Internet a klasifikační metody, 6. přednáška

Tým zvládne víc než jedinec

volitelný předmět pro magisterské studium

Martin Holeňa

O čem to bude?

- Spojování více klasifikátorů do týmu
  - zahrnutí různé důvěry různým členům týmu
  - týmy klasifikátorů různých a stejných typů
  - hlavní metody spojování: bagging, boosting, stacking
- Hierarchické uspořádání týmů klasifikátorů
- Týmy klasifikačních stromů − klasifikační náhodné lesy

Spojování klasifikátorů do týmu

- *Motivace*: máme-li tým klasifikátorů, nepotřebujeme

  aby každý klasifikoval správně jakýkoliv vstup

  - *možnost specializace* na část prostoru příznaků X
  - + nutnost odhadnout důvěry jednotlivým členů týmu

- Pro r-členný tým: $F: X \to \{C_1, \ldots, C_m\}, F(x) = A(F_1(x), \ldots, F_r(x), \tau_1(x), \ldots, \tau_r(x))$

  - A – *agregační funkce*, τ_i – *důvěra* (trust) klasifikátoru F_i, $\tau_i: X \to [0,1]$

Jak měřit důvěru klasifikátoru

1. *Globální* důvěra klasifikátoru F – nezávisí na $x, (\forall x \in X)\tau(x) = \tau_F \in [0,1]$
   - příklad $\tau_F = \frac{1}{|V|}\sum_{v \in V} \| F(v) = C_v \|$, kde $v \in C_v, V$ – daná validační množina
2. *Lokální* důvěra – závisí na x, nejznámější příklad:

   Euclidean Local Accuracy – používá eukleidovskou vzdálenost d_E

   a k nejbližších sousedů z $V, \tau^{ELA}(x) = \frac{1}{k}\sum_{i=1}^{k} \left\| F\left(v^{(x,i)}\right) = C_{v^{(x,i)}} \right\|$
   - $v^{(x,1)}, \ldots, v^{(x,|V|)}$ – uspořádání prvků V dle vzdálenosti $d_E(v, x)$ od x

Týmy a soubory klasifikátorů

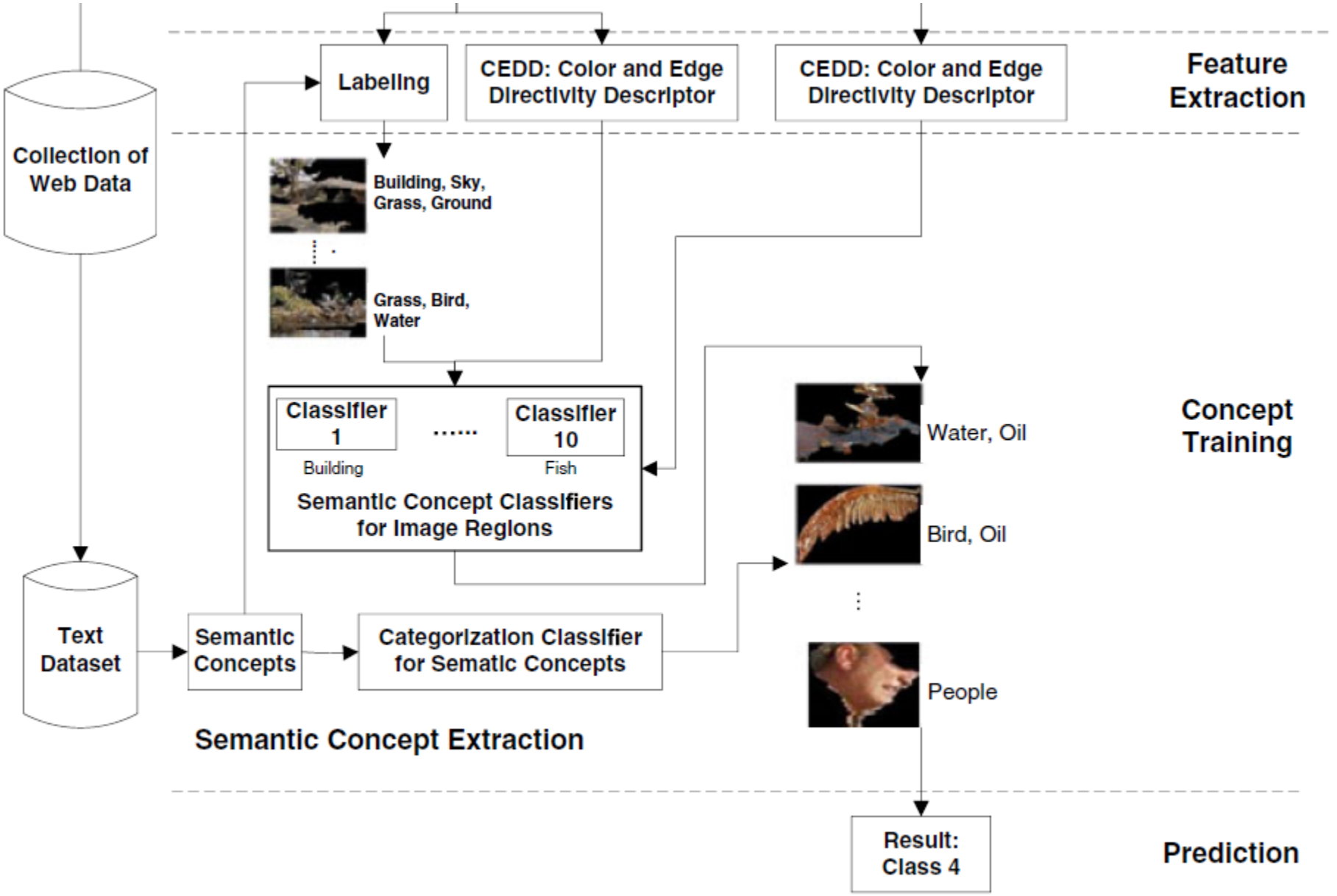
- *Jak* vnést mezi členy týmu *nepodobnost*?
  - proč potřeba? jinak tým = kterýkoliv člen

1. Klasifikátory *různých druhů* (k-NN, Bayesův, SVM, …)
2. Klasifikátory 1 druhu s různými hodnotami parametrů
   - k-NN: k, SVM: c, q v $\kappa(x, y) = (x^{\mathsf{T}} y + c)^q$, MLP: počet neuronů
   - takový tým nazýváme soubor (*ensemble*) *klasifikátorů*

Týmová klasifikace multimedií

- K plnému využití informací klasifikujeme každé

  použité medium: obrázky, text, videa, zvuky, …


Collection of Web Data
Labeling
CEDD: Color and Edge Directivity Descriptor
CEDD: Color and Edge Directivity Descriptor
Feature Extraction
Building, Sky, Grass, Ground
Grass, Bird, Water
Classifier 1
Building
Classifier 10
Fish
Semantic Concept Classifiers for Image Regions
Concept Training
Water, Oil
Bird, Oil
People
Text Dataset
Semantic Concepts
Categorization Classifier for Sematic Concepts
Semantic Concept Extraction
Result: Class 4
Prediction

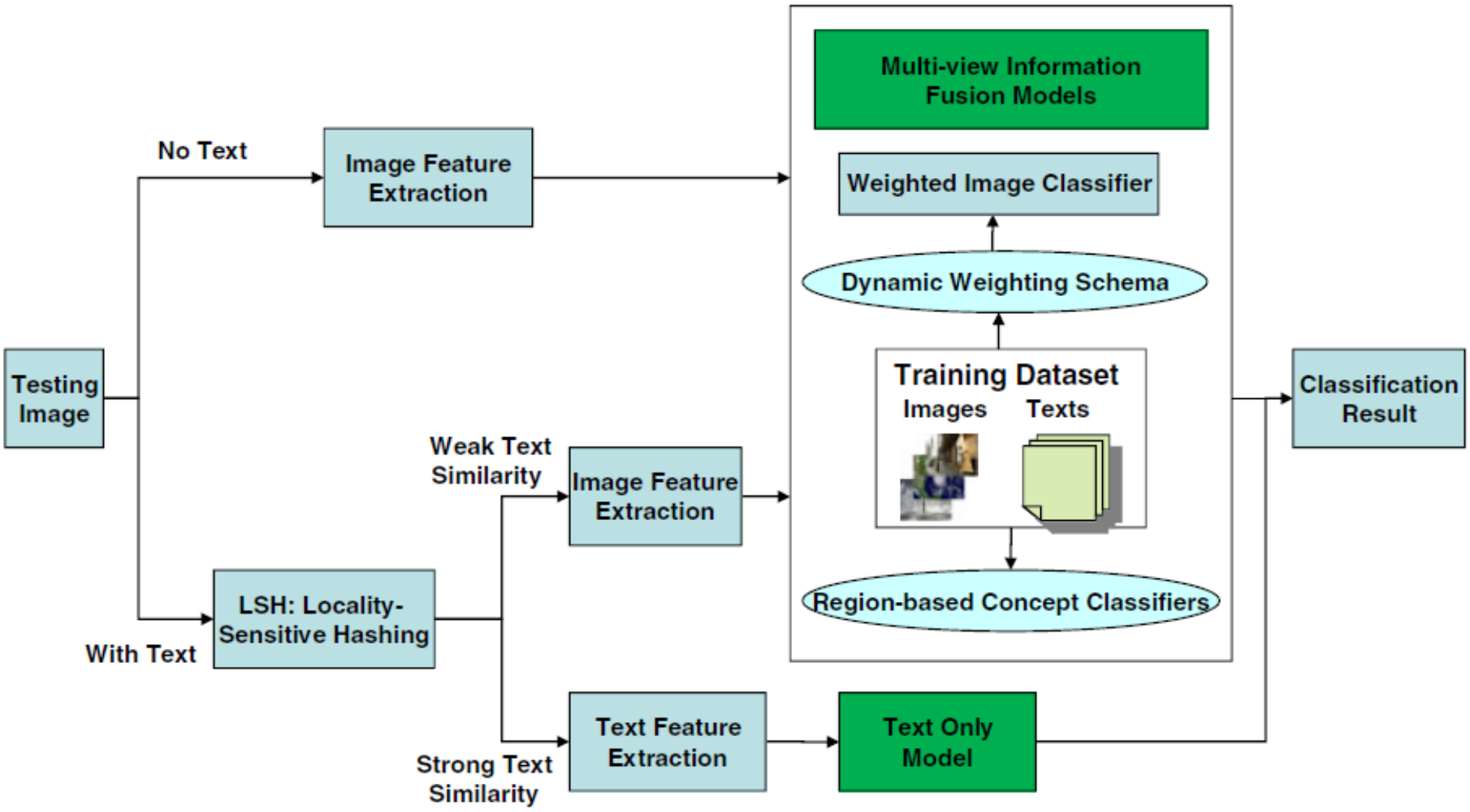
Týmová klasifikace multimedií

- K plnému využití informací klasifikujeme každé

  použité medium: obrázky, text, videa, zvuky, …

- Různá media nejlépe klasifikována různými typy

  klasifikátorů → potřebujeme tým *klasifikátorů různých typů*


Multi-view Information Fusion Models
Weighted Image Classifier
Dynamic Weighting Schema
Training Dataset
Images
Texts
Region-based Concept Classifiers
Testing Image
No Text
Image Feature Extraction
With Text
LSH: Locality-Sensitive Hashing
Weak Text Similarity
Image Feature Extraction
Strong Text Similarity
Text Feature Extraction
Text Only Model
Classification Result

