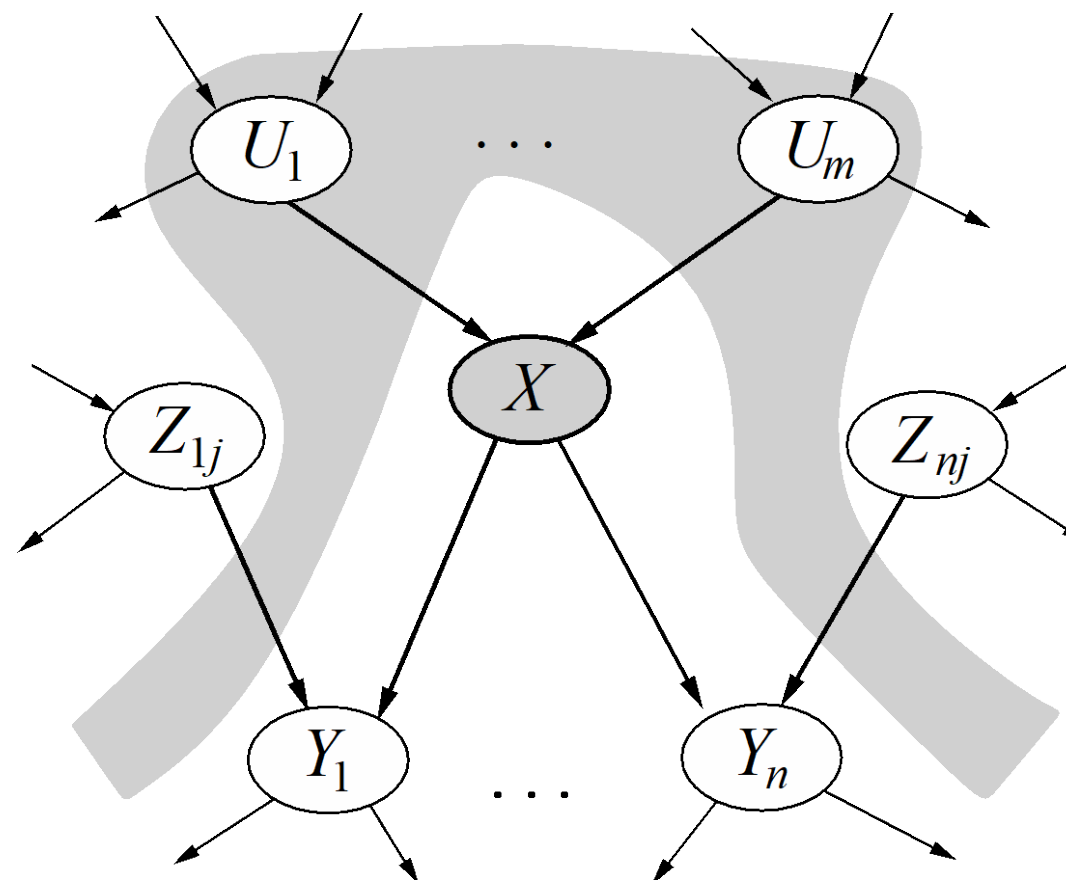
$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid X_1, \dots, X_{i-1})$$



- Asumir (sin pérdida de generalidad) que X_1, \dots, X_n están ordenadas en orden topológico según el grafo (i.e., padres antes que los hijos), por lo que $\text{Padres}(X_i) \subseteq X_1, \dots, X_{i-1}$
- Por tanto, la red de Bayes expresa independencias condicionales $P(X_i \mid X_1, \dots, X_{i-1}) = P(X_i \mid \text{Padres}(X_i))$
 - Para garantizar su validez, elegimos padres para el nodo X_i que lo "cubran" de otros predecesores

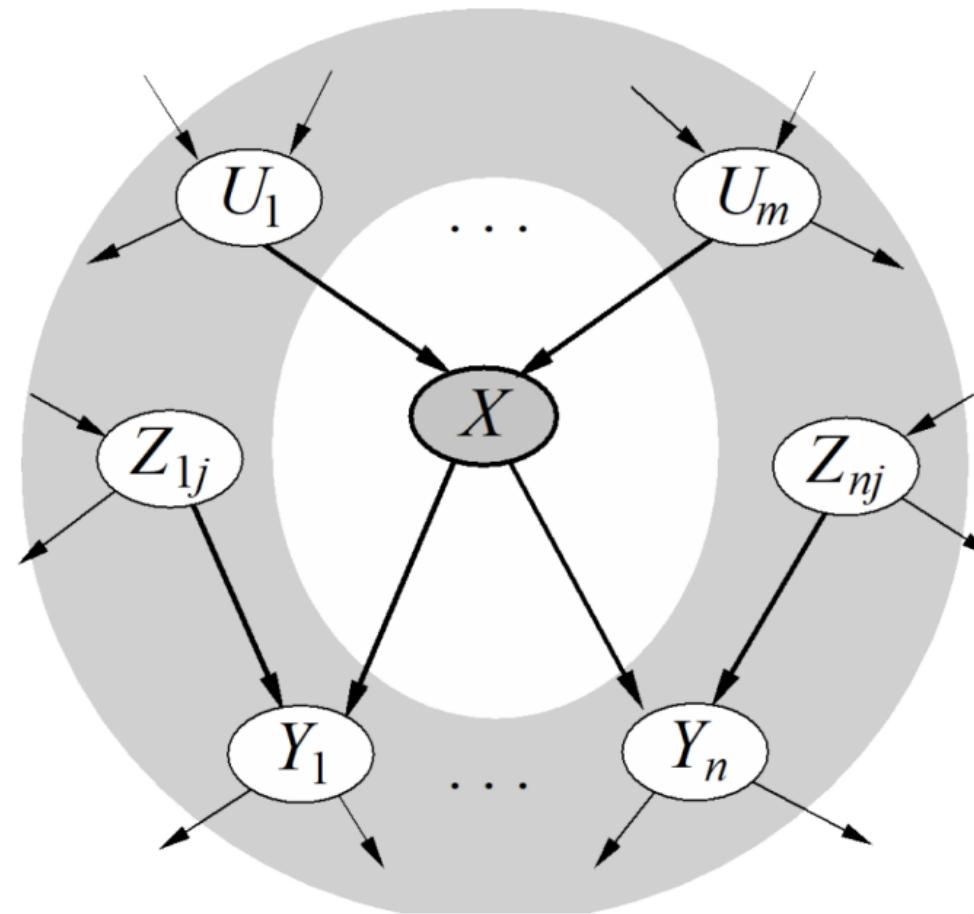
Estructura de las redes bayesianas

- Toda variable es condicionalmente independiente de sus no descendientes dados sus padres

