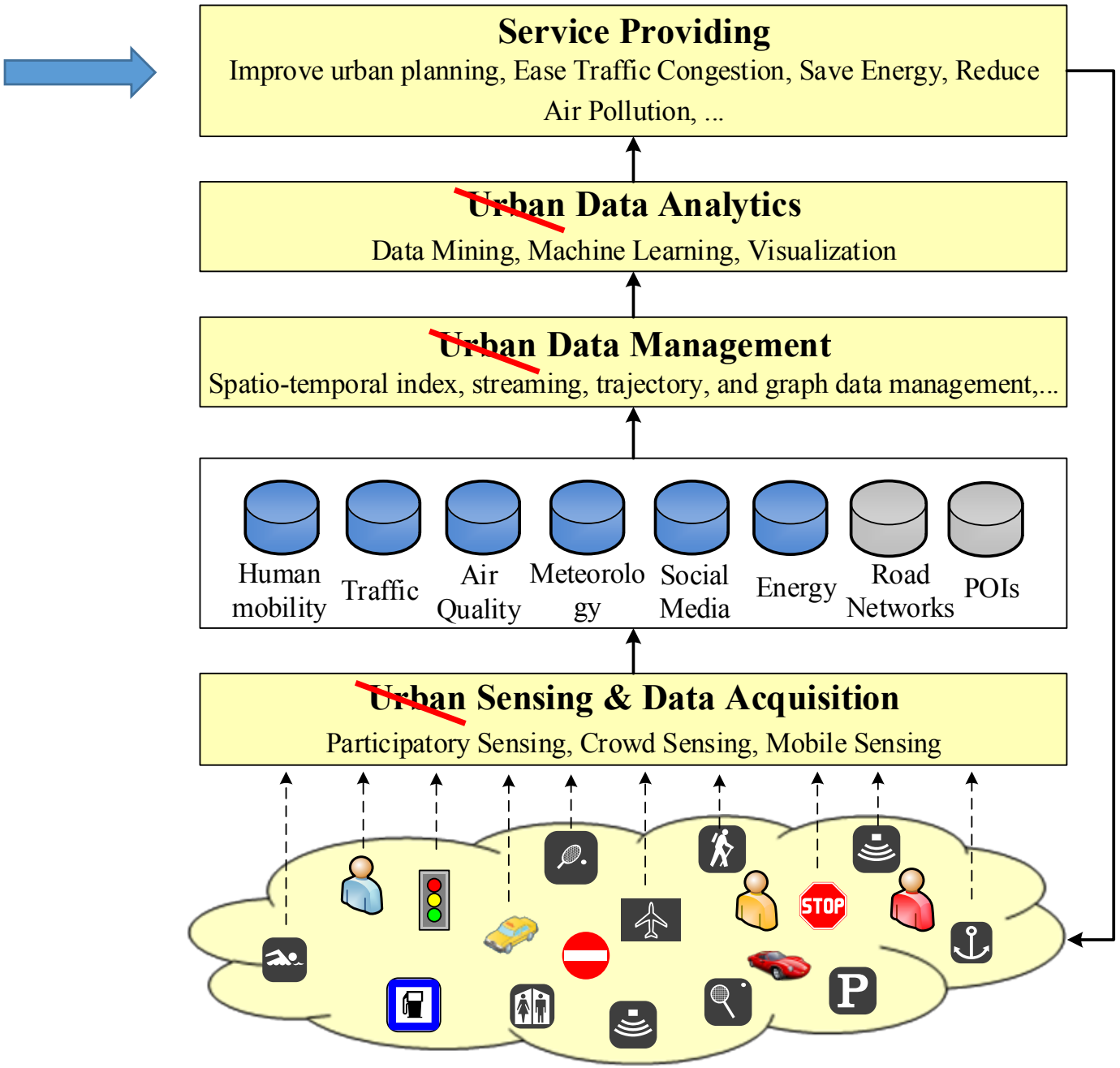
Welcome to

DS3010:
DS-III: Computational Data Intelligence
Ensemble

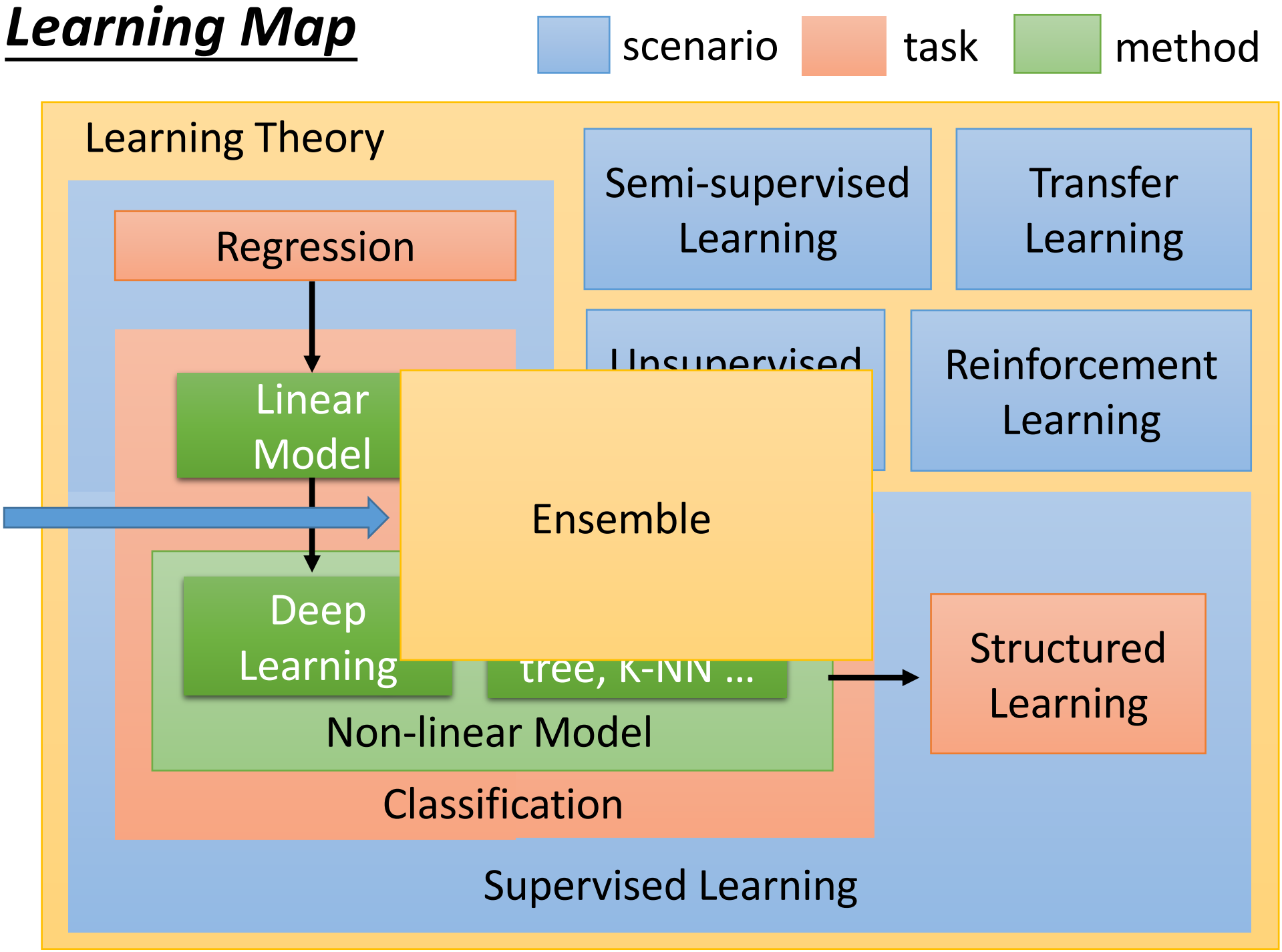
Prof. Yanhua Li

Time: 11:00am – 12:50pm M & R
Location: HL 114
D-term 2022

Data pipeline

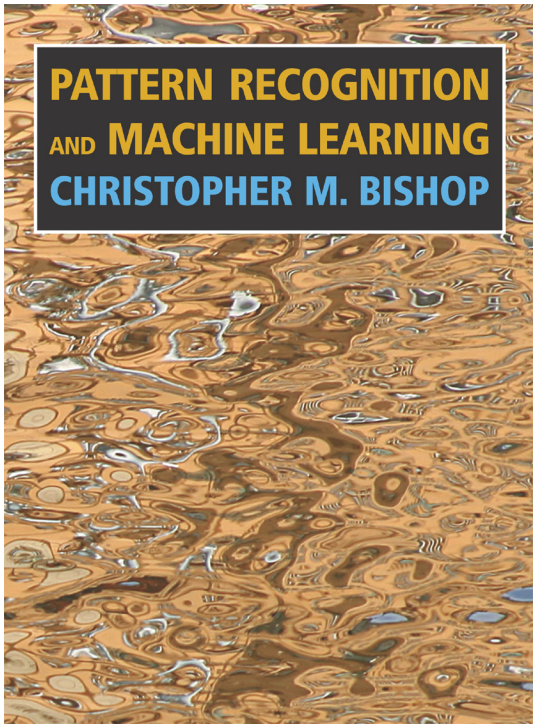


Urban Computing: concepts, methodologies, and applications.
Zheng, Y., et al. *ACM transactions on Intelligent Systems and Technology*.


Learning Map
scenario
task
method
Learning Theory
