

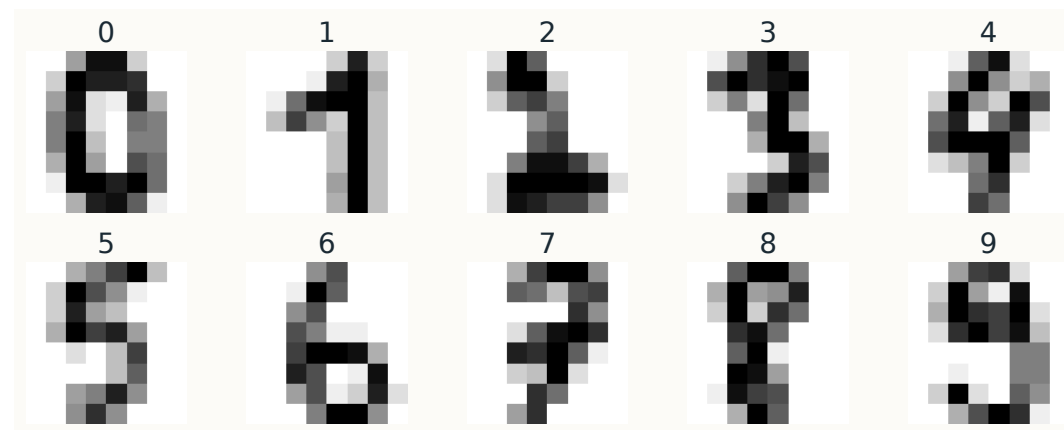
NI-ROZ / ANI-ROZ

Volba metody pro rozpoznávání

Devět algoritmů a jeden kontrolovaný experiment.

Osm úvodních slidů a 90 minut praktické práce.

[Zadání cvičení](#) [Notebook](#) [Python](#)



Skutečné trénovací obrázky z `load_digits`, 8×8 pixelů.

Devět metod, různé úlohy

Metoda	Úloha zde	Princip
Linear regression	regrese	Přímka v příznacích, predikce číselné hodnoty
Logistic regression	klasifikace	Lineární skóre a model pravděpodobností tříd
Decision tree	klasifikace	Posloupnost dělení, zde hloubka 8
Random forest	klasifikace	Průměrování randomizovaných stromů
SVM (SVC)	klasifikace	Hranice s marginem, zde RBF kernel
k-NN	klasifikace	Hlasování tří nejbližších trénovacích vzorů
k-means	shlukování	Deset shluků, počet K dodáváme
Gaussian naive Bayes	klasifikace	Model hustot s podmíněnou nezávislostí příznaků
Gradient boosting	klasifikace	Postupně přidáváme malé stromy podle ztráty

Role v tabulce odpovídá zvolené variantě. Například k-NN a stromy mají i regresní varianty.