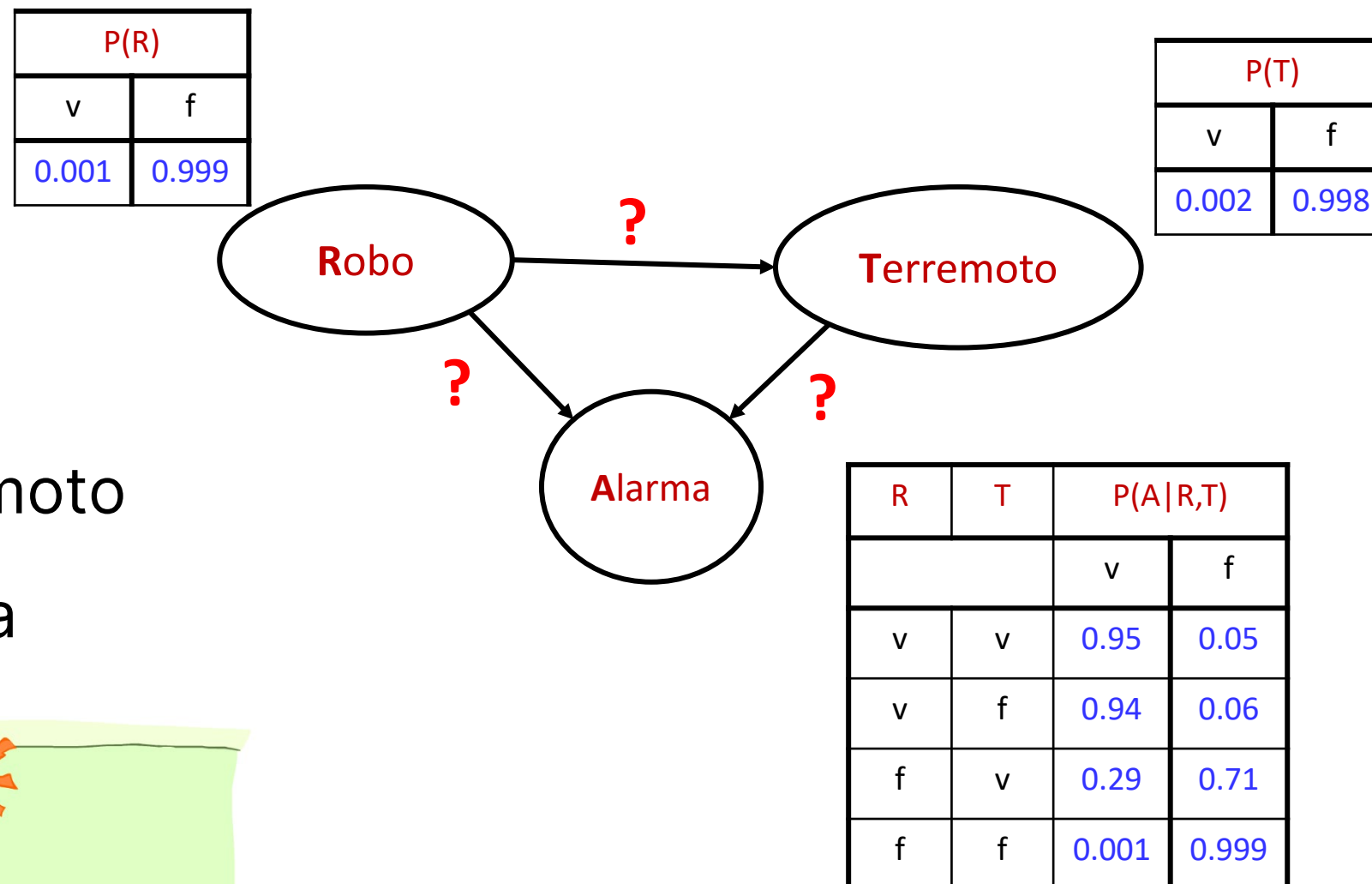


Estructura de las redes bayesianas

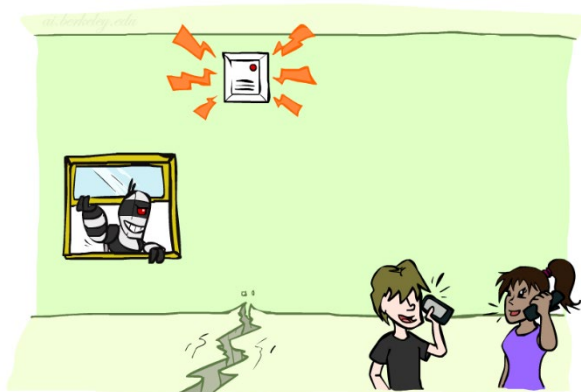
- Cada nodo es condicionalmente independiente de todas las demás variables dado su manto de Markov
- El manto de Markov de una variable comprende los padres, los hijos, y los otros padres de los hijos.



Ejemplo: Robo



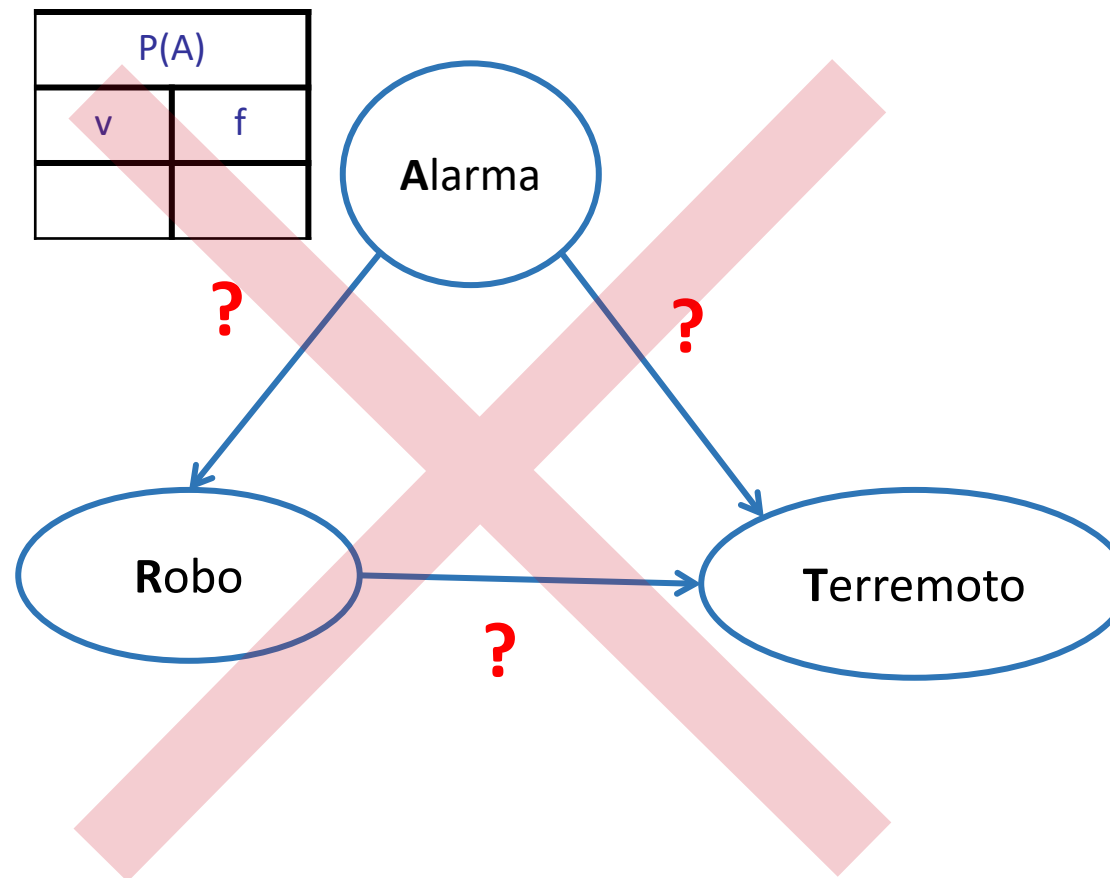
- Robo
- Terremoto
- Alarma



Ejemplo: Robo

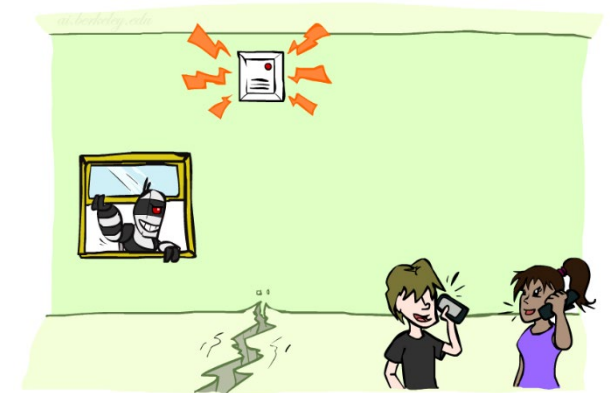
- Alarma
- Robo
- Terremoto

A	P(R A)	
	v	f
v	?	
f		



P(A)	
v	f

A	B	P(T A,R)	
		V	F
v	v	?	
v	f		
f	v		
f	f		



Inferencia por enumeración en r. bayesianas

- Recordatorio de la inferencia por enumeración:

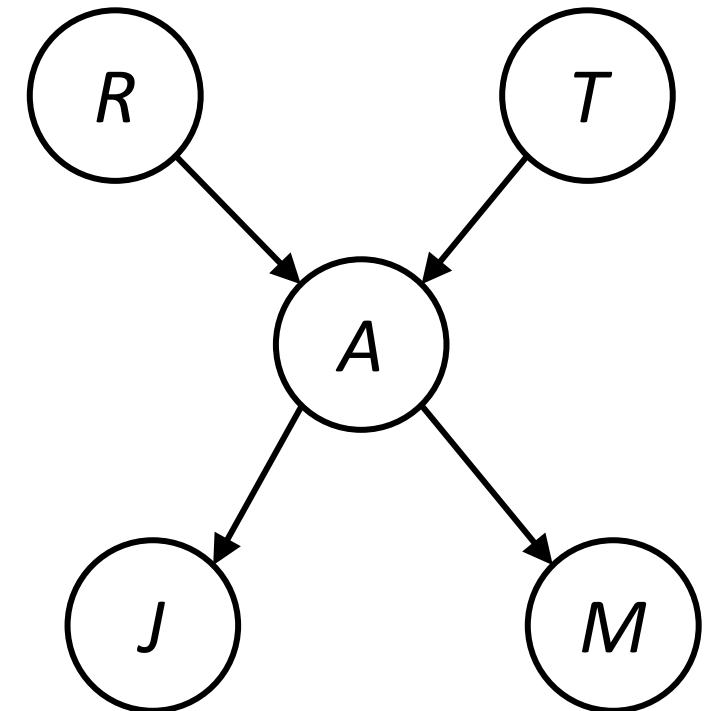
- Cualquier probabilidad de interés puede calcularse sumando las entradas de la distribución conjunta:

$$P(\mathbf{Q} \mid \mathbf{e}) = \alpha \sum_{\mathbf{h}} P(\mathbf{Q}, \mathbf{h}, \mathbf{e})$$

- Las entradas de la distribución conjunta pueden obtenerse a partir de una RB multiplicando las probabilidades condicionales correspondientes

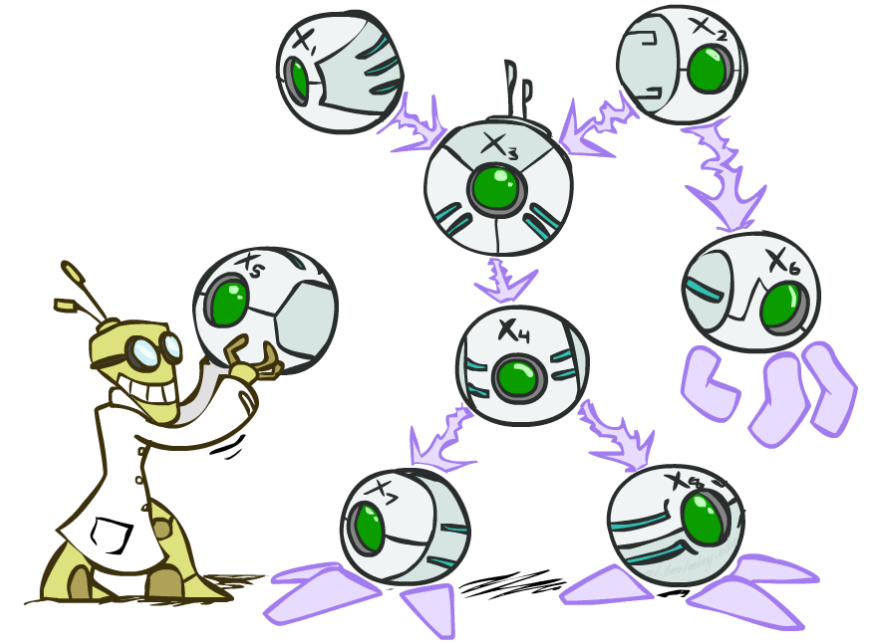
$$\begin{aligned} P(R \mid j, m) &= \alpha \sum_{t,a} P(R, t, a, j, m) \\ &= \alpha \sum_{t,a} P(R) P(t) P(a \mid R, t) P(j \mid a) P(m \mid a) \end{aligned}$$

- Así que la inferencia en las redes bayesianas significa calcular sumas de productos de números: parece sencillo
- Problema: suma de un **número exponencialmente grande** de productos



Resumen

- La independencia y la independencia condicional son formas importantes de conocimiento probabilístico
- Las redes bayesianas expresan distribuciones conjuntas de forma eficiente aprovechando la independencia condicional
 - Probabilidad conjunta global = producto de condicionales locales
- Inferencia exacta = sumas de productos de las probabilidades condicionales de la red



Inteligencia Artificial

Redes bayesianas: inferencia exacta



[Transparencias adaptadas de Dan Klein and Pieter Abbeel: CS188 Intro to AI, UC Berkeley (ai.berkeley.edu)]

Inferencia por enumeración en r. bayesianas

- Recordatorio de la inferencia por enumeración:

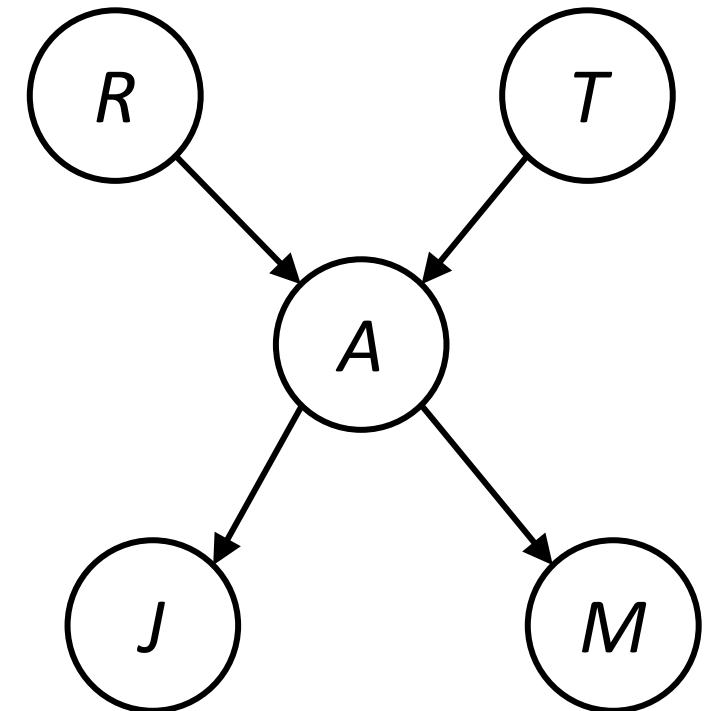
- Cualquier probabilidad de interés puede calcularse sumando las entradas de la distribución conjunta:

$$P(\mathbf{Q} \mid \mathbf{e}) = \alpha \sum_{\mathbf{h}} P(\mathbf{Q}, \mathbf{h}, \mathbf{e})$$

- Las entradas de la distribución conjunta pueden obtenerse a partir de una RB multiplicando las probabilidades condicionales correspondientes

$$\begin{aligned} P(R \mid j, m) &= \alpha \sum_{t,a} P(R, t, a, j, m) \\ &= \alpha \sum_{t,a} P(R) P(t) P(a \mid R, t) P(j \mid a) P(m \mid a) \end{aligned}$$

- Así que la inferencia en las redes bayesianas significa calcular sumas de productos de números: parece sencillo
- Problema: suma de un **número exponencialmente grande** de productos



¿Podemos hacerlo mejor?

- Sea **$uwy + uwz + uxy + uxz + vwy + vwz + vxy + vxz$**

- 16 multiplicaciones, 7 sumas
- Muchas expresiones repetidas

- Reescribir como **$(u+v)(w+x)(y+z)$**

- 2 multiplicaciones, 3 sumas

- $\sum_{t,a} P(R) P(t) P(a|R,t) P(j|a) P(m|a)$

$$= P(R)P(t)P(a|R,t)P(j|a)P(m|a) + P(R)P(\neg t)P(a|R,\neg t)P(j|a)P(m|a)$$

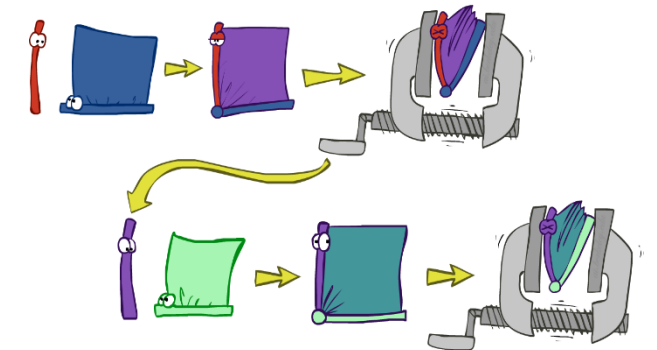
$$+ P(R)P(t)P(\neg a|R,t)P(j|\neg a)P(m|\neg a) + P(R)P(\neg t)P(\neg a|R,\neg t)P(j|\neg a)P(m|\neg a)$$

Hay muchas subexpresiones repetidas...