Semi-supervised Learning
Transfer Learning
Regression
Unsupervised
Reinforcement Learning
Linear Model
Ensemble
Deep Learning
tree, K-NN ...
Structured Learning
Non-linear Model
Classification
Supervised Learning

References

Ensemble



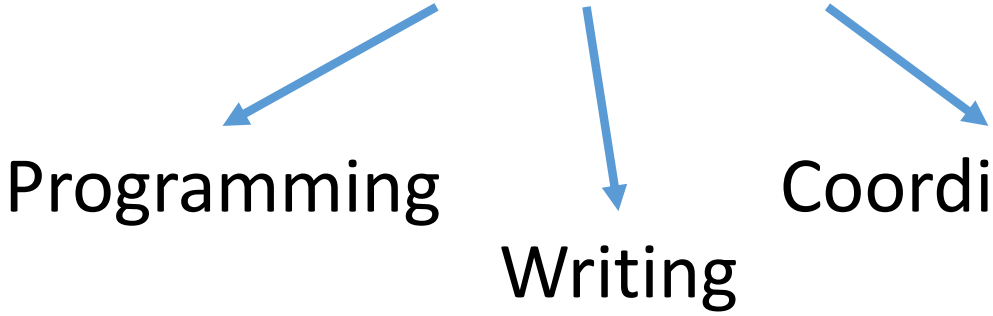
Bishop: Chapter 14.3-14.4

Ensemble

Framework of Ensemble

- Get a set of classifiers
  - $f_1(x), f_2(x), f_3(x), \ldots\ldots$

They should be diverse.



Members in a team.

- Aggregate the classifiers (*properly*)

