REGRESSION TREES

Predictive Analytics

- Classification
- Regression
- Clustering

CART models

- Classification Trees
- Regression Trees

Example

Salary	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns
475	81	7	24	38	39	14	3449	835	69	321
480	130	18	66	72	76	3	1624	457	63	224
500	141	20	65	78	37	11	5628	1575	225	828
91.5	87	10	39	42	30	2	396	101	12	48
750	169	4	74	51	35	11	4408	1133	19	501
70	37	1	23	8	21	2	214	42	1	30
100	73	0	24	24	7	3	509	108	0	41
75	81	6	26	32	8	2	341	86	6	32
1100	92	17	49	66	65	13	5206	1332	253	784
517.143	159	21	107	75	59	10	4631	1300	90	702
512.5	53	4	31	26	27	9	1876	467	15	192
550	113	13	48	61	47	4	1512	392	41	205
700	60	0	30	11	22	6	1941	510	4	309
240	43	7	29	27	30	13	3231	825	36	376
775	158	20	89	75	73	15	8068	2273	177	1045
175	46	2	24	8	15	5	479	102	5	65
135	32	8	16	22	14	8	727	180	24	67
100	92	16	72	48	65	1	413	92	16	72

Example

Salary	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns
475	81	7	24	38	39	14	3449	835	69	321
480	130	18	66	72	76	3	1624	457	63	224
500	141	20	65	78	37	11	5628	1575	225	828
91.5	87	10	39	42	30	2	396	101	12	48
750	169	4	74	51	35	11	4408	1133	19	501
70	37	1	23	8	21	2	214	42	1	30
100	73	0	24	24	7	3	509	108	0	41
75	81	6	26	32	8	2	341	86	6	32
1100	92	17	49	66	65	13	5206	1332	253	784
517.143	159	21	107	75	59	10	4631	1300	90	702
512.5	53	4	31	26	27	9	1876	467	15	192
550	113	13	48	61	47	4	1512	392	41	205
700	60	0	30	11	22	6	1941	510	4	309
240	43	7	29	27	30	13	3231	825	36	376
775	158	20	89	75	73	15	8068	2273	177	1045
175	46	2	24	8	15	5	479	102	5	65
135	32	8	16	22	14	8	727	180	24	67
100	92	16	72	48	65	1	413	92	16	72

Example

`AtBat`
: Number of times at bat in 1986

`Hits`
: Number of hits in 1986

`HmRun`
: Number of home runs in 1986

`Runs`
: Number of runs in 1986

`RBI`
: Number of runs batted in in 1986

`Walks`
: Number of walks in 1986

`Years`
: Number of years in the major leagues

`CAtBat`
: Number of times at bat during his career

`CHits`
: Number of hits during his career

`CHmRun`
: Number of home runs during his career

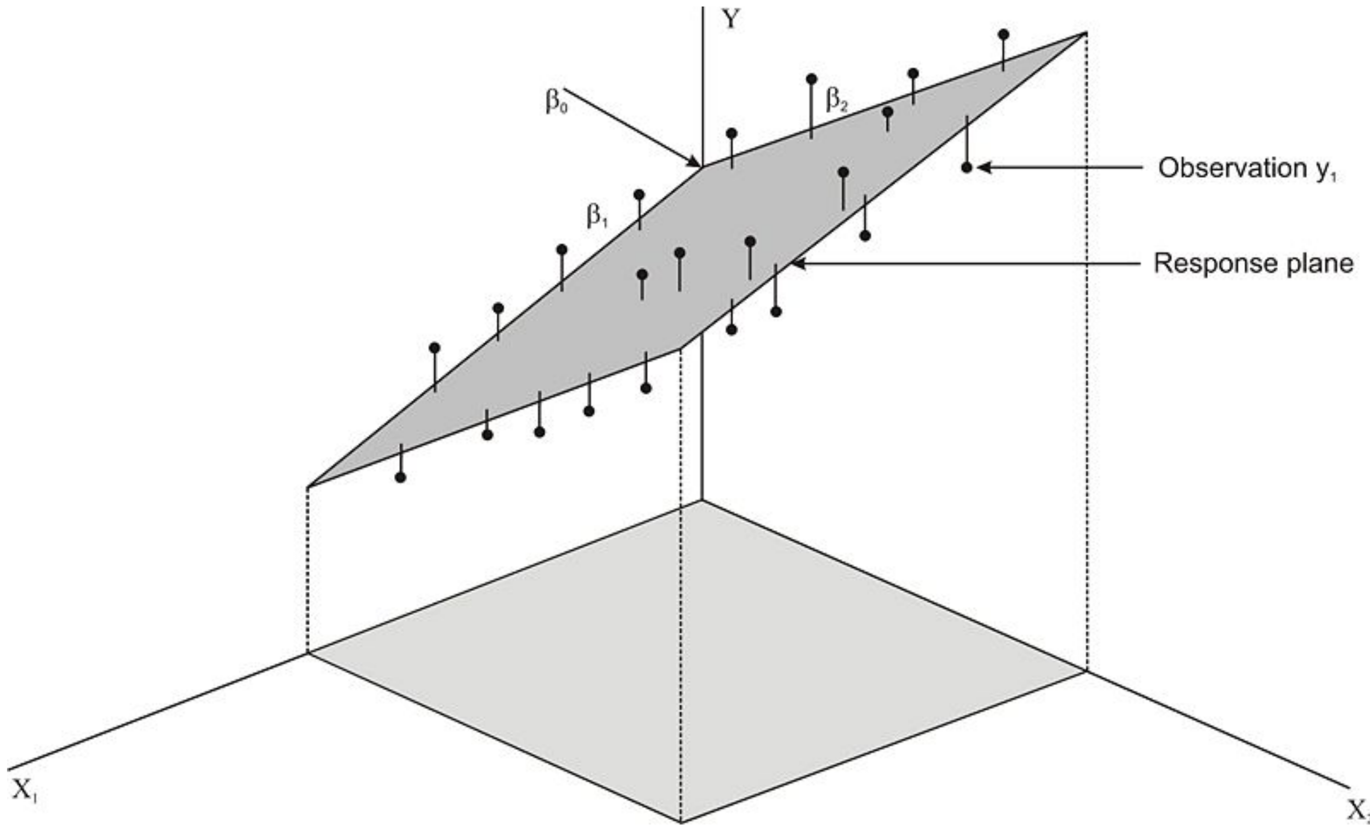
Example

Salary	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns
475	81	7	24	38	39	14	3449	835	69	321
480	130	18	66	72	76	3	1624	457	63	224
500	141	20	65	78	37	11	5628	1575	225	828
91.5	87	10	39	42	30	2	396	101	12	48
750	169	4	74	51	35	11	4408	1133	19	501
70	37	1	23	8	21	2	214	42	1	30
100	73	0	24	24	7	3	509	108	0	41
75	81	6	26	32	8	2	341	86	6	32
1100	92	17	49	66	65	13	5206	1332	253	784
517.143	159	21	107	75	59	10	4631	1300	90	702
512.5	53	4	31	26	27	9	1876	467	15	192
550	113	13	48	61	47	4	1512	392	41	205
700	60	0	30	11	22	6	1941	510	4	309
240	43	7	29	27	30	13	3231	825	36	376
775	158	20	89	75	73	15	8068	2273	177	1045
175	46	2	24	8	15	5	479	102	5	65
135	32	8	16	22	14	8	727	180	24	67
100	92	16	72	48	65	1	413	92	16	72

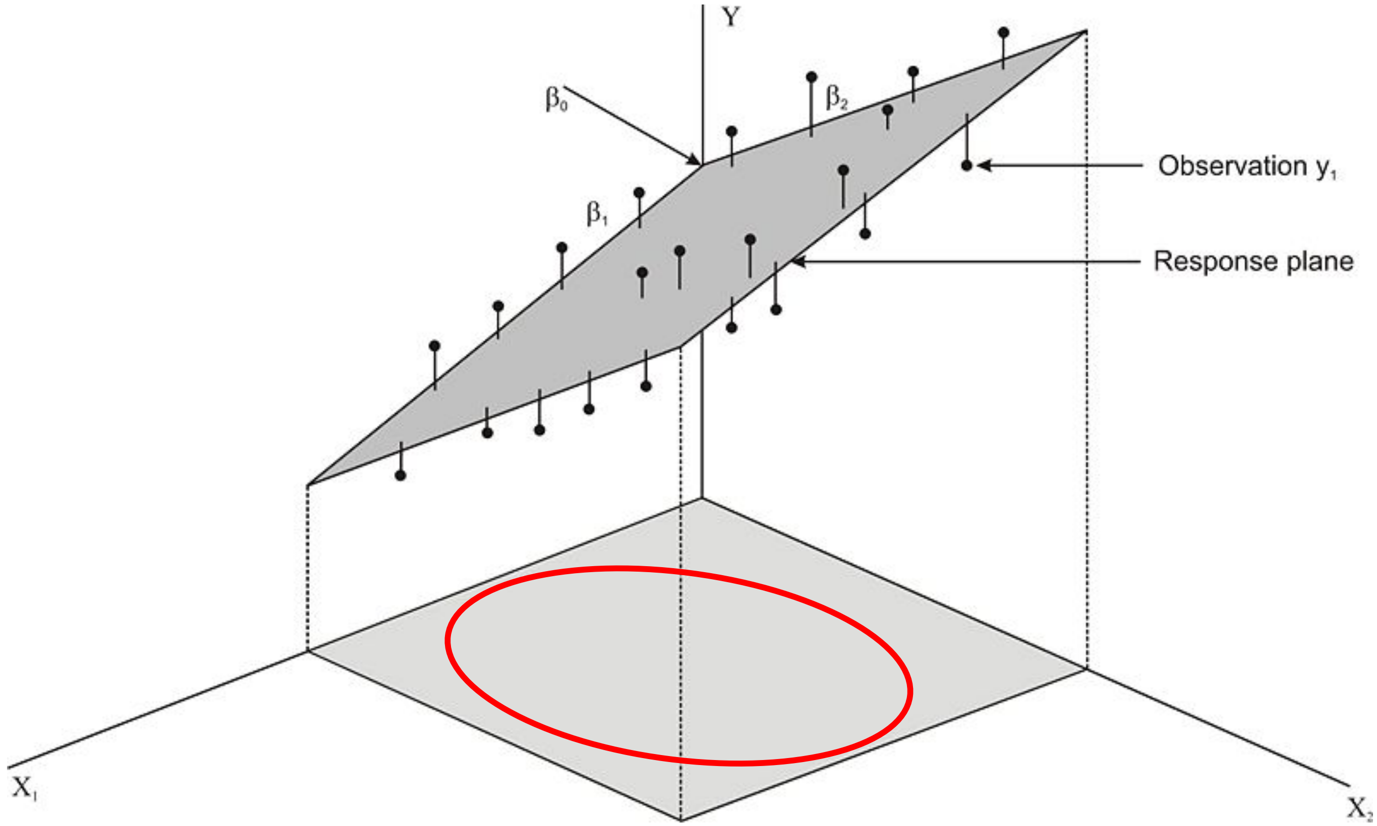
Example

y	x1	x2									
Salary	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	
475	81	7	24	38	39	14	3449	835	69	321	
480	130	18	66	72	76	3	1624	457	63	224	
500	141	20	65	78	37	11	5628	1575	225	828	
91.5	87	10	39	42	30	2	396	101	12	48	
750	169	4	74	51	35	11	4408	1133	19	501	
70	37	1	23	8	21	2	214	42	1	30	
100	73	0	24	24	7	3	509	108	0	41	
75	81	6	26	32	8	2	341	86	6	32	
1100	92	17	49	66	65	13	5206	1332	253	784	
517.143	159	21	107	75	59	10	4631	1300	90	702	
512.5	53	4	31	26	27	9	1876	467	15	192	
550	113	13	48	61	47	4	1512	392	41	205	
700	60	0	30	11	22	6	1941	510	4	309	
240	43	7	29	27	30	13	3231	825	36	376	
775	158	20	89	75	73	15	8068	2273	177	1045	
175	46	2	24	8	15	5	479	102	5	65	
135	32	8	16	22	14	8	727	180	24	67	
100	92	16	72	48	65	1	413	92	16	72	