Eliminación de variables: ideas básicas

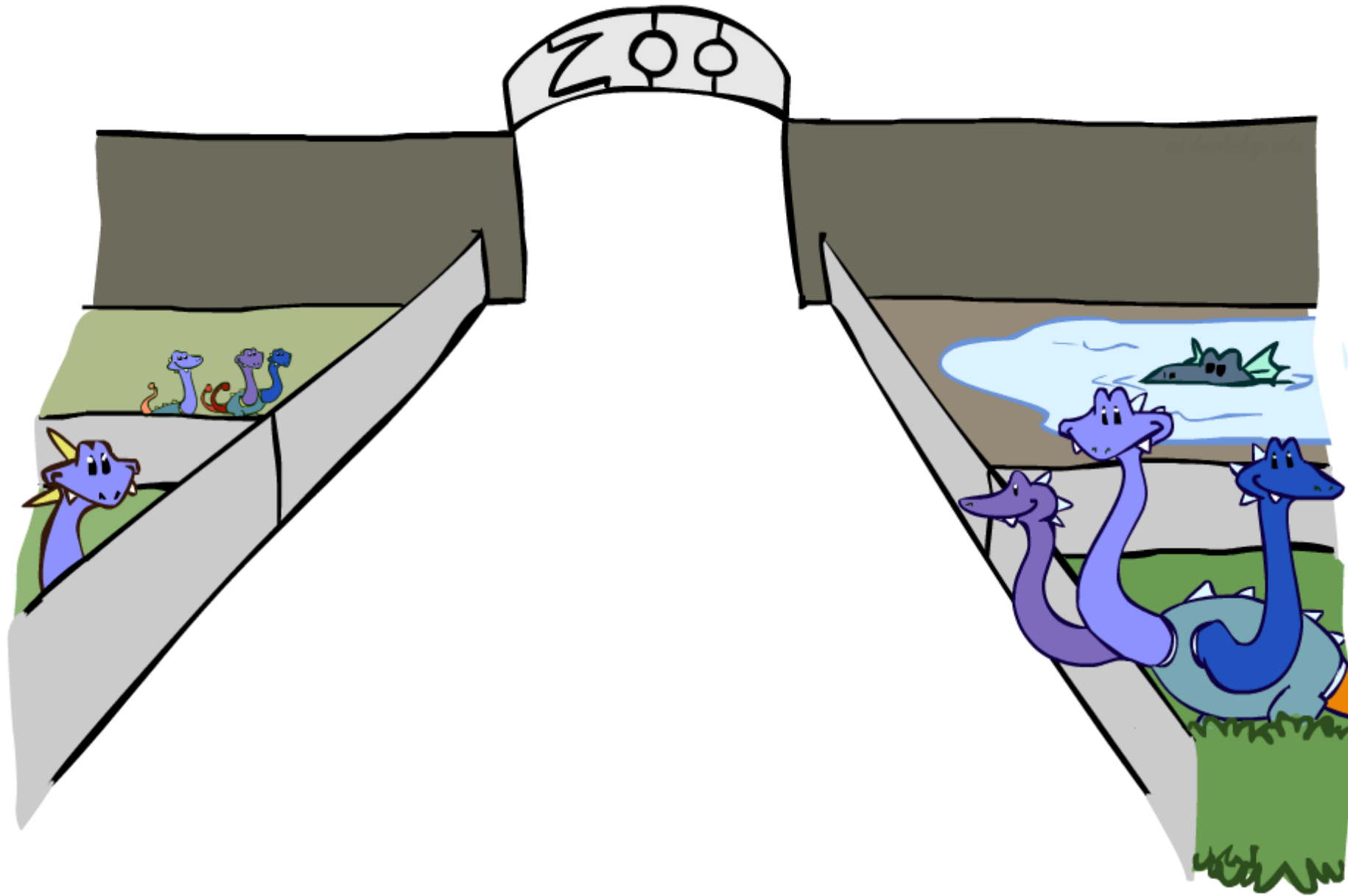
- Desplazar las sumas hacia el interior en la medida de lo posible

$$\begin{aligned} P(R \mid j, m) &= \alpha \sum_{t,a} P(R) P(t) P(a \mid R, t) P(j \mid a) P(m \mid a) \\ &= \alpha P(R) \sum_t P(t) \sum_a P(a \mid R, t) P(j \mid a) P(m \mid a) \end{aligned}$$



- Hacer el cálculo de dentro hacia fuera
 - Es decir, sumar primero sobre a , luego sumar sobre t
 - Problema: $P(a \mid R, t)$ no es un único número, sino muchos números diferentes en función de los valores de R y t
 - Solución: utilizar matrices de números (de varias dimensiones) con operaciones apropiadas sobre ellas; se denominan **factores**

Zoo de factores



Zoo de factores I

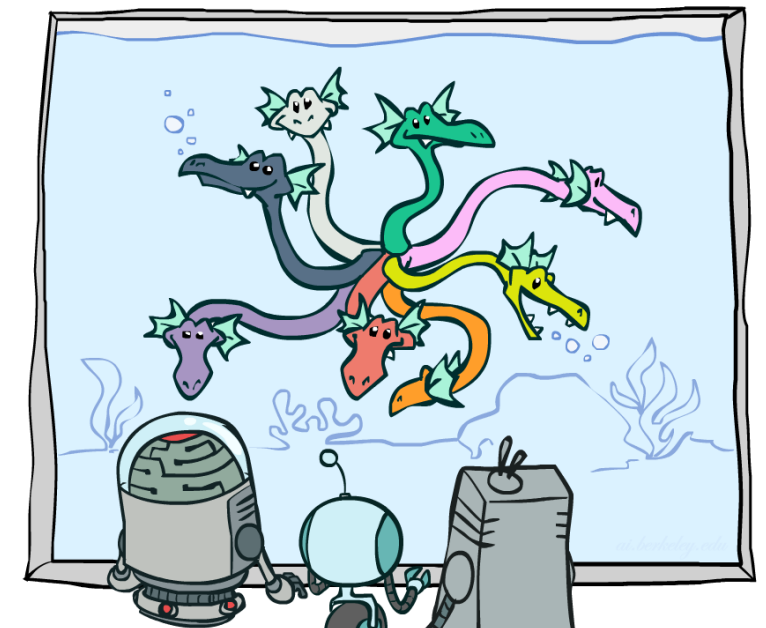
- Distribución conjunta: $P(X,Y)$
 - Entradas $P(x,y)$ para todo x, y
 - Matriz $|X| \times |Y|$
 - Suman 1
- Conjunta proyectada: $P(x,Y)$
 - Una porción de la distribución conjunta
 - Entradas $P(x,y)$ para una x particular y todas las y
 - Vector de $|Y|$ elementos
 - Suman $P(x)$

$$P(A,J)$$

$A \setminus J$	v	f
v	0.09	0.01
f	0.045	0.855

$$P(a,J) = P_a(J)$$

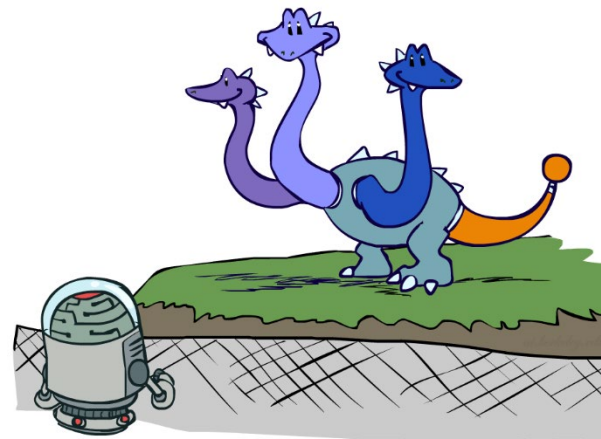
$A \setminus J$	v	f
v	0.09	0.01



Número de variables (mayúsculas) = dimensionalidad de la tabla

Zoo de factores II

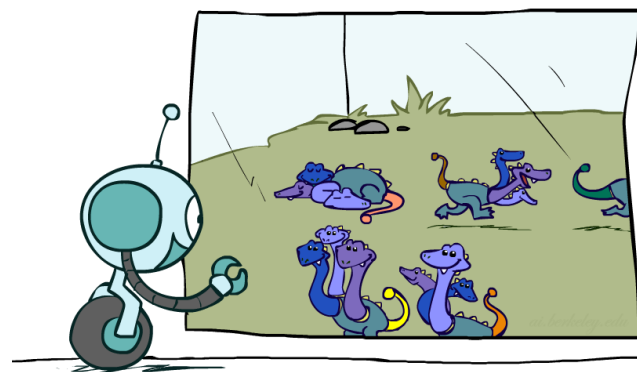
- Condicional único : $P(Y \mid x)$
 - Entradas $P(y \mid x)$ para x fija, todas las y
 - Suman 1