Týmová klasifikace multimedií

- K plnému využití informací klasifikujeme každé

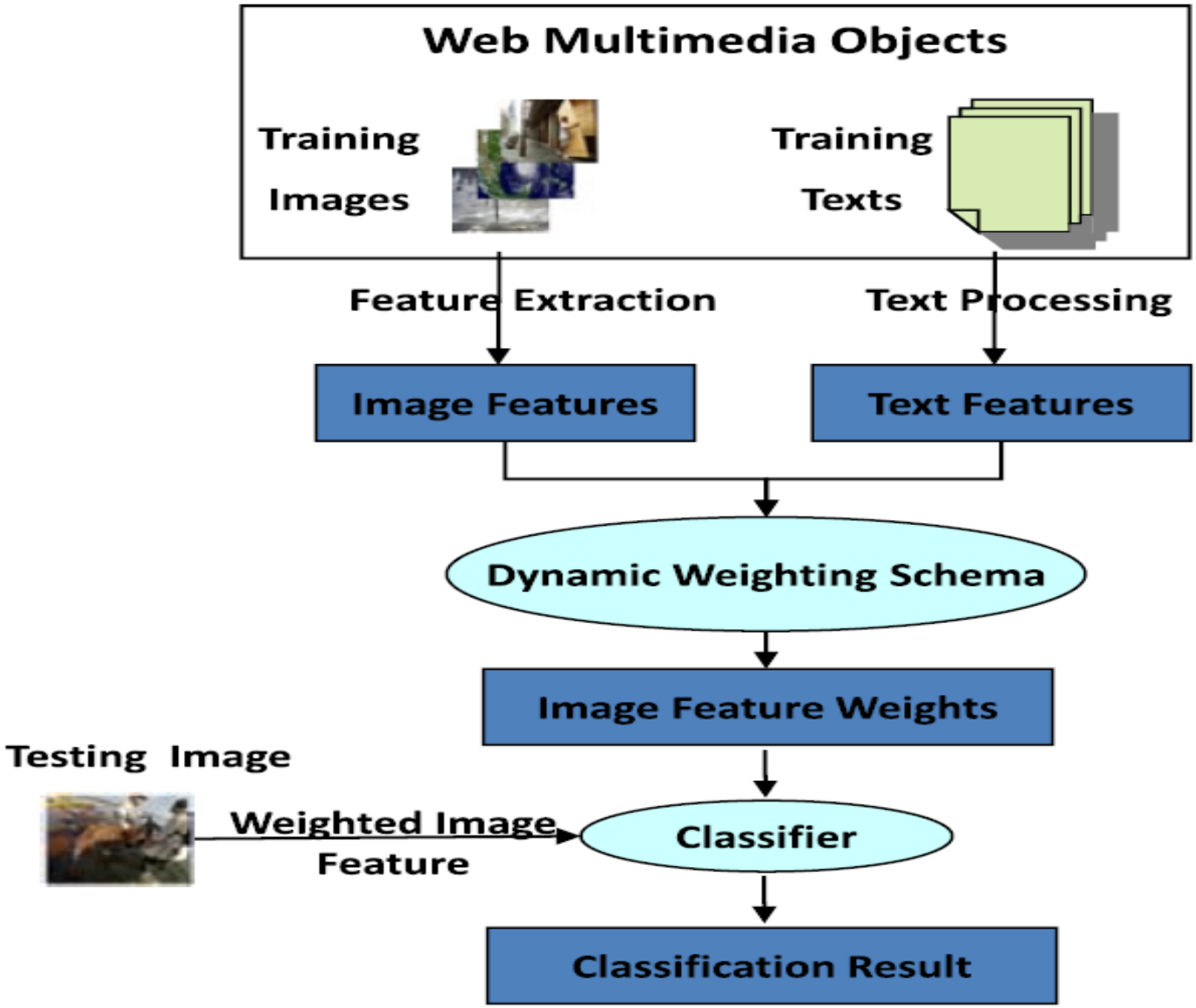
  použité medium: obrázky, text, videa, zvuky, …

- Různá media nejlépe klasifikována různými typy

  klasifikátorů → potřebujeme tým *klasifikátorů různých typů*

- Agregační funkce: klasifikátory odpovídající jednotlivým mediím

  dynamicky vážíme důvěrou (dynamicky – lokální důvěrou)


Web Multimedia Objects
Training Images
Training Texts
Feature Extraction
Text Processing
Image Features
Text Features
Dynamic Weighting Schema
Image Feature Weights
Testing Image
Weighted Image Feature
Classifier
Classification Result

Jak tým klasifikátorů vytvořit?

- *Různodruhové týmy*: nemnoho členů (nejčastěji jednotky)
  - každý člen naučen nezávisle, specifickou metodou
- *Soubory*: desítky − tisíce členů, principiálně vytvářeny

učením různými podmnožinami dat − bagging, boosting

Bagging

- ***B**ootstrap **agg**regati**ng***: agregované $F_k, k = 1, \ldots, r$ trénovány na $\{x_1^{(k)}, \ldots, x_n^{(k)}\}$ získaných bootstrapem = *resamplováním s vracením*
  - bootstrap z n dat: počet různých $\rightarrow n \lim_{n\to\infty} P\left((\exists j) x_j^{(k)} = x_i\right) = n\left(1 - \frac{1}{e}\right)$
- *Validace* pro F_k: daty $x \notin \{x_1^{(k)}, \ldots, x_n^{(k)}\}$ („out of bag“)
  - baggingová alternativa křížové validace: $q \geq 2$ částí (q-fold),

    každá validuje klasifikátor naučený $q - 1$ zbývajícími, + zprůměrování

Boosting

- Členové učeni všemi učicími daty U, *vliv* $u \in U$

  vážen neúspěšností klasifikace u některými jinými členy

  - algoritmy: AdaBoost (nejužívanější), LPBoost, LogitBoost, BrownBoost

- *AdaBoost* (Adaptive Boosting): klasifikátory $F_i: X \to \{1,-1\}, i = 1,\ldots,r$

  - váhy: $(\forall u \in U) w_1(u) = \frac{1}{|U|}, w_{i+1}(u) = w_i(u) e^{\tau_i (2\|F_i(u) \neq C_u\| - 1)}$

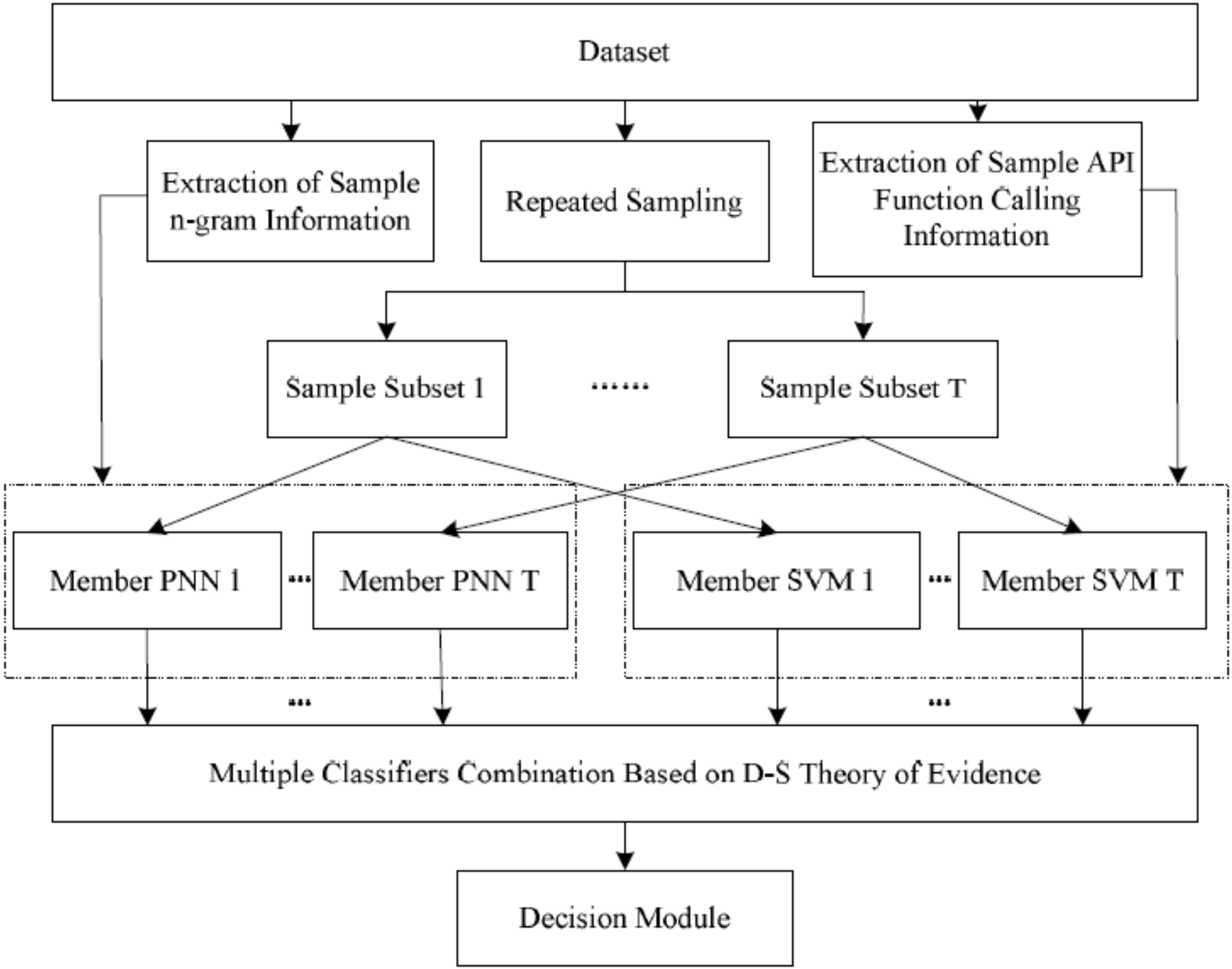
  - agregace: $A(F_1,\ldots,F_r,\tau_1,\ldots,\tau_r) = \text{sign}(\sum_{i=1}^{r} \tau_i F_i), \tau_i = \frac{1}{2} \ln \frac{\sum_{u \in U} w_i(u) \|F_i(u) = C_u\|}{\sum_{u \in U} w_i(u) \|F_i(u) \neq C_u\|}$

Jak tým klasifikátorů vytvořit?