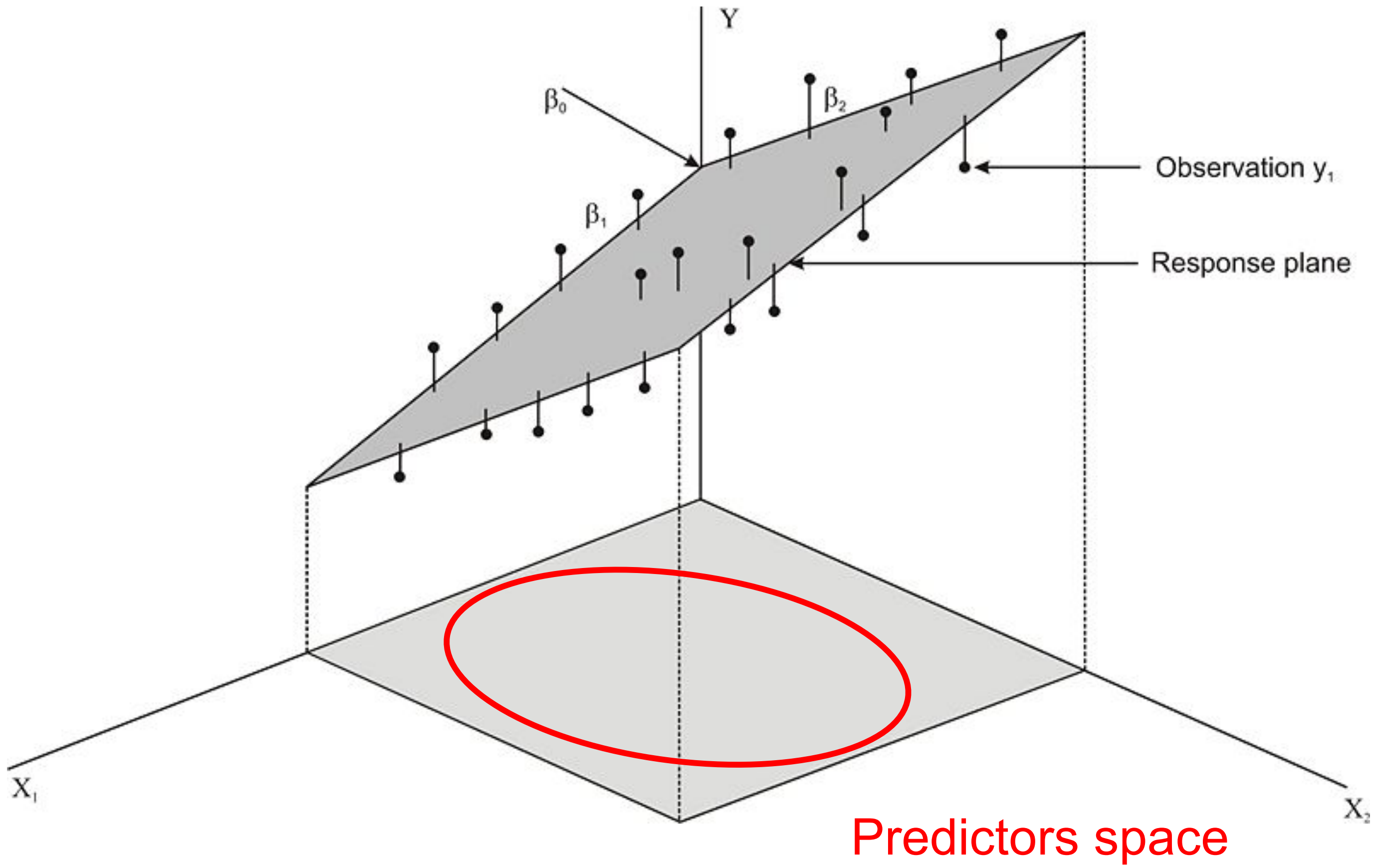
Linear Regression vs. Regression Tree



Linear Regression vs. Regression Tree



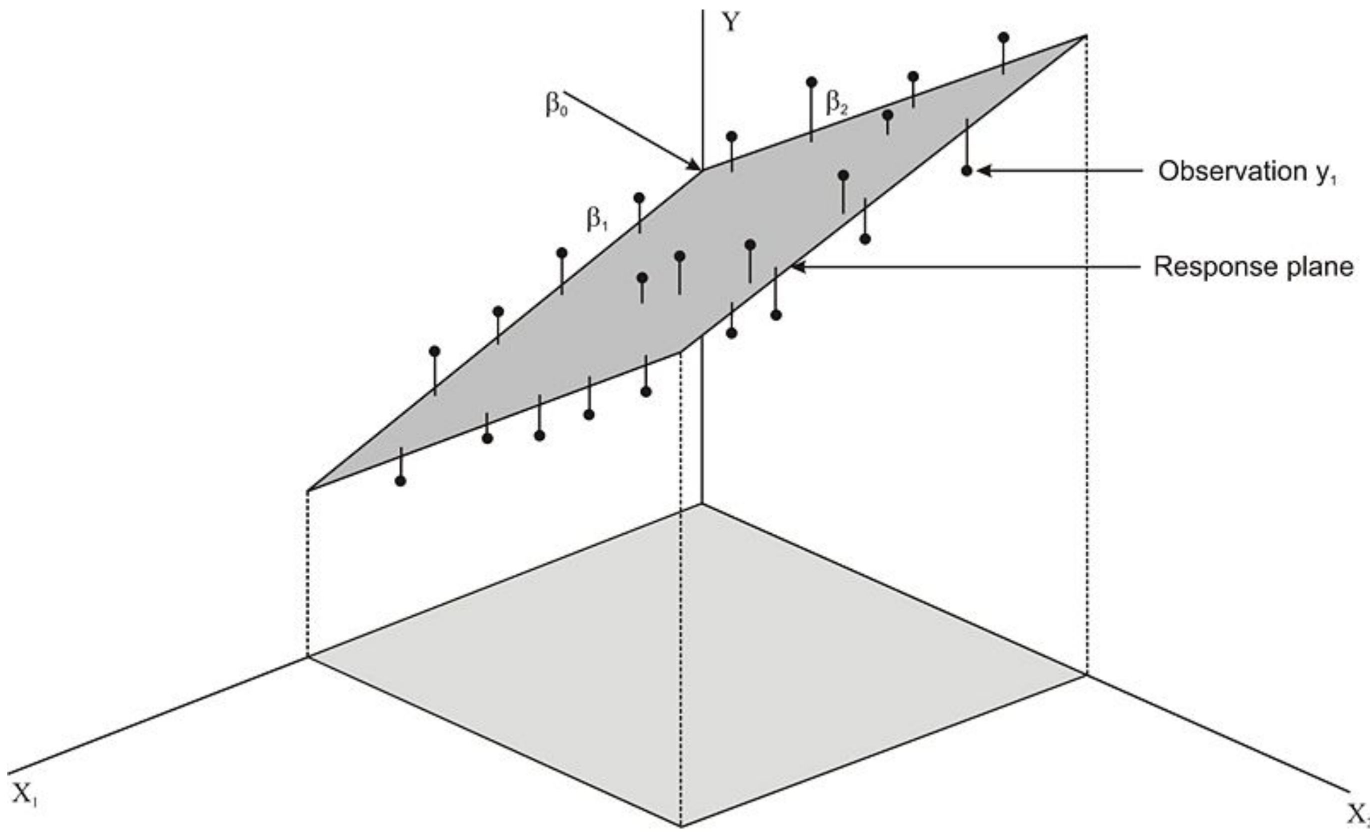
Linear Regression vs. Regression Tree



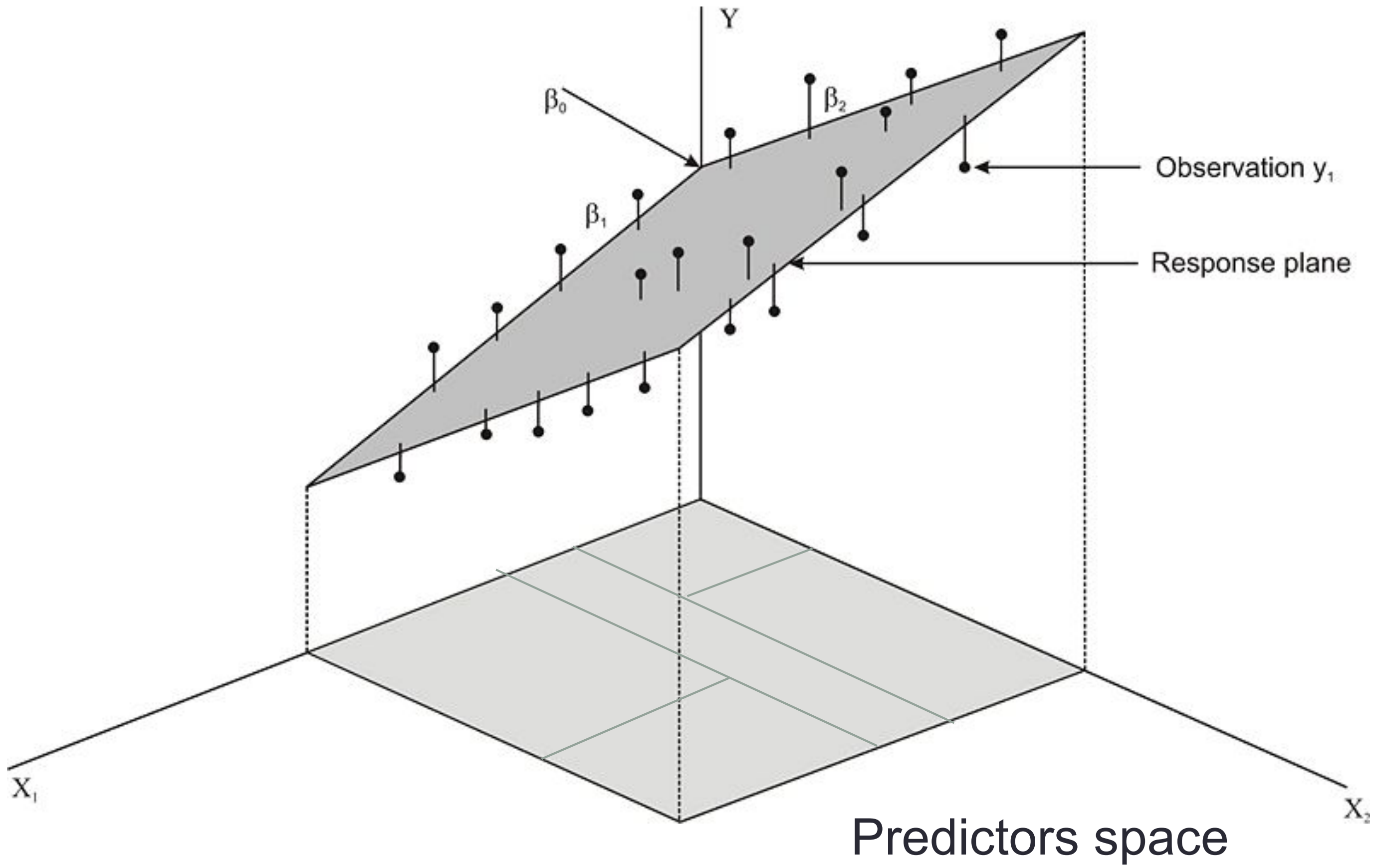
Partitioning Up the Predictor Space

- Split the predictors space into non-overlapping regions R_1, R_2,…, R_k

Linear Regression vs. Regression Tree



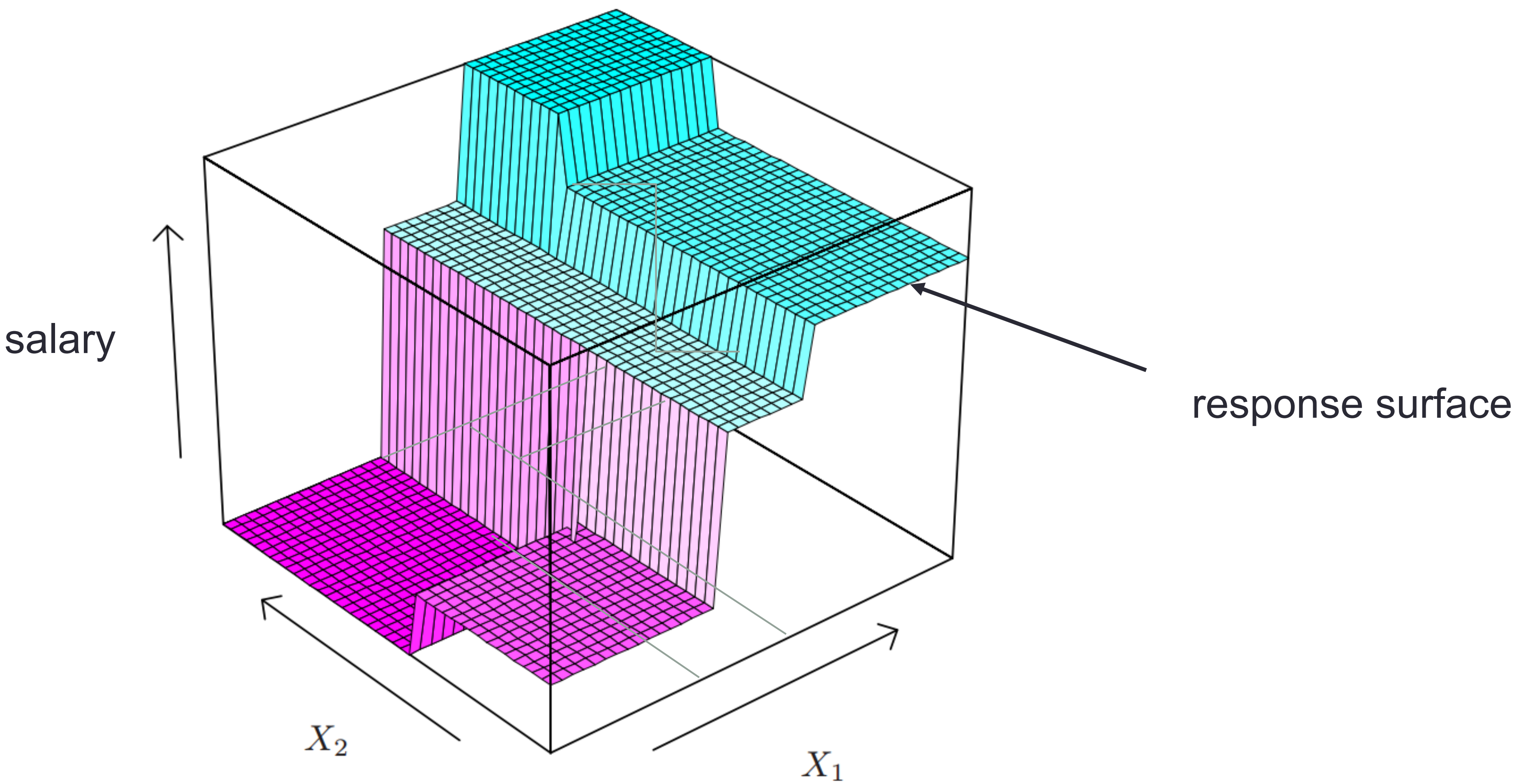
Linear Regression vs. Regression Tree



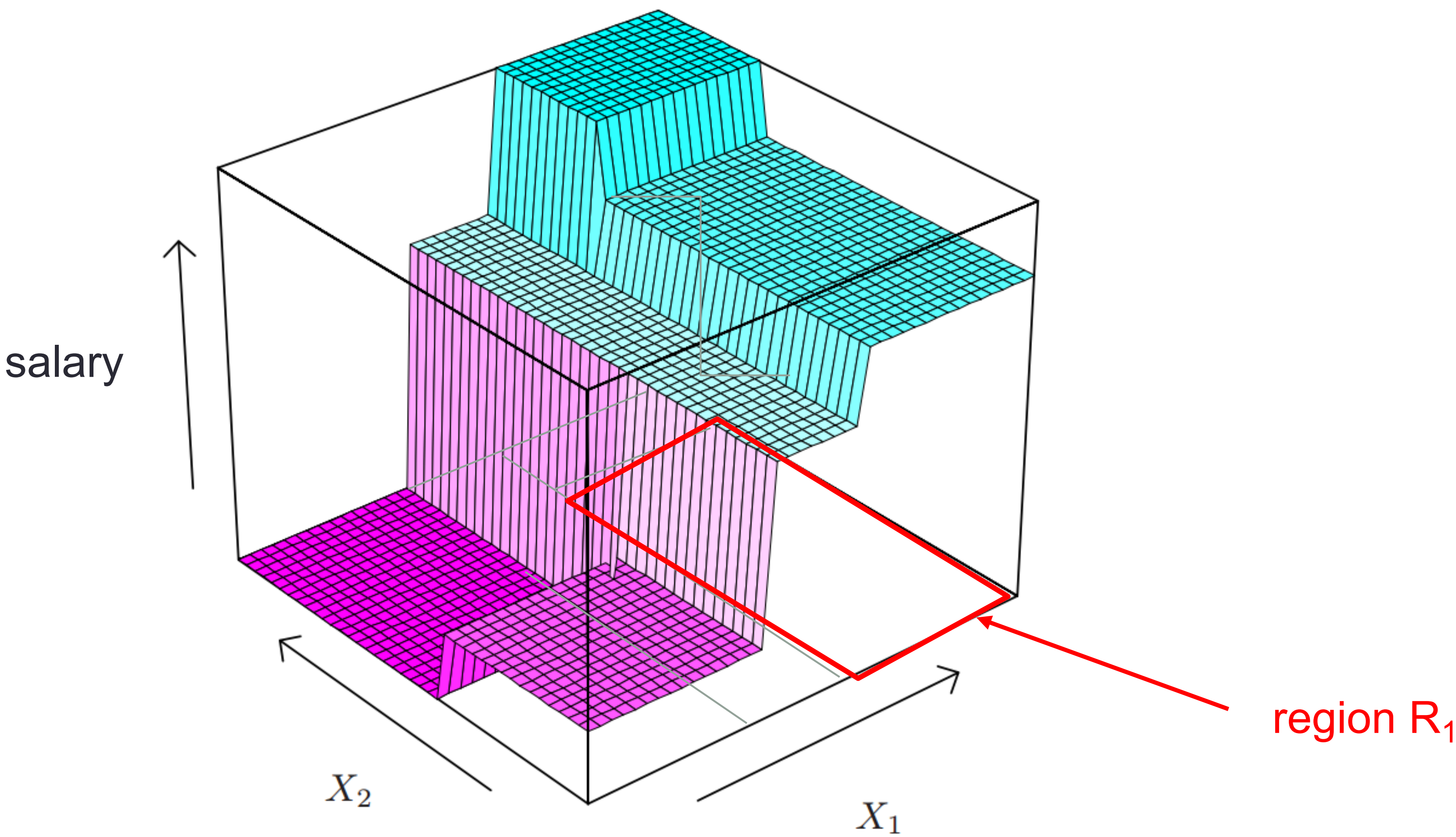
Partitioning Up the Predictor Space

- Split the predictors space into non-overlapping regions R_1, R_2,…, R_k
- For each region, the prediction is the mean response of all observations in that region