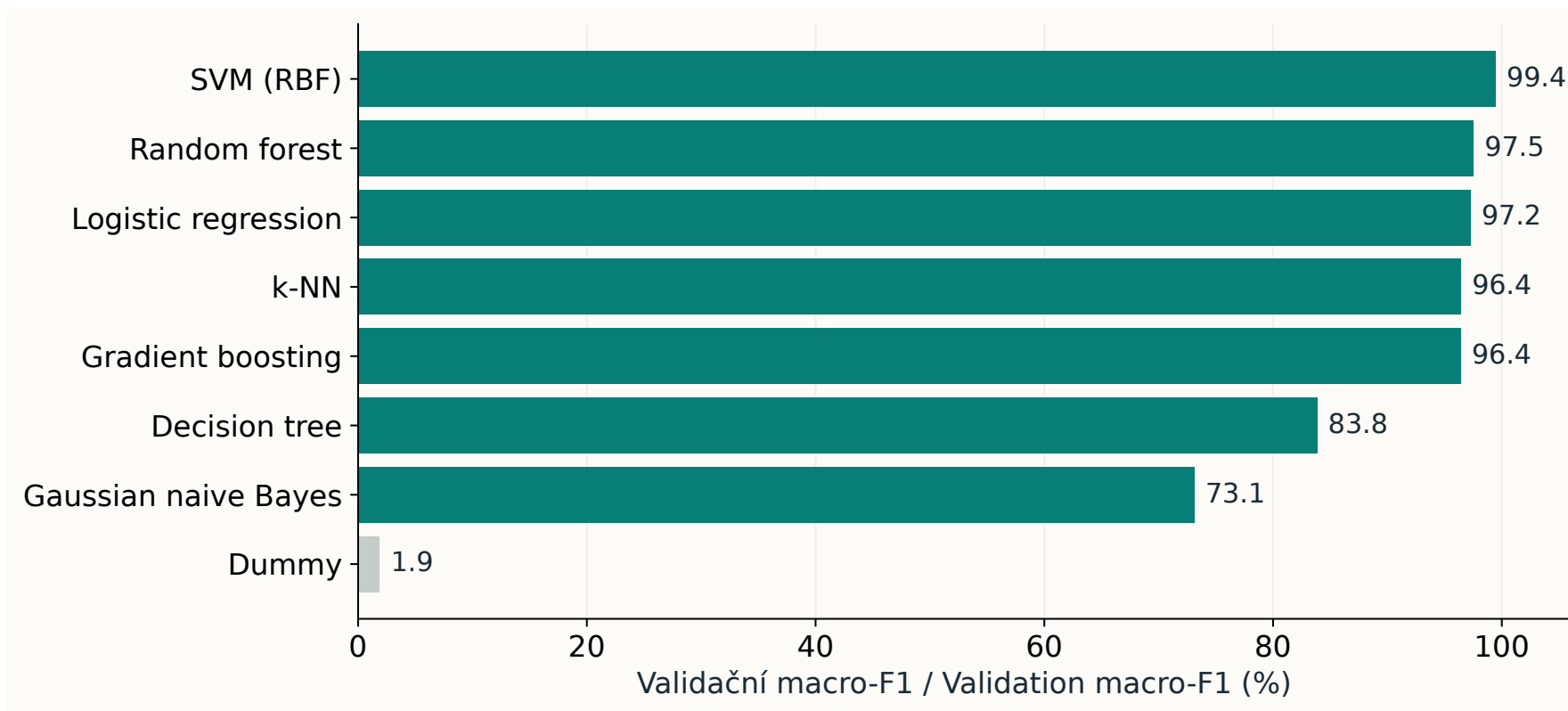
Trénink, validace a test

Část	Počet	Použití
Train	1077	Učení parametrů i scaleru
Validation	360	Volba metody, příznaků a nastavení
Test	360	Závěrečné posouzení předem vybrané varianty

StandardScaler se učí jen na tréninku. Výběr podle validačního macro-F1 předchází jedinému závěrečnému testu.

Dělení je po obrázcích. V tomto příkladu neověřujeme zobecnění na nové pisatele.

