

# **Aplikace bayesovských sítí**

**Jiří Vomlel**

**Ústav teorie informace a automatizace  
Akademie věd České republiky**

**Tato prezentace je k dispozici na:**

**<http://www.utia.cas.cz/vomlel/>**

# Obsah přednášky

- Podmíněná pravděpodobnost, Bayesův vzorec
- Nezávislost a podmíněná nezávislost
- Řetězcové pravidlo
- Definice bayesovské sítě
- Výhody reprezentace bayesovskou sítí
- Aplikace 1: modelování dědičných nemocí
- Aplikace 2: podpora rozhodování
- Aplikace 3: technická diagnostika
- Aplikace 4: adaptivní testování znalostí

## Podmíněná pravděpodobnost

Podmíněná pravděpodobnost veličiny  $A$  dáno veličina  $B$  je pravděpodobnostní distribuce  $P(A|B)$  splňující vztah

$$P(A|B) \cdot P(B) = P(A, B). \text{ Jestliže } P(B) \text{ je nenulové pak}$$
$$P(A|B) = \frac{P(A, B)}{P(B)}.$$

Například, mějme dvě binární veličiny:

- délka vlasů s hodnotami dlouhé a krátké
- pohlaví s hodnotami muž, žena

$$P(\text{muž} \mid \text{dlouhé vlasy}) = \frac{P(\text{muž, dlouhé vlasy})}{P(\text{dlouhé vlasy})}$$

$$P(\text{žena} \mid \text{dlouhé vlasy}) = \frac{P(\text{žena, dlouhé vlasy})}{P(\text{dlouhé vlasy})}$$

# Bayesův vzorec

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{\sum_A P(A, B)} = \frac{P(B|A) \cdot P(A)}{\sum_A P(B|A) \cdot P(A)}$$

Například mějme dvě veličiny:

- *R - Rain* ... v noci pršelo s hodnotami  $y = yes$  a  $n = no$ .
- *W - Wet grass* ... tráva je mokrá s hodnotami  $y = yes$  a  $n = no$ .

Víme, že:

- jestliže v noci pršelo pak je tráva mokrá s pravděpodobností  $\frac{3}{4}$ ,
- jestliže v noci nepršelo pak je tráva mokrá s pravděpodobností  $\frac{1}{8}$ ,
- pravděpodobnost, že bude v noci pršet je  $\frac{1}{3}$ .

Ráno vidíme, že tráva je mokrá. Jaká je pravděpodobnost, že v noci pršelo?

## Bayesův vzorec

$$\begin{aligned} P(R = y|W = y) &= \frac{P(W = y|R = y) \cdot P(R = y)}{P(W = y|R = y) \cdot P(R = y) + P(W = y|R = n) \cdot P(R = n)} \\ &= \frac{\frac{3}{4} \cdot \frac{1}{3}}{\frac{3}{4} \cdot \frac{1}{3} + \frac{1}{8} \cdot \frac{2}{3}} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{12}} = \frac{\frac{1}{4}}{\frac{4}{12}} = \frac{3}{4} \end{aligned}$$

Pravděpodobnost, že v noci přelo je  $\frac{3}{4}$ .

# Nezávislost



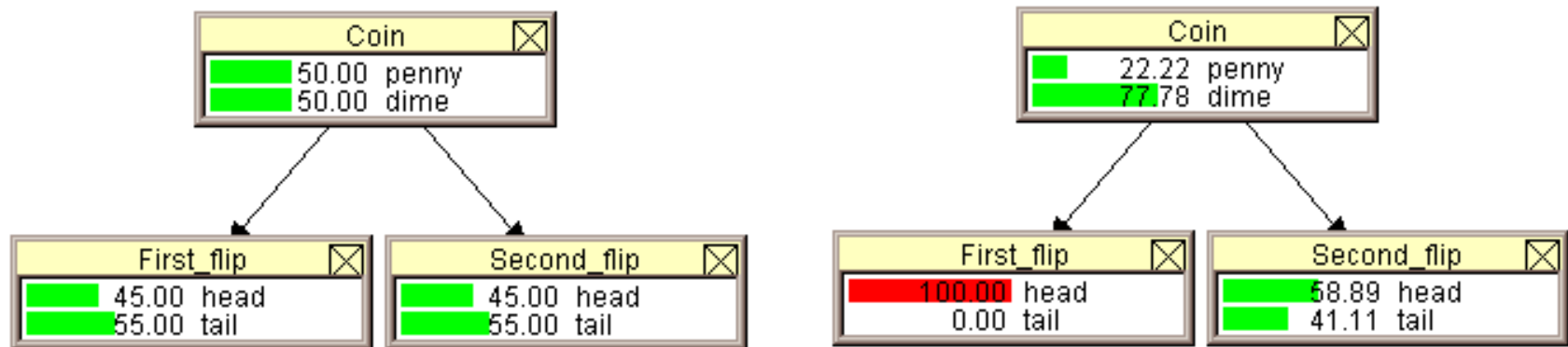
Pravděpodobnost společného výskytu hodnot veličin je rovna součinu pravděpodobností hodnot jednotlivých veličin.

$$P(Dime = head, Penny = head) = P(Dime = head) \cdot P(Penny = head)$$

Též, zjistíme-li hodnotu jedné veličiny, nemá to vliv na hodnotu druhé veličiny.

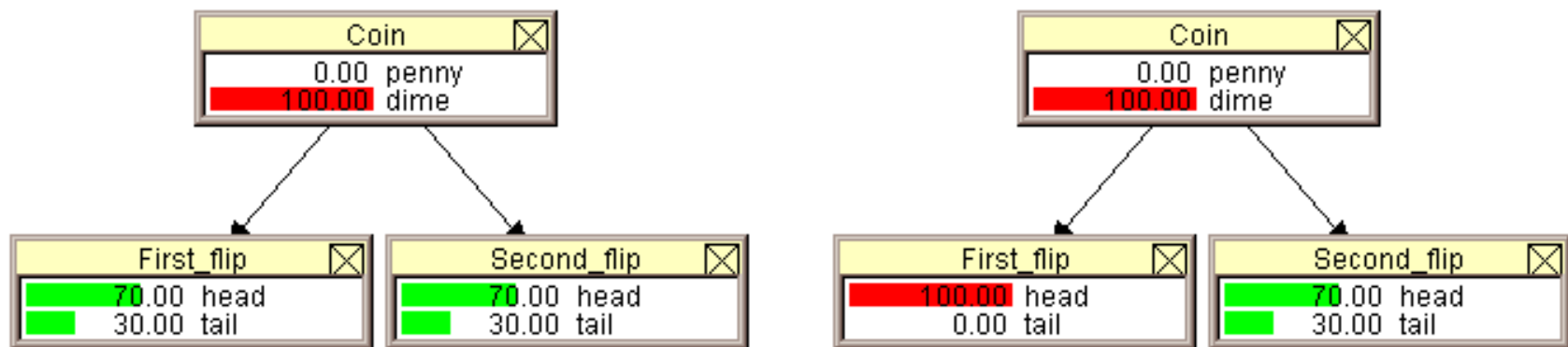
$$P(Dime = head | Penny = head) = P(Dime = head)$$

# Náhodně vybereme jednu minci pro dva hody.



První hod **má vliv** na pravděpodobnost výsledku **druhého** hodu.

Nyní, předpokládejme, že známe vybranou minci.



Jestliže víme,  **která mince**  bude použita pak výsledek  **prvního**  hodů  **nemá vliv**  na pravděpodobnost výsledku  **druhého**  hodů.

# Podmíněná nezávislost

Pravděpodobnost společného výskytu hodnot veličin při dané hodnotě třetí veličiny je rovna součinu podmíněných pravděpodobností jednotlivých veličin dáno třetí veličina:

$$\begin{aligned} &P(\textit{First\_flip} = \textit{head}, \textit{Second\_flip} = \textit{head} | \textit{Coin} = \textit{dime}) \\ &= P(\textit{First\_flip} = \textit{head} | \textit{Coin} = \textit{dime}) \cdot P(\textit{Second\_flip} = \textit{head} | \textit{Coin} = \textit{dime}) \end{aligned}$$

Jestliže neznáme minci, výsledek **prvního** hodů **má vliv** na pravděpodobnost výsledku **druhého** hodů.