$P(J \mid a)$

$A \setminus J$	v	f
v	0.9	0.1

- Familia de condicionales:
 $P(X \mid Y)$
 - Múltiples condicionales
 - Entradas $P(x \mid y)$ para todas las x, y
 - Suman $|Y|$



$P(J \mid A)$

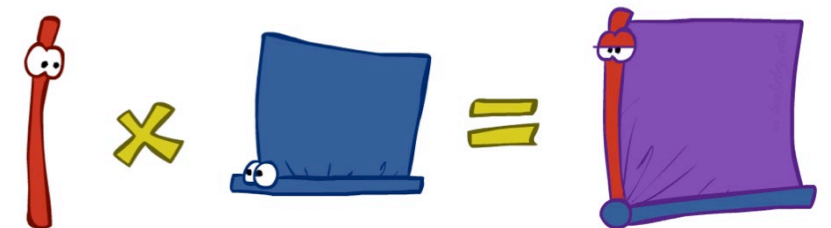
$A \setminus J$	v	f
v	0.9	0.1
f	0.05	0.95

$\} - P(J \mid a)$
 $\} - P(J \mid \neg a)$

Operación 1: Producto punto a punto

- Primera operación básica: **producto punto a punto** (*pointwise product*) de factores (**no** es multiplicación de matrices).

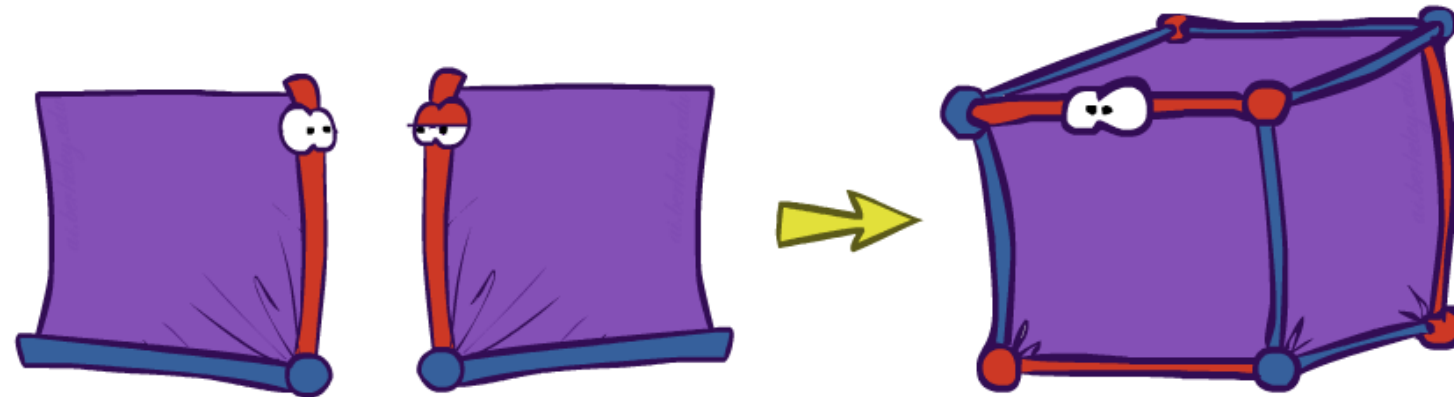
- El nuevo factor tiene la **unión** de las variables de los dos factores originales
- Cada entrada es el producto de las entradas correspondientes de los factores originales



- Ejemplo: $P(J|A) \times P(A) = P(A,J)$

$P(A)$		\times	$P(J A)$			$=$	$P(A,J)$		
			$A \setminus J$	v	f		$A \setminus J$	v	f
v	0.1		v	0.9	0.1		v	0.09	0.01
f	0.9		f	0.05	0.95		f	0.045	0.855

Ejemplo: Aumentar factores



- Ejemplo: $P(A,J) \times P(A,M) = P(A,J,M)$

$P(A,J)$

$A \setminus J$	v	f
v	0.09	0.01
f	0.045	0.855

×

$P(A,M)$

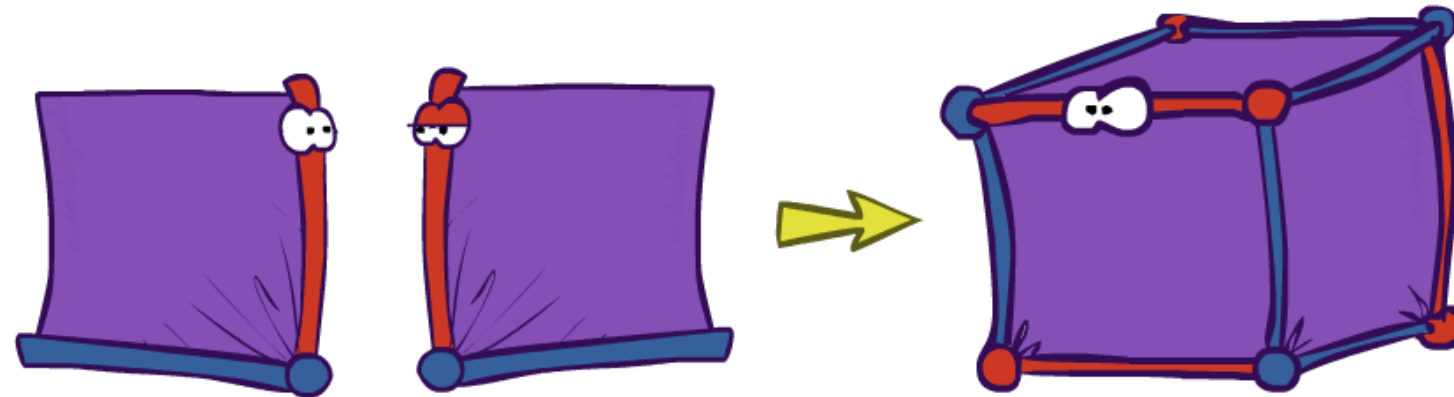
$A \setminus M$	v	f
v	0.07	0.03
f	0.009	0.891

=

$P(A,J,M)$

$J \setminus M$	v	f	
v			18
f		.0003	A=v

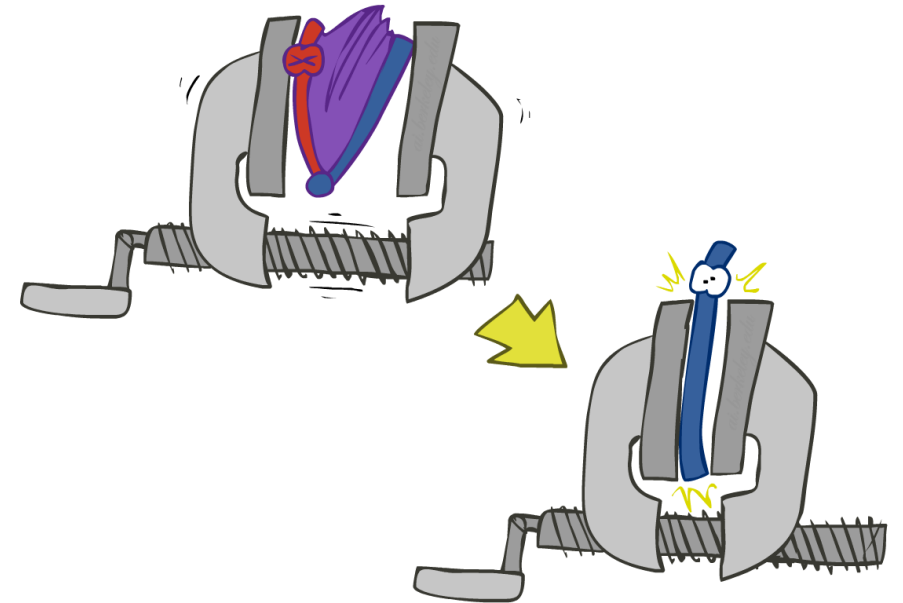
Ejemplo: Aumentar factores



- Ejemplo: $P(U,V) \times P(V,W) \times P(W,X) = P(U,V,W,X)$
- Tamaños: $[10,10] \times [10,10] \times [10,10] = [10,10,10,10]$
- Es decir, 300 números aumentan (*explotan*) a 10 000 números
- La explosión de factores puede hacer la eliminación de variables muy intensiva en el uso de memoria

