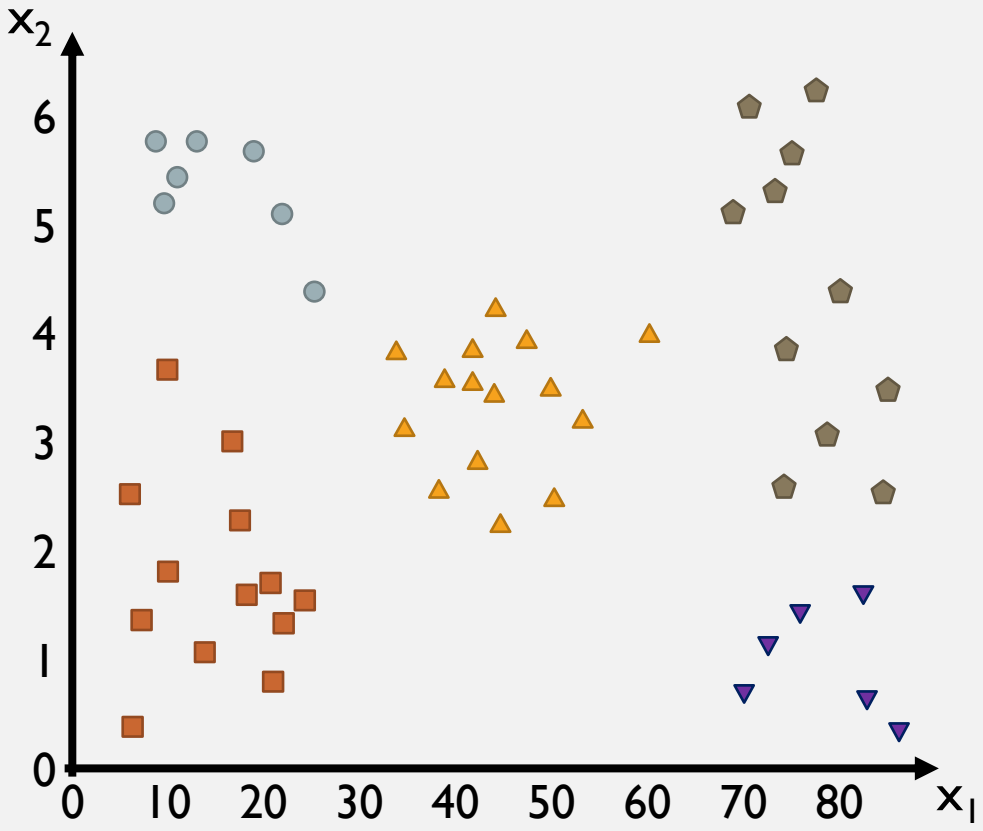
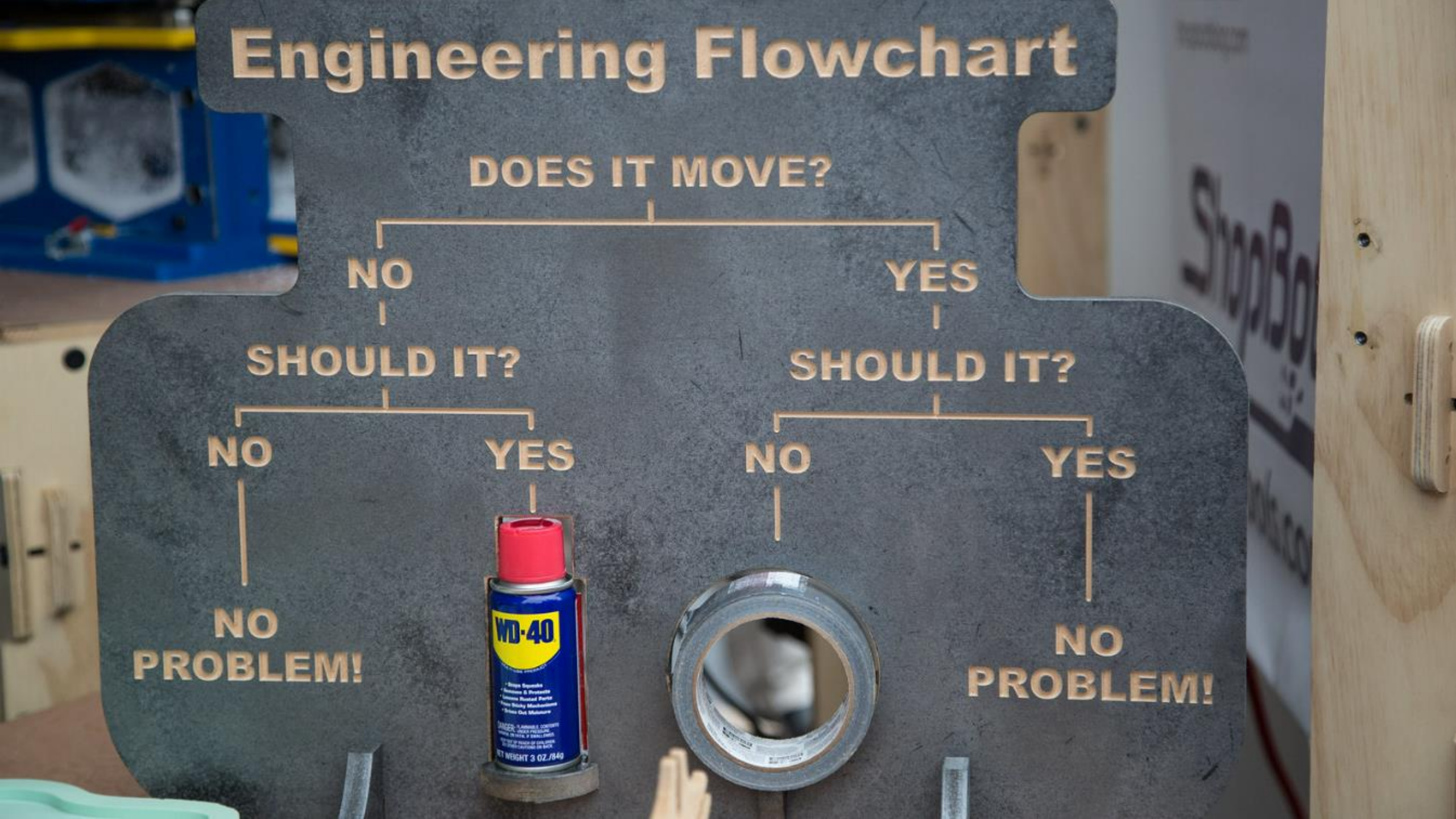
TREE-BASED MODELS

