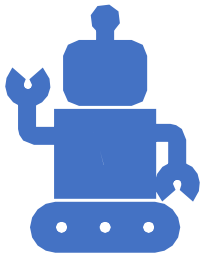
UiO : University of Oslo

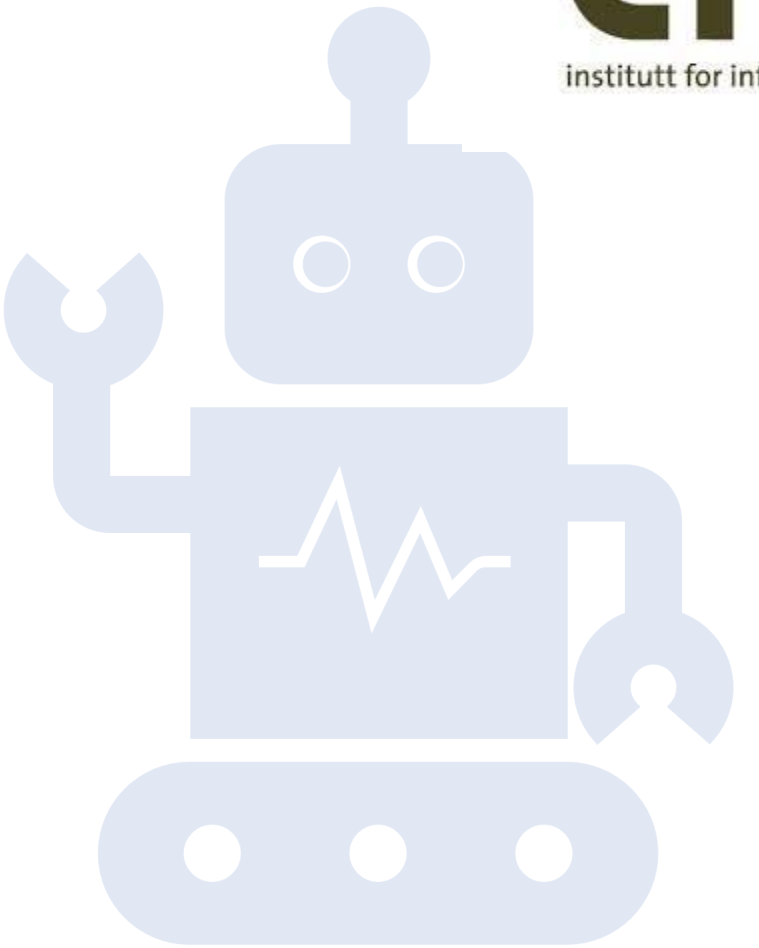
institutt for informatikk

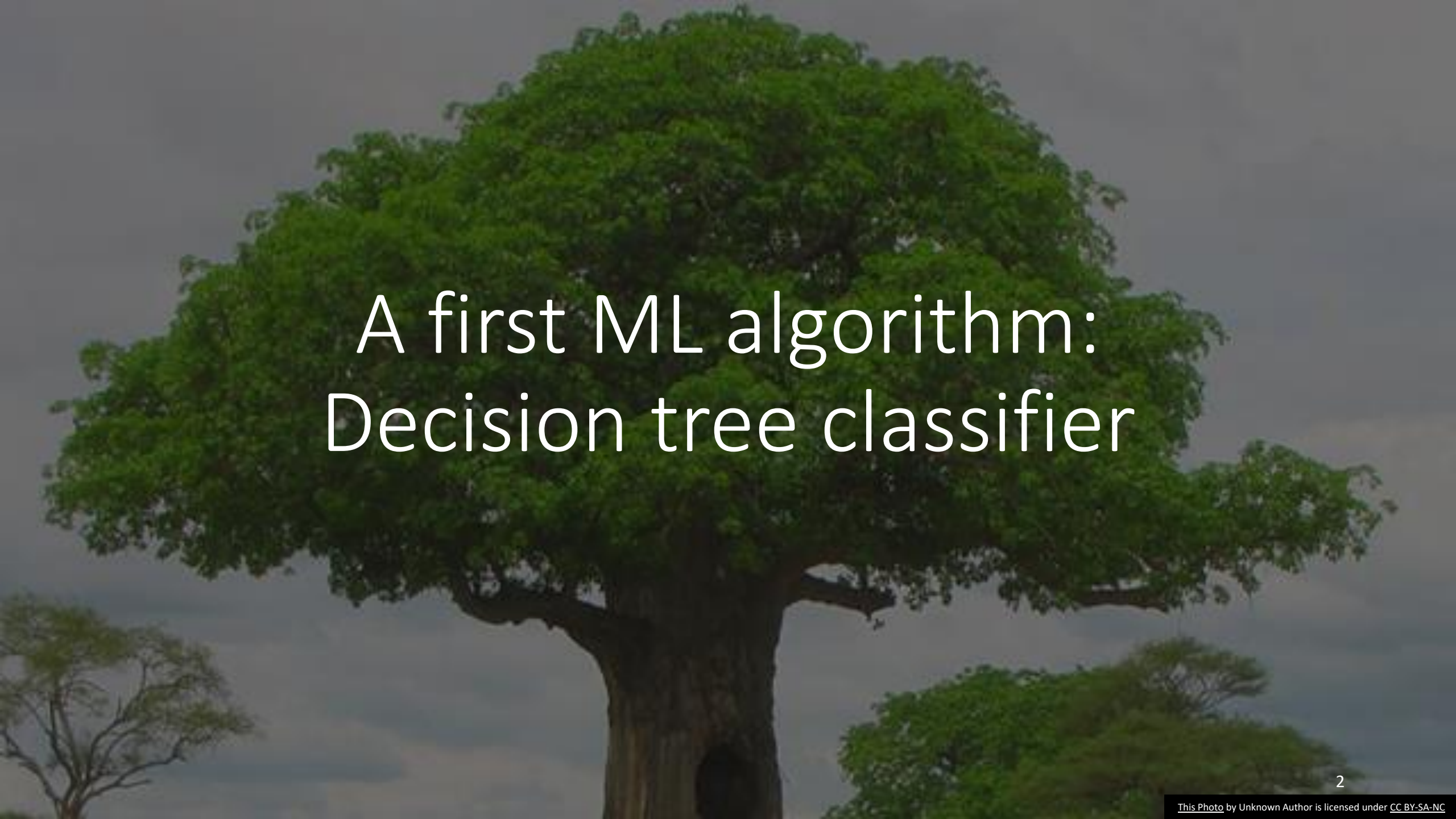


IN3050/IN4050 - Introduction to Artificial Intelligence and Machine Learning

Decision Trees

Jan Tore Lønning



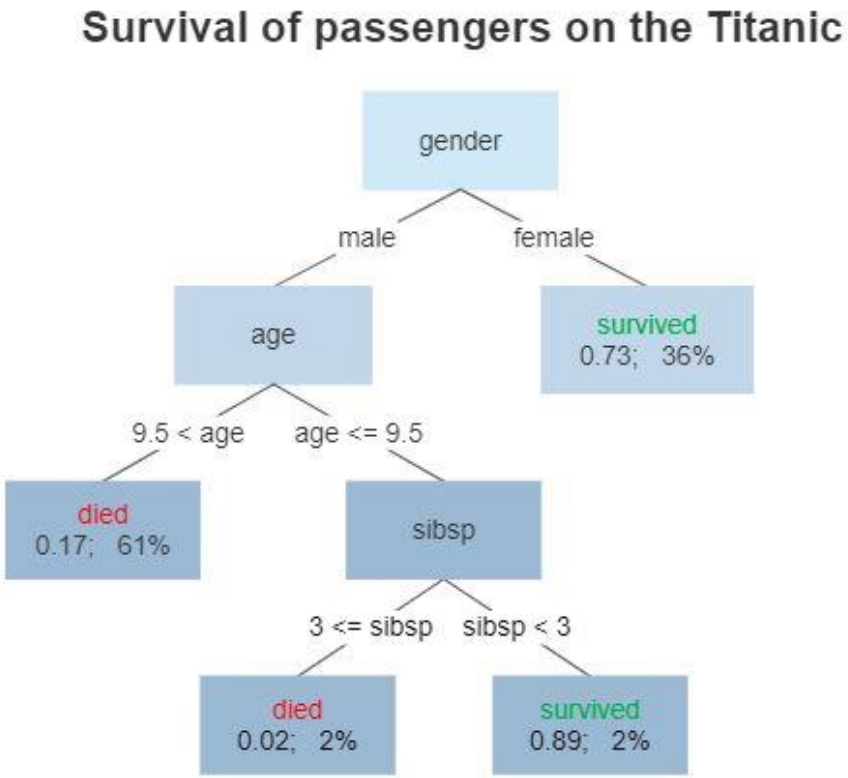


A first ML algorithm: Decision tree classifier

This Photo by Unknown Author is licensed under CC BY-SA-NC

Decision trees

- Decisions are based on some features (questions/answers)
- One answers one question at a time
- The answer to one question decides the next question
- After answering all the relevant questions, you reach a leave node with a conclusion