Operación 2: Suma de una variable

- Segunda operación básica: **sumar eliminando** una variable desde un factor
 - Encoge un factor a uno más pequeño
- Ejemplo: $\sum_j P(A,J) = P(A,j) + P(A,\neg j) = P(A)$

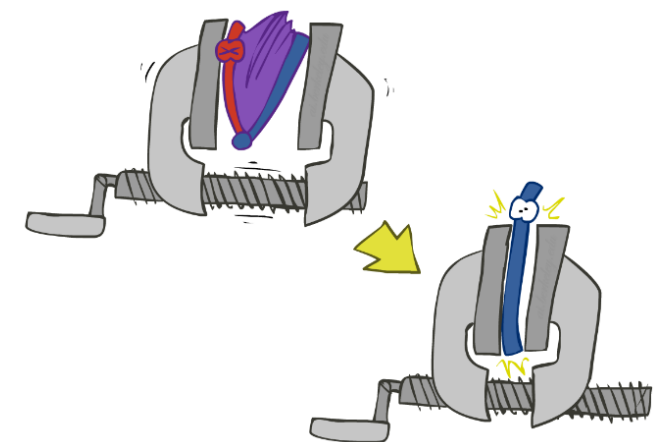


$P(A,J)$						$P(A)$		
$A \setminus J$	v	f				v	f	
v	0.09	0.01	Sumar eliminando J			0.1		
f	0.045	0.855				0.9		

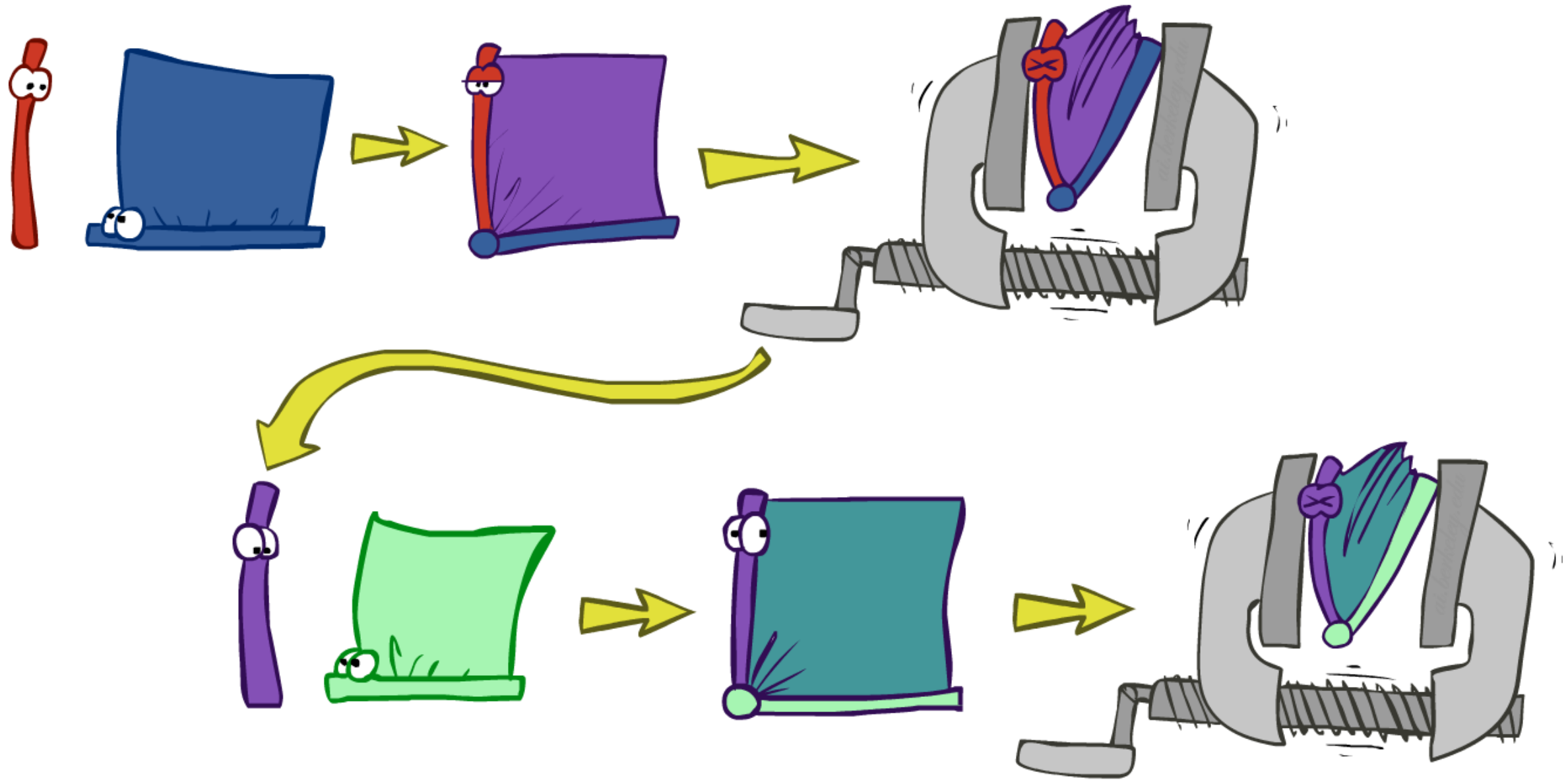
Sumar eliminando en un producto de factores

- Proyectar los factores en cada sentido primero, luego sumar los productos

- Ejemplo: $\sum_a P(a | R, t) \times P(j | a) \times P(m | a)$
 $= P(a | R, t) \times P(j | a) \times P(m | a) +$
 $P(\neg a | R, t) \times P(j | \neg a) \times P(m | \neg a)$

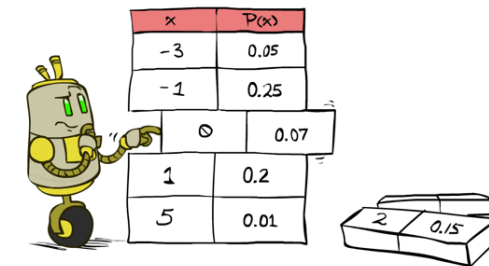


Eliminación de variables

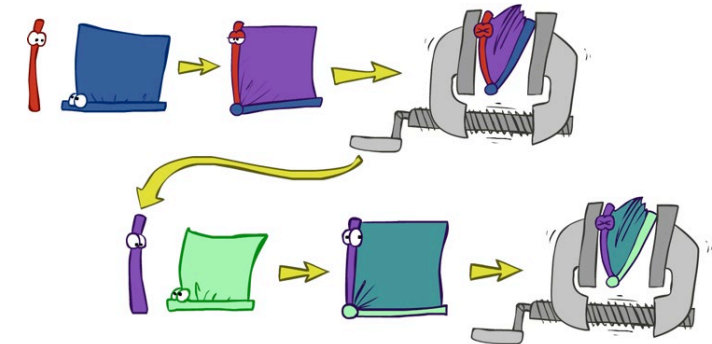


Eliminación de variables

- Consulta: $P(Q|E_1=e_1, \dots, E_k=e_k)$
- Empezar con factores iniciales:
 - CPT locales (pero instanciadas por evidencia)
- Mientras siga habiendo variables ocultas (no consulta o evidencias):
 - Elegir una variable oculta H_j
 - Eliminar (sumar eliminando) H_j del producto de todos los factores que mencionan H_j
- Unir todos los factores restantes y normalizar



x	P(x)
-3	0.05
-1	0.25
0	0.07
1	0.2
5	0.01

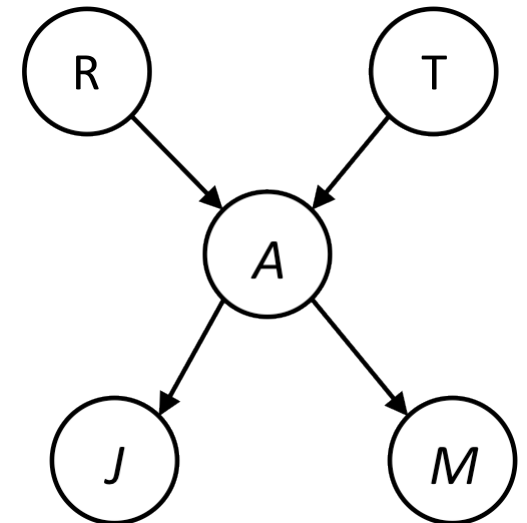



$$\text{stick figure} \times \text{blue square} = \text{purple square} \times \alpha$$