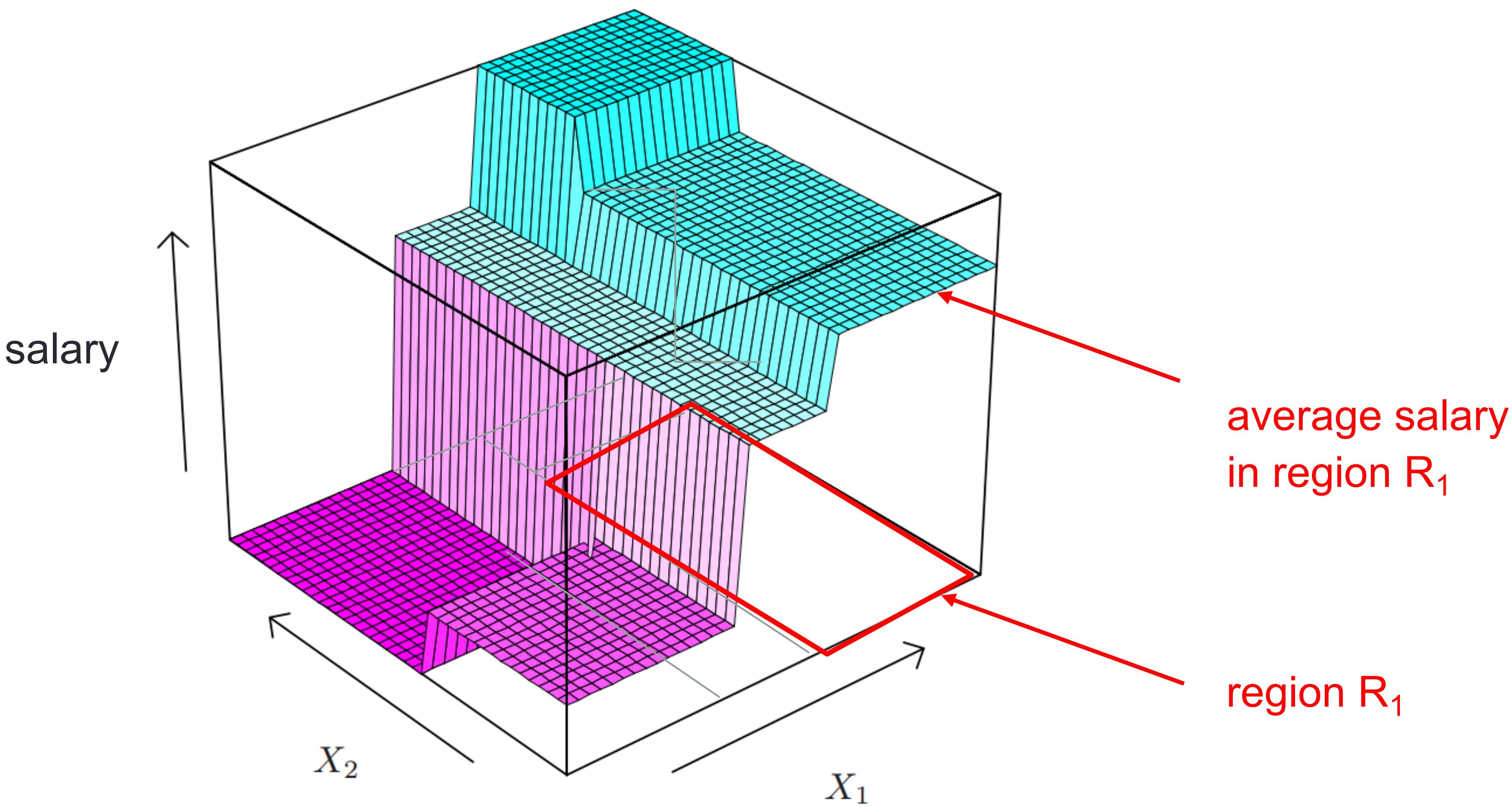
Linear Regression vs. Regression Tree



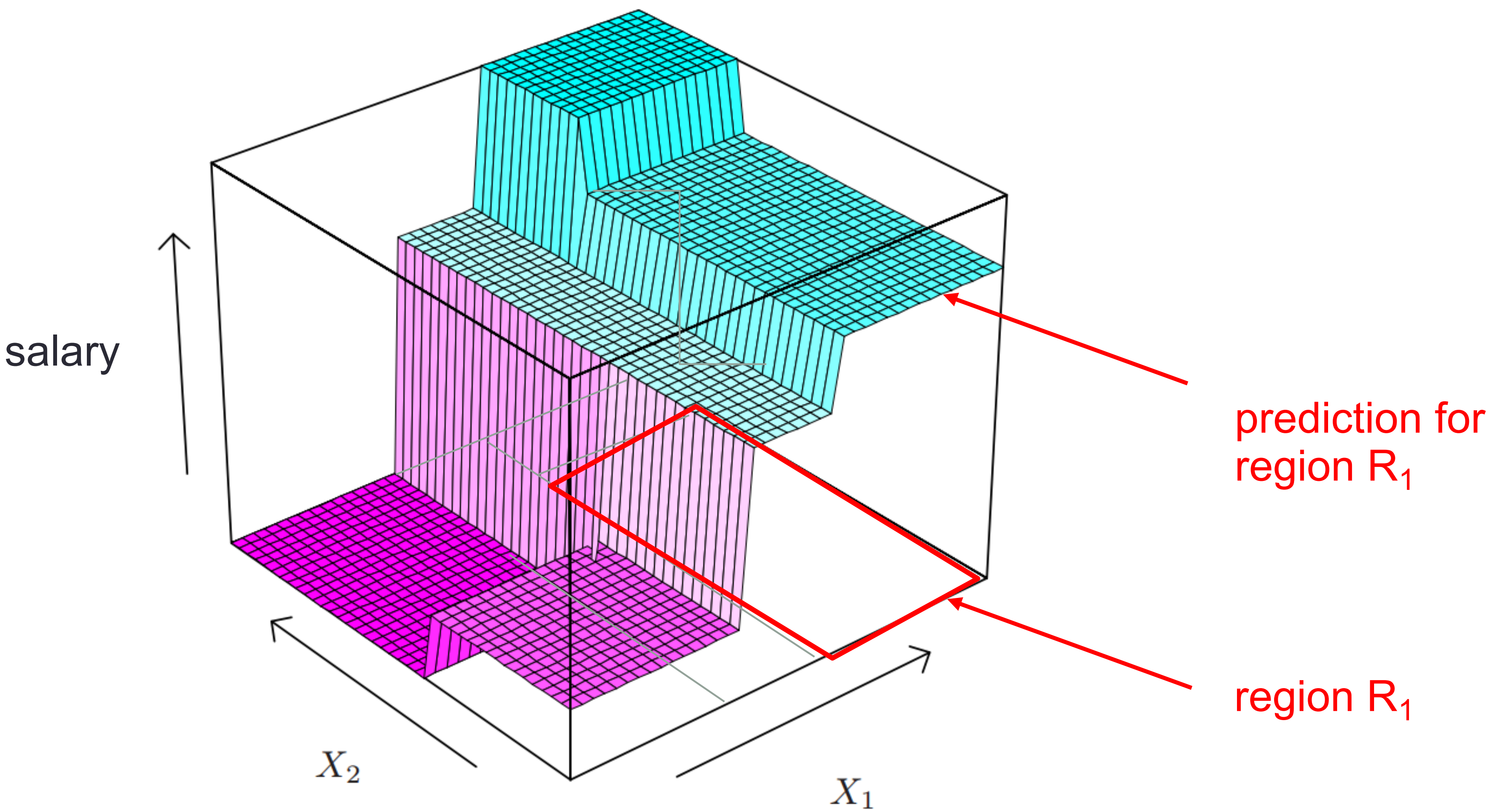
Linear Regression vs. Regression Tree



Linear Regression vs. Regression Tree



Linear Regression vs. Regression Tree



Regression Trees

If the observations in Region R_1

have mean response 100,

we would predict 100,

for any new observation in R_1

Regression Trees

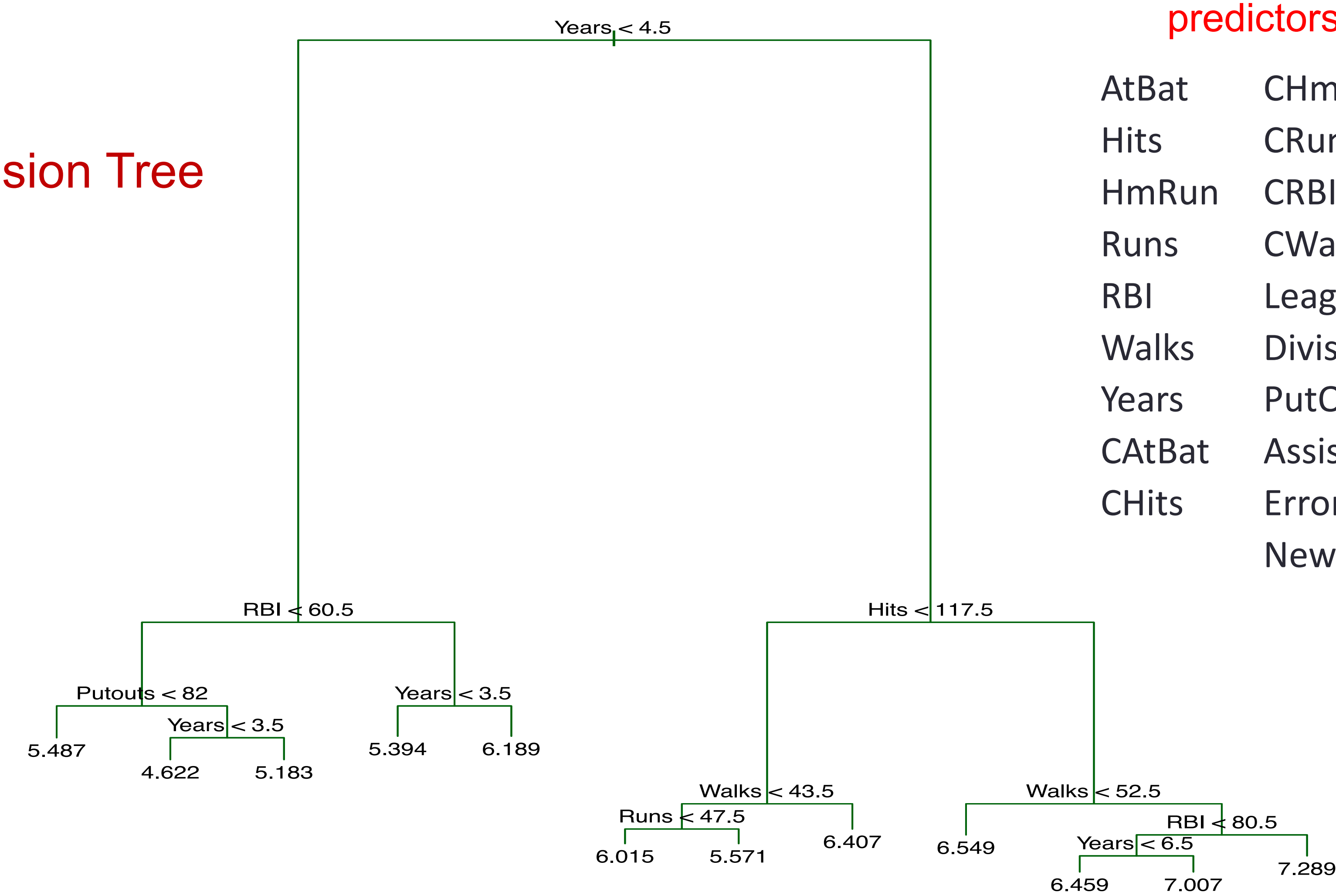
By splitting the predictors region
we obtain a Decision Tree

Example: Baseball Players' Salaries

Salary	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns
475	81	7	24	38	39	14	3449	835	69	321
480	130	18	66	72	76	3	1624	457	63	224
500	141	20	65	78	37	11	5628	1575	225	828
91.5	87	10	39	42	30	2	396	101	12	48
750	169	4	74	51	35	11	4408	1133	19	501
70	37	1	23	8	21	2	214	42	1	30
100	73	0	24	24	7	3	509	108	0	41
75	81	6	26	32	8	2	341	86	6	32
1100	92	17	49	66	65	13	5206	1332	253	784
517.143	159	21	107	75	59	10	4631	1300	90	702
512.5	53	4	31	26	27	9	1876	467	15	192
550	113	13	48	61	47	4	1512	392	41	205
700	60	0	30	11	22	6	1941	510	4	309
240	43	7	29	27	30	13	3231	825	36	376
775	158	20	89	75	73	15	8068	2273	177	1045
175	46	2	24	8	15	5	479	102	5	65
135	32	8	16	22	14	8	727	180	24	67
100	92	16	72	48	65	1	413	92	16	72

Example: Baseball Players' Salaries

Decision Tree

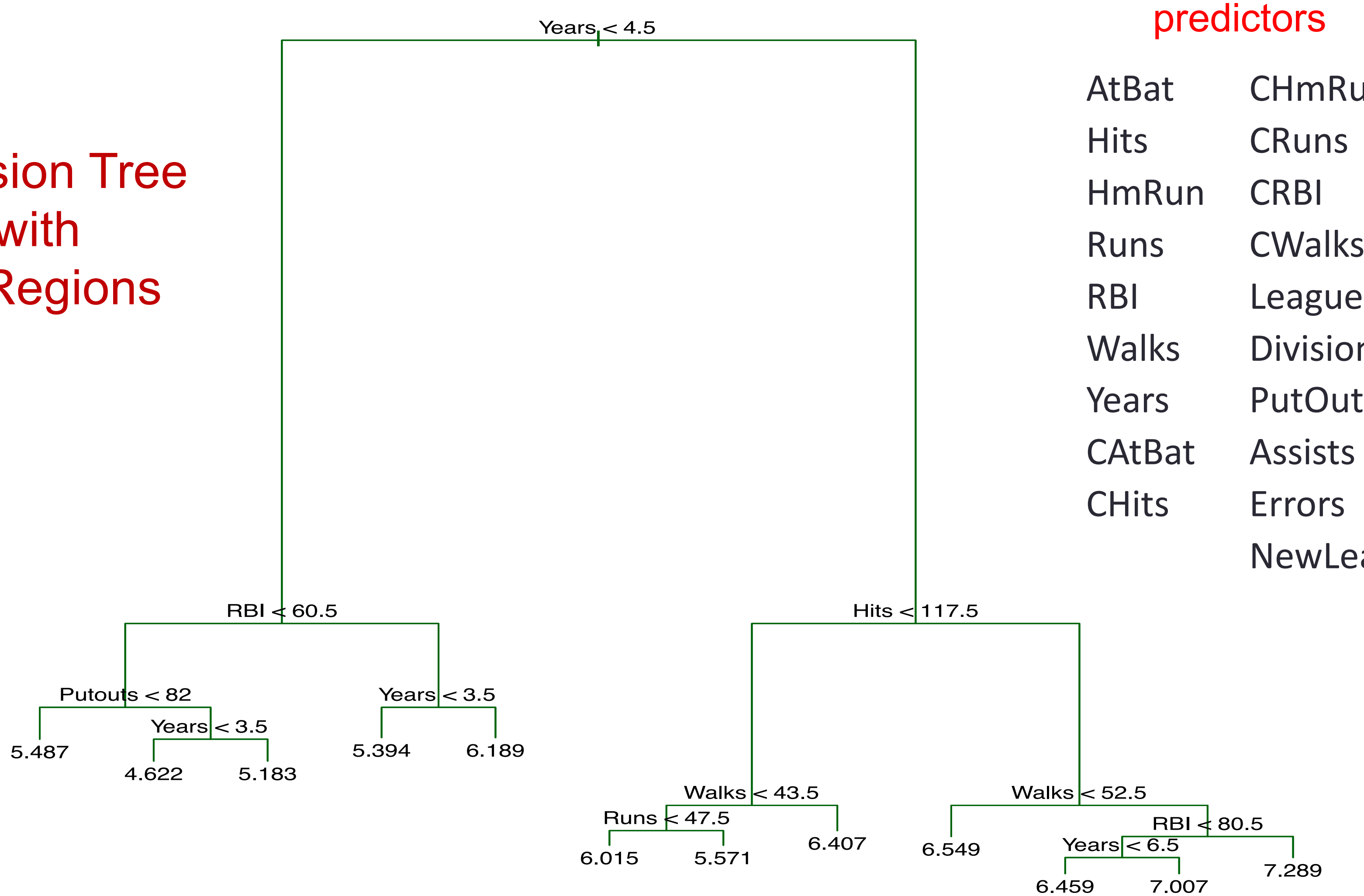


predictors

AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

Example: Baseball Players' Salaries

Decision Tree
with
12 Regions



predictors

AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

Example: Baseball Players' Salaries

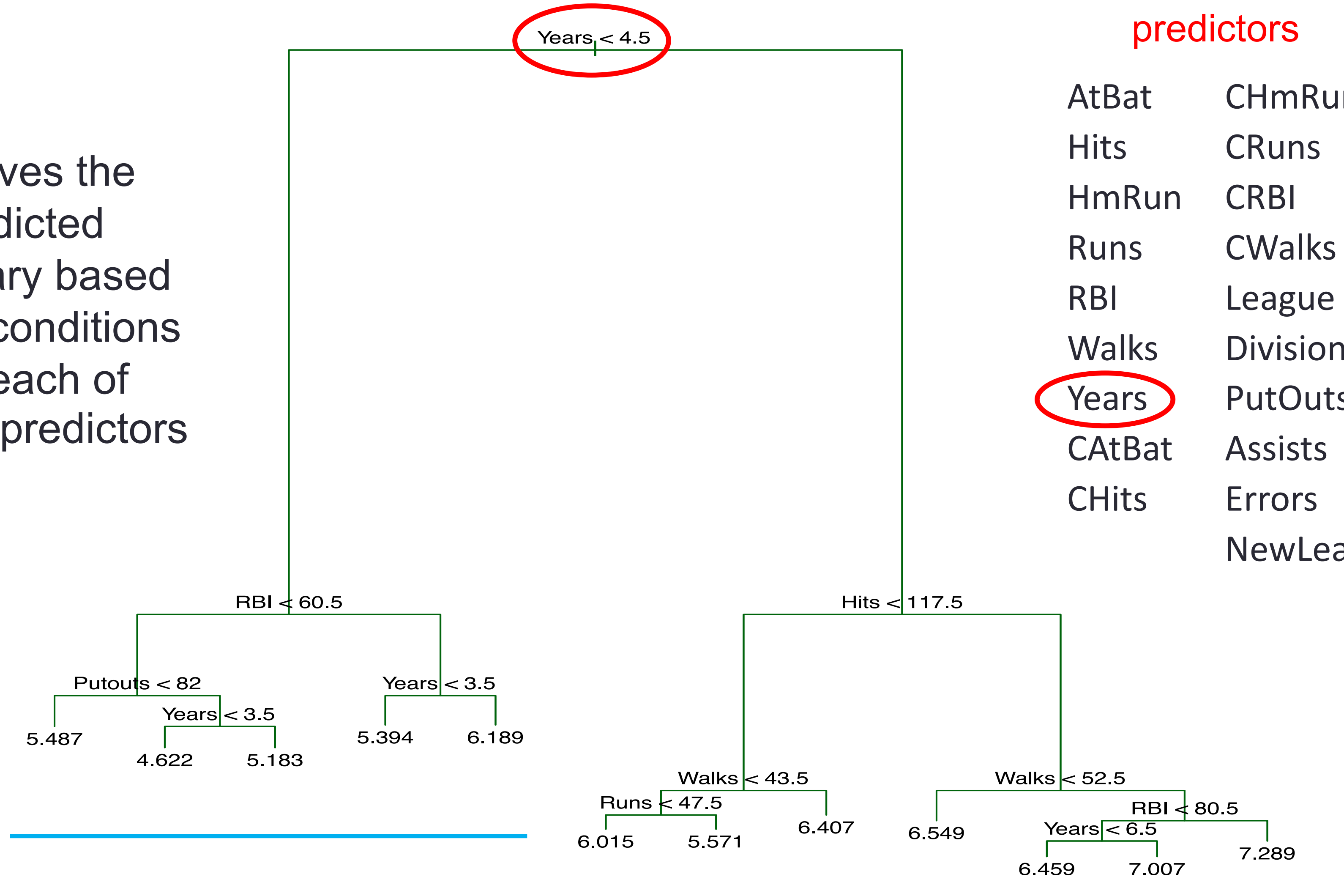
It gives the predicted salary based on conditions on each of the predictors