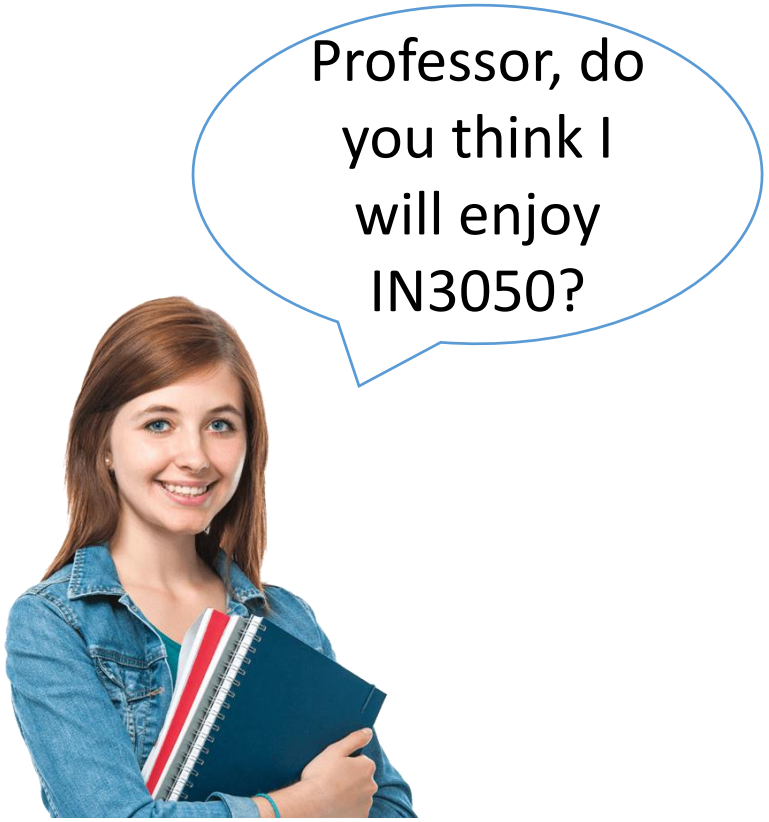


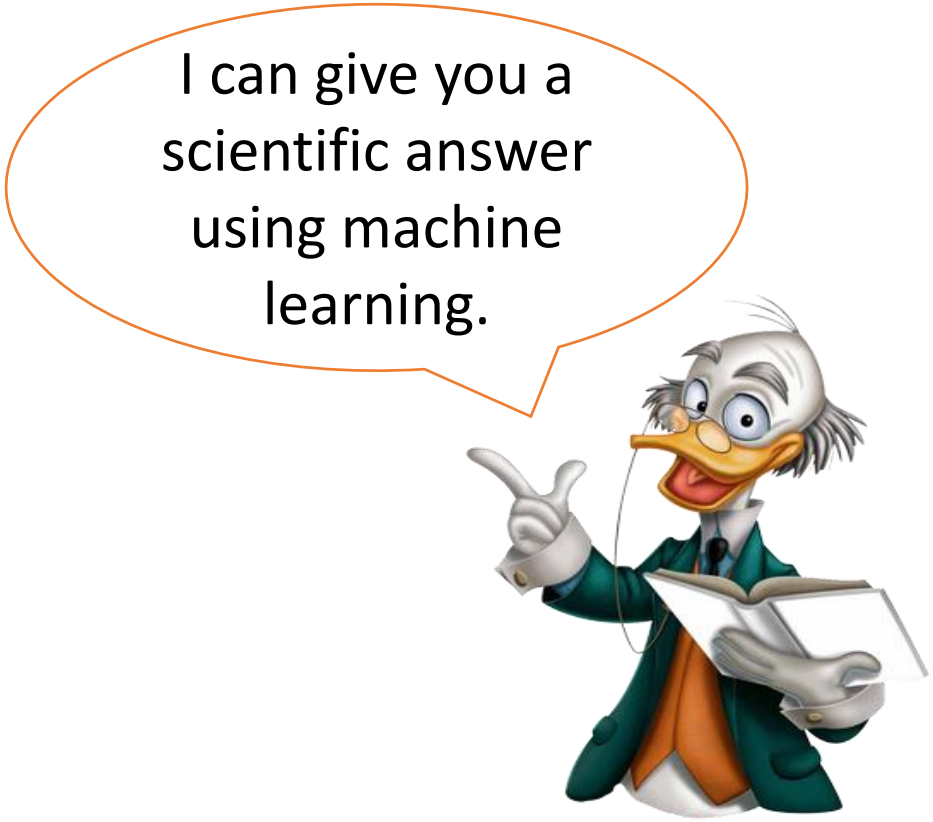
Source: https://en.wikipedia.org/wiki/Decision_tree_learning

Decision trees in AI

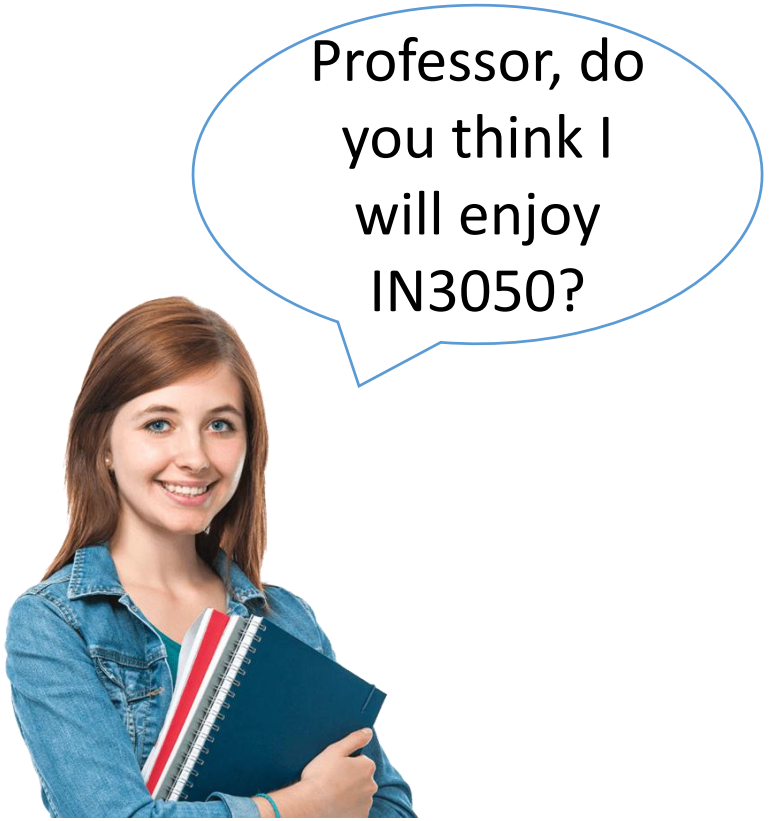
- Decision trees can be used in both rule-based and ML approaches
- Rule-based:
  - A series of if-then-else statements
- Typical architecture for traditional expert systems
  - E.g., medical diagnosis, one symptom at a time
  - Constructed with the knowledge of domain experts
- Decision trees can also be learned from data
  - NEXT

Example: (the numbers are constructed)





Example: (the numbers are constructed)



Baseline

- Survey
  - Asked all the students of 2020
  - 200 answered:
    - 130 yes
    - 70 no
- From this construct a baseline classifier:
  - Choose the majority class
  - Accuracy 0.65=65% (= the fraction of correct answers)
    - Calculated on the 2020 class (the training set)
    - Correct evaluation: test it on a different test set, e.g., the 2021 class.

