# 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**

# Announcements

- Lab will be completely remote today, see CLE for ZOOM link (same link as class)
- Stick around, this lab is MUCH easier when done with a group (and more fun)

# Outline

- Cross-validation
- Ensemble methods
  - Bagging: *Random Forests*
  - Boosting: *Adaboost*
  - Stacking
- Feature importance

# Cross-validation

# Why cross-validation?

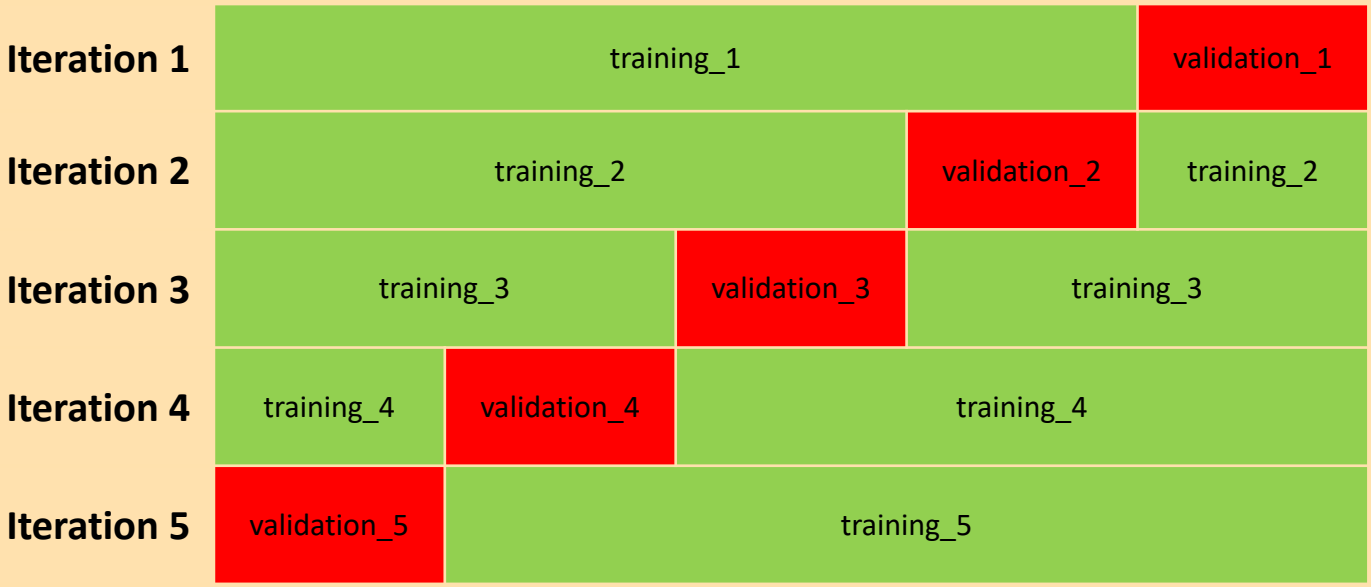
- The purpose of cross validation is to get an unbiased assessment of model performance for the purpose of model selection or hyperparameter tuning in small data settings
- Train/Validate/Test: 200 subjects, requires 100-50-50 split.
- Can I somehow maintain more data for model training, while still validating properly?
- Yes, cross-validation! Replaces the Train/Validate split but NOT the Test split

# E.g. 5-fold cross-validation

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

| Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|

- 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)