Jestliže známe minci, pak výsledek **prvního** hodů **nemá vliv** na pravděpodobnost výsledku **druhého** hodů.

$$\begin{aligned} &P(\textit{Second\_flip} = \textit{head} | \textit{Coin} = \textit{dime}, \textit{First\_flip} = \textit{head}) \\ &= P(\textit{Second\_flip} = \textit{head} | \textit{Coin} = \textit{dime}) \end{aligned}$$

# Řetězcové pravidlo

S definice podmíněné pravděpodobnosti plyne, že můžeme psát:

$$\begin{aligned}P(A, B, C, D) &= P(A|B, C, D) \cdot P(B, C, D) \\ &= P(A|B, C, D) \cdot P(B|C, D) \cdot P(C, D) \\ &= P(A|B, C, D) \cdot P(B|C, D) \cdot P(C|D) \cdot P(D)\end{aligned}$$

# Proč je Holmesův trávník mokrý?

*Holm* Je Holmesův trávník mokrý?

*Rn* Pršelo v noci?

*Sprnk* Byl Holmesův postřikovač zapnutý?

*Wat* Je Watsonův trávník mokrý?

- Holmesův trávník může být mokrý buď protože pršelo, nebo protože měl zapnutý postřikovač.
- Watsonův trávník může být mokrý protože pršelo. Holmesův postřikovač nemá vliv na Watsonův trávník.
- Déšť nesouvisí s tím, jestli má Holmes zapnutý postřikovač.

Řetězcové pravidlo a podmíněné nezávislosti uvedené výše dávají:

$P(\text{Holm}, \text{Wat}, \text{Rn}, \text{Sprnk})$

$$= P(\text{Holm} | \text{Wat}, \text{Rn}, \text{Sprnk}) \cdot P(\text{Wat} | \text{Rn}, \text{Sprnk}) \cdot P(\text{Rn} | \text{Sprnk}) \cdot P(\text{Sprnk})$$

$$= P(\text{Holm} | \text{Rn}, \text{Sprnk}) \cdot P(\text{Wat} | \text{Rn}) \cdot P(\text{Rn}) \cdot P(\text{Sprnk})$$

# Proč je Holmesův trávník mokrý?

$P(\text{Holm}, \text{Wat}, \text{Rn}, \text{Sprnk})$

$$= P(\text{Holm}|\text{Rn}, \text{Sprnk}) \cdot P(\text{Wat}|\text{Rn}) \cdot P(\text{Rn}) \cdot P(\text{Sprnk})$$

Holmes?				
Edit	Functions	View		
Sprinkler?	yes		no	
Rain?	yes	no	yes	no
yes	1	0.9	1	0
no	0	0.1	0	1

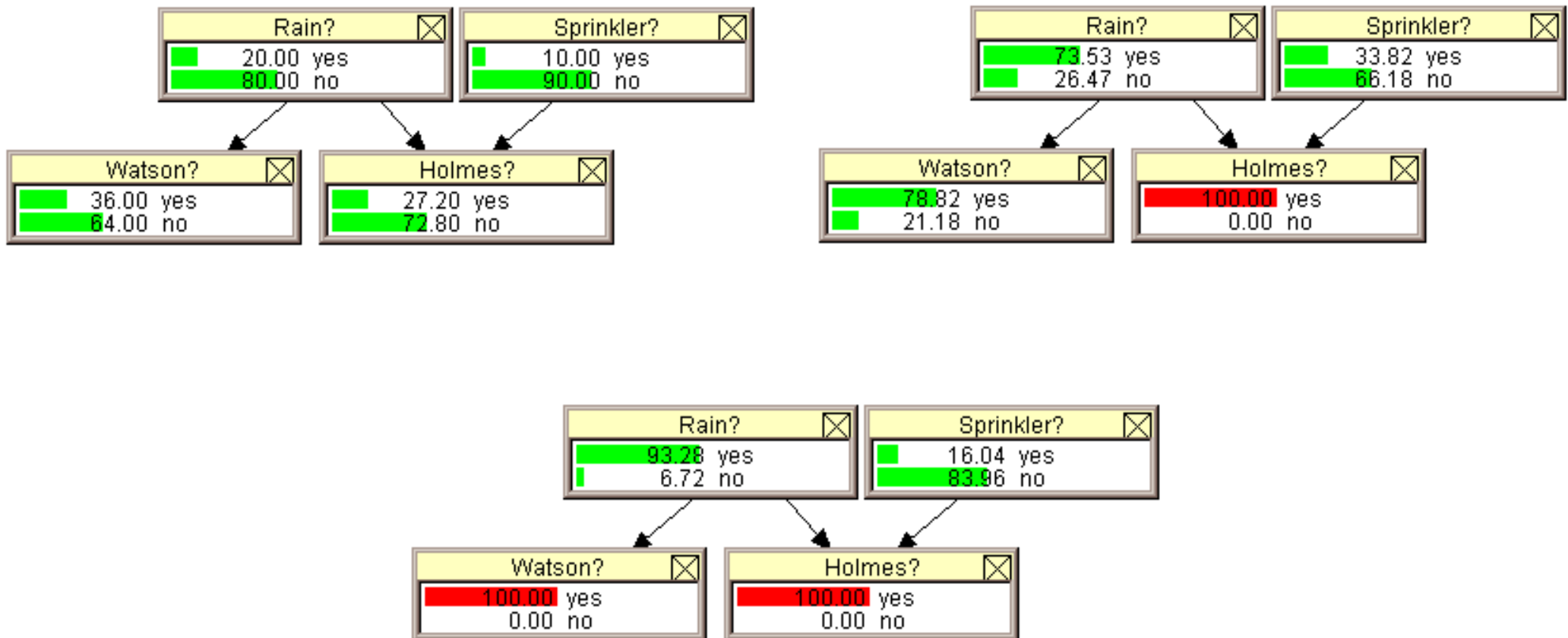
Watson?		
Edit	Functions	View
Rain?	yes	no
yes	1	0.2
no	0	0.8

Sprinkler?	
Edit	View
yes	0.1
no	0.9

Rain?	
Edit	View
yes	0.2
no	0.8

```
graph TD; Rain((Rain?)) --> Watson((Watson?)); Rain((Rain?)) --> Holmes((Holmes?)); Sprinkler((Sprinkler?)) --> Holmes((Holmes?));
```

# Byl Holmesův postřikovač zapnutý?



## Definice bayesovské sítě

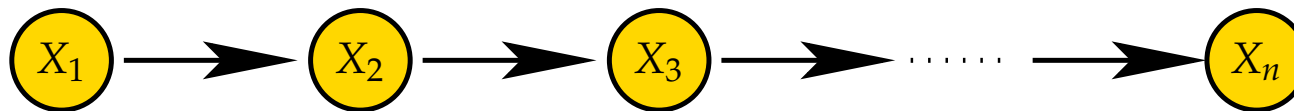
- acyklický orientovaný graf (DAG)  $G = (V, E)$
- každý uzel  $i \in V$  odpovídá jedné náhodné veličině  $X_i$  s konečným počtem navzájem disjunktních hodnot  $\mathbb{X}_i$
- $pa(i)$  bude označovat množinu rodičů uzlu  $i$  v grafu  $G$
- ke každému uzlu  $i \in V$  odpovídá podmíněná pravděpodobnostní distribuce  $P(X_i \mid (X_j)_{j \in pa(i)})$
- acyklický orientovaný graf (DAG) reprezentuje podmíněné nezávislostní vztahy mezi veličinami  $(X_i)_{i \in V}$

# Výhoda reprezentace bayesovskou sítí

Předpokládejme, že máme problém, který budeme modelovat pomocí  $n$  veličin a každá veličina může nabývat dvou hodnot.

Použijeme-li reprezentaci pomocí jedné tabulky potřebujeme pro uložení v paměti počítače distribuce  $2^n - 1$  hodnot.