- *Různodruhové týmy*: nemnoho členů (nejčastěji jednotky)
  - každý člen naučen nezávisle, specifickou metodou
- *Soubory*: desítky − tisíce členů, principiálně vytvářeny učením různými podmnožinami dat − bagging, boosting
- *Dodatečná metoda* vytváření týmů: použít jako člen tým (např. soubor) → *hierarchické* týmy

Soubory klasifikátorů a malware

- Detekce malware na základě statických vlastností
  - klasifikace *n-gramů* souborem specifických *neuronových sítí*
- Detekce malware na základě dynamických vlastností
  - klasifikace *volání API* souborem *SVM* klasifikátorů
- Různí členové souborů trénováni různými podmnožinami dat

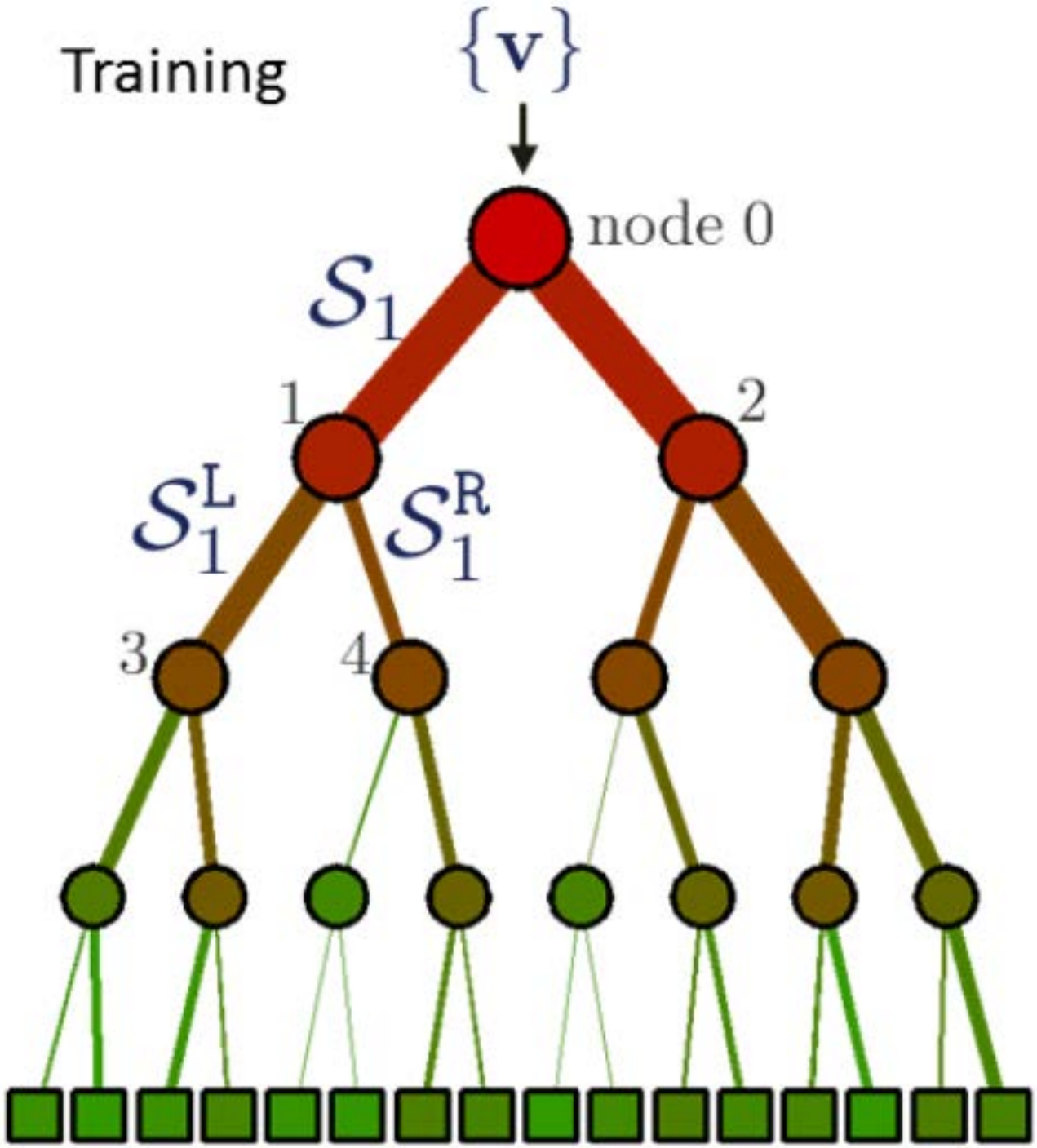

Dataset
Extraction of Sample n-gram Information
Repeated Sampling
Extraction of Sample API Function Calling Information
Sample Subset 1
.....
Sample Subset T
Member PNN 1
...
Member PNN T
Member SVM 1
...
Member SVM T
...
...
Multiple Classifiers Combination Based on D-S Theory of Evidence
Decision Module

Soubory klasifikátorů a malware

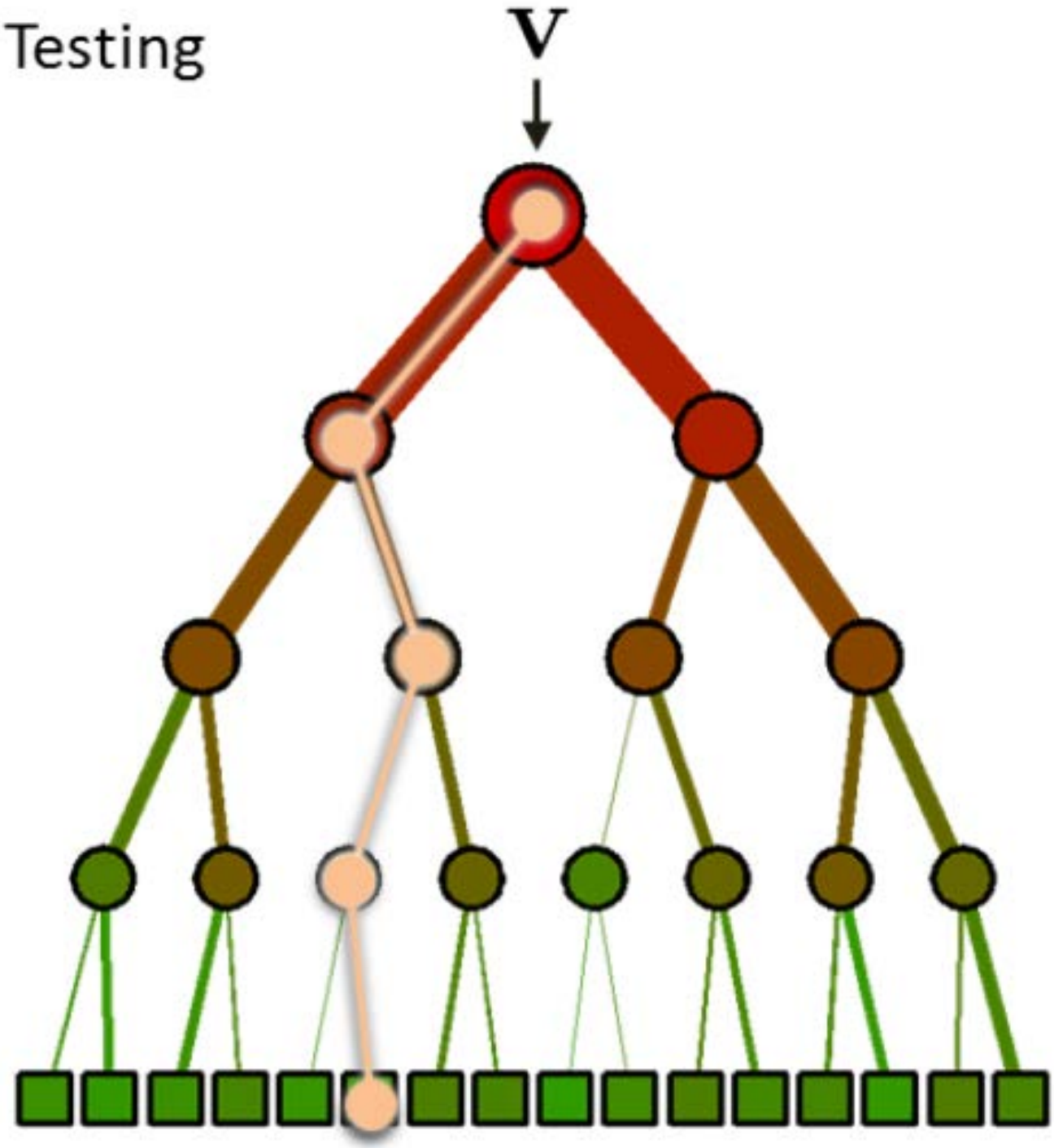
- Detekce malware na základě statických vlastností
  - klasifikace *n-gramů* souborem specifických *neuronových sítí*
- Detekce malware na základě dynamických vlastností
  - klasifikace *volání API* souborem *SVM* klasifikátorů
- Různí členové souborů trénováni různými podmnožinami dat
- Princip agregace: teorie evidence (Dempster-Shafer)

