

Biostat 202: Opportunities and Challenges of “Big Data”:

Machine Learning 4:

Cross-Validation, Ensembles, and Feature Importance

Aaron Wolfe Scheffler

Division of Biostatistics

Department of Epidemiology and Biostatistics

Outline

- Cross-validation
- Ensemble methods
 - Bagging: *Random Forests*
 - Boosting: *Adaboost*
 - Stacking
- Parameter tuning, continued
- Feature importance

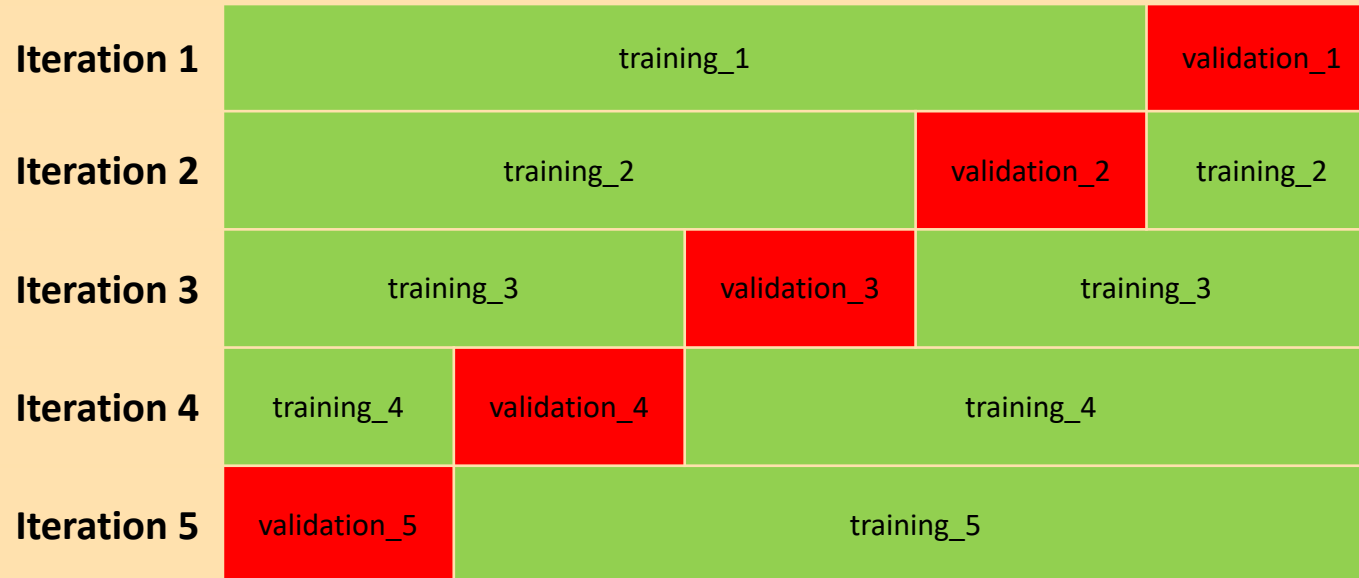
Cross-validation

E.g. 5-fold cross-validation

- 1. Break up data into 5 “folds” of the same size



- 2. Set aside one fold for validation and use the rest to build model, repeat this process for each fold