Předpokládejme, bayesovskou sít' mající též  $n$  veličin nabývajících dvou hodnot s grafem následující struktury:

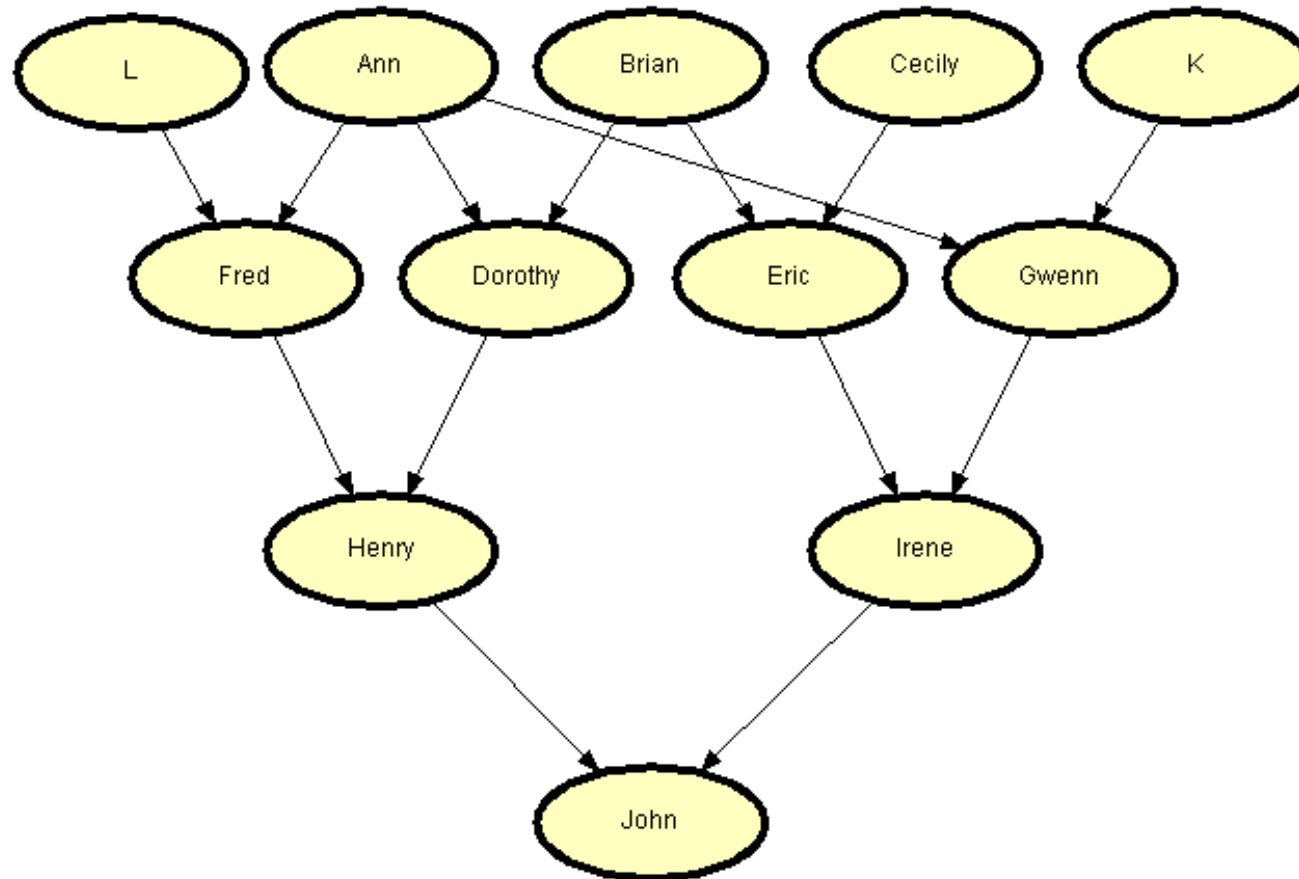


Pro její uložení v paměti počítače potřebujeme  $1 + (n - 1) \cdot 2 = 2n - 1$  hodnot.

# Výhoda reprezentace bayesovskou sítí

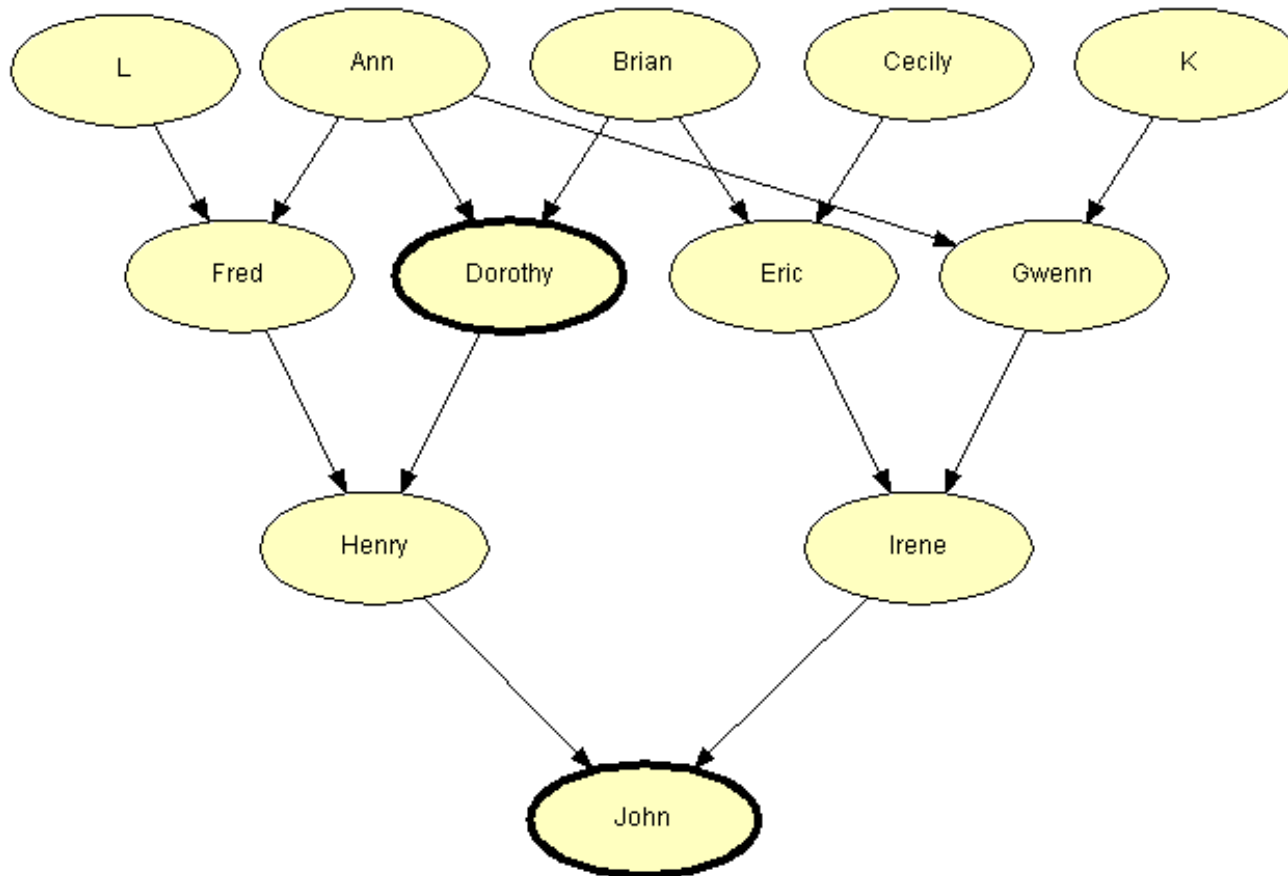
$n$	$2^n - 1$	$2n - 1$
2	3	3
3	7	5
4	15	7
10	1023	19
100	$1.27 \cdot 10^{30}$	199
1000	$1.07 \cdot 10^{301}$	1999

