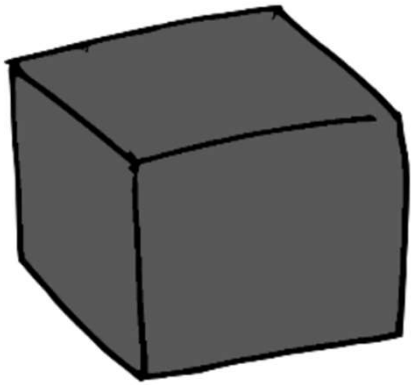
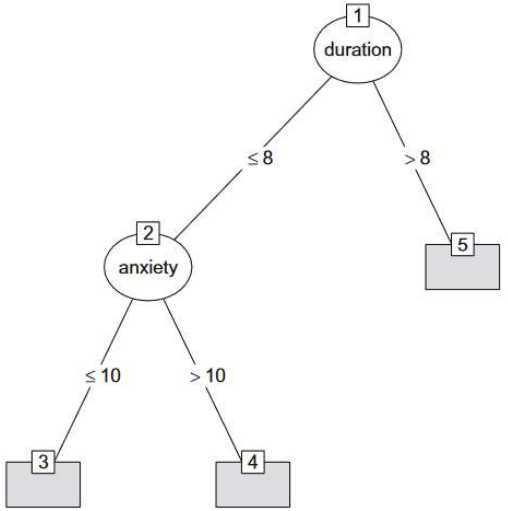
X-AI/ML: Two approaches





1. Fit black box, use explanation tools.

+ Predictive accuracy
- Depend on permutation of predictors
- ‘Explanations’ not understood by users
- Can explain effects of many variables?

2. Fit interpretable model.

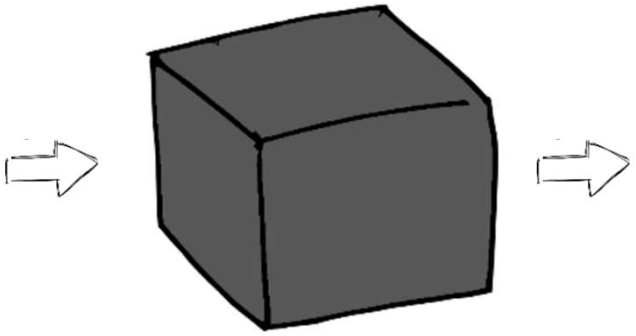
- Predictive accuracy
+ Better understood
+ Less variables in model

Bridging the two approaches

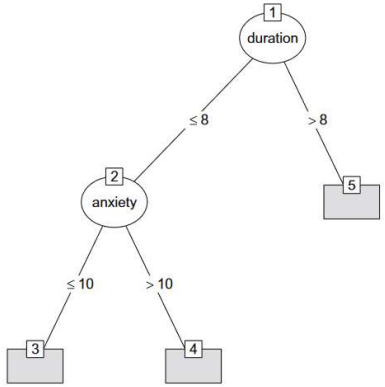
Fit tree on predictions of black box, instead of original outcome *y*

(Breiman & Shang, 1996: Born again trees)

