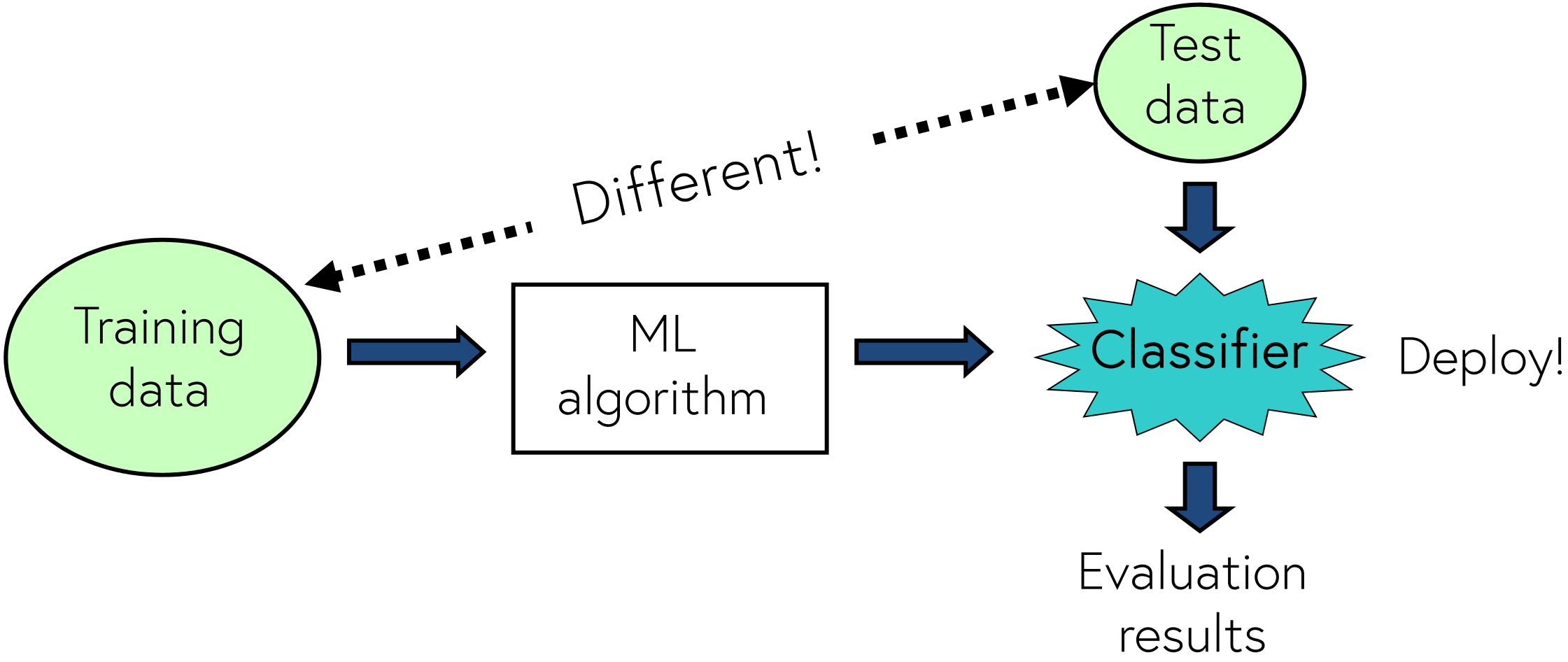
Data Mining with Weka

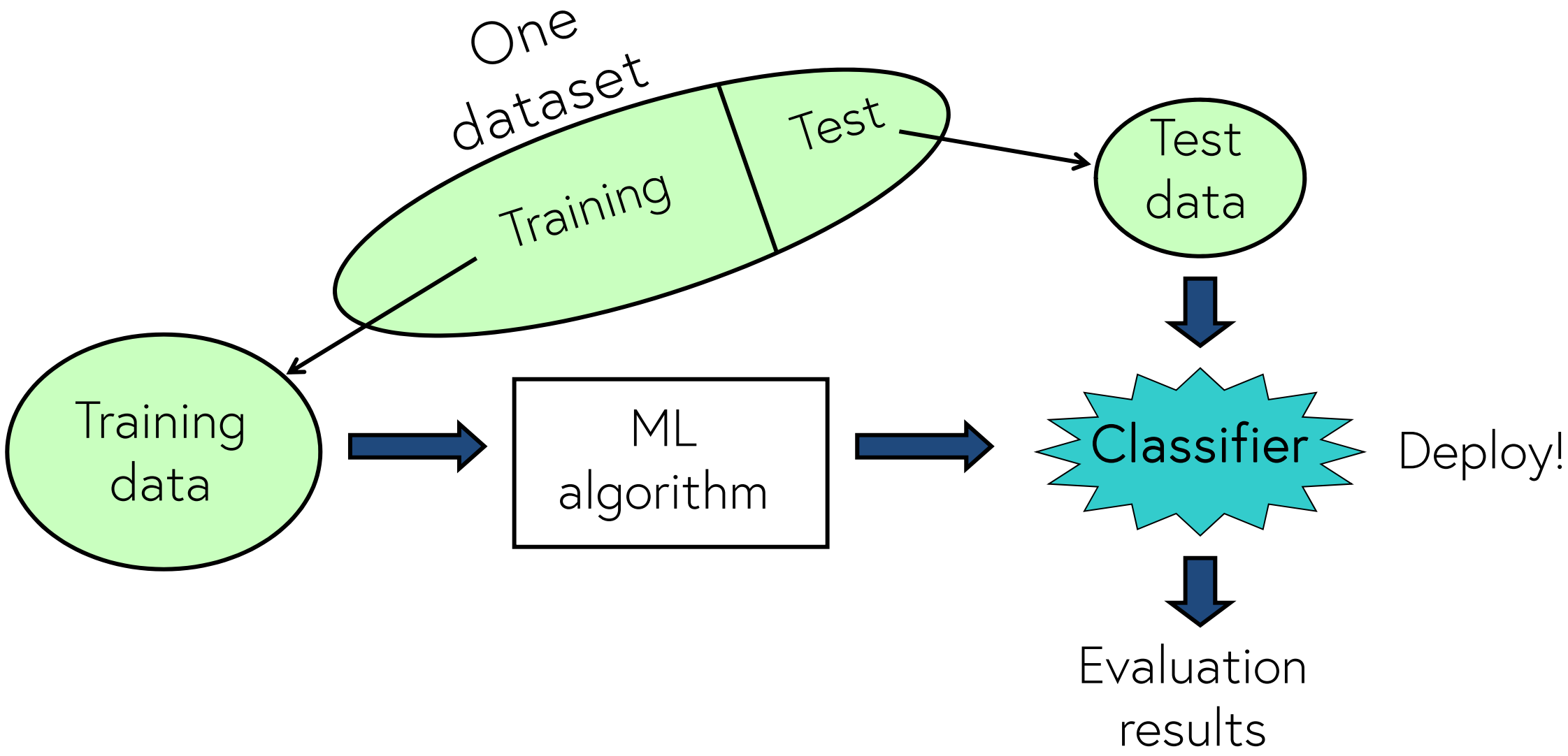
Training and testing

Ian H. Witten

Training and testing



Training and testing



Basic assumption: training and test sets produced by independent sampling from an infinite population

Training and testing

Use J48 to analyze the segment dataset

- Open file segment-challenge.arff
- Choose J48 decision tree learner (trees>J48)
- Supplied test set segment-test.arff
- Run it: 96% accuracy
- Evaluate on training set: 99% accuracy
- Evaluate on percentage split: 95% accuracy
- Do it again: get exactly the same result!

Training and testing

- Basic assumption:
  training and test sets sampled independently from an infinite population
- Just one dataset? — hold some out for testing
- Expect slight variation in results
- … but Weka produces same results each time
- J48 on segment-challenge dataset