Example: take 2

The 2020 survey

- Did you enjoy the course?
  - Yes/no
- Do you like mathematics?
  - Yes/no
- Do you have programming experience?
  - None/some (1 or 2 courses) /good (= 3 or more courses)
- Have you taken advanced machine learning courses?
  - Yes/no
- And many more questions, but we must simplify here

Results of the 2020 survey: a data set

Cand no	Enjoy maths	Programming	Adv. ML	Enjoy
1	Y	Good	N	Y
2	Y	Some	N	Y
3	N	Good	Y	N
4	N	None	N	N
5	N	Good	N	Y
6	N	Good	Y	Y
….				

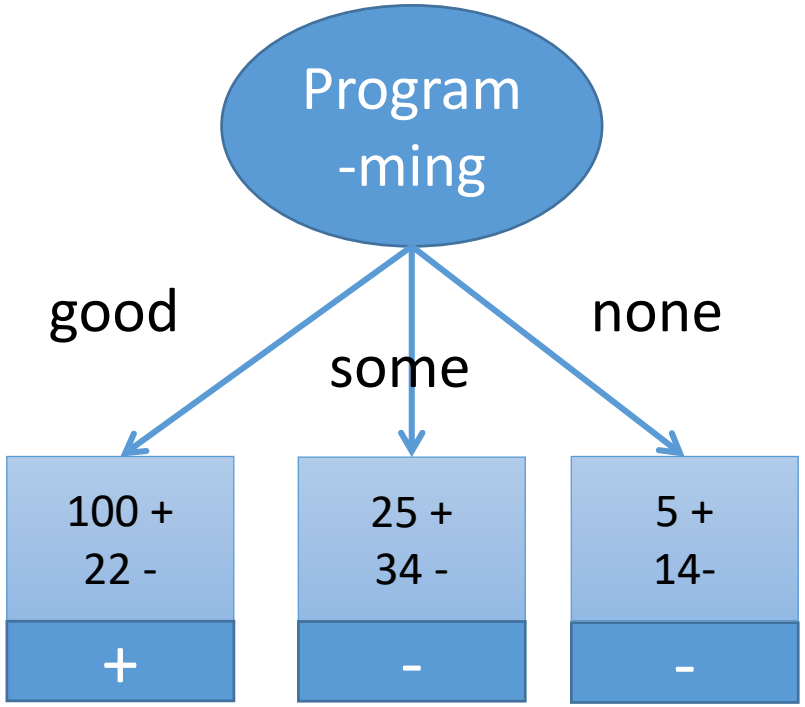
Summary of the 2020 survey

	A	B	C	D	E	F
1	programing	AdvML-course	Like maths	enjoyed	not enjoye	sum
2	good	yes	yes	3	10	13
3	good	yes	no	7	4	11
4	good	no	yes	50	4	54
5	good	no	no	40	4	44
6	some	yes	yes	4	1	5
7	some	yes	no	0	0	0
8	some	no	yes	11	9	20
9	some	no	no	10	24	34
10	none	yes	yes	1	2	3
11	none	yes	no	0	0	0
12	none	no	yes	2	5	7
13	none	no	no	2	7	9
14				130	70	200

Goal: Build a decision tree from the survey

- More precisely: we want a small and efficient tree
  - We want to find the answer with as few questions as possible
  - (With 3 features, we could have asked all the questions, but not if there are many more questions/features).
- To find the best tree is NP-complete.
- Instead, we use a greedy approach
  - Ask the single question which is most informative
    - Then for each answer, ask the question which is most informative given the first answer
      - etc