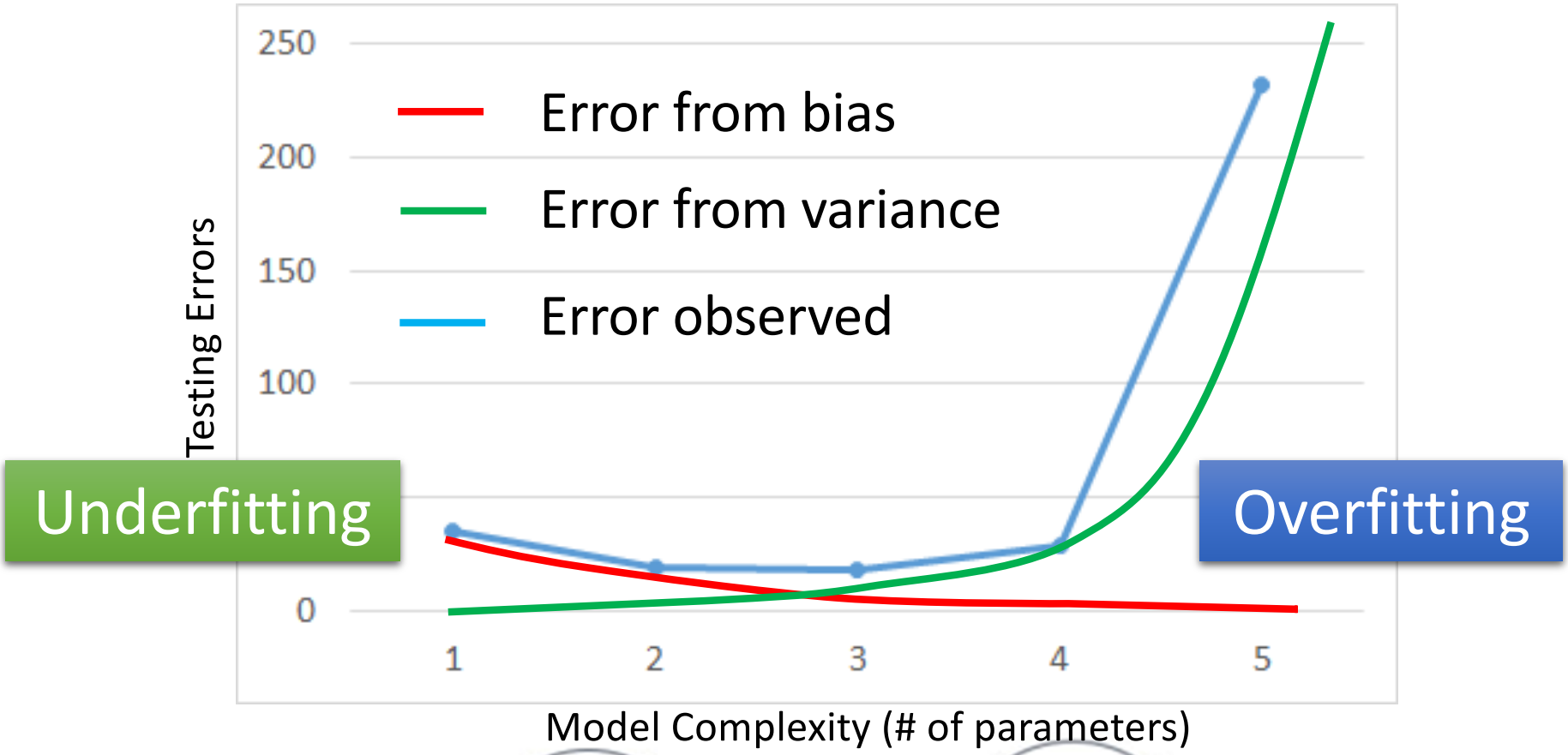
Outline

Ensemble: Bagging

Ensemble: Stacking

Ensemble: Bagging

Review: Bias v.s. Variance



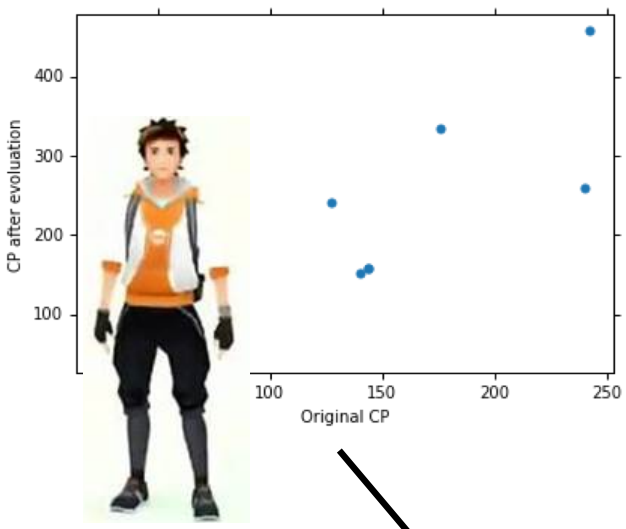
Large Bias

Small Variance

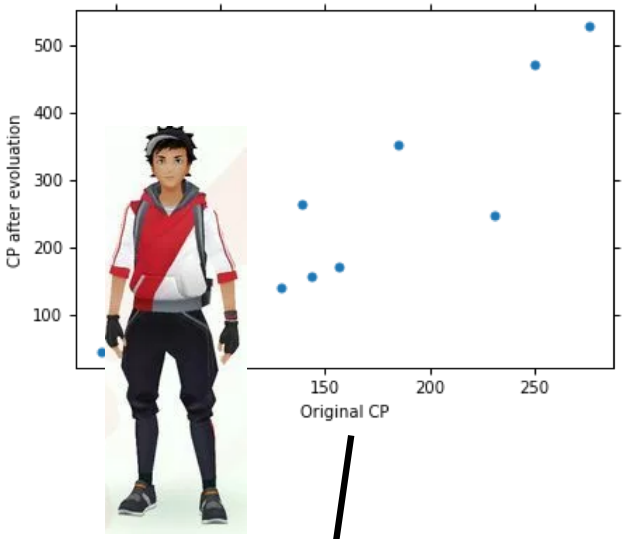
Small Bias

Large Variance

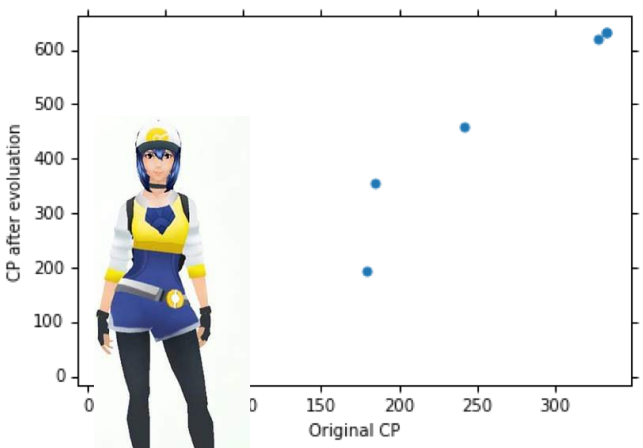
Universe 1



Universe 2

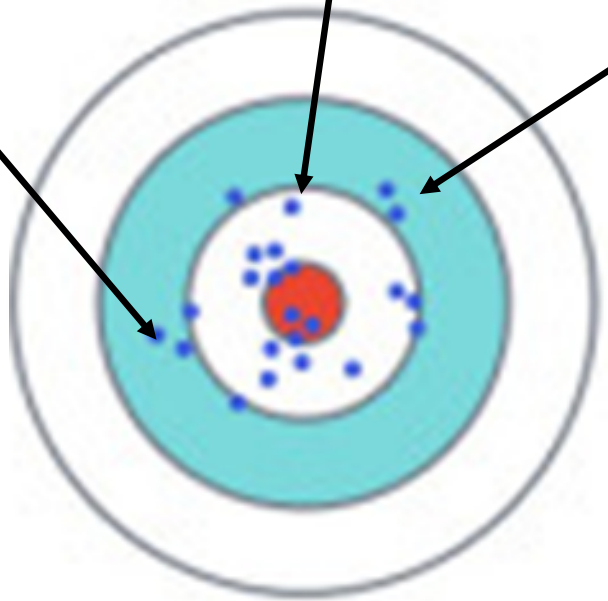


Universe 3



A complex model will have large variance.

