Rozhodovací stromy

Monika Stambolidis
Jindra Nováková
2010/2011

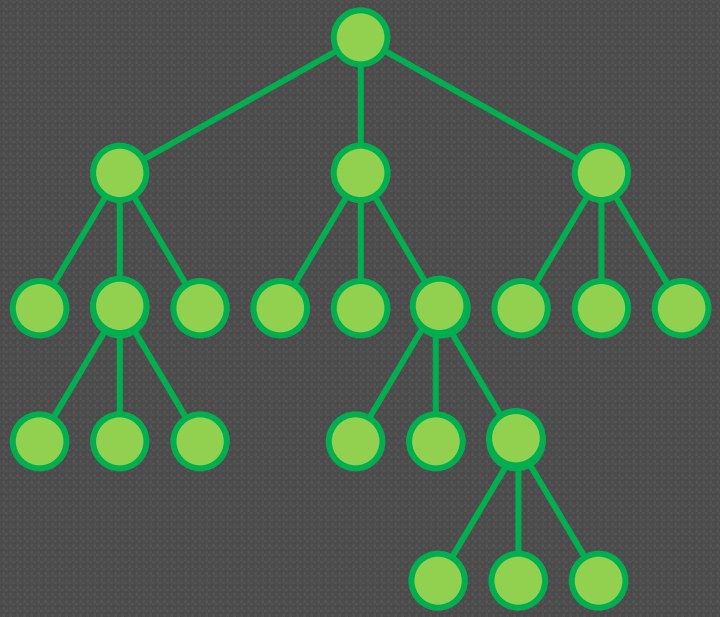
Osnova

- Co jsou stromy?
- K čemu jsou dobré?
- Jak se používají?
- Příklad
- Algoritmy pro vytváření rozhodovacích stromů

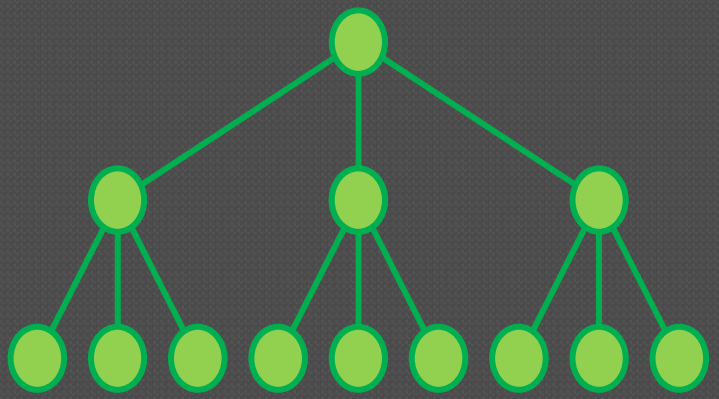
Co je strom?

- Souvislý graf bez kružnic
- Struktura je podobná stromu
- Stromy jsou
  - Neorientované
  - Orientované
  - Kořenové
    - Má kořen, vnitřní uzly, listy (nemají potomky)

Neorientované stromy



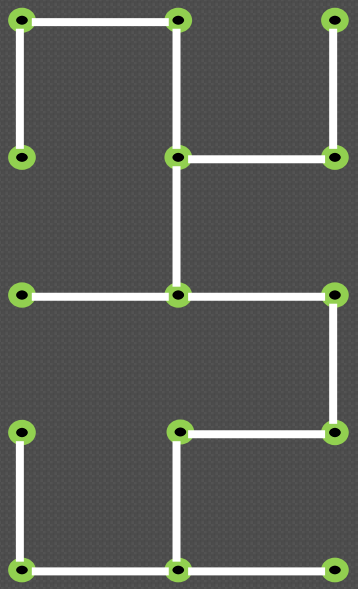
stupeň 3
hloubka 4



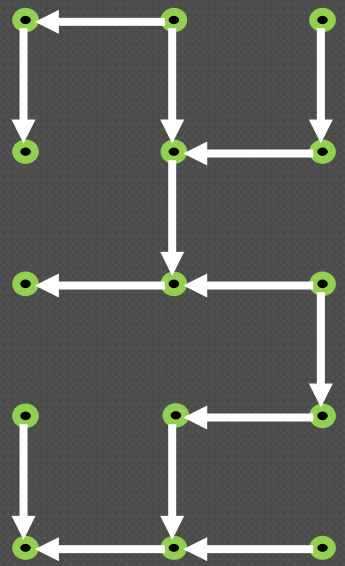
stupeň 3 (ternární strom)