Decision stumps

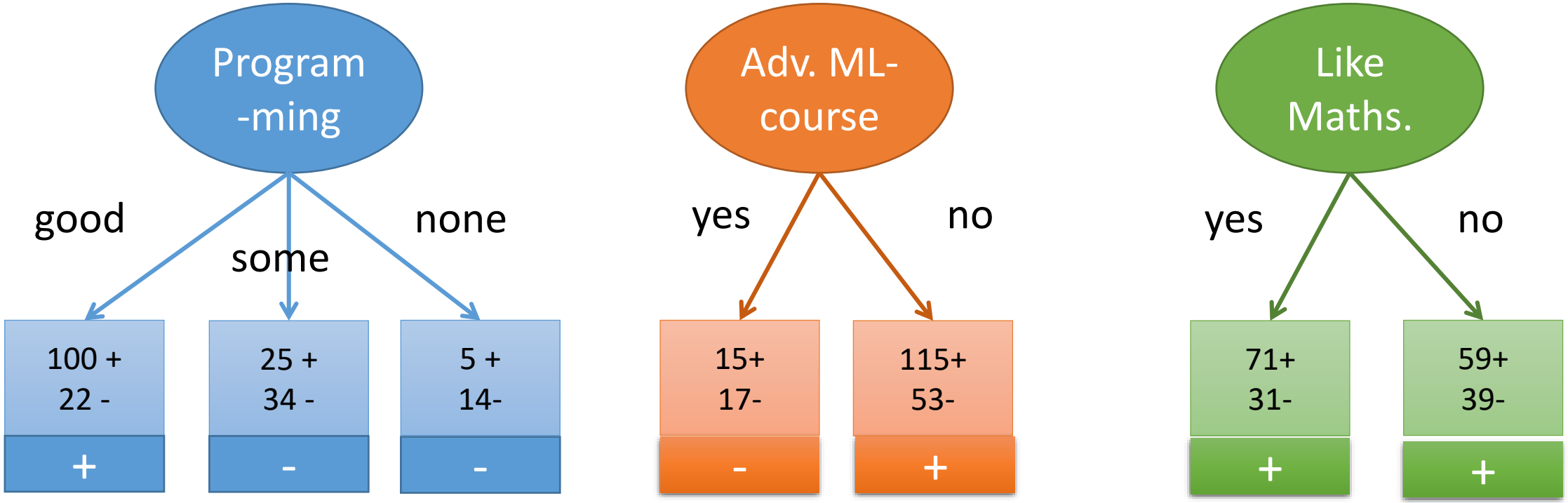


	A	B	C	D	E	F
1	programing	AdvML-course	Like maths	enjoyed	not enjoye	sum
2	good	yes	yes	3	10	13
3	good	yes	no	7	4	11
4	good	no	yes	50	4	54
5	good	no	no	40	4	44
6	some	yes	yes	4	1	5
7	some	yes	no	0	0	0
8	some	no	yes	11	9	20
9	some	no	no	10	24	34
10	none	yes	yes	1	2	3
11	none	yes	no	0	0	0
12	none	no	yes	2	5	7
13	none	no	no	2	7	9
14				130	70	200

We consider the majority class and count how many are correct

	Choose class	correct	incorrect
Good	yes	100	22
Some	no	34	25
None	no	14	5
Sum		148	52

