

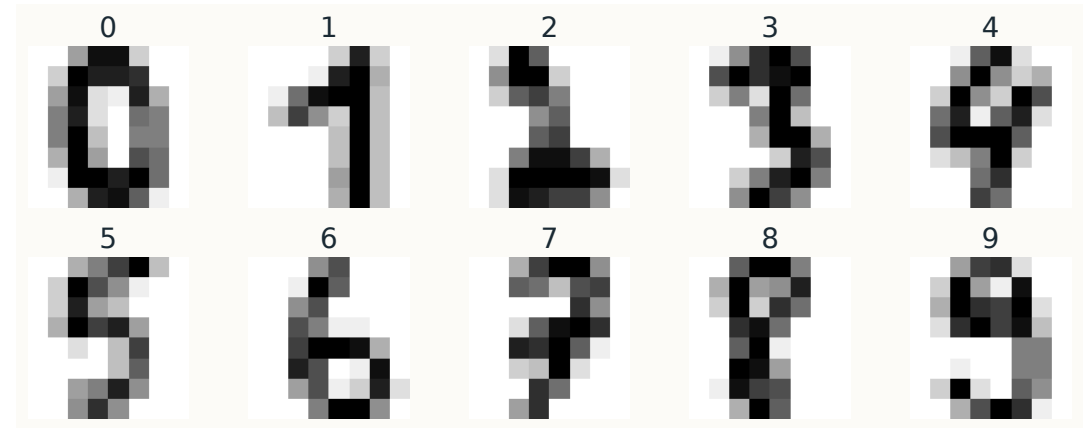
NI-ROZ / ANI-ROZ

Choosing a recognition method

Nine algorithms and one controlled experiment.

Eight introductory slides and a 90-minute practical session.

[Open the worksheet](#) [Notebook](#) [Python](#)



Actual training images from load_digits, 8×8 pixels.

Nine methods, different tasks

Method	Task here	Idea
Linear regression	regression	A linear combination of features predicts a number
Logistic regression	classification	Linear scores and a model of class probabilities
Decision tree	classification	A sequence of splits, depth 8 here
Random forest	classification	An average of randomized trees
SVM (SVC)	classification	A margin-based boundary, using an RBF kernel here
k-NN	classification	A vote of the three nearest training examples
k-means	clustering	Ten clusters, with K supplied as an input
Gaussian naive Bayes	classification	Density models with conditional feature independence
Gradient boosting	classification	Small trees added sequentially using a loss function

The table describes the chosen variants. For example, k-NN and trees also have regression variants.