hloubka 2

Neorientované a orientované stromy

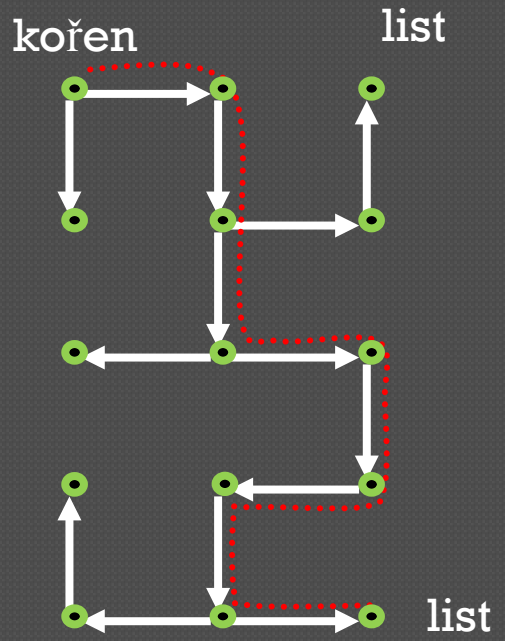


strom



orientovaný strom

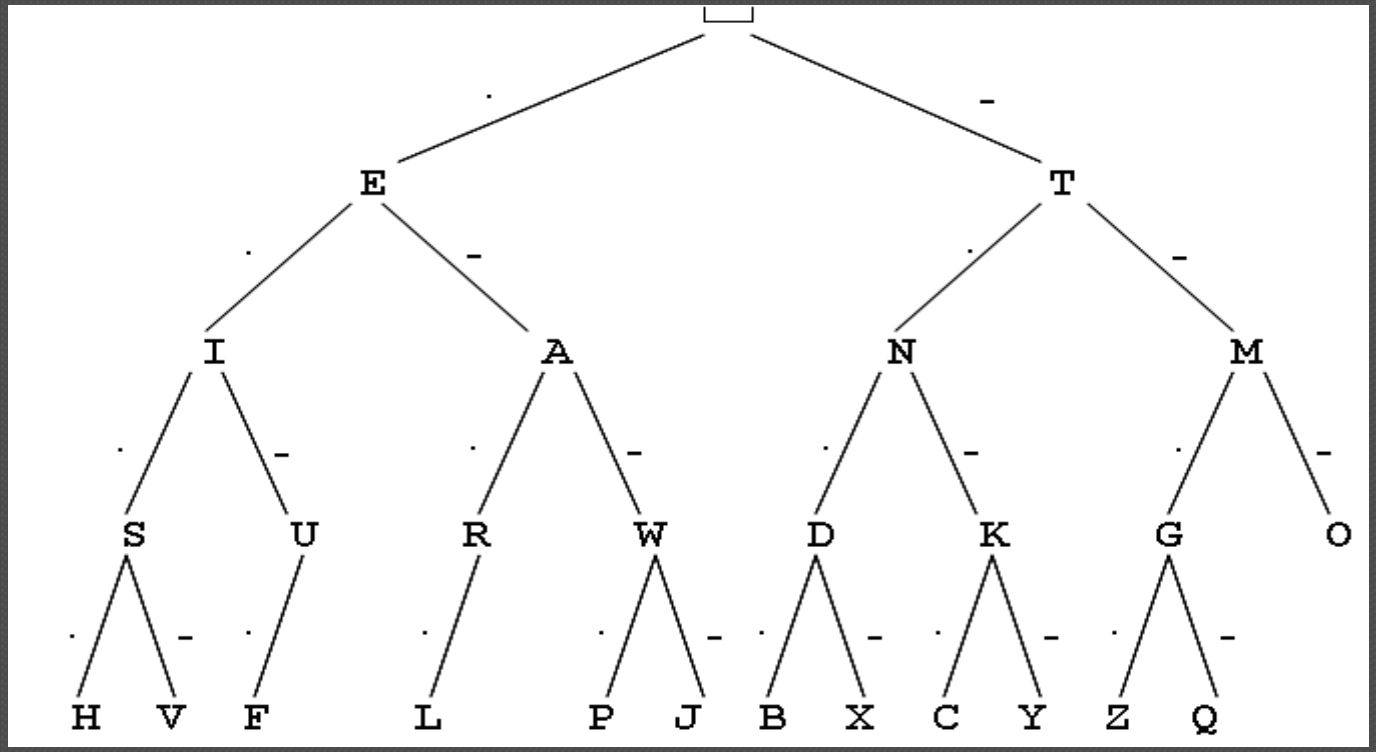
Kořen – nemá rodiče
vstupní stupeň kořene
$d^-(\text{kořen}) = 0$



kořenový strom

List – nemá potomky
výstupní stupeň listu
$d^+(\text{list}) = 0$

Rozpoznávání pomocí stromu Morseova abeceda



Výhoda: počet kroků rozpoznání znaku je roven délce znaku v morseově abecedě (např. od metody prohledávání seznamu, tabulky)

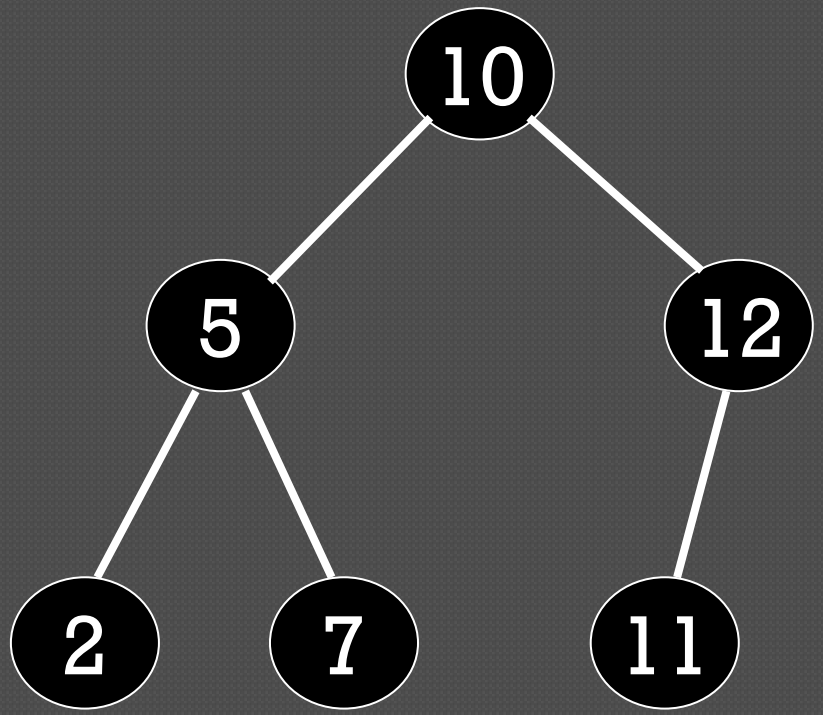
Druhy stromů

- Uspořádaný
  - následníci (podstromy) každého uzlu jsou pevným způsobem uspořádáni (první, druhý, …)
- Pravidelný stupně *k*
  - Každý uzel má právě *k* potomků nebo žádného
- Úplný pravidelný stupně *k*
  - pravidelný strom stupně *k*, kde všechny listy jsou ve stejné hloubce od kořene
- Binární
  - Každý uzel má maximálně dva potomky

K čemu jsou stromy dobré?