Decision stumps



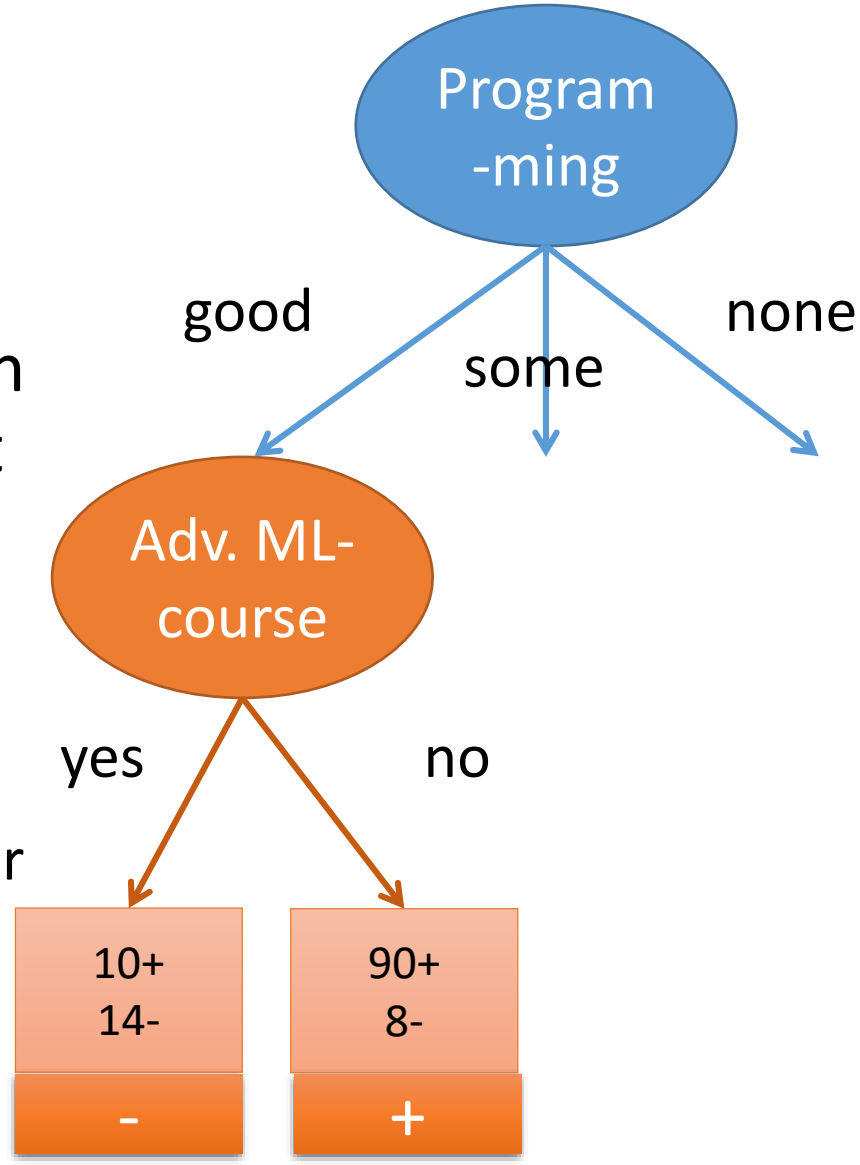
Similarly for the other two stumps

Programming: 148/200=0.74 Adv. ML: 132/200=0.66 Like maths: 0.65

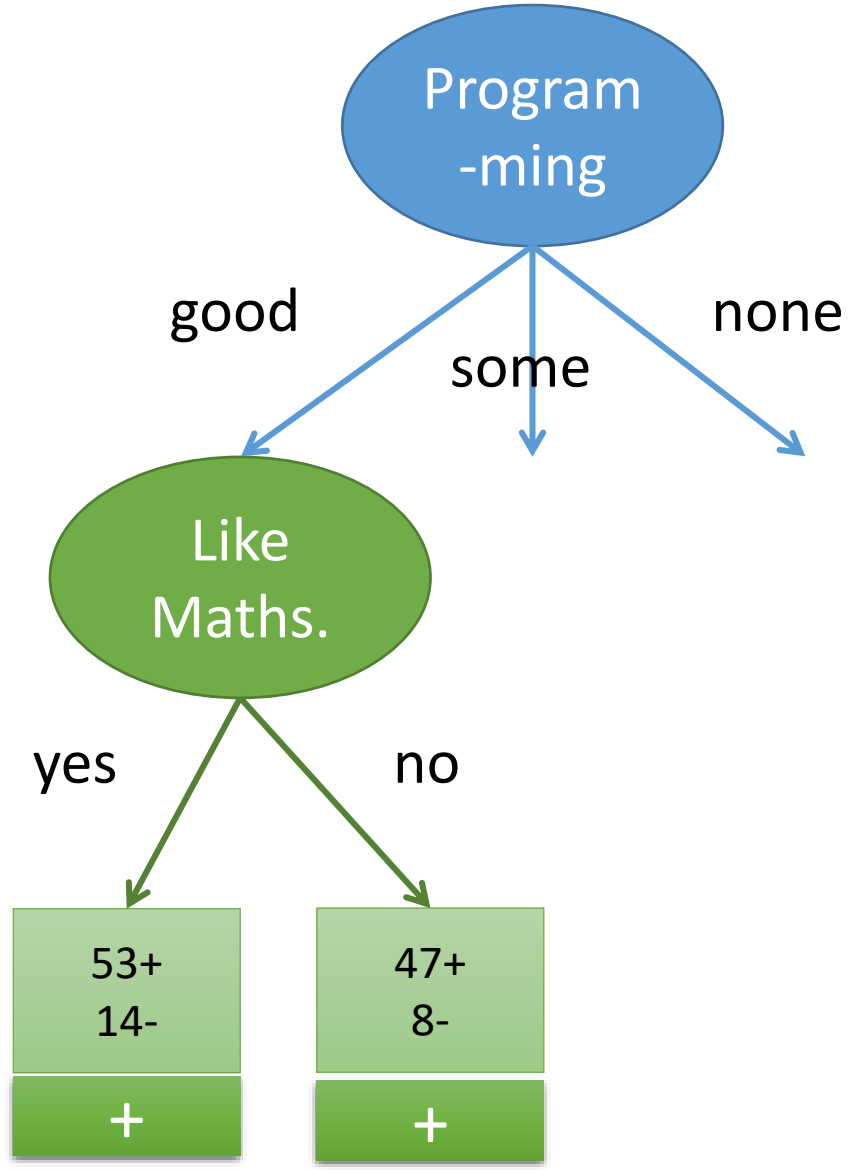
Choose the attribute which is most informative, here: Programming

Next step

- Compare which feature is most informative given the answer to the first feature:
  - Count number of correct
- Her choose <u>Adv ML</u>