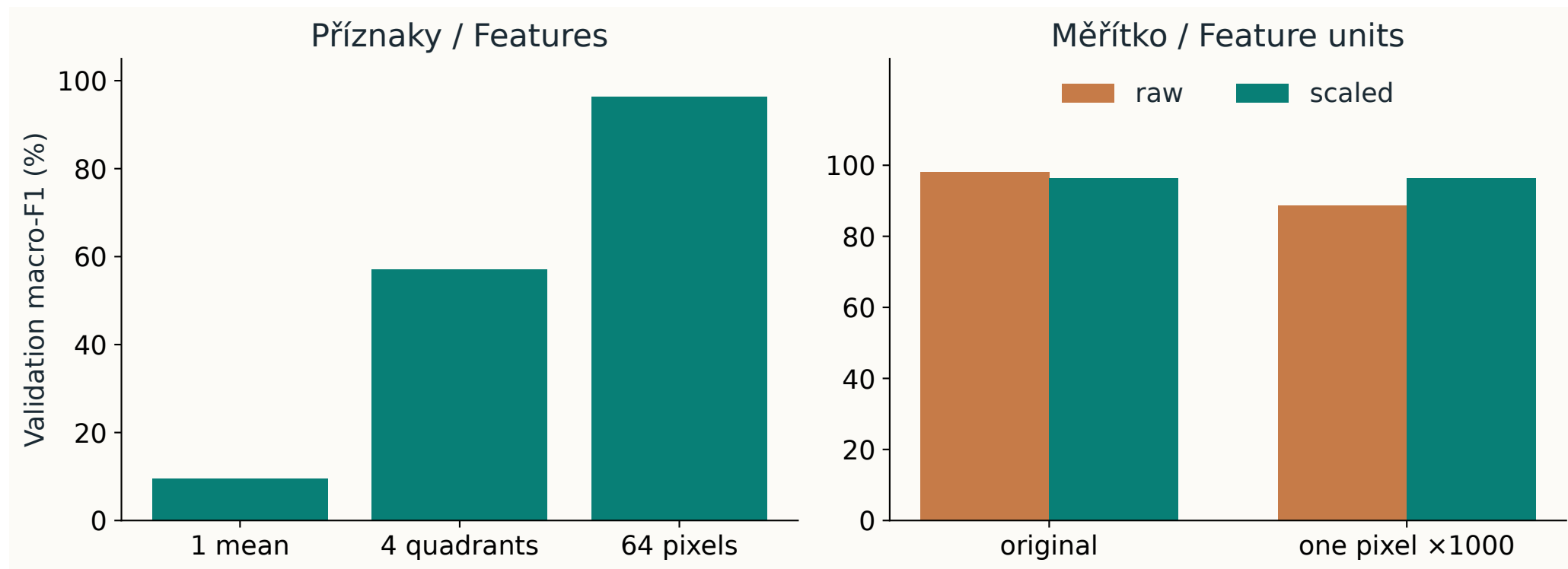
Porovnání klasifikátorů



Vzorový běh na stejné validaci. Dummy je kontrolní základ, nikoli jedna z devíti metod příspěvku.

Které chyby se skrývají za jedním číslem? Prohlédněte matici záměn.