predictors

AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

Example: Baseball Players' Salaries

It gives the predicted salary based on conditions on each of the predictors

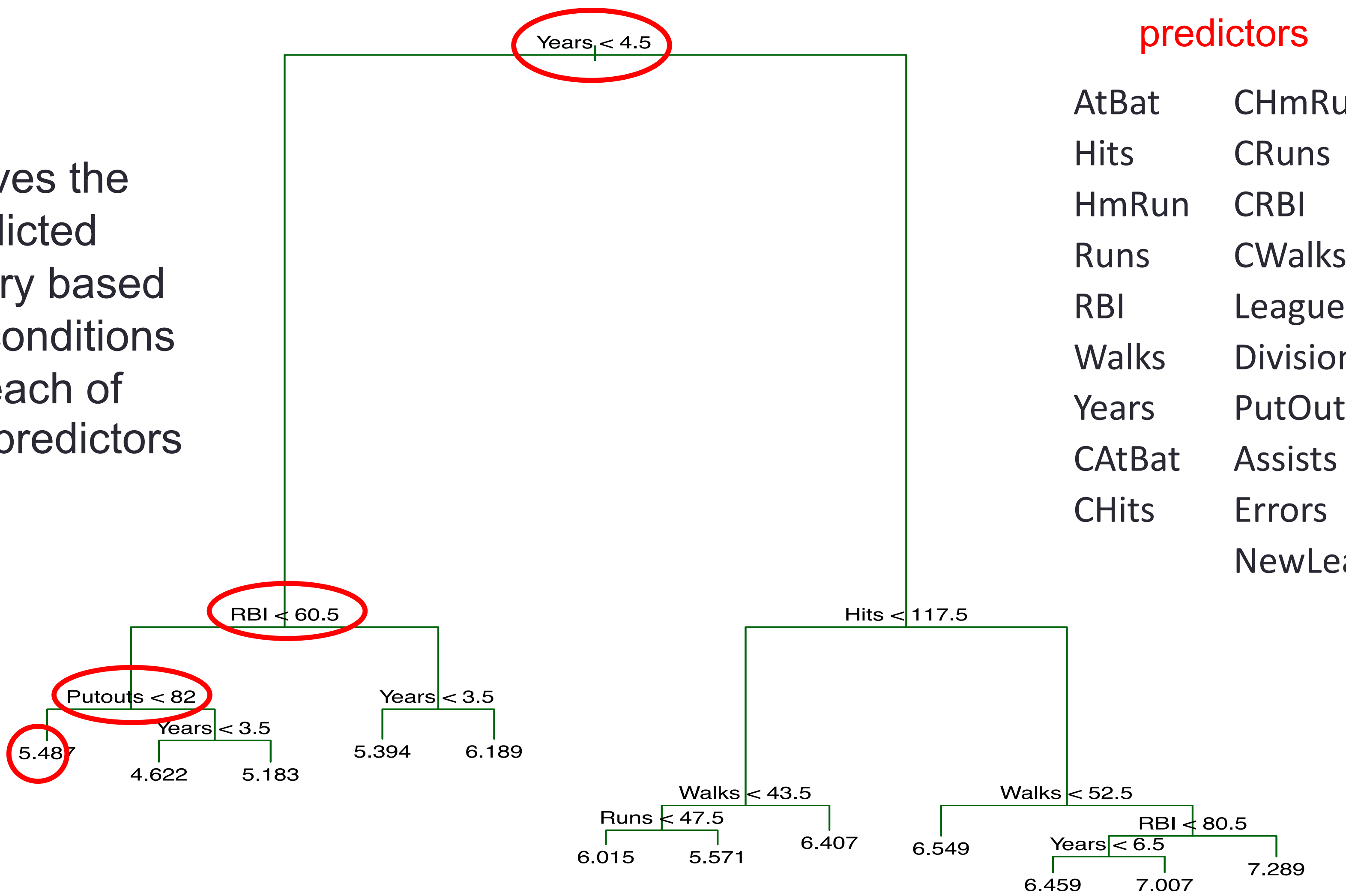


predictors

AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

Example: Baseball Players' Salaries

It gives the predicted salary based on conditions on each of the predictors

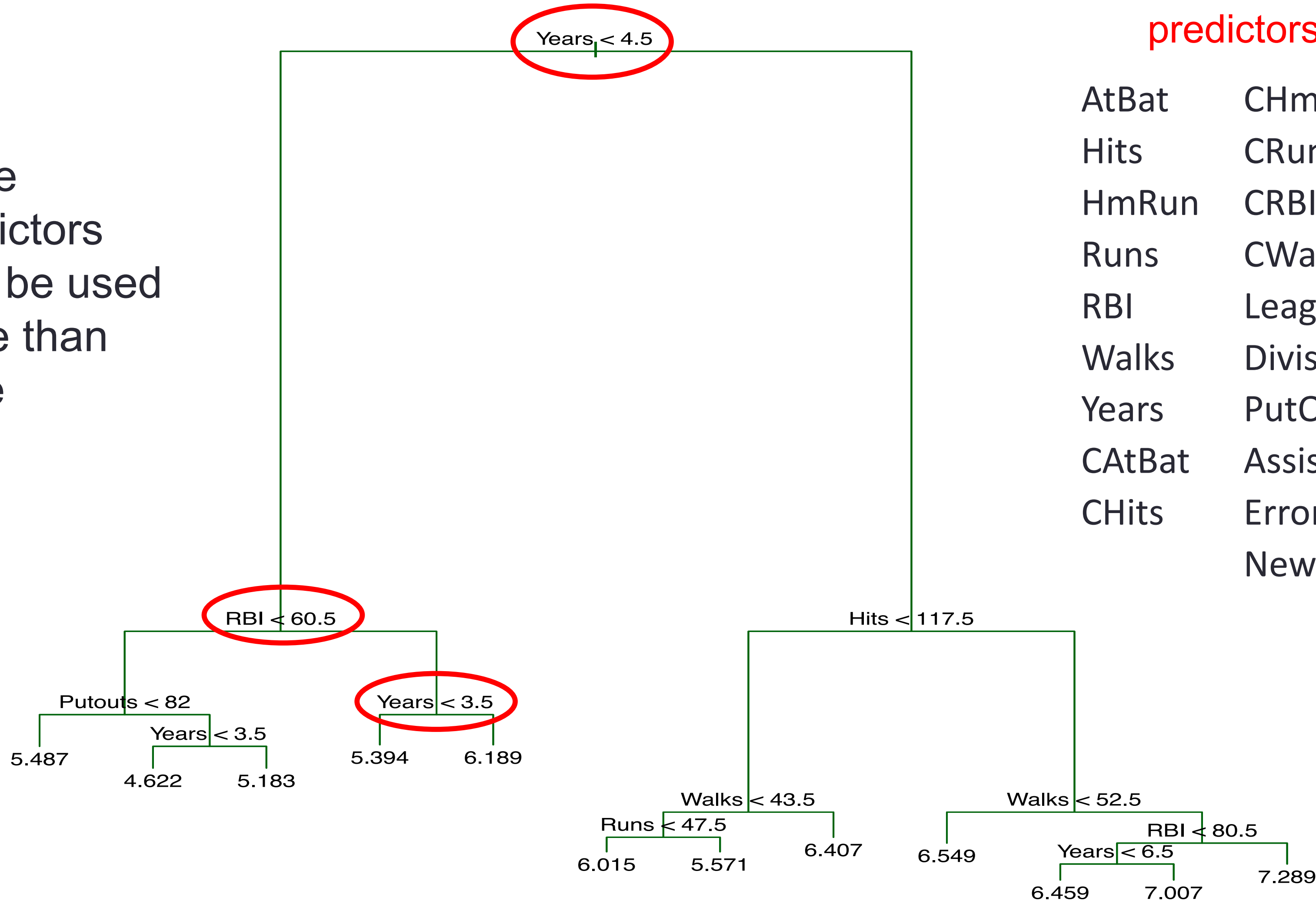


predictors

AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

Example: Baseball Players' Salaries

some predictors may be used more than once

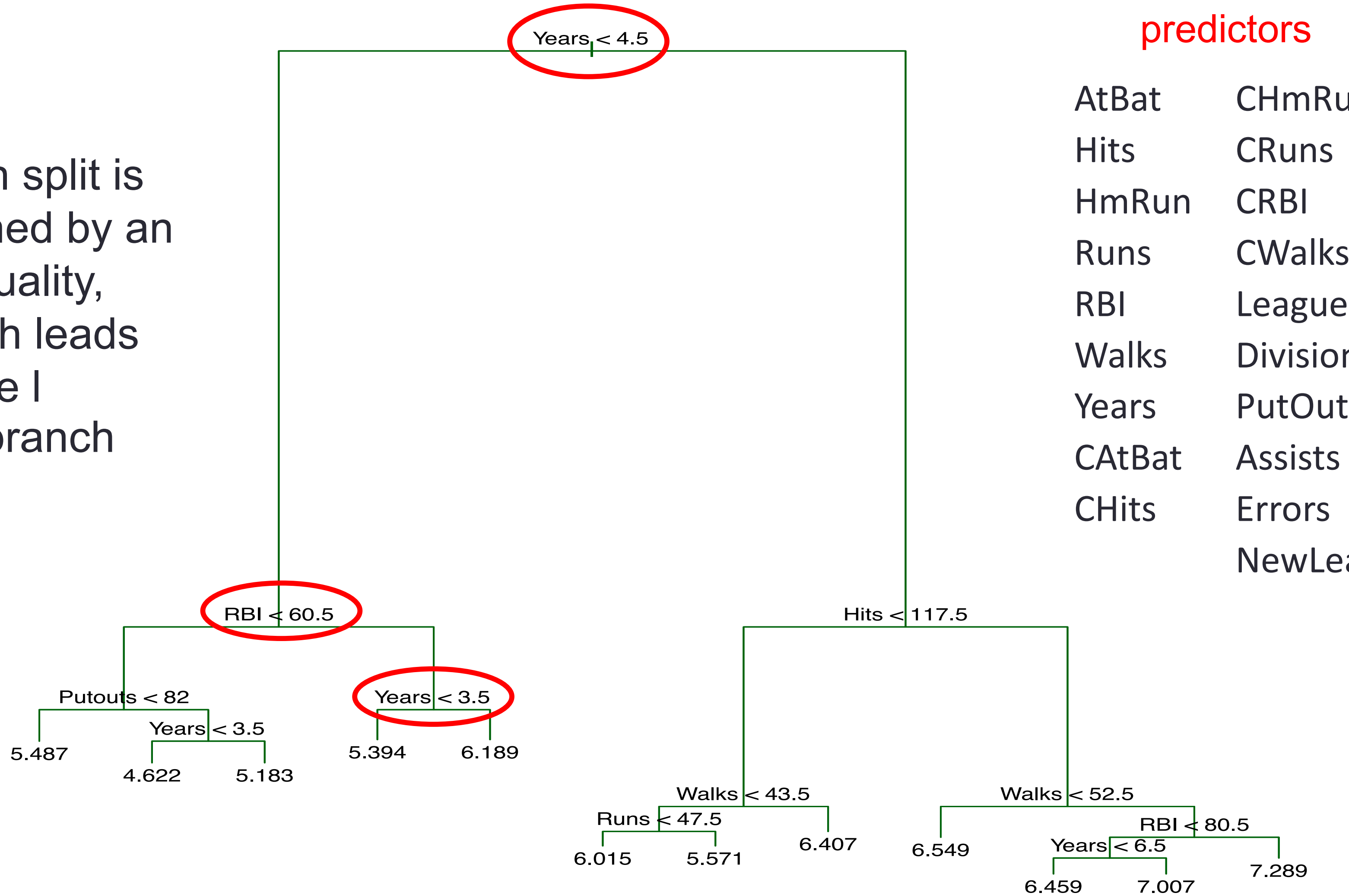


predictors

AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

Example: Baseball Players' Salaries

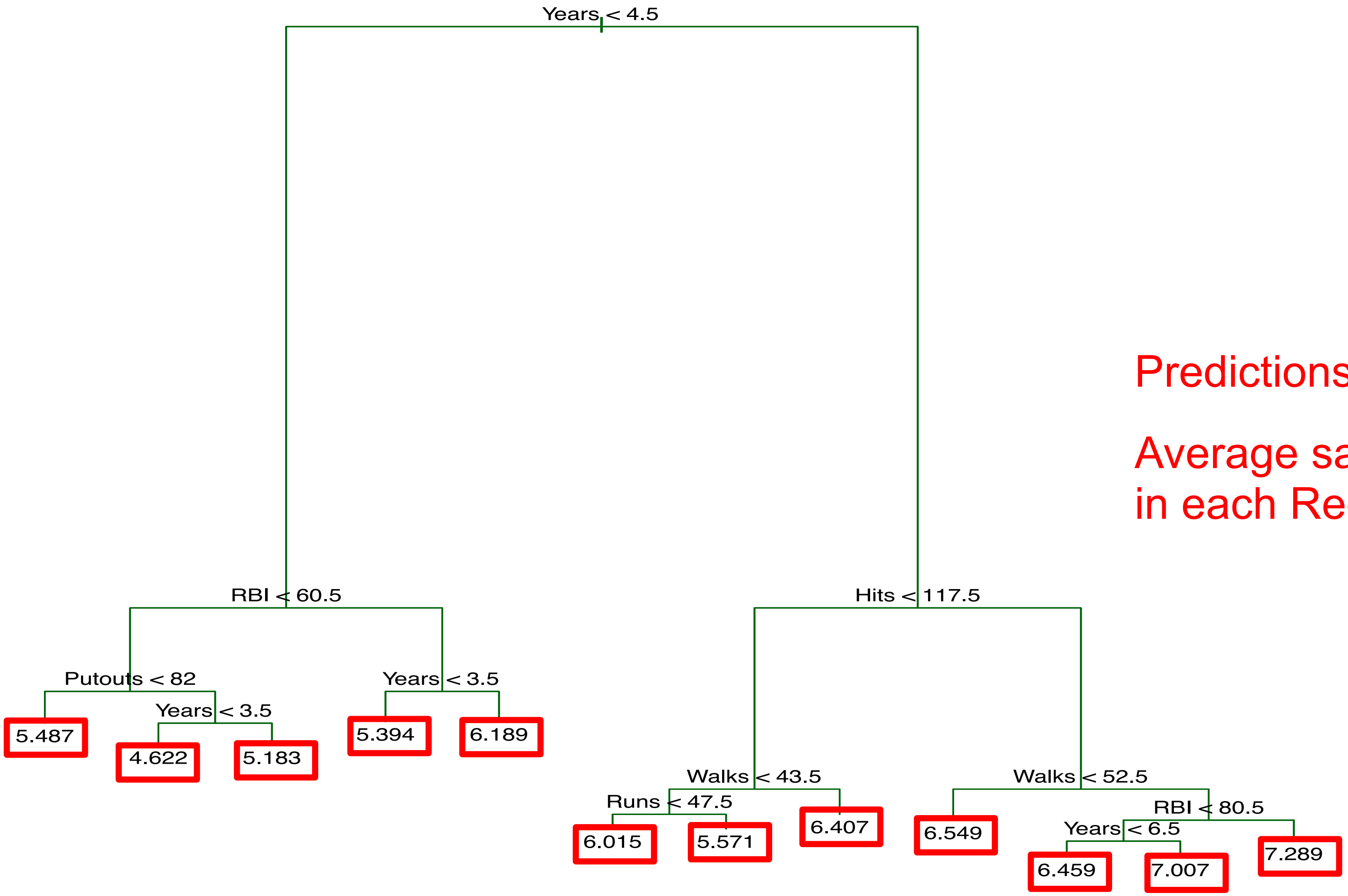
Each split is defined by an inequality, which leads to the l left branch



predictors

AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

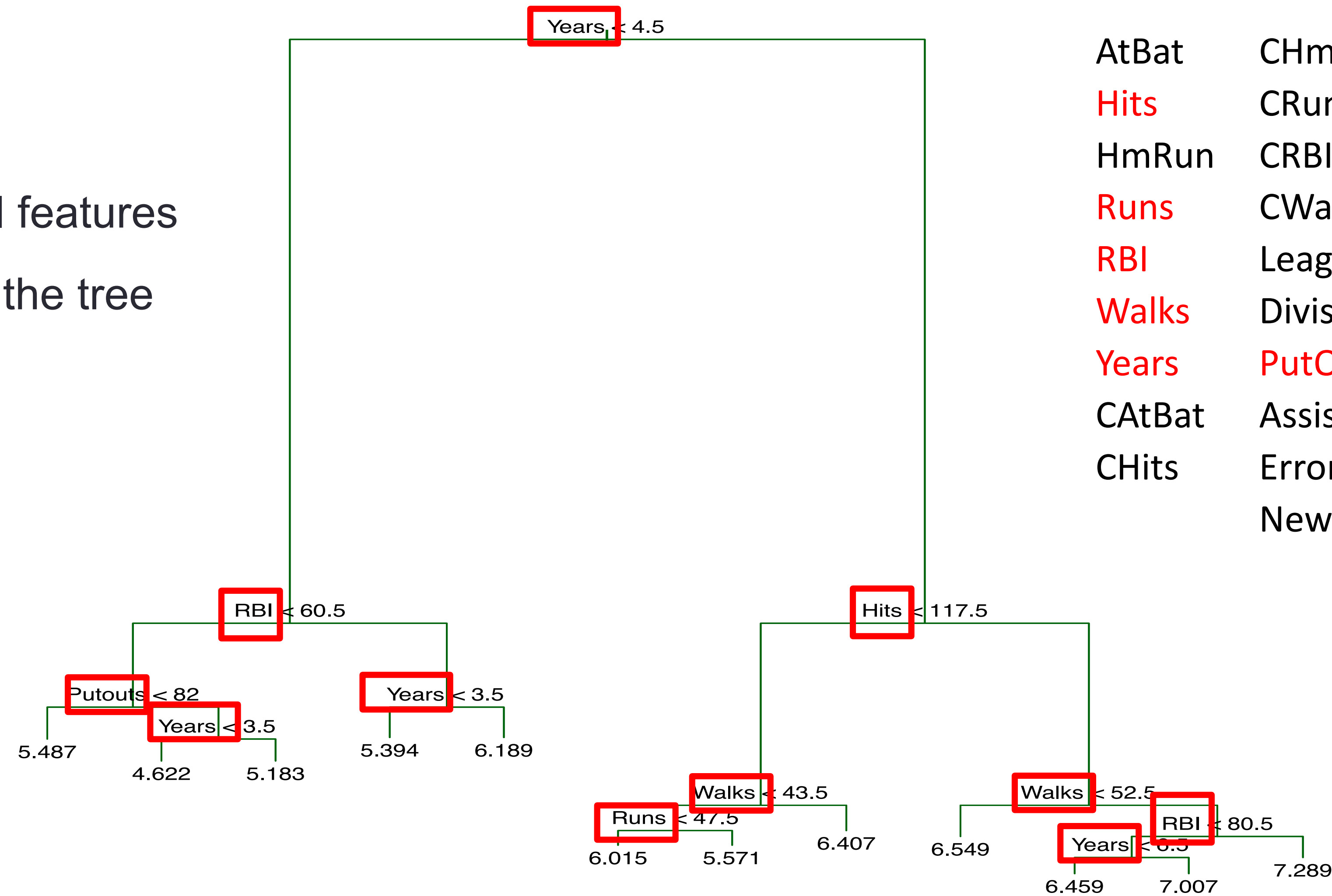
Example: Baseball Players' Salaries



Predictions are

Average salary in each Region

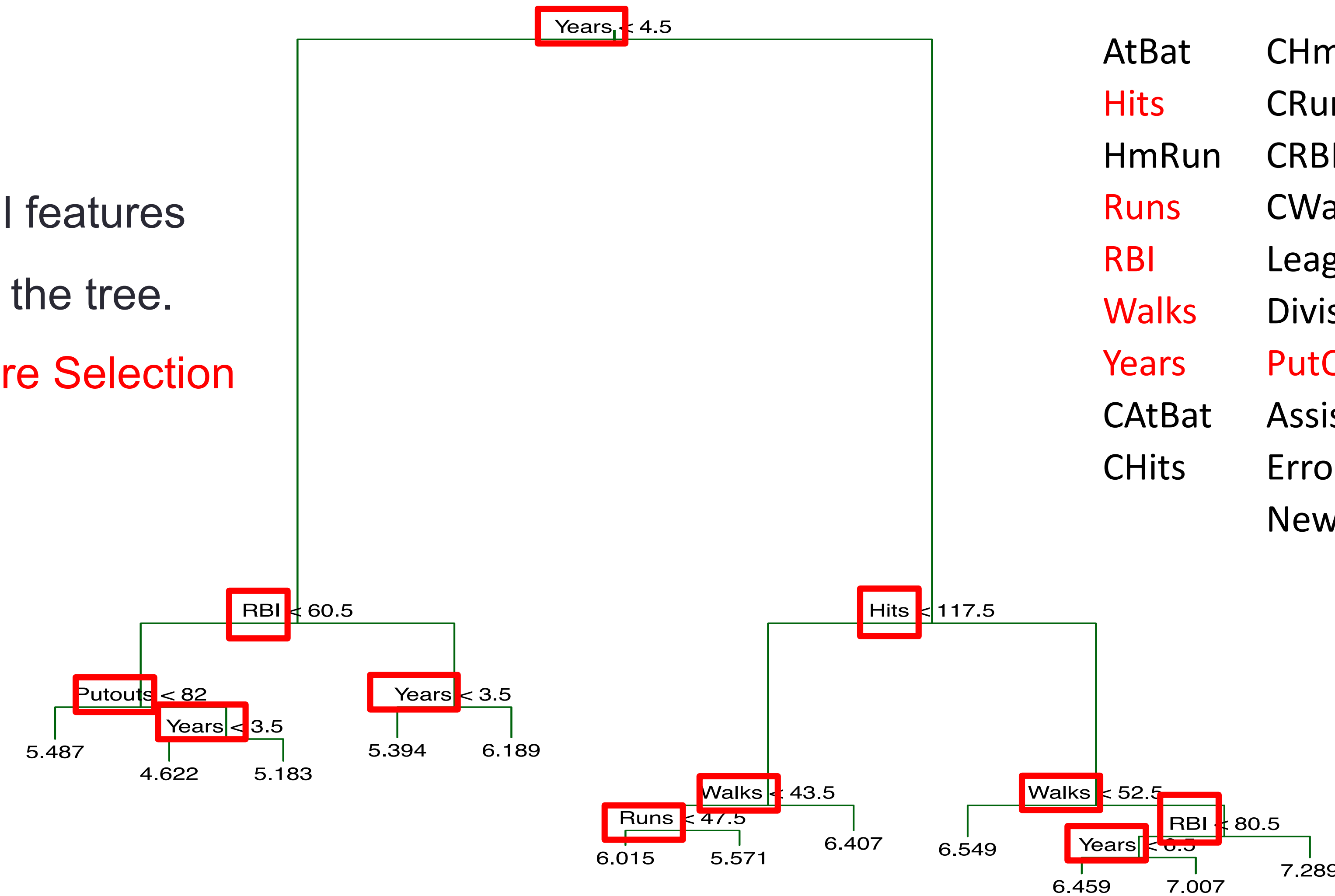
Example: Baseball Players' Salaries



Not all features are in the tree

AtBat
Hits
HmRun
Runs
RBI
Walks
Years
CAtBat
CHits
CHmRun
CRuns
CRBI
CWalks
League
Division
PutOuts
Assists
Errors
NewLeague

Example: Baseball Players' Salaries



AtBat	CHmRun
Hits	CRuns
HmRun	CRBI
Runs	CWalks
RBI	League
Walks	Division
Years	PutOuts
CAtBat	Assists
CHits	Errors
	NewLeague

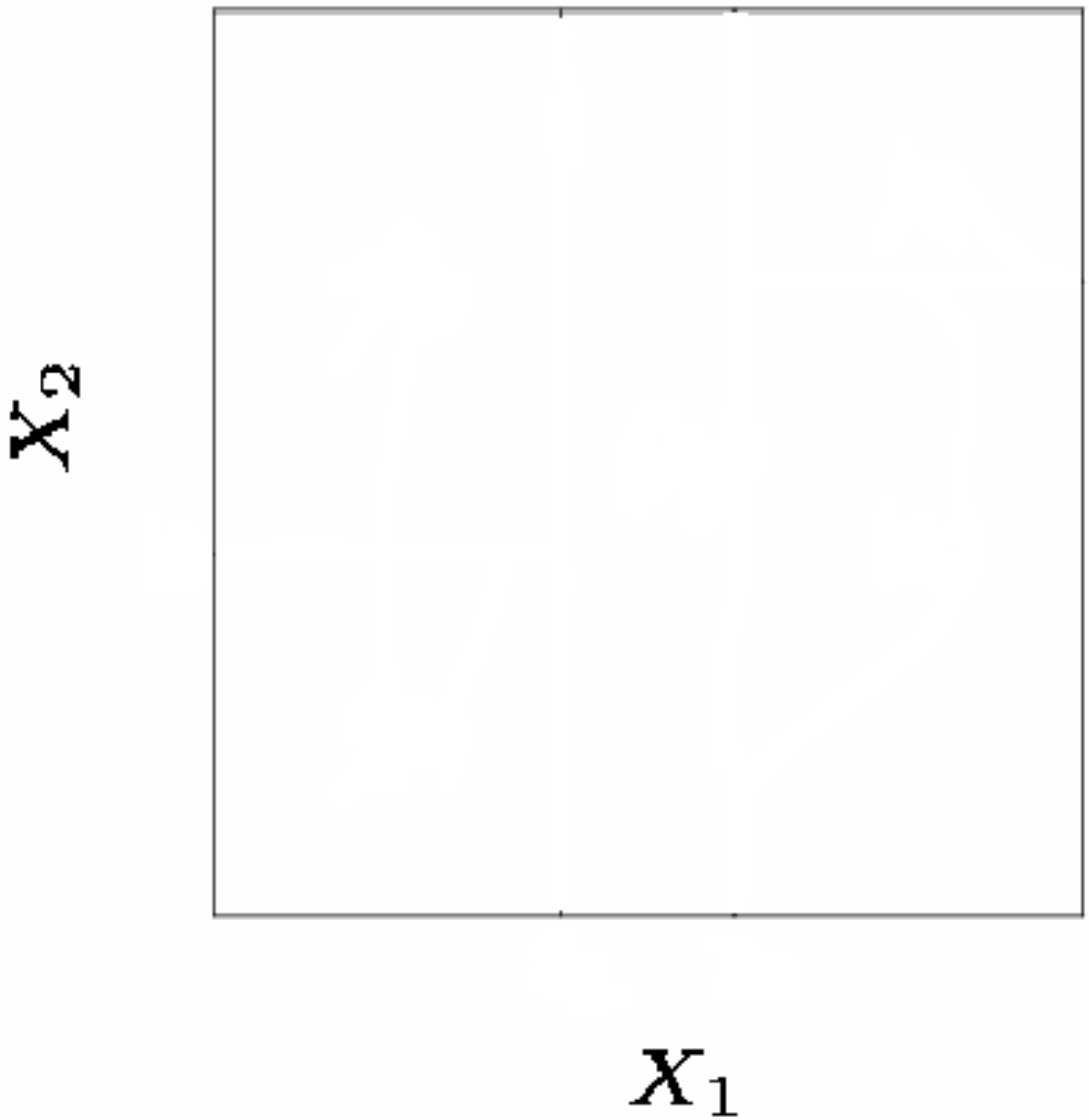
Not all features

are in the tree.