	x1	x2	y
1	5	2	8
2	9	0	3
3	2	1	6



	x1	x2	$\hat{y}$
1	5	2	7
2	9	0	4
3	2	1	5

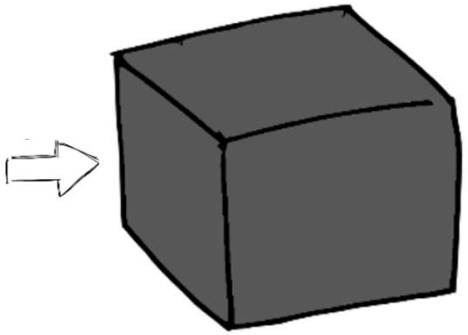


-> More stable and accurate tree

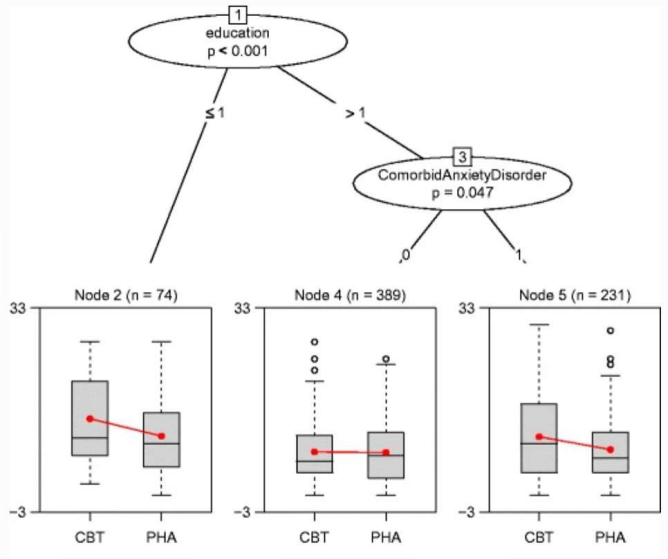
Treatment outcomes

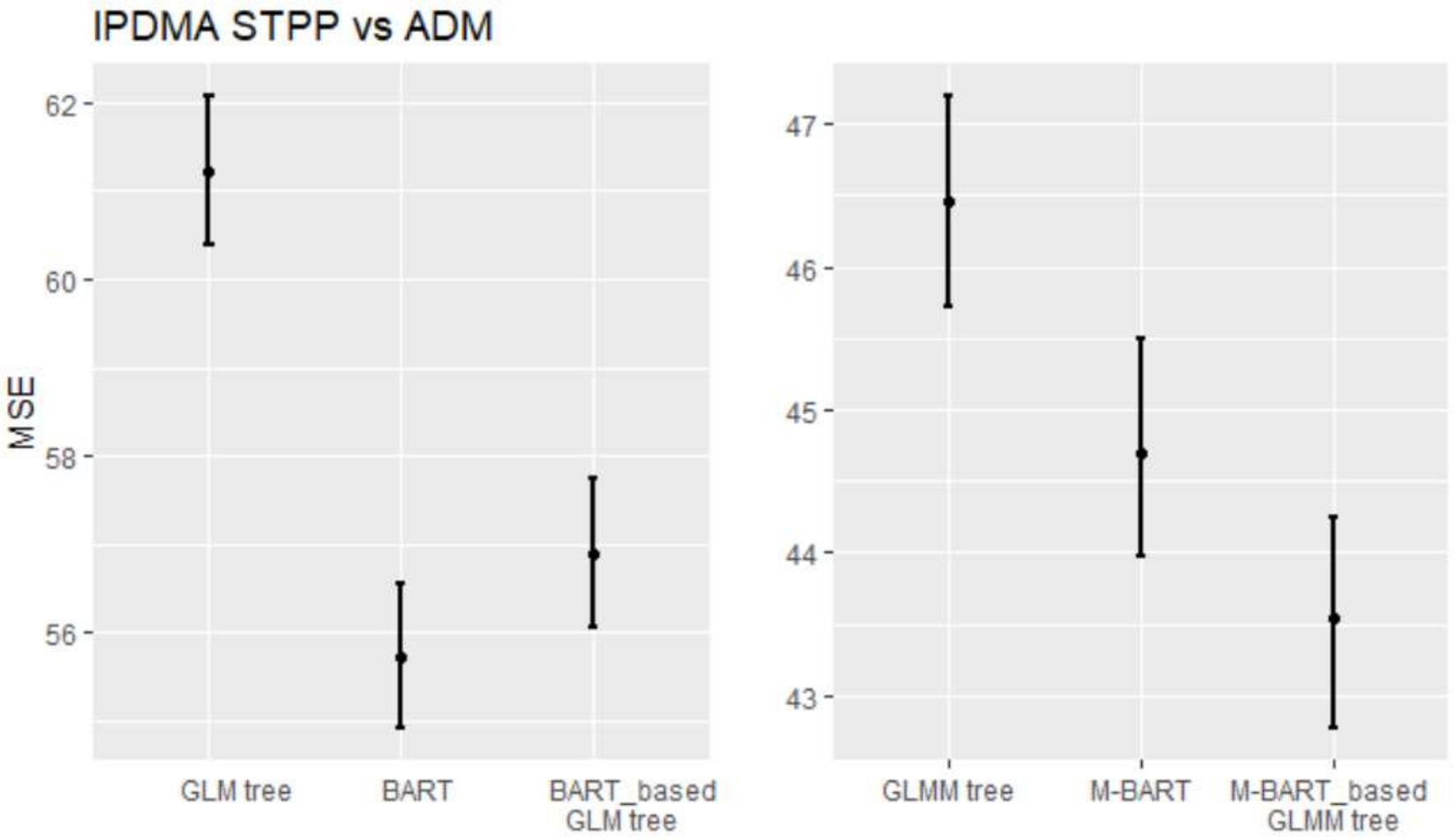
Fit tree on predictions of black box, instead of original outcome *y*

	x1	x2	y
1	5	2	8
2	9	0	3
3	2	1	6



	x1	x2	$\hat{y}$
1	5	2	7
2	9	0	4
3	2	1	5





Fixed effects

Mixed effects

Discussion points

Human behavior is infinitely complex, should predictive models be?

- When would black box (+ explanation tool) be preferable?
- When would simpler, interpretable method be preferable?

What would be a relevant gain in accuracy when predicting treatment outcomes? Can we quantify?