Lecture 3

MAL1, 2024

DECISION TREES




Engineering Flowchart
DOES IT MOVE?
NO
YES
SHOULD IT?
SHOULD IT?
NO
YES
NO
YES
NO PROBLEM!
NO PROBLEM!
WD-40

HOW DECISION TREES WORK

- Step 1
- Step 2
- Step 3
- Stop when

A LOAN IN THE BANK

A FICTITIOUS EXAMPLE

id	salary	savings	debt	class
1	Low	High	True	Approved
2	Low	Low	False	Declined
3	High	Low	False	Approved
4	Low	Low	True	Declined
5	High	Low	True	Approved
6	High	High	False	Approved
7	High	Low	False	Approved
8	Low	Low	True	Declined
9	High	High	True	Approved
10	Low	Low	False	Declined
11	Low	High	False	Approved
12	Low	Low	True	Declined

How do we decide which feature to branch off on?

A LOAN IN THE BANK

A FICTITIOUS EXAMPLE

id	salary	savings	debt	class
1	Low	High	True	Approved
2	Low	Low	False	Declined
3	High	Low	False	Approved
4	Low	Low	True	Declined
5	High	Low	True	Approved
6	High	High	False	Approved
7	High	Low	False	Approved
8	Low	Low	True	Declined
9	High	High	True	Approved
10	Low	Low	False	Declined
11	Low	High	False	Approved
12	Low	Low	True	Declined

The Gini impurity index

A LOAN IN THE BANK

A FICTITIOUS EXAMPLE

id	salary	savings	debt	class
1	Low	High	True	Approved
2	Low	Low	False	Declined
3	High	Low	False	Approved
4	Low	Low	True	Declined
5	High	Low	True	Approved
6	High	High	False	Approved
7	High	Low	False	Approved
8	Low	Low	True	Declined
9	High	High	True	Approved
10	Low	Low	False	Declined
11	Low	High	False	Approved
12	Low	Low	True	Declined