- Rozpoznávání
  - Viz Morseova abeceda
- Klasifikace
  - při diskrétní proměnné, viz dále
- Predikce
  - při spojité proměnné
- Různé vědecké a průmyslové aplikace, programování (efektivní hledání), počítačová grafika

Programování
Hledání v uspořádaném stromě



- Hledáme, je-li ve stromě číslo:

A) 7

B) 13

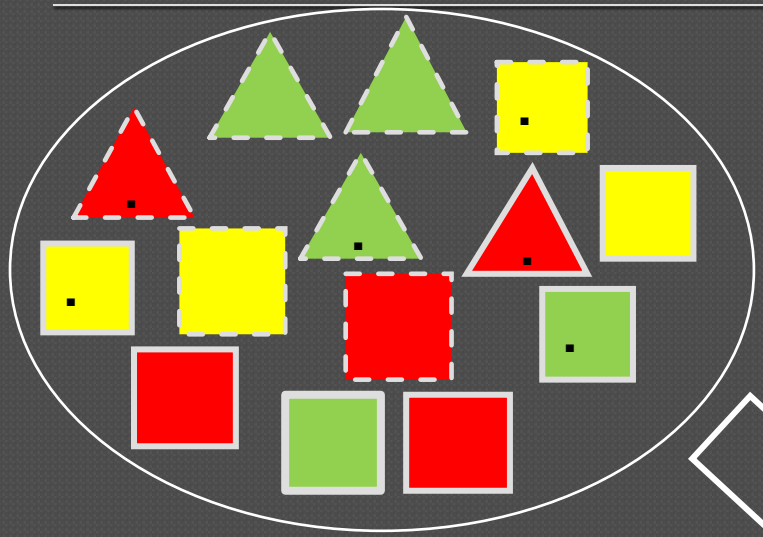
Stromy v rozhodování

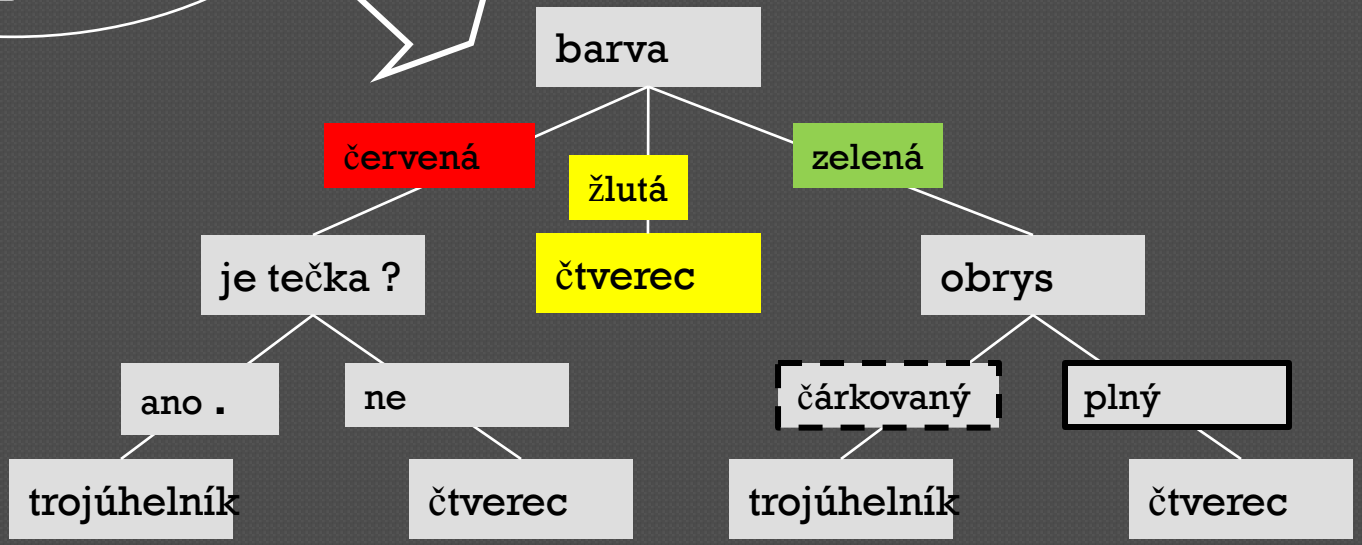
- Vnitřní vrcholy
  - reprezentují testy proměnných
- Větve (hrany)
  - reprezentují výsledky testů
- List
  - reprezentuje třídu, tj. výsledek rozpoznání

Vlastnosti

- Schopnost zpracovávat málo i veloobjemová data (škálovatelnost)
- Srozumitelnost
- Jednoduchost
- Přesnost závisí na datech nastavení stromu
- Schopnost zpracovávat data s chybějícími hodnotami nebo zašumělá data

Příklad rozhodovacího stromu





O Co jde?