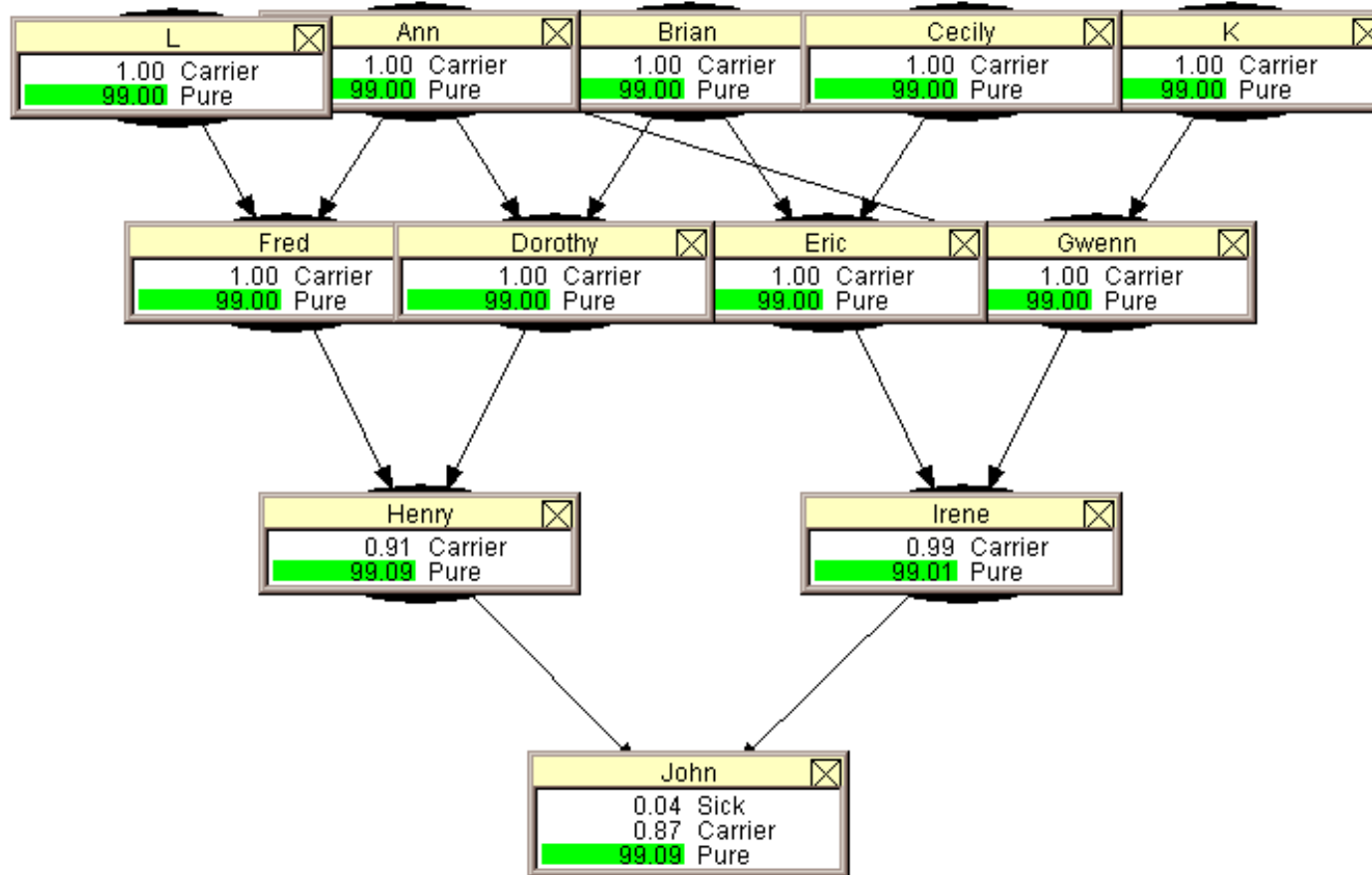
## Aplikace 1: modelování dědičných nemocí

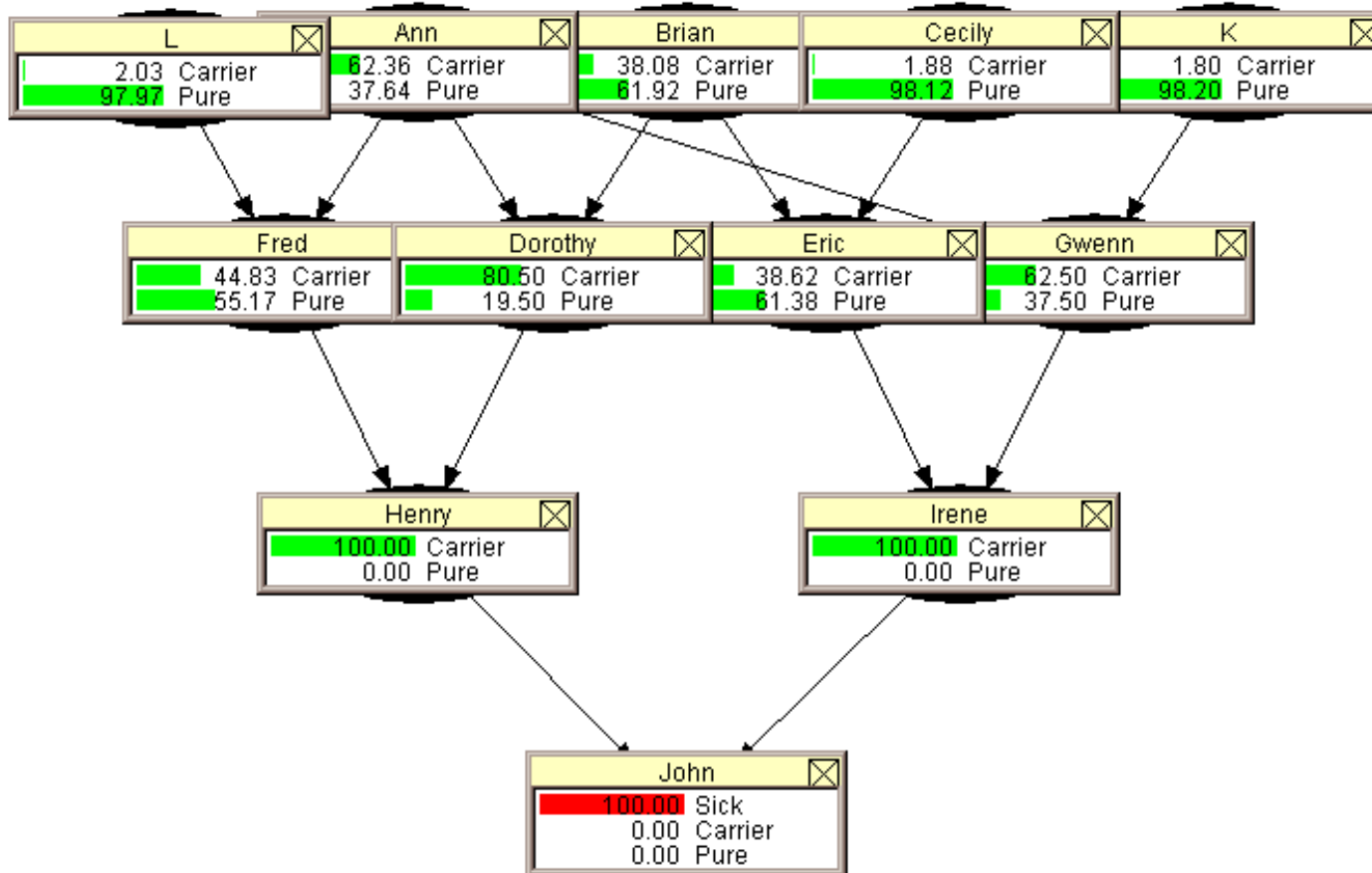


Dorothy				
Ann	Carrier		Pure	
Brian	Carrier	Pure	Carrier	Pure
Carrier	0.6666...	0.5	0.5	0
Pure	0.3333...	0.5	0.5	1

John				
Henry	Carrier		Pure	
Irene	Carrier	Pure	Carrier	Pure
Sick	0.25	0	0	0
Carrier	0.5	0.5	0.5	0
Pure	0.25	0.5	0.5	1

