Vytěžování dat, cvičení 4: Bayesovské sítě

Miroslav Čapek, Michael Anděl

October 15, 2013

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Modelový příklad

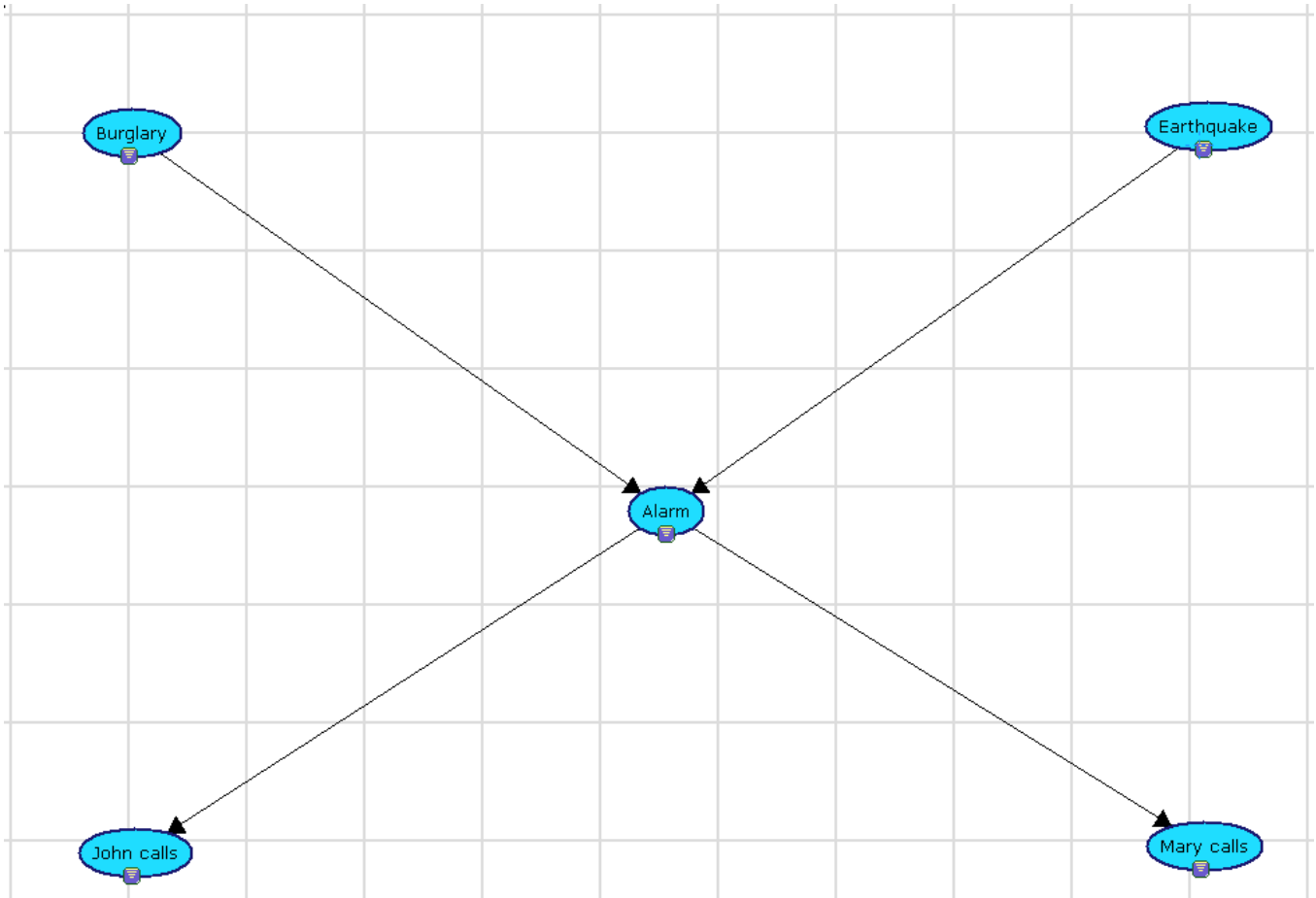
Consider the following situation. You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes.[1] *You also have two neighbors, John and Mary, who have promised to call you at work when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too. Mary, on the other hand, likes rather loud music and sometimes misses the alarm altogether. Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.* Zdroj: AIMA

[1] This example is due to Judea Pearl, a resident of Los Angeles; hence the acute interest in earthquakes.

