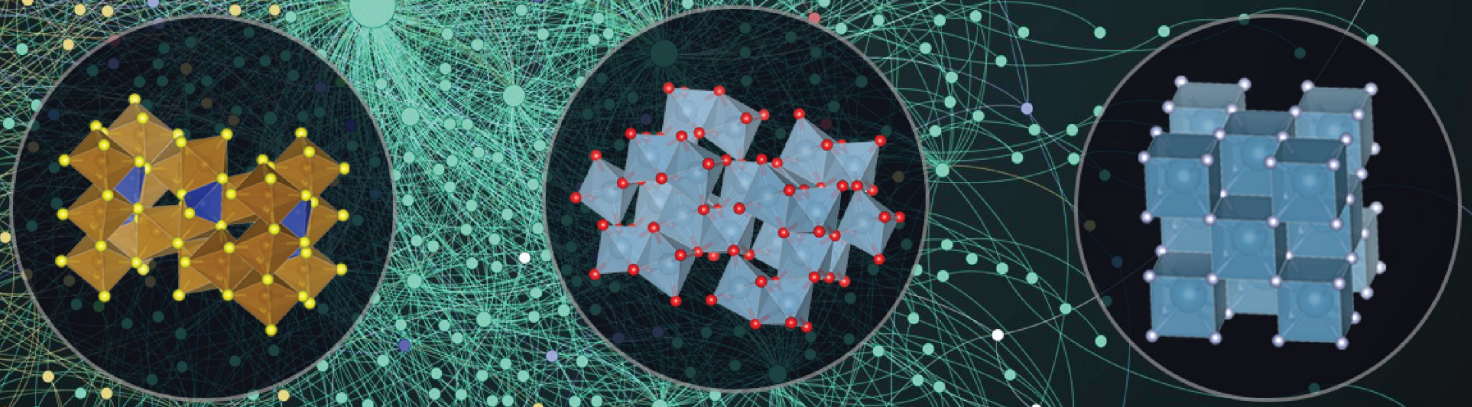
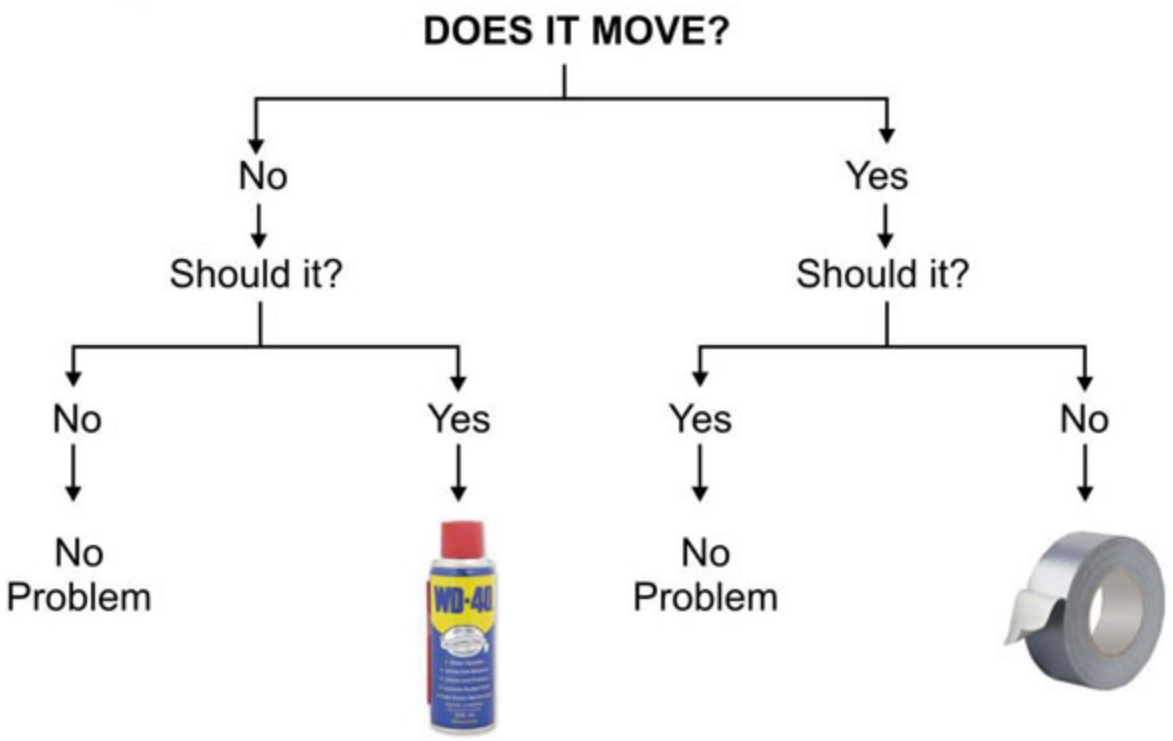
ensemble techniques



Ensembling: many weak learners can accomplish what a few strong learners cannot

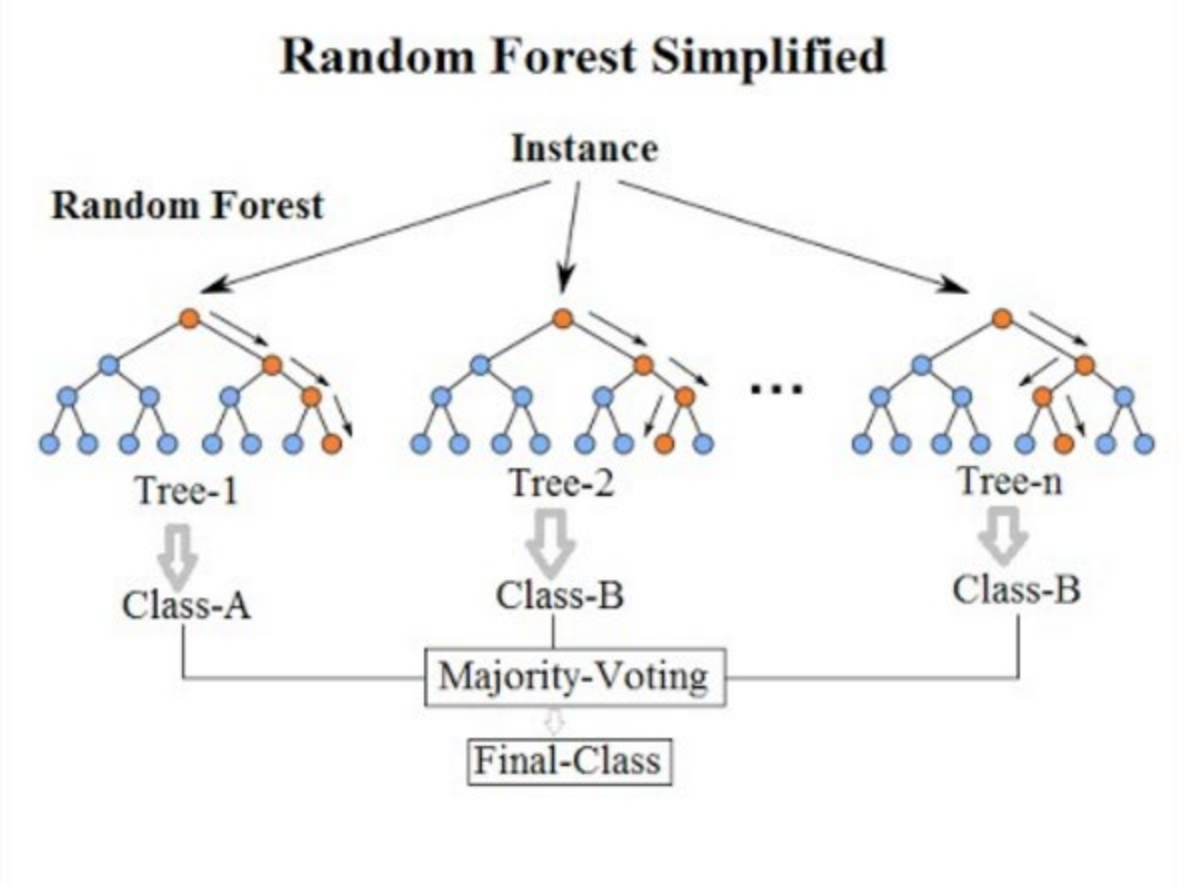
https://knowyourmeme.com/memes/zerg-rush, https://www.youtube.com/watch?v=jNMocWgp9oE&ab_channel=catboi333