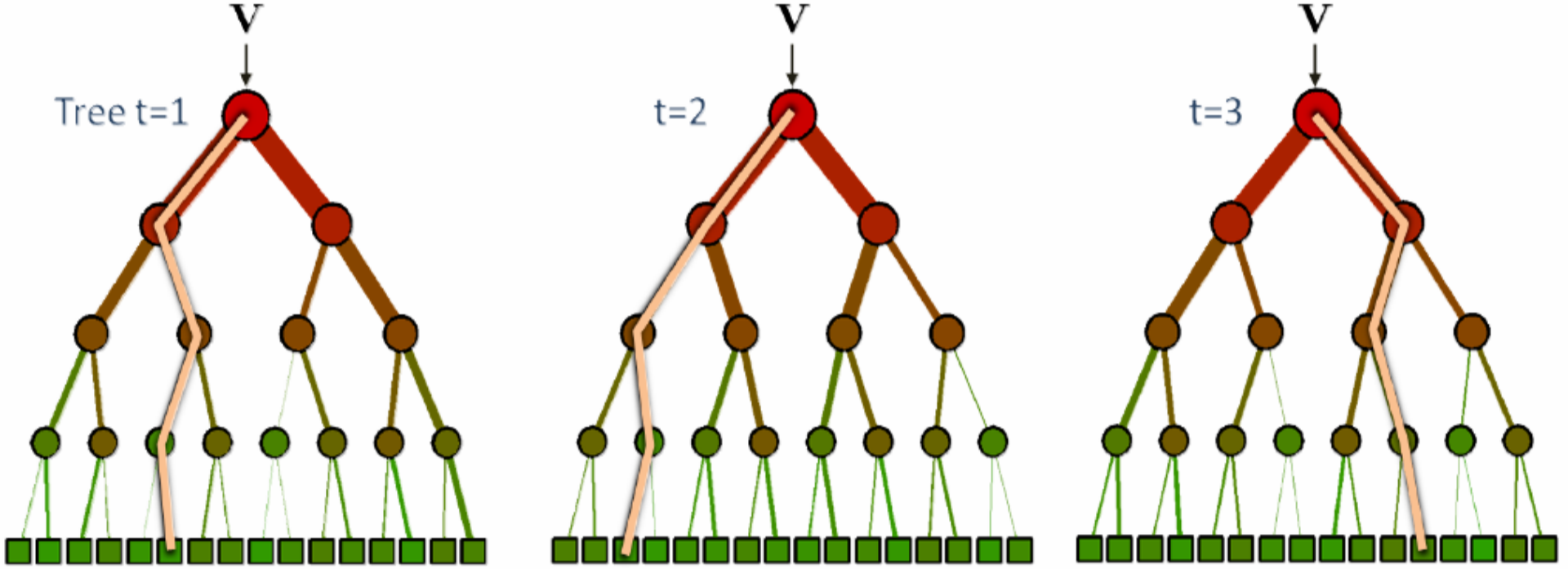
Náhodné lesy

- Náhodný les: *soubor stromů*, trénován každý zvlášť, při klasifikaci jsou *predikce sčítány*


Training
{v}
node 0
$\mathcal{S}_1$
1
2
$\mathcal{S}_1^{\mathrm{L}}$
$\mathcal{S}_1^{\mathrm{R}}$
3
4

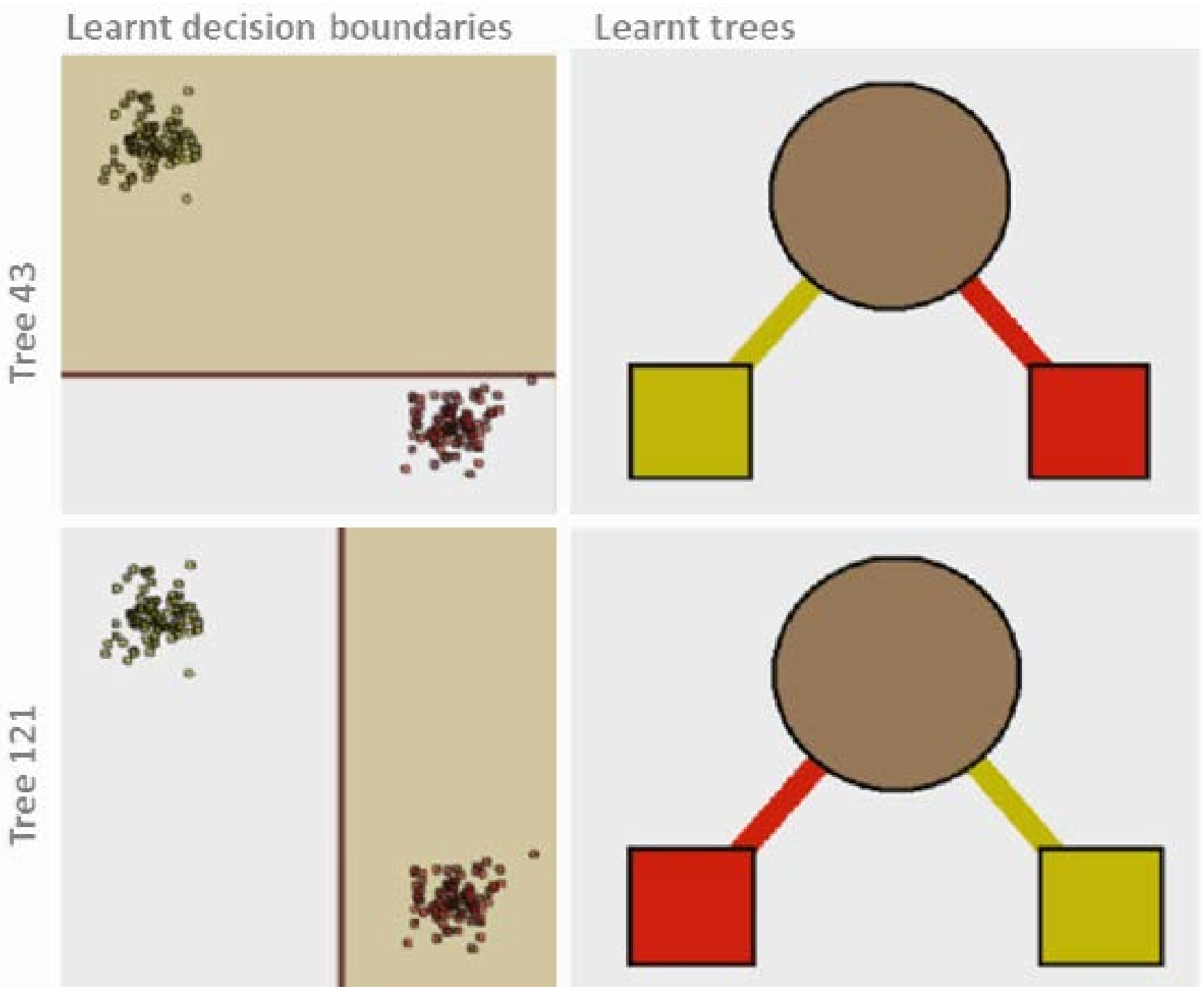
Testing




v
Tree t=1
v
t=2
v
t=3

Náhodné lesy

- Náhodný les: *soubor stromů*, trénován každý zvlášť, při klasifikaci jsou *predikce sčítány*
  - hranice: stromy – ostré, les – postupně přecházející


Learnt decision boundaries
Learnt trees
Tree 43
Tree 121

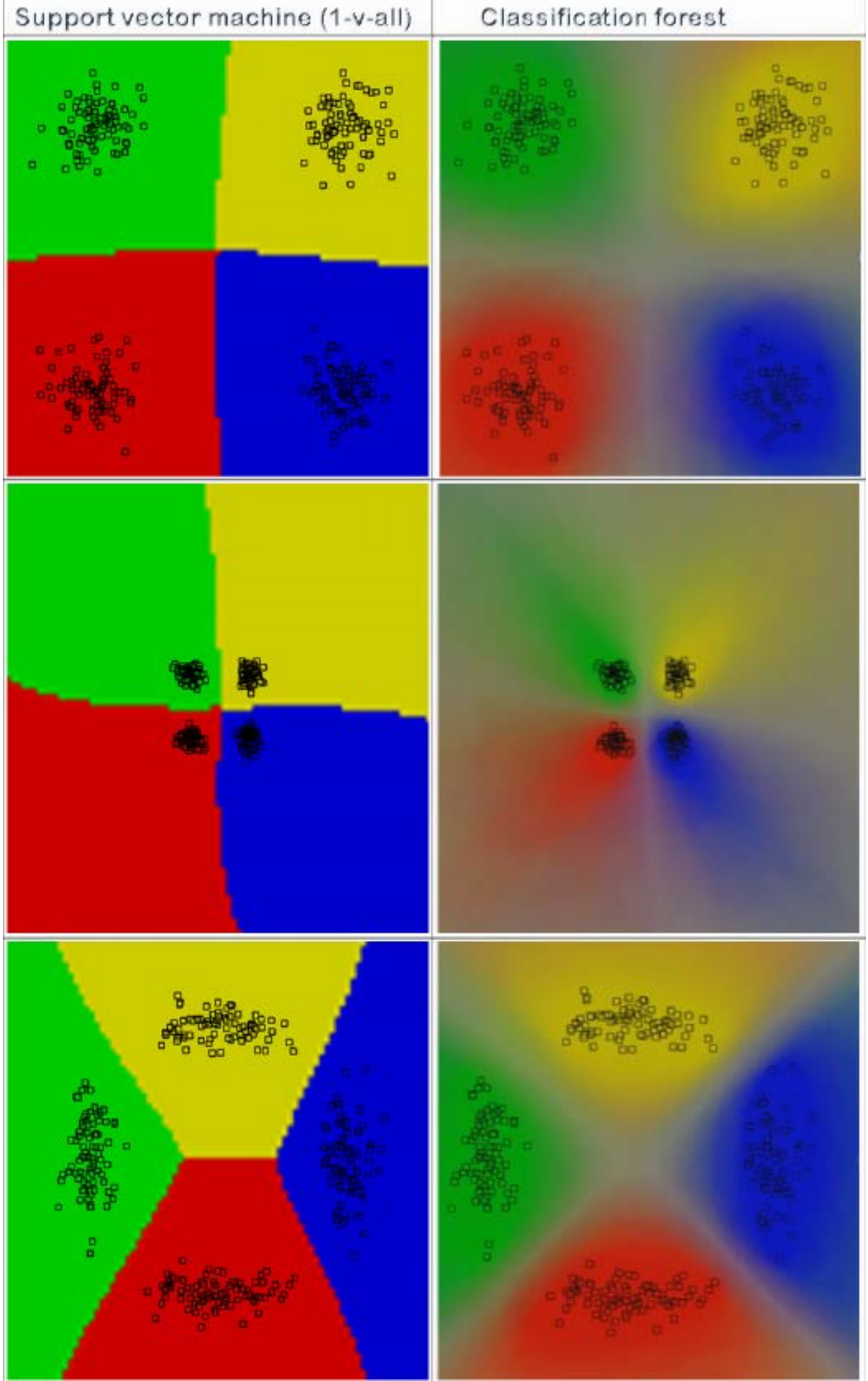
T=8
c_2

T=200
c_3

Náhodné lesy

- Náhodný les: *soubor stromů*, trénován každý zvlášť, při klasifikaci jsou *predikce sčítány*
  - hranice: stromy – ostré, les – postupně přecházející