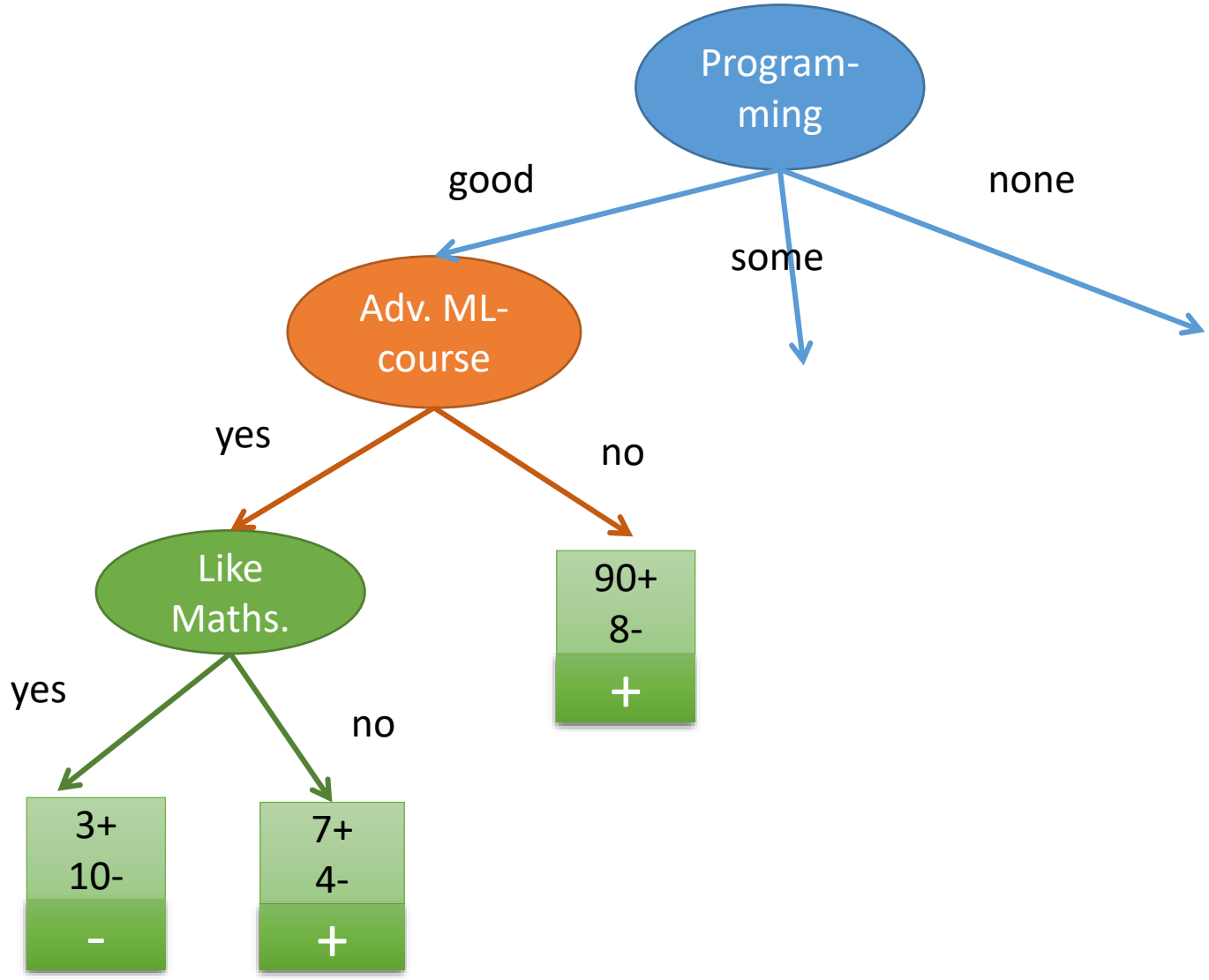
Adv. ML: 104/122



Like maths: 100/122

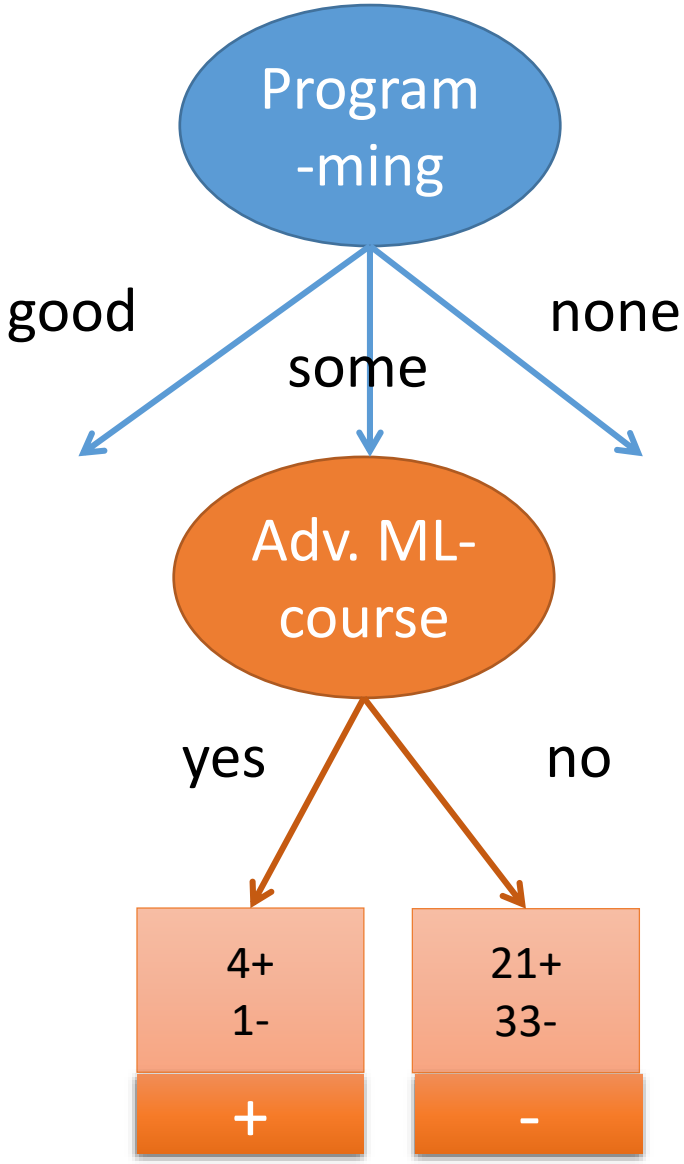
Next step

- We repeat for they who have taken Adv. ML
- For they who have not taken Adv. ML, most of them enjoyed the course whether they like maths. or not