Ensemble techniques rely on a large number of learning events



https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

A forest of individual trees is powerful for several reasons



Multiple small trees can be calculated in parallel*

Different trees can sample different features

*depending on algorithm!

https://en.wikipedia.org/wiki/Random_forest#/media/File:Random_forest_diagram_complete.png

Ensemble methods include strong learners made up of weak learners

Weak learners (sometimes called base models):

- Simple
- Building blocks for harder models
- Do poorly alone due to high bias (low complexity) or high variance (high complexity)

Strong learner (ensemble model):

- Combines several weak learners
- Overall improved performance

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Which weak learners should be combined and how?

Which?

- Combining weak learners of the same type is most common and most simple
  - “homogeneous” ensemble
- Base models of different types are also possible, but less common
  - “heterogeneous” ensemble

How?

- Guiding principle is “coherent aggregation”
  - If we choose base models with low bias but high variance, it should be with an aggregating method that tends to reduce variance, whereas if we choose base models with low variance but high bias, it should be with an aggregating method that tends to reduce bias.

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

There are 3 common model combination options

Bagging
- Homogeneous weak learners
- Learning is independent and in parallel
- Combined via deterministic averaging

Boosting
- Homogeneous weak learners
- Learning is sequential and adaptive (depends on previous)
- Combined via deterministic strategy

Stacking
- Heterogeneous weak learners
- Learning is independent and in parallel
- Combines them with a meta-model using weak model outputs as inputs to the meta model

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

There are 3 common model combination options

Bagging
- Homogeneous weak learners
- Learning is independent and in parallel
- Combined via deterministic averaging

Bagging tends to reduce variance

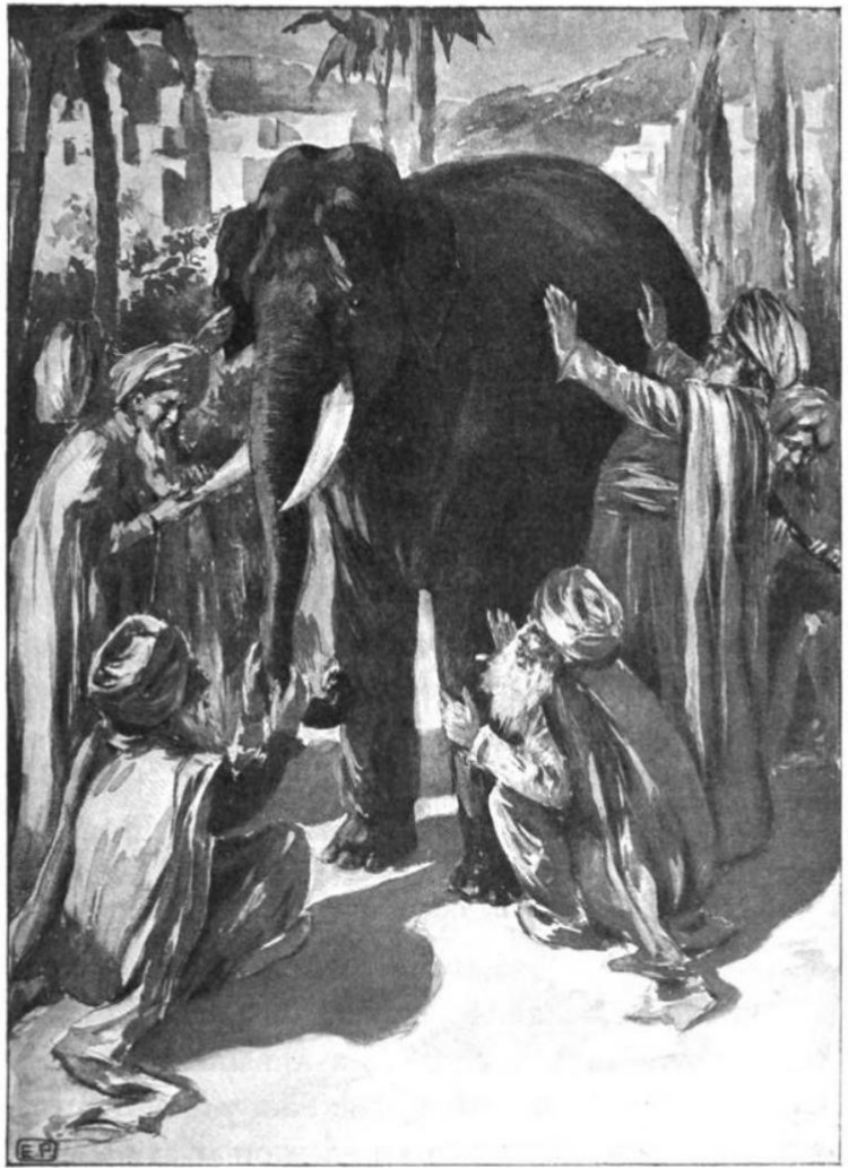
Boosting
- Homogeneous weak learners
- Learning is sequential and adaptive (depends on previous)
- Combined via deterministic strategy

Boosting/Stacking tends to reduce bias but may also reduce variance

Stacking
- Heterogeneous weak learners
- Learning is independent and in parallel
- Combines them with a meta-model using weak model outputs as inputs to the meta model

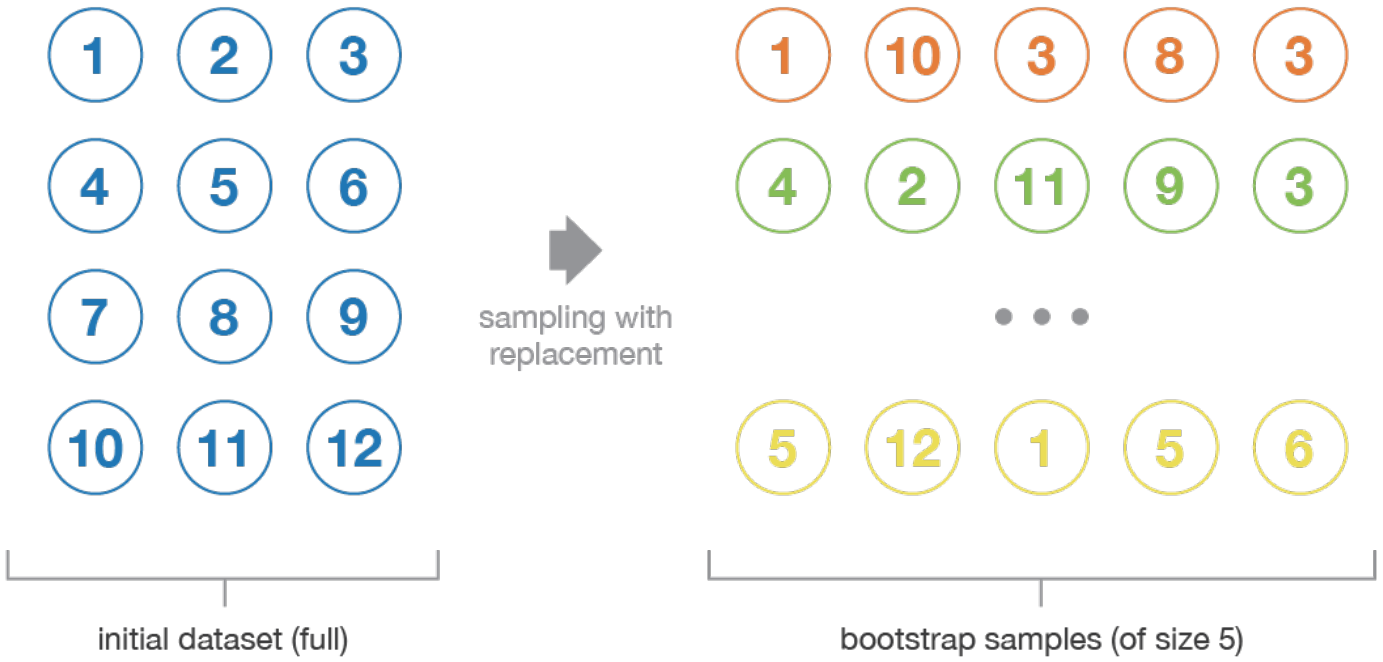
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Each learner has their own perspective and biases



https://en.wikipedia.org/wiki/Blind_men_and_an_elephant#/media/File:Blind_men_and_elephant4.jpg