Ejemplo

Consulta $P(R \mid j, m)$

$P(R)$	$P(T)$	$P(A \mid R, T)$	$P(j \mid A)$	$P(m \mid A)$
--------	--------	------------------	---------------	---------------



Elegir A

$P(A \mid R, T)$	\times	\rightarrow	Σ	\rightarrow	$P(j, m \mid R, T)$
$P(j \mid A)$					
$P(m \mid A)$					

$P(R)$	$P(T)$	$P(j, m \mid R, T)$
--------	--------	---------------------

Ejemplo

$P(R)$	$P(T)$	$P(j,m R,T)$
--------	--------	--------------

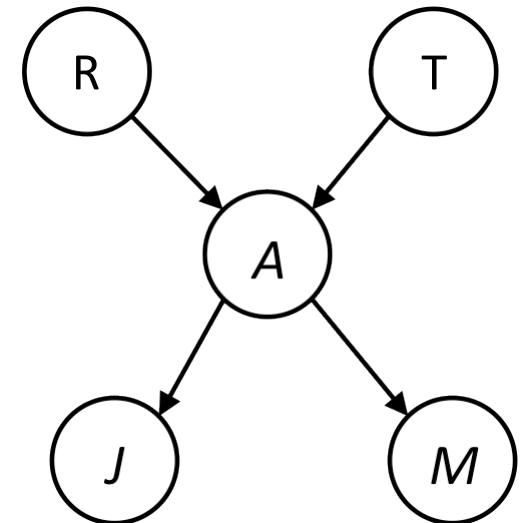
Elegir T

$$\begin{array}{c} P(T) \\ P(j,m|R,T) \end{array} \xrightarrow{\times} \xrightarrow{\Sigma} P(j,m|R)$$

$P(R)$	$P(j,m R)$
--------	------------

Acabar con R

$$\begin{array}{c} P(R) \\ P(j,m|R) \end{array} \xrightarrow{\times} P(j,m,R) \xrightarrow{\text{Normalizar}} P(R | j,m)$$



El orden importa

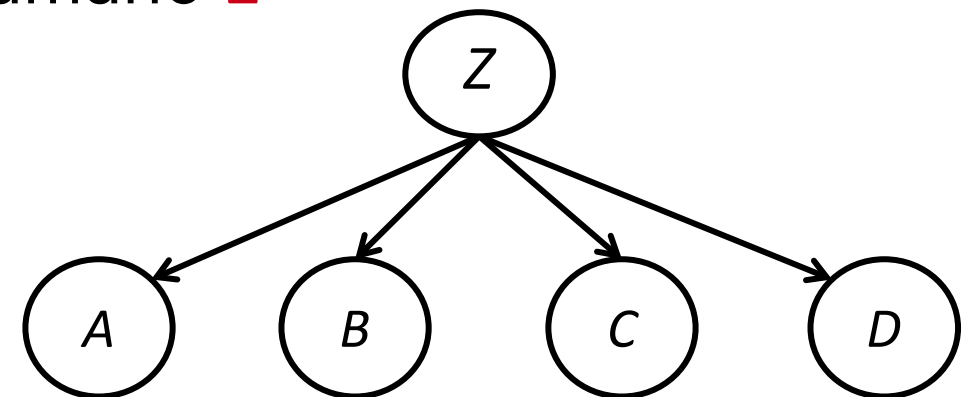
- Ordenando los términos como Z, A, B C, D

$$\begin{aligned} P(D) &= \alpha \sum_{z,a,b,c} P(z) P(a|z) P(b|z) P(c|z) P(D|z) \\ &= \alpha \sum_z P(z) \sum_a P(a|z) \sum_b P(b|z) \sum_c P(c|z) P(D|z) \end{aligned}$$

- El factor más grande tiene 2 variables (D,Z)
- Ordenando los términos A, B, C, D, Z

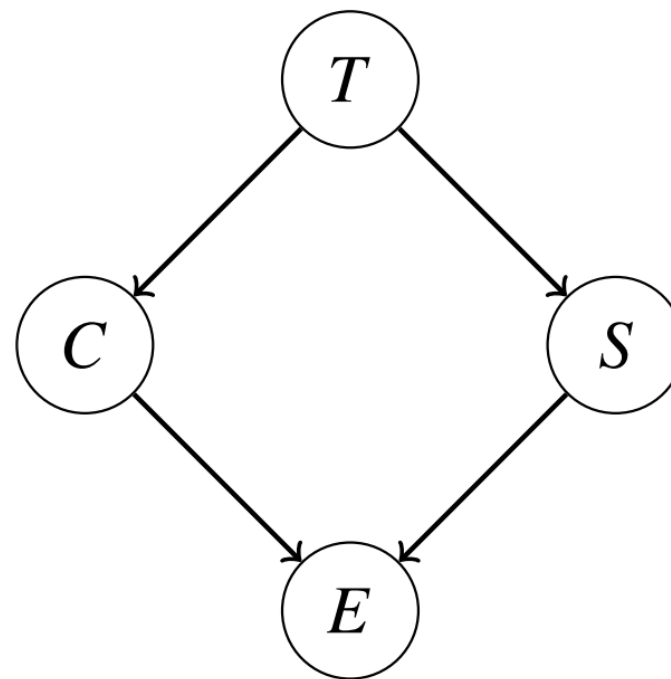
$$\begin{aligned} P(D) &= \alpha \sum_{a,b,c,z} P(a|z) P(b|z) P(c|z) P(D|z) P(z) \\ &= \alpha \sum_a \sum_b \sum_c \sum_z P(a|z) P(b|z) P(c|z) P(D|z) P(z) \end{aligned}$$

- El factor más grande tiene 4 variables (A,B,C,D)
- En general, con *n* hojas, factor de tamaño 2^n



Ejemplo

T representa la probabilidad de que un aventurero coja un tesoro, **C** representa la probabilidad de que una jaula caiga sobre el aventurero dado que coge el tesoro, **S** representa la probabilidad de que se suelten serpientes si un aventurero coge el tesoro, y **E** representa la probabilidad de que el aventurero escape dada la información sobre el estado de la jaula y las serpientes.