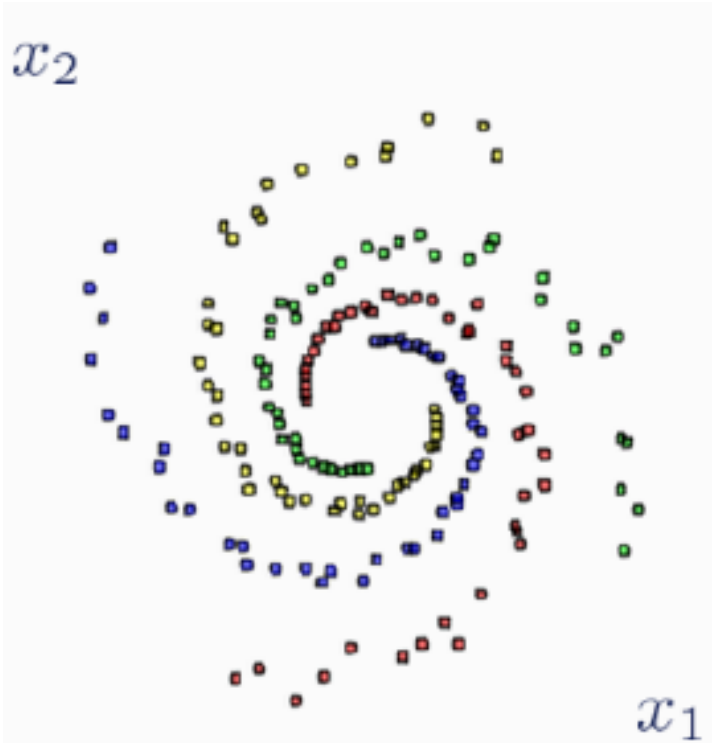
1. *výhoda*: pokud třídy ve skutečnosti fuzzy, popisují je přesněji než ostatní klasifikátory

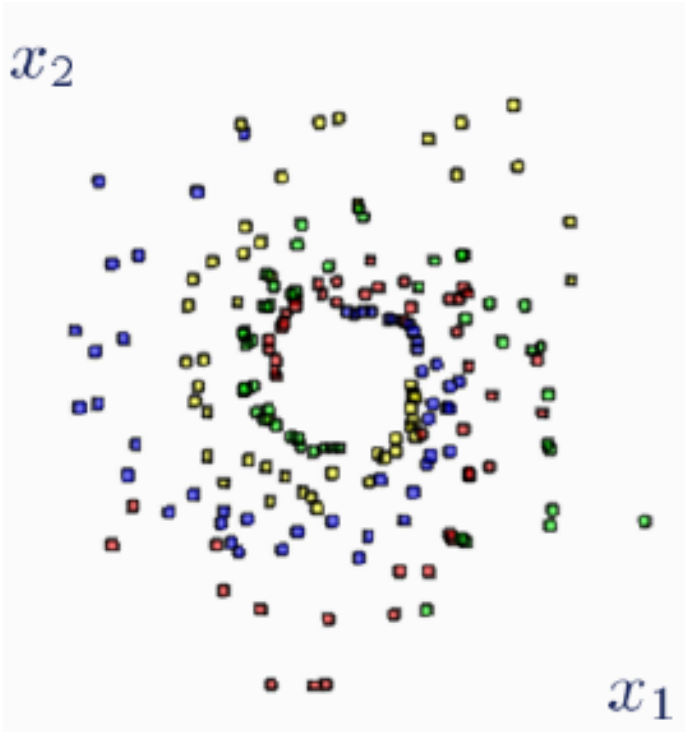

Support vector machine (1-v-all)
Classification forest

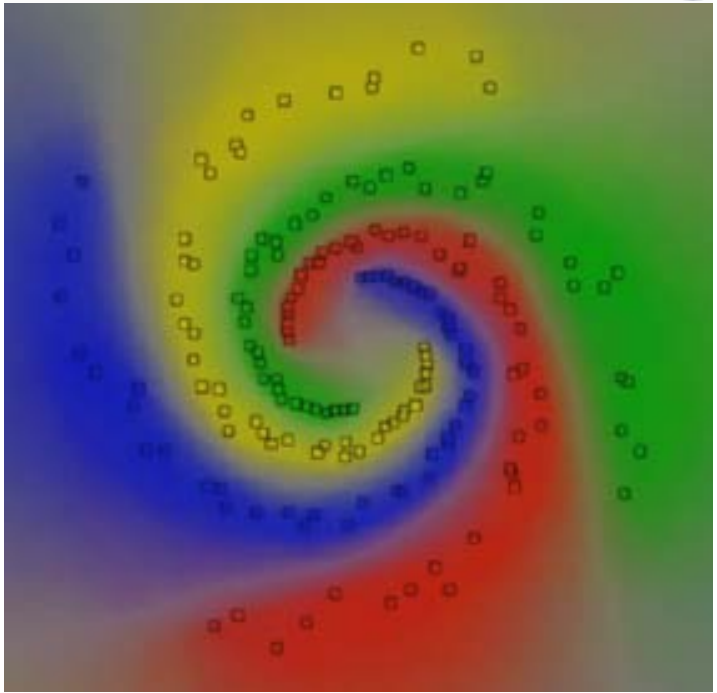
Náhodné lesy

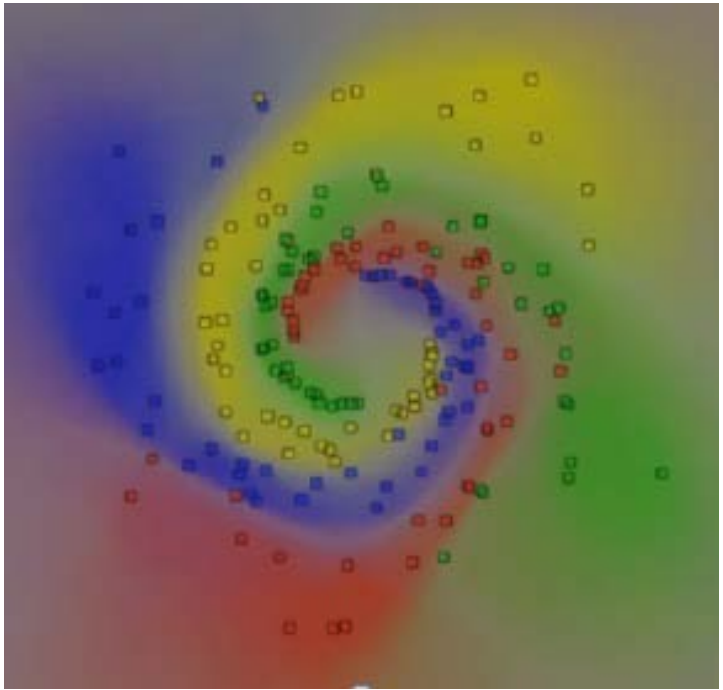
- Náhodný les: *soubor stromů*, trénován každý zvlášť, při klasifikaci jsou *predikce sčítány*
  - hranice: stromy – ostré, les – postupně přecházející

1. *výhoda*: pokud třídy ve skutečnosti fuzzy, popisují je přesněji než ostatní klasifikátory
2. *výhoda*: sčítané predikce méně ovlivněny šumem