- Úkol: zkonstruovat nějakým algoritmem rozhodovací strom
- Jak:
  - Vytvořit strom, aby rozpoznávání bylo co nejrychleší
  - strom co nejmenší -> Vybírat takové "nejlepší" atributy (při rozhodování), které **nesou největší množství informace**
- Jak měřit množství informace:
  - Např. pomocí entropie, informačního zisku

Používané algoritmy

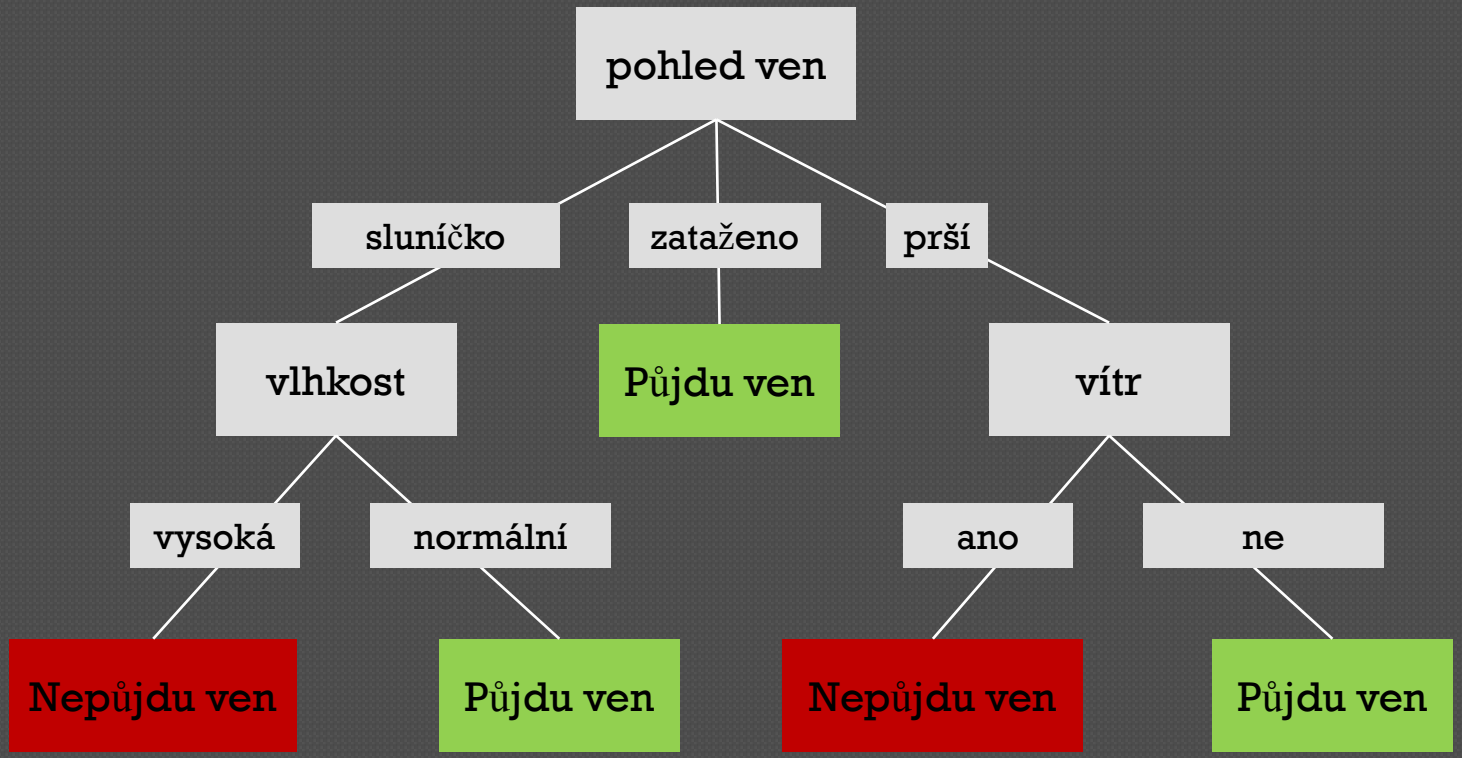
- ID3 nebo TDIDT (Top-Down Induction of Decision Trees)
- CHAID (Chi square Automatic Interaction Detector)
- C&RT (Classifikation and Regression Tree)
- QUEST (Quick Unbiased Efficient Statistical Test)
- C4.5 (i spojité proměnné)
- C5.0 (komerční)

ID3

- **Algoritmus ID3:** Dáno: S … množina klasifikovaných příkladů
- **1. Nalezneme "nejlepší" atribut, x_j (atribut, který nese největší množství informace)**
- 2. Rozdělíme množinu S na podmnožiny $S_1, S_2, \ldots, S_n$, kterých je tolik, kolik je hodnot daného atributu. Dělení probíhá tak, že do podmnožiny S_i patří právě ty příklady, jejichž hodnota atributu x_j je v_i. Takto vzniklými množinami jsou ohodnoceny nové uzly vytvářeného rozhodovacího stromu.
- 3. Pro každý uzel ohodnocený podmnožinu S_i platí: Jestliže všechny příklady v S_i patří do téže klasifikace (všechny jsou pozitivní nebo všechny jsou negativní), pak uzel ohodnocený S_i je prohlášen za list vytvářeného rozhodovacího stromu (a tedy se už dále nevětví), jinak jdi na bod 1 s tím, že $S = S_i$.