Bagging is short for "bootstrap aggregating"



Bootstrapping:
generating samples of size B (called bootstrap samples) from an initial dataset of size N by randomly drawing with replacement B observations.

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Is bootstrapping an acceptable way to create a dataset?

In probability theory and statistics, a collection of random variables is independent and identically distributed (I.I.D.) if each random variable has the same probability distribution as the others and all are mutually independent.

We are both flipping the same coin (IID)	We are both flipping the same coin, but we scratch it every 1,000 flips (not I ID)
We are flipping coins that are imbalanced differently (I not ID)	We are working with differently weighted coins and scratch them every 1,000 flips (not I not ID)

Example courtesy of Jason Hattrick-Simpers and Brian DeCost

https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables

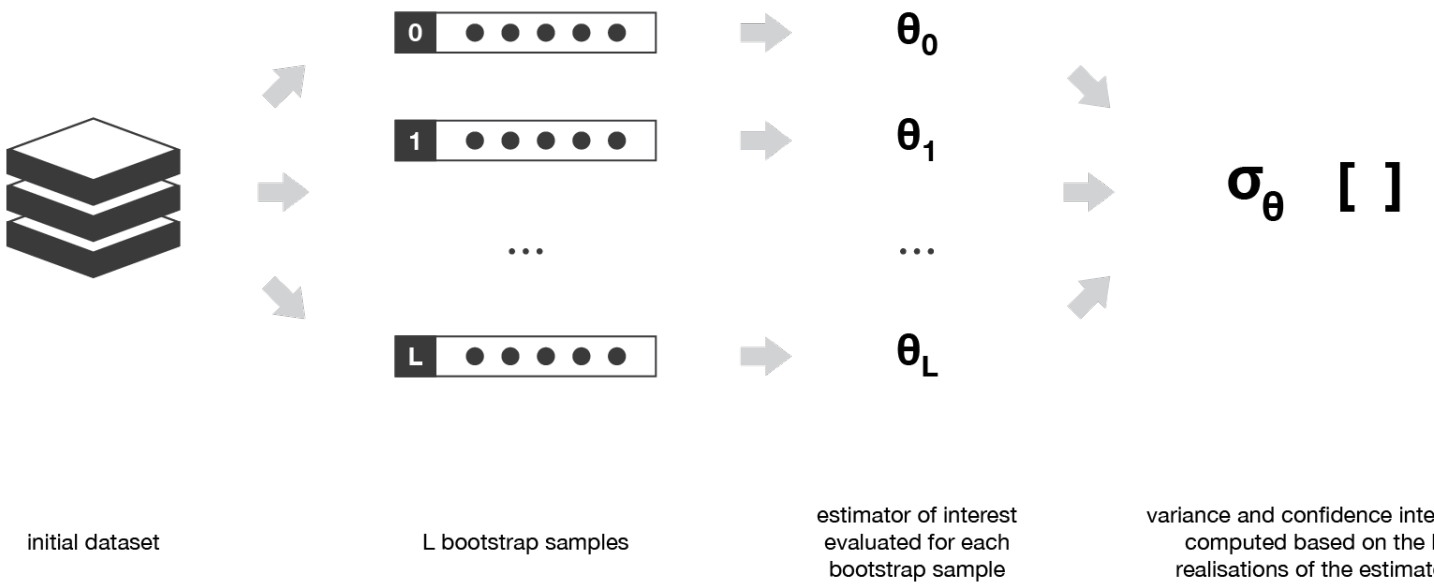
What does I.I.D. look like in materials data example?

We use an oxide database to predict oxide material phases	We use a time-varying oxide database to predict oxide phases (not I ID)
We use an oxide database to predict NITRIDE phases (I not ID)	We use a time-varying oxide database to predict NITRIDE phases (not I not ID)

For bootstrapping to approximately adhere to I.I.D. we should follow a few guidelines

First, the size N of the initial dataset should be large enough to capture most of the complexity of the underlying distribution so that sampling from the dataset is a good approximation of sampling from the real distribution (**representativity**)

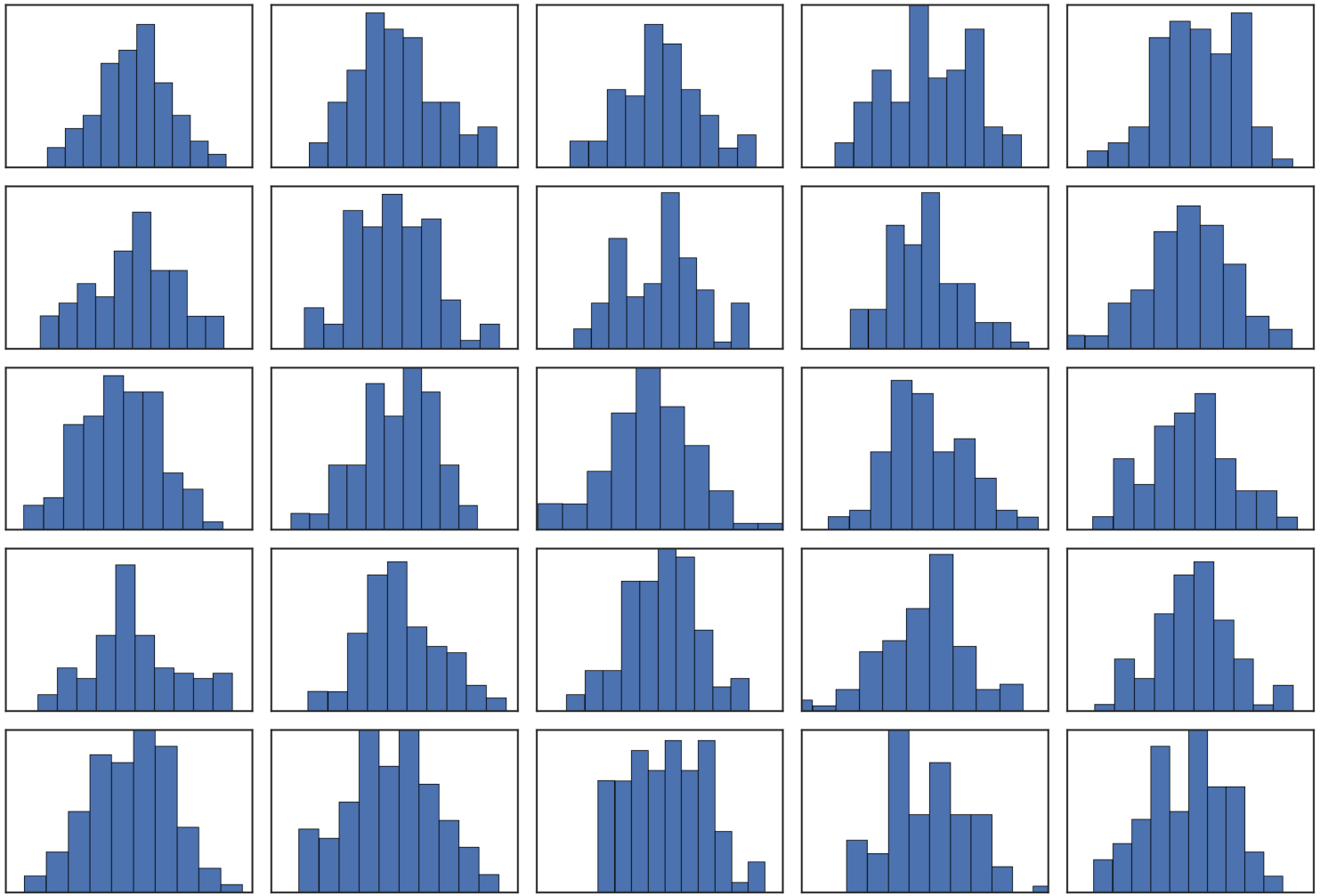
Second, the size N of the dataset should be large enough compared to the size B of the bootstrap samples so that samples are not too much correlated (**independence**).