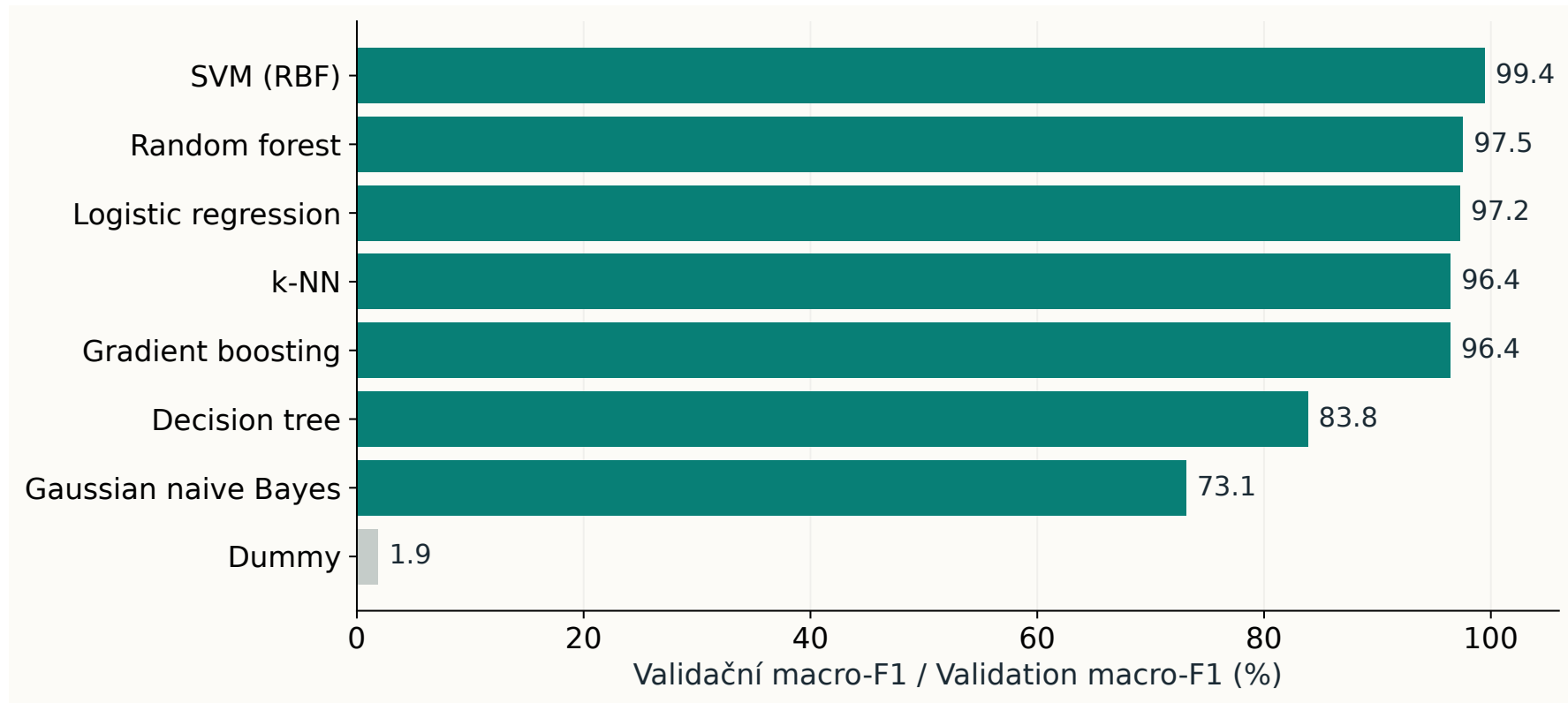
Training, validation and test

Subset	Count	Purpose
Train	1077	Fit parameters and scalers
Validation	360	Choose method, features and settings
Test	360	Final evaluation of a previously selected configuration

StandardScaler learns from training data only. Selection by validation macro-F1 precedes the final test.

The split is image-based. This example does not test generalization to new writers.

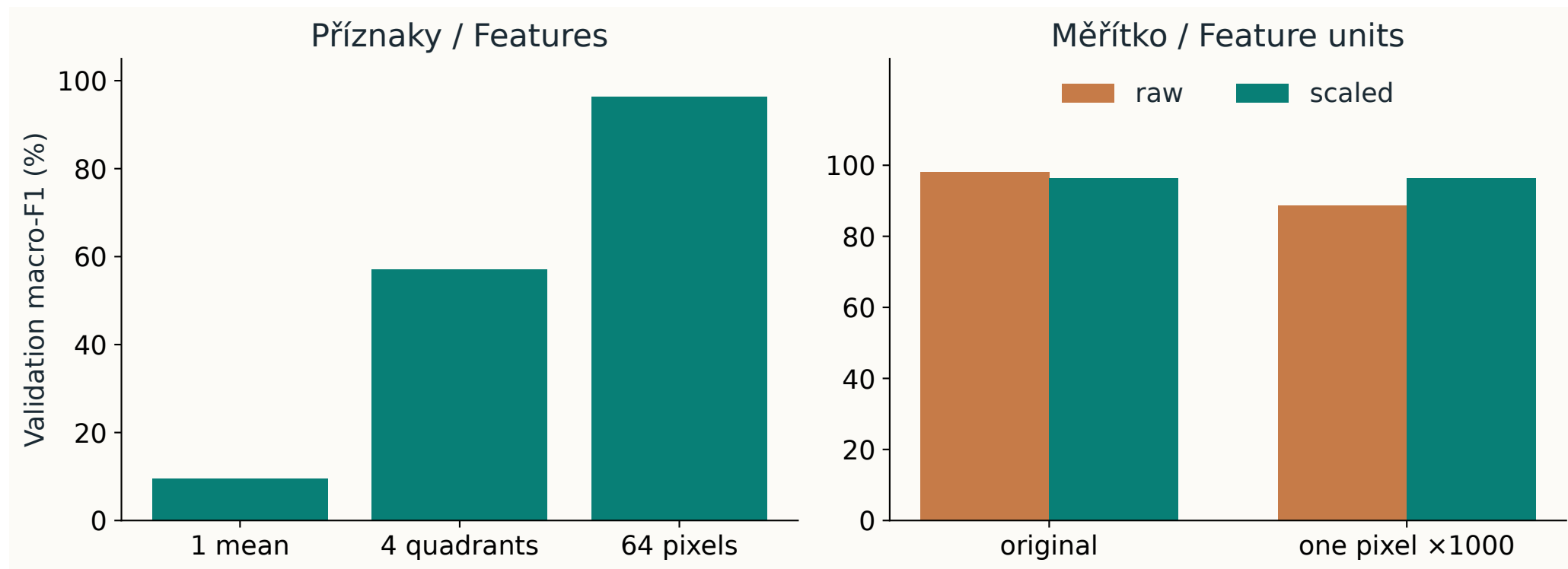
Comparing classifiers



Example run on the same validation split. Dummy is an added control baseline, not one of the post's nine methods.