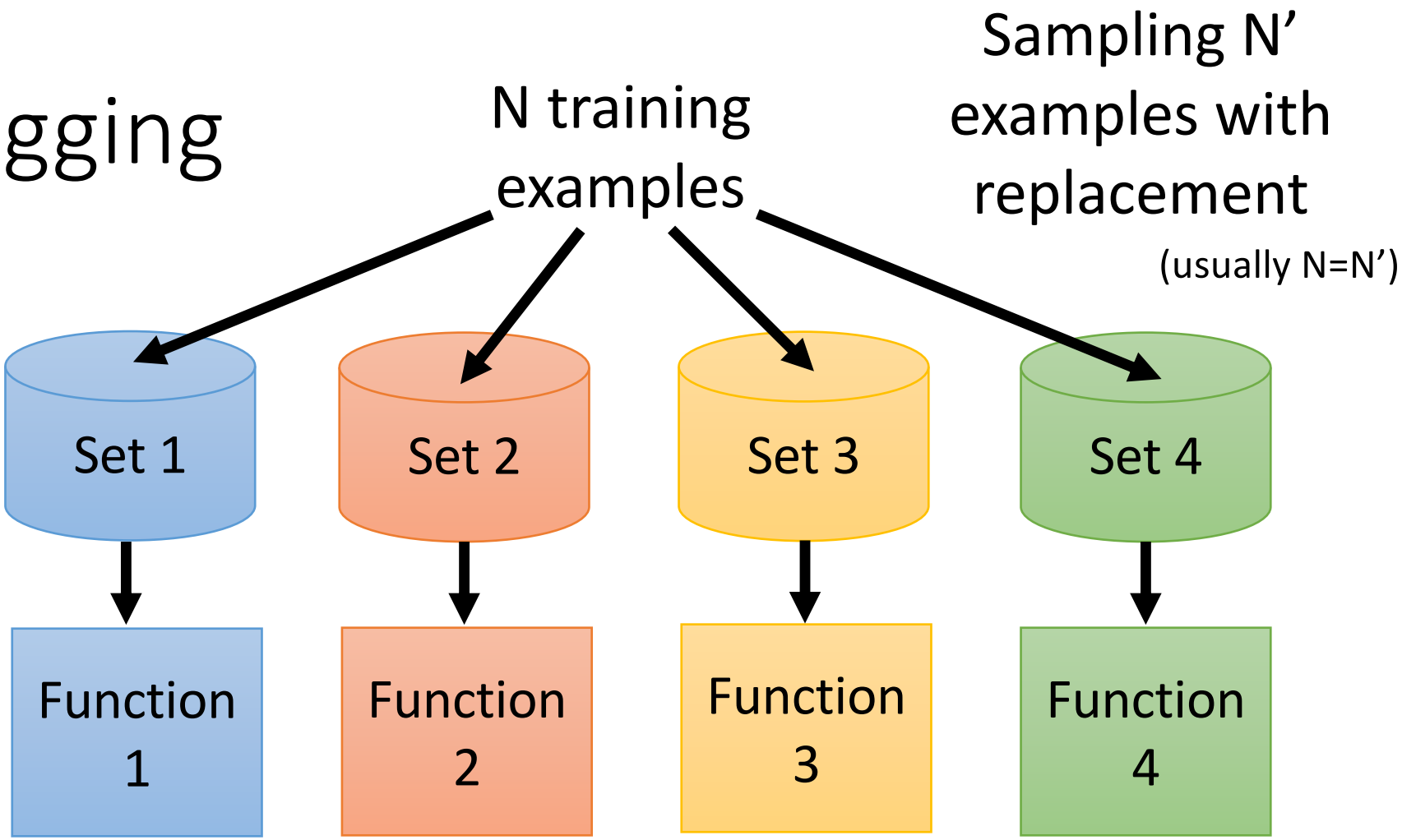
We can average complex models to reduce variance.



If we average all the f^*, is it close to $\hat{f}$

$$E[f^*] = \hat{f}$$

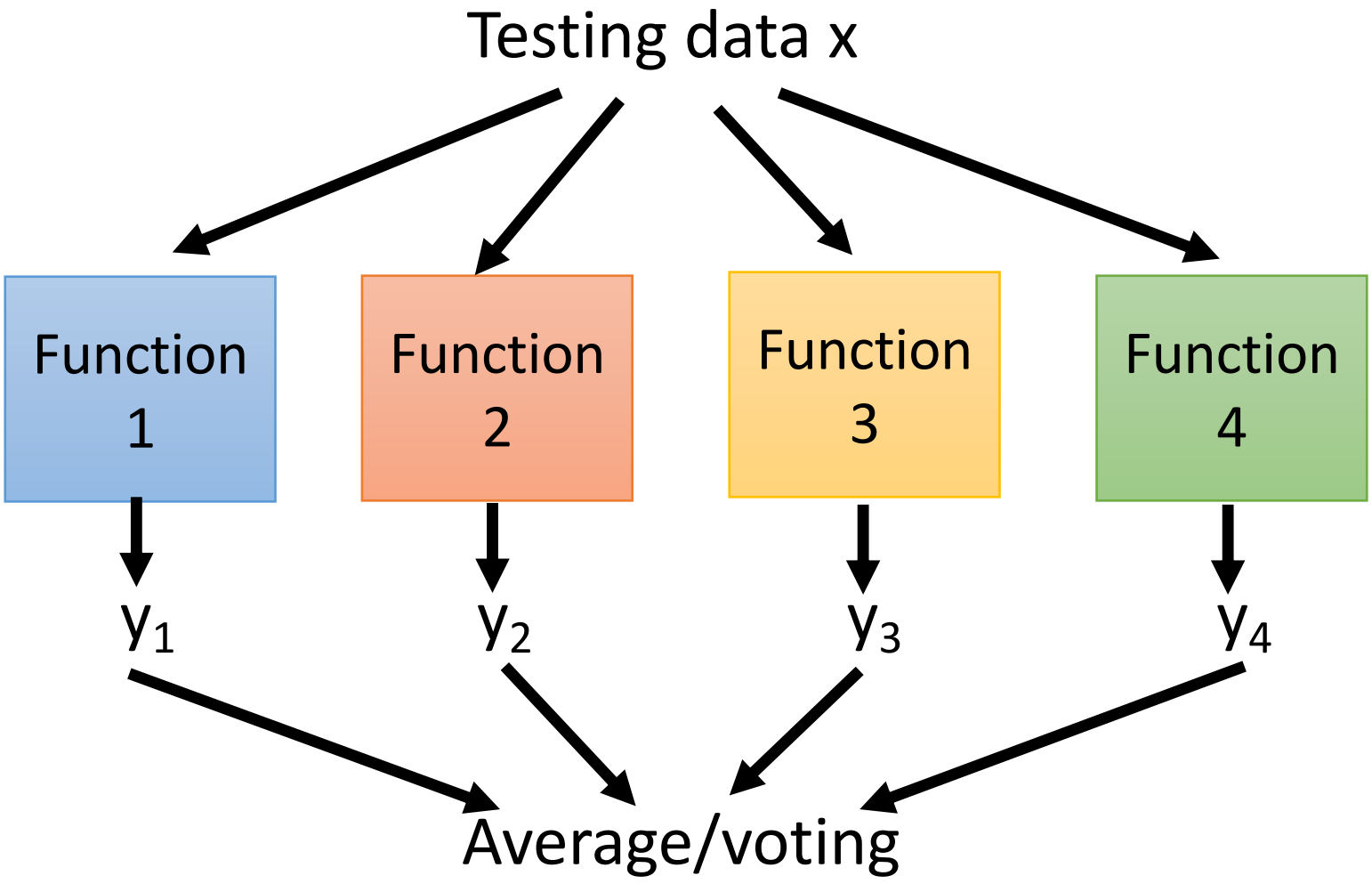
Bagging


N training examples
Sampling N' examples with replacement
(usually N=N')
Set 1
Set 2
Set 3
Set 4
Function 1
Function 2
Function 3
Function 4