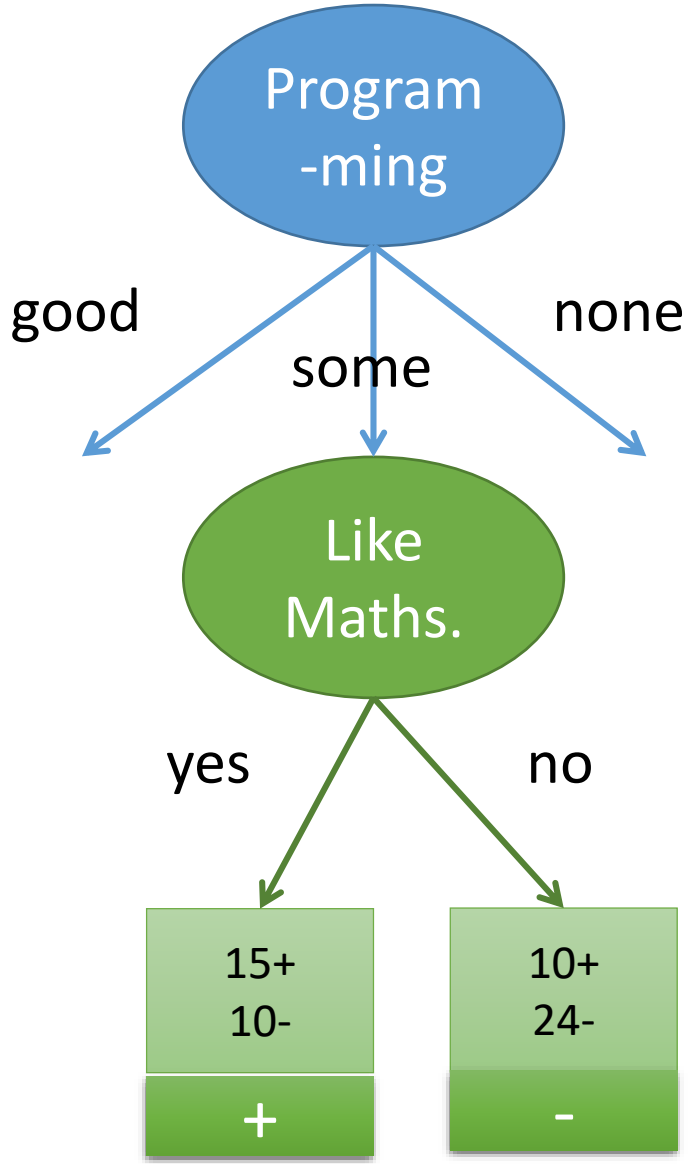


Next step

- We proceed to they with some programming
- What is the next feature?
- Like maths is the best.

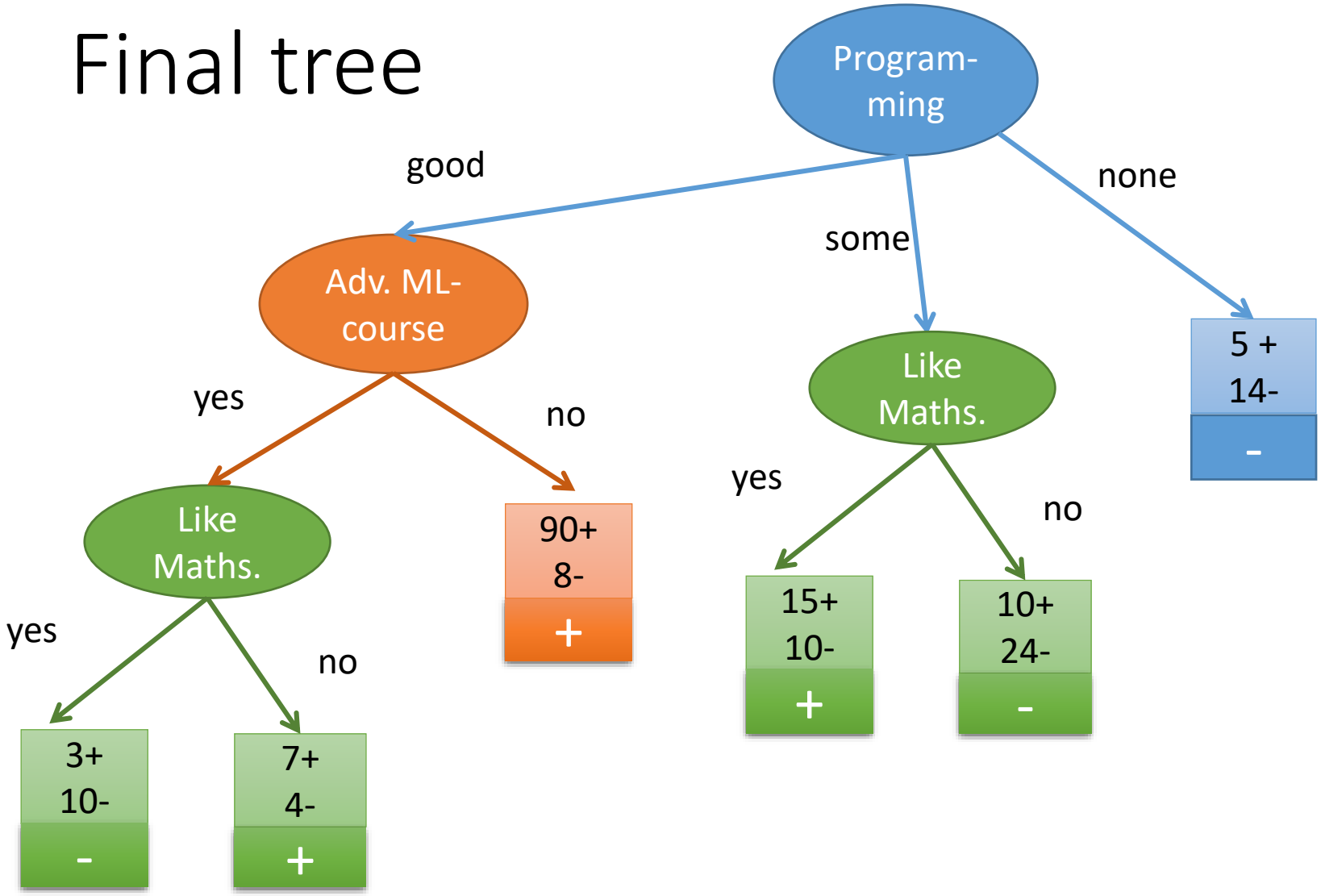


Adv. ML: 37/59