Feature Selection

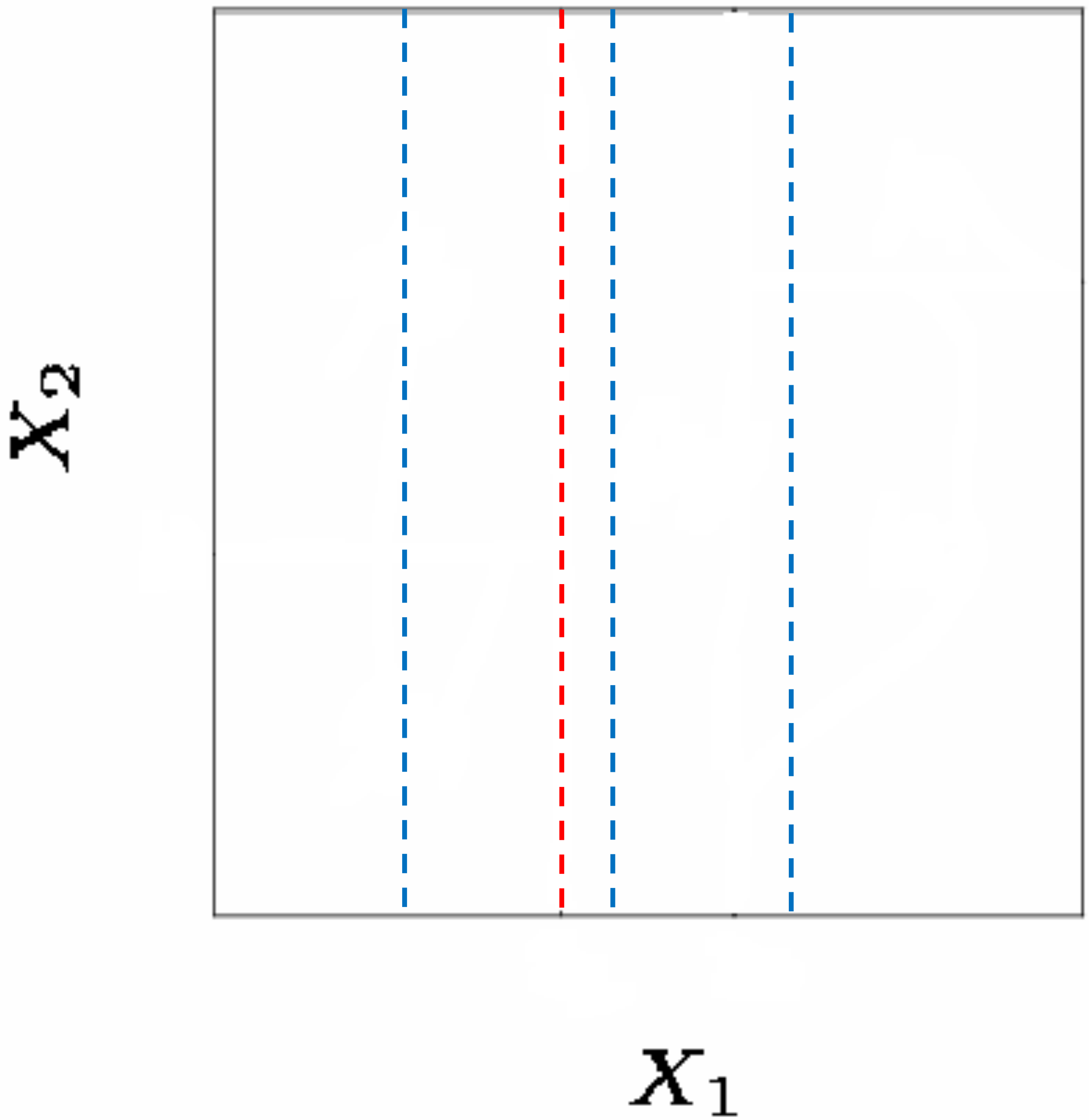
Partitioning Up the Predictor Space

- Regions are created by iteratively splitting one of the X-variable axes into two segments

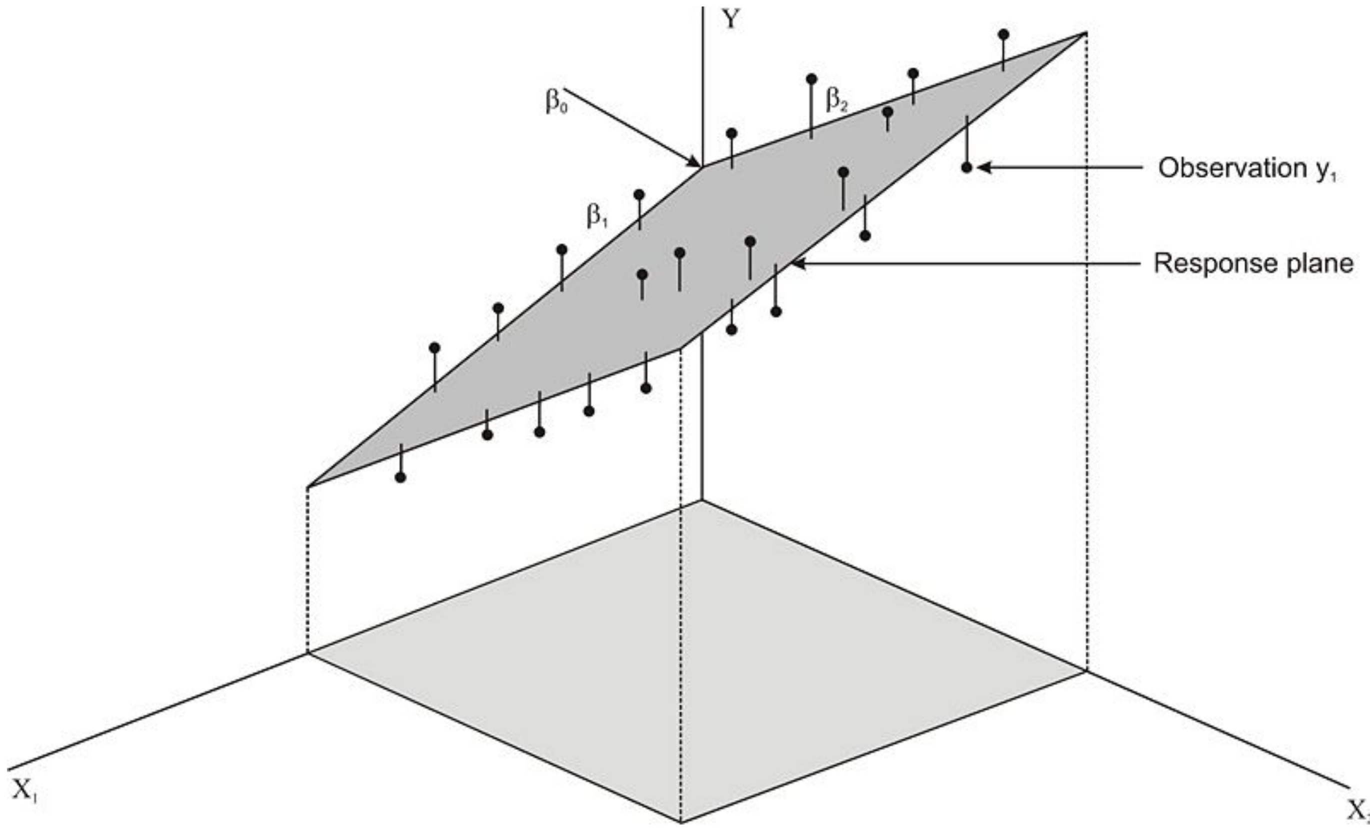


Partitioning Up the Predictor Space

- For each variable select the boundary that results in the largest MSE reduction

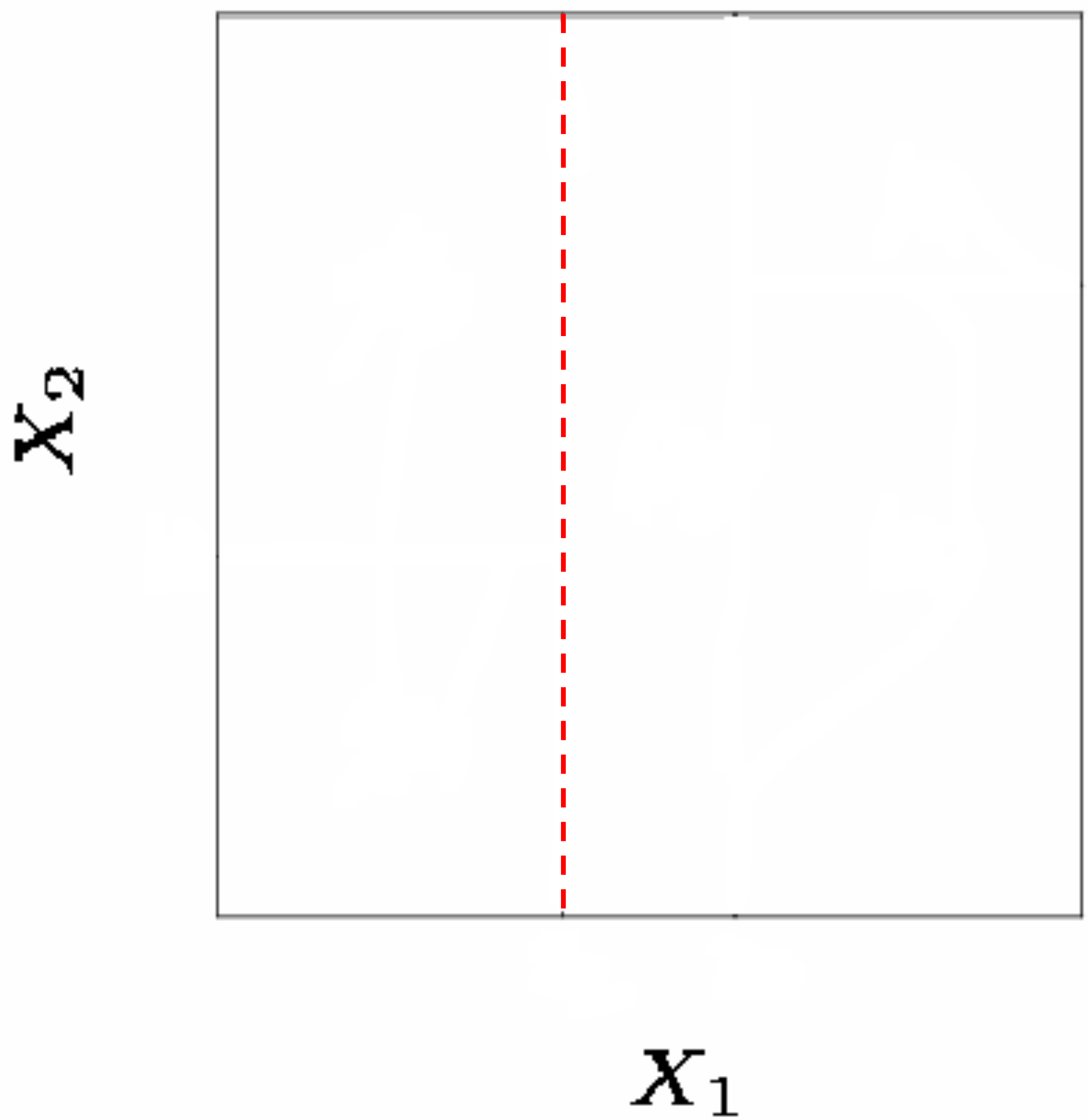


Linear Regression vs. Regression Tree



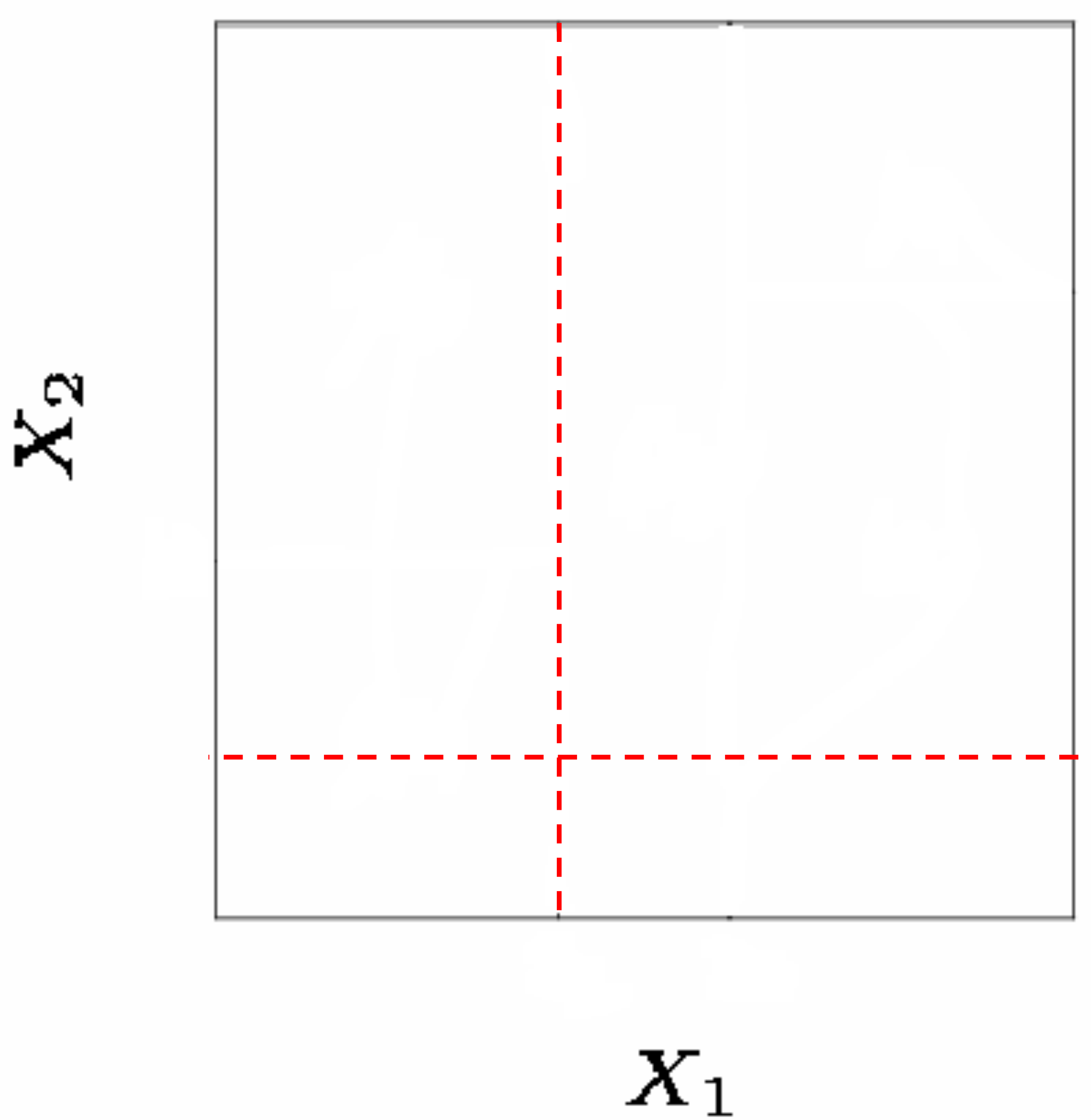
Partitioning Up the Predictor Space

- For each variable select the boundary that results in the largest MSE reduction



Partitioning Up the Predictor Space

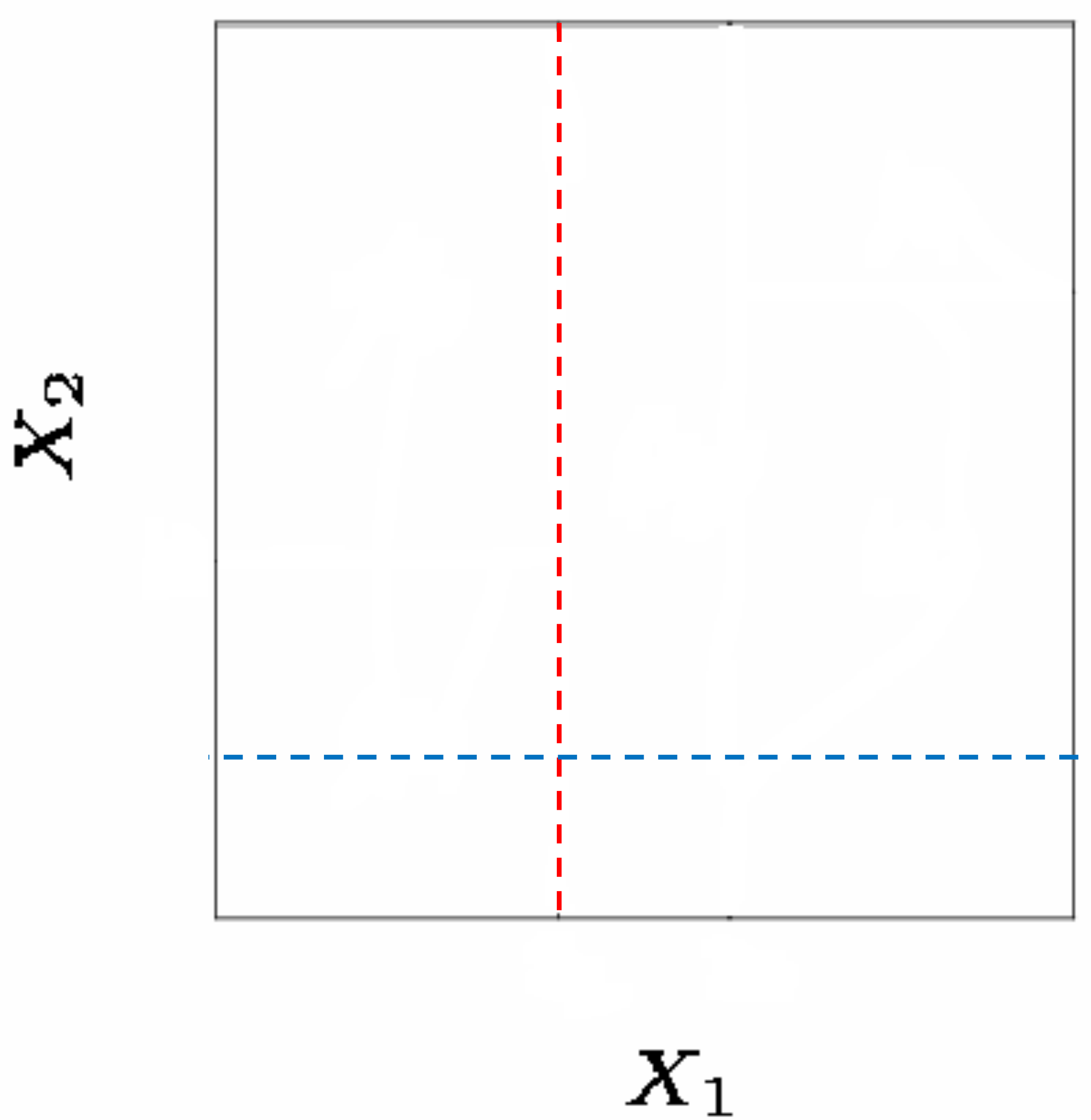
- For each variable select the boundary that results in the largest MSE reduction