x_2
x_1


x_2
x_1

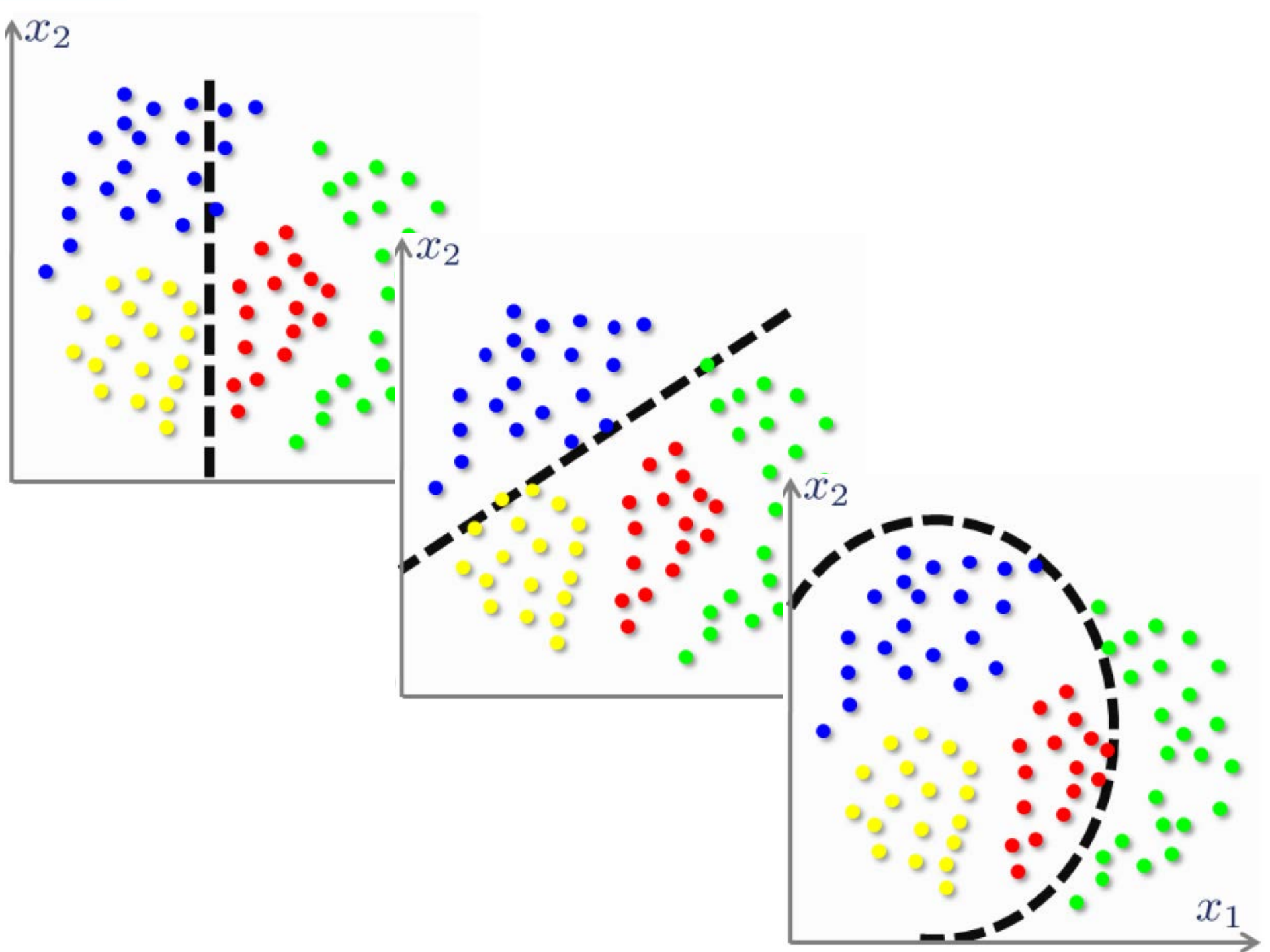


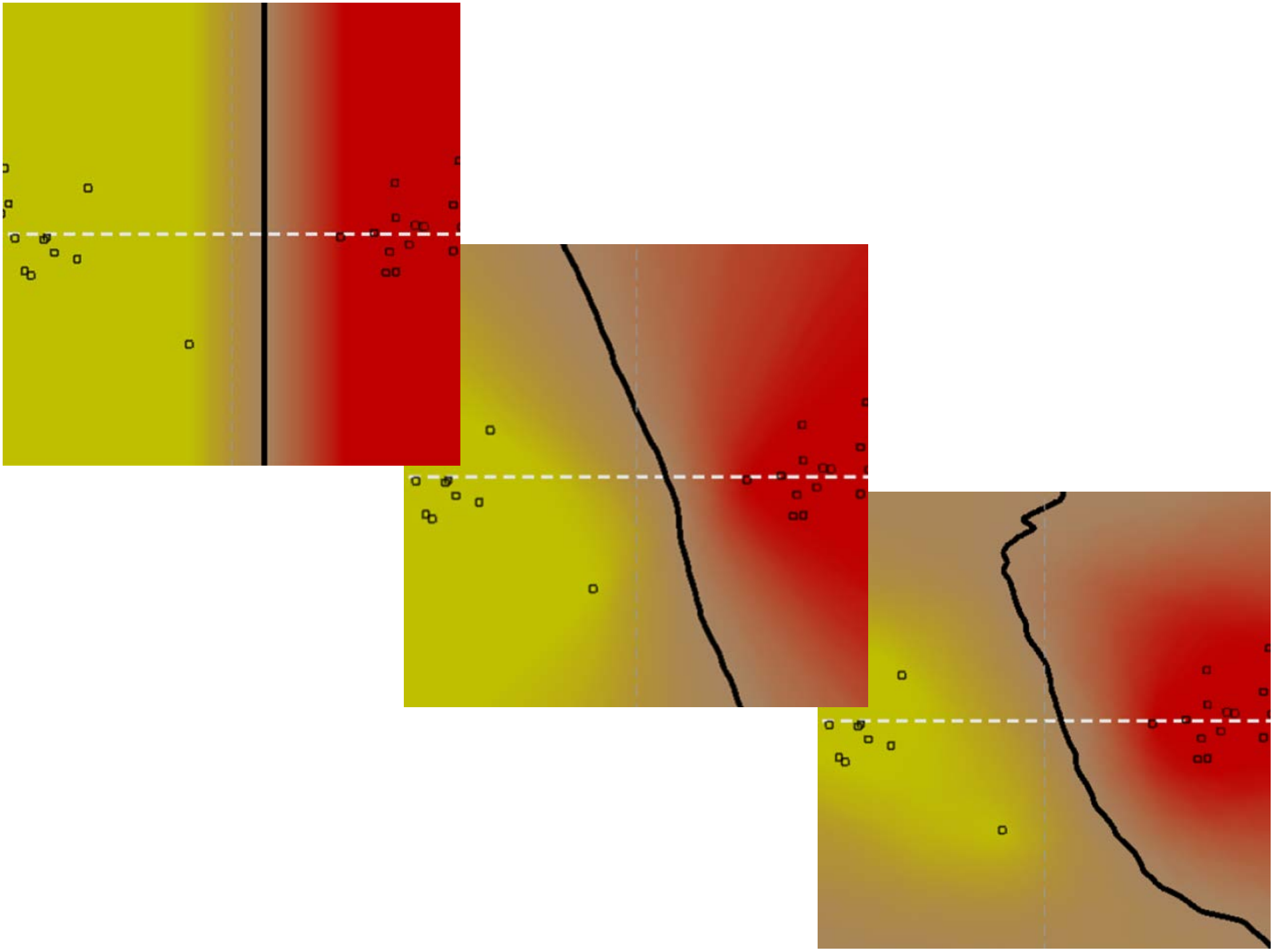


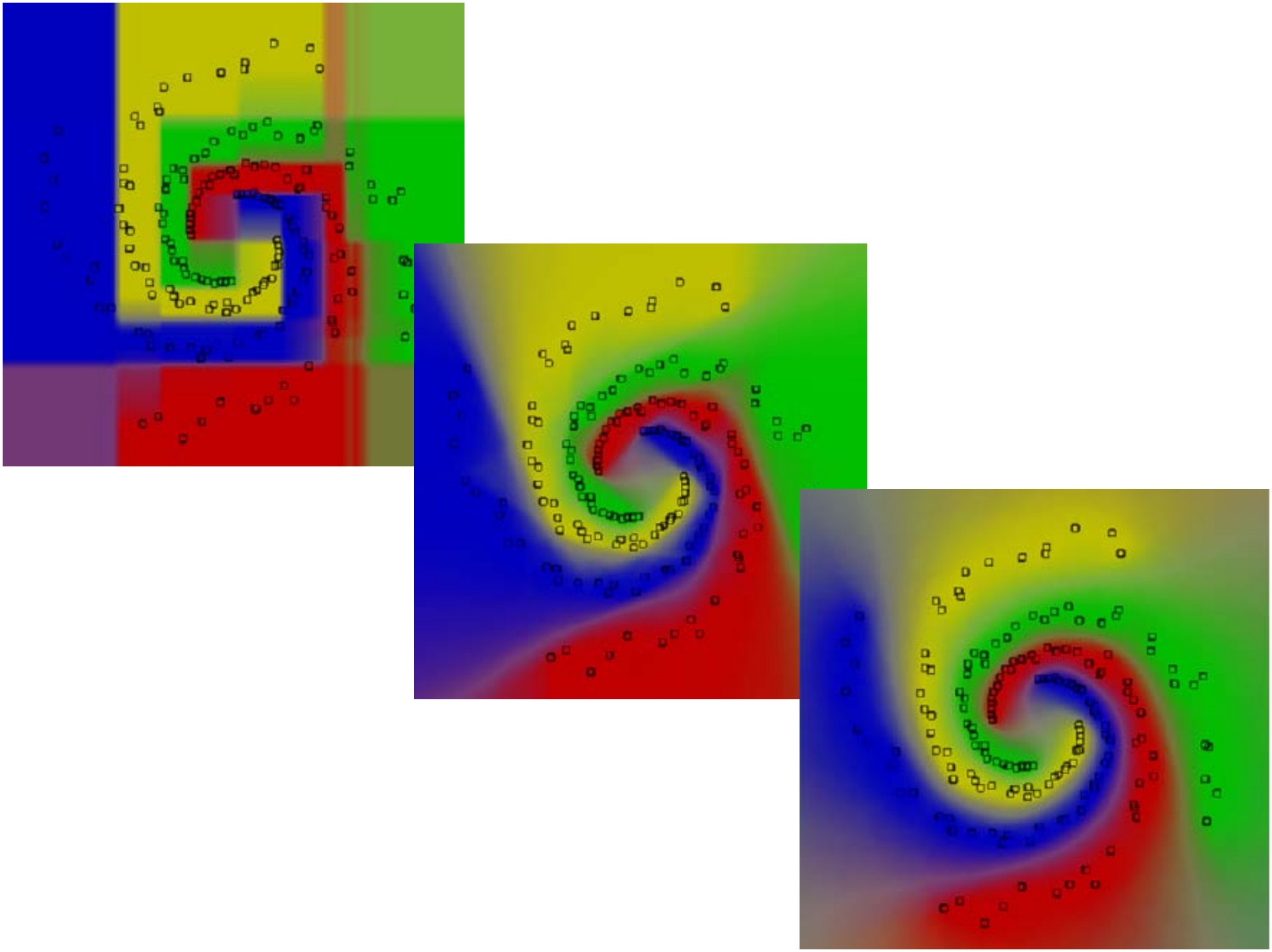
Typy náhodných lesů

1. Diferenciace stromů učením různých podmnožin dat

2. Diferenciace stromů zahrnutím různých množin parametrů


x_2
x_2
x_2
x_1

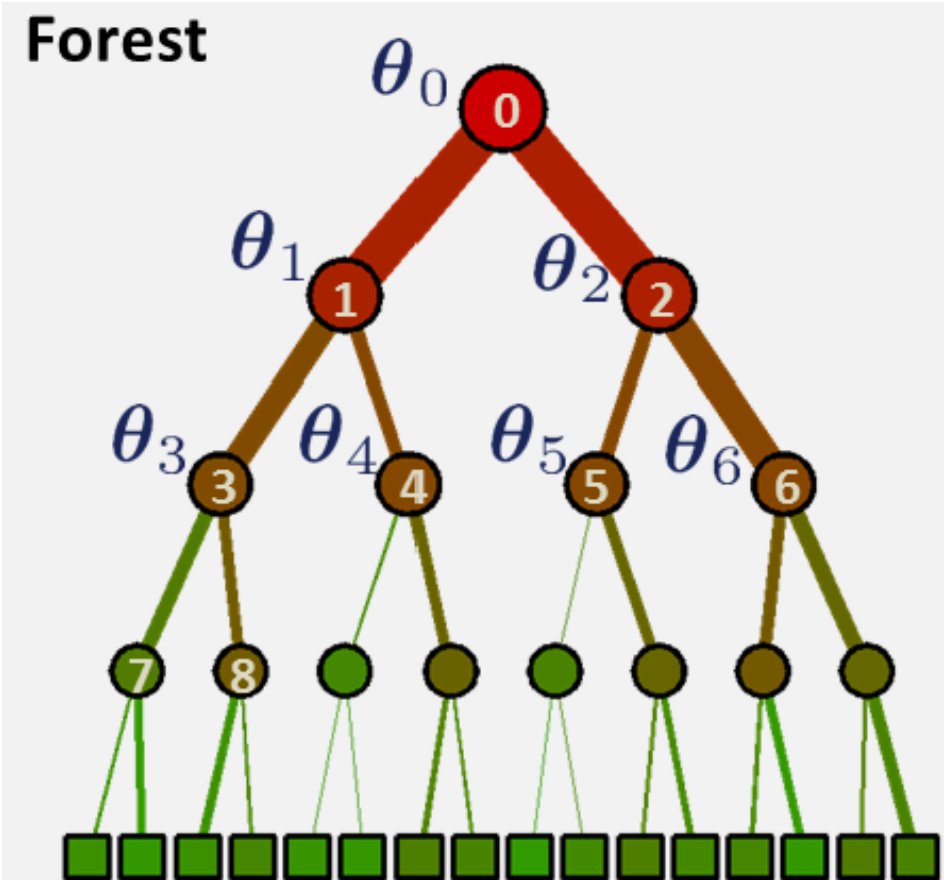


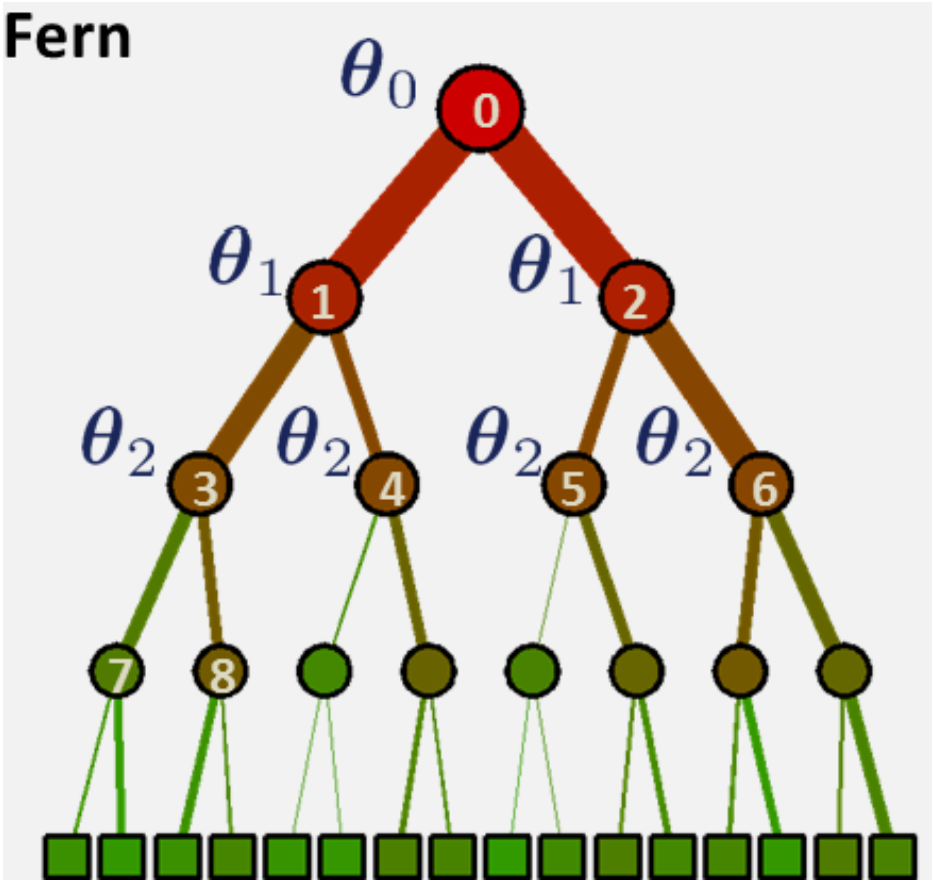


Typy náhodných lesů

1. Diferenciace stromů zahrnutím různých množin parametrů
2. Diferenciace stromů učením různých podmnožin dat
3. Diferenciace stromů kombinací 1. a 2.