Which errors does a single score hide? Inspect the confusion matrix.

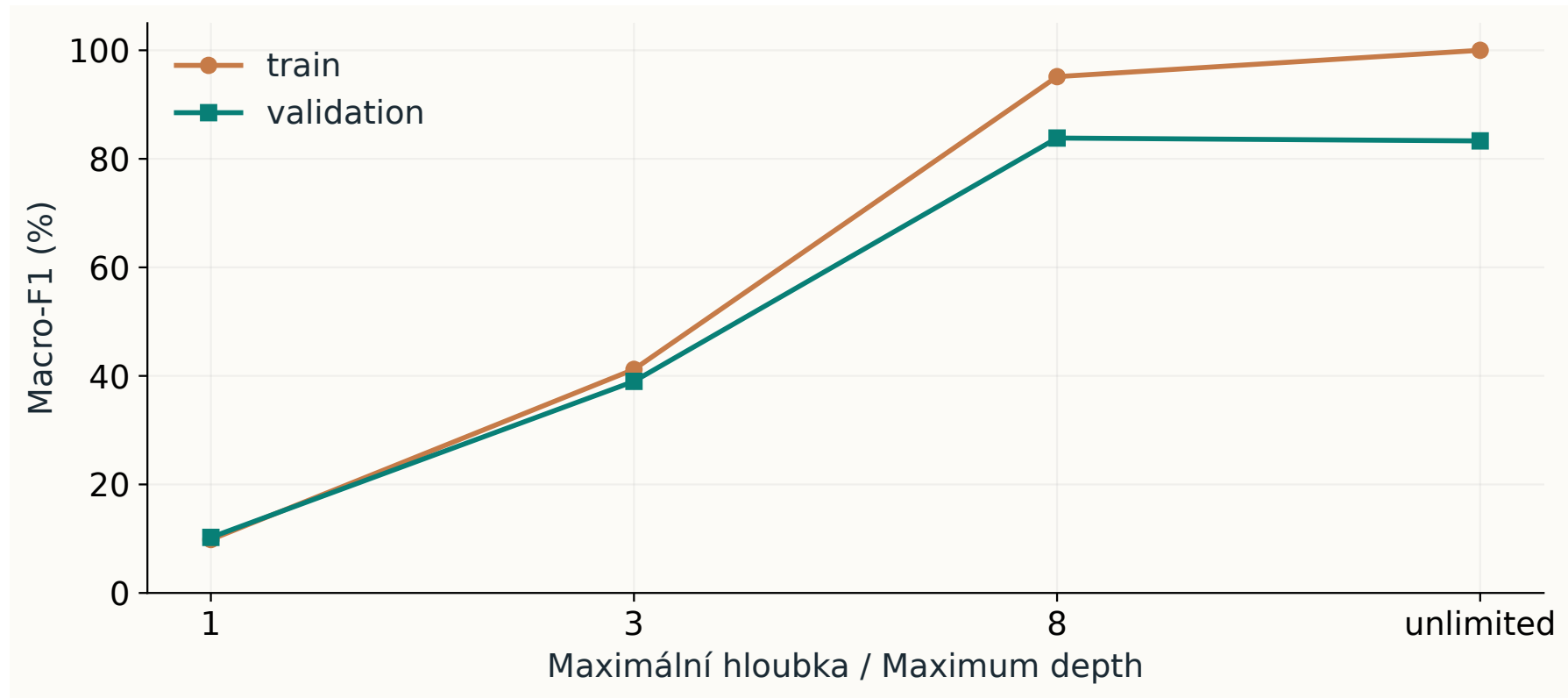
Features and feature units



Left: only the features change. Right: one pixel changes units. All runs use k-NN, $k = 3$.

A single mean loses pixel layout. Scaling changes the distances used to find neighbours.