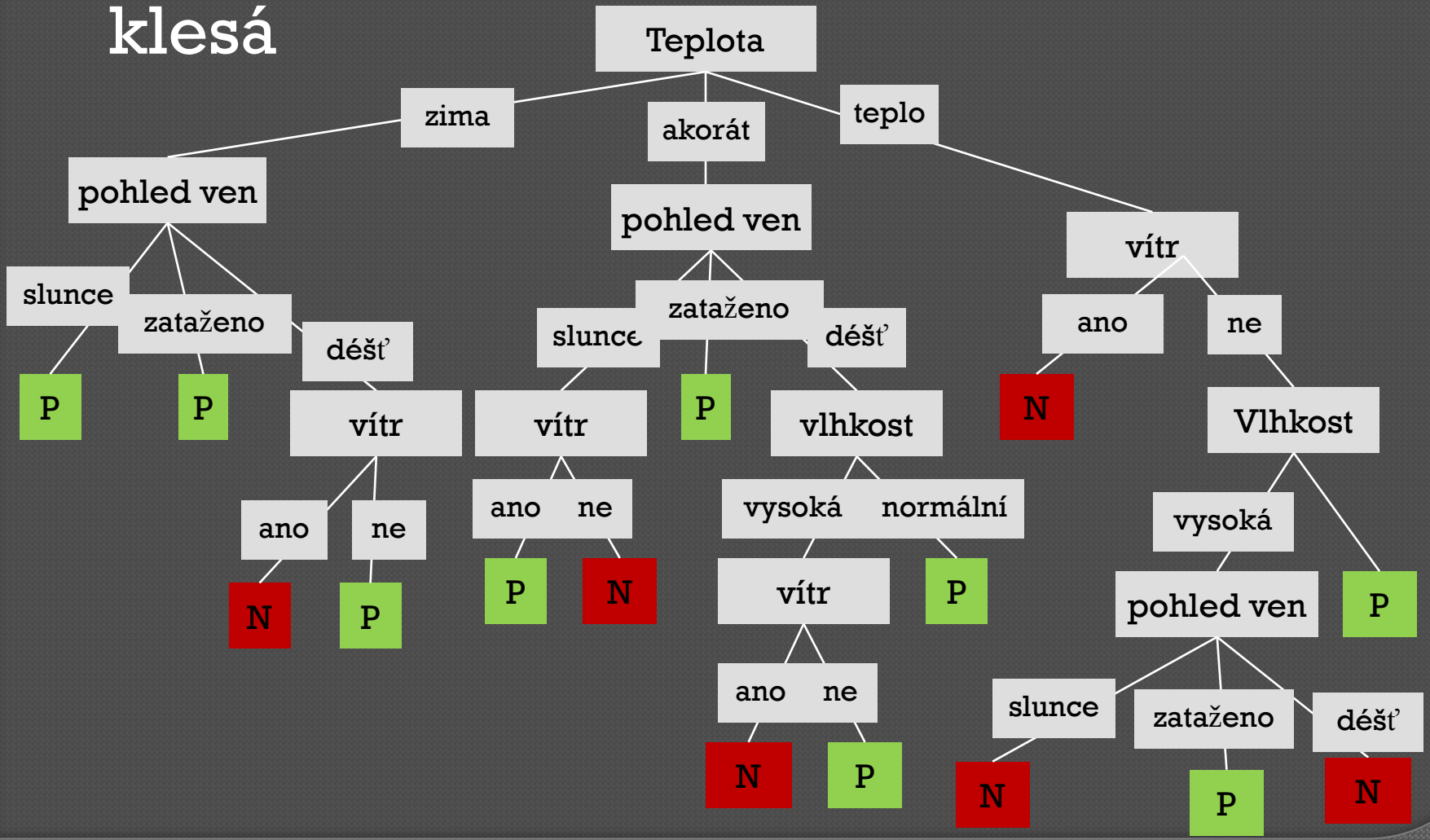
Jednoduchý strom

- Informační zisk nízký -> míra entropie vysoká



Složitější strom

- Informační zisk vyšší -> míra entropie klesá

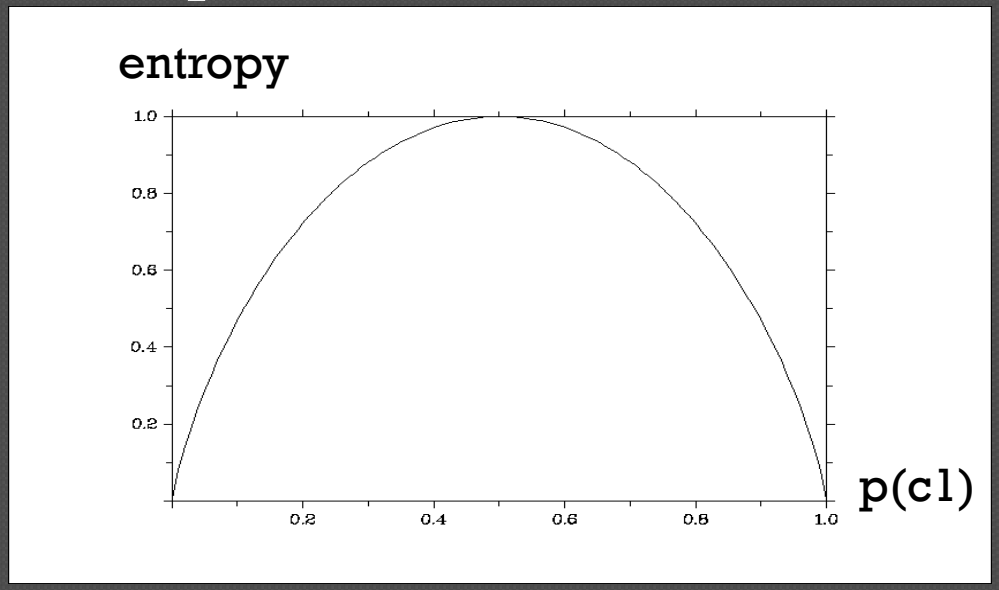


Entropie

- Průměrné množství informací, které potřebuji abych klasifikovala objekt udává měřítko entropie:

$$I = -\sum_{c} p(c) \log_2 p(c)$$

- For a two-class problem:



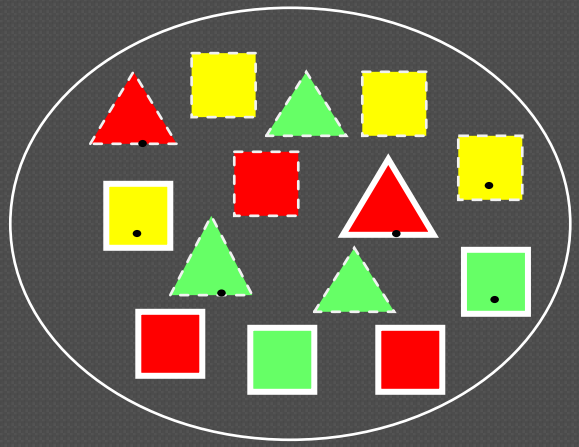
Informační zisk

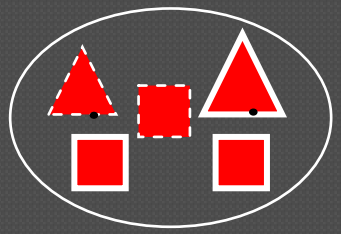
$$I(s_1, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