♦ Kapradí − les specifických stromů: uzlům stejné úrovně odpovídají stejné proměnné + hodnoty parametrů


Forest
θ_0
0
θ_1
1
θ_2
2
θ_3
3
θ_4
4
θ_5
5
θ_6
6
7
8

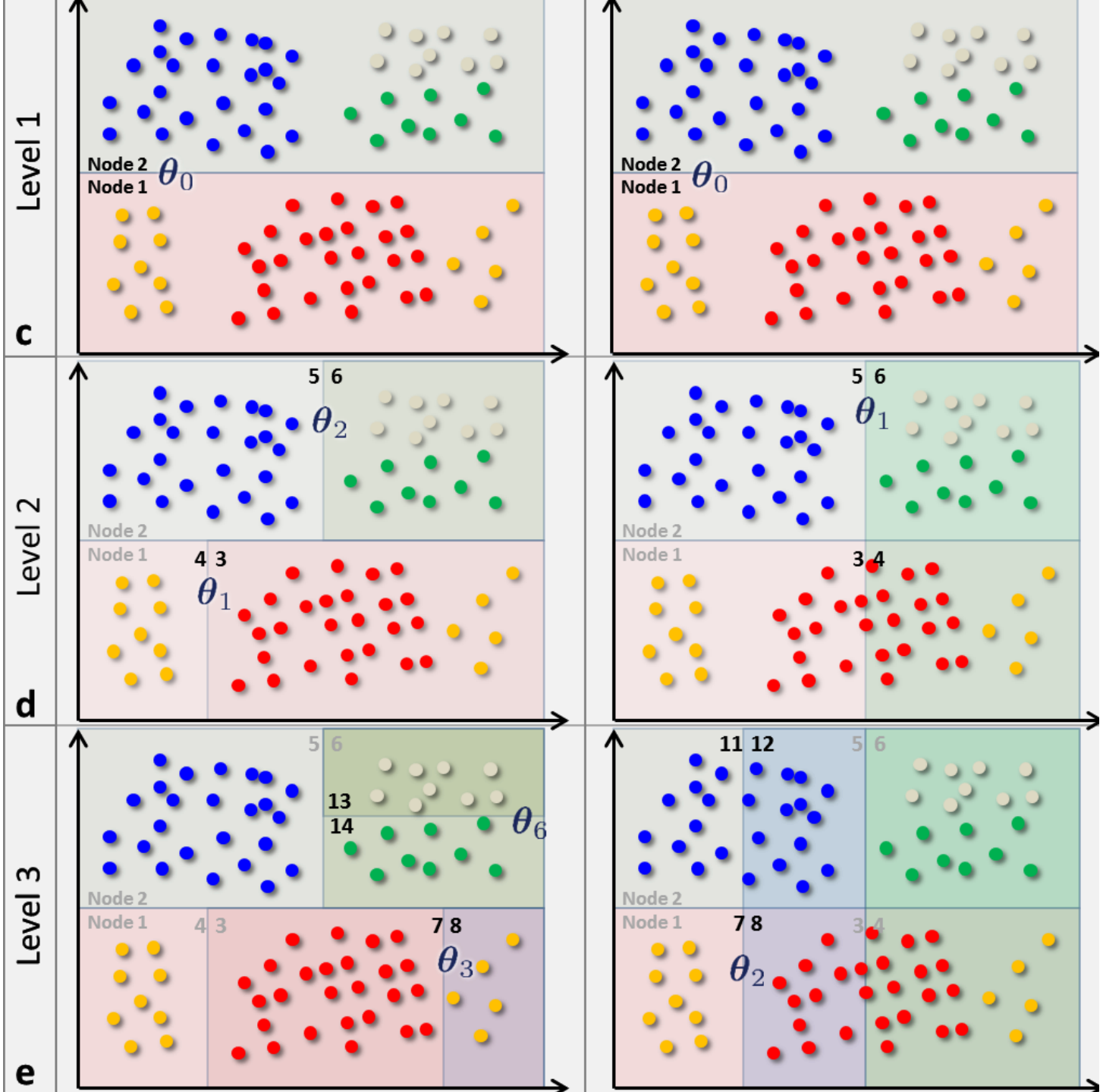

Fern
θ_0
0
θ_1
1
θ_1
2
θ_2
3
θ_2
4
θ_2
5
θ_2
6
7
8

Typy náhodných lesů

1. Diferenciace stromů zahrnutím různých množin parametrů
2. Diferenciace stromů učením různých podmnožin dat
3. Diferenciace stromů kombinací 1. a 2.

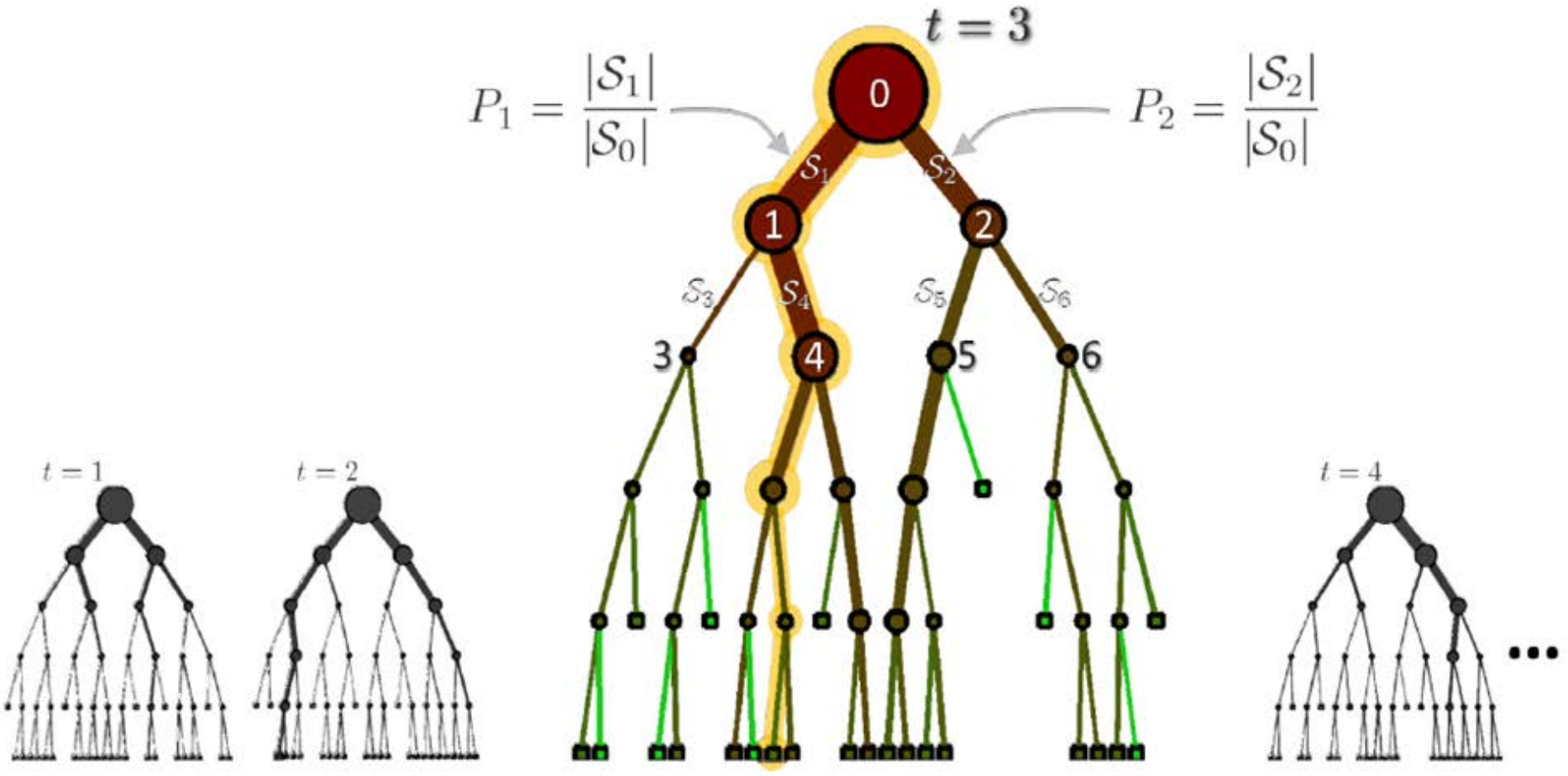
♦ Kapradí − les specifických stromů: uzlům stejné úrovně odpovídají stejné proměnné + hodnoty parametrů

