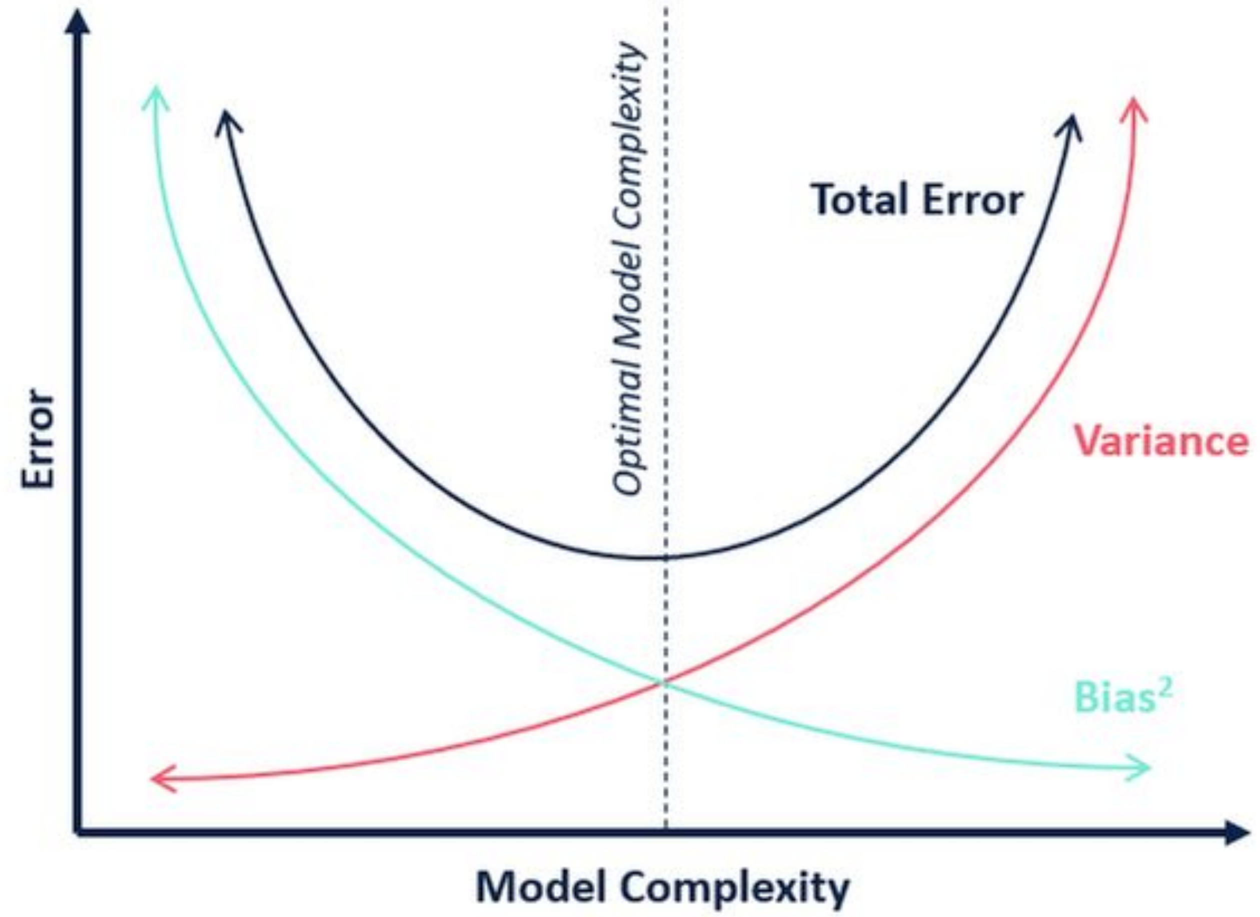
# Random Forests

Bagging, ensembling

COGNIBOT
AI meets Industry

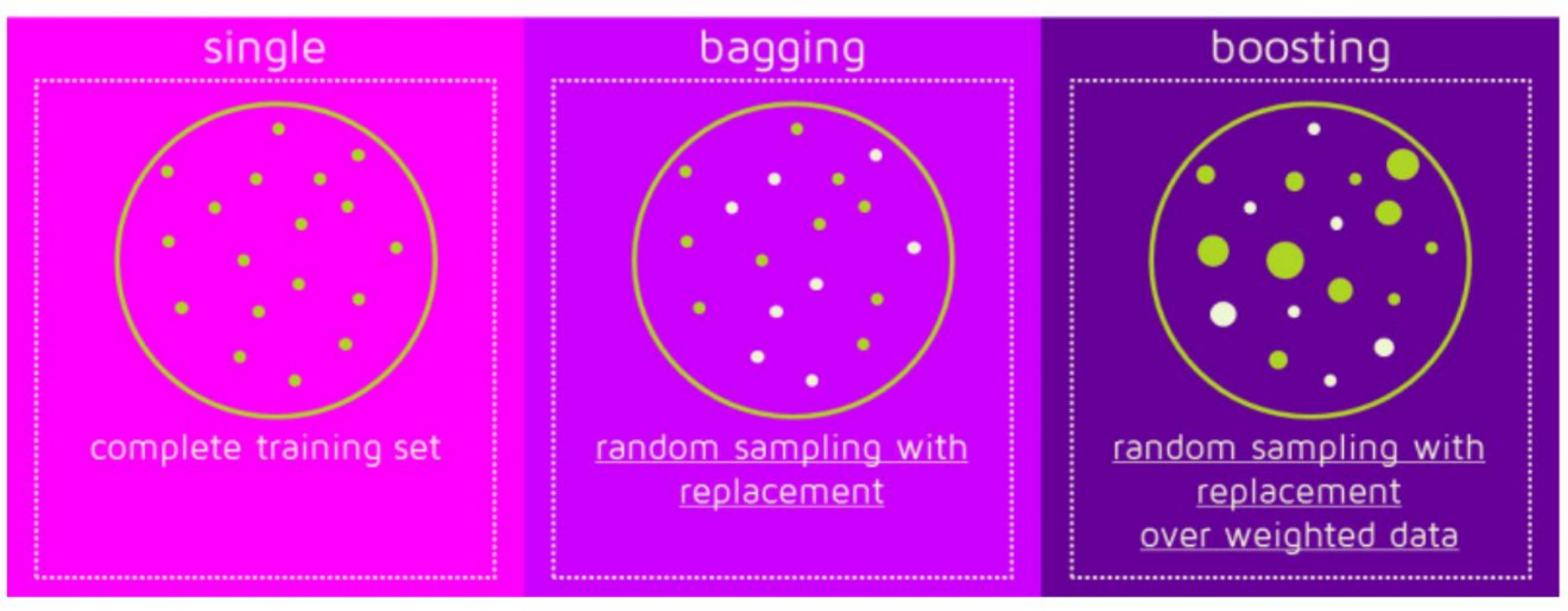



GNIBOT
AI meets Industry

# Ensemble models

- Voting or Averaging of predictions of multiple pre-trained models
- Instead of training different models on same data, train same model multiple times on different data sets, and “combine” these “different” models
- We can use some simple/weak model as the base model
- How do we get multiple training data sets (in practice, we only have one data set at training time)?