The Gini impurity index

$$G_{\text{salary}}(D) = 0.24$$
$$G_{\text{savings}}(D) = 0.31$$
$$G_{\text{debt}}(D) = 0.47$$

A LOAN IN THE BANK

A FICTITIOUS EXAMPLE

id	salary	savings	debt	class
1	Low	High	True	Approved
2	Low	Low	False	Declined
3	High	Low	False	Approved
4	Low	Low	True	Declined
5	High	Low	True	Approved
6	High	High	False	Approved
7	High	Low	False	Approved
8	Low	Low	True	Declined
9	High	High	True	Approved
10	Low	Low	False	Declined
11	Low	High	False	Approved
12	Low	Low	True	Declined

Beginning to draw the tree

A LOAN IN THE BANK

A FICTITIOUS EXAMPLE

id	salary	savings	debt	class
1	Low	High	True	Approved
2	Low	Low	False	Declined
3	High	Low	False	Approved
4	Low	Low	True	Declined
5	High	Low	True	Approved
6	High	High	False	Approved
7	High	Low	False	Approved
8	Low	Low	True	Declined
9	High	High	True	Approved
10	Low	Low	False	Declined
11	Low	High	False	Approved
12	Low	Low	True	Declined

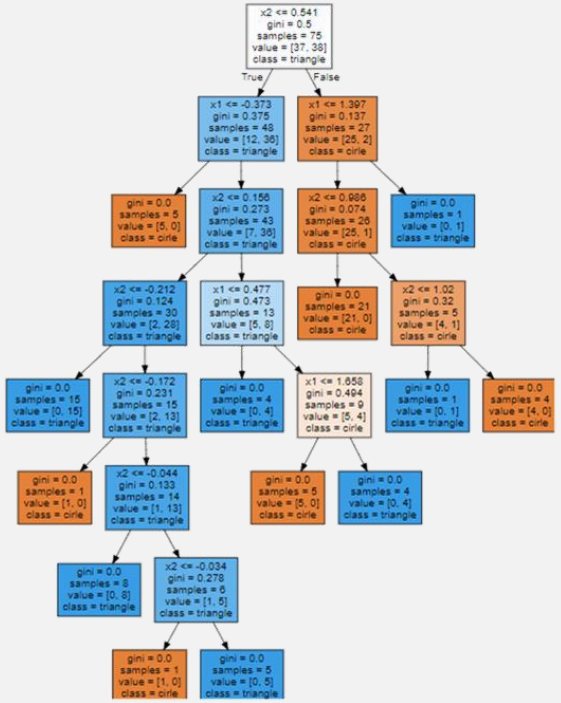
Finishing the tree

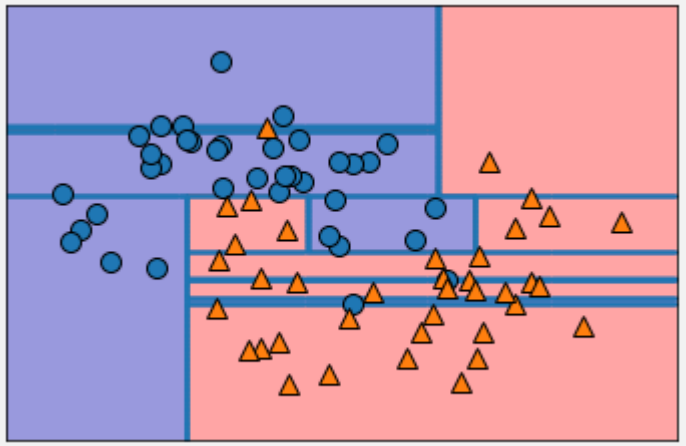
LEARNING DECISION TREES

VISUALIZATION

Jupyter Notebook **Decision Trees I: Visualization and hyperparameters**

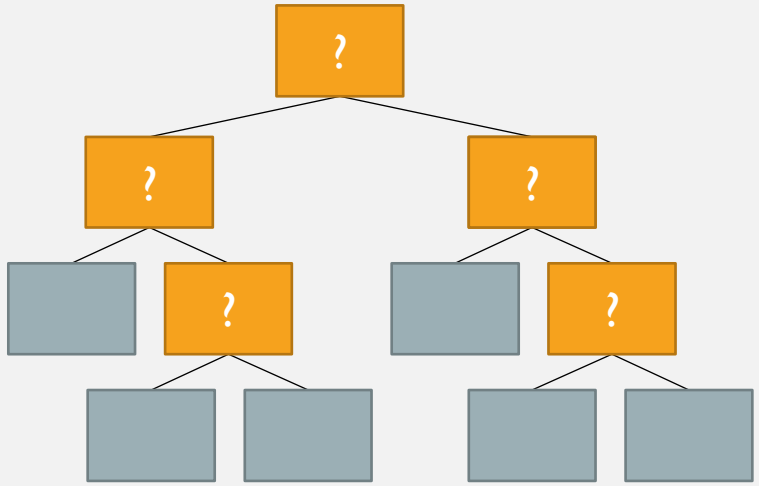
VISUALIZATION





OVERFITTING AND HYPERPARAMETERS

```
Accuracy on training data: 1.0
Accuracy on testing data: 0.92
```



max_depth

max_leaf_notes

min_samples_split

(**criterion**)