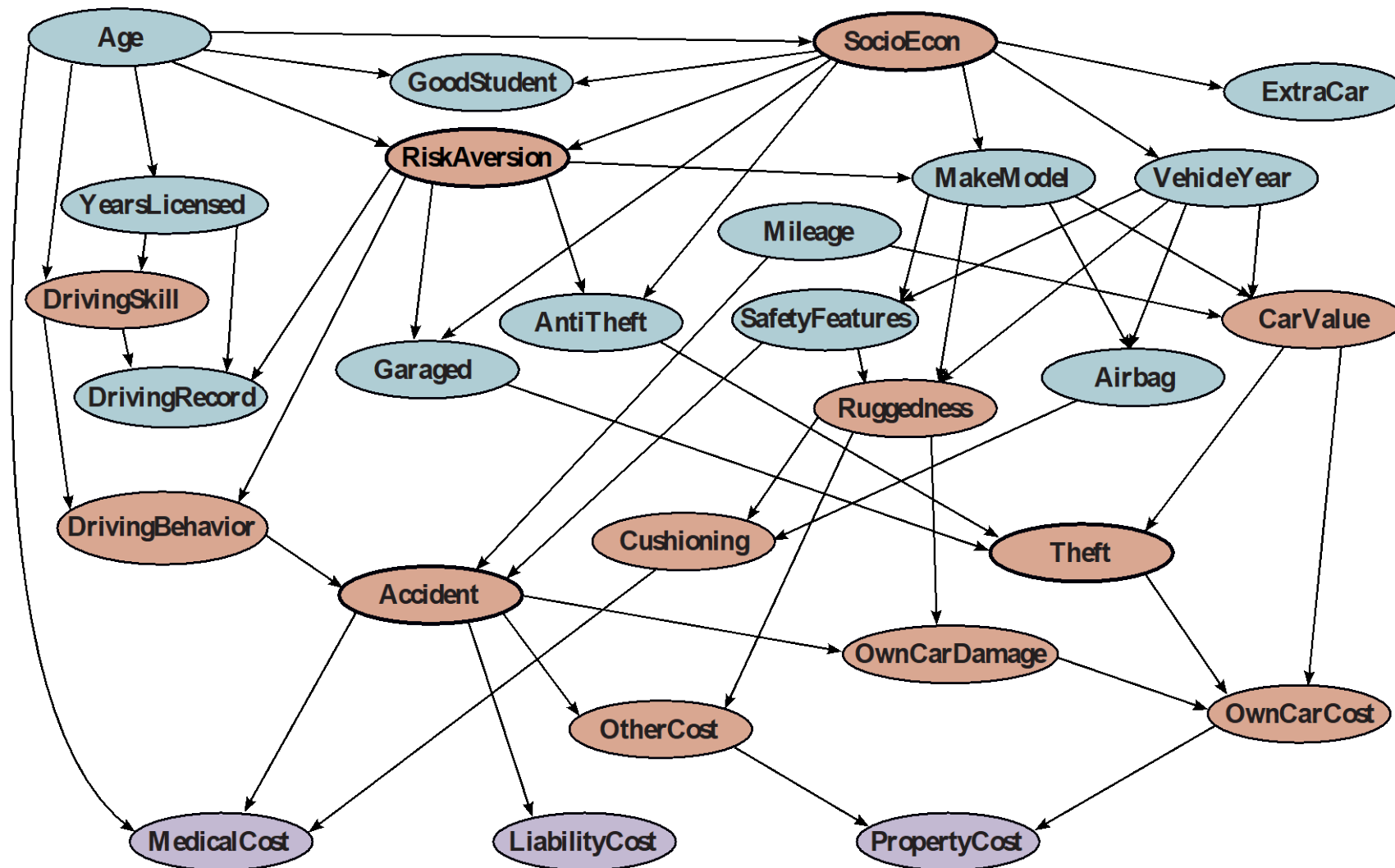
Ejemplo

- En este caso, tenemos los factores $P(T)$, $P(C|T)$, $P(S|T)$ y $P(E|C,S)$.
- Supongamos que queremos calcular $P(T|+e)$.
- El método de inferencia por enumeración consistiría en formar la PDF conjunta de 16 filas $P(T,C,S,E)$, seleccionar sólo las filas correspondientes a $+e$, luego sumar C y S y finalmente normalizar.

Ejemplo

- En este caso, tenemos los factores $P(T)$, $P(C|T)$, $P(S|T)$ y $P(E|C,S)$, y queremos calcular $P(T|+e)$.
- El enfoque alternativo consiste en eliminar C, luego S, una variable cada vez:
 1. Unir (multiplicar) todos los factores que implican a C, formando $f1(C,+e,T,S) = P(C|T) \cdot P(+e|C,S)$
 2. Sumar eliminando C de este nuevo factor, lo que nos deja con un nuevo factor $f2(+e,T,S)$
 3. Unir todos los factores que implican a S, formando $f3(+e,S,T) = P(S|T) \cdot f2(+e,T,S)$
 4. Sumar eliminando S, dando $f4(+e,T)$
 5. Une los factores restantes, lo que da $f5(+e,T) = f4(+e,T) \cdot P(T)$
 6. Podemos calcular fácilmente $P(T|+e)$ normalizando.

Ejemplo red bayesiana: seguro de coche



Enumeración: **227M** operaciones

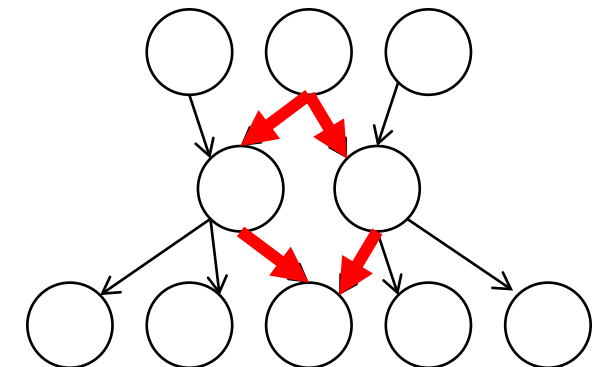
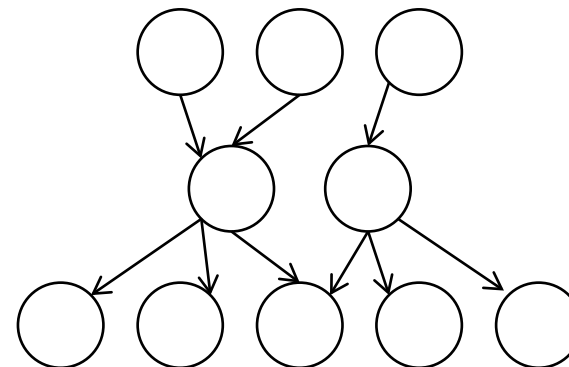
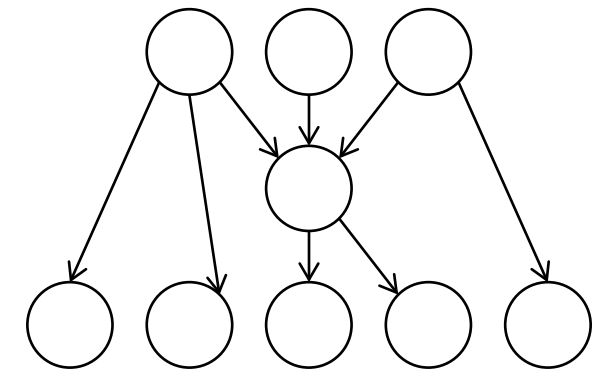
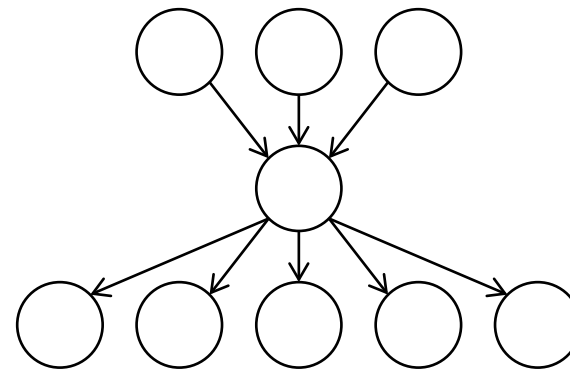
Eliminación: **221K** operaciones

Complejidad computacional y espacial

- La complejidad computacional y espacial de la eliminación de variables está determinada por el factor más grande (y es el espacio *lo que nos mata*).
- El orden de eliminación puede afectar en gran medida al tamaño del factor más grande.
 - P.ej., ejemplo ZABCD 2^n vs. 2
- ¿Existe siempre una ordenación que sólo da lugar a factores pequeños?
 - **No**

Poliárboles

- Un poliárbol es un grafo dirigido sin ciclos no dirigidos
 - Es decir, aquellos en los que no aparecen bucles cuando quitamos las flechas
- Para los poliárboles, la complejidad de la eliminación de variables es **lineal en el tamaño de la red** si se va eliminando desde las hojas a las raíces



Resumen

- Inferencia exacta = sumas de productos de probabilidades condicionales de la red
- La enumeración es siempre exponencial
- La eliminación de variables lo reduce evitando el recálculo de subexpresiones repetidas
 - Aceleración masiva en la práctica
 - Tiempo lineal para poliárboles
- La inferencia exacta puede ser extremadamente compleja en términos de cálculo

