Data Mining with Weka

Cross-validation

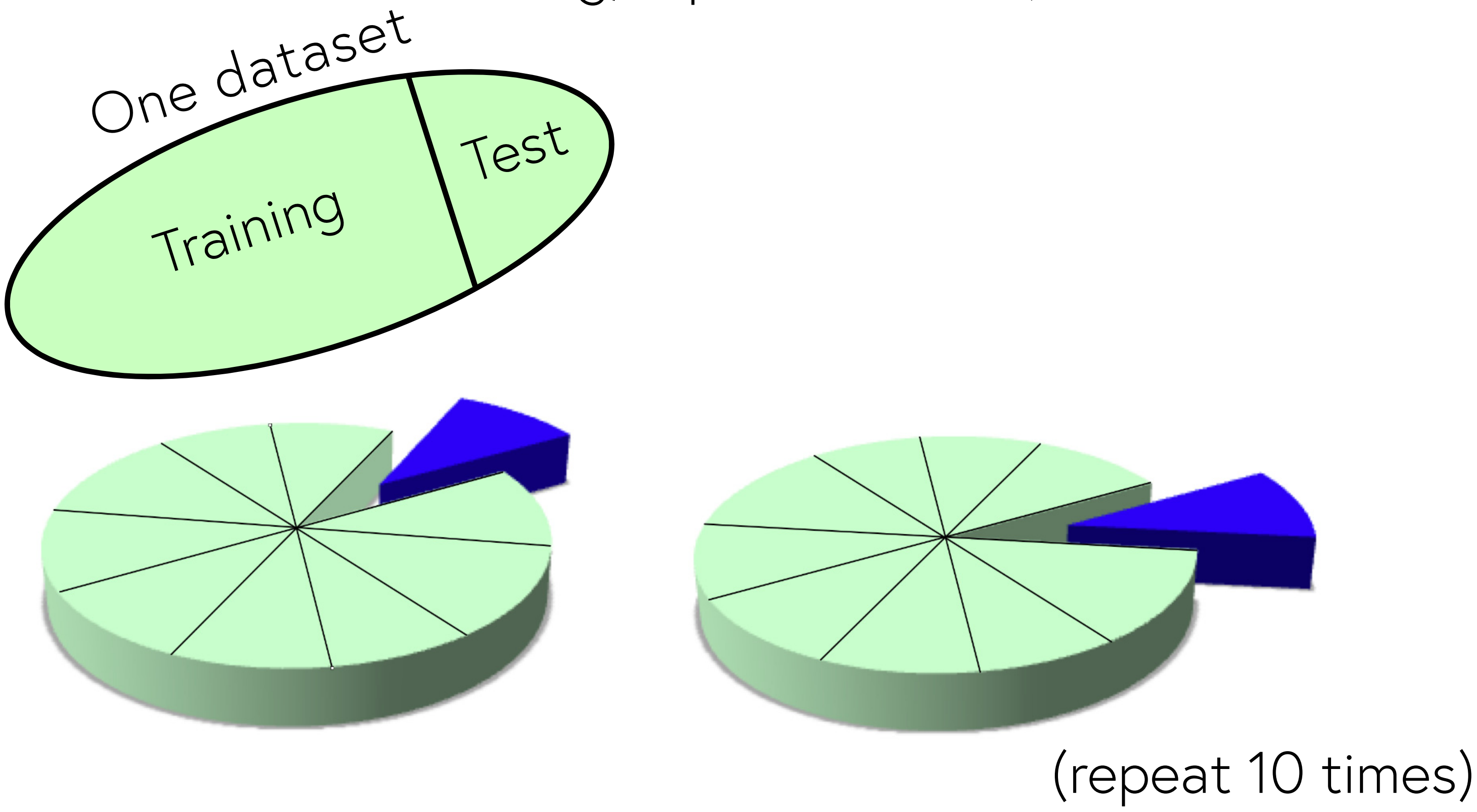
Ian H. Witten

Cross-validation

- Can we improve upon repeated holdout?
  (i.e. reduce variance)
- Cross-validation
- Stratified cross-validation

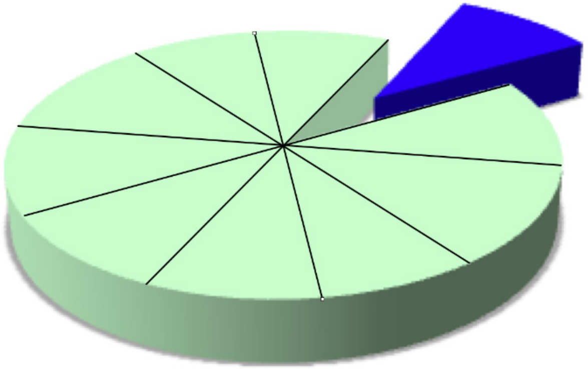
Cross-validation

- Repeated holdout
(in "Repeated training and testing" lesson,
hold out 10% for testing, repeat 10 times)



Cross-validation

10-fold cross-validation



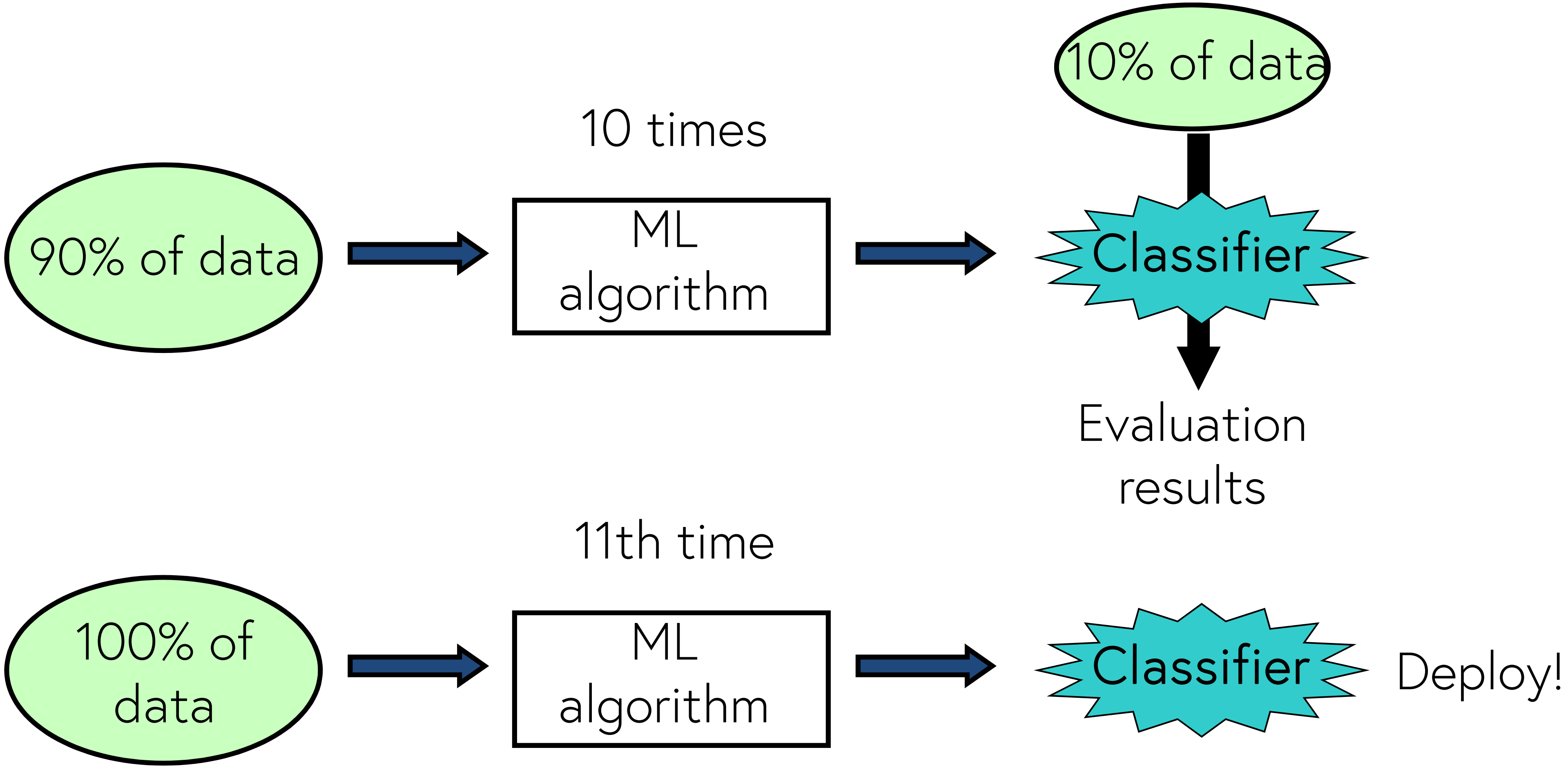
- Divide dataset into 10 parts (folds)
- Hold out each part in turn
- Average the results
- Each data point used once for testing, 9 times for training

Stratified cross-validation

- Ensure that each fold has the right proportion of each class value

Cross-validation

After cross-validation, Weka outputs an extra model built on the entire dataset



Cross-validation

- Cross-validation better than repeated holdout
- Stratified is even better
- With 10-fold cross-validation, Weka invokes the learning algorithm 11 times
- Practical rule of thumb:
  - Lots of data? – use percentage split
  - Else stratified 10-fold cross-validation