Bagging

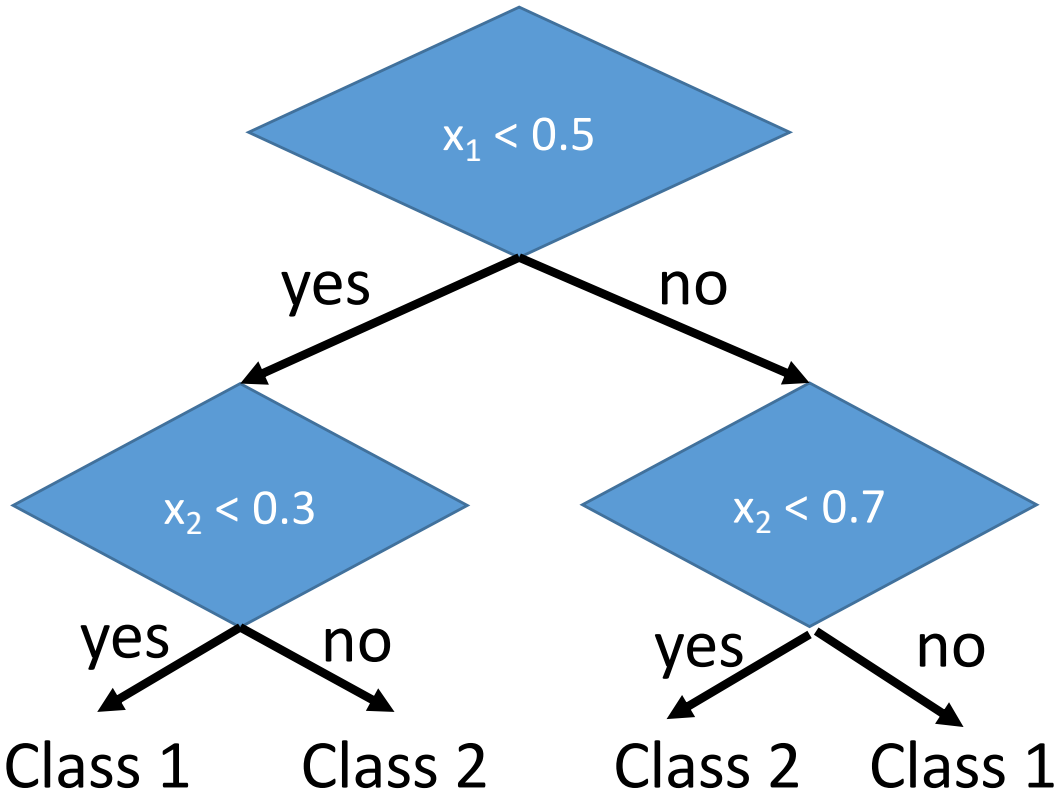
This approach would be helpful when your model is complex, easy to overfit.

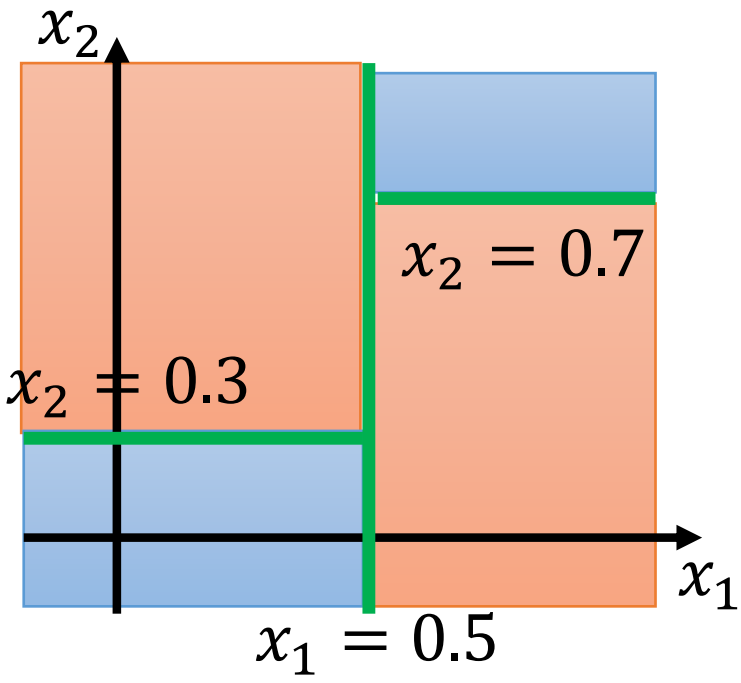
e.g. decision tree



Decision Tree

Assume each object x is represented by a 2-dim vector $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$



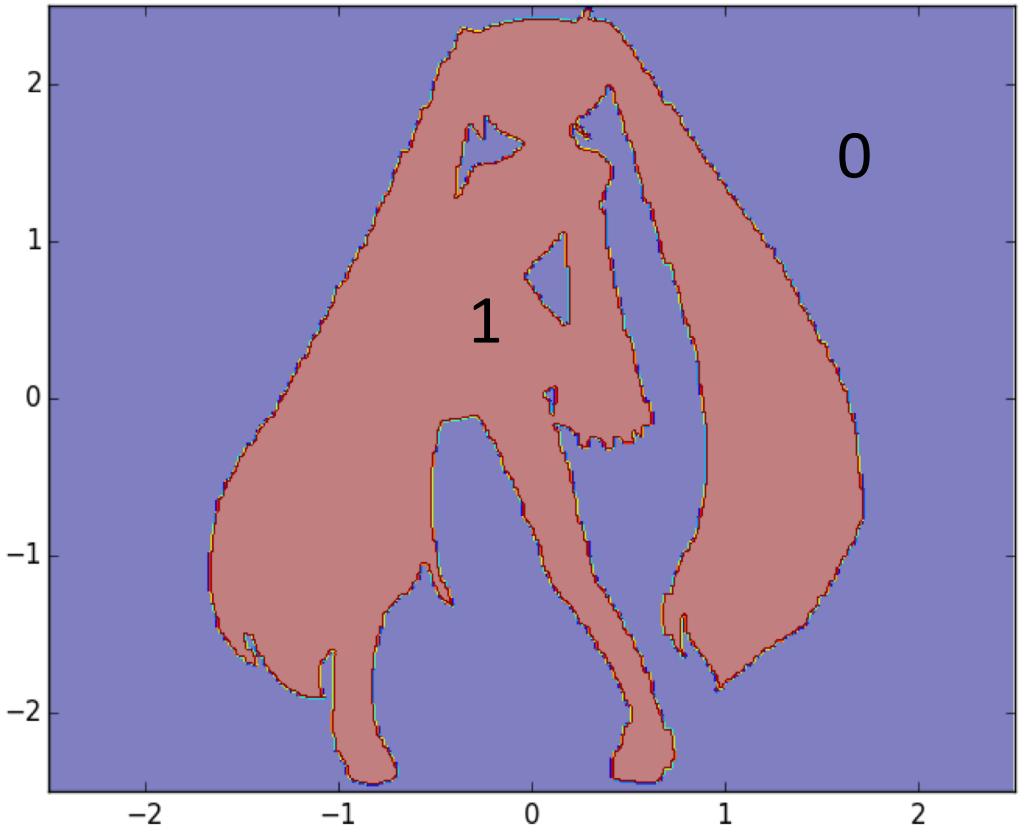


The questions in training .....