Partitioning Up the Predictor Space

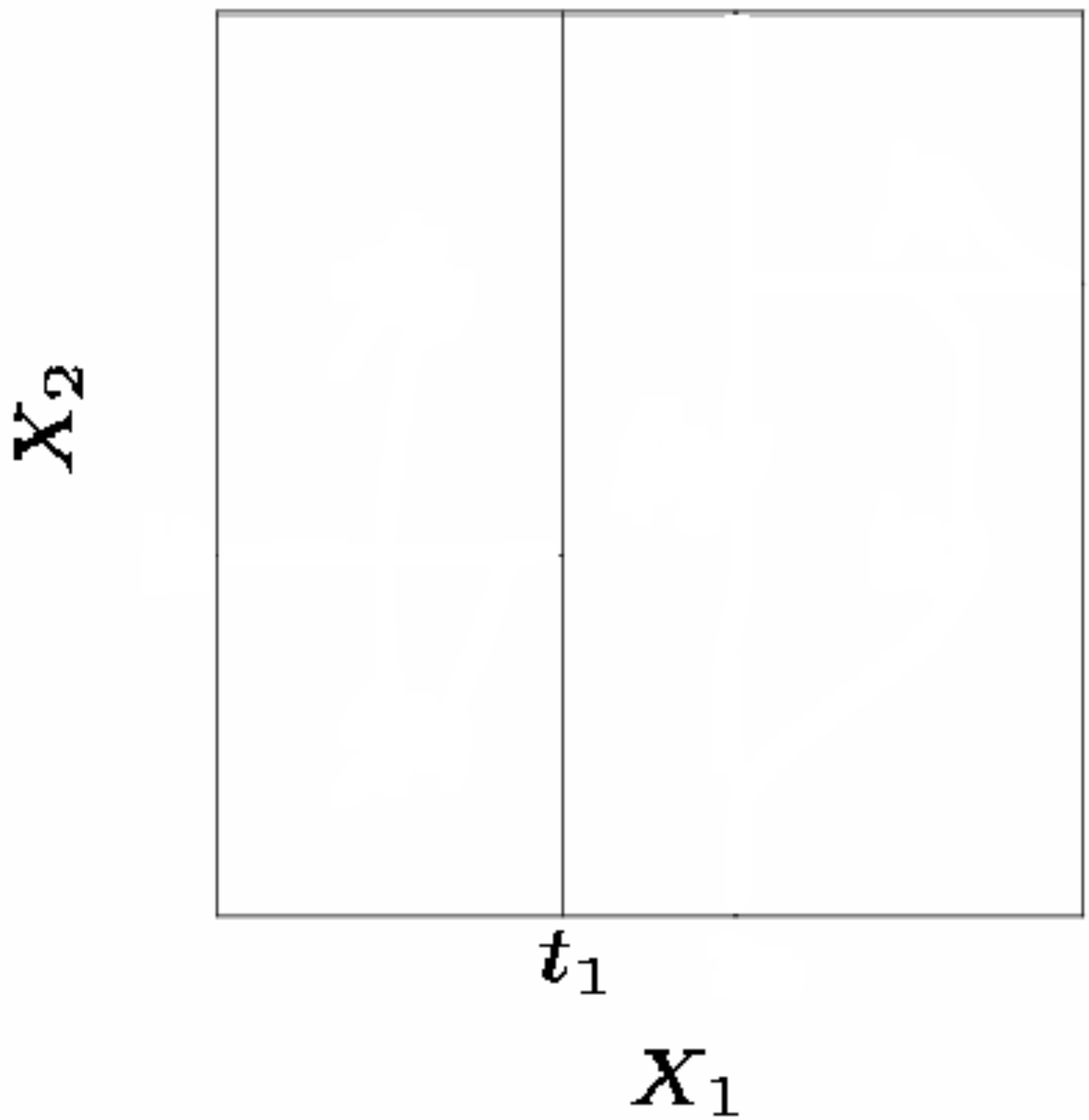
- For each variable select the boundary that results in the largest MSE reduction

- Choose that variable with the largest reduction



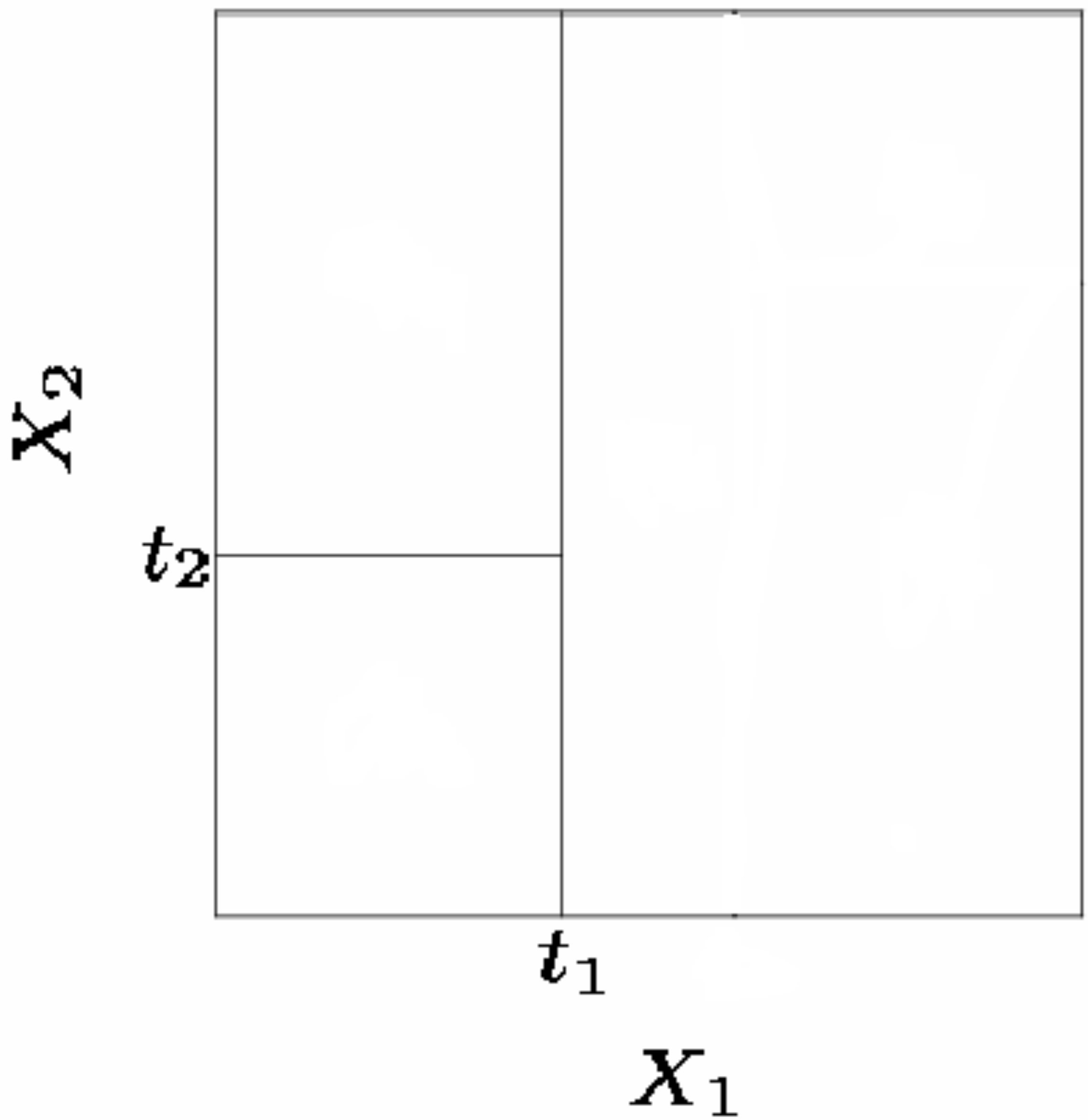
Partitioning Up the Predictor Space

1. First split on $X_1 = t_1$



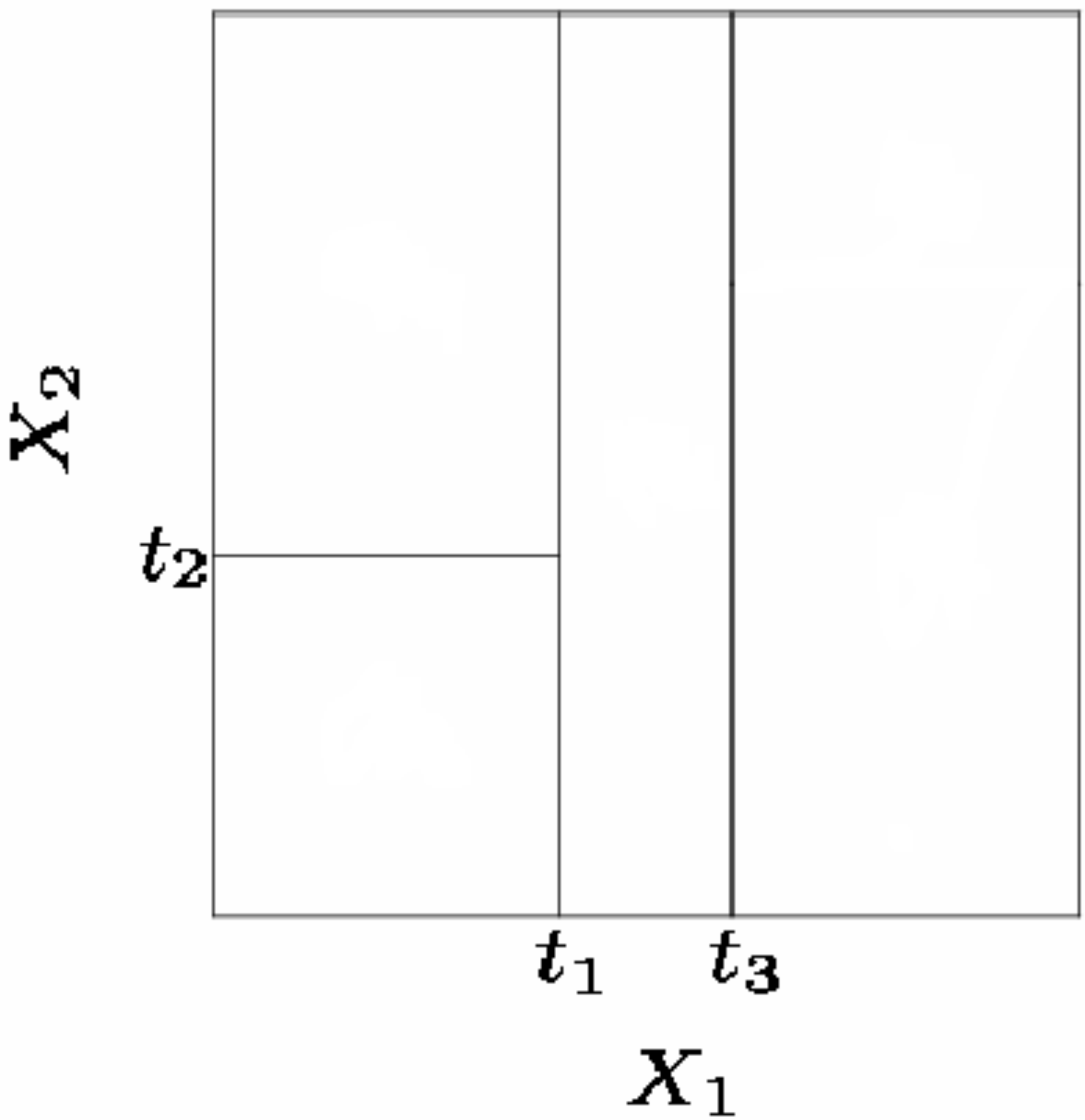
Partitioning Up the Predictor Space

1. First split on $X_1 = t_1$
2. If $X_1 < t_1$, split on $X_2 = t_2$



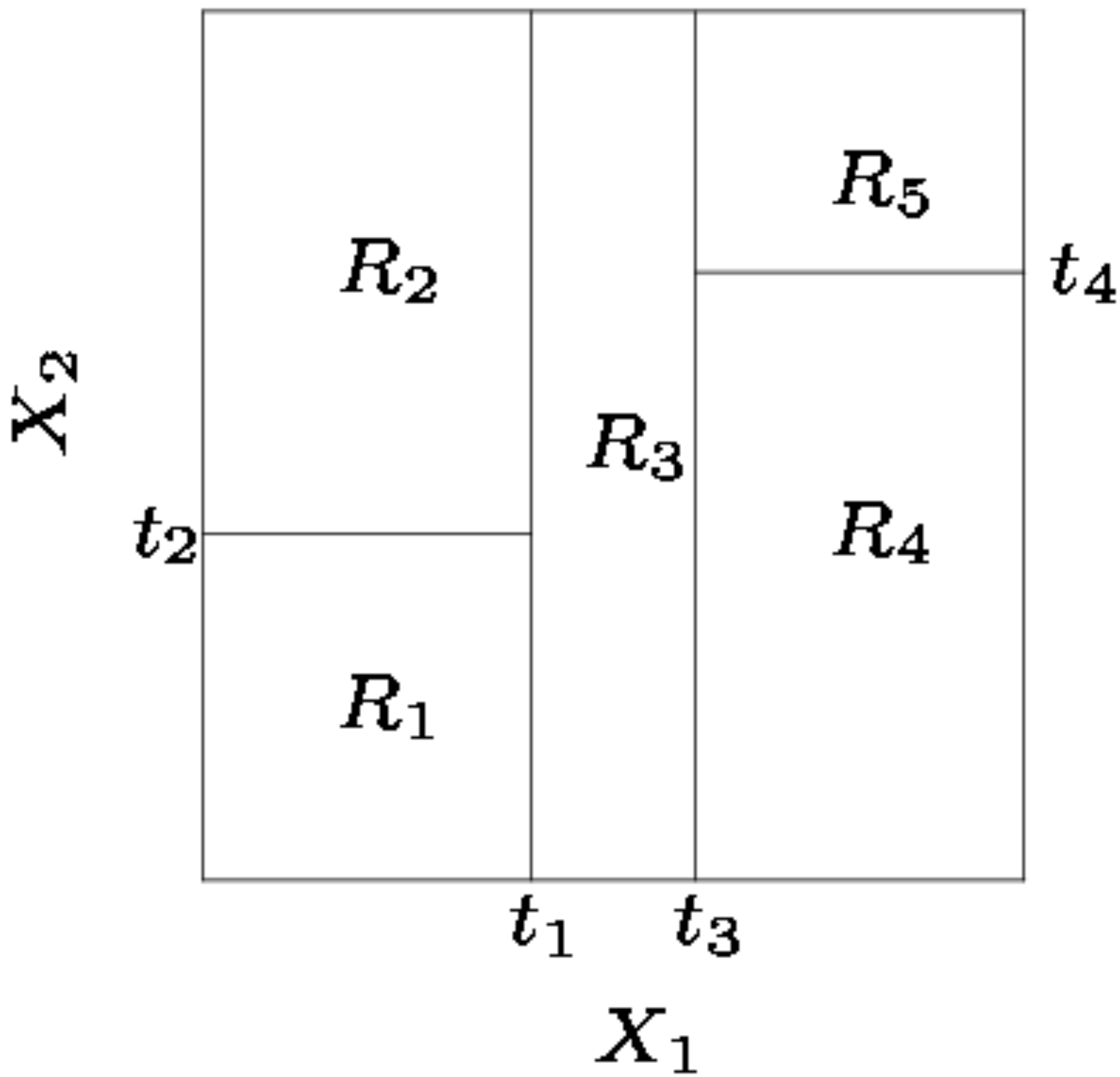
Partitioning Up the Predictor Space

1. First split on $X_1=t_1$
2. If $X_1<t_1$, split on $X_2=t_2$
3. If $X_1>t_1$, split on $X_1=t_3$



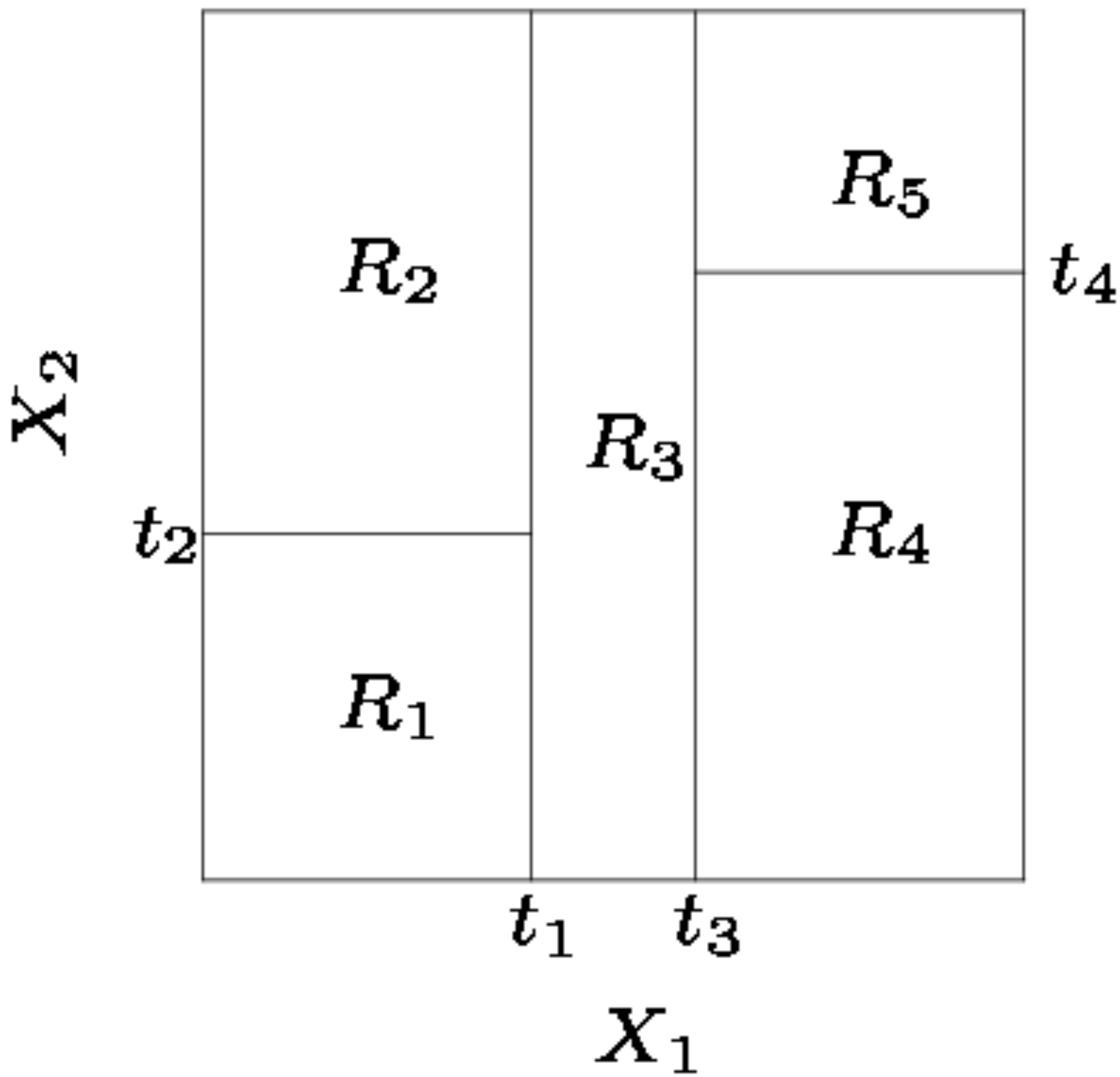
Partitioning Up the Predictor Space

1. First split on $X_1 = t_1$
2. If $X_1 < t_1$, split on $X_2 = t_2$
3. If $X_1 > t_1$, split on $X_1 = t_3$
4. If $X_1 > t_3$, split on $X_2 = t_4$



Partitioning Up the Predictor Space

1. First split on $X_1=t_1$
2. If $X_1<t_1$, split on $X_2=t_2$
3. If $X_1>t_1$, split on $X_1=t_3$
4. If $X_1>t_3$, split on $X_2=t_4$
5. stop