Krok 1: Vytvoření struktury Bayesovské sítě

- Mějme 4 náhodné proměnné:
  Burglary, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*.
- Základní princip: Šipka vede od X do Y právě tehdy když X má přímý vliv na Y.
- Navrhněte vazby mezi proměnnými!
- Bayesian Network tools in Java (BNJ):
  `http://bnj.sourceforge.net/`

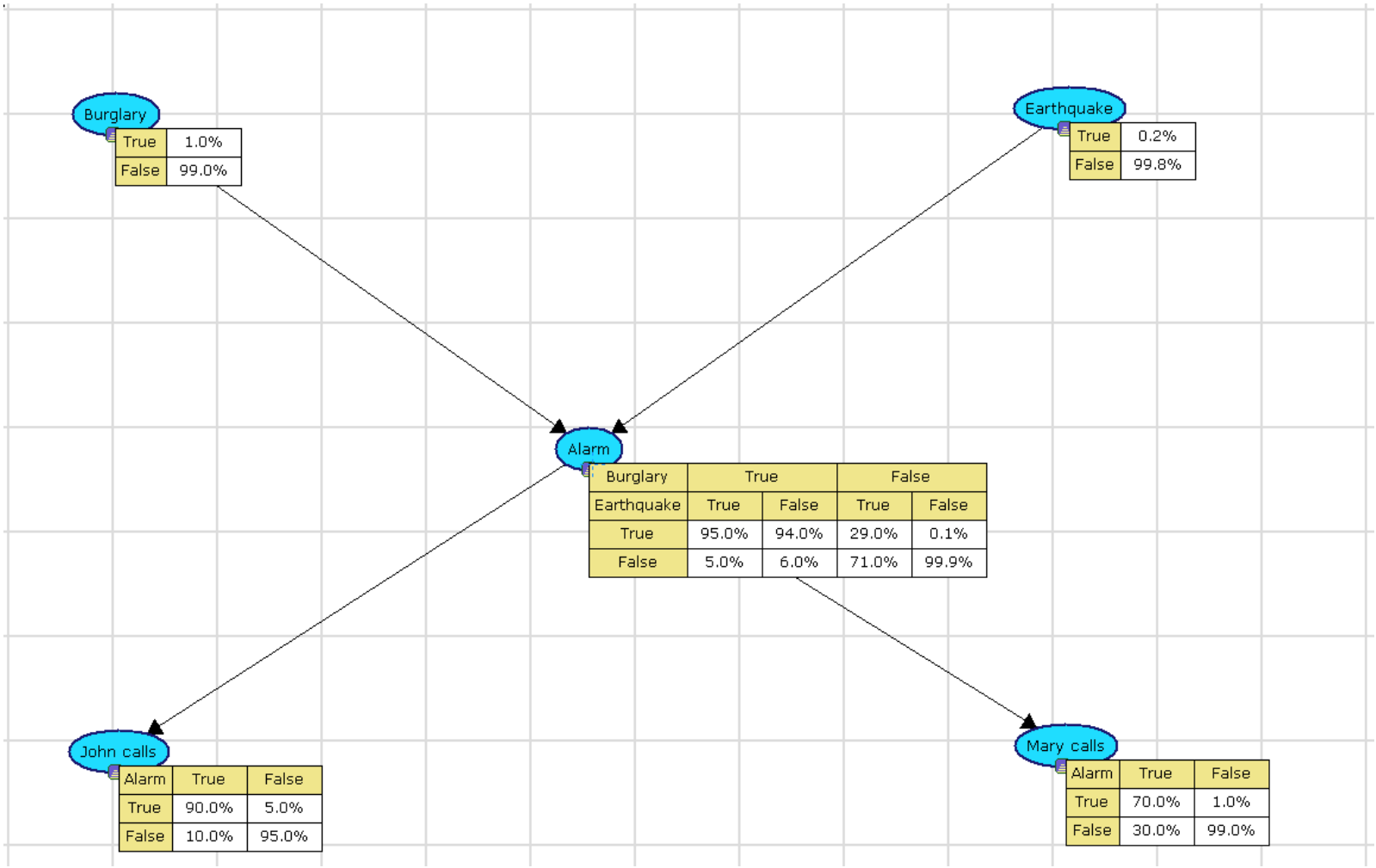
Krok 1: Správná odpověď



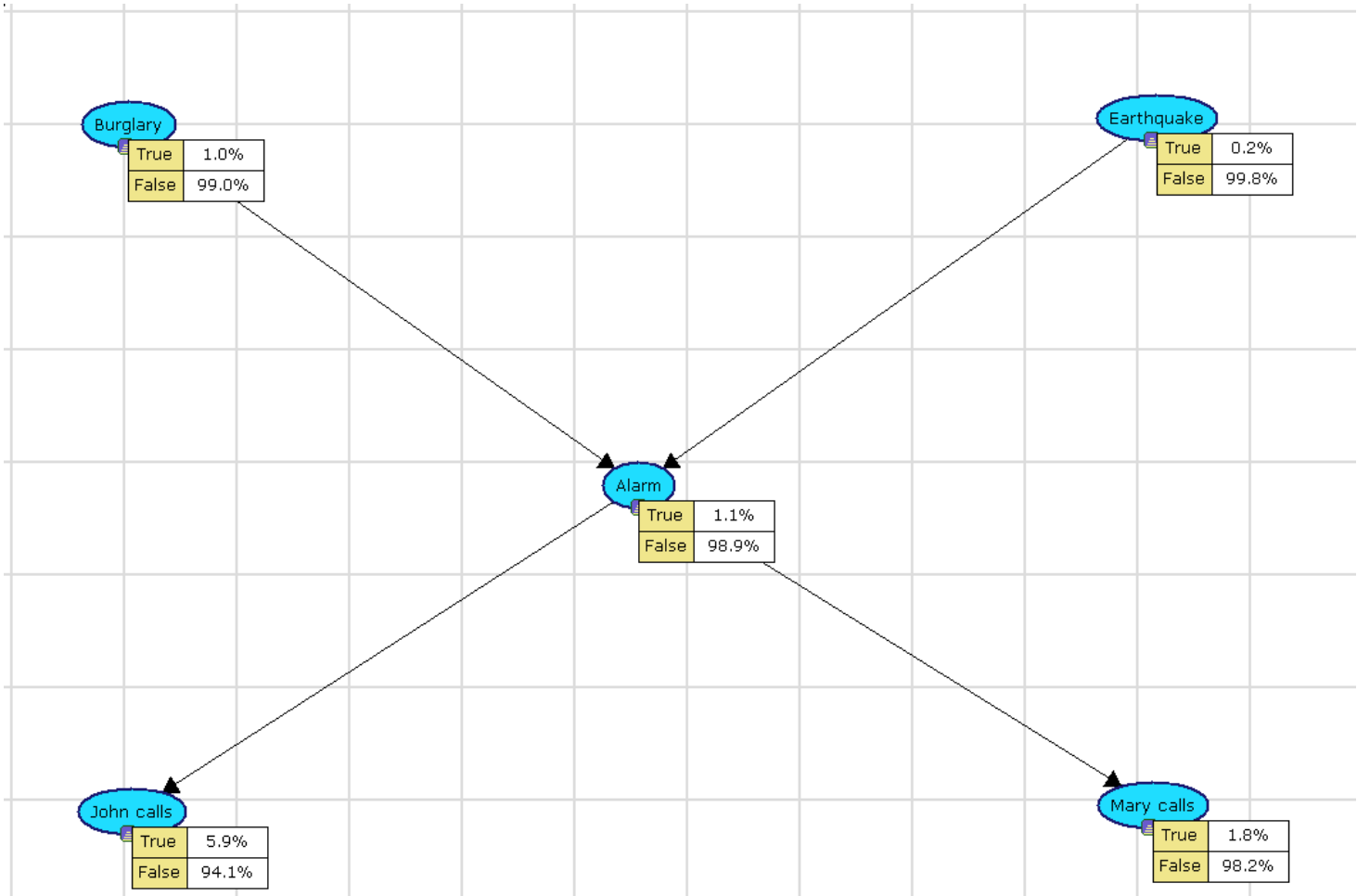
Krok 2: Data

- $\mathrm{P}(\mathit{Burglary} = \mathit{true}) = 1\%$
- $\mathrm{P}(\mathit{Earthquake} = \mathit{true}) = 0.2\%$
- $\mathrm{P}(\mathit{Alarm} = \mathit{true} \,|\, \mathit{Burglary} = \mathbf{true}, \mathit{Earthquake} = \mathbf{true}) = 95\%$
- $\mathrm{P}(\mathit{Alarm} = \mathit{true} \,|\, \mathit{Burglary} = \mathbf{true}, \mathit{Earthquake} = \mathbf{false}) = 94\%$
- $\mathrm{P}(\mathit{Alarm} = \mathit{true} \,|\, \mathit{Burglary} = \mathbf{false}, \mathit{Earthquake} = \mathbf{true}) = 29\%$
- $\mathrm{P}(\mathit{Alarm} = \mathit{true} \,|\, \mathit{Burglary} = \mathbf{false}, \mathit{Earthquake} = \mathbf{false}) = 0.1\%$
- $\mathrm{P}(\mathit{JohnCalls} = \mathit{true} \,|\, \mathit{Alarm} = \mathbf{true}) = 90\%$
- $\mathrm{P}(\mathit{JohnCalls} = \mathit{true} \,|\, \mathit{Alarm} = \mathbf{false}) = 5\%$
- $\mathrm{P}(\mathit{MaryCalls} = \mathit{true} \,|\, \mathit{Alarm} = \mathbf{true}) = 70\%$
- $\mathrm{P}(\mathit{MaryCalls} = \mathit{true} \,|\, \mathit{Alarm} = \mathbf{false}) = 1\%$