number of branches,
Branching criteria,
termination criteria,
base hypothesis

Can have more complex questions

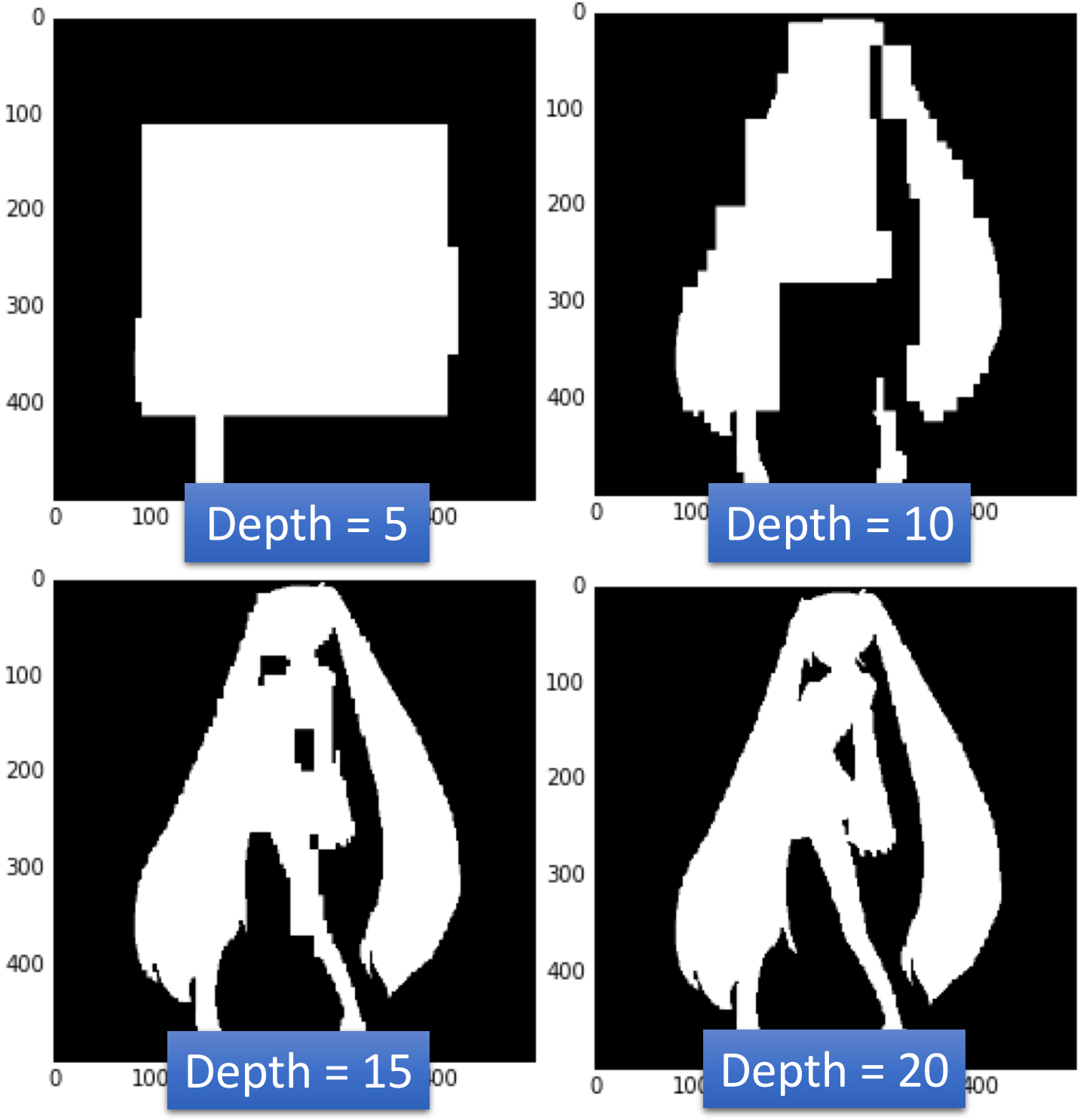
Experiment: Function of Miku



(Data: 1st column: x, 2nd column: y, 3rd column: output (1 or 0))

Experiment: Function of Miku

Single Decision Tree



Random Forest

train	f_1	f_2	f_3	f_4
x^1	O	X	O	X
x^2	O	X	X	O
x^3	X	O	O	X
x^4	X	O	X	O

- Decision tree:
  - Easy to achieve 0% error rate on training data
    - If each training example has its own leaf ……
- Random forest: Bagging of decision tree
  - Resampling training data is not sufficient
  - Randomly restrict the features/questions used in each split
- Out-of-bag validation for bagging
  - Using RF = f_2+f_4 to test x^1
  - Using RF = f_2+f_3 to test x^2
  - Using RF = f_1+f_4 to test x^3
  - Using RF = f_1+f_3 to test x^4