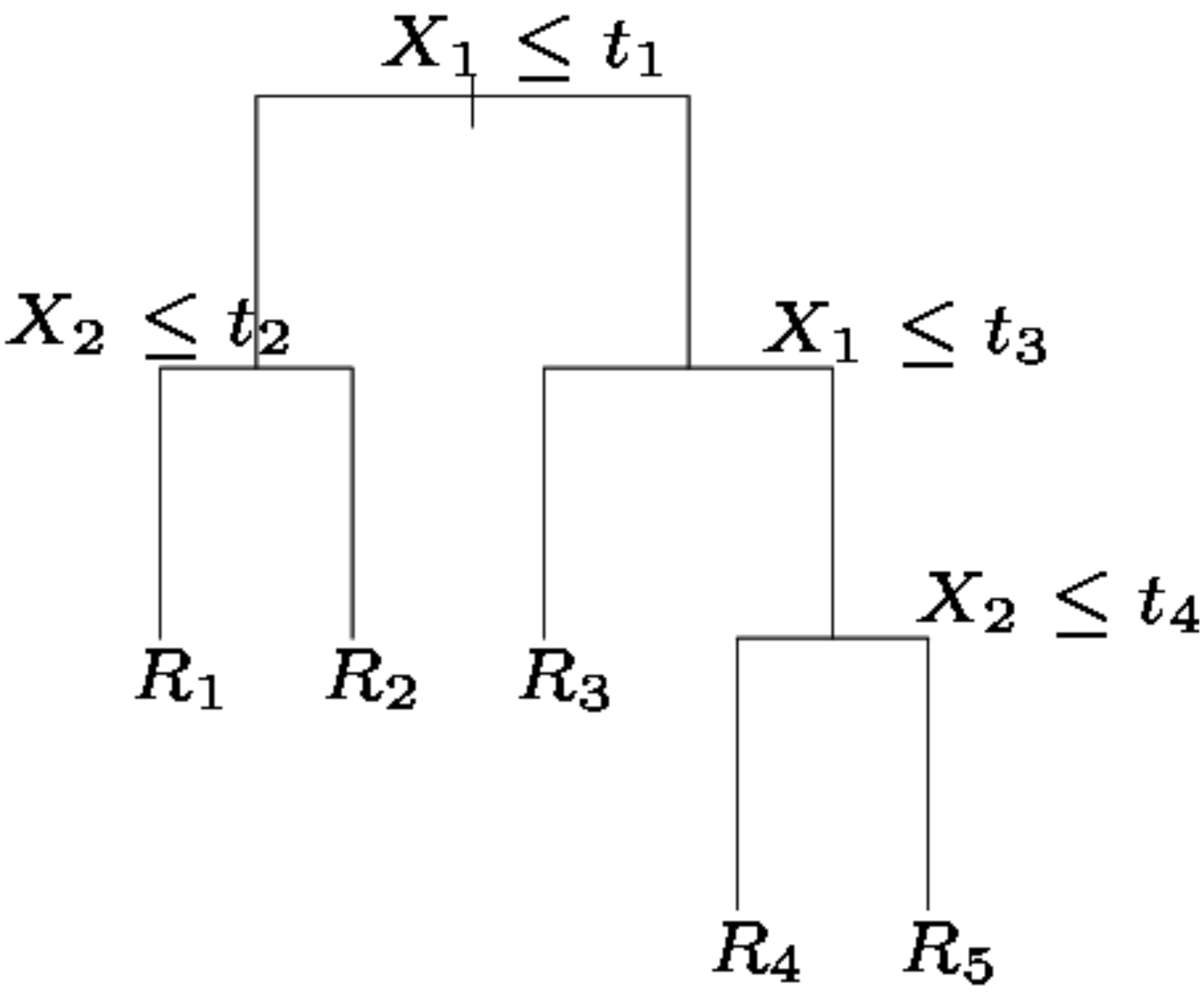
Stopping criteria

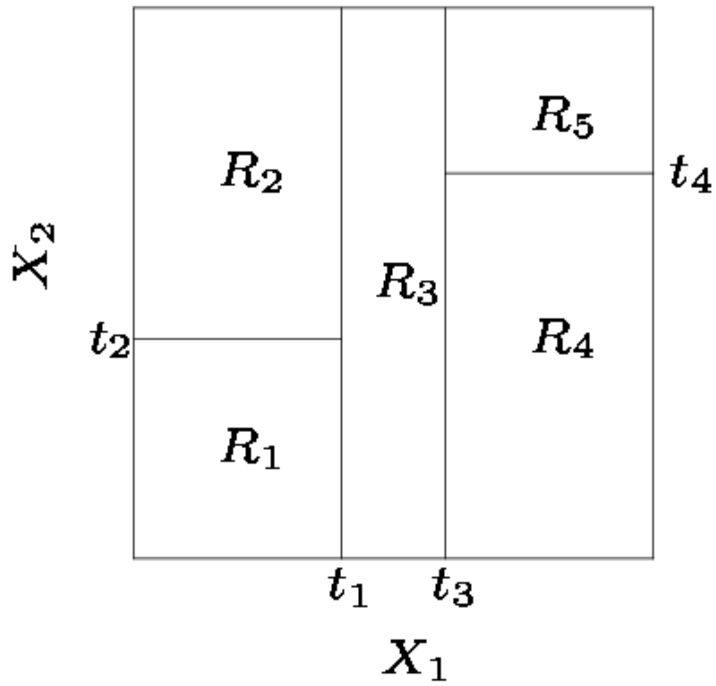
- As the number of splits increase,

  the number of observations in the split regions decrease

- Criteria 1: Stop when the max number of observations falls below threshold
- Criteria 2: Stop when the resulting MSE decrease is small enough
- Criteria 3: Fix the number of splits