Bootstrapping allows us to create a large amount of data based on a single distribution where each dataset is approximately i.i.d.

https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables

Each dataset is a **sample** coming from a true, unknown distribution



Fitted models are trained on datasets, not the underlying true, unknown distribution!

So each model is subject to variability

Bootstrapping gives us many nearly independent model datasets

http://work.thaslwanter.at/Stats/html/statsDistributions.html

Mathematically, bootstrapping is…

If we have L bootstrap samples of size B

$$\{z_1^1, z_2^1 \dots z_B^1\}, \{z_1^2, z_2^2 \dots z_B^2\}, \dots \{z_1^L, z_2^L \dots z_B^L\}$$

Fit each bootstrap sample to get L weak learners

$$w_1, w_2, \dots w_L$$

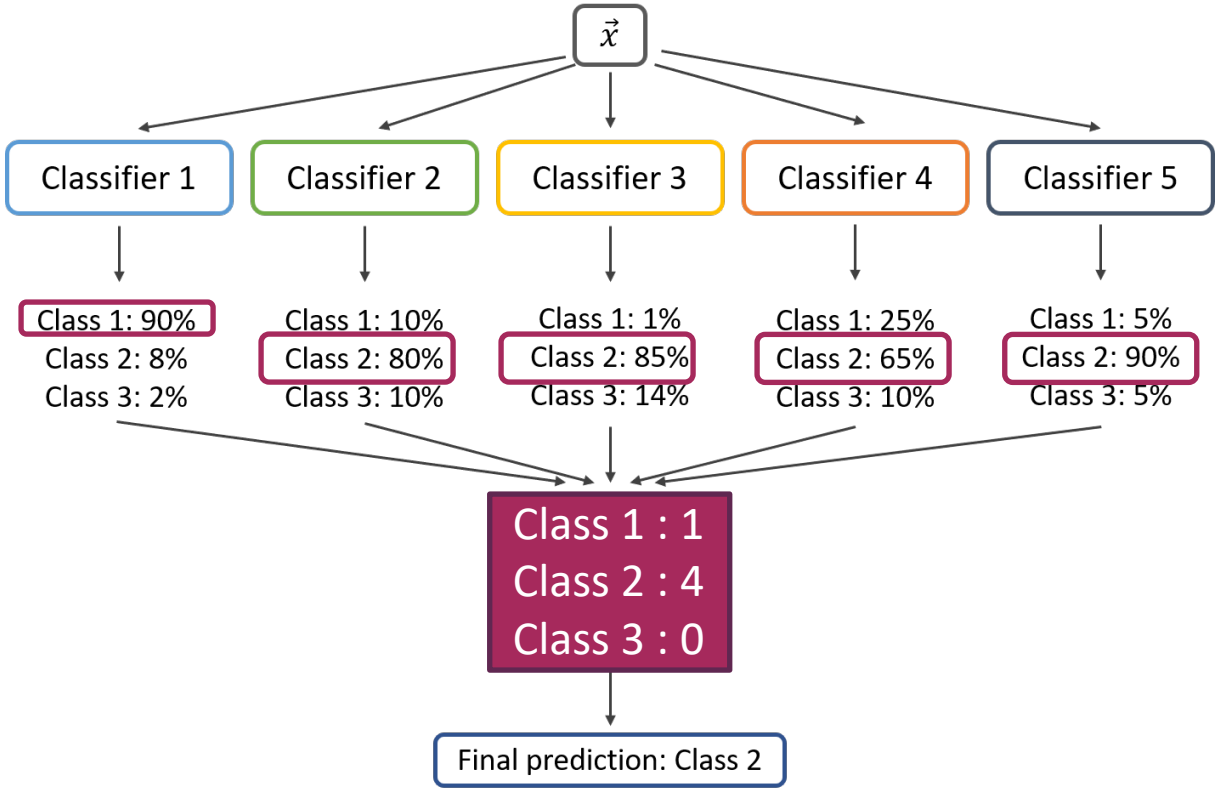
And then aggregate to get an ensemble with lower variance

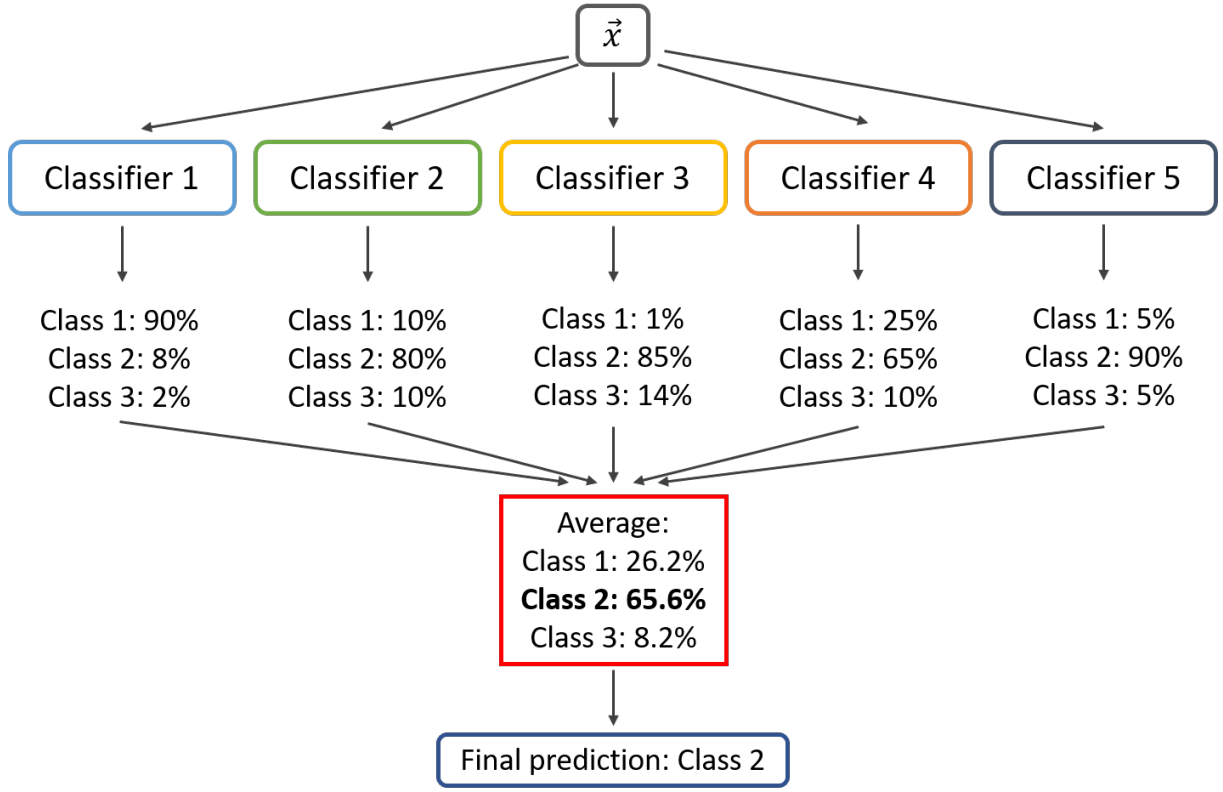
$$regression\ (average):\ \ s_L = \frac{1}{L}\sum_{i=1}^{L} w_i$$