Out-of-bag (OOB) error

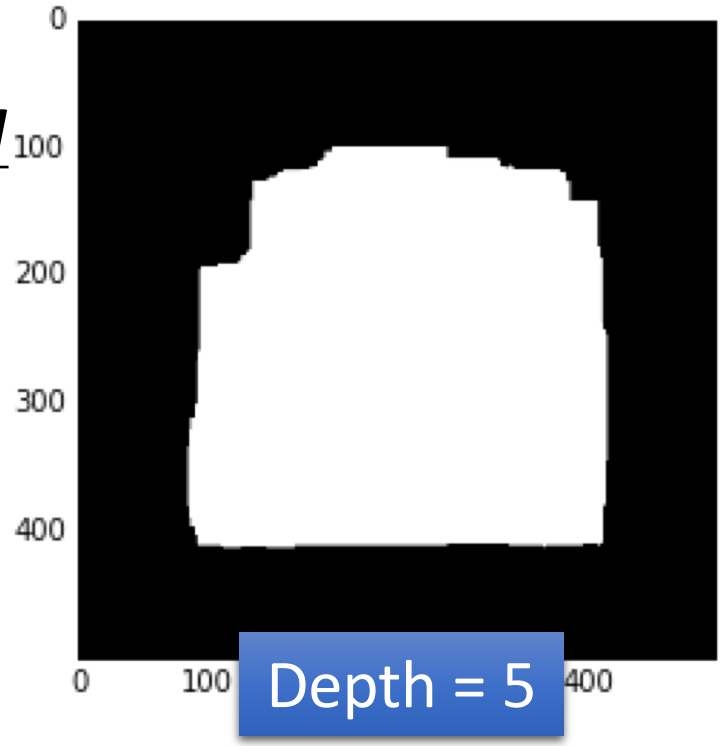
Good error estimation of testing set

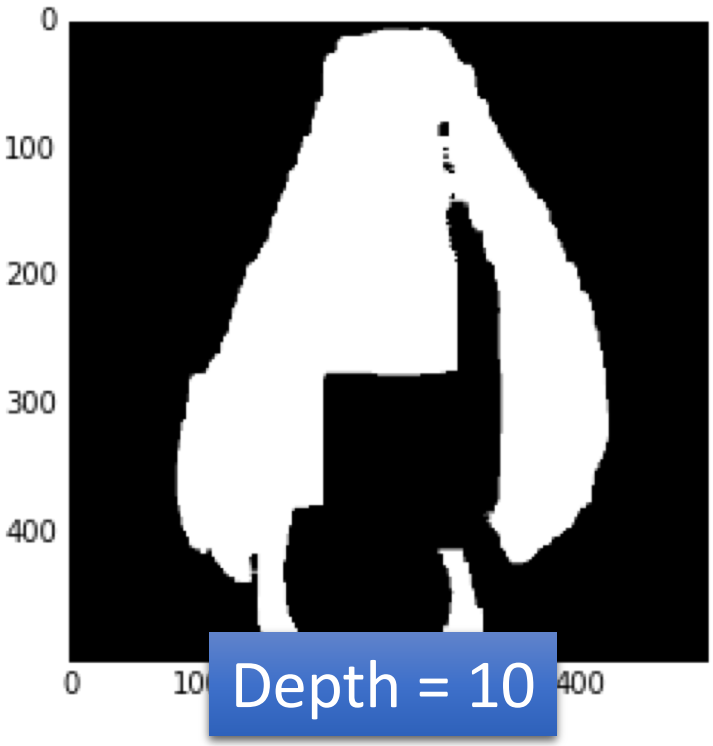
Experiment:

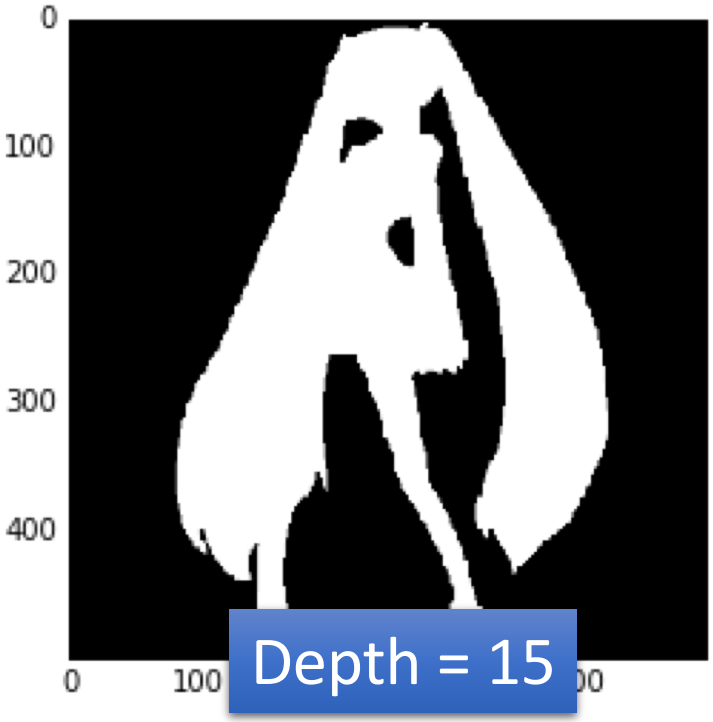
Function of Miku

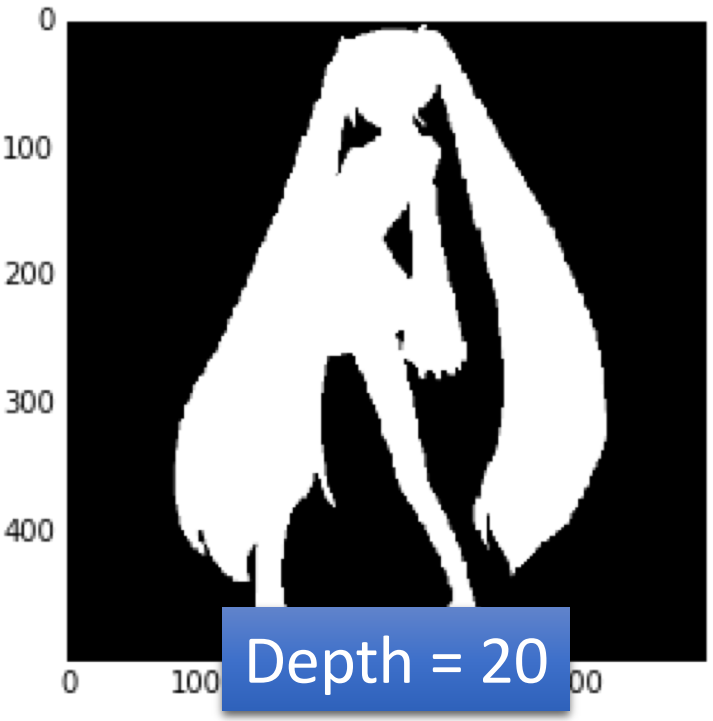
Random Forest

(100 trees)