COGNIBOT
AI meets Industry

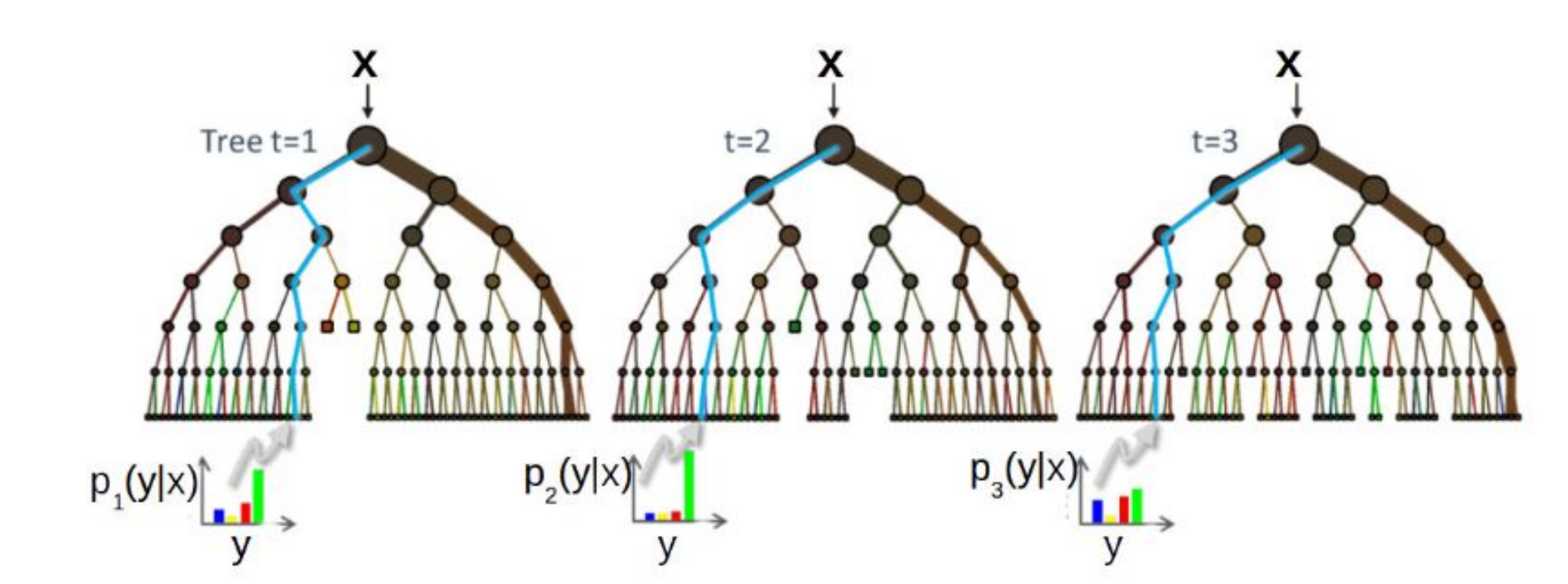



COGNIBOT
AI meets Industry

# Bagging

- Bagging stands for Bootstrap Aggregation
- Takes original data set D with N training examples
- Creates M copies:
  - Each ˜Dm is generated from D by sampling with replacement
  - Each data set ˜Dm has the same number of examples as in data set D
- Train models h1,...,hm using ˜D1,...,˜DM, respectively
- Use an averaged model $h = \frac{1}{M}\sum_{m=1}^{M} h_m$ as the final model

# Random Forests

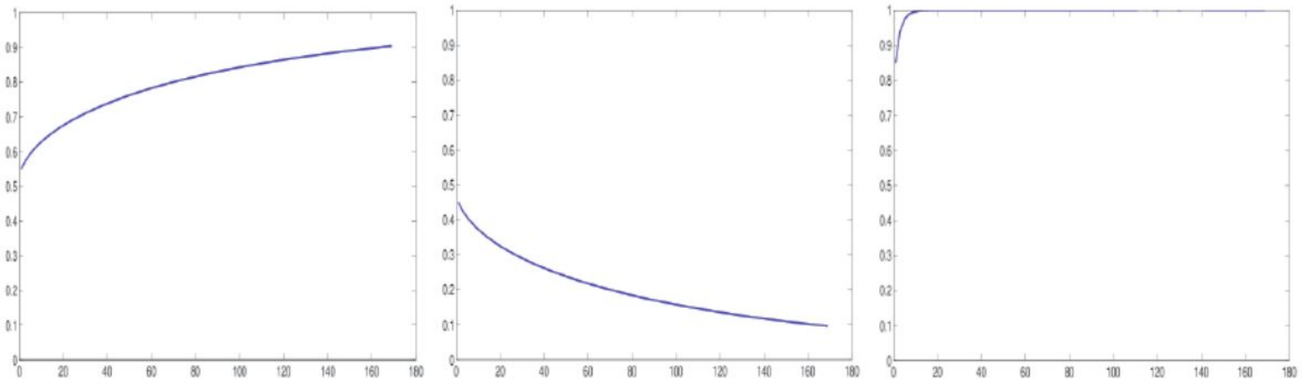



COGNIBOT
AI meets Industry

- An ensemble of decision tree (DT) classifiers
- Uses bagging on features (each DT will use a random set of features)
  - Given a total of D features, each DT uses $\sqrt{D}$ randomly chosen features
- All DTs usually have the same depth
- Each DT will split the training data differently at the leaves
- Prediction for a test example votes on/averages predictions from all the DTs

COGNIBOT
AI meets Industry

- Given n voters, the probability the majority makes the right choice: $$\Pr(\text{majority correct}) \sum_{j=\frac{m+1}{2}}^{m} \frac{m!}{j!(m-j)!} p^{j} (1-p)^{m-j}$$



$p = 0.55$ $p = 0.45$ $p = 0.85$

# Bagging example