Trees, trees, trees

An introduction to classification tree, random forest, bagging, and boosting models

Huiwen Goy, June 2023

Dataset and problem

Why does someone decide to buy hearing aids?

	Age	Sex	Hearing test	Reported handicap	Stigma	x = 28	Purchased aids ⅕
Case 1	76	M	65	32	3	...	**1**
Case 2	61	M	45	26	4	...	**0**
Case 3	68	F	50	24	4	...	**0**
n = 753	...	...	...	...	...	...	...

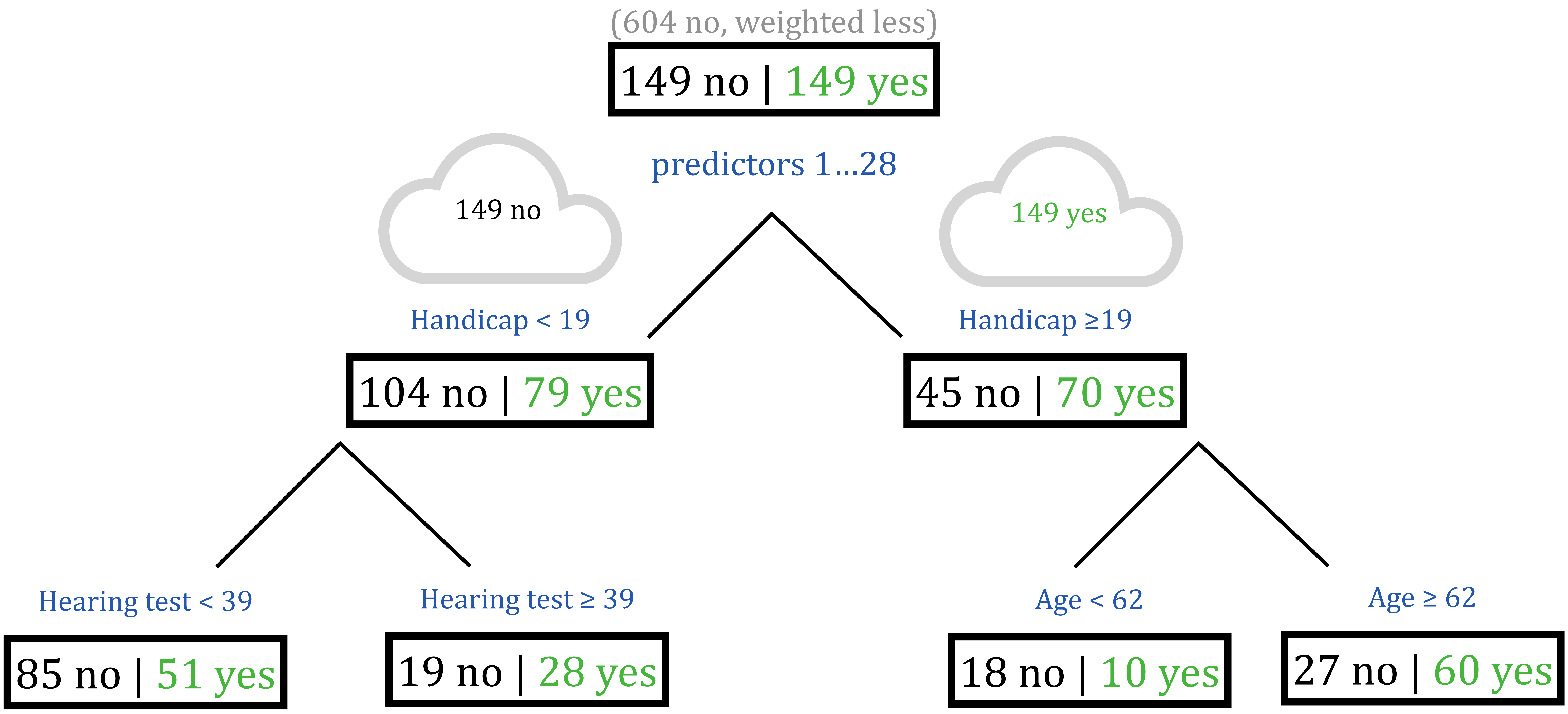
Initial analysis: Logistic regression

> glmnet (Purchased ~ Age + Sex + Hearing + Handicap… , data = df, family = “binomial”)