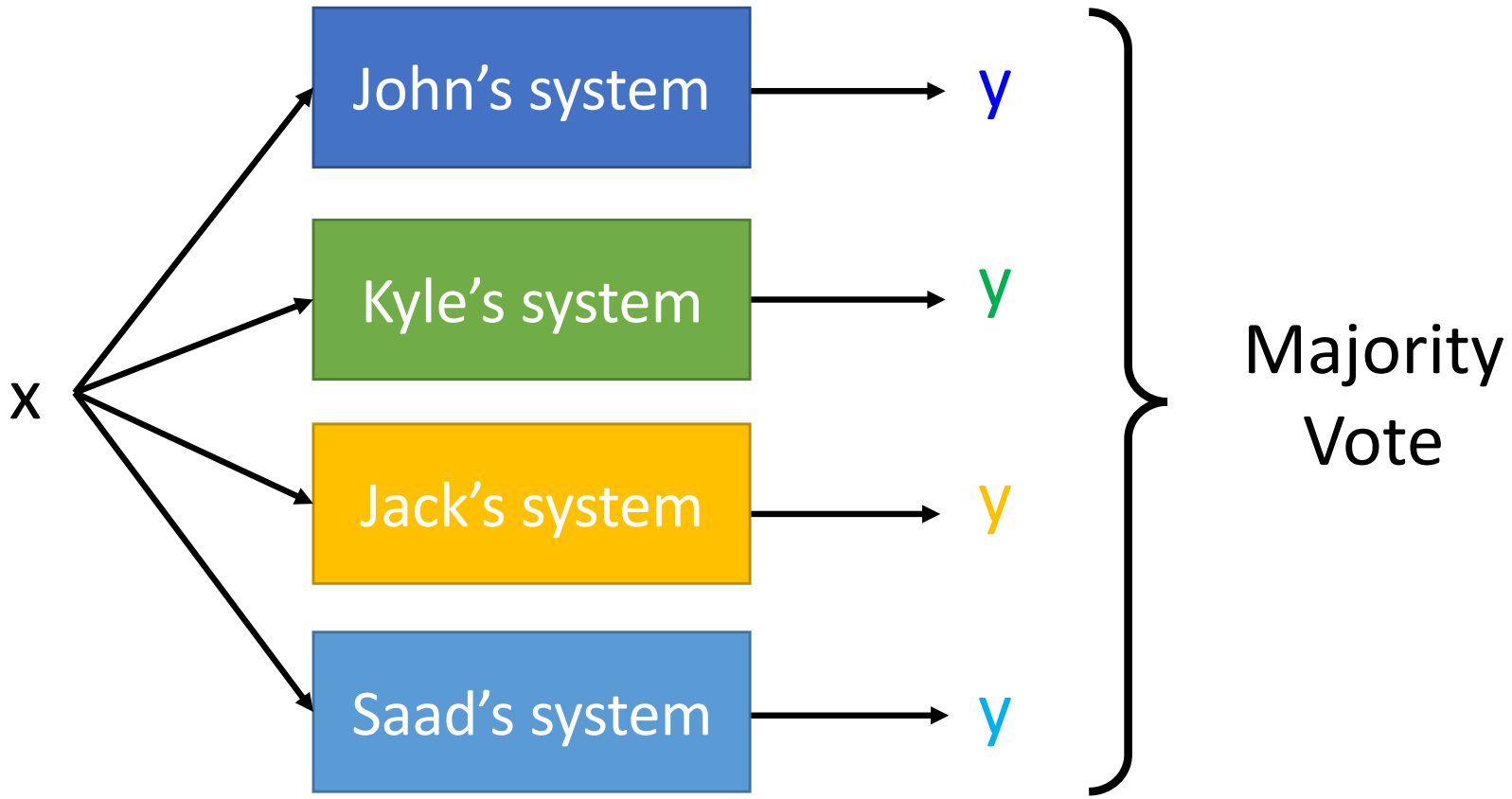
Outline

Ensemble: Bagging

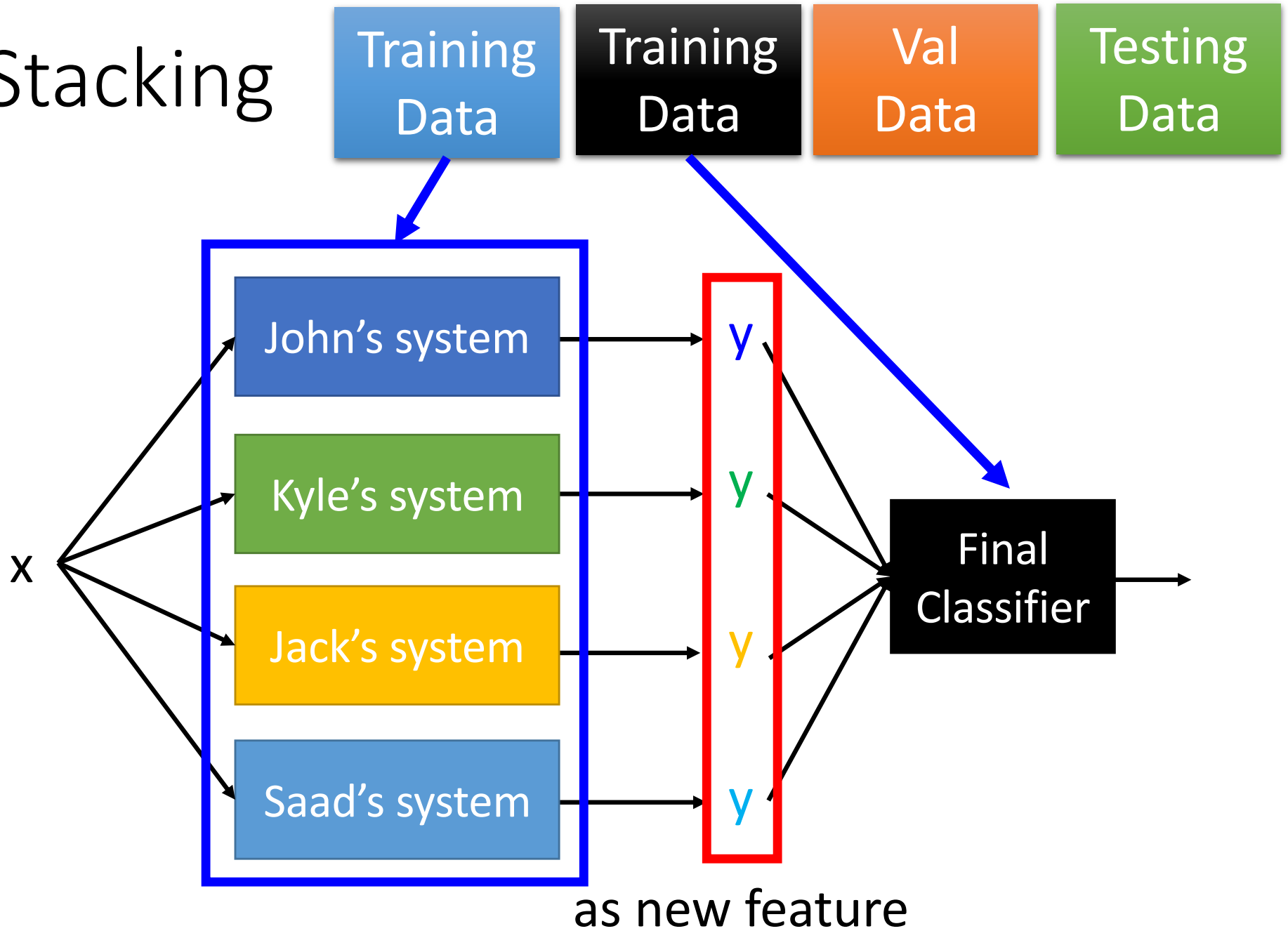
Ensemble: Stacking

Ensemble: Stacking

Voting


x
John's system
Kyle's system
Jack's system
Saad's system
y
y
y
y
Majority
Vote

Stacking


Training Data
Training Data
Val Data
Testing Data
John's system
Kyle's system
Jack's system
Saad's system
x
y
y
y
y
Final Classifier
as new feature

Questions