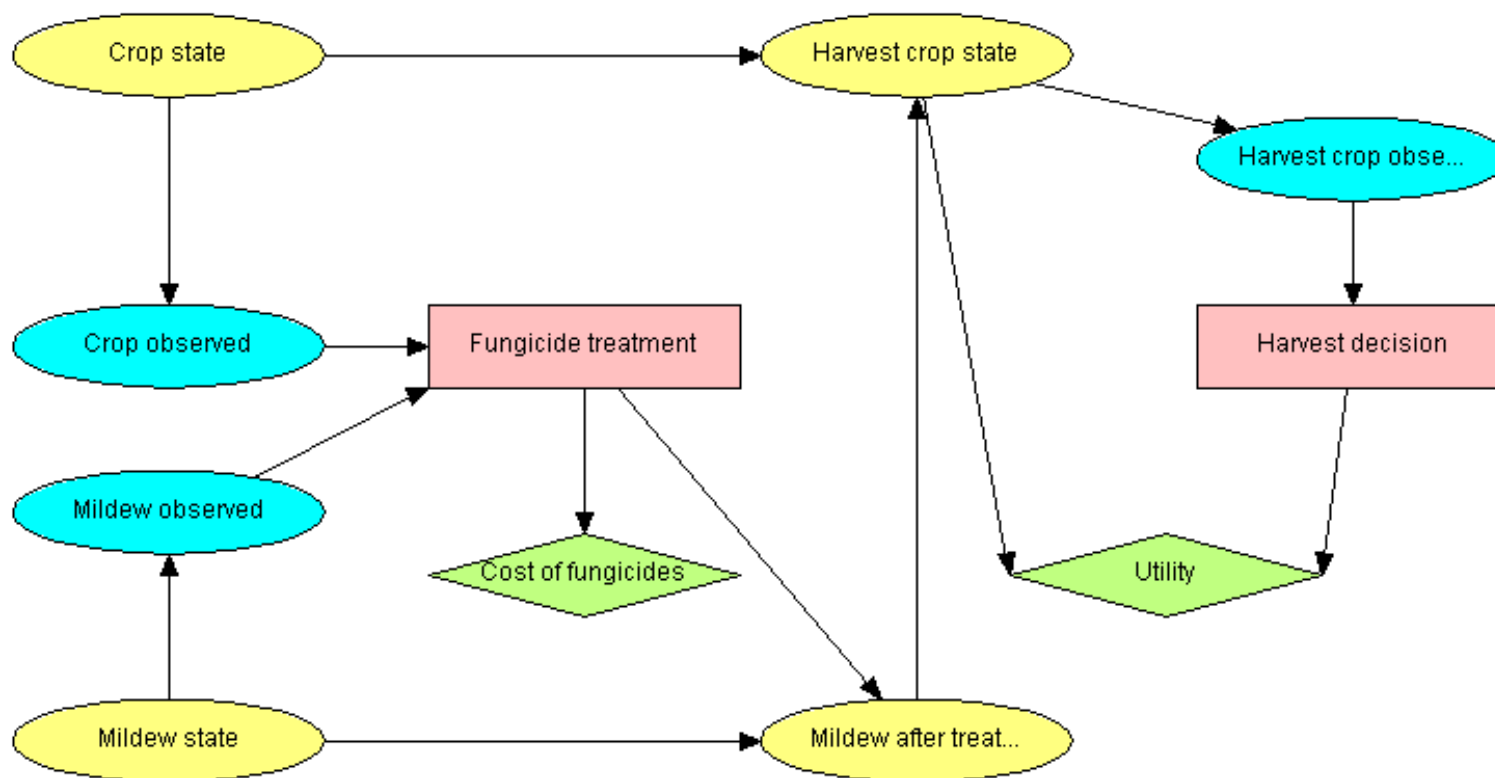


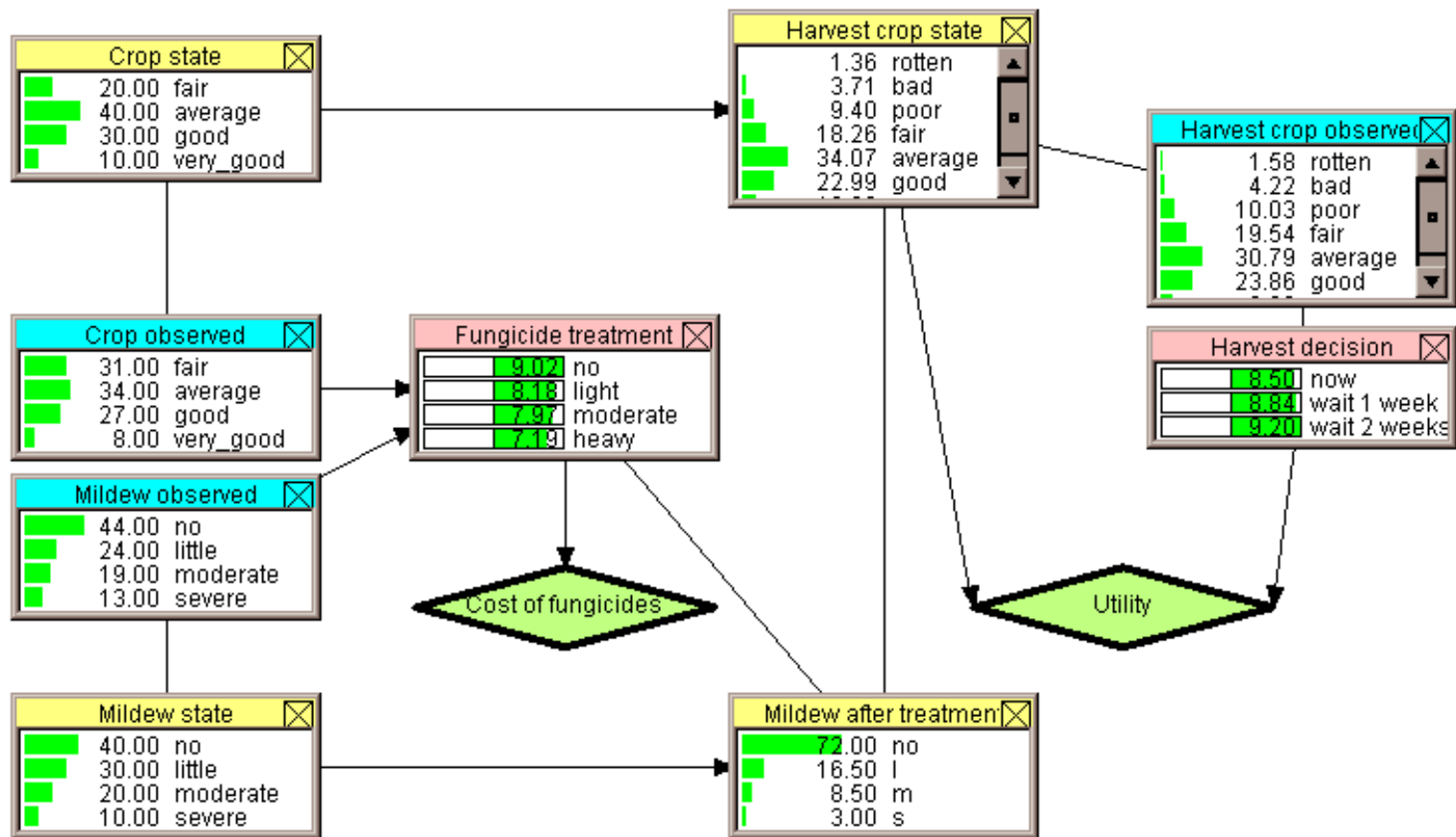


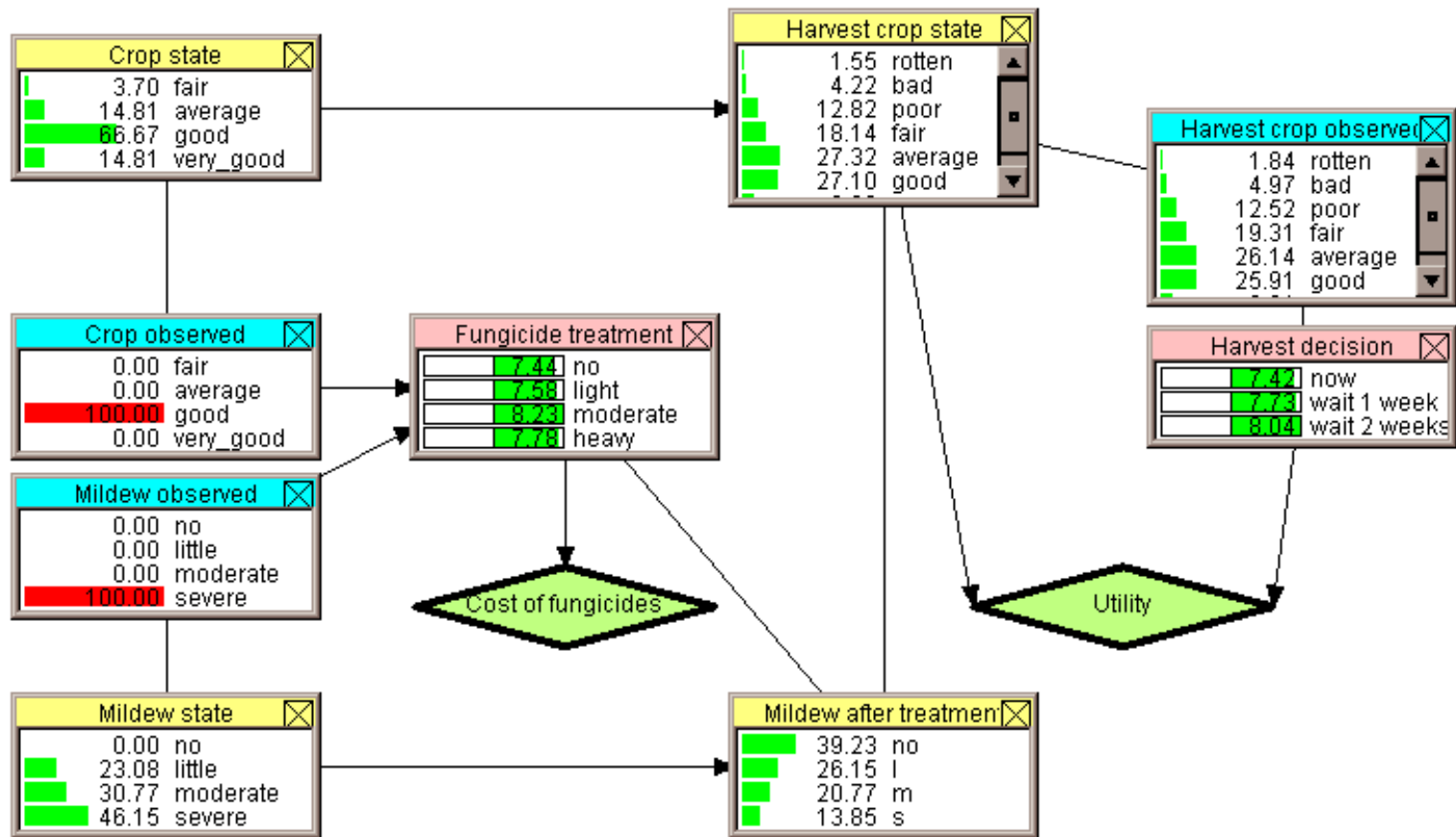


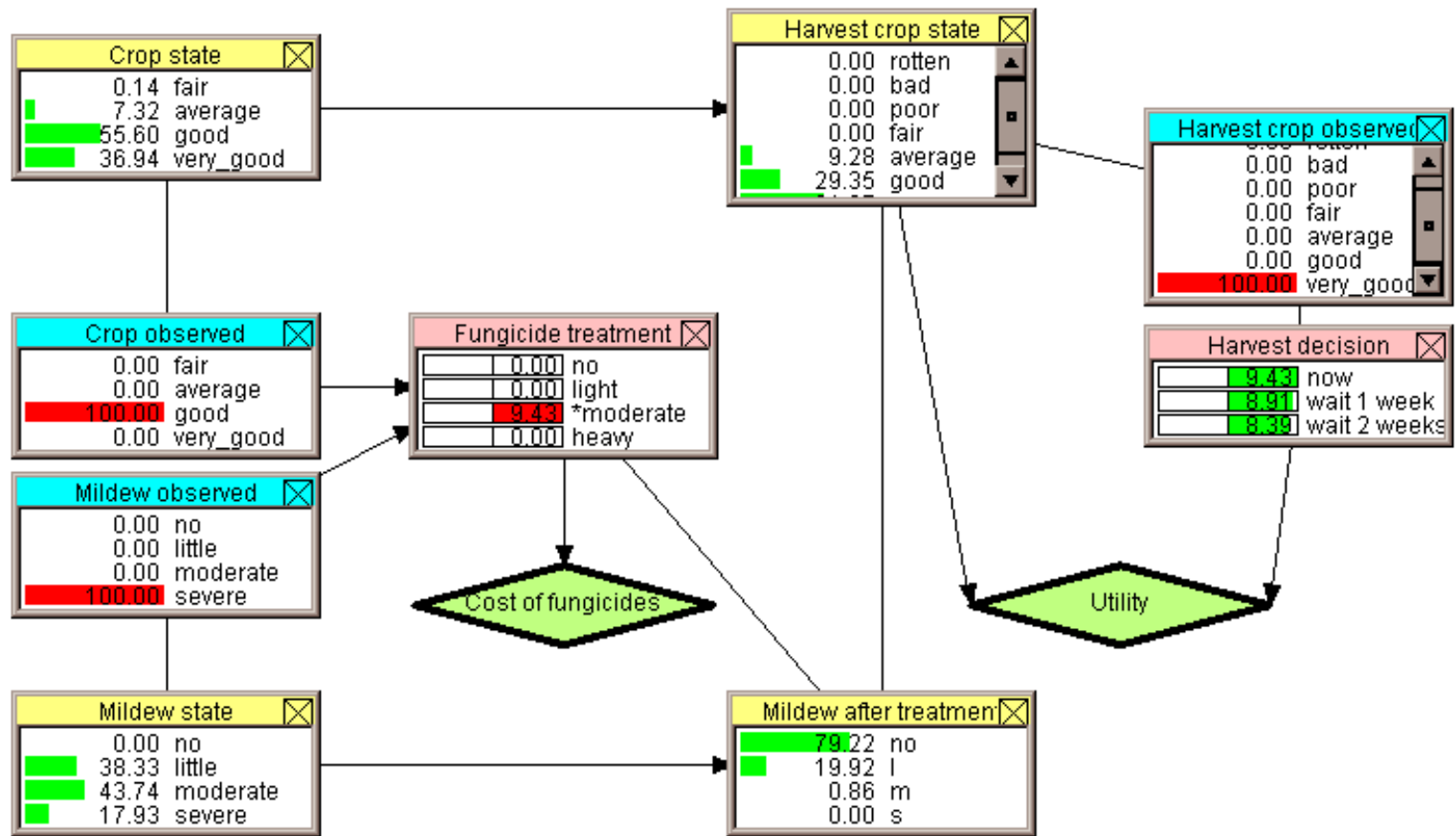
## Aplikace 2: podpora rozhodování

Cíl: maximalizace očekávaného užitku









## Aplikace 3: Technická diagnostika - popis problému

- Příčiny problému (závady)  $C \in \mathcal{C}$ .
- Akce  $A \in \mathcal{A}$  - opravné kroky, které mohou odstranit závadu.