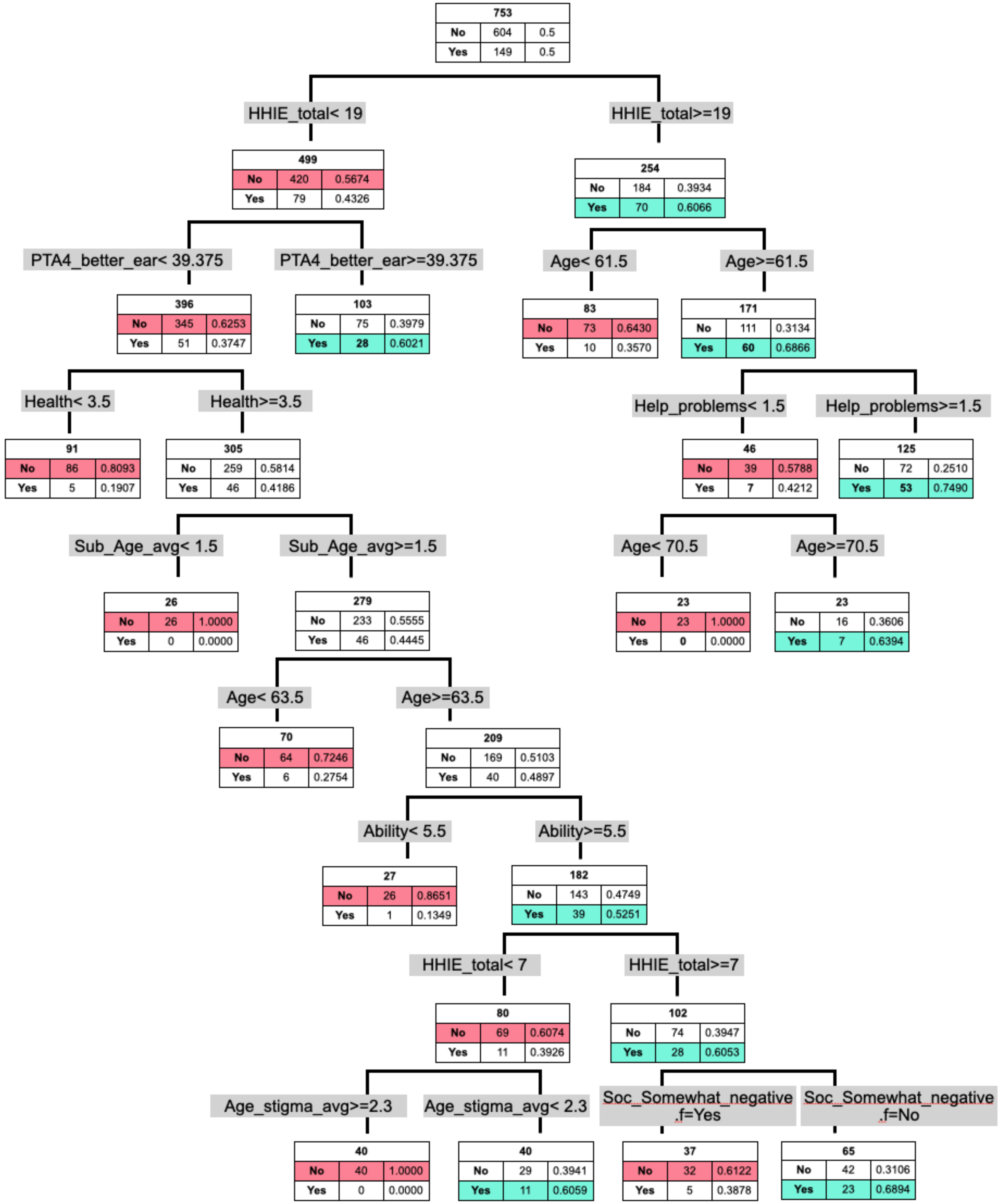
	Odds ratio	**CI lower**	**CI upper**	**p-value**
Age	1.046	1.024	1.069	<0.0001
Handicap	1.047	1.027	1.068	<0.0001
Stigma	0.85	0.71	1.01	0.065
Know someone	2.10	1.12	4.30	0.029

	LR
Accuracy	**63.5**
Sensitivity	**59.7**
Specificity	**64.4**

Classification tree (CART)



Complexity Parameter	Number of splits in tree	Overall accuracy	Sensitivity	Specificity	Area Under Curve
0.2	0	0.50	0.000	1.000	0.5000
0.1	1	0.6507	0.46980	0.69536	0.5826
0.05	3	0.672	0.5906	0.6921	0.6413
0.03	3	0.672	0.5906	0.6921	0.6413
0.025	3	0.672	0.5906	0.6921	0.6413
0.02	3	0.672	0.5906	0.6921	0.6413
0.015	5	0.7025	0.5906	0.7301	0.6604
0.013	12	0.6534	0.8188	0.6126	0.7157
0.012	12	0.6534	0.8188	0.6126	0.7157
0.011	15	0.6454	0.8792	0.5877	0.7335
0.01	22	0.7211	0.8725	0.6838	0.7781
0.005	31	0.745	0.9262	0.7003	0.8133
0	41	0.7822	0.9195	0.7483	0.8339



	LR	**Tree**
Accuracy	63.5	**67.5**
Sensitivity	59.7	**78.5**
Specificity	64.4	**64.7**

Model stability

Changes in model metrics (%) across data subsets, CP = 0.013

metric	subset_1	subset_2	subset_3	subset_4	subset_5
Accuracy	69.32	70.27	73.3	69.60	75.08
Sensitivity	73.95	63.03	72.5	67.23	82.35
Specificity	68.18	72.05	73.5	70.19	73.29
AUC	71.07	67.54	73.0	68.71	77.82

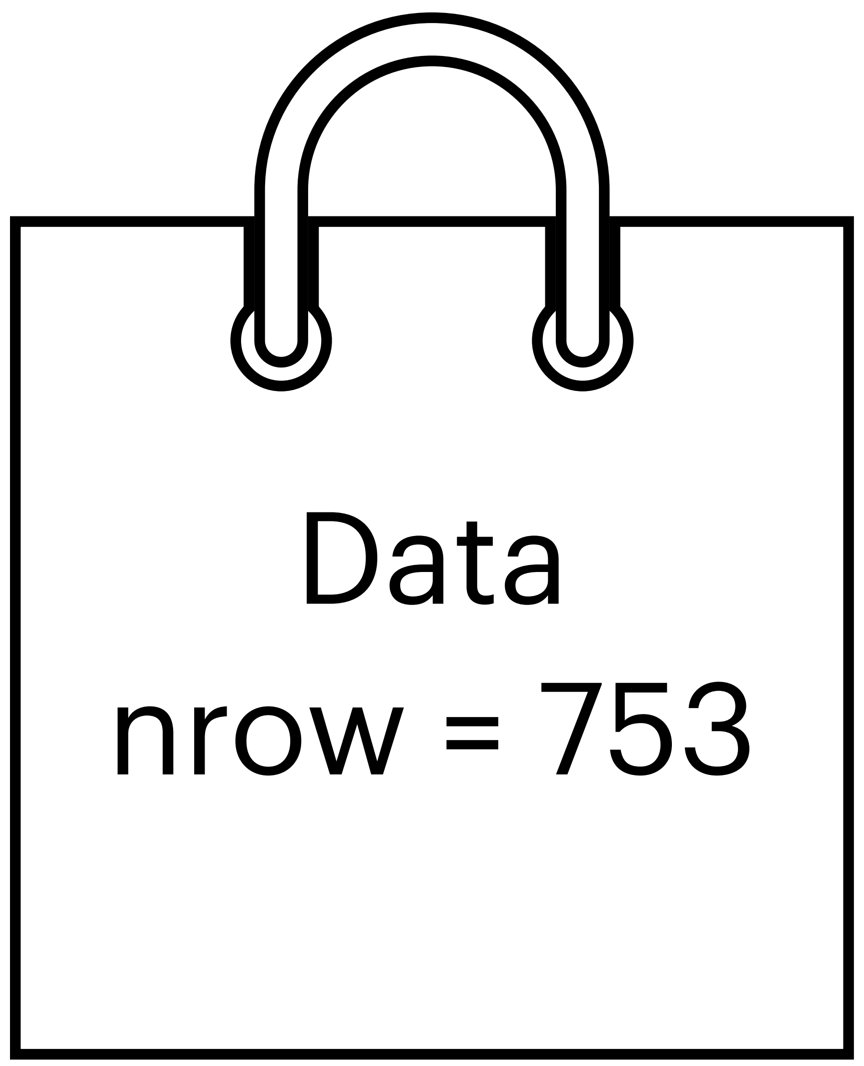
Changes in variable importance (%) across data subsets, CP = 0.013