$$classification\ (majority\ vote): \underset{k}{\operatorname{argmax}}[card(i|w_i = k)]$$

http://work.thaslwanter.at/Stats/html/statsDistributions.html

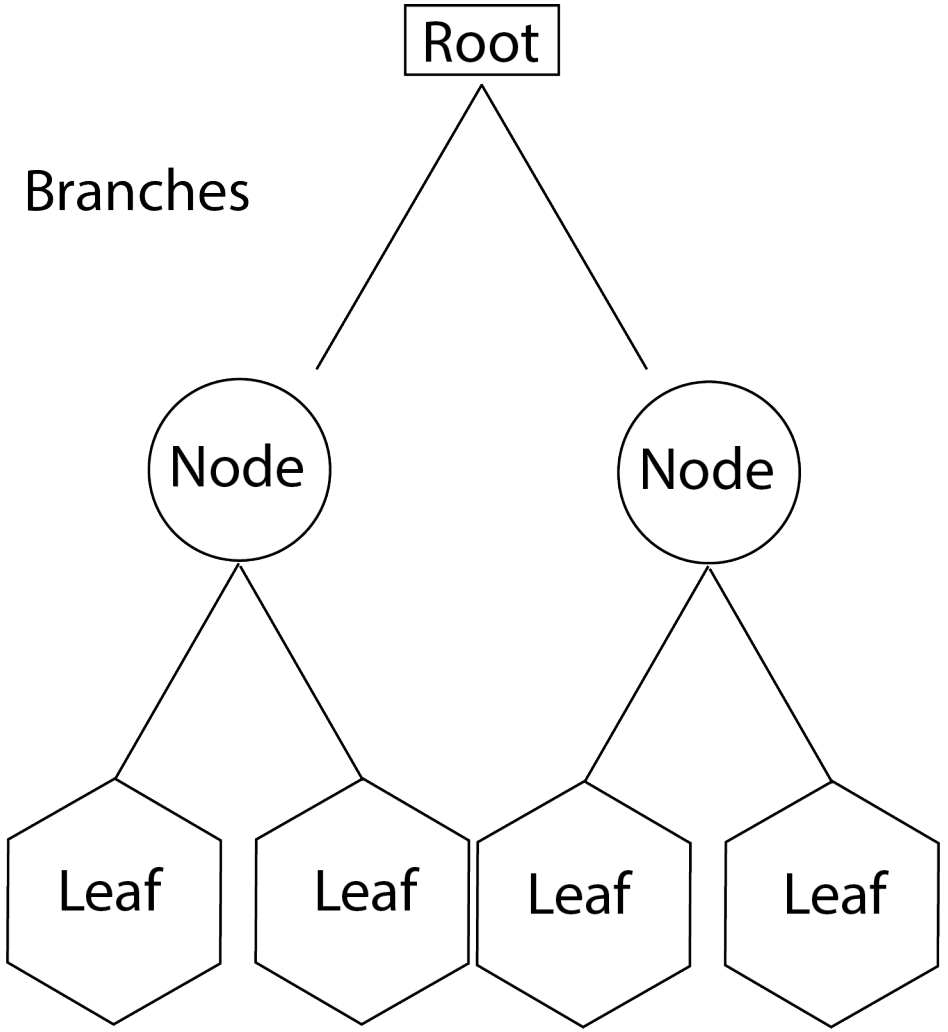
Classification can occur via hard voting or soft voting





http://work.thaslwanter.at/Stats/html/statsDistributions.html

There are a number of hyperparameters to select from for decision trees

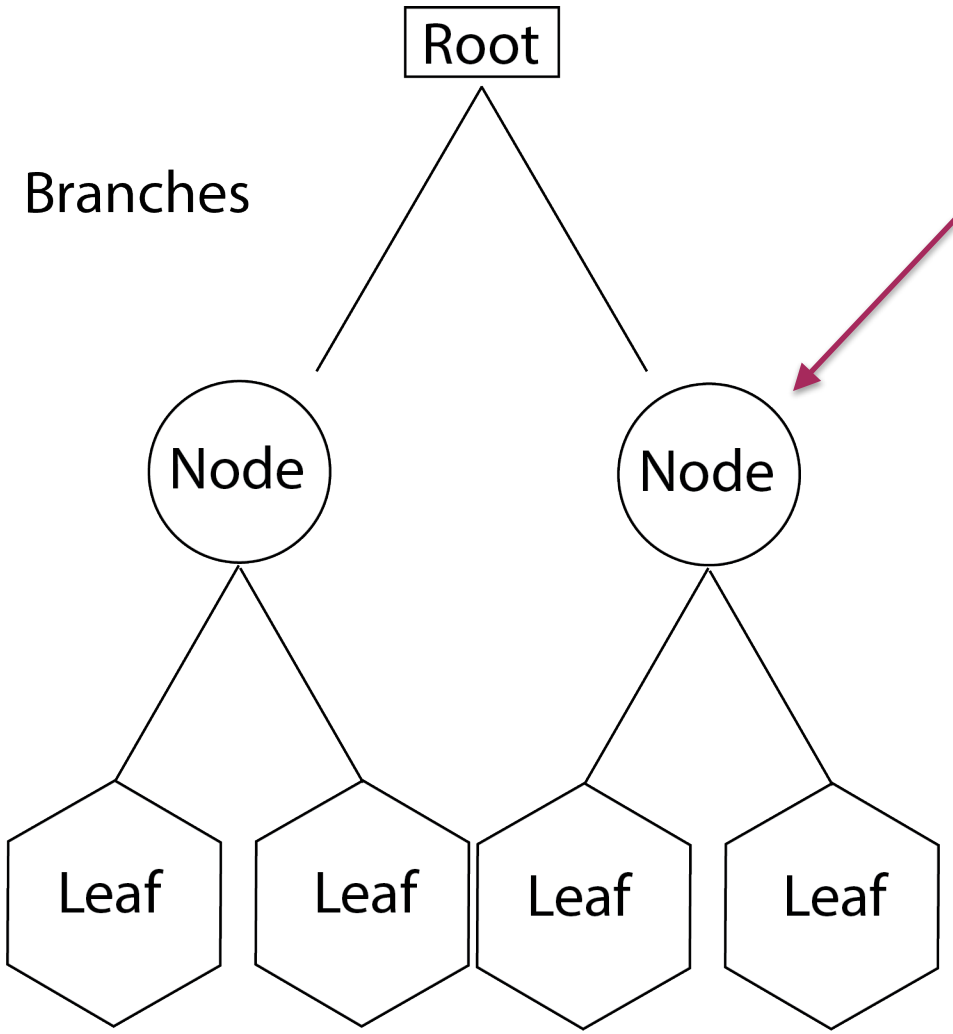


Less variance, more bias
(better for sequential methods)

Increasing tree depth

More variance, less bias
(better for parallel methods)