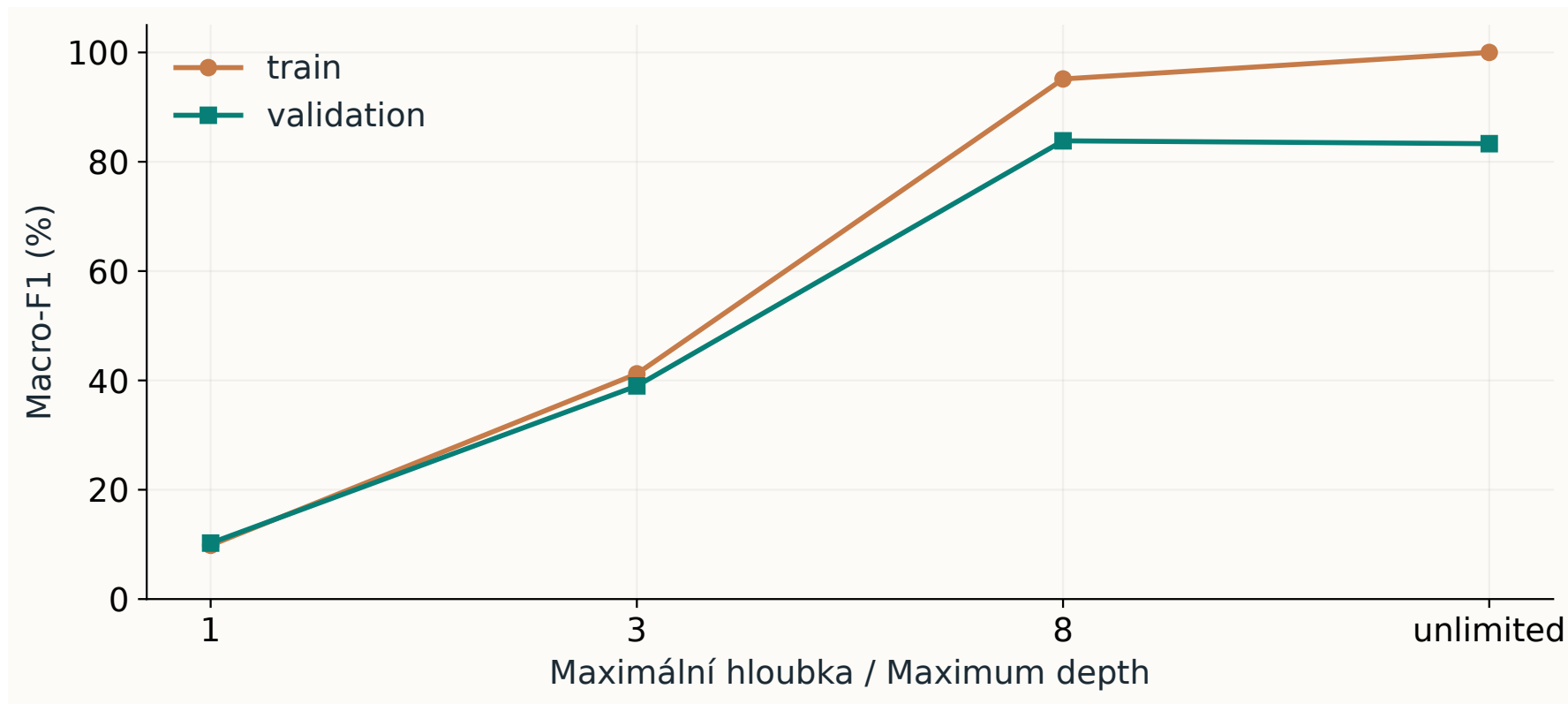
Příznaky a měřítko



Vlevo měníme pouze příznaky. Vpravo přepočítáme jeden pixel na jiné jednotky. Všude používáme k-NN, $k = 3$.

Jeden průměr ztratí uspořádání pixelů. Škálování mění vzdálenosti, podle kterých hledáme sousedy.