Tuning these parameters is called *pre-pruning*

PRE-PRUNING

Jupyter Notebook **Decision Trees I: Visualization and hyperparameters**

PROS AND CONS OF DECISION TREES

Pros

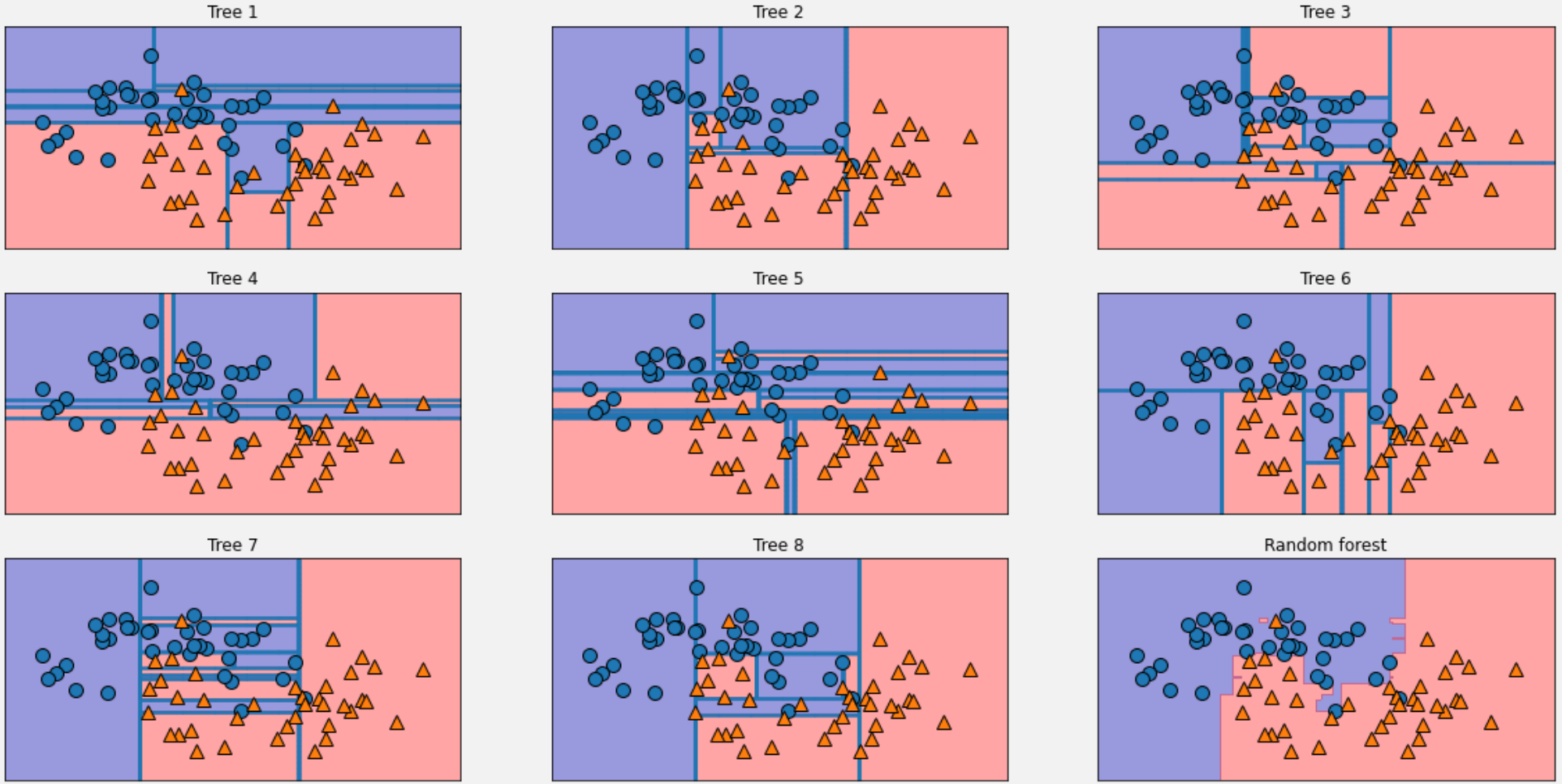
Cons

ENSEMBLES OF DECISION TREES

- **Random forests** *(bagging)*

- **Gradient boosted decision trees** *(boosting)*

RANDOM FORESTS



RANDOMIZATION I: BOOTSTRAPPING

	f_1	f_2	f_3	f_4	f_5	f_6
x_1	45	5	21	45	15	1
x_2	87	2	12	44	64	2
x_3	24	8	15	43	36	3
x_4	67	7	17	44	87	2
x_5	13	5	12	44	65	3
x_6	87	4	16	42	34	1
x_7	89	7	13	42	2	2
x_8	68	3	14	43	54	3
x_9	35	6	11	41	63	2

7
9
4
8
7
2
3
3
8

A bootstrap dataset

	f_1	f_2	f_3	f_4	f_5	f_6
x_7	89	7	13	42	2	2
x_9	35	6	11	41	63	2
x_4	67	7	17	44	87	2
x_8	68	3	14	43	54	3
x_7	89	7	13	42	2	2
x_2	87	2	12	44	64	2
x_3	24	8	15	43	36	3
x_3	24	8	15	43	36	3
x_8	68	3	14	43	54	3

RANDOMIZATION I: BOOTSTRAPPING

Dataset for tree 1

	f_1	f_2	f_3	f_4	f_5	f_6