5 trojúhelníků, 9 čtverců

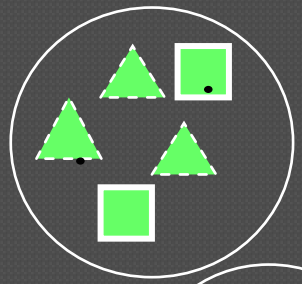
4 žluté, 5 červených, 5 zelených

$$H(tvar) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.9410\,bits$$

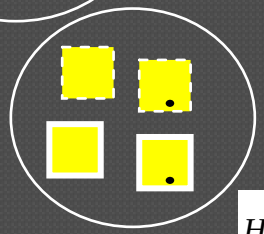




$$H(cervena) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971\,bits$$



$$H(zelena) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971\,bits$$



$$H(zluta) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0.0\,bits$$

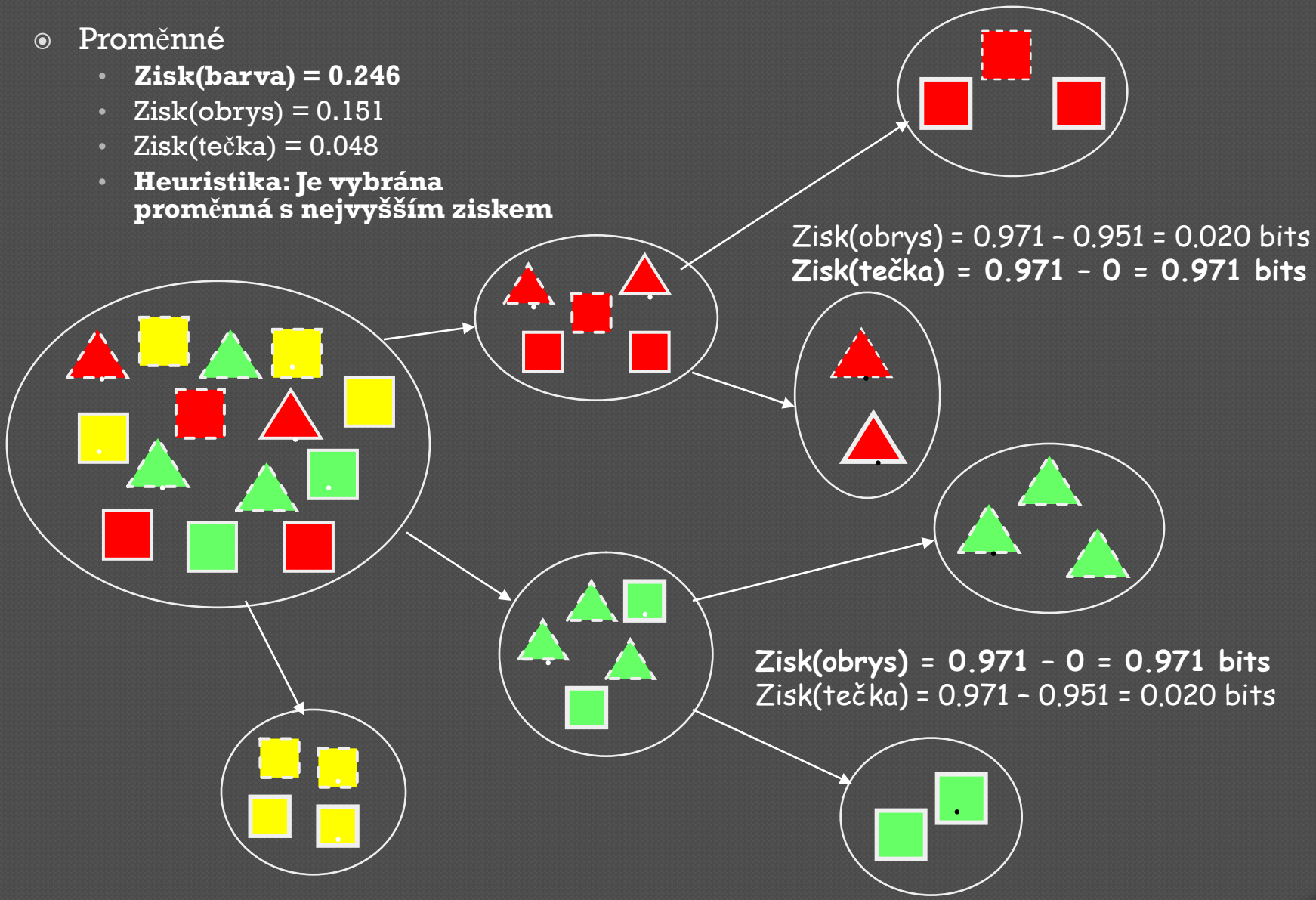
$$H_{res} = \sum_{v} p(v) \sum_{c} p(c \mid v) \log_2 p(c \mid v)$$

$$H_{res}(barva) = \sum_{barva} p(barva)\, H(barva) = \frac{5}{14}0.971 + \frac{5}{14}0.971 + \frac{4}{14}0.0 = 0.691\,bits$$