- Otázky  $Q \in \mathcal{Q}$  - kroky, které mohou pomoci identifikovat, kde je závada.
- Ke každé akci i otázce je přiřazena cena ( $c_A$  značí cenu akce  $A$ ,  $c_Q$  cenu otázky  $Q$ ). Cena může znamenat:
  - dobu potřebnou k provedení akce či otázky,
  - cenu za náhradní díl, který použijeme
  - rizikovost akce
  - nějaká kombinace výše uvedených.

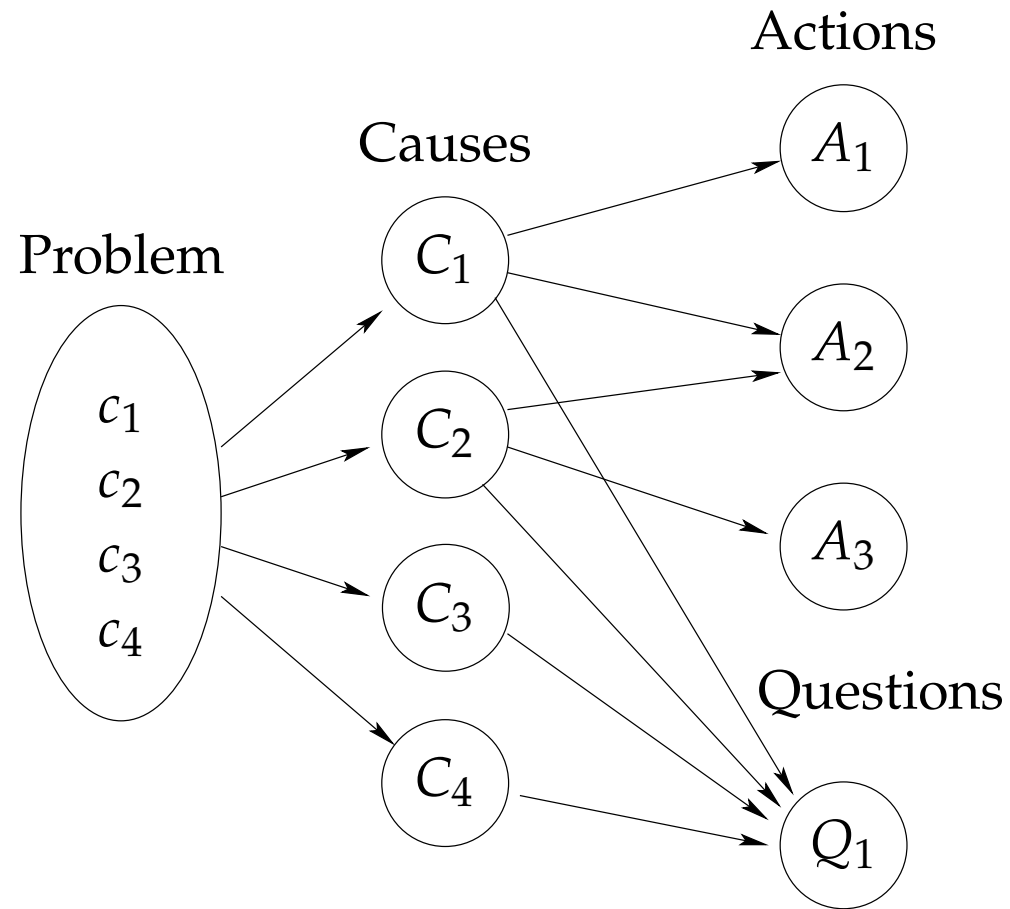
## Příklad technické diagnostiky tiskárny