x_7	89	7	13	42	2	2
x_9	35	6	11	41	63	2
x_4	67	7	17	44	87	2
x_8	68	3	14	43	54	3
x_7	89	7	13	42	2	2
x_2	87	2	12	44	64	2
x_3	24	8	15	43	36	3
x_3	24	8	15	43	36	3
x_8	68	3	14	43	54	3

Dataset for tree 2

	f_1	f_2	f_3	f_4	f_5	f_6
x_6	87	4	16	42	34	1
x_8	68	3	14	43	54	3
x_2	87	2	12	44	64	2
x_2	87	2	12	44	64	2
x_3	24	8	15	43	36	3
x_7	89	7	13	42	2	2
x_4	67	7	17	44	87	2
x_2	87	2	12	44	64	2
x_8	68	3	14	43	54	3

Dataset for tree 3

	f_1	f_2	f_3	f_4	f_5	f_6
x_3	24	8	15	43	36	3
x_3	24	8	15	43	36	3
x_8	68	3	14	43	54	3
x_7	89	7	13	42	2	2
x_1	45	5	21	45	15	1
x_1	45	5	21	45	15	1
x_6	87	4	16	42	34	1
x_5	13	5	12	44	65	3
x_7	89	7	13	42	2	2

RANDOMIZATION II: FEATURE SELECTION

Dataset for tree 1

	f_1	f_2	f_3	f_4	f_5	f_6
x_7	89	7	13	42	2	2
x_9	35	6	11	41	63	2
x_4	67	7	17	44	87	2
x_8	68	3	14	43	54	3
x_7	89	7	13	42	2	2
x_2	87	2	12	44	64	2
x_3	24	8	15	43	36	3
x_3	24	8	15	43	36	3
x_8	68	3	14	43	54	3