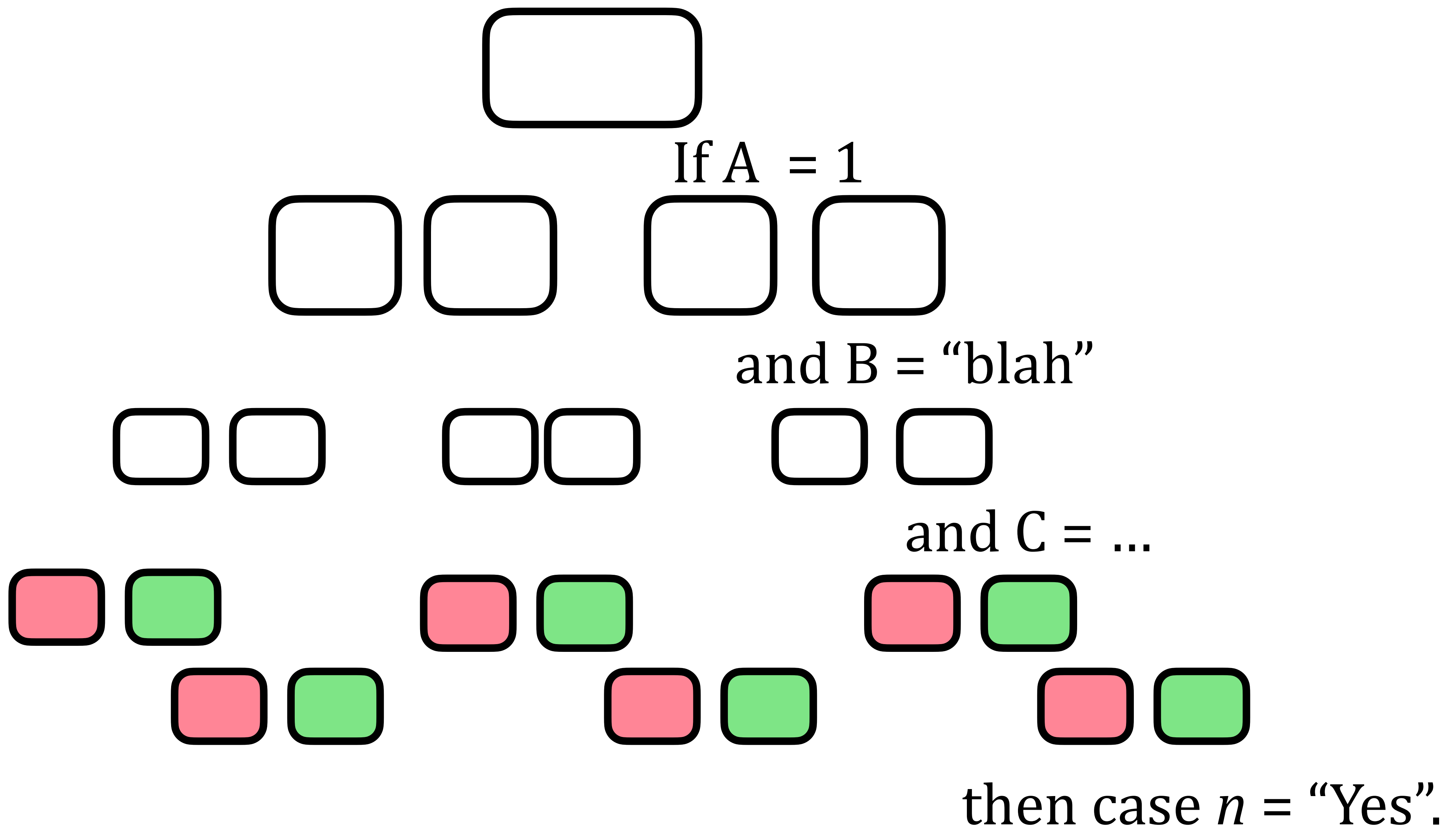
variable	subset_1	subset_2	subset_3	subset_4	subset_5
Ability	*	*	9.15	*	*
Accomp.f	*	*	*	*	*
Age	25.18	40.85	21.86	29.31	13.59
Age_stigma_avg	20.41	*	*	12.02	6.59
Concern	*	*	4.48	*	*
Edu	*	*	*	*	4.15
HA_stigma_avg	*	*	*	*	8.22
Health	*	*	*	*	4.61
Help_neighbours	*	*	*	*	*
Help_problems	11.92	11.64	*	9.96	*
HHIE_total	21.5	26.56	13.63	22.81	9.63
Lonely	*	*	*	*	5.19
Married.f	*	*	*	*	8.42
PTA4_better_ear	3.48	20.96	22.59	25.91	12.87
QoL	7.94	*	*	*	3.73