Kapacita stromu a soubory modelů

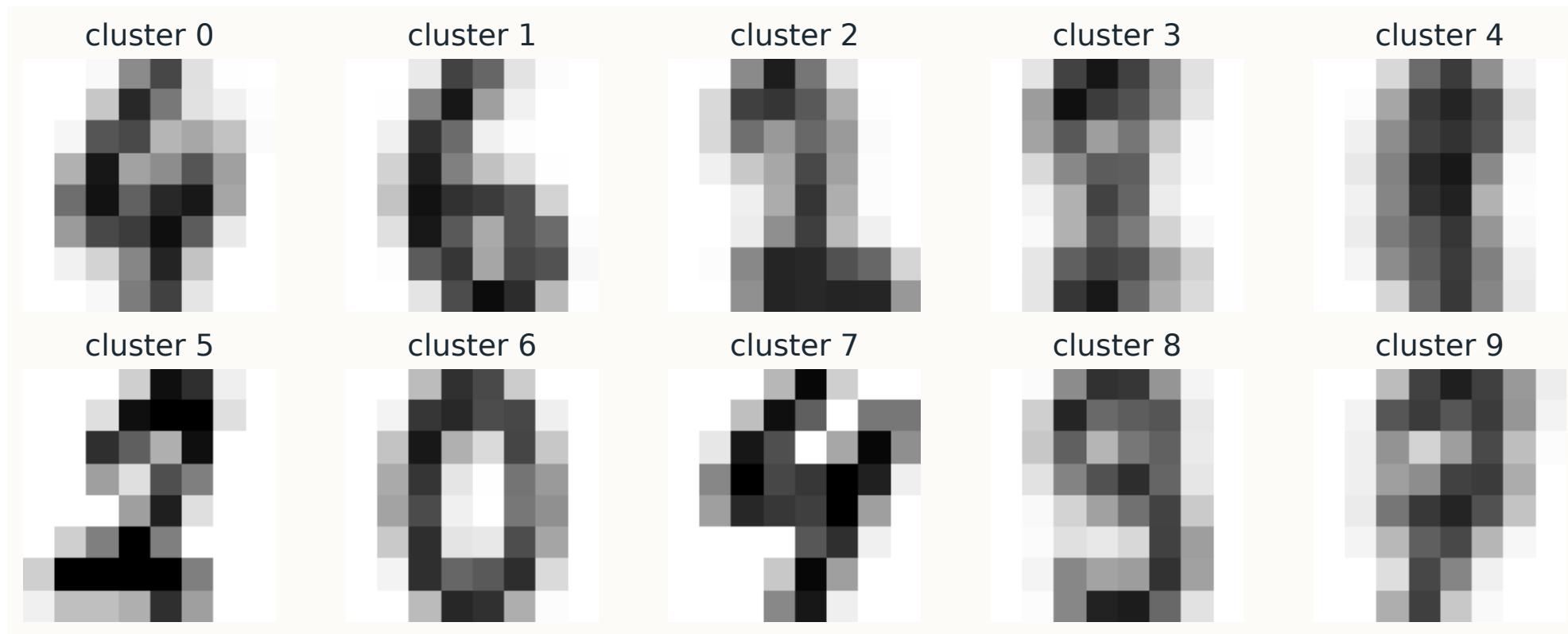


Skutečný běh stejného stromového algoritmu s různou maximální hloubkou.

Random forest: randomizované stromy a průměrování. **Gradient boosting:** postupné přidávání modelů podle ztráty.

Pro klasifikační log-loss používáme záporný gradient. Obyčejná rezidua $y - \hat{y}$ odpovídají speciálnímu případu kvadratické regrese.

Shluky a třídy

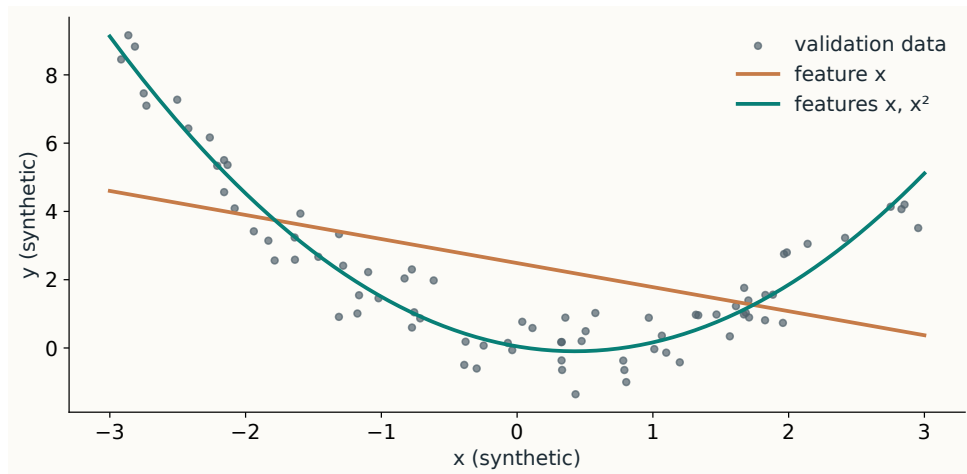


Centra k-means naučená bez třídních štítků. $K = 10$ dodáváme předem.

ARI na validaci: **0.490**. Po přejmenování shluků: **0.490**.

Číslo shluku nemá samo o sobě význam číslice.

Regrese a samostatný experiment



Syntetická regresní úloha. Oba modely používají lineární regresi, liší se příznaky.

Přidání x^2 umožní popsat zakřivení. Regresi hodnotíme pomocí MAE a RMSE.

V závěru si zvolíte jednu klasifikační variantu podle validace a popíšete její chyby. Test otevřete až po této volbě.

[Závěrečné zadání](#)