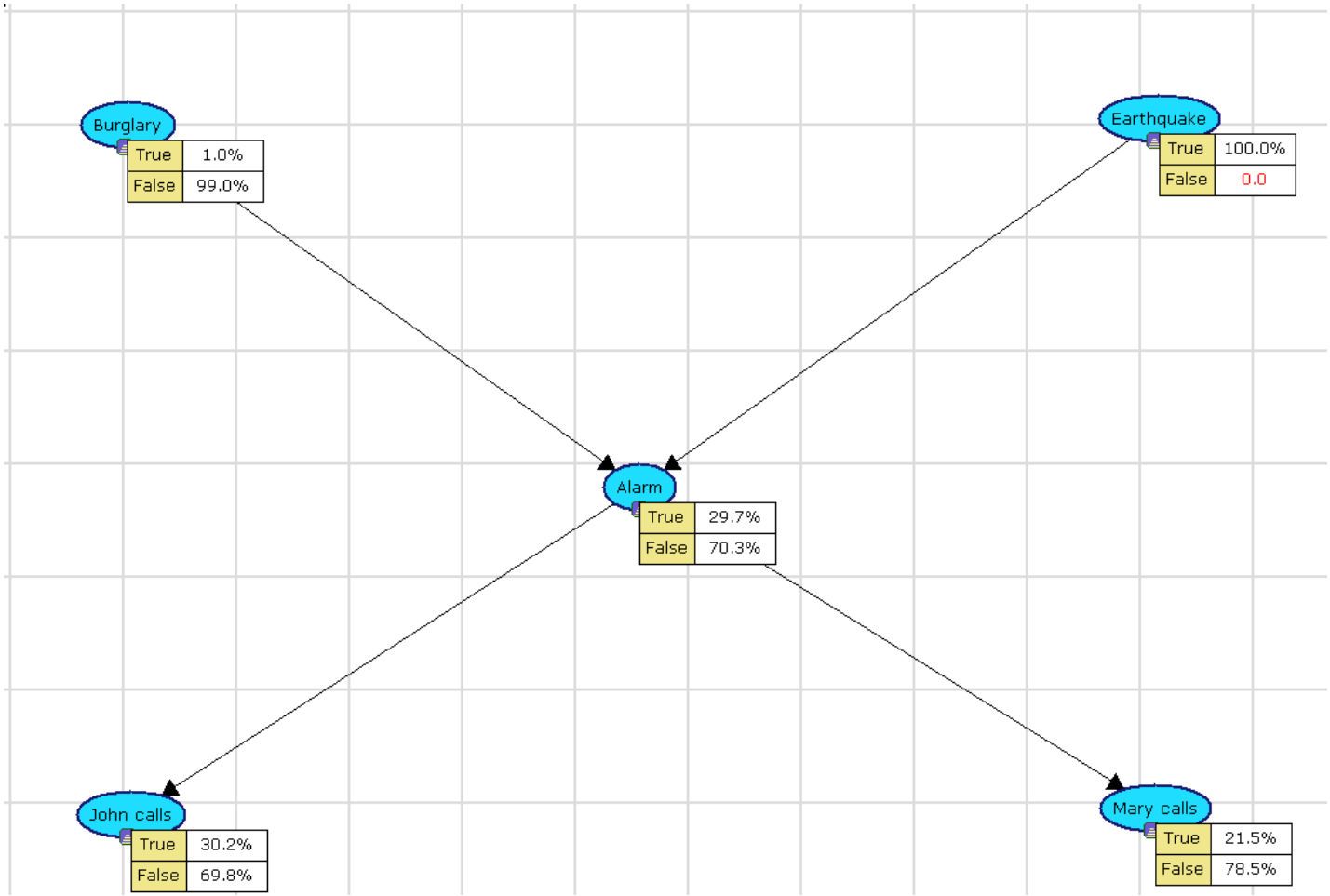
Krok 3: Zadání parametrů



Krok 4: Evidence



Krok 4: Výpočet podm. p.



Krok 5: Ruční výpočet (1/3)

$$\mathrm{P}(\mathit{MaryCalls} \mid \mathit{Earthquake} = \mathrm{true}) = \mathrm{P}(M \mid E) = \frac{\mathrm{P}(M, E)}{\mathrm{P}(E)} = \ldots \tag{1}$$

$$\begin{aligned}
\mathrm{P}(M, E) &= \sum_A \sum_B \sum_J \mathrm{P}(B, E, A, M, J) && (2)\\
&= \sum_A \sum_B \sum_J \mathrm{P}(B) \cdot \mathrm{P}(E) \cdot \mathrm{P}(A \mid B, E) \cdot \mathrm{P}(M \mid A) \cdot \mathrm{P}(J \mid A) && (3)\\
&= \sum_A \mathrm{P}(M \mid A) \sum_B \sum_J \mathrm{P}(B) \cdot \mathrm{P}(E) \cdot \mathrm{P}(A \mid B, E) \cdot \mathrm{P}(J \mid A) && (4)\\
= \cdots &= \mathrm{P}(E) \sum_A \mathrm{P}(M \mid A) \sum_B \mathrm{P}(B) \cdot \mathrm{P}(A \mid B, E) \cdot \sum_J \mathrm{P}(J \mid A) && (5)\\
&= \mathrm{P}(E) \sum_A \mathrm{P}(M \mid A) \sum_B \mathrm{P}(B) \cdot \mathrm{P}(A \mid B, E) && (6)
\end{aligned}$$

Tahák: $\sum_X \mathrm{P}(X) = 1$; $\sum_X \mathrm{P}(X, Y) = \mathrm{P}(Y)$;
$\mathrm{P}(M, E) = \mathrm{P}(E) \cdot \mathrm{P}(M \mid E)$; $\sum_Y f(X) \cdot g(Y) = f(X) \sum_Y g(Y)$.

Krok 5: Ruční výpočet (2/3)