Pros & cons

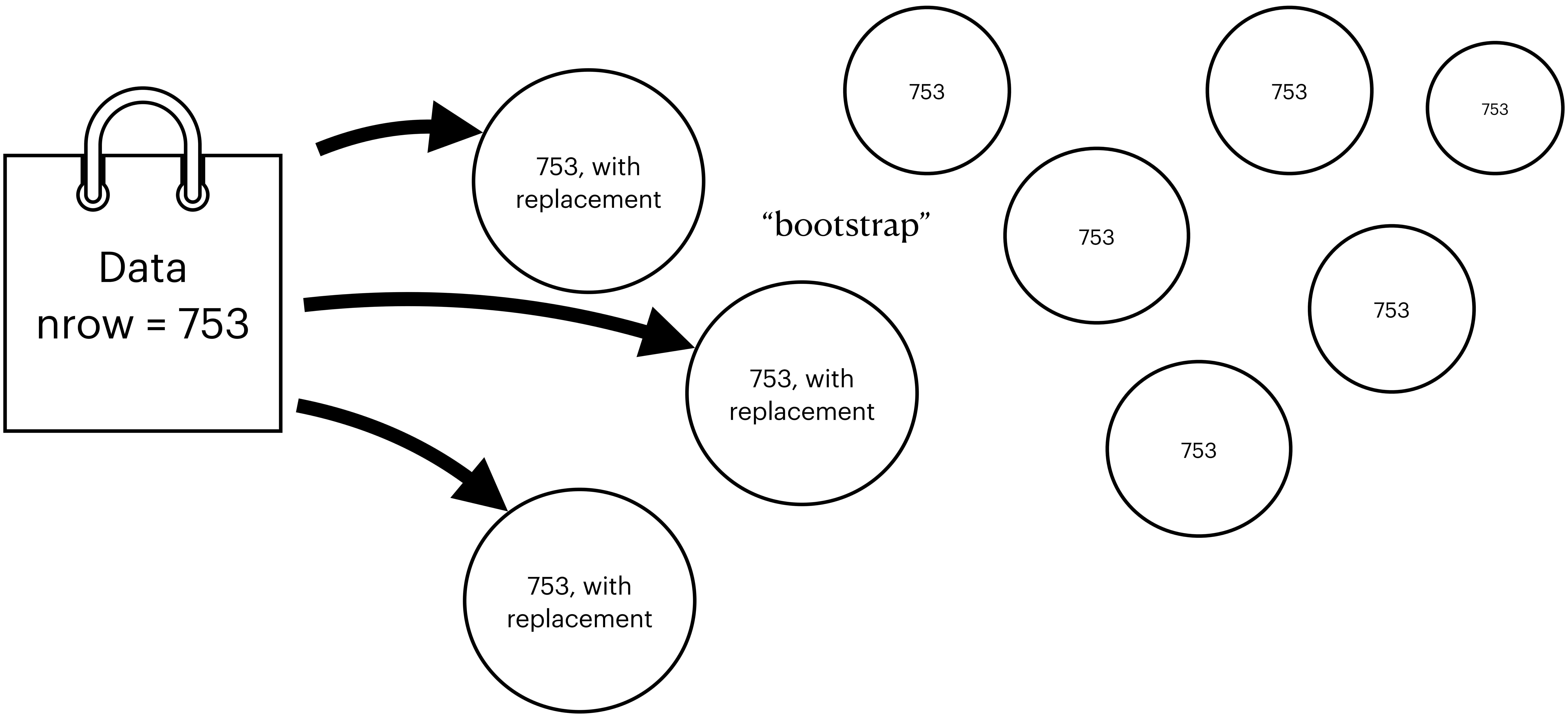
LR	Single tree				
Interpretation?	Easy to interpret High variance				