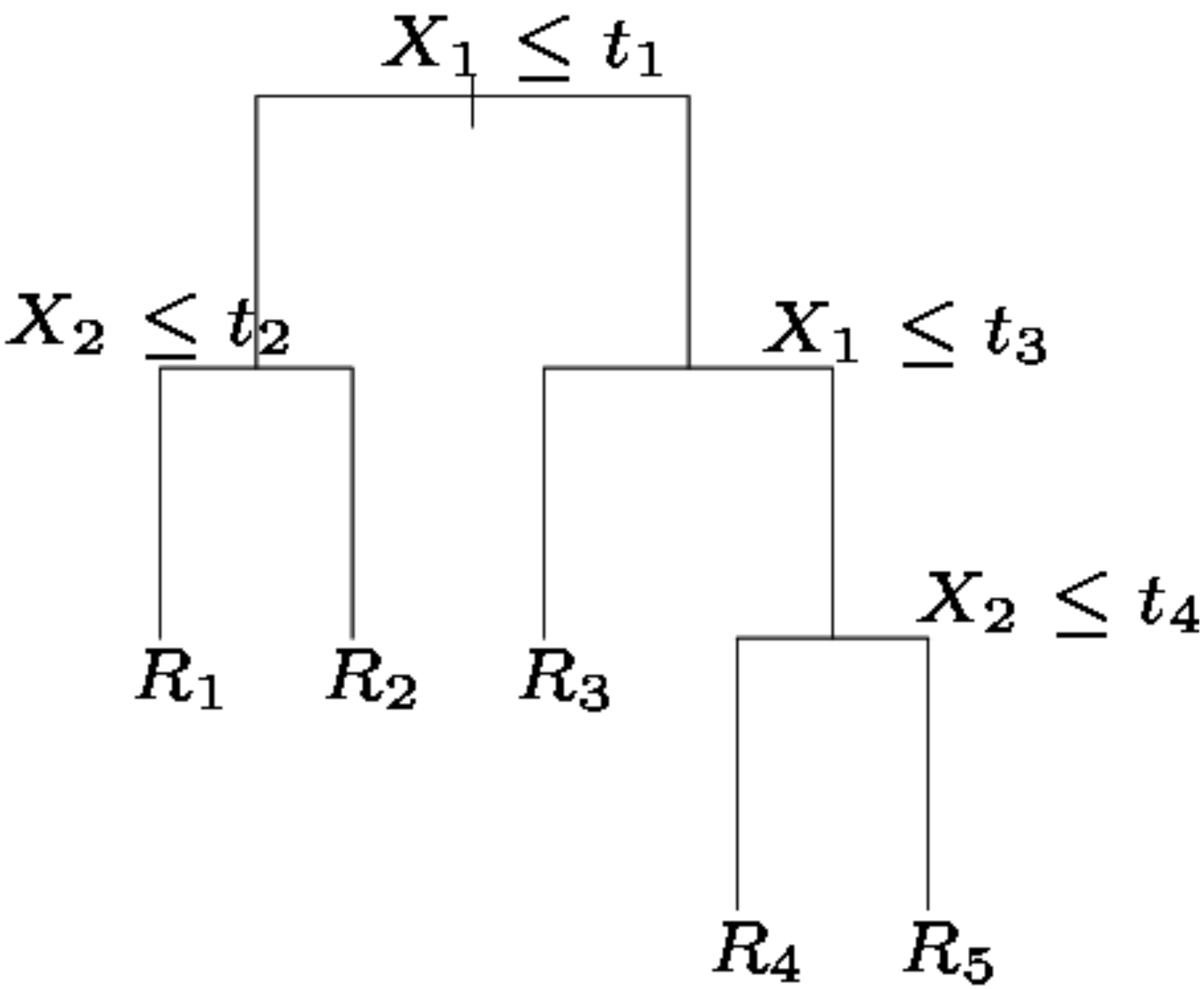
Decision Tree

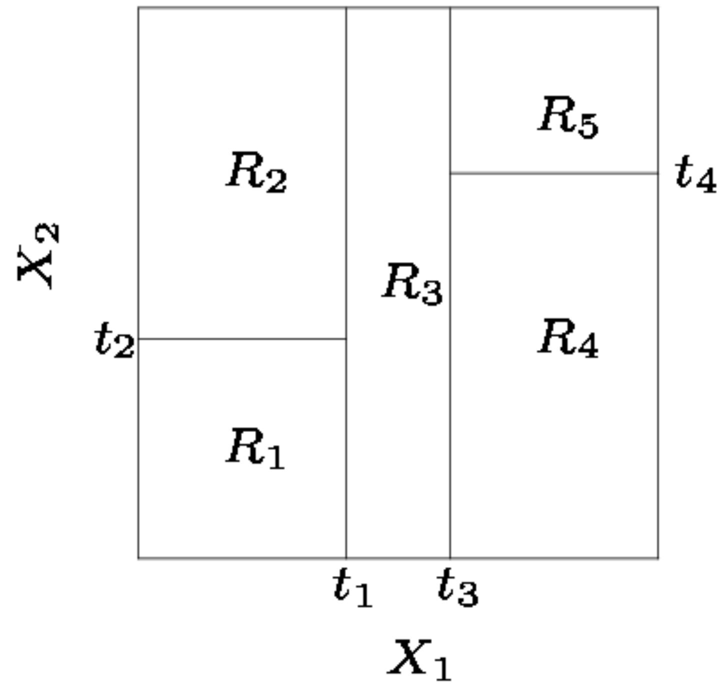




graphical representation of the splits

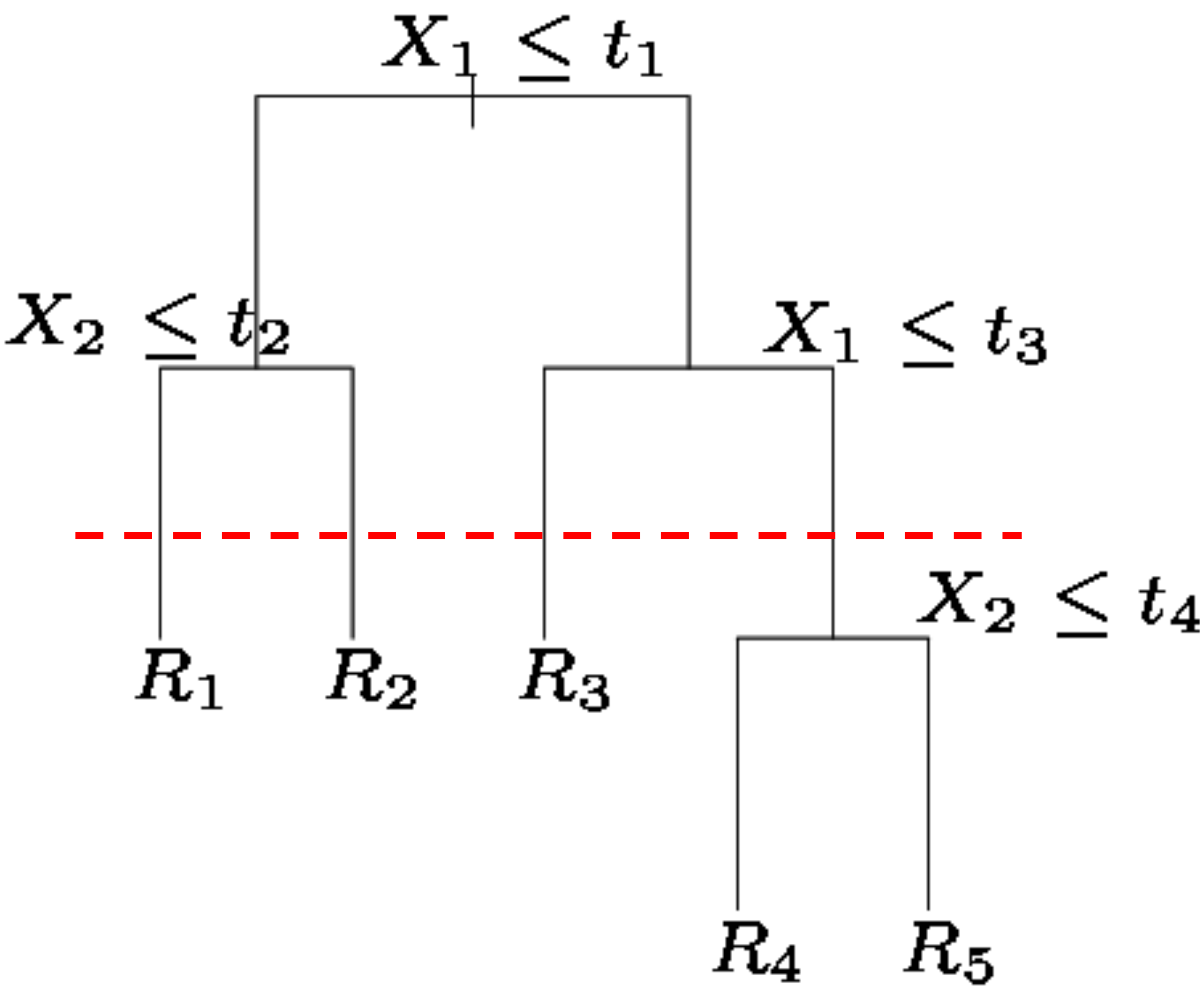
Decision Tree

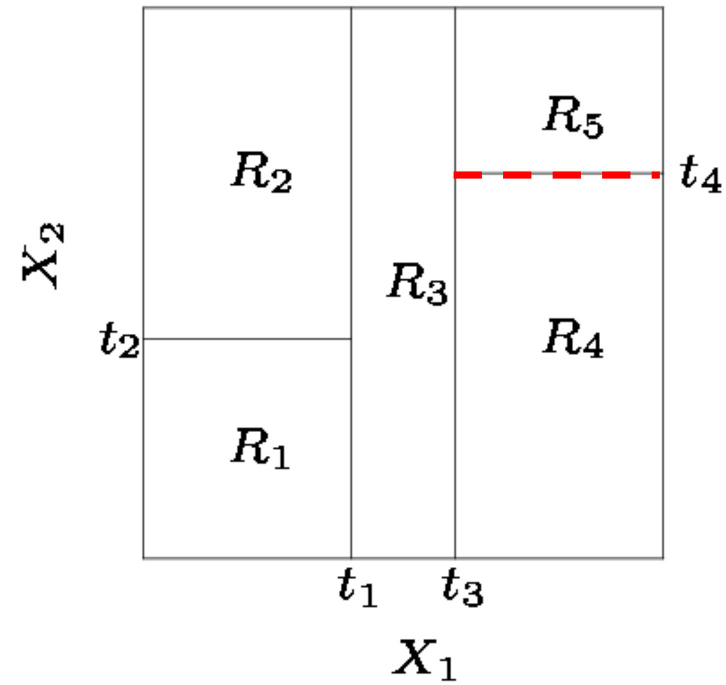




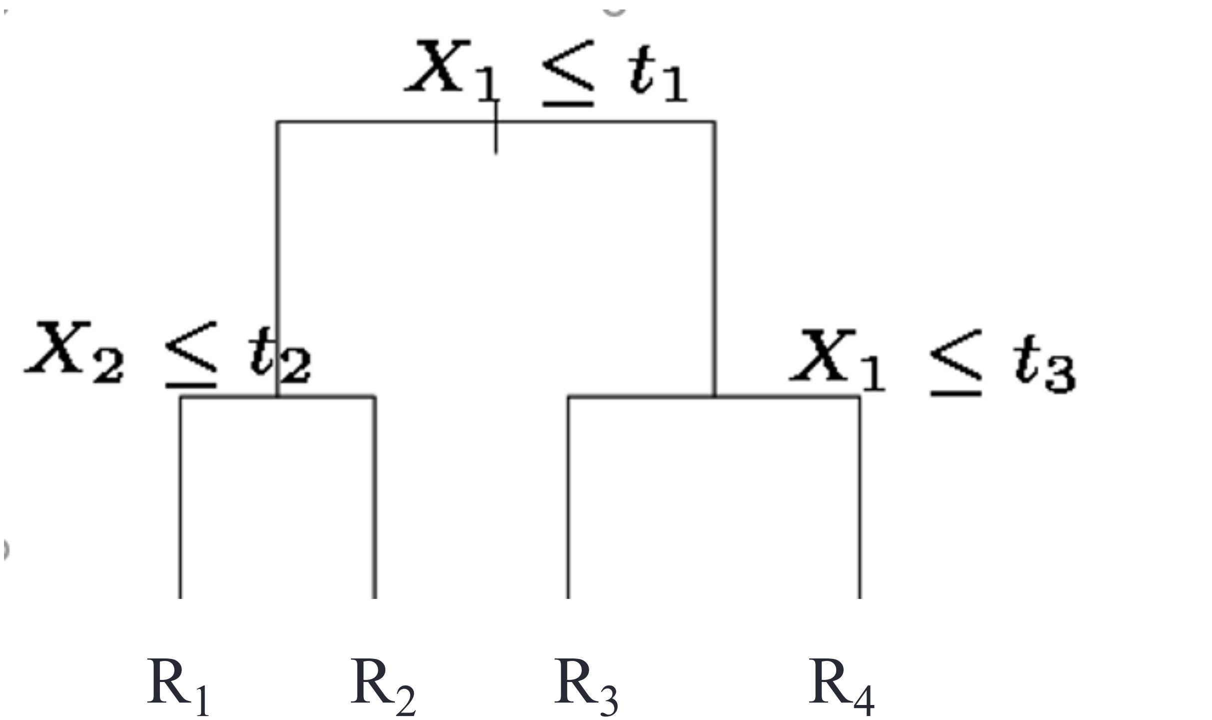
n. regions
=
n. terminal nodes

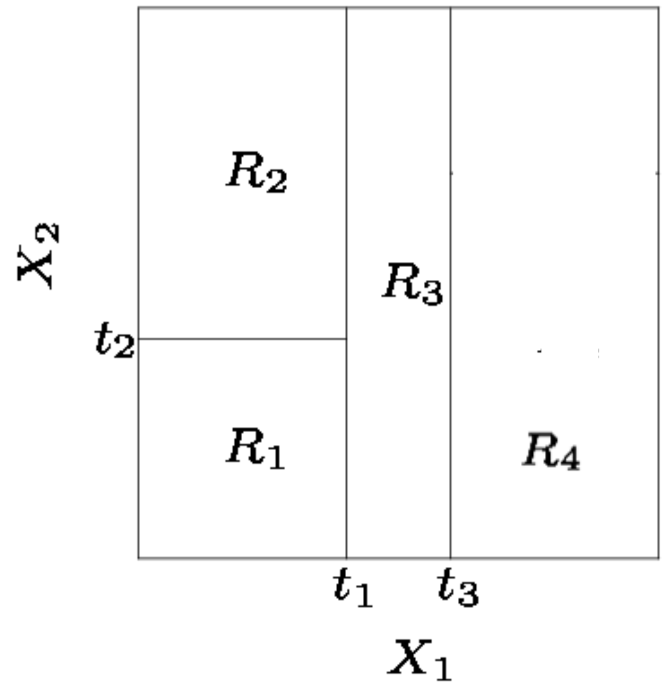
Decision Tree



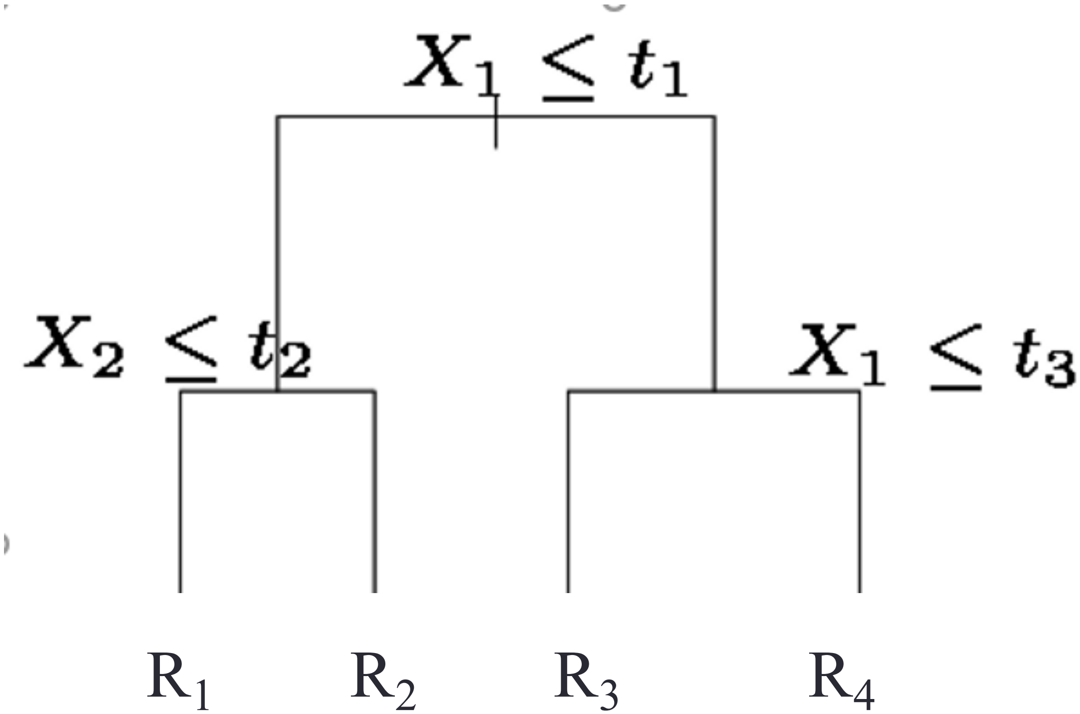


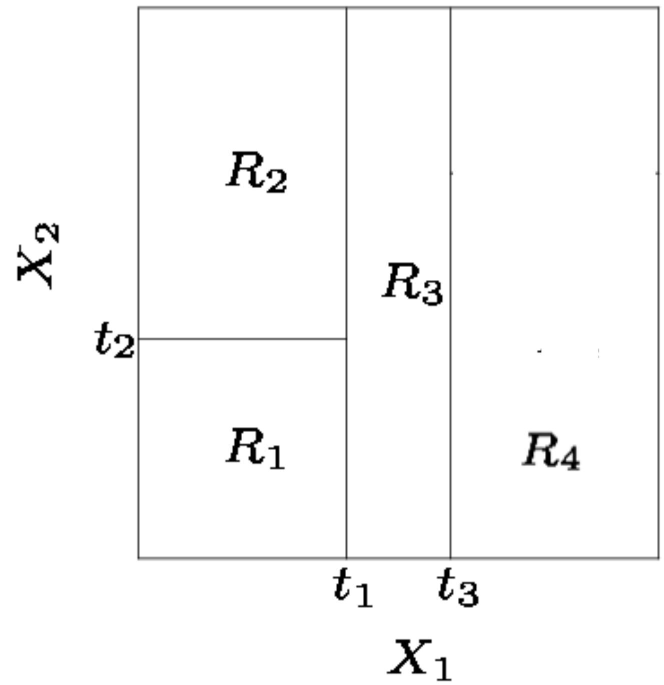
Pruned Tree





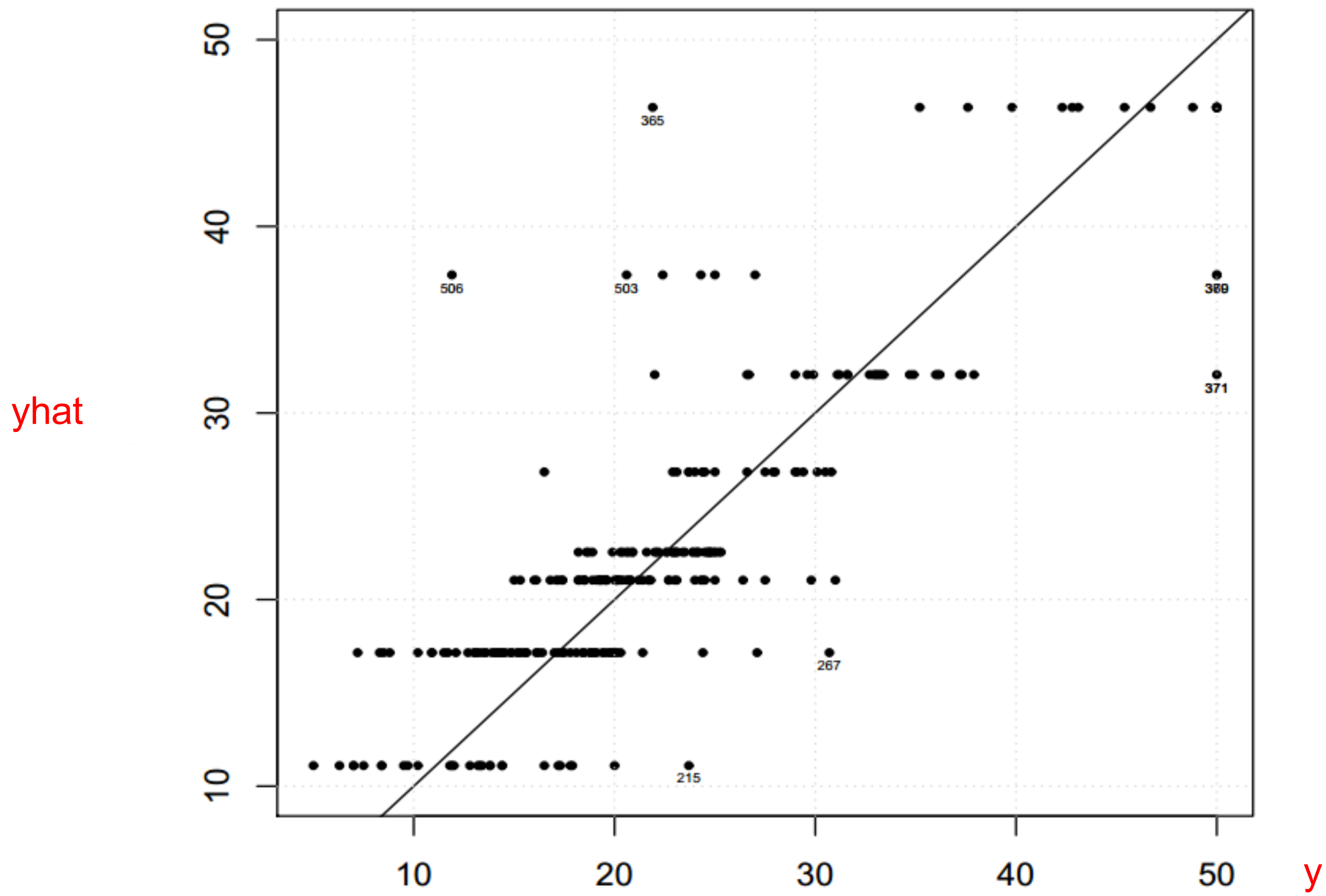
Decision Tree





n. of terminal regions = n. of predicted values

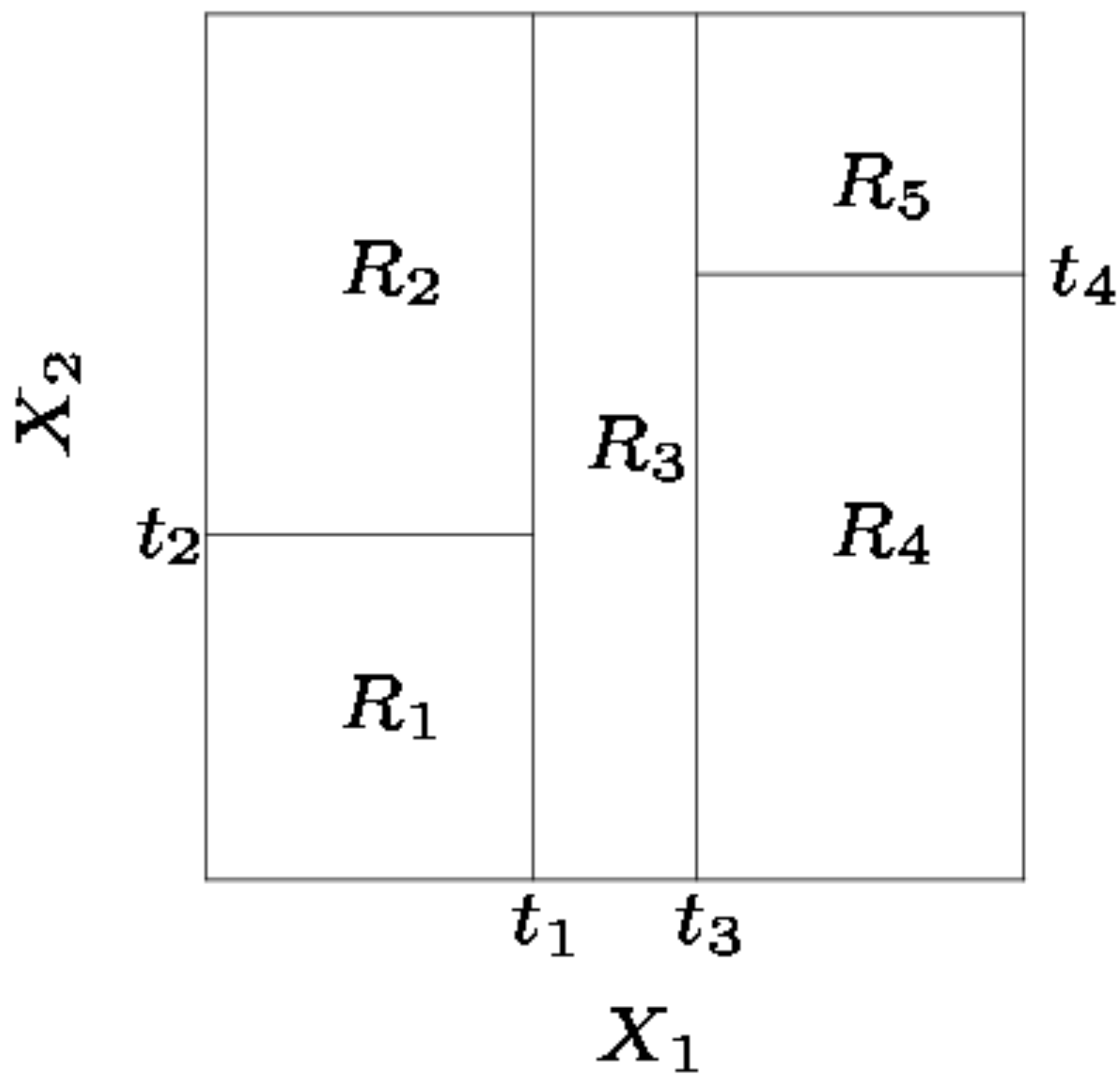
Decision Tree accuracy



n. of terminal regions = n. of predicted values

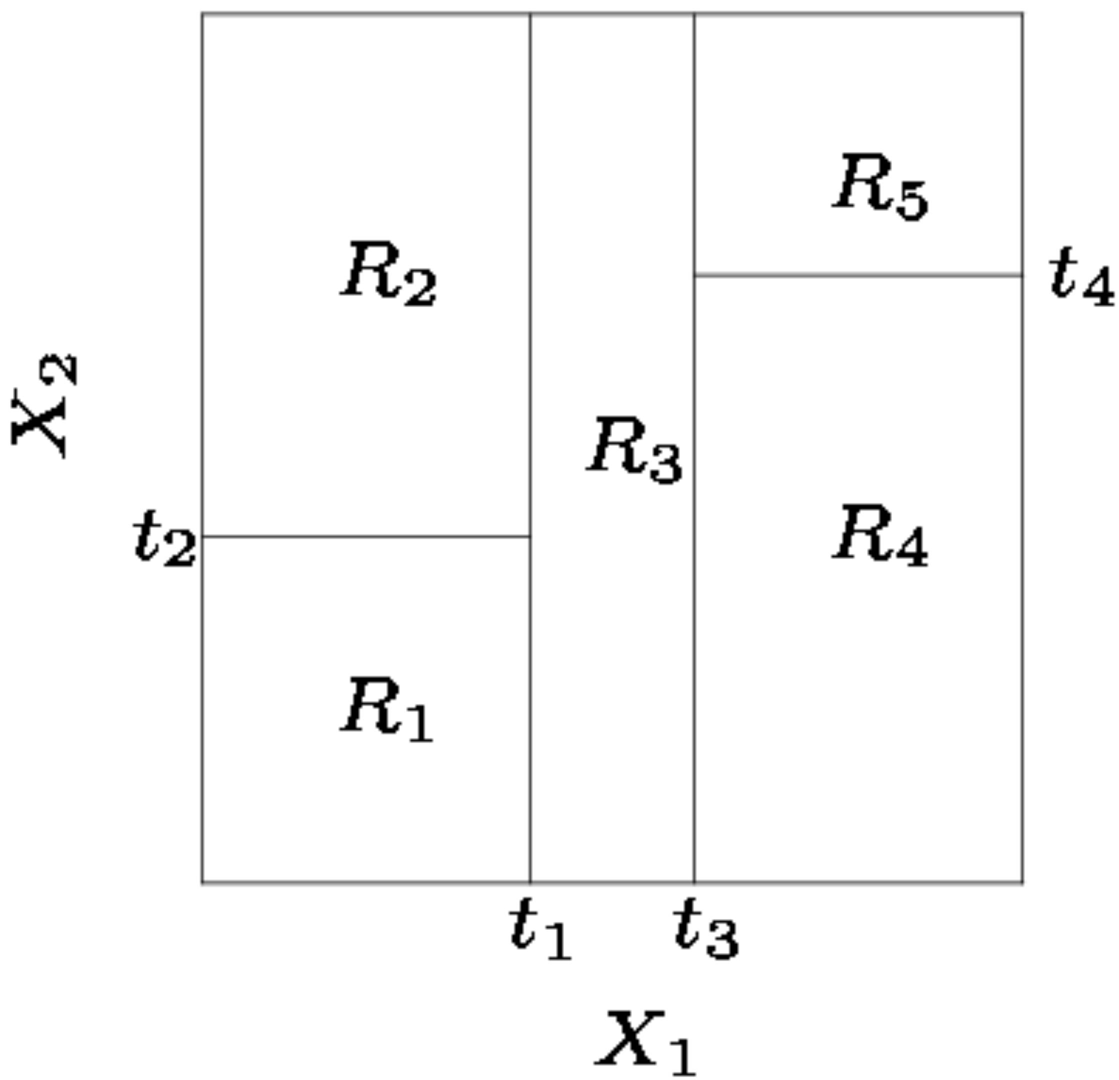
Rectangular regions

- CART models partition the predictor space into regions with special shape



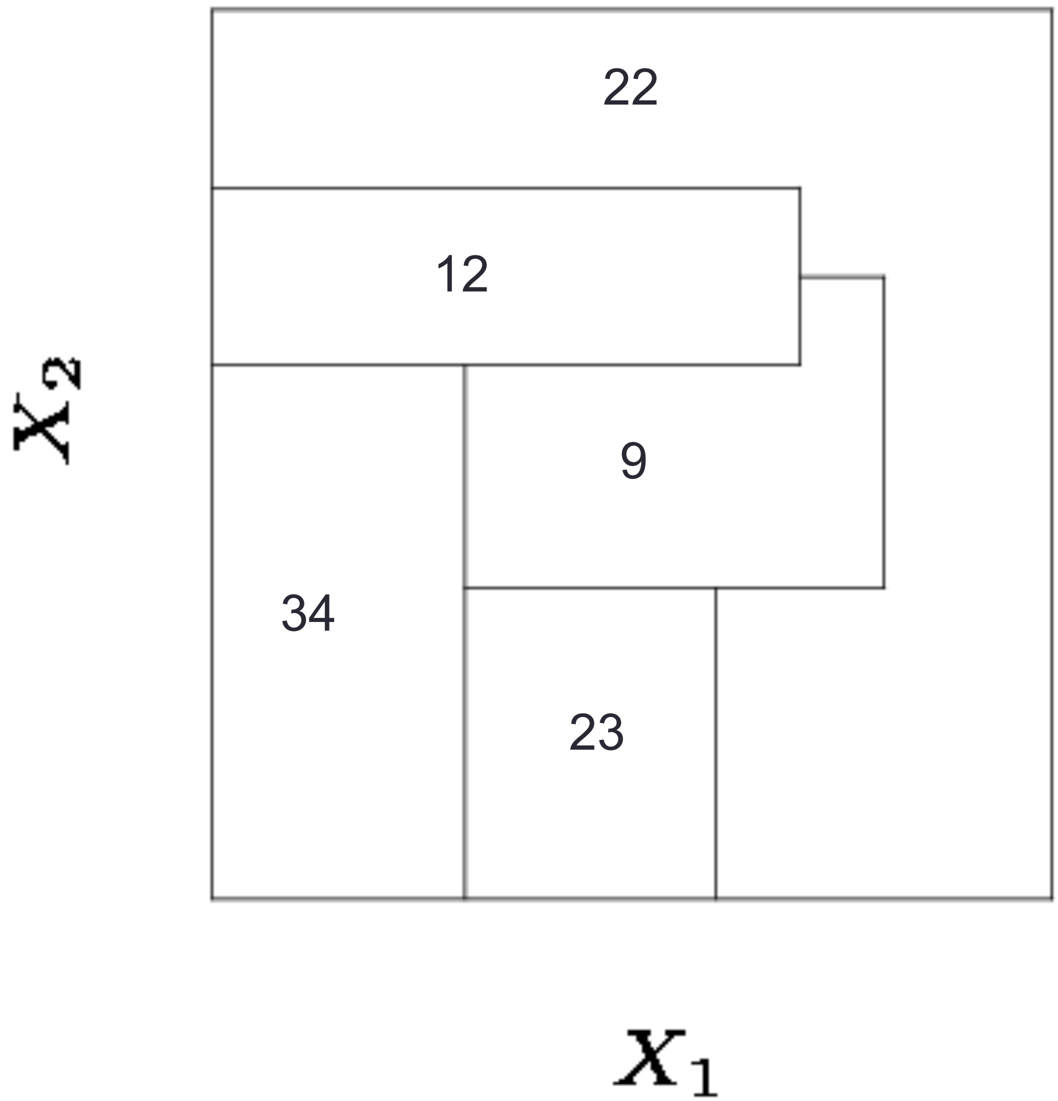
Rectangular regions

- CART models partition the predictor space into regions with special shape
- Regions are always rectangular and disjoint



Not possible

- This partitioning cannot result from a regression tree



Not possible

- This partitioning cannot result from a regression tree
- Region 9 is not rectangular