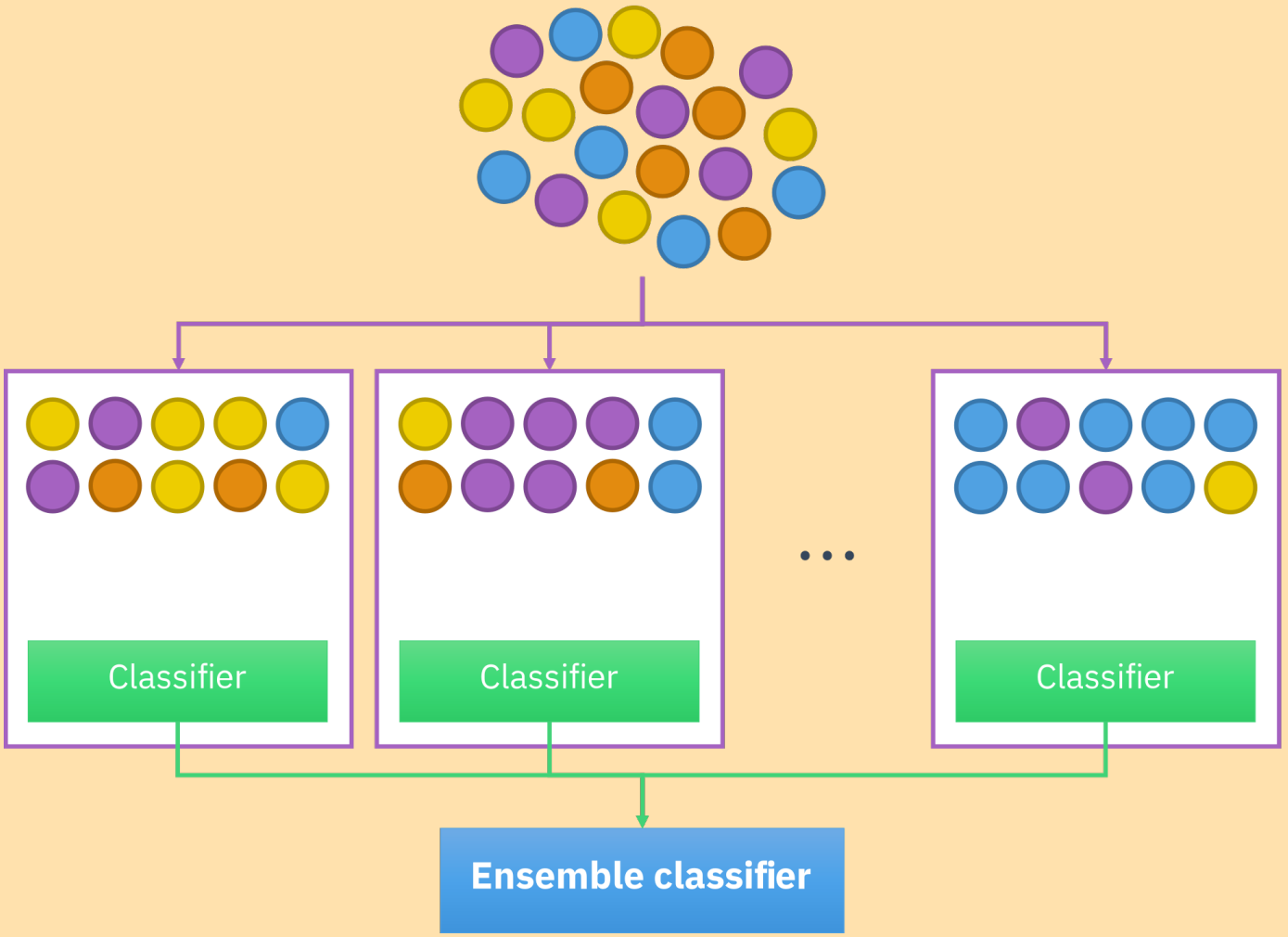
# Validation comparison

- **Train/Test**: to assess performance of a single model
- **Train/Validation/Test**: to compare multiple models for model selection or hyperparameter tuning (more data)
- **Cross-validation/Test**: to compare multiple models for model selection or hyperparameter tuning (less data)
- We always assess our final model choice on TEST data!!!