Classification tree

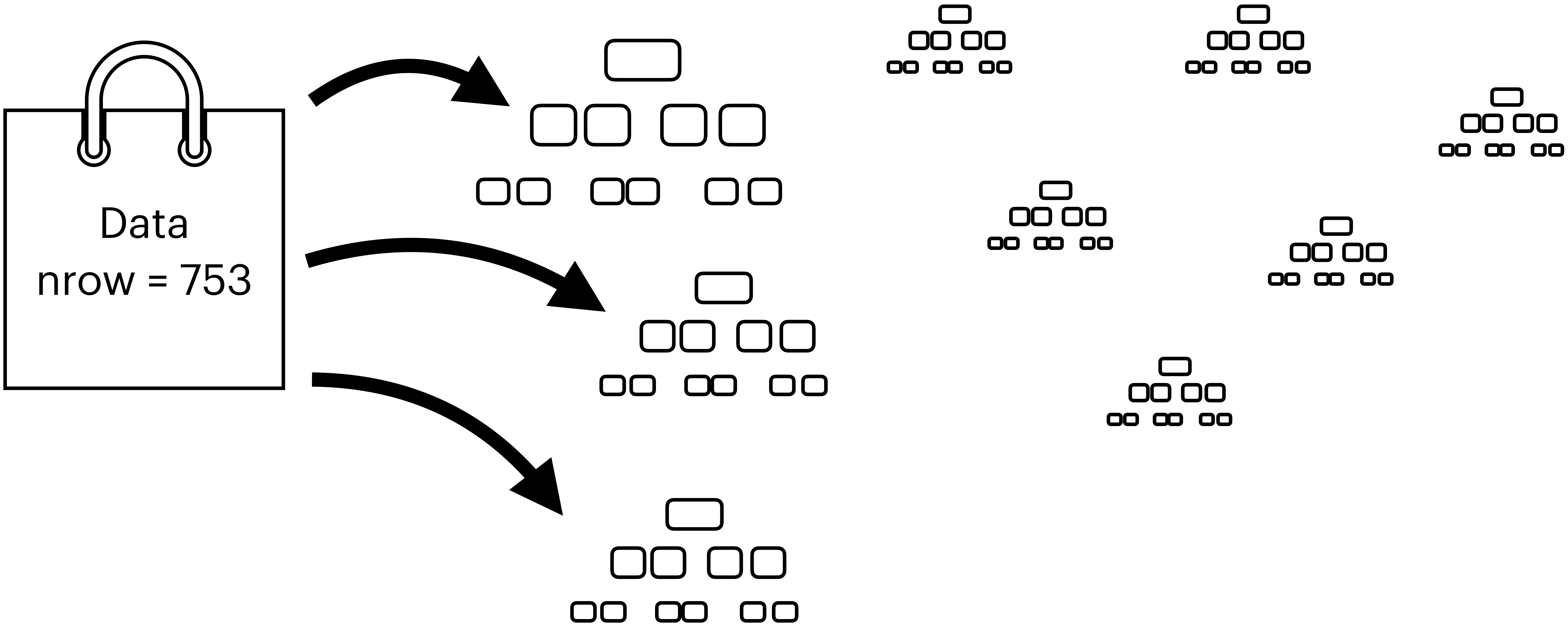




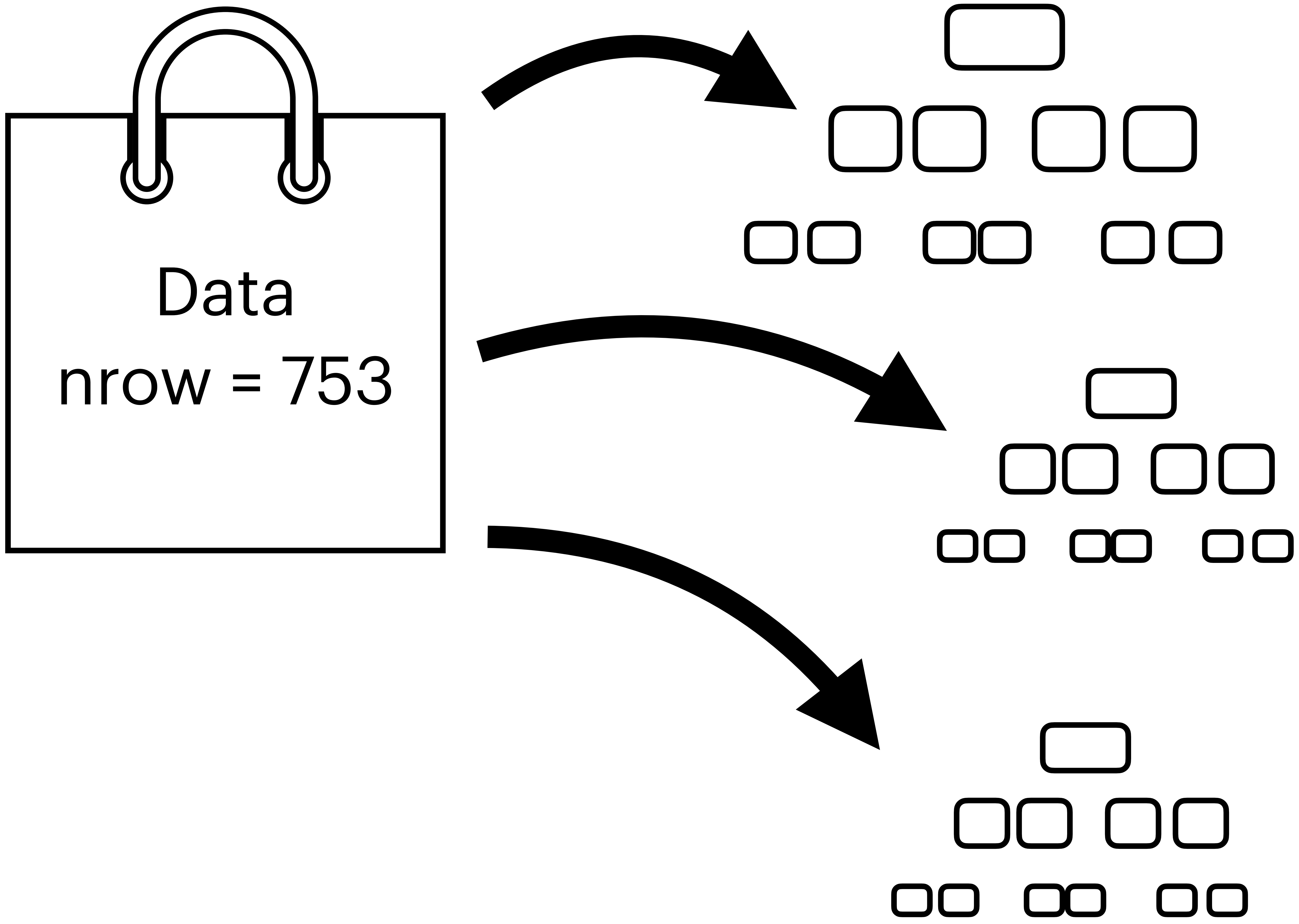
Bagging: Bootstrap aggregating



Bagging: Bootstrap aggregating



Bagging: Bootstrap aggregating



"aggregating"

	Tree 1	Tree 2	Tree ...	**Majority vote**
Case 1	Yes	No	No	**No**
Case 2	No	No	Yes	**No**
Case 3	Yes	No	Yes	**Yes**
...	No	Yes	No	**No**
Case 753	No	No	No	**No**

Bagging: Bootstrap aggregating

$ntrees = 250$

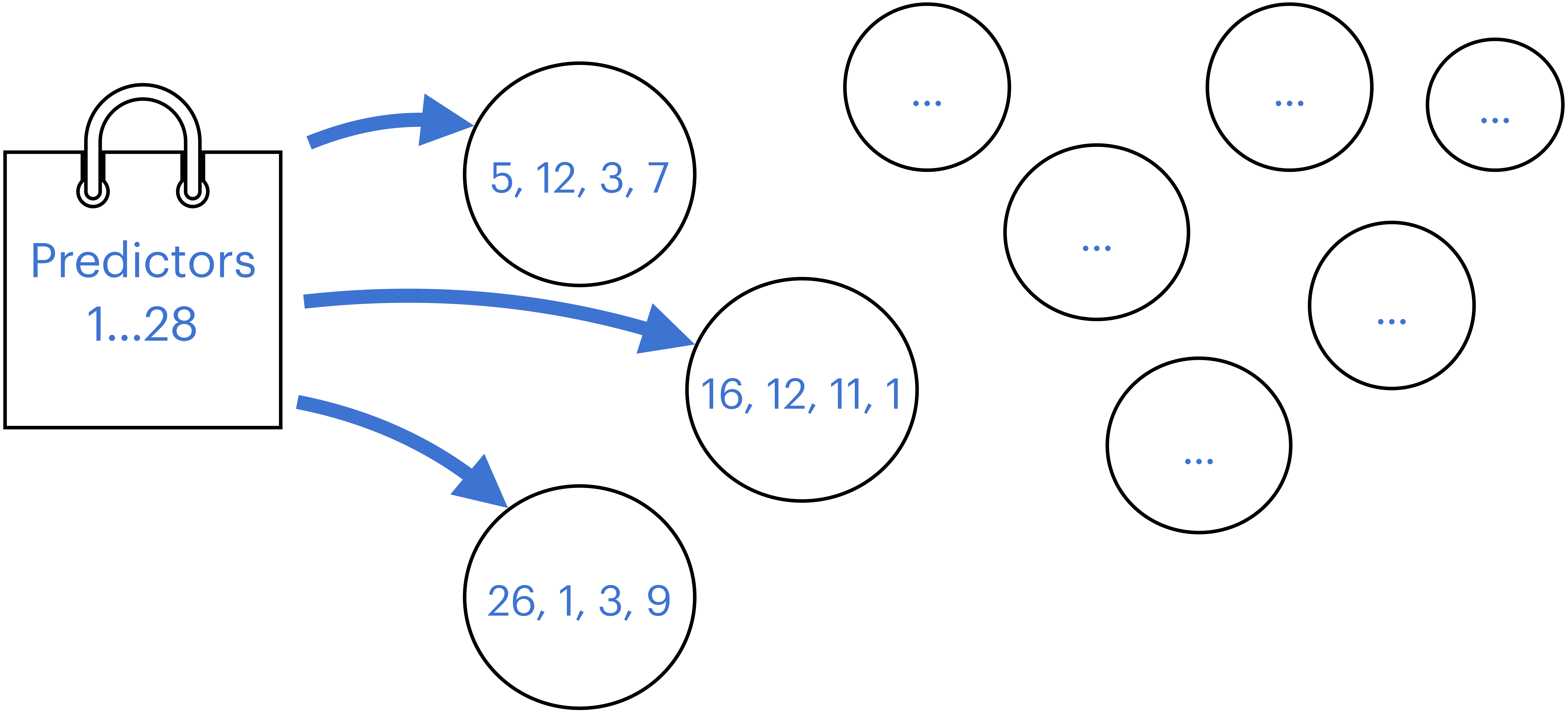
	LR	Tree	**Bag***
Accuracy	64	65	**77**
Sensitivity	59	81	**32**
Specificity	64	61	**89**