$$Zisk(barva) = H(tvar) - H_{res}(barva) = 0.940 - 0.694 = 0.246\,bits$$

Informační zisk

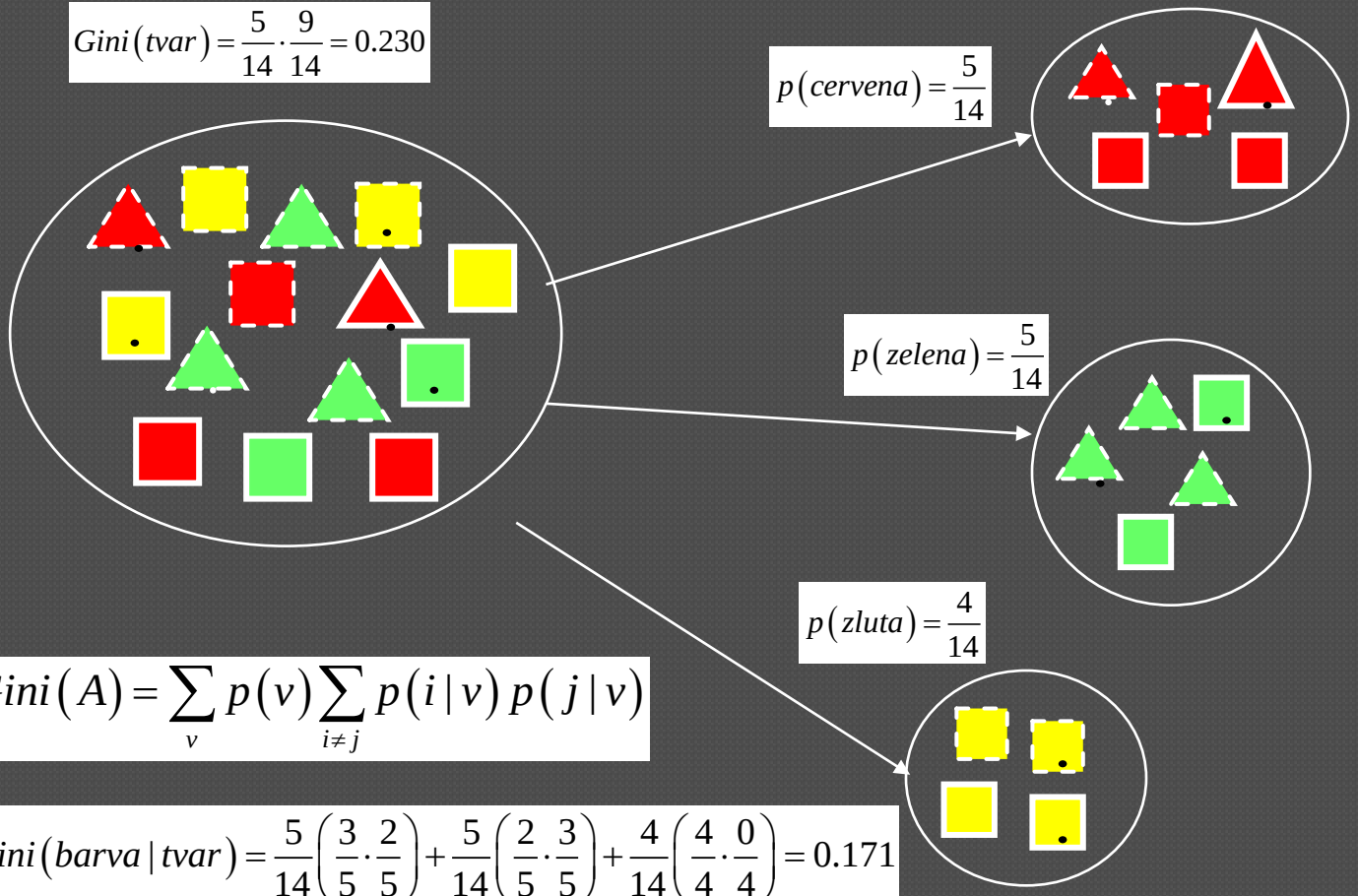
- Proměnné
  - **Zisk(barva) = 0.246**
  - Zisk(obrys) = 0.151
  - Zisk(tečka) = 0.048
  - **Heuristika: Je vybrána proměnná s nejvyšším ziskem**