$$\mathrm{P}(\mathit{MaryCalls} \mid \mathit{Earthquake} = \mathrm{true}) = \frac{\mathrm{P}(M, E = \mathrm{true})}{\mathrm{P}(E = \mathrm{true})} =$$

$$= \frac{\mathrm{P}(E = \mathrm{true}) \sum_A \mathrm{P}(M \mid A) \sum_B \mathrm{P}(B) \cdot \mathrm{P}(A \mid B, E = \mathrm{true})}{\mathrm{P}(E = \mathrm{true})}$$

$$= \sum_A \mathrm{P}(M \mid A) \sum_B \mathrm{P}(B) \cdot \mathrm{P}(A \mid B, E = \mathrm{true})$$

$$= \sum_A \mathrm{P}(M \mid A) \sum_B \left(\begin{array}{c|c} B = & \\ \hline \mathrm{true} & 0.01 \\ \mathrm{false} & 0.99 \end{array} \right) \times \left(\begin{array}{c|cc} & \multicolumn{2}{c}{\mathrm{A}} \\ & \mathrm{true} & \mathrm{false} \\ \hline B = \mathrm{true} & 0.95 & 0.05 \\ \mathrm{false} & 0.29 & 0.71 \end{array} \right)$$

$$= \sum_A \mathrm{P}(M \mid A) \sum_B \left(\begin{array}{c|cc} & \multicolumn{2}{c}{\mathrm{A}} \\ & \mathrm{true} & \mathrm{false} \\ \hline B = \mathrm{true} & 0.0095 & 0.0005 \\ \mathrm{false} & 0.2871 & 0.7029 \end{array} \right)$$

$$= \sum_A \mathrm{P}(M \mid A) \left(\begin{array}{cc} \multicolumn{2}{c}{\mathrm{A}} \\ \mathrm{true} & \mathrm{false} \\ \hline 0.2966 & 0.7034 \end{array} \right)$$