*Downsampled majority class

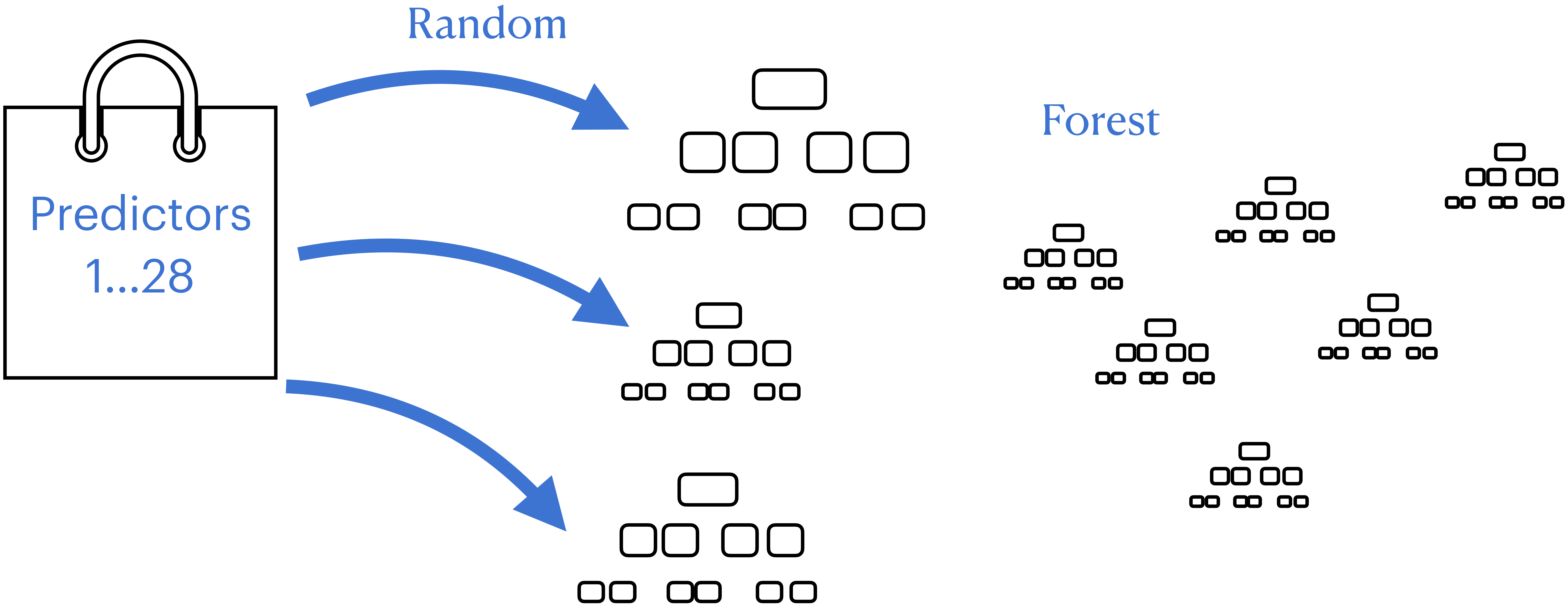
Pros & cons

LR	Single tree	Bagging		
Interpretation?	Easy to interpret High variance	Less easy to interpret Less variance All trees look alike... less accurate		