There are a number of hyperparameters to select from for decision trees

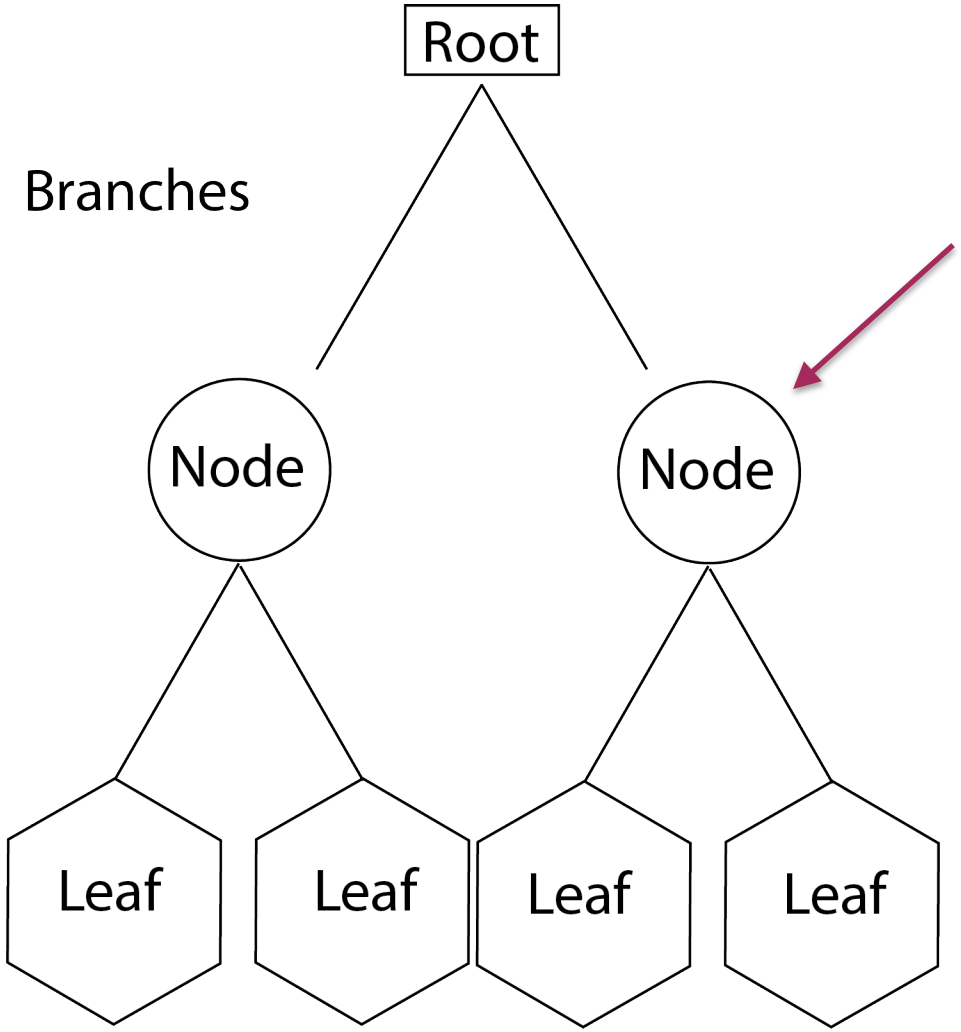


Decision made at each node based on features

Not all features are used
Each tree gets random subset of features
Best if $\sqrt{N}$ where N is total number of features

Makes model more robust to missing data because regression or classification can be done only on trees having required features!

The value of the feature used for splitting data must be learned



Based on minimized residual

$$RSS = \sum_{left} (y_i - \bar{y}_{left})^2 + \sum_{right} (y_i - \bar{y}_{right})^2$$

Based on maximizing split proportions

$$GINI = n_{left} \sum_{k=1..K} p_{kleft}(1 - p_{kleft}) + n_{right} \sum_{k=1..K} p_{kright}(1 - p_{kright})$$

There are a number of hyperparameters to select from for decision trees

```
from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor(random_state = 42)

from pprint import pprint

# Look at parameters used by our current forest
print('Parameters currently in use:\n')
pprint(rf.get_params())
```

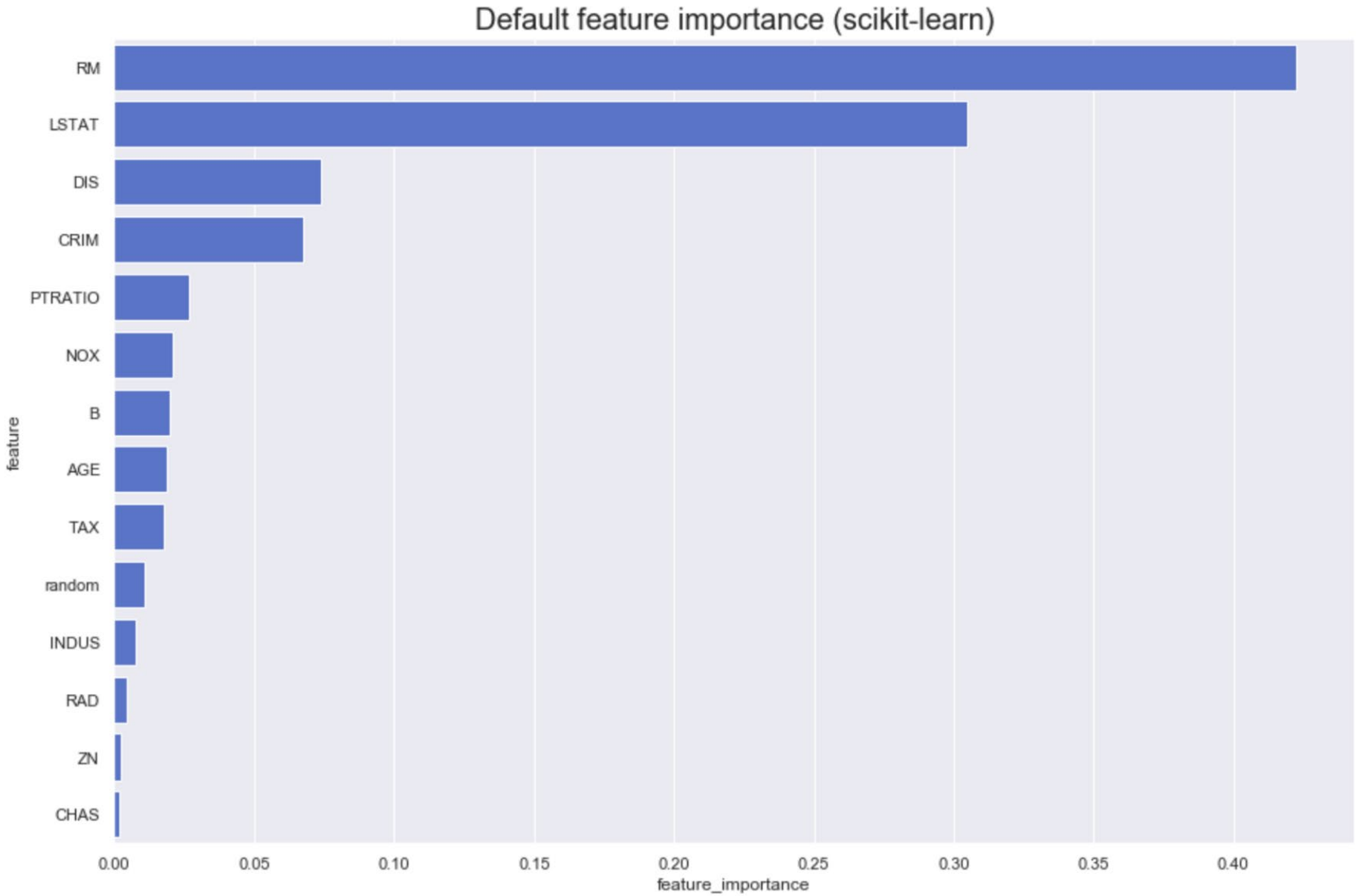
```
Parameters currently in use:

{'bootstrap': True,
 'criterion': 'mse',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 10,
 'n_jobs': 1,
 'oob_score': False,
 'random_state': 42,
 'verbose': 0,
 'warm_start': False}
```

We can read all about the different model parameters (hyperparameters) through the documentation

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Random forests can easily give us information on feature importance!