- → rovnoměrnější rozdělení vstupního prostoru (někdy +, jindy -)


Level 1
c
Node 2
Node 1
θ_0
Level 2
d
5 6
θ_2
Node 2
Node 1
4 3
θ_1
5 6
θ_1
Node 2
Node 1
3 4
Level 3
e
5 6
13
14
θ_6
Node 2
Node 1
4 3
7 8
θ_3
11 12
5 6
Node 2
Node 1
7 8
θ_2
3 4

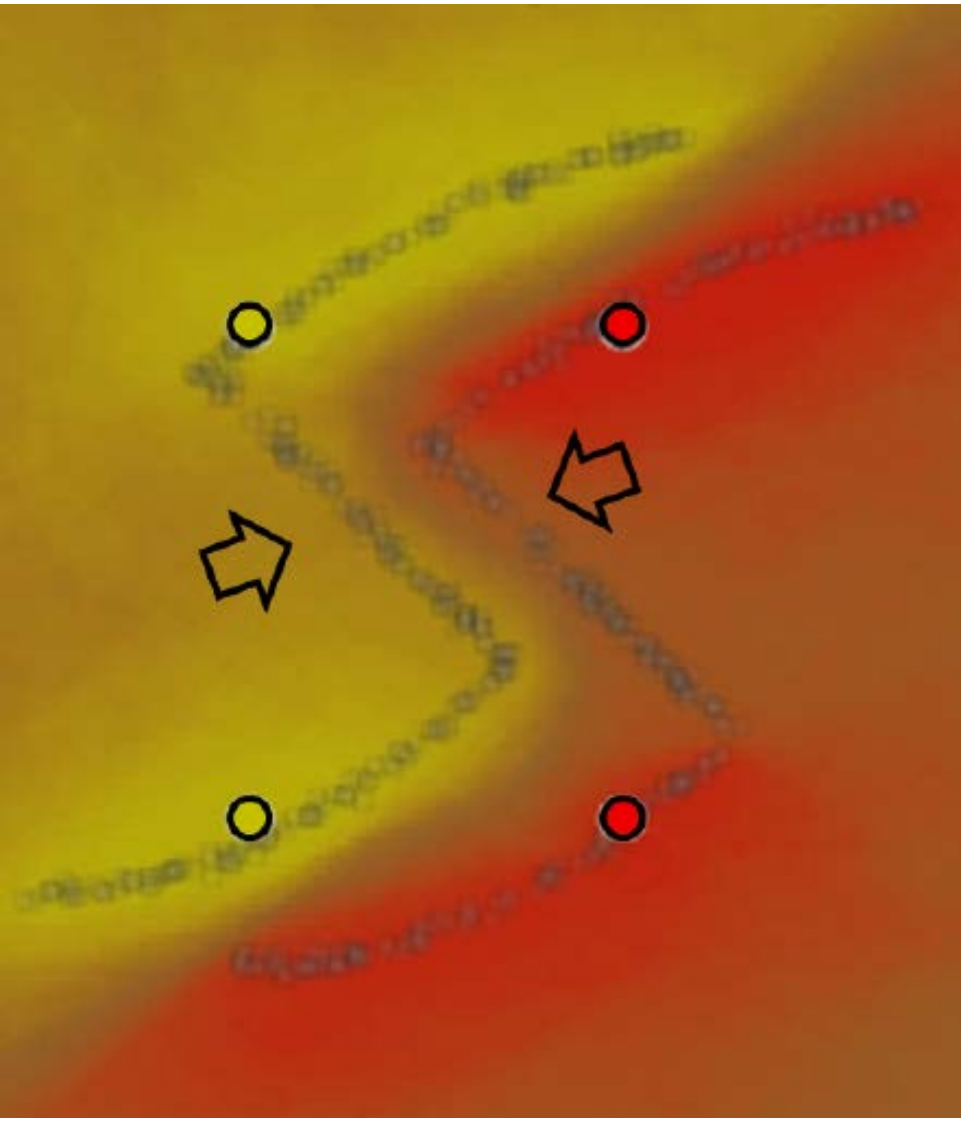
Učení náhodných lesů

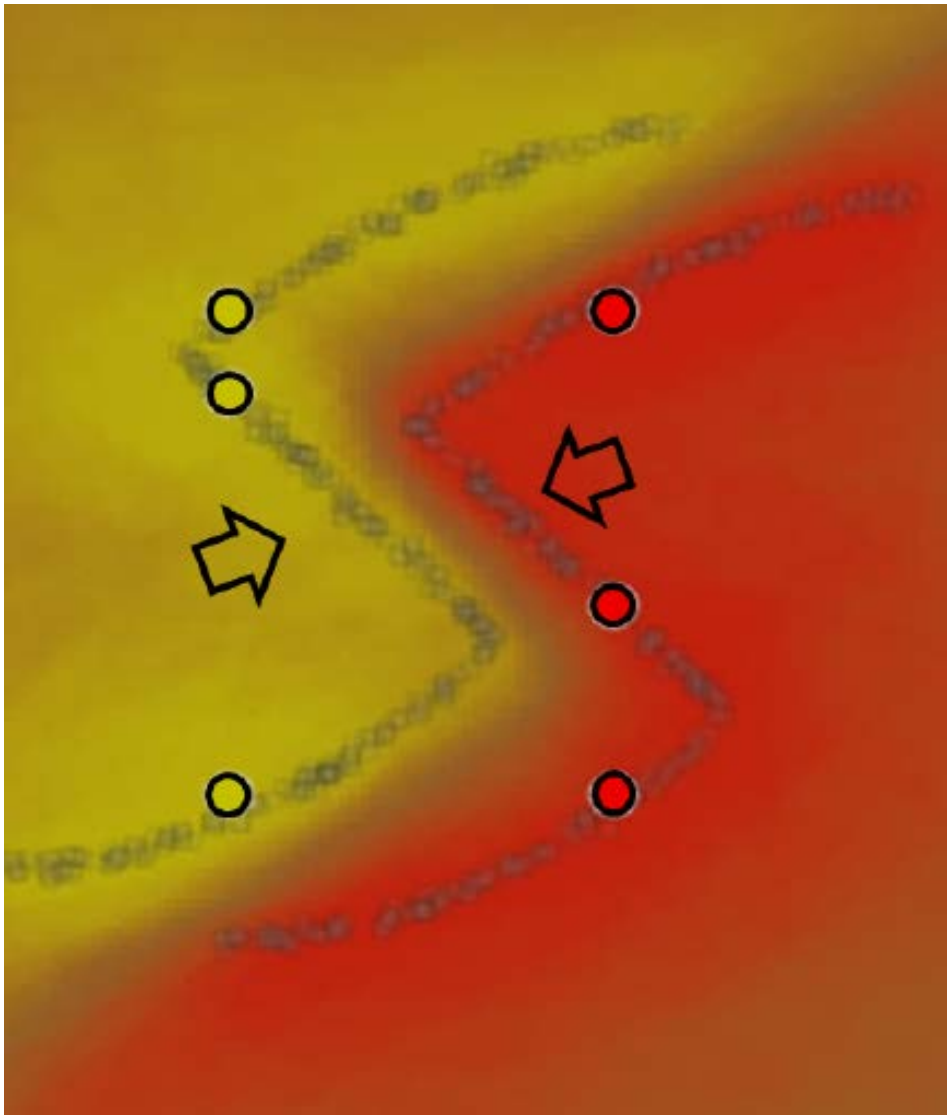
- Každý strom zvlášť (informační zisk | Gini)
  - pouze parametry specificky pro něj zahrnuté
  - pouze na jemu vybrané podmnožině dat
- Výsledné pravděpodobnosti klasifikace lesem: $\sum$ všechny stromy


$t = 3$
$P_1 = \frac{|\mathcal{S}_1|}{|\mathcal{S}_0|}$
$P_2 = \frac{|\mathcal{S}_2|}{|\mathcal{S}_0|}$
0
$\mathcal{S}_1$
$\mathcal{S}_2$
1
2
$\mathcal{S}_3$
$\mathcal{S}_4$
$\mathcal{S}_5$
$\mathcal{S}_6$
3
4
5
6
$t = 1$
$t = 2$
$t = 4$

Učení náhodných lesů

- Každý strom zvlášť (informační zisk | Gini)
  - pouze parametry specificky pro něj zahrnuté
  - pouze na jemu vybrané podmnožině dat
- Výsledné pravděpodobnosti klasifikace lesem: $\sum$ všechny stromy
- Aktivní učení: drahé zjištění správné klasifikace
  - po SVM 2. nejčastější případ, kdy používáno





Náhodné lesy při doporučování

- Klasifikování uživatelů do různých demografických profilů
  - *demografický profil*: kombinace věku, pohlaví, vzdělání,…
  - motivace: profilům odpovídají *specifické strategie doporučování*
- *Učení*: dotázáním některých uživatelů (např. registrovaných)
  - neriskujeme odrazení ostatních vyptáváním − zdržující, vlezlé
- Hlavní důvod lesů: diskrétní + spojité vstupy