RANDOMIZATION II: FEATURE SELECTION

Dataset for tree 1

	f_1	f_2	f_3	f_4	f_5	f_6
x_7	89	7	13	42	2	2
x_9	35	6	11	41	63	2
x_4	67	7	17	44	87	2
x_8	68	3	14	43	54	3
x_7	89	7	13	42	2	2
x_2	87	2	12	44	64	2
x_3	24	8	15	43	36	3
x_3	24	8	15	43	36	3
x_8	68	3	14	43	54	3

max_features

max_features = n_features

max_features = 1

RANDOMIZATION II: FEATURE SELECTION

Dataset for tree 1

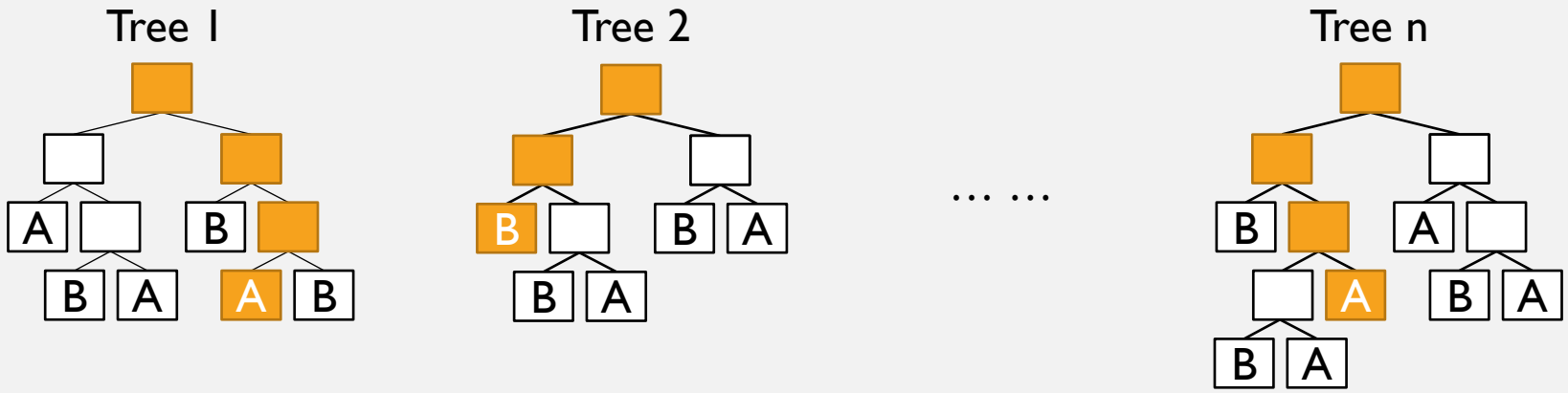
	f_1	f_2	f_3	f_4	f_5	f_6
x_7	89	7	13	42	2	2
x_9	35	6	11	41	63	2
x_4	67	7	17	44	87	2
x_8	68	3	14	43	54	3
x_7	89	7	13	42	2	2
x_2	87	2	12	44	64	2
x_3	24	8	15	43	36	3
x_3	24	8	15	43	36	3
x_8	68	3	14	43	54	3

A low value of `max_features`

A high value of `max_features`

A rule of thumb

PREDICTIONS USING RANDOM FORESTS



PROS AND CONS OF RANDOM FORESTS

Pros

Cons

FEATURE IMPORTANCE

Jupyter Notebook **Decision Trees 2: Feature importance and ensembles of trees**

GRADIENT BOOSTED DECISION TREES

OR GRADIENT BOOSTED REGRESSION TREES *OR* GRADIENT BOOSTING MACHINES

HYPERPARAMETERS

n_estimators

max_depth

learning_rate

HOW DOES IT WORK?

- First tree
- Second tree
- *n*th tree

CODING BOOSTED TREES

Jupyter Notebook **Decision Trees 2: Feature importance and ensembles of trees**

PROS AND CONS OF GRADIENT BOOSTED DECISION TREES

Pros

Cons

WHEN TO USE WHAT

kortlink.dk/2gtzg

WHEN TO USE WHAT

Tree **Forest** **Boosted tree**


WHERE DOES A DATA SCIENTIST CAMP?
IN A RANDOM FOREST