Feature weight or importance can be calculated in several different ways

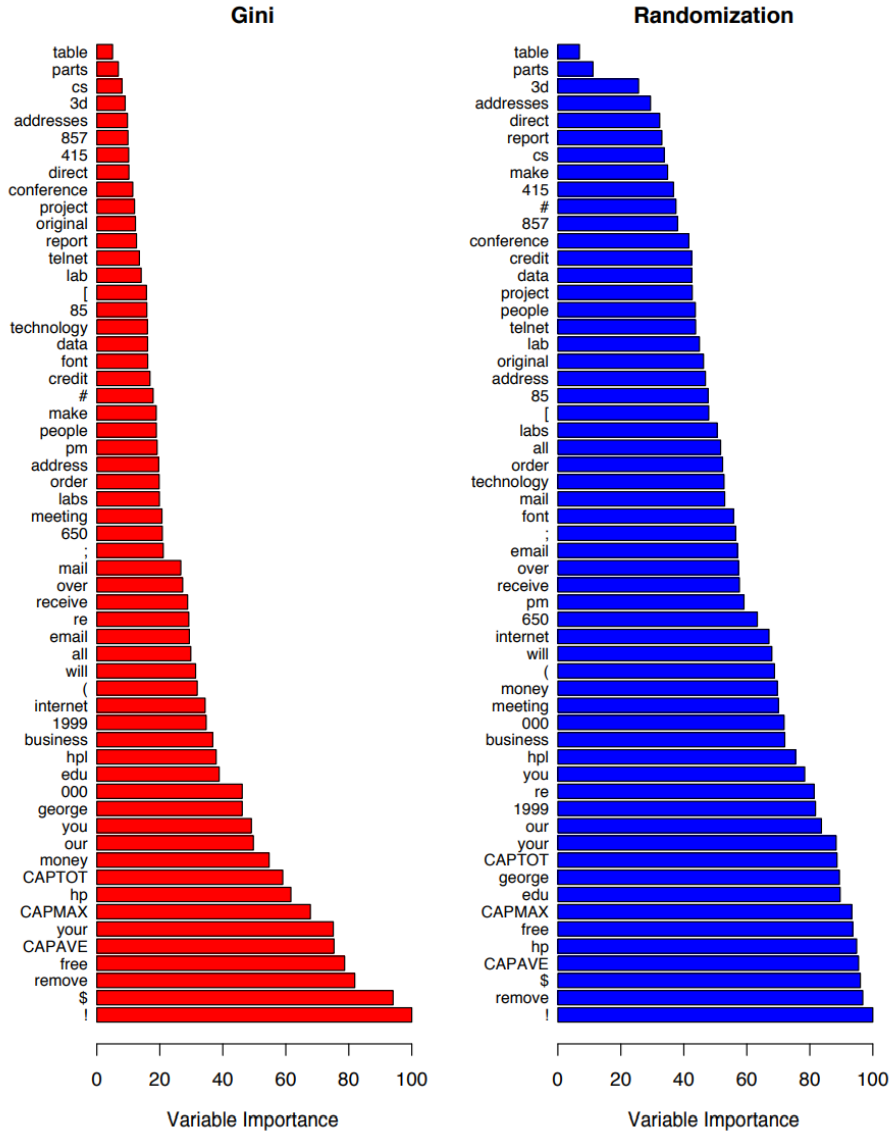
GINI importance (mean decrease in impurity)

- "Defined as the total decrease in node impurity (weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node)) averaged over all trees of the ensemble."

Permutation importance (mean decrease in accuracy)

- Random forests also use the OOB samples to construct a different variable-importance measure, apparently to measure the prediction strength of each variable. When the b-th tree is grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded. Then the values for the j-th variable are randomly permuted in the OOB samples, and the accuracy is again computed. The decrease in accuracy as a result of this permuting is averaged over all trees.

Comparison of two feature weighting techniques for spam features



https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf

Scikit-learn calculates feature weight using GINI importance

For each decision tree, we calculate all node importance values, ni_j

$$ni_j = w_j C_j + w_{left(j)} C_{left(j)} + w_{right(j)} C_{right(j)}$$

Where C_j is the impurity of node j

$$C_j = \sum_{i=1}^{C} f_i(1 - f_i)$$

Importance of feature on a specific tree becomes

$$f_{i_i} = \frac{\sum_{j,node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k\ all\ nodes} ni_k}$$

Normalize by dividing by sum of all features

$$norm\ f_{i_i} = \frac{f_{i_i}}{\sum_{j\ all\ features} fi_j}$$

Then average feature weight across all trees

$$RF f_{i_i} = \frac{\sum_{j\ all\ trees} norm\ fi_{ij}}{T}$$

Boosting for sequential ensemble methods

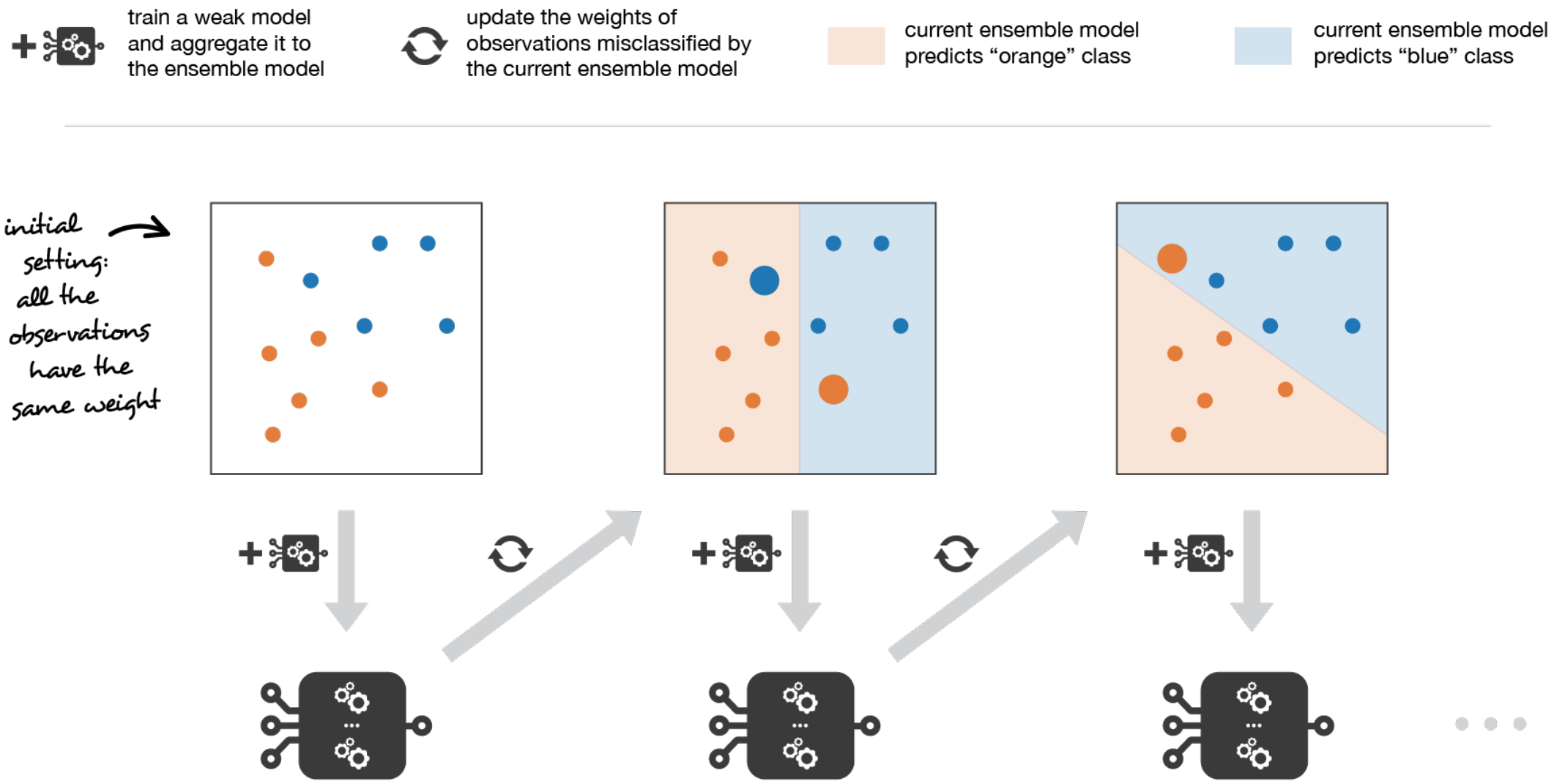
Iterative model creation where model *i* is based on previous models.