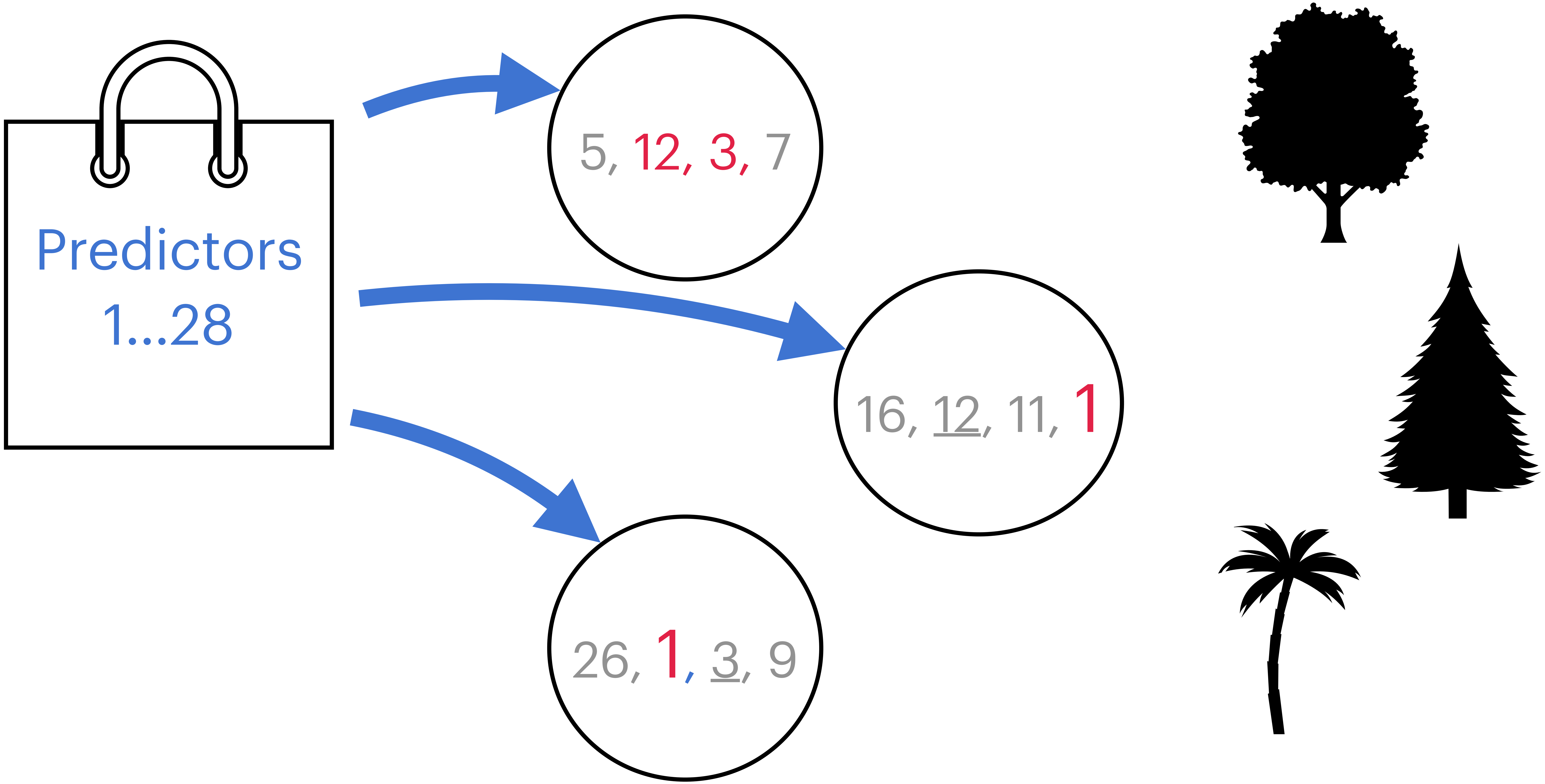
Random forest



Random forest



Random forest



Random forest

$ntrees = 200$
$m = 4$

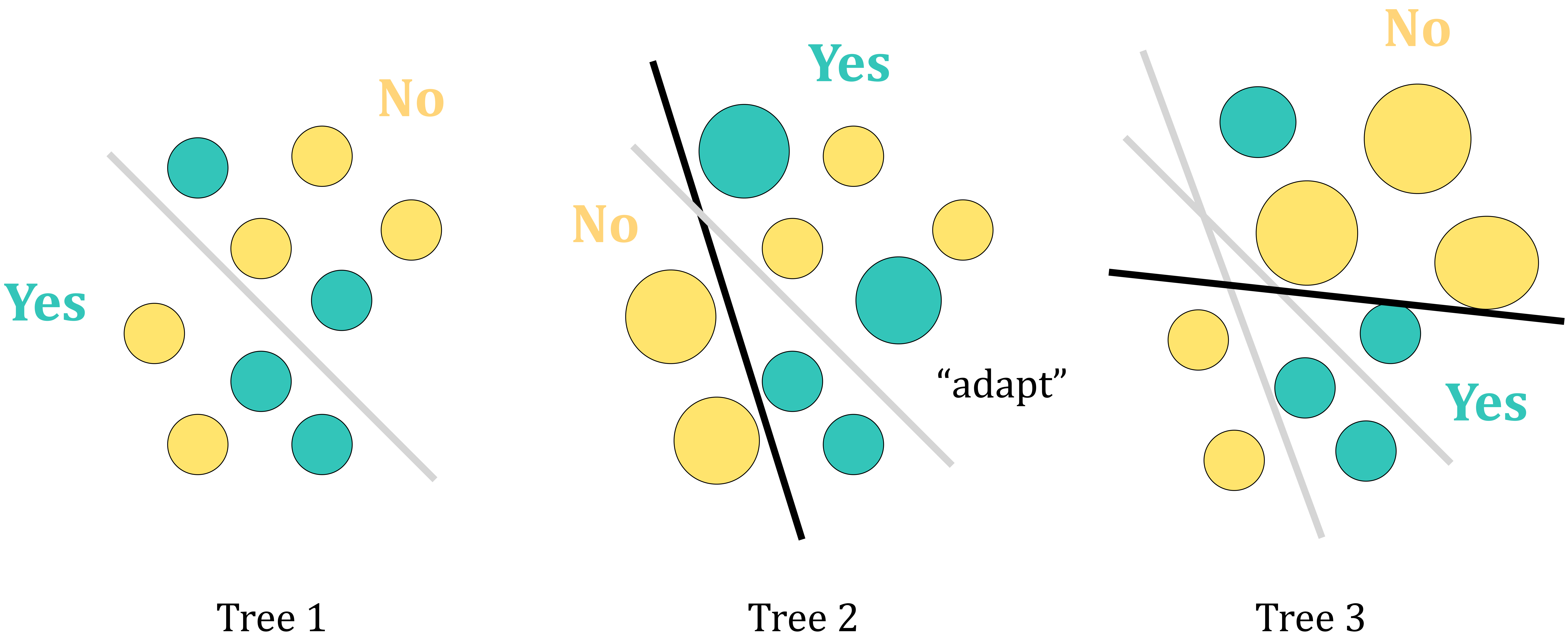
	LR	Tree	Bag	**RF***
Accuracy	64	65	77	**98**
Sensitivity	60	81	32	**100**
Specificity	64	61	89	**98**