# 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



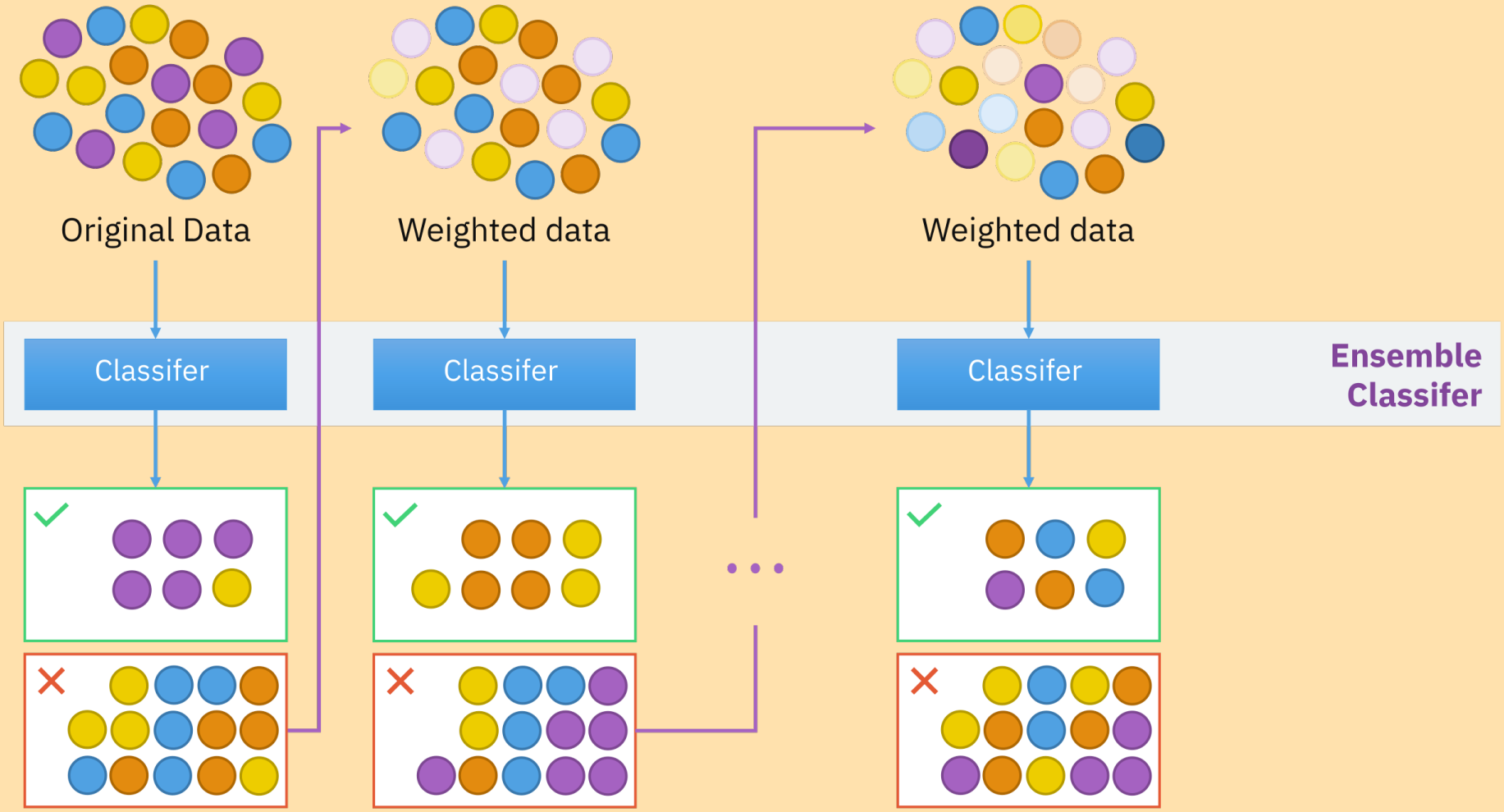

Original Data

**Bootstrapping**

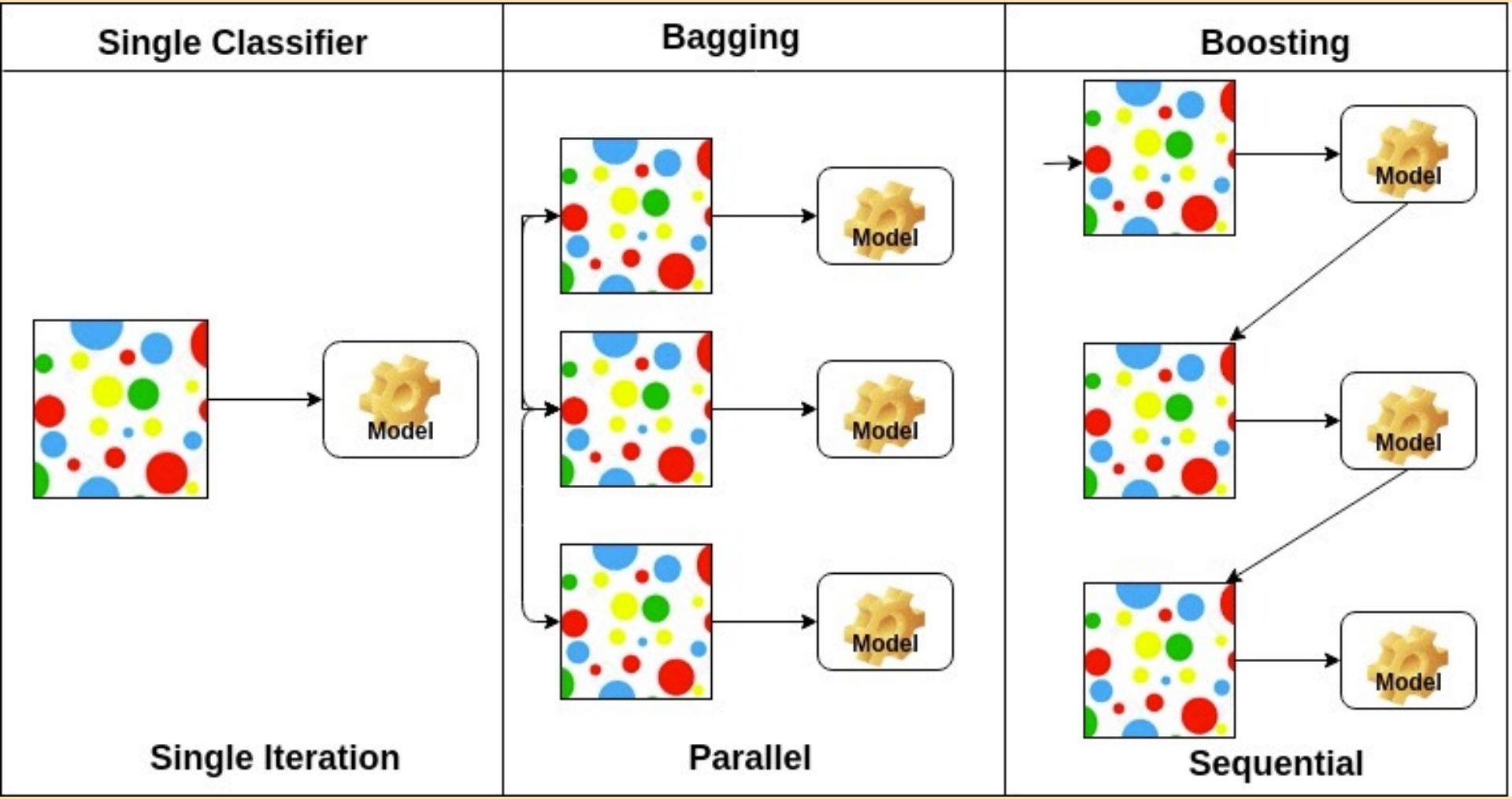
**Aggregating**

**Bagging**

# Boosting


Original Data
Weighted data
Weighted data
Classifer
Classifer
Classifer
Ensemble Classifer

# Boosting vs Bagging



# Stacking

initial dataset

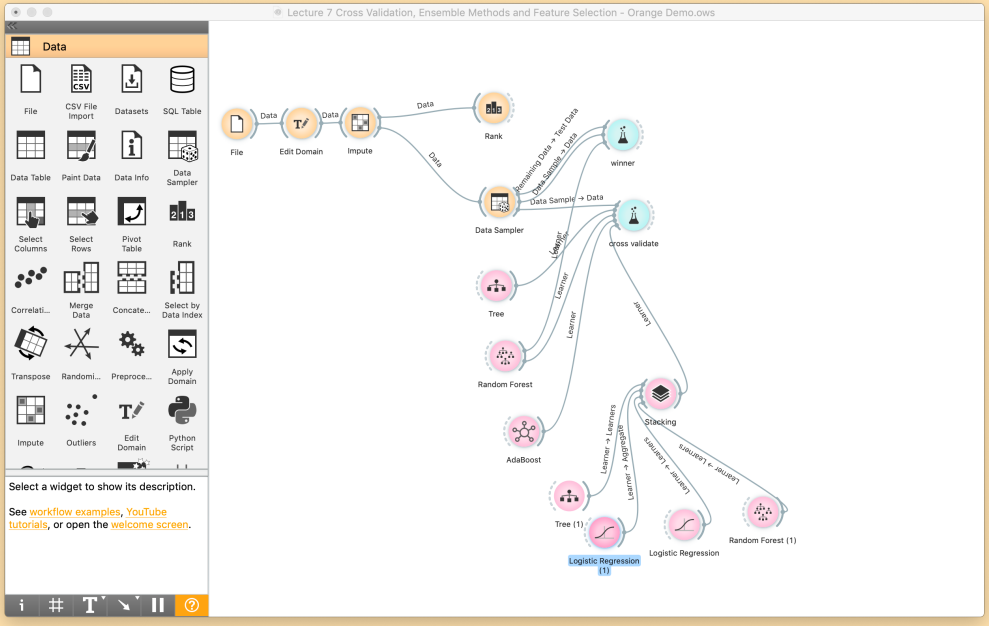
L weak learners
(that can be non-homogeneous)

meta-model
(**trained** to output predictions based on weak learners predictions)

# 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?
- The define this, you can develop measures of **feature importance**
- Feature importance
  - **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*)
  - **Permutation**: shuffle the data among subjects, one feature at a time, and track changes in predictive performance. A big drop indicates the shuffled feature was important

# Orange Demo of Topics



# Next time

- Clustering
  - k-means clustering
  - Hierarchical clustering
- 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