- More attention given to observations badly handled in previous iteration
- Focus on hardest observations to fit
- Implement with shallow trees (low variance, high bias)

How will sequential models be fit?

- What info from previous model is taken into account?
- How is current and previous model aggregated?

Adaptive boosting AKA adaboost is one way to select next model



Adaptive boosting AKA adaboost is one way to select next model

Ensemble model as weighted sum of weak learners

$$s_L = \sum_{l=1}^{L} c_l w_l$$

Where c_l is coefficient, w_l is a weak learner. Super hard to solve this equation….

So break it up and add new learners iteratively

$$s_l = s_{l-1} + c_l w_l$$

Select c_l and w_l so that s_l fits training data best. (max improvement over s_{l-1})

$$c_l, w_l = argmin\ E\ (s_{l-1} + cw) = argmin \sum_{n=1}^{N} e\big(y_n, s_{l-1}(x_n) + cw(x_n)\big)$$

E is model error, e is the loss function so optimization takes place on most recent model, not all models.

Adaboost in plain English

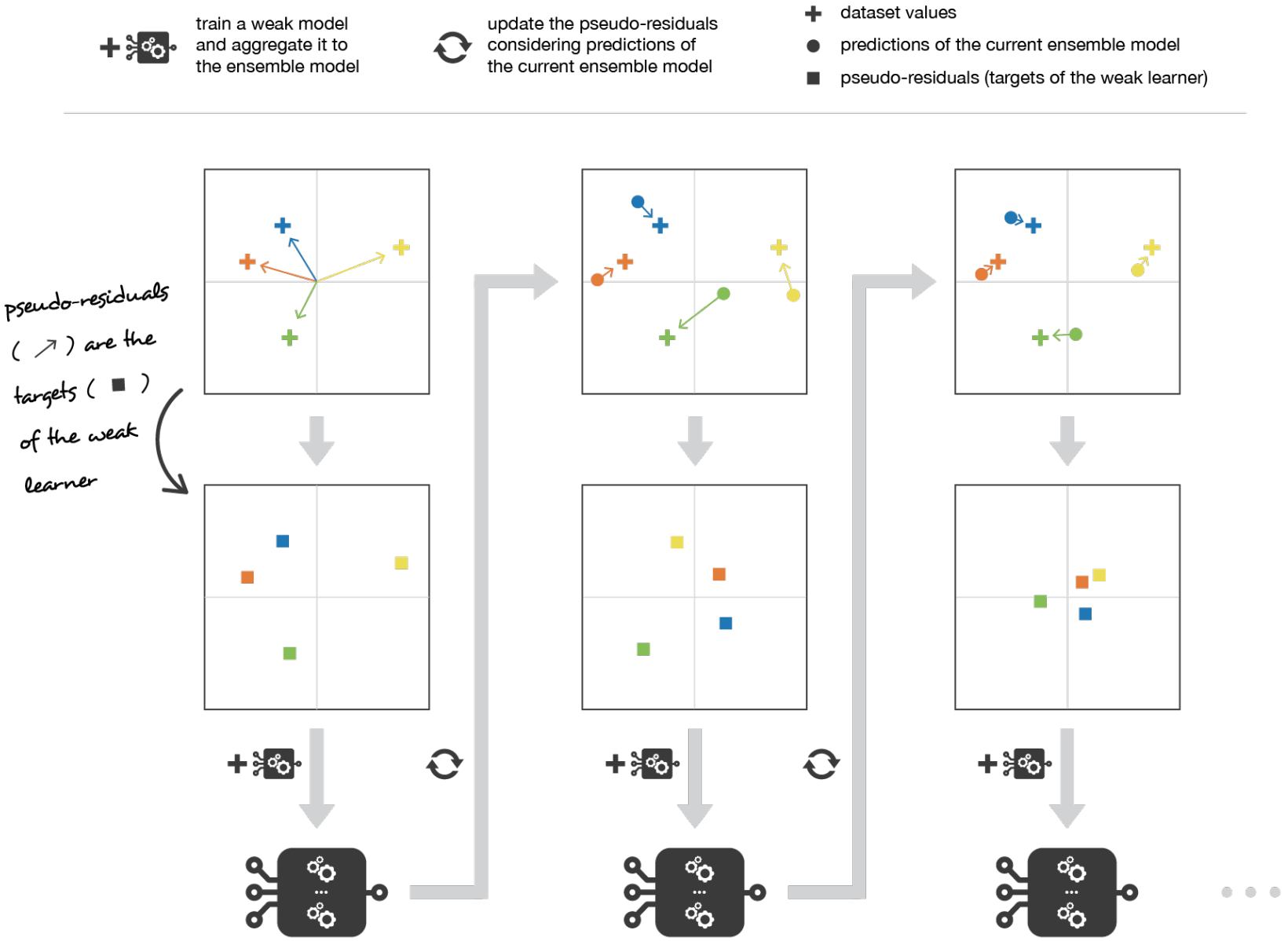
First, it updates the observations weights in the dataset and train a new weak learner with a special focus given to the observations misclassified by the current ensemble model. Second, it adds the weak learner to the weighted sum according to an update coefficient that express the performances of this weak model: the better a weak learner performs, the more it contributes to the strong learner.

Consider N observations

- Fit the best possible weak model with the current observations weights (1/N)
- Compute the value of the update coefficient
- Update the strong learner (add new weak learner times its coefficient)
- Compute new observations and repeat

Variations like LogitBoost, L2Boost available with slightly different loss functions

Gradient boosting is another way to select next model



Gradient boosting is another way to select next model

Ensemble model as weighted sum of weak learners

$$s_L = \sum_{l=1}^{L} c_l w_l$$

At each iteration we fit a weak learner to the opposite of the gradient of the fitting error.

$$s_l = s_{l-1} - c_l * \nabla_{s_{l-1}} E(s_{l-1})$$

Where *E* is the fitting error of the given model, c_l is coefficient corresponding to step size, the final term is the gradient of the fitting error with respect to the previous model s_{l-1}.

In essence, each observation has a pseudo-residual describing how well the model fits it. So we fit a weak learner to the pseudo-residuals themselves!

Gradient boosting in plain English

Each observation has pseudo-residual equal to observation and then…

- Fit weak learner to residuals
- Compute the value of the optimal step size that defines by how much we update the ensemble model in the direction of the new weak learner
- Update the ensemble model by adding the new weak learner multiplied by the step size (make a step of gradient descent)
- Compute new pseudo-residuals that indicate, for each observation, in which direction we would like to update next the ensemble model predictions

Thus, gradient boosting can be considered as a generalization of adaboost to arbitrary differentiable loss functions.

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Stacking allows us to combine heterogeneous weak learners

"A model made up of models"

Basic steps

- We split data into two folds
- Choose weak learners, fit to one fold of data
- Use weak learners to make predictions for second fold observations
- Fit the meta-model on the second fold, using predictions made by weak learners as inputs

Data used to train weak learners is not used again for training meta-model so we lose ½ of our data unless we do k-fold splitting and training

Multi level stacking is possible with more than one meta-model.

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205