Krok 5: Ruční výpočet (2/3)

$$
\begin{aligned}
&= \sum_A \mathrm{P}(M \mid A) \times \left(\begin{array}{c|cc} & \multicolumn{2}{c}{\mathsf{A}} \\ & \text{true} & \text{false} \\ \hline & 0.2966 & 0.7034 \end{array} \right) \\
&= \sum_A \left(\begin{array}{cc|cc} & & \multicolumn{2}{c}{\mathsf{A}} \\ & & \text{true} & \text{false} \\ \hline M = & \text{true} & 0.70 & 0.01 \\ & \text{false} & 0.30 & 0.99 \end{array} \right) \times \left(\begin{array}{c|cc} & \multicolumn{2}{c}{\mathsf{A}} \\ & \text{true} & \text{false} \\ \hline & 0.2966 & 0.7034 \end{array} \right) \\
&= \sum_A \left(\begin{array}{cc|cc} & & \multicolumn{2}{c}{\mathsf{A}} \\ & & \text{true} & \text{false} \\ \hline M = & \text{true} & 0.20762 & 0.007034 \\ & \text{false} & 0.08898 & 0.696366 \end{array} \right) \\
&= \underline{\underline{\left(\begin{array}{cc|c} M = & \text{true} & 0.214654 \\ & \text{false} & 0.785346 \end{array} \right)}}
\end{aligned}
$$

Úloha: Popis dat

Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	Wait
Yes	No	No	Yes	Some	3	No	Yes	French	0-10	Yes
Yes	No	No	Yes	Full	1	No	No	Thai	>30	No
No	Yes	No	No	Some	1	No	No	Burger	0-10	Yes
Yes	No	Yes	Yes	Full	1	No	No	Thai	10-30	Yes
Yes	No	Yes	No	Full	3	No	Yes	French	>30	No
No	Yes	No	Yes	Some	2	Yes	Yes	Italian	0-10	Yes
No	Yes	No	No	None	1	Yes	No	Burger	0-10	No
No	No	No	Yes	Some	2	Yes	Yes	Thai	0-10	Yes
No	Yes	Yes	No	Full	3	No	Yes	Italian	>30	No
Yes	Yes	Yes	Yes	Full	3	No	Yes	Italian	>30	No
No	No	No	No	None	1	No	No	Thai	0-10	No
Yes	Yes	Yes	Yes	Full	1	No	No	Burger	>30	Yes

Úloha: Legenda

1. Alt: whether there is a suitable alternative restaurant nearby.
2. Bar: whether the restaurant has a comfortable bar area to wait in.
3. Fri: true on Fridays and Saturdays.
4. Hun: whether we are hungry.
5. Pat: how many people are in the restaurant (values are None, Some, and Full).
6. Price: the restaurant's price range ($, $$, $$$).
7. Rain: whether it is raining outside.
8. Res: whether we made a reservation.
9. Type: the kind of restaurant (French, Italian, Thai, or Burger).
10. Est: the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).
11. Wait: whether we decided to wait

Úloha: Zadání (1/2)

1. Navrhněte strukturu Bayesovské sítě na základě přiložených dat. Snažte se respektovat kauzální vazby mezi náhodnými proměnnými. Složitost sítě by měla odpovídat množství dostupných pozorování.
2. Z dodaných dat vypočtěte tabulky podmíněných pravděpodobností, které odpovídají struktuře vaší BS.
3. Pomocí počítače vypočtěte následující pravděpodobnosti z Bayesovské sítě:
   (1) $\mathrm{P}(\mathit{Est})$
   (2) $\mathrm{P}(\mathit{Est} \mid \mathit{Pat})$
   (3) $\mathrm{P}(\mathit{Rain})$
   (4) $\mathrm{P}(\mathit{Rain} \mid \mathit{Fri})$

Úloha: Zadání (2/2)

4. Podmíněnou pravděpodobnost (4) vypočtěte navíc ručně.
5. Porovnejte výsledky (1) s (2) a dále (3) s (4). Ovlivňuje dotazovanou proměnou informace o proměnné v podmínce?
6. Proč nelze počítat podmíněné pravděpodobnosti přímo z dat (marginalizací) a je dobré využít strukturu BS?