Giniho index

$$Gini = \sum_{i \neq j} p(i)\,p(j)$$

$$Gini(tvar) = \frac{5}{14} \cdot \frac{9}{14} = 0.230$$



$$p(cervena) = \frac{5}{14}$$

$$p(zelena) = \frac{5}{14}$$

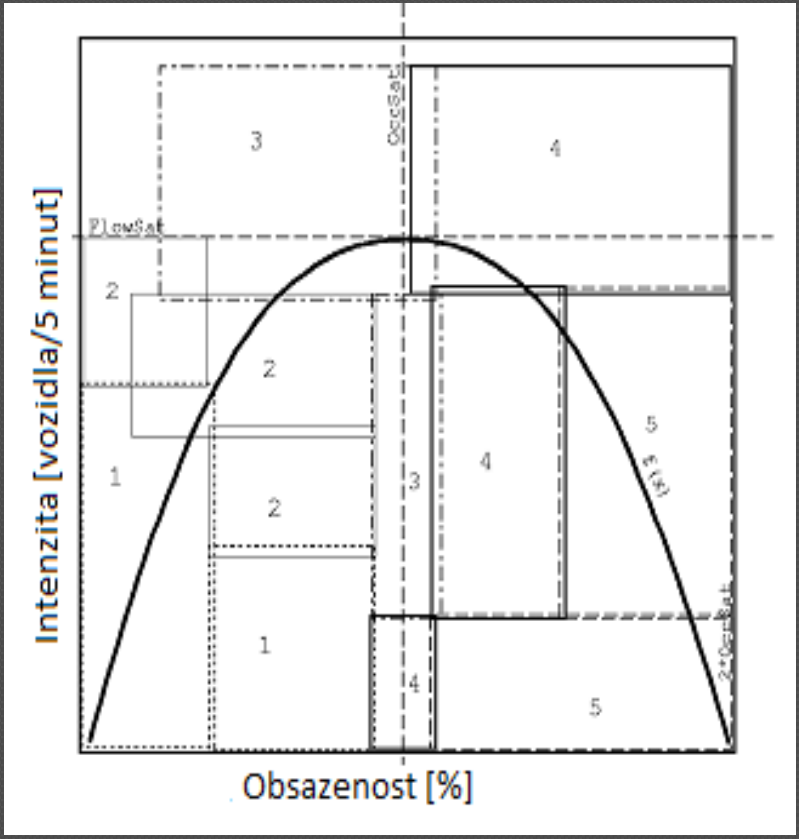
$$p(zluta) = \frac{4}{14}$$

$$Gini(A) = \sum_{v} p(v) \sum_{i \neq j} p(i \mid v)\,p(j \mid v)$$

$$Gini(barva \mid tvar) = \frac{5}{14}\left(\frac{3}{5} \cdot \frac{2}{5}\right) + \frac{5}{14}\left(\frac{2}{5} \cdot \frac{3}{5}\right) + \frac{4}{14}\left(\frac{4}{4} \cdot \frac{0}{4}\right) = 0.171$$

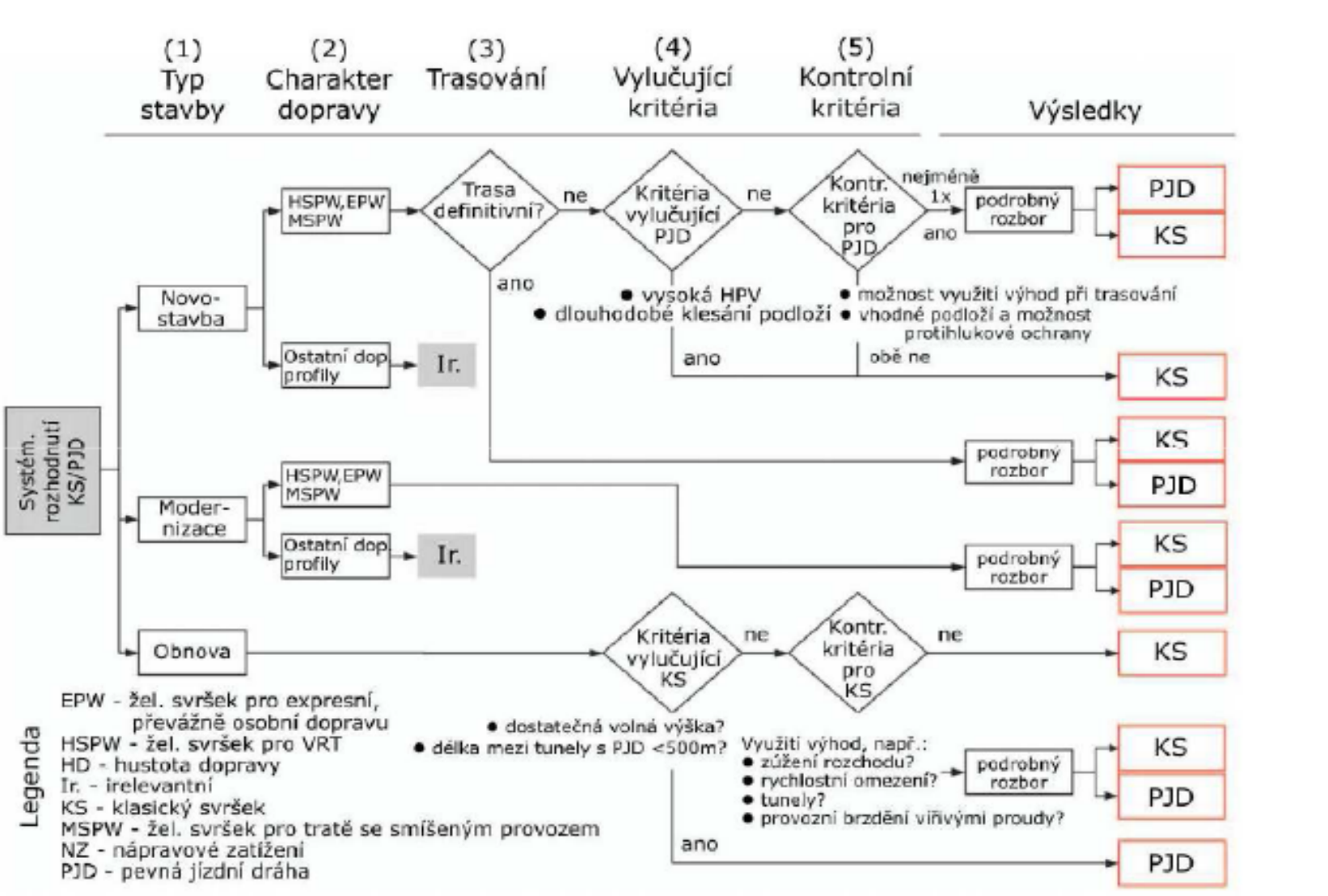
$$GainGini(barva) = Gini(tvar) - Gini(barva \mid tvar) = 0.230 - 0.171 = 0.058$$

Použití v dopravě



- Viz. excel (vytisknuté)

Použití v dopravě



Zdroje

- Blaž Zupan, Ivan Bratko - Induction of Decision Trees (uisp02-ecTrees.ppt)
- RNDr. Miroslav Pavelka, Ph.D. – Rozhodovací stromy (Stromy_prezentace1.ppt)
- Ing. Vít Fábera, Ph.D. - Stromy a kostry (XTI-Prednaska-06.ppt)
- www.fce.vutbr.cz/zel/plasek.o/studium/12_PJD_rozhodovaci_proces.pdf
- www.wikipedia.org

Děkujeme

- Ing. Vítu Fáberovi Ph.D. za poskytnutí materiálů a pomoc
- Ing. Ondřeji Přibylovi Ph.D. za pomoc a ochotné konzultace
- Vám za pozornost