*Oversampling of minority class
*Beware inflation; OOB accuracy = 77

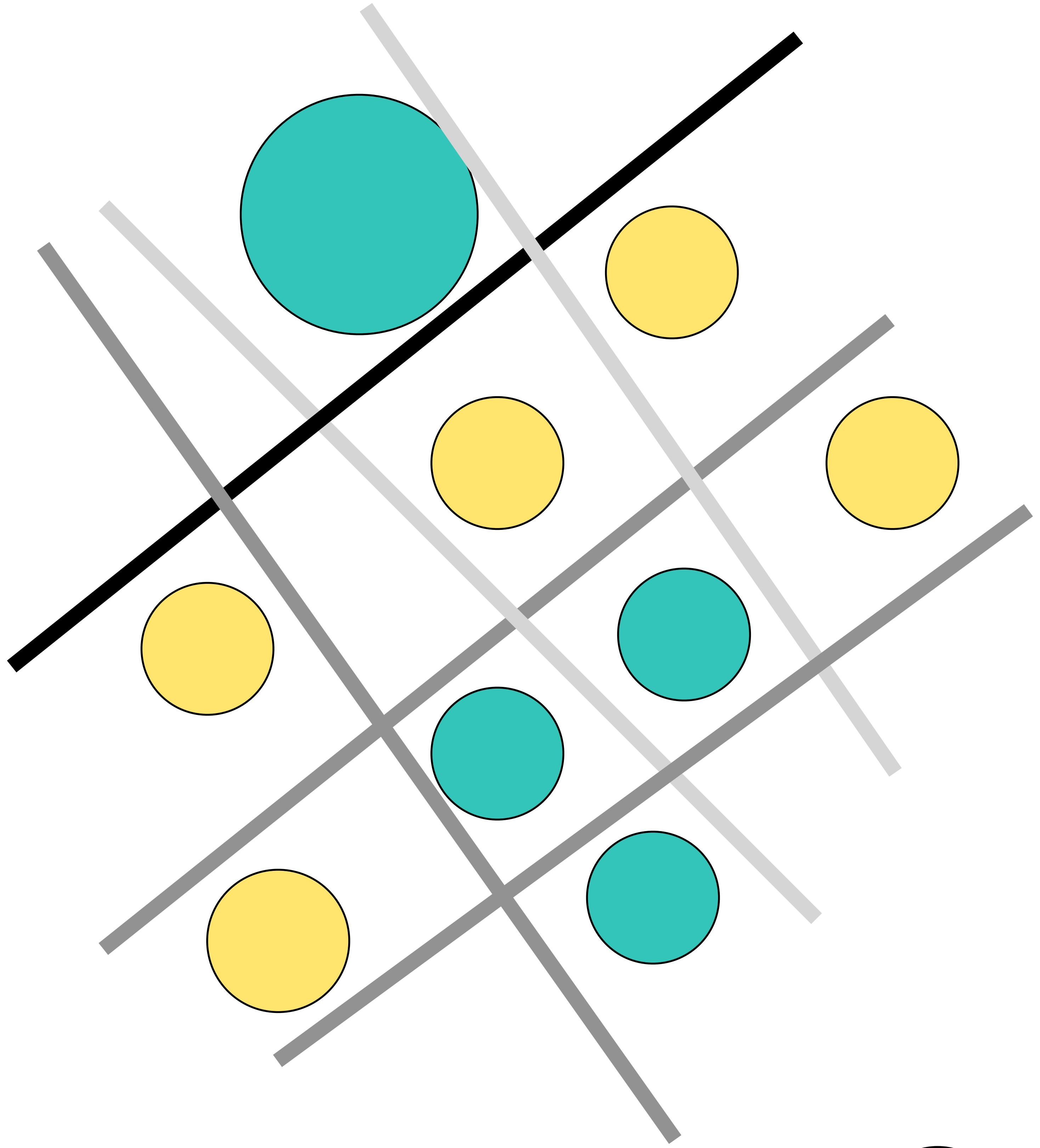
Pros & cons

LR	Single tree	Bagging	Random forest	
Interpretation?	Easy to interpret High variance	Less easy to interpret Less variance All trees look alike… less accurate	Less easy to interpret Less variance De-correlated trees; more accurate	

AdaBoost: Adaptive boosting



AdaBoost: Adaptive boosting



Final model = (W1)Tree 1 + (W2)Tree 2 + (W3)Tree 3

Tree *Gazillion!* Over-fitting

AdaBoost: Adaptive boosting

ntrees = 5
maxdepth = 1
ν = 0.25

	LR	Tree	Bag	RF	**Ada***	
Accuracy	64	65	77	98	**80**	
Sensitivity	60	81	32	100	**0**	
Specificity	64	61	89	98	**1**	

*No case weights!

Gradient boosting

"Tree" 1

	Age	Sex	Hearing test	…	Initial predict	Actual	Error so far
Case 1	76	M	65	…	**-0.482**	**1**	1.482
Case 2	61	M	45	…	**-0.482**	**0**	0.482
Case 3	68	F	50	…	**-0.482**	**0**	0.482
…	…	…	…	…	…	…	…

Gradient boosting

	Error so far	Tree2 predict	Error so far	Tree3 predict	Error so far	TreeX predict	Final error
Case 1	1.482		1.263		0.994		-0.05
Case 2	0.482		0.388		0.268		0.03
Case 3	0.482		0.425		0.379		0.01
...	...						

Gradient boosting

Final model = Tree 1 + W * Tree 2 + W * Tree 3 + …

(Individual cases are not weighted, like in AdaBoost)

(Weight for all trees is the same, the “learning rate”)

Gradient boosting

ntrees = 10
maxdepth = 3
ν = 0.1

	LR	Tree	Bag	RF	Ada	**Gra***
Accuracy	64	65	77	98	80	**68**
Sensitivity	60	81	32	100	0	**50**
Specificity	64	61	89	98	1	**72**

*Case weights, but conditional inference trees

Pros & cons

LR	Single tree	Bagging	Random forest	Boosting
Interpretation?	Easy to interpret	Less easy to interpret	Less easy to interpret	Less easy to interpret
	High variance	Less variance	Less variance	Less variance
		All trees look alike… less accurate	De-correlated trees; more accurate	Build on previous trees; accurate??
			Can’t overfit	Possible to overfit

Trees, trees, trees

- Ensembles are usually better than a single tree; “wisdom of the crowd”
- Consider the quirks of the dataset
- How useful are the results from that method?
- How open is the field to that method?
- Each method has parameters to tune; see GitHub