**Trouble:** světlý tisk.

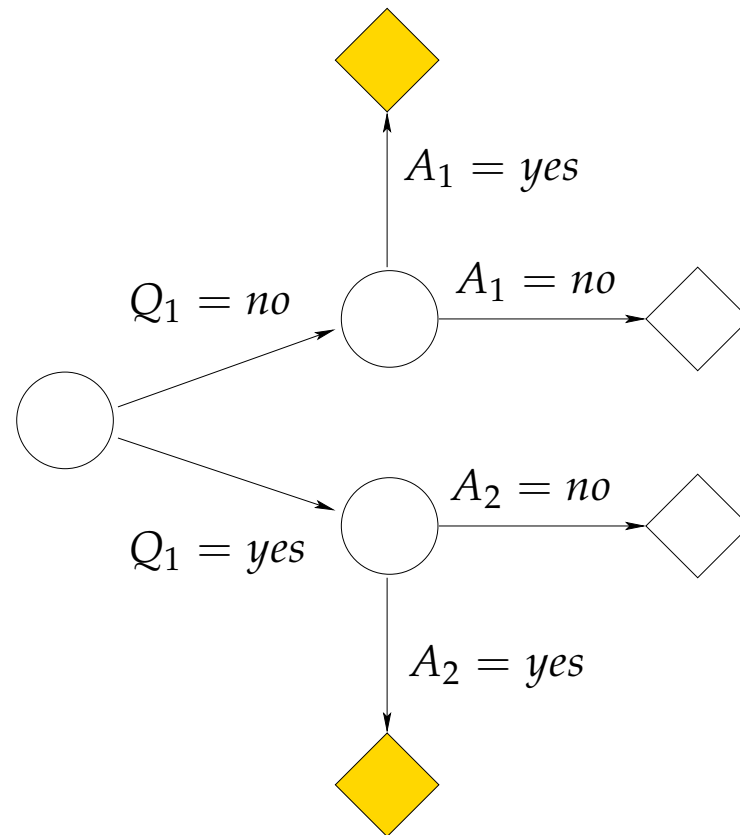
**Troubeshooter:** doporučí kroky, které pomohou odstranit “trouble”

Akce a otázky	cena
$A_1$ : Remove, shake and reseal toner	5
$A_2$ : Try another toner	15
$A_3$ : Cycle power	1
$Q_1$ : Is the printer configuration page printed light?	2
Možné závady při světlém tisku	$P(C_i)$
$C_1$ : Toner low	0.4
$C_2$ : Defective toner	0.3
$C_3$ : Corrupted dataflow	0.2
$C_4$ : Wrong driver setting	0.1

# Light Print Problem - Bayesian Network



## Světlý tisk - strategie odstranění závady



## Application 4: Adaptivní testování znalostí

Příklady úloh:

$$T_1: \left(\frac{3}{4} \cdot \frac{5}{6}\right) - \frac{1}{8} = \frac{15}{24} - \frac{1}{8} = \frac{5}{8} - \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$

$$T_2: \frac{1}{6} + \frac{1}{12} = \frac{2}{12} + \frac{1}{12} = \frac{3}{12} = \frac{1}{4}$$

$$T_3: \frac{1}{4} \cdot 1\frac{1}{2} = \frac{1}{4} \cdot \frac{3}{2} = \frac{3}{8}$$

$$T_4: \left(\frac{1}{2} \cdot \frac{1}{2}\right) \cdot \left(\frac{1}{3} + \frac{1}{3}\right) = \frac{1}{4} \cdot \frac{2}{3} = \frac{2}{12} = \frac{1}{6} .$$

## Základní a operační dovednosti

---

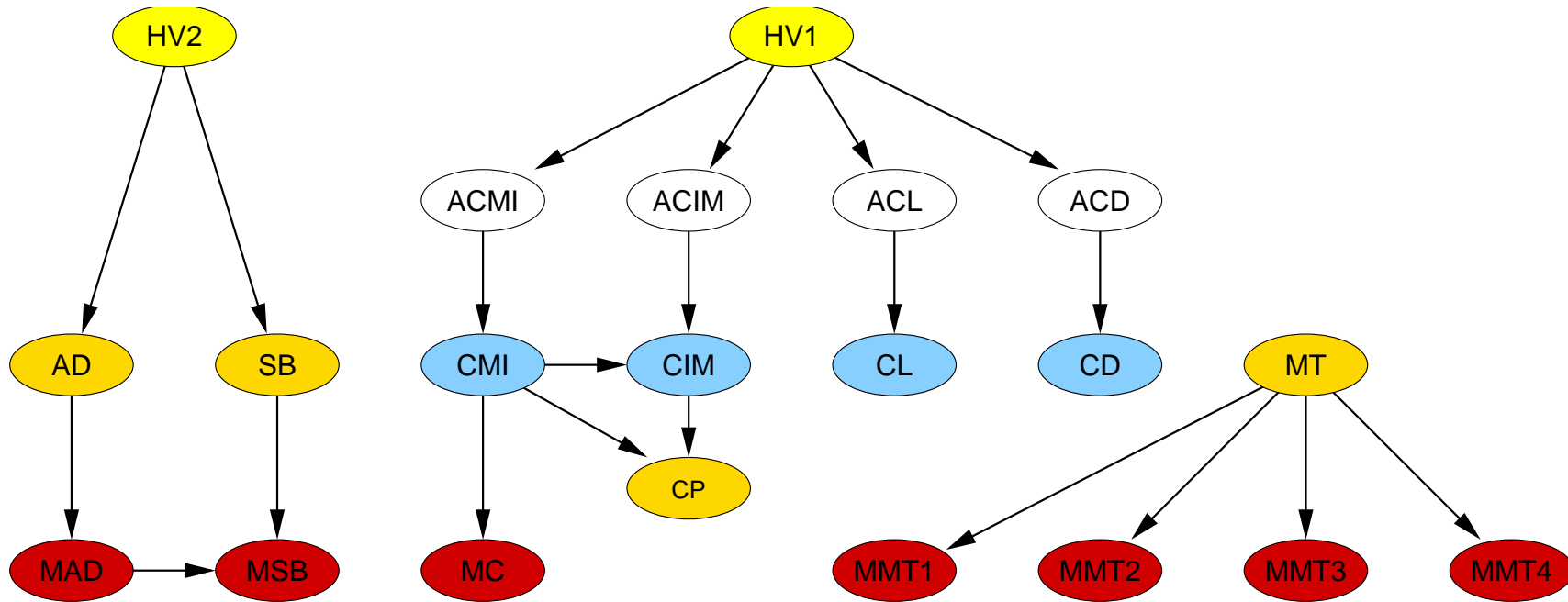
<b>CP</b>	Porovnávání (spol. čítec nebo jmenovatel)	$\frac{1}{2} > \frac{1}{3}, \frac{2}{3} > \frac{1}{3}$
<b>AD</b>	Sčítání (spol. jmenovatel)	$\frac{1}{7} + \frac{2}{7} = \frac{1+2}{7} = \frac{3}{7}$
<b>SB</b>	Odečítání (spol. jmenovatel)	$\frac{2}{5} - \frac{1}{5} = \frac{2-1}{5} = \frac{1}{5}$
<b>MT</b>	Násobení	$\frac{1}{2} \cdot \frac{3}{5} = \frac{3}{10}$
<b>CD</b>	Spol. jmenovatel	$\left(\frac{1}{2}, \frac{2}{3}\right) = \left(\frac{3}{6}, \frac{4}{6}\right)$
<b>CL</b>	Krácení	$\frac{4}{6} = \frac{2 \cdot 2}{2 \cdot 3} = \frac{2}{3}$
<b>CIM</b>	Konv. na slož. zlomek	$\frac{7}{2} = \frac{3 \cdot 2 + 1}{2} = 3\frac{1}{2}$
<b>CMI</b>	Konv. na nepravý zlomek	$3\frac{1}{2} = \frac{3 \cdot 2 + 1}{2} = \frac{7}{2}$

---

# Špatné postupy

Označení	Popis	Výskyt
<b>MAD</b>	$\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d}$	14.8%
<b>MSB</b>	$\frac{a}{b} - \frac{c}{d} = \frac{a-c}{b-d}$	9.4%
<b>MMT1</b>	$\frac{a}{b} \cdot \frac{c}{b} = \frac{a \cdot c}{b}$	14.1%
<b>MMT2</b>	$\frac{a}{b} \cdot \frac{c}{b} = \frac{a+c}{b \cdot b}$	8.1%
<b>MMT3</b>	$\frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot d}{b \cdot c}$	15.4%
<b>MMT4</b>	$\frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot c}{b+d}$	8.1%
<b>MC</b>	$a \frac{b}{c} = \frac{a \cdot b}{c}$	4.0%

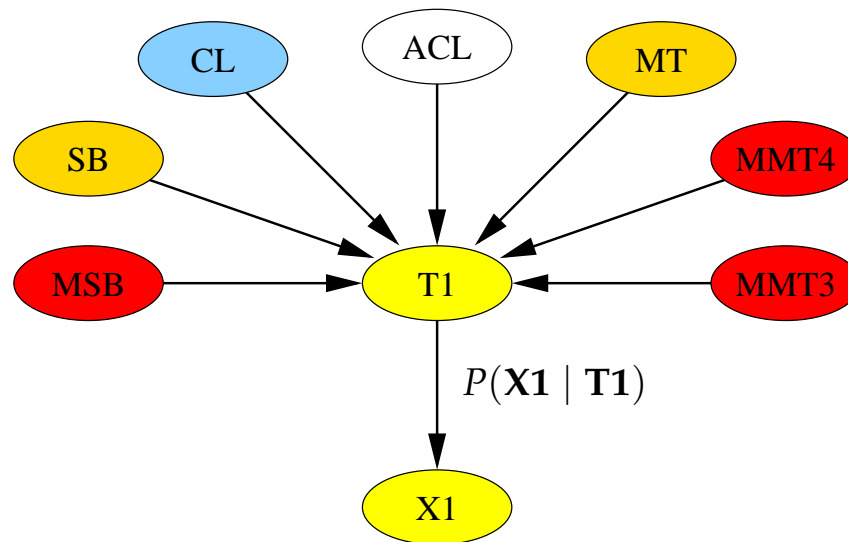
# Model studenta



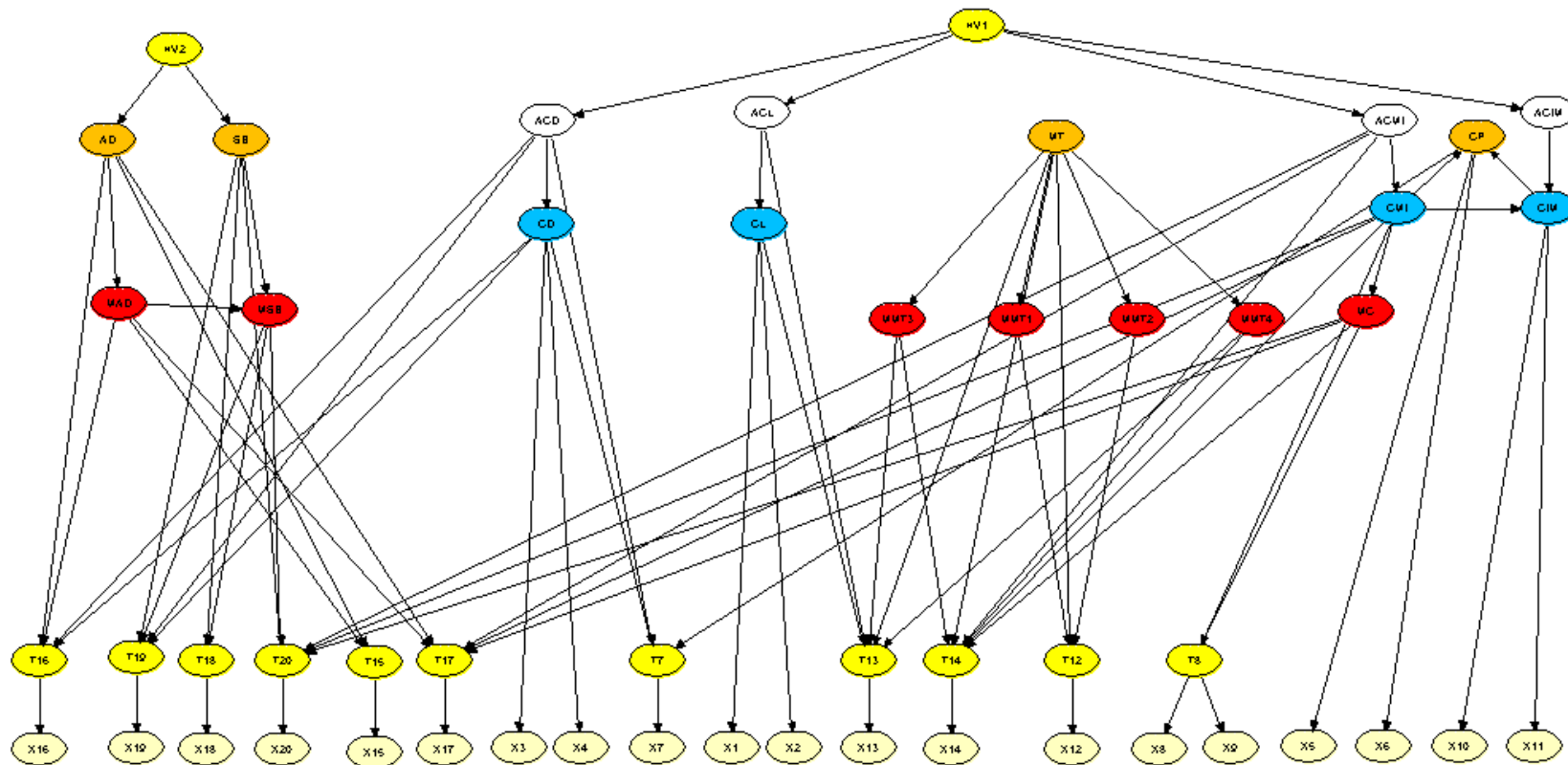
# Model úlohy T1

$$\left(\frac{3}{4} \cdot \frac{5}{6}\right) - \frac{1}{8} = \frac{15}{24} - \frac{1}{8} = \frac{5}{8} - \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$

$T1 \Leftrightarrow MT \ \& \ CL \ \& \ ACL \ \& \ SB \ \& \ \neg MMT3 \ \& \ \neg MMT4 \ \& \ \neg MSB$



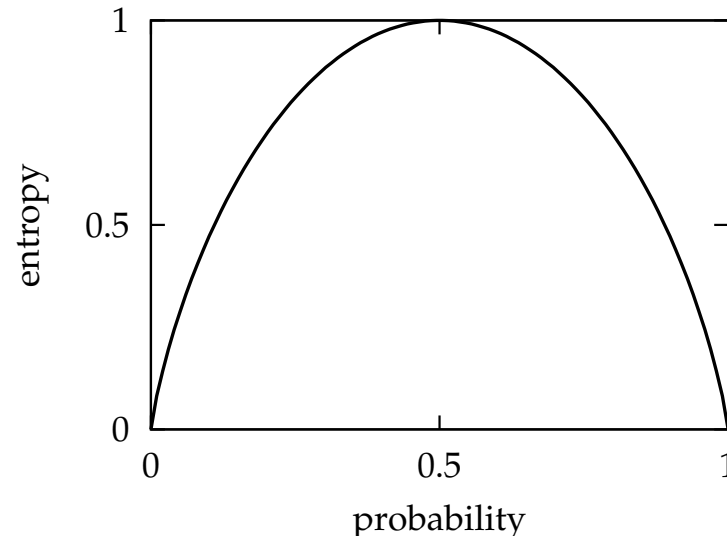
# Model studenta spojený s modely otázek



# Užitkovou funkcí je informační zisk

*“Čím nižší je entropie, tím více o studentovi víme.”*

$$H(P(\mathbf{X})) = - \sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}) \cdot \log P(\mathbf{X} = \mathbf{x})$$



Informační zisk v uzlu  $n$  strategie

$$IG(\mathbf{e}_n) = H(P(\mathbf{S})) - H(P(\mathbf{S} | \mathbf{e}_n))$$

# Skill Prediction Quality