Tree capacity and ensembles

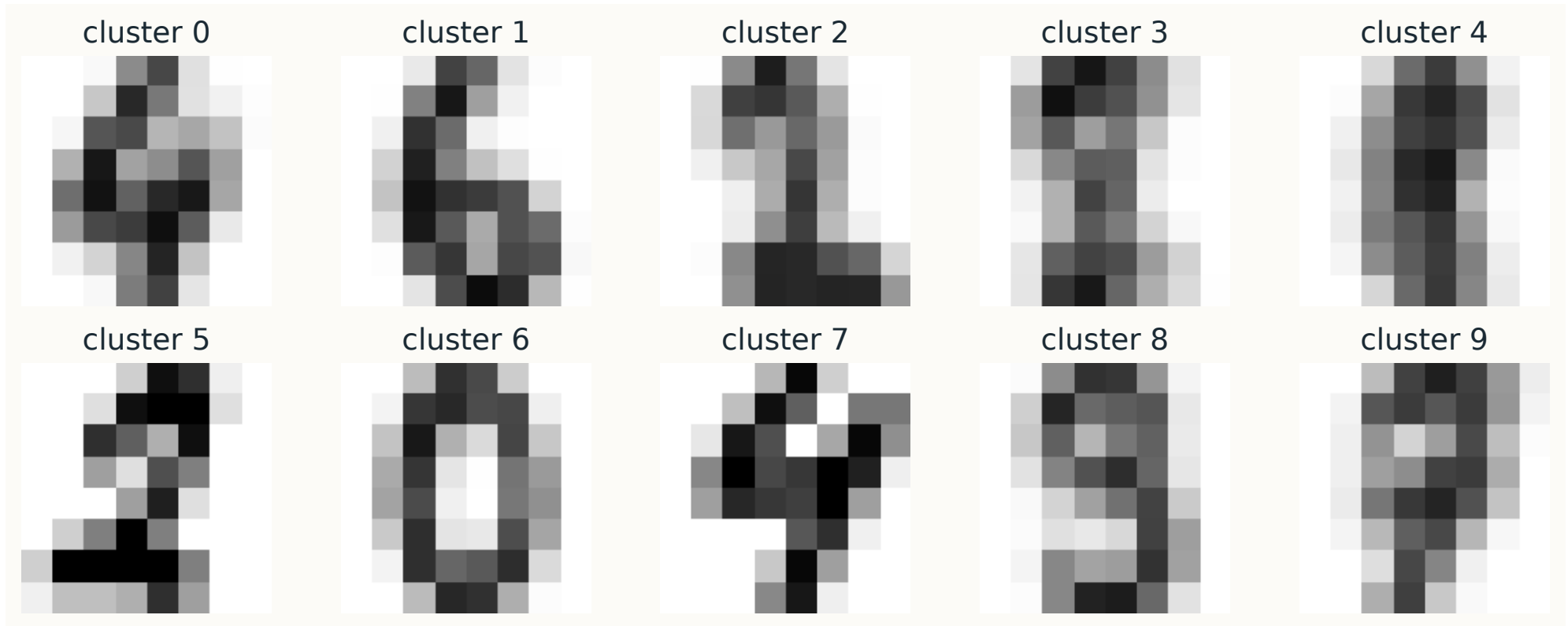


Actual runs of the same tree algorithm at different maximum depths.

Random forest: randomized trees and averaging. **Gradient boosting:** sequential additions guided by the loss.

Classification with log-loss uses the negative gradient. Ordinary residuals $y - \hat{y}$ describe the special case of squared-error regression.

Clusters and classes

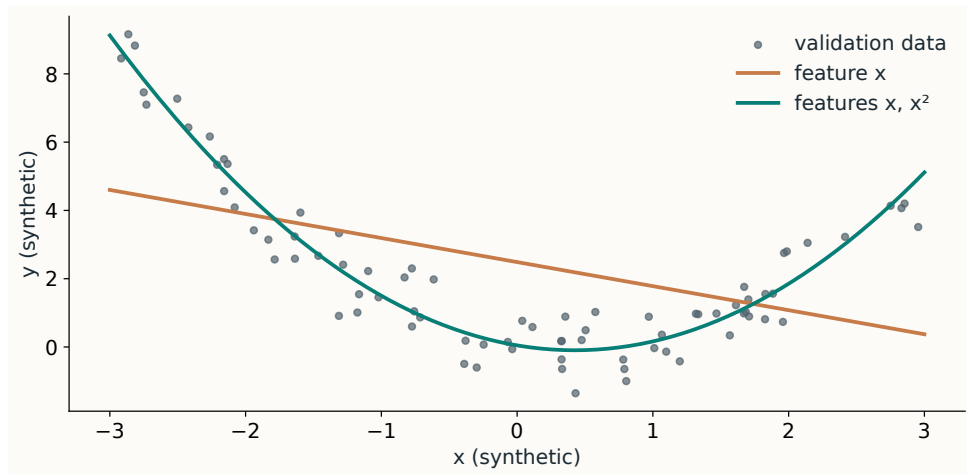


K-means centres learned without class labels. $K = 10$ is supplied in advance.

Validation ARI: **0.490**. After renaming clusters: **0.490**.

A cluster ID does not by itself denote a digit.

Regression and an independent experiment



Synthetic regression task. Both models use linear regression with different features.

Adding x^2 represents curvature. We evaluate regression with MAE and RMSE.

At the end, choose one classification configuration using validation and explain its errors. Open the test only after choosing.

[Final assignment](#)