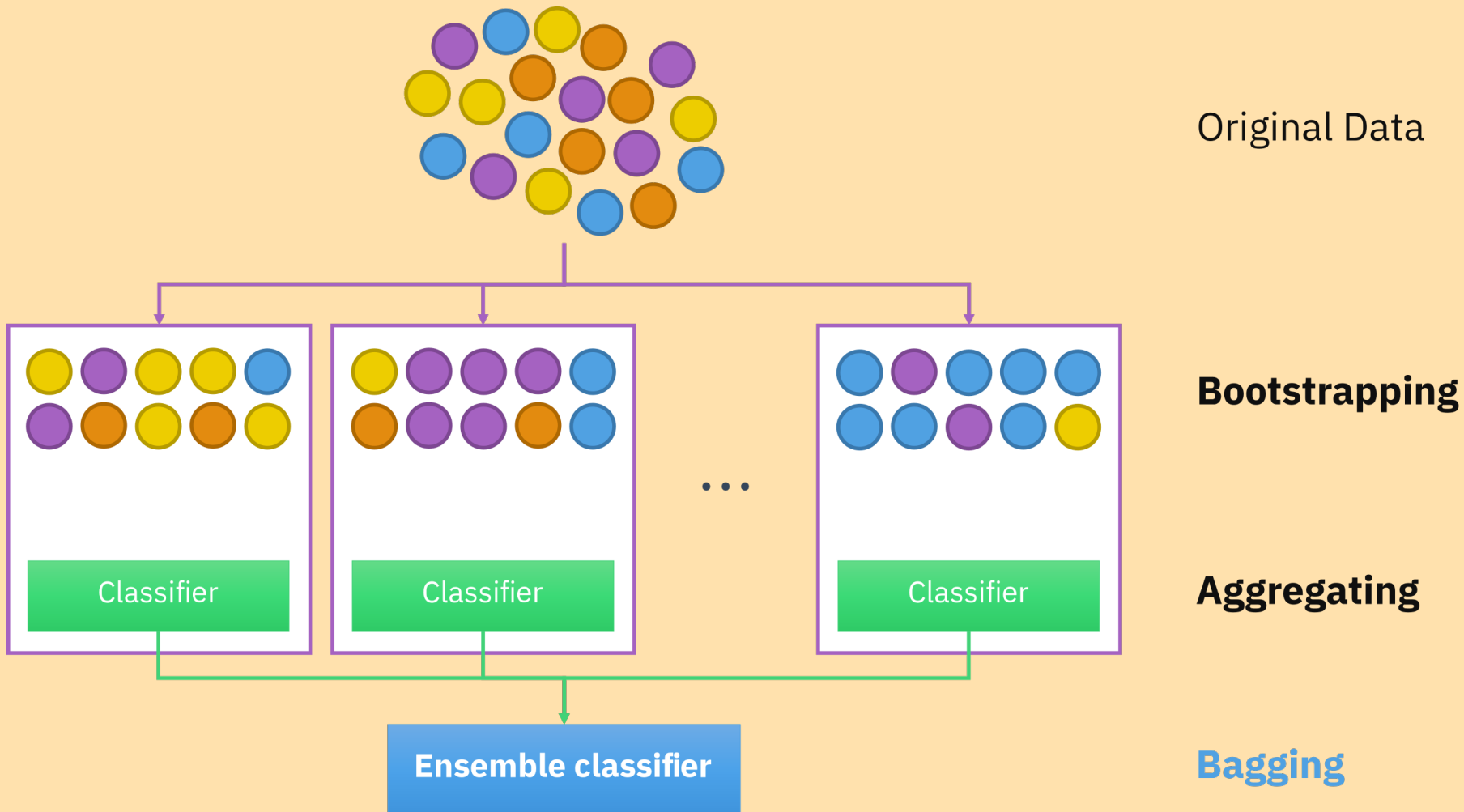


- Estimate the predictive performance of our model on the combined validation data (i.e. the validation_1-5 sets)

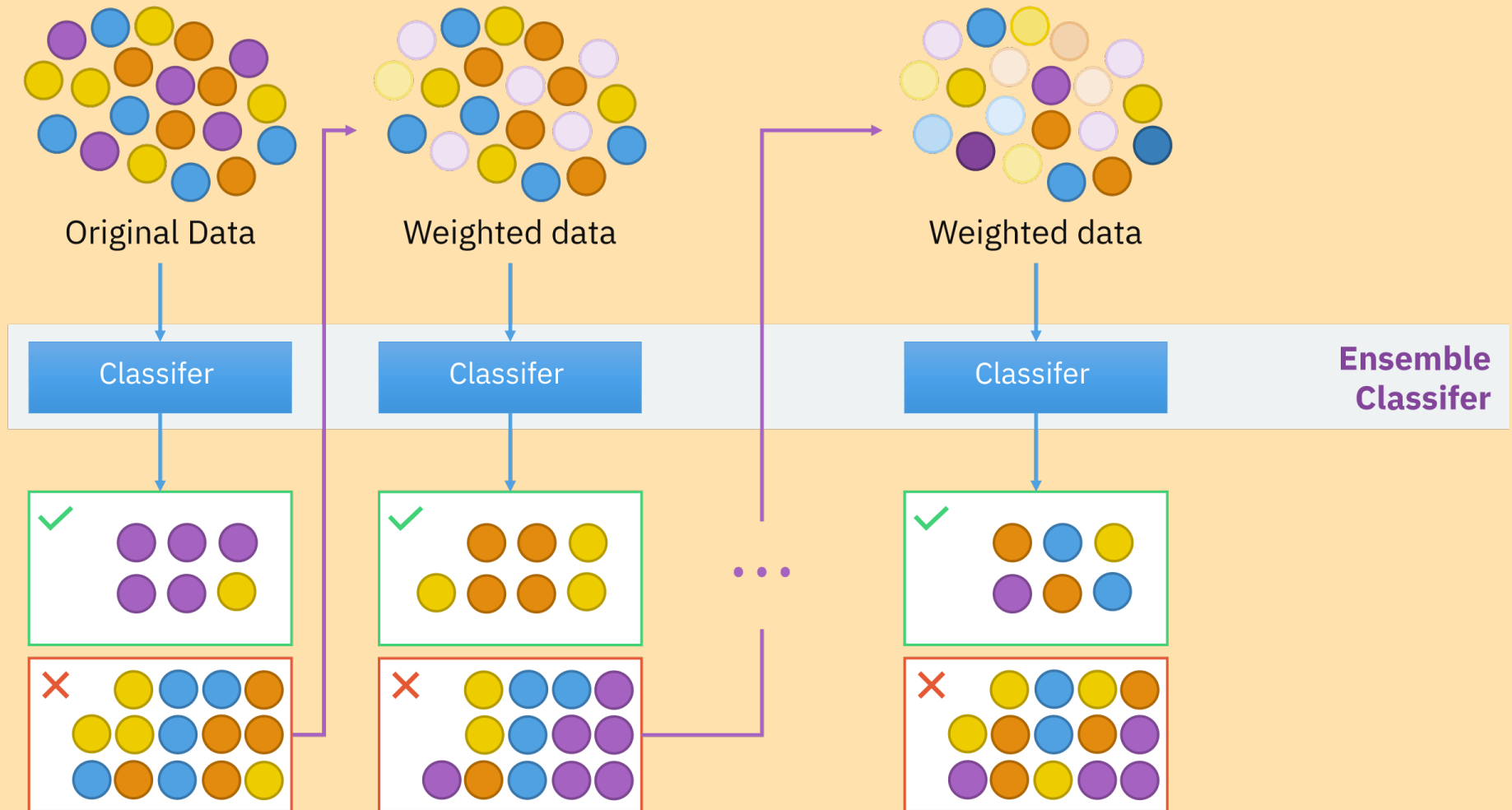
Ensemble methods

- Train a bunch of predictive models and pool them to form a final prediction.
- **Advantage:**
 - improvement in predictive accuracy
- **Disadvantages:**
 - more computation
 - it is difficult to understand an ensemble of classifiers
- Examples: bagging, boosting, and stacking

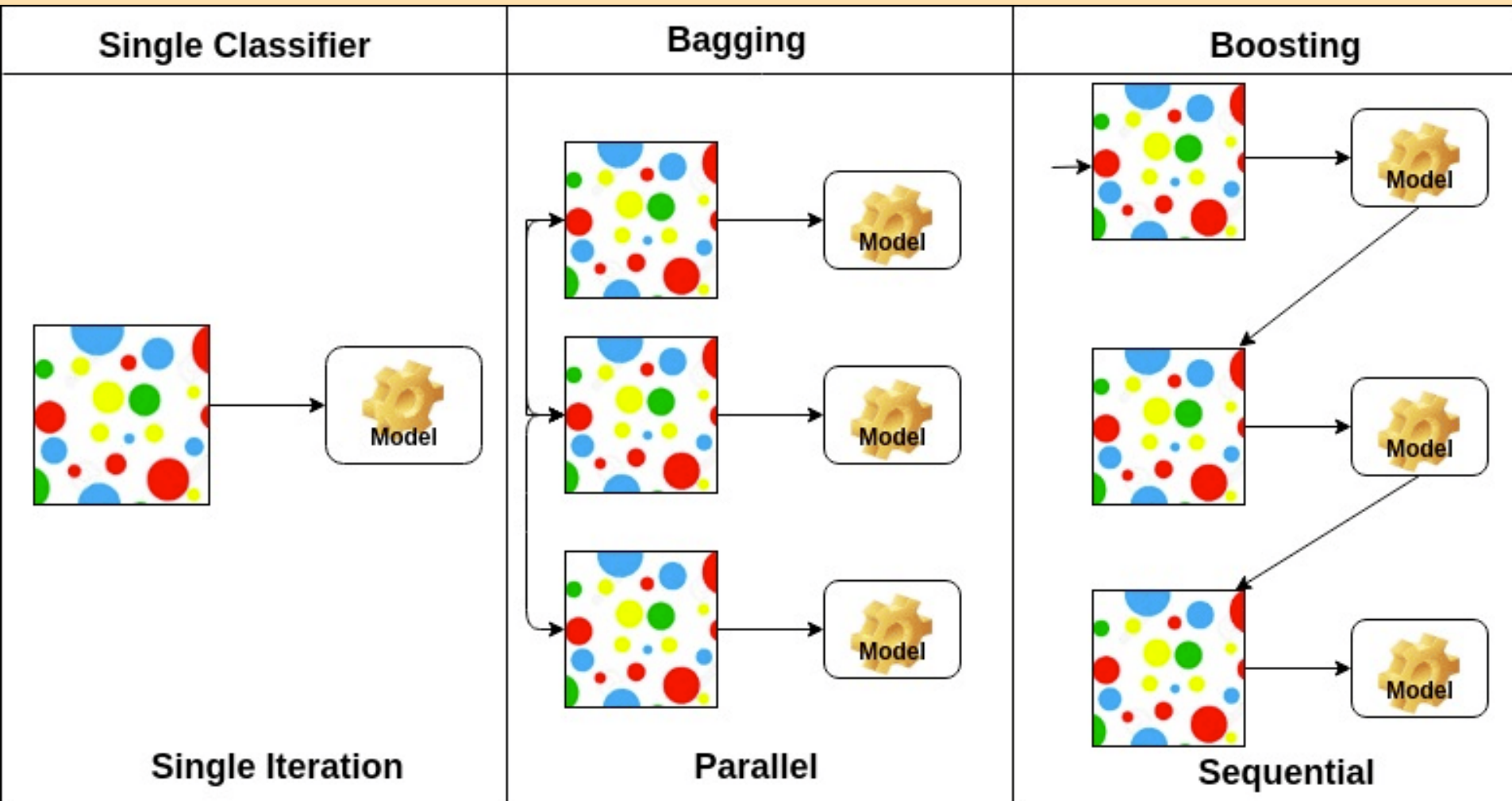
Bagging



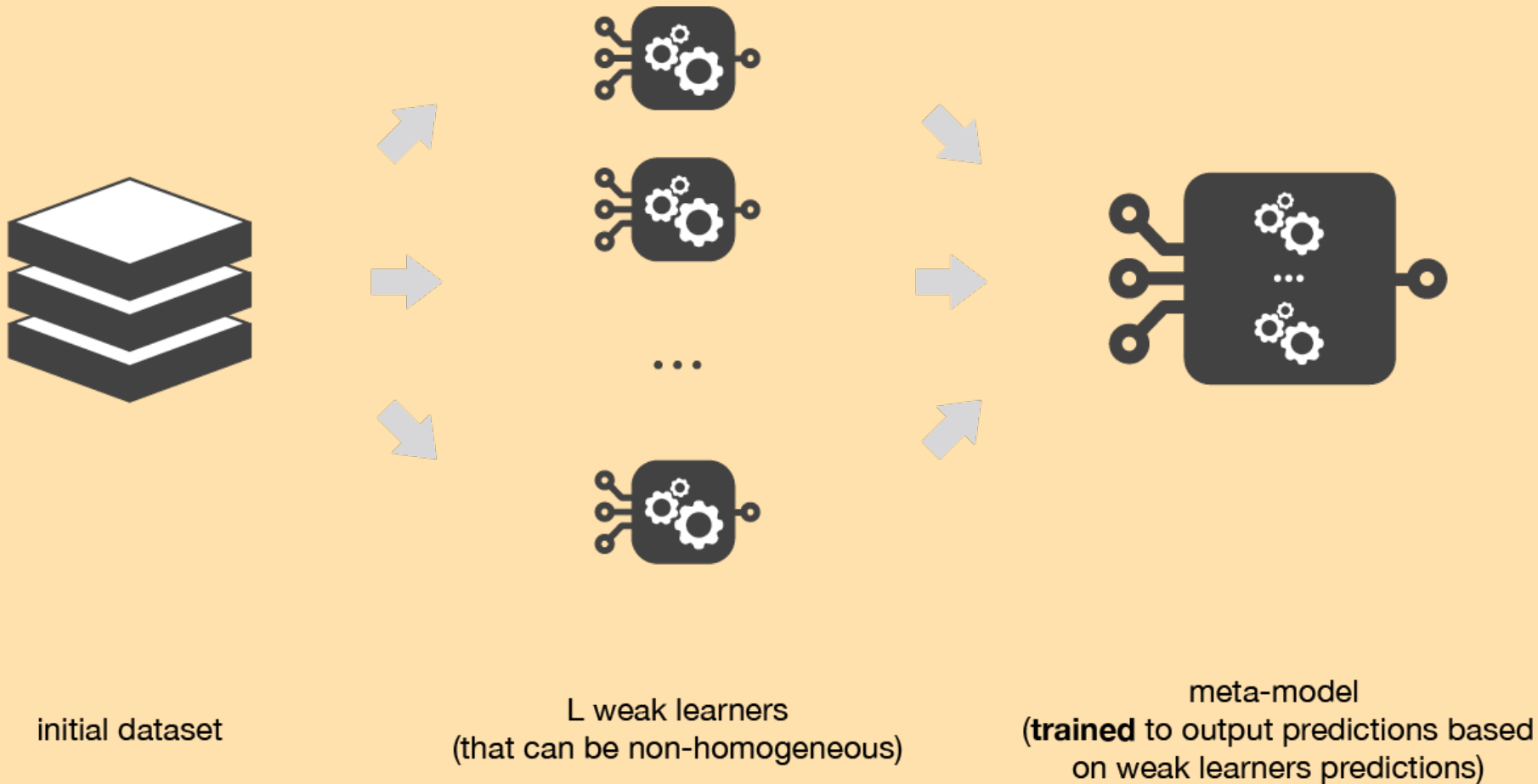
Boosting



Boosting vs Bagging



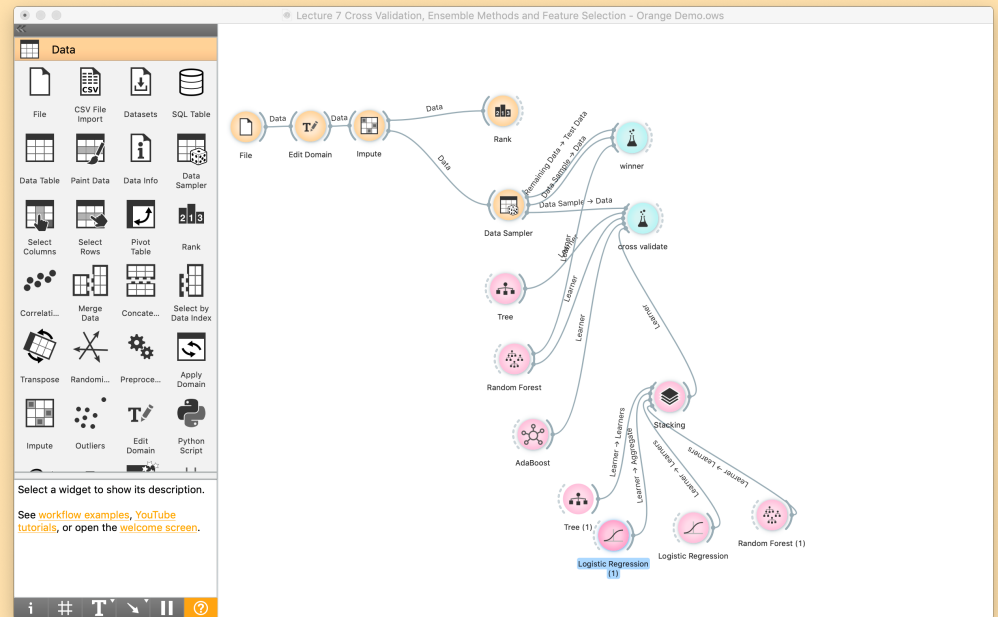
Stacking



Feature importance

- Even though we fool ourselves into only being concerned with prediction, we often want to understand which features impact our prediction
- Given a target variable, which variables are important for prediction?
- To define this, you can develop measures of **feature importance**
- Feature importance can either be pair wise or model based
 - **Pairwise:** based directly on measures of association between target and individual features (*e.g. correlation*)
 - **Model based:** based on machine learning algorithms (*e.g. regression coefficients, classification tree splits*)

Orange Demo of Topics



Next time

- Clustering
 - k-means clustering
 - TwoStep clustering
 - Kohonen self-organizing map (SOM)
- Data reduction
 - Principal components analysis (PCA)
 - t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Guidance for choosing machine learning algorithms

Lab- Prediction Competition



Predict time to readmission
among diabetic patients



Prepare reproducible
models



Collaborate as teams,
submit work individually



You will have one hour
(max) to perform this task