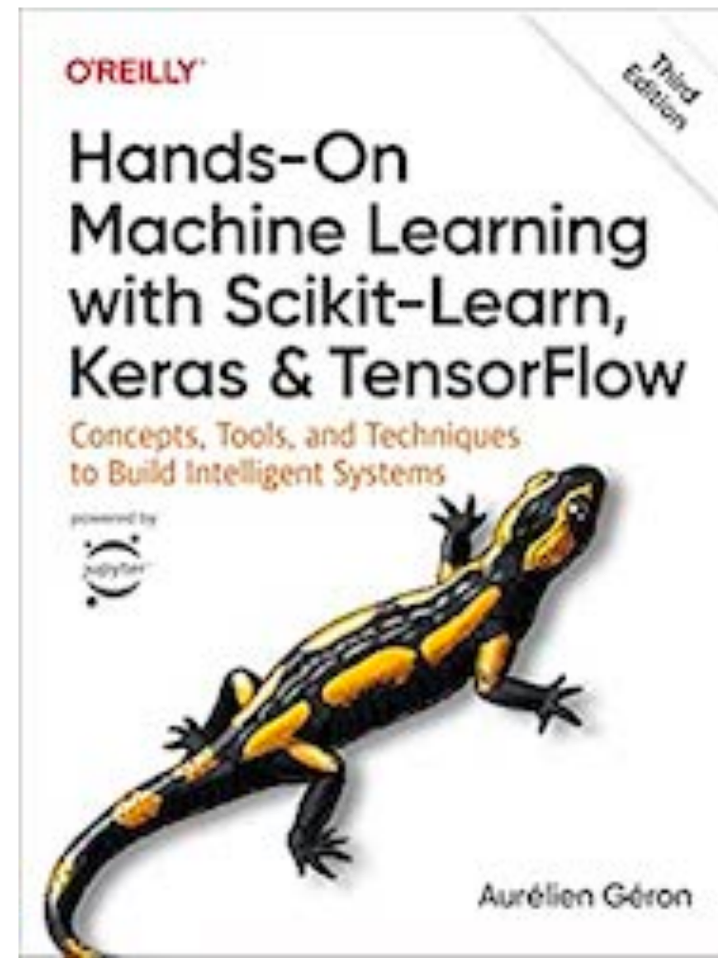
Machine Learning Security

7 Ensemble Learing and Random Forests



Made Oct 1, 2023

Topics

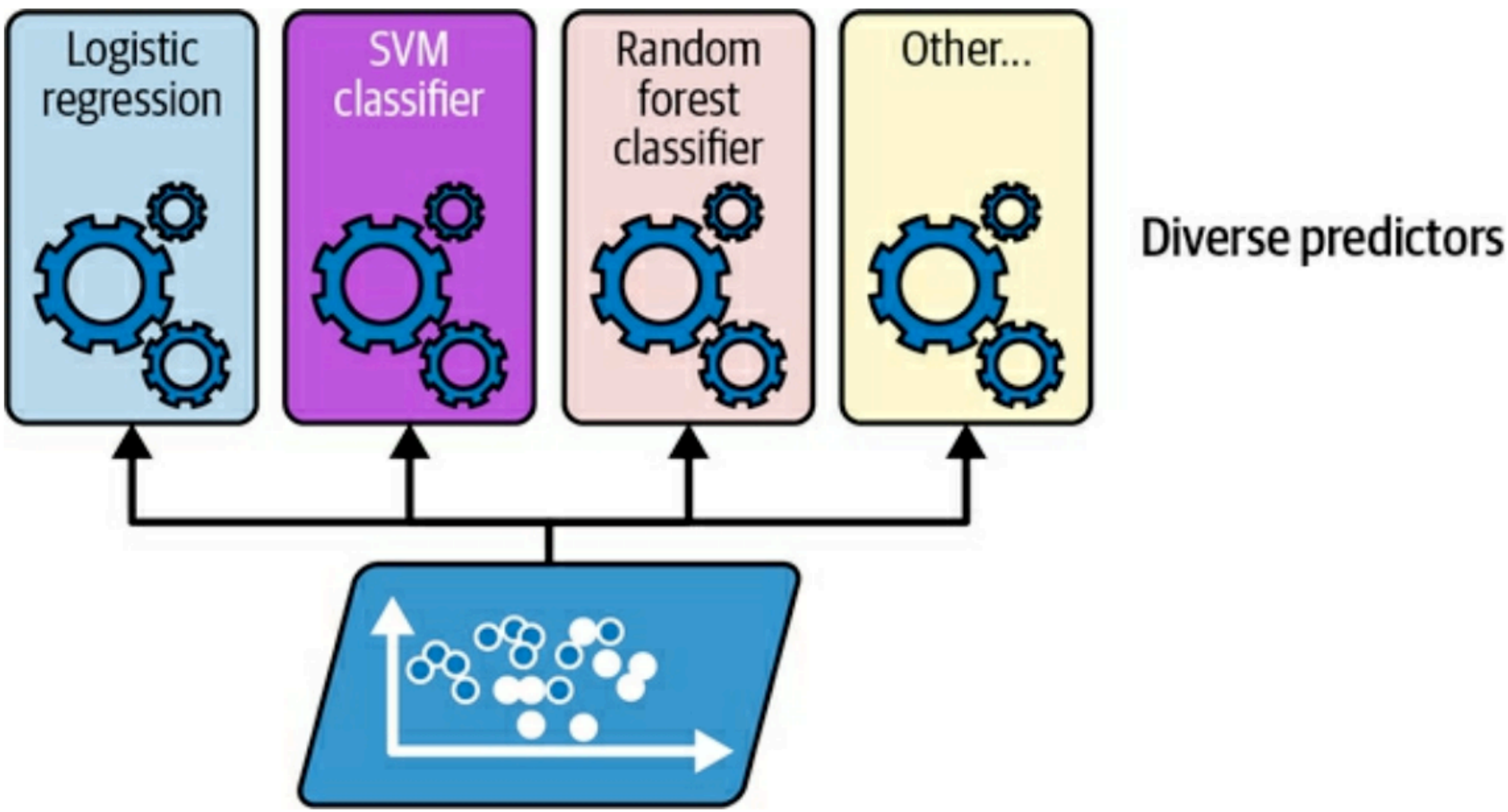
- Voting Classifiers
- Bagging and Pasting
- Random Forests
- Boosting
- Stacking

Ensemble Learning

- Aggregate the predictions of several different models
  - An ***ensemble***
  - Using ***the wisdom of the crowd***
- **Random forest**
  - A group of decision tree classifiers
    - Trained on different subsets of the data
  - One of the most powerful ML algorithms

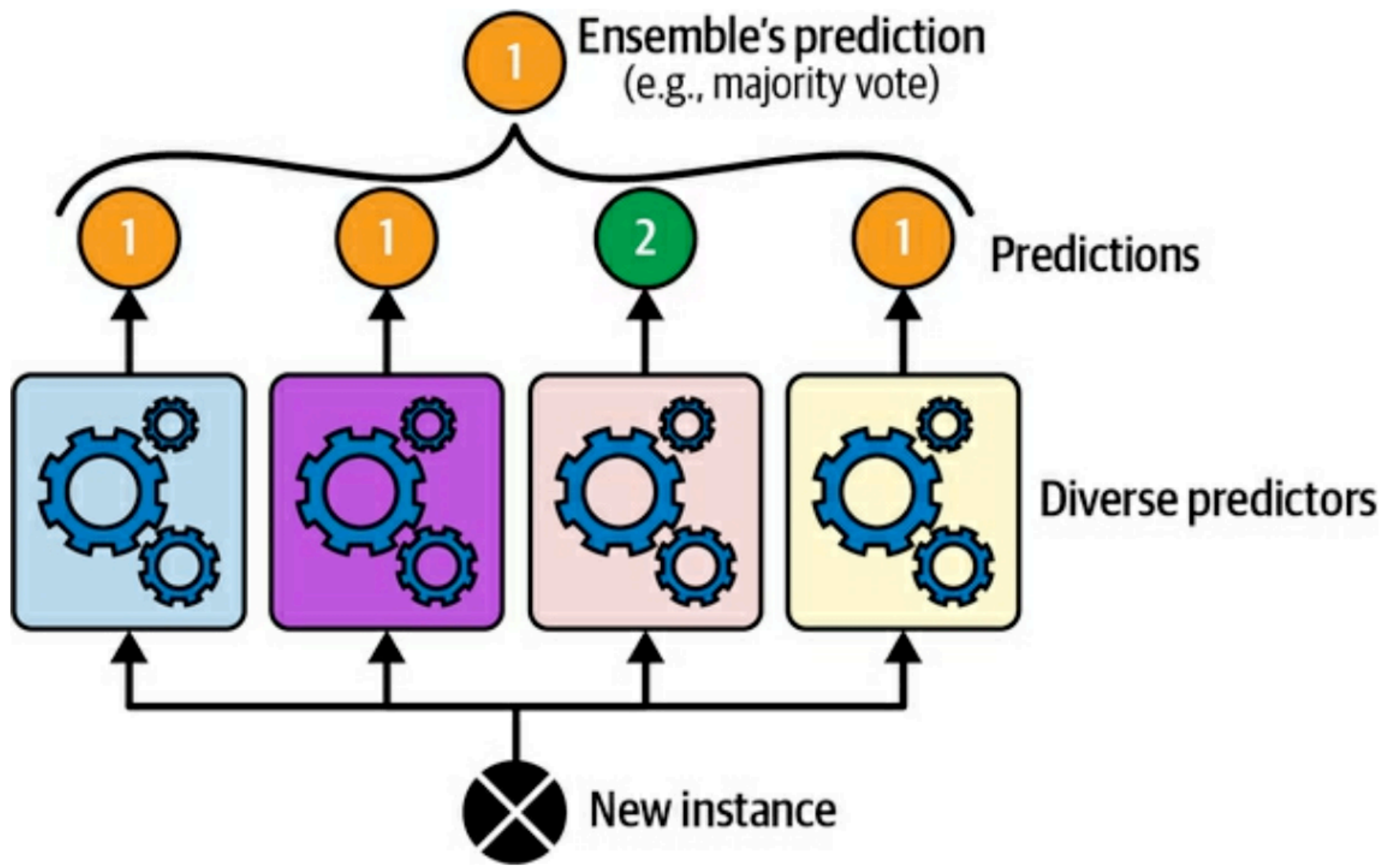
Voting Classifiers

Diverse Classifiers



Hard Voting

- Often more accurate than the best classifier in the ensemble
- Like measuring more data to reduce noise
- Works best if the predictors are independent
    - Not making the same errors



Coin Tosses

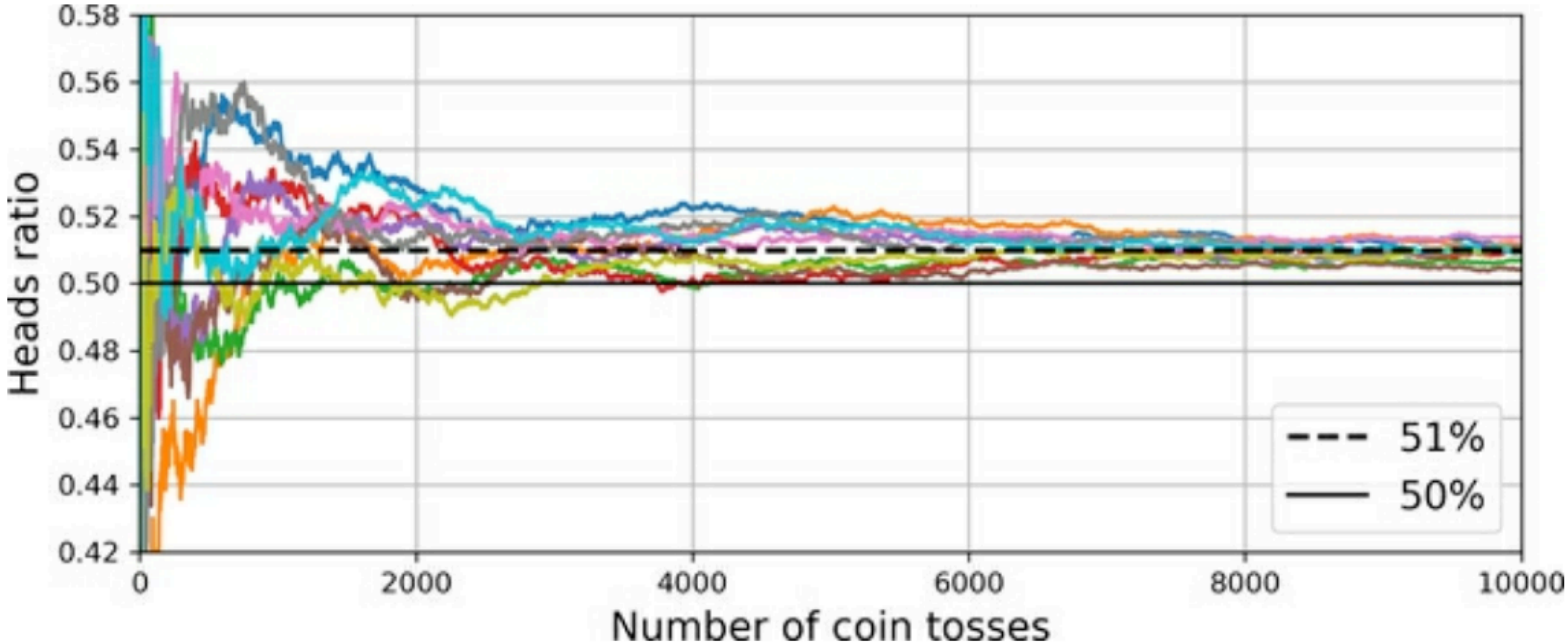


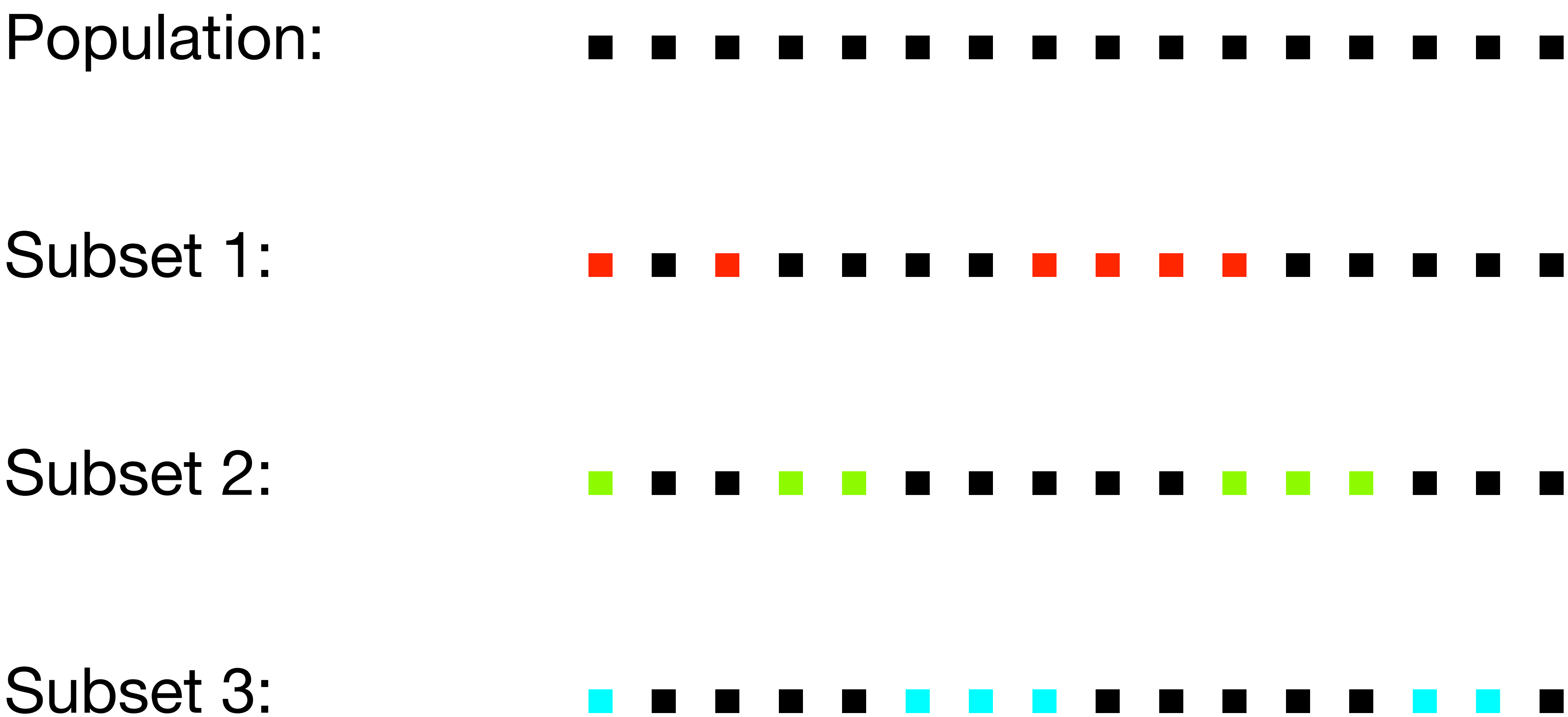
Figure 7-3. The law of large numbers

Bagging and Pasting

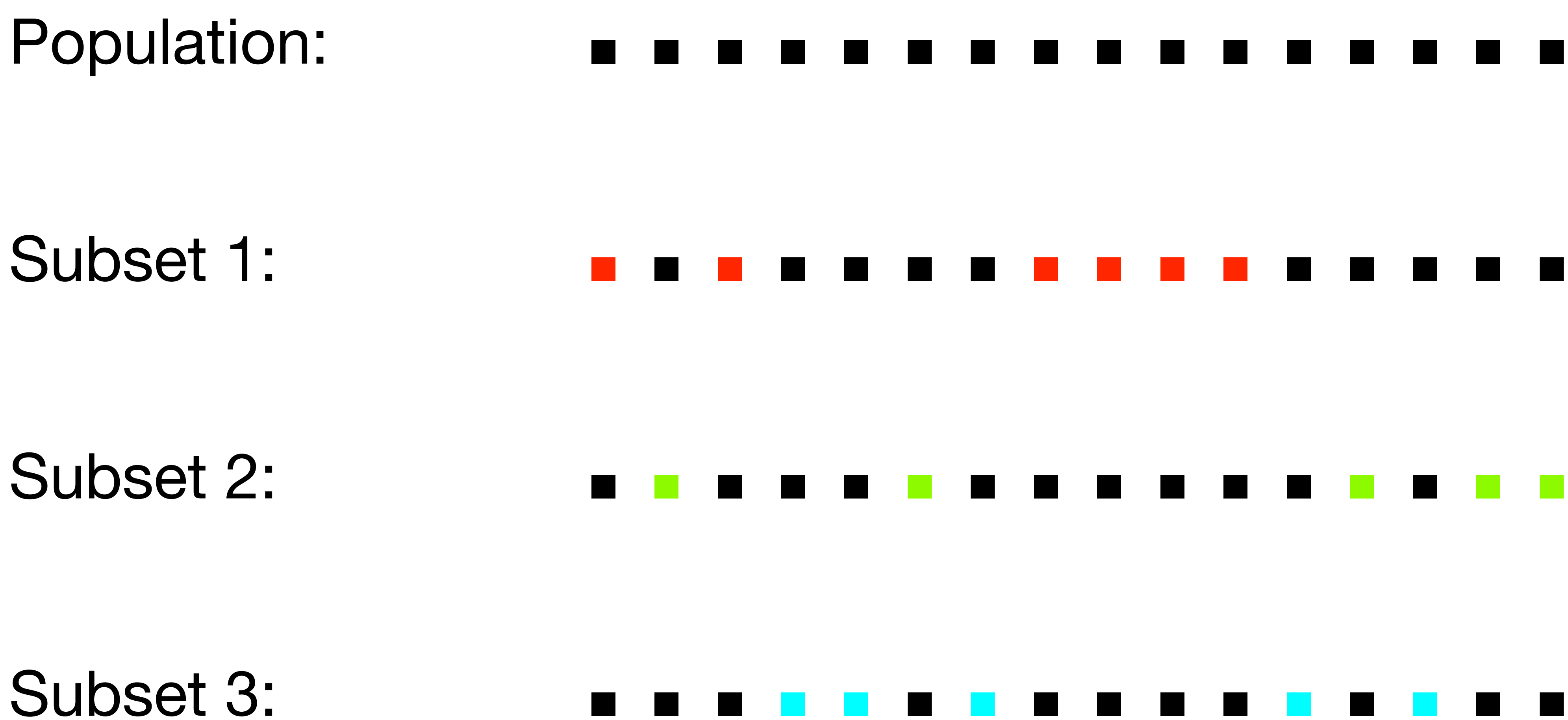
Achieving Diversity

- Use different training algorithms, or
- Use same algorithem every time, but
  - Train on different subsets of the same data
- **Bagging** (short for ***bootstrap aggregating***)
  - Sampling with replacement
- **Pasting**
  - Sampling without replacement

Sampling With Replacement "Bagging"



Sampling Without Replacement "Pasting"



Ensemble

- Predictors can be trained in parallel

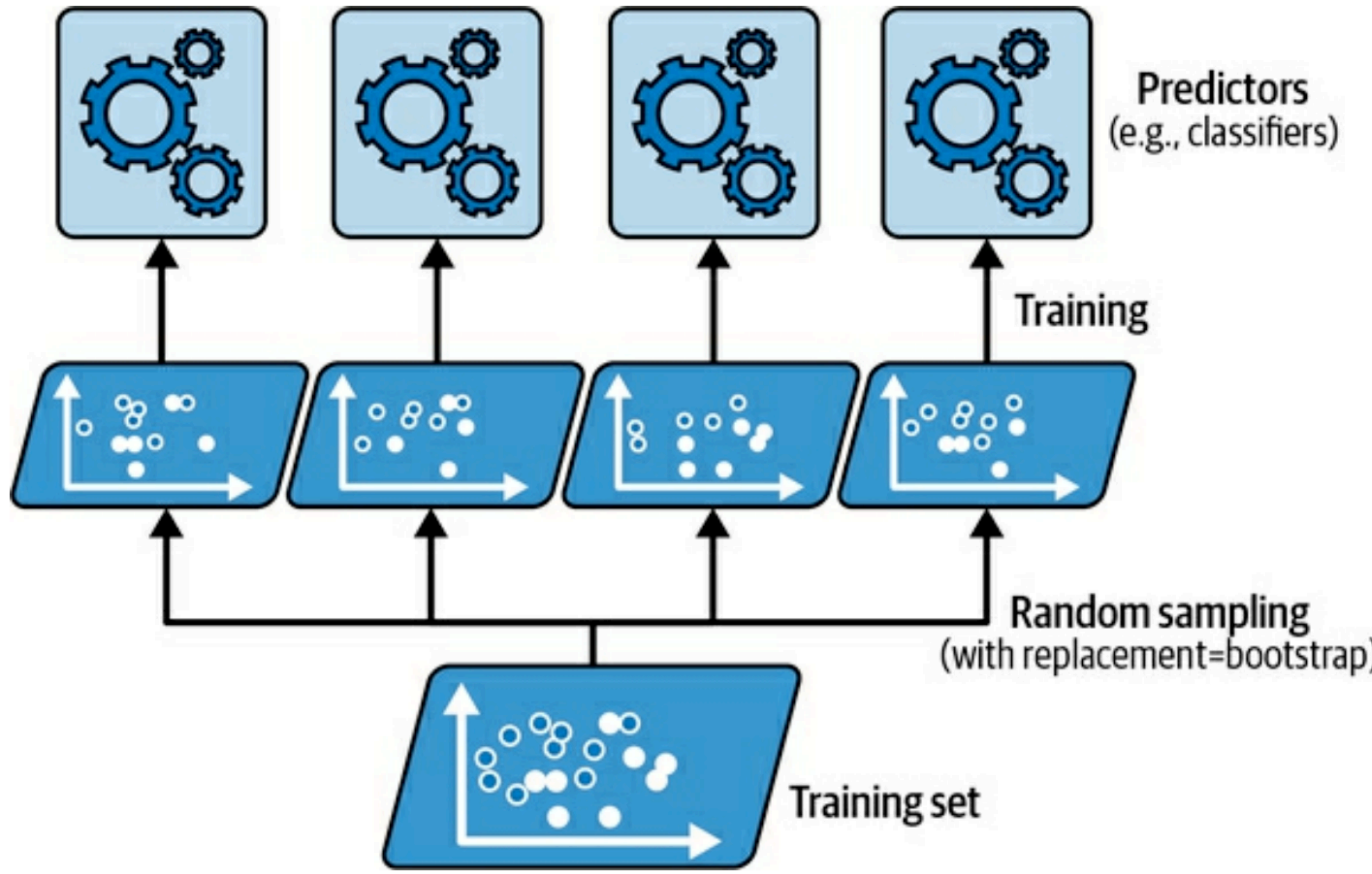
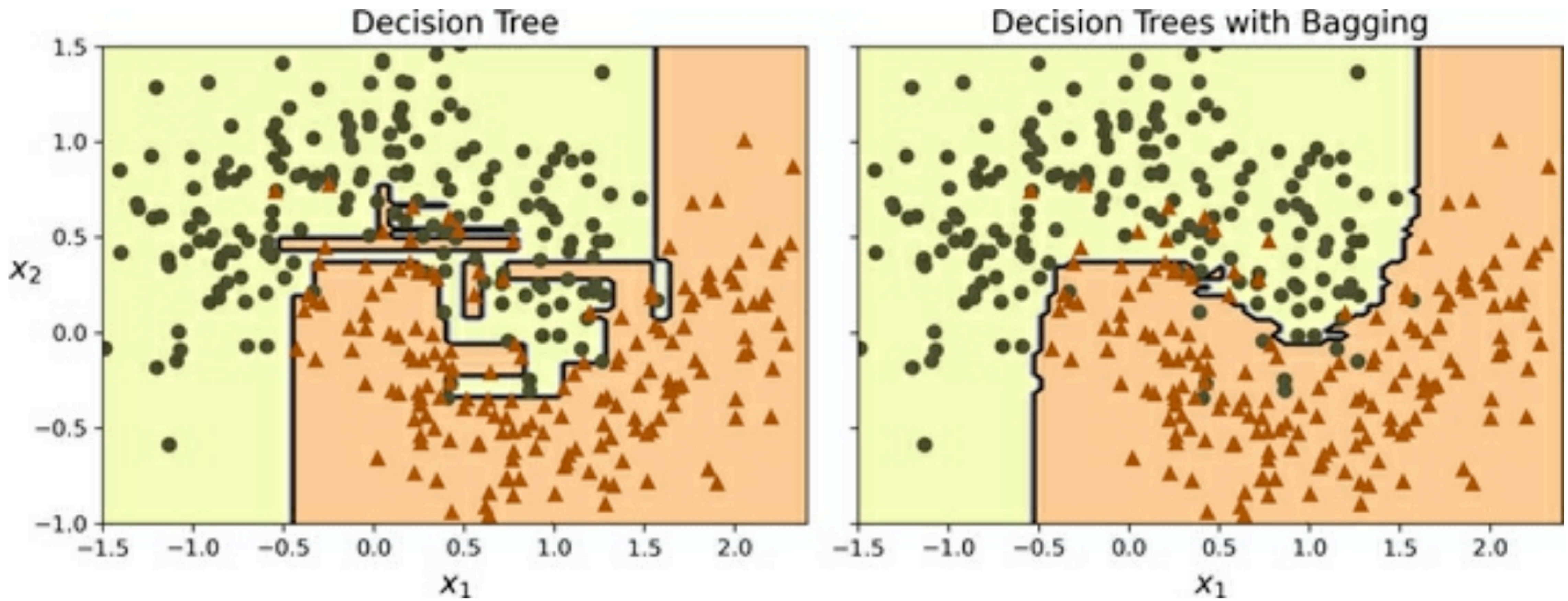


Figure 7-4. Bagging and pasting involve training several predictors on different random samples of the training set

Ensemble of 500 Decision Trees

- x



Bagging Statistics

- Training set contains ***m*** instances
- Each predictor draws ***m*** with replacement
  - So it only uses 63% of the samples
  - Drawing some samples twice or more times
- The remaining 37% not uses are called ***out-of-bag*** (OOB)
- You can use them as the test set

Random Patches and Random Subspaces

- **Random patches**
  - Sample both training instances and features
- **Random subspaces**
  - Keep all training instances but sample features

Random Forest

Random Forest

- An ensemble of decision trees
    - Generally trained by bagging
        - With ***m*** samples from a training set of ***m*** instances
- Uses a random sqrt(***n***) sample of the ***n*** features
    - To increase tree diversity

Extra-Trees

- Uses a random threshold value for each node
    - Instead of searching for the best possible threshold
- This forest is called **extremely randomized trees**
    - or **extra-trees**
- Increases variance and makes training much faster
- Sometimes extra-trees perform better, not always

Feature Importance

- Examine a random forest
    - Look at how much nodes using a feature reduce impurity
    - Averaging across all trees in the forest

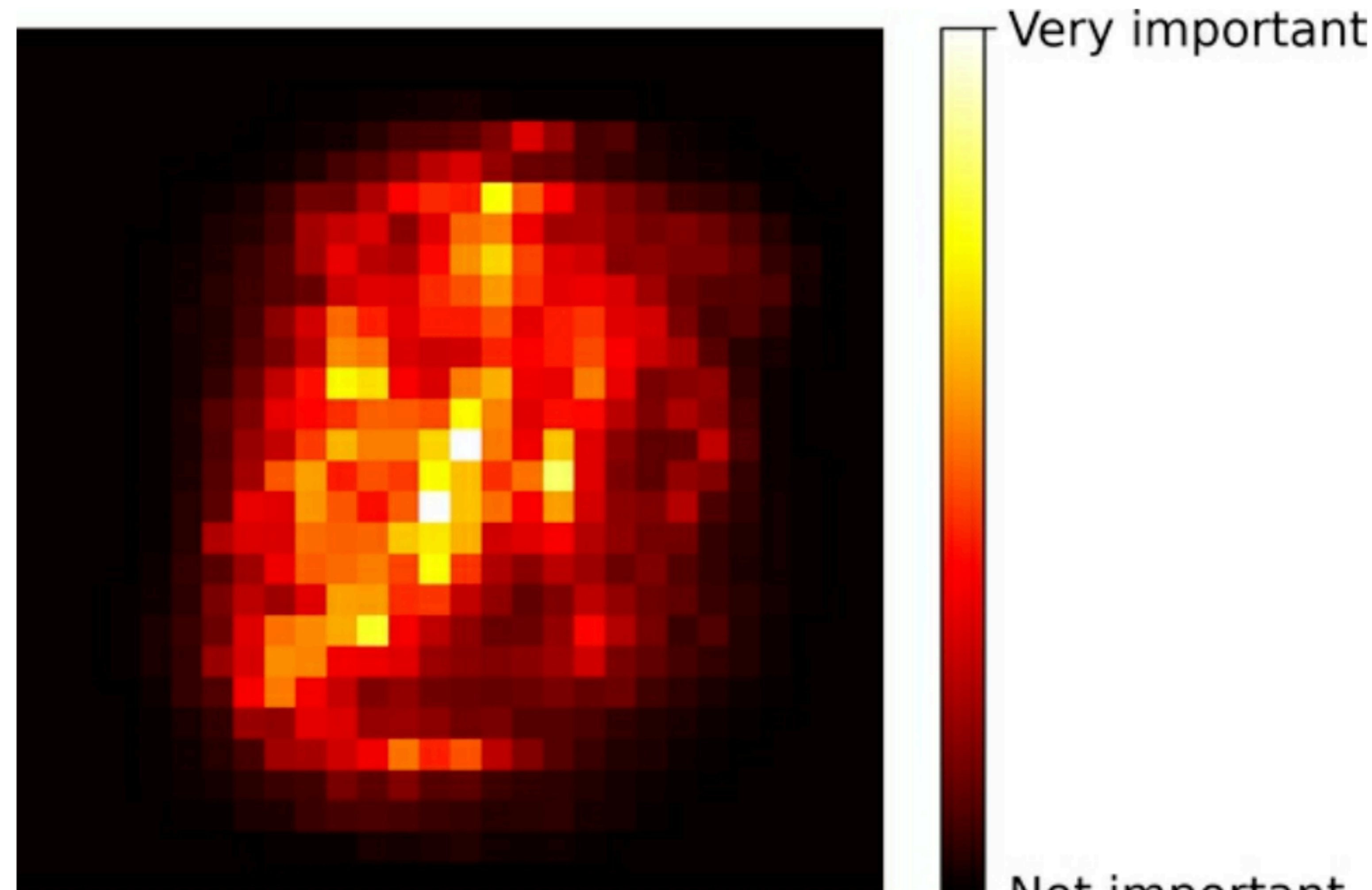


Figure 7-6. MNIST pixel importance (according to a random forest classifier)

Kahoot!

Ch 7a

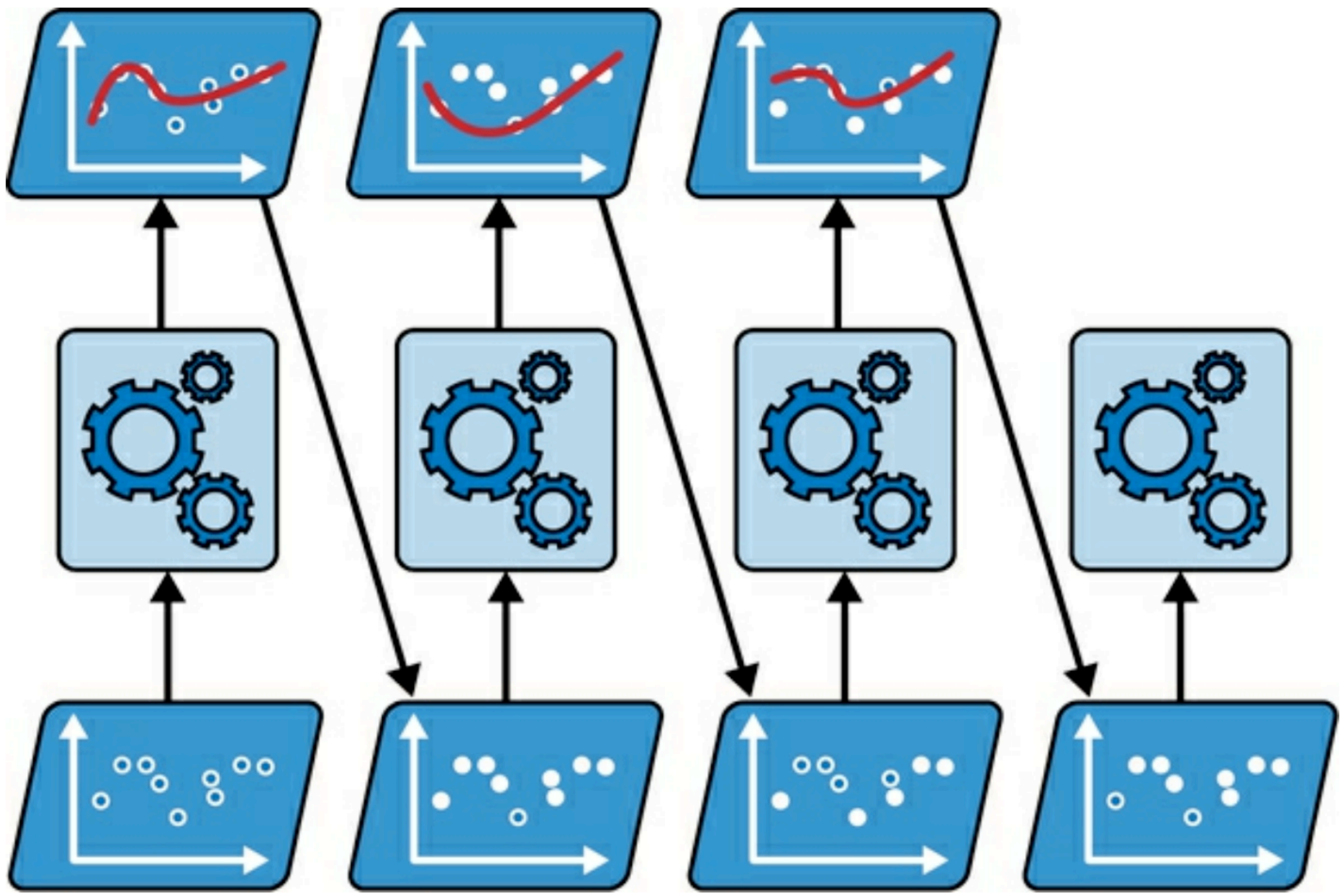
Boosting

Boosting

- Originally called *hypothesis boosting*
- Any ensemble that combines weak learners into a strong learner
- Train predictors sequentially
    - Each trying to correct its predecessor
- Two popular methods
    - ***AdaBoost*** (adaptive boosting)
    - ***Gradient Boosting***

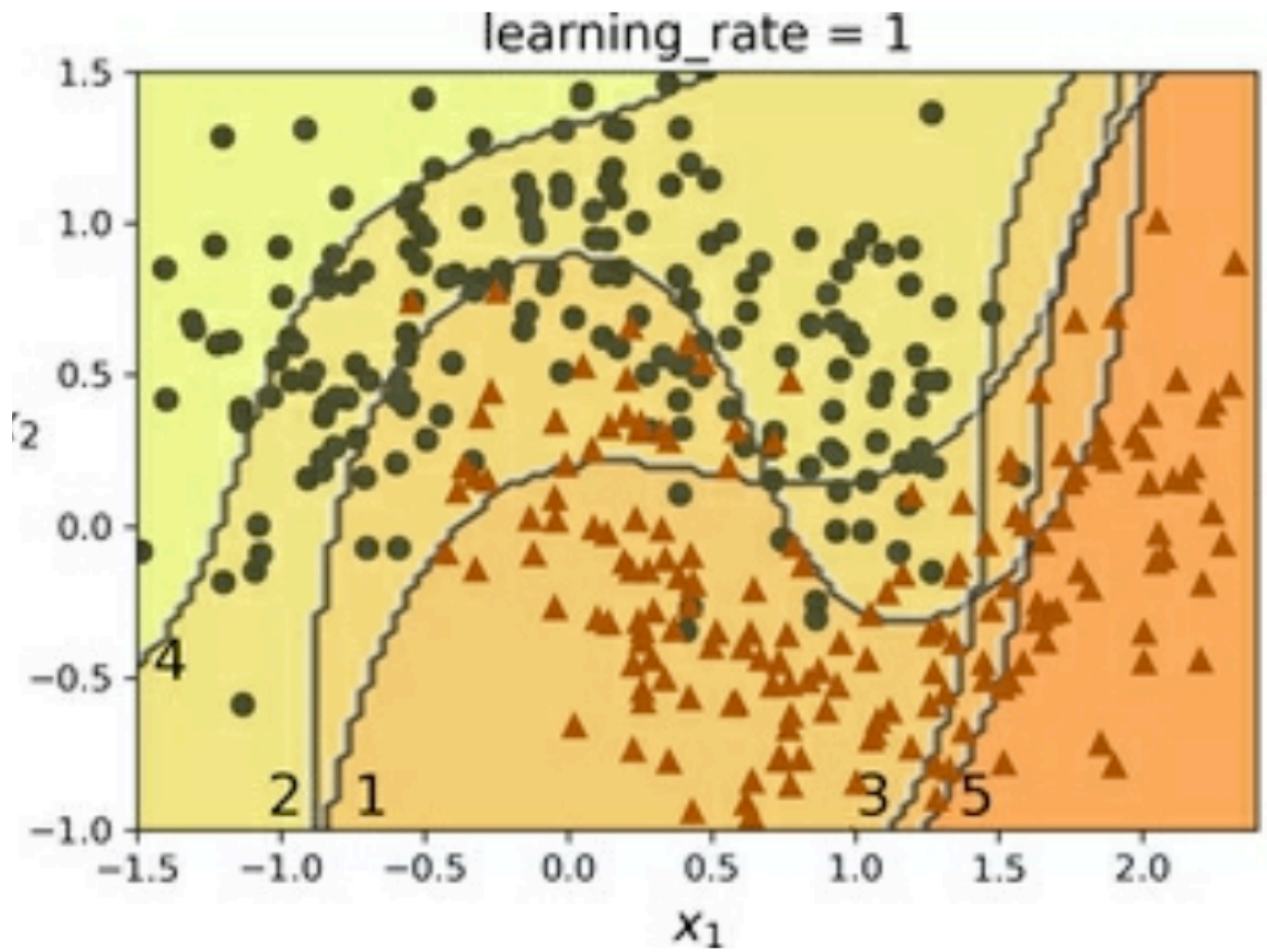
AdaBoost

- Each predictor pays more attention to the training instances its predecessor underfit



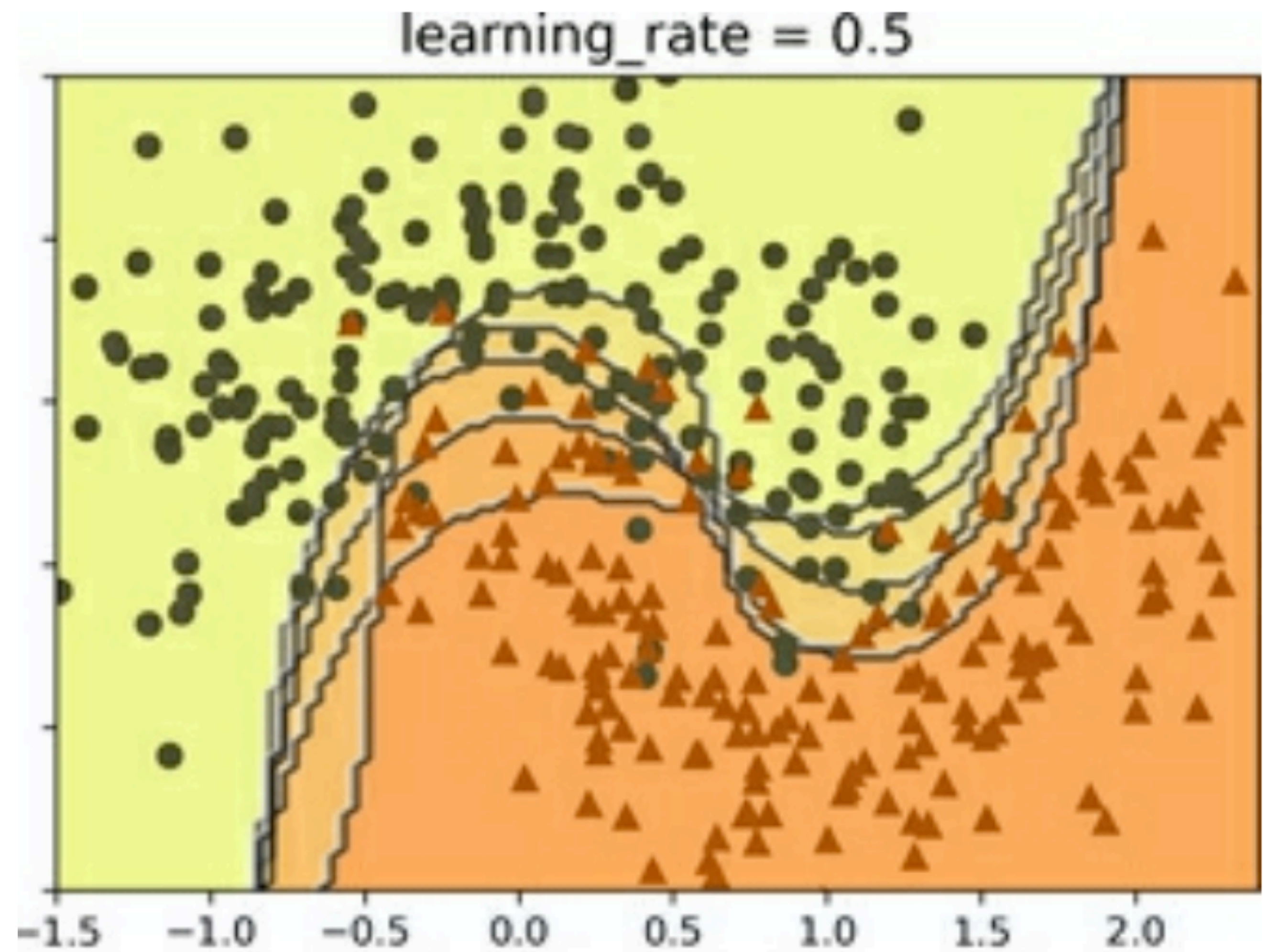
Decision Boundaries

- Models jerk from one set of instances to another
- As in the previous slide



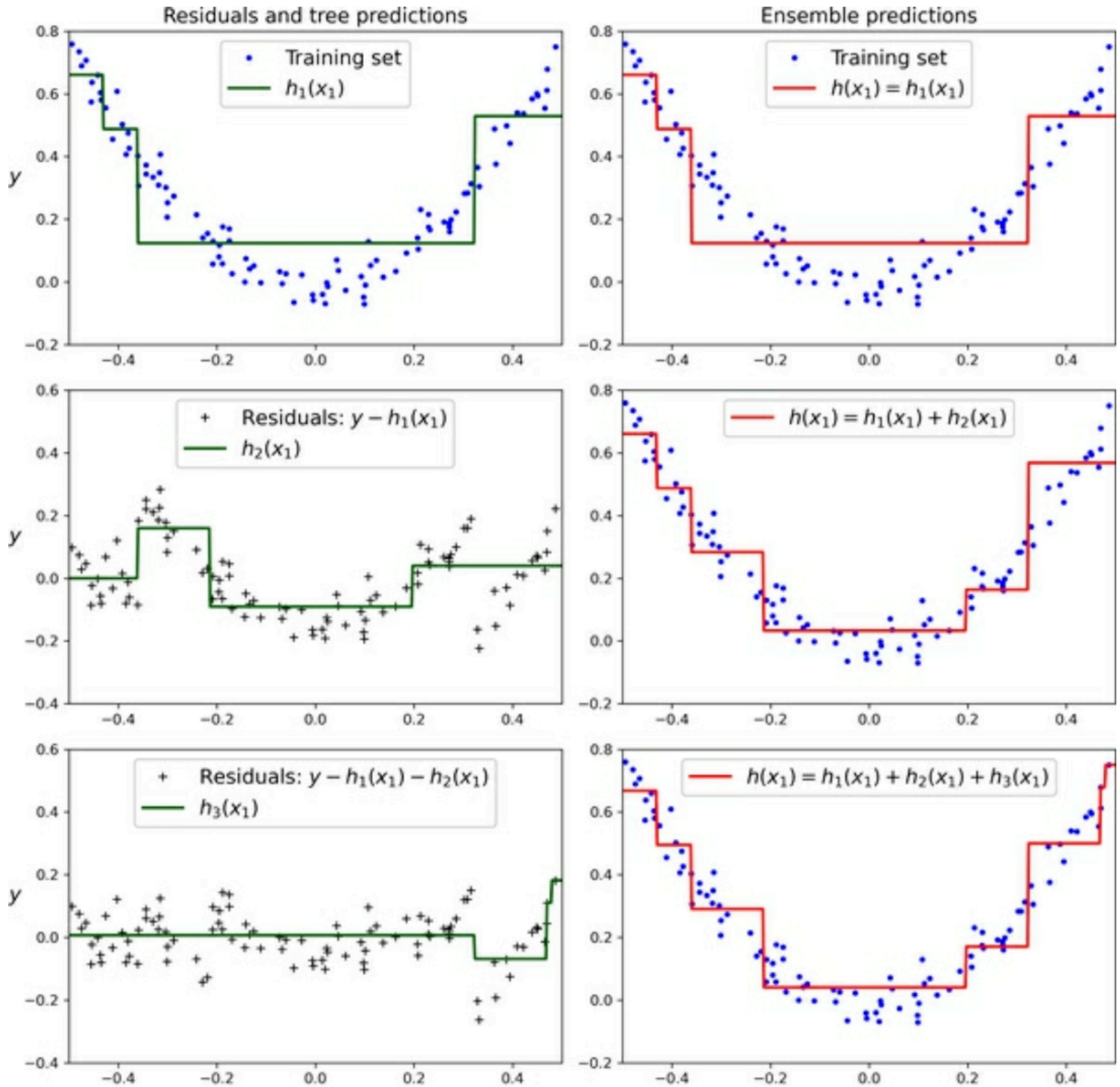
Decision Boundaries

- With slower learning, the AdaBoost model converges to a good fit
- Like gradient descent



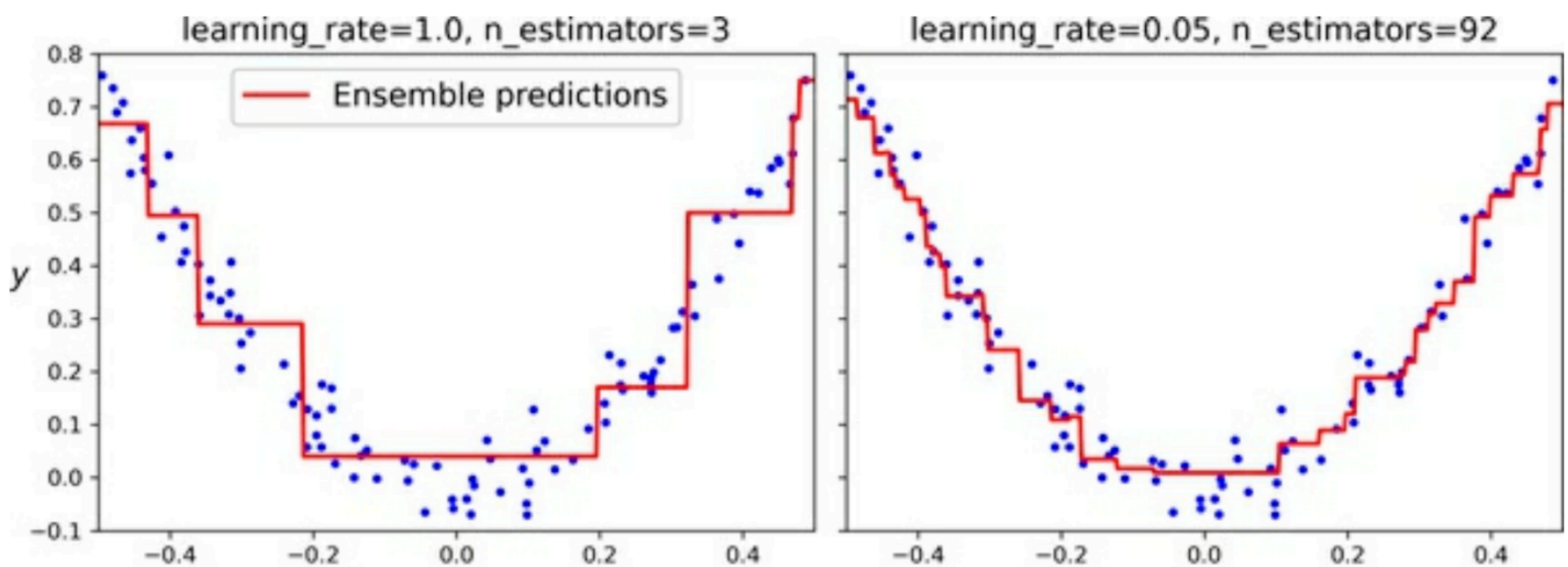
Gradient Boosting

- Tries to fit each predictor to the ***residual errors*** of its predecessor



Hyperparameters: Learning Rate and Number of Trees

- Low learning rate requires more trees, but generalizes better
  - This regularization technique is called ***shrinkage***
- Early stopping helps to find the best number of trees
  - Hyperparameter **n_iter_no_change** set to a value, such as 10
  - Stop when the last 10 trees didn't help



Histogram-Based Gradient Boosting

- Optimized for large datasets
- Bins the input features into ***b*** bins (<=255)
  - Replacing them by integers
- Greatly reduces the number of threshold values to explore
- Can use more efficient integer data structures
- Makes training much faster (hundreds of times faster)
- Computational complexity $\boldsymbol{O(b \times m)}$ instead of $\boldsymbol{O(n \times m \times \log(m))}$
  - ***n*** features, ***m*** instances
- Precision loss acts as a regularizer

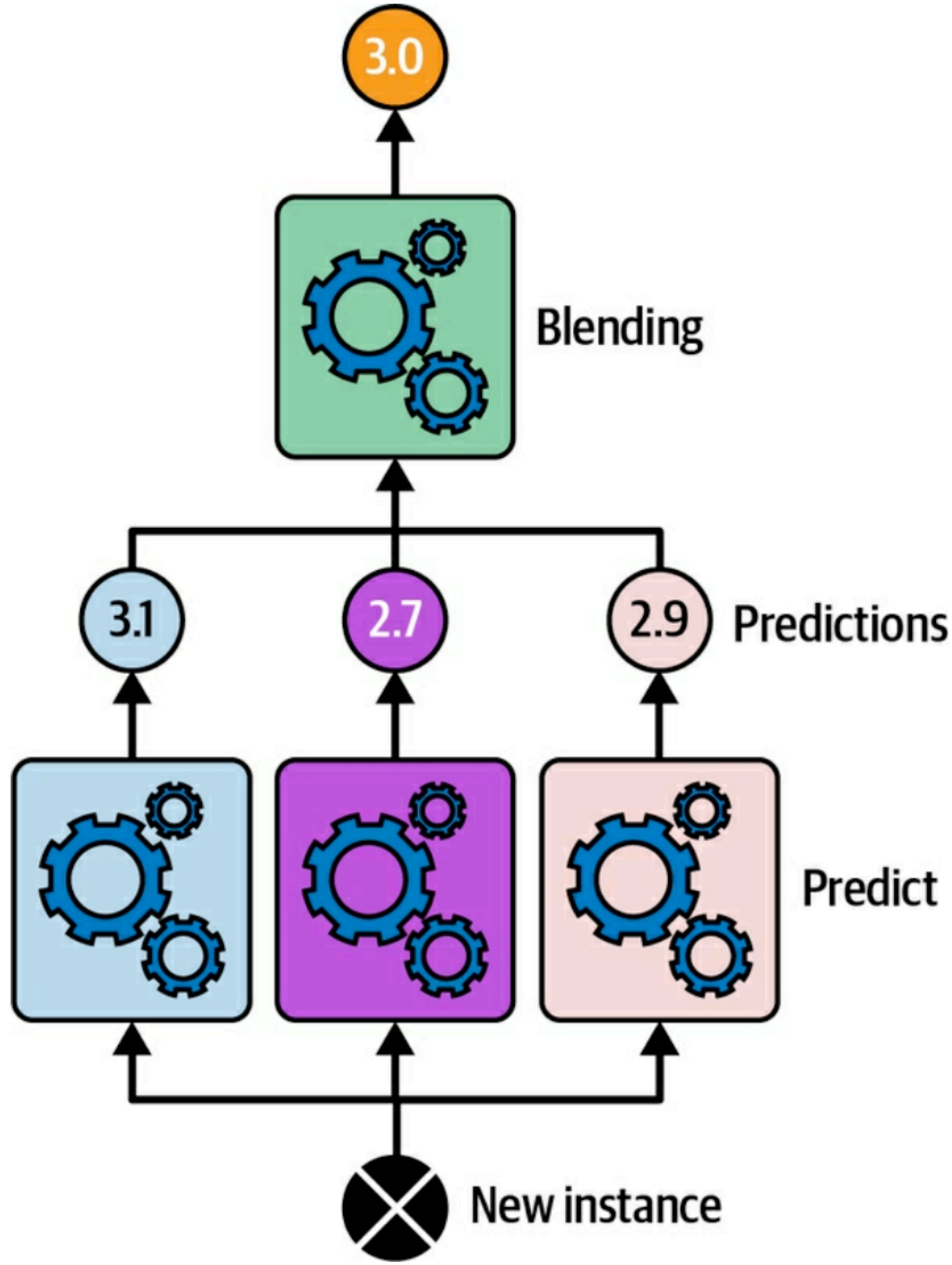
Stacking

Stacking

- Short for ***stacked generalization***
- Instead of using trivial functions (like hard voting)
  - To aggregate the predictions in an ensemble
- Train a model to perform aggregation
- Final predictor is called a ***blender*** or a ***meta learner***

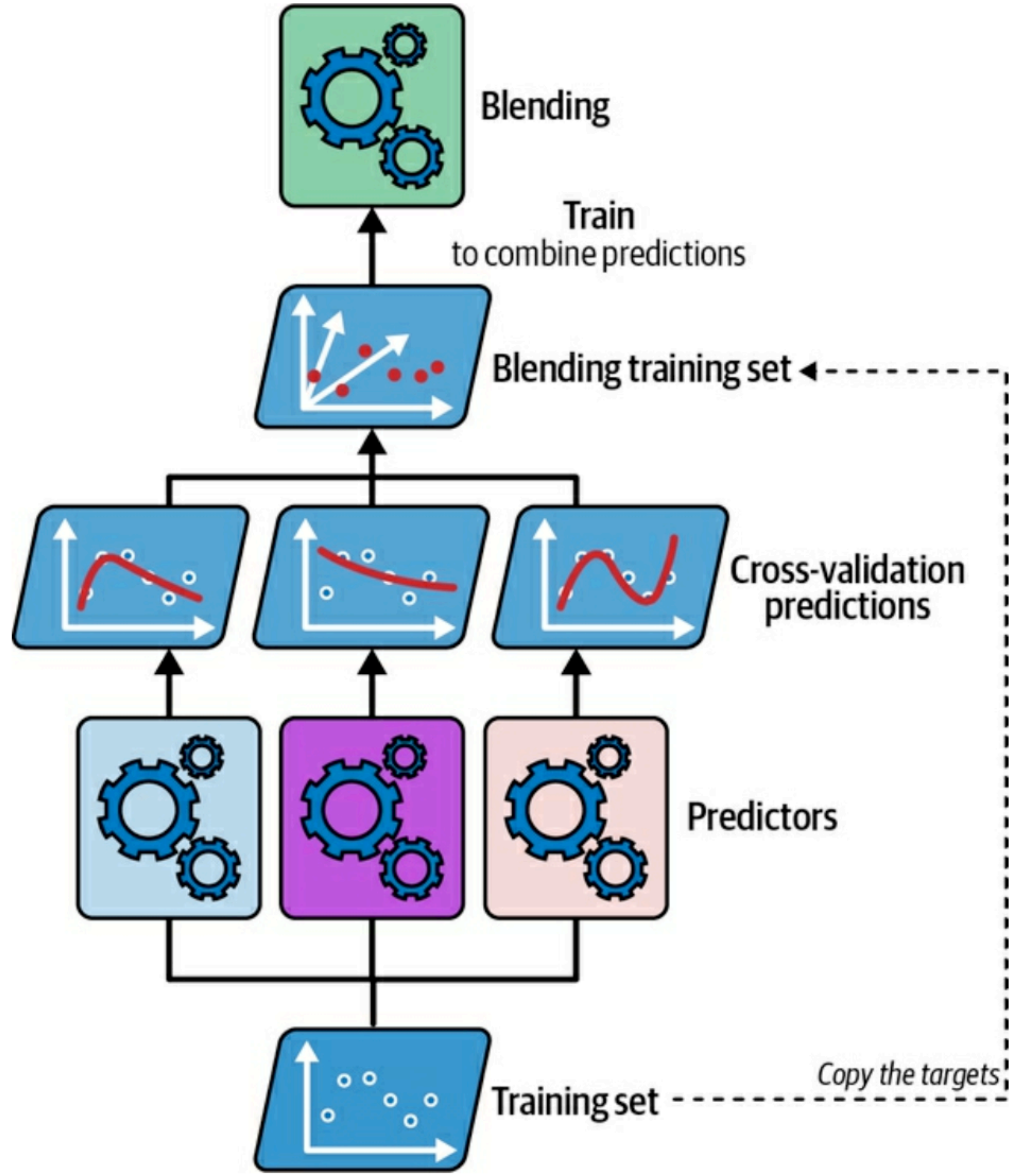
Stacking

- Regardles of how many features are input to the predictors
- They only have one output value each



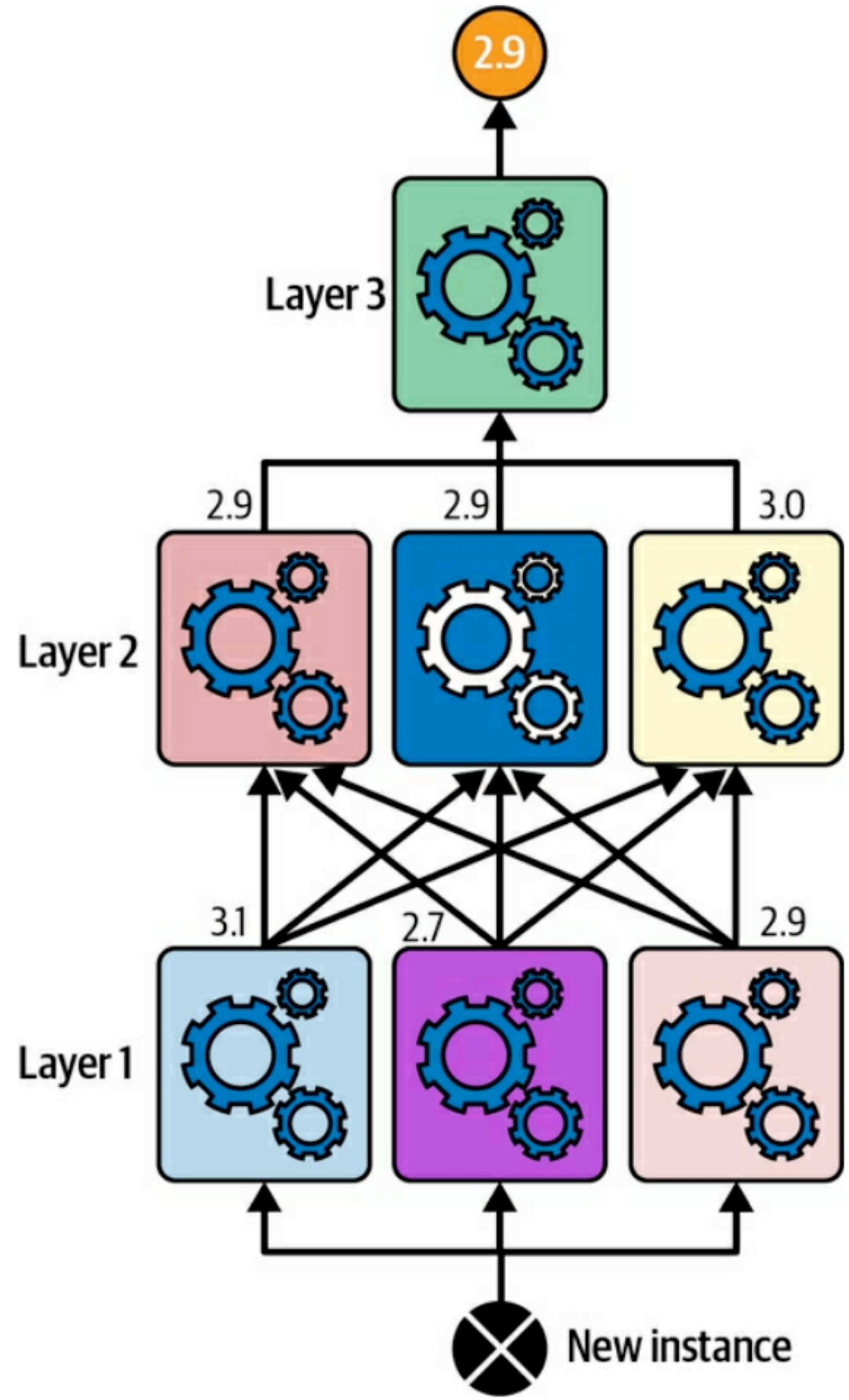
Training the Blender

- Use cross-validation to make predictions
- Feed those predictions into the blender



Multilayer Stacking

- Two layers of blenders
- May perform better
- But increases training time and system complexity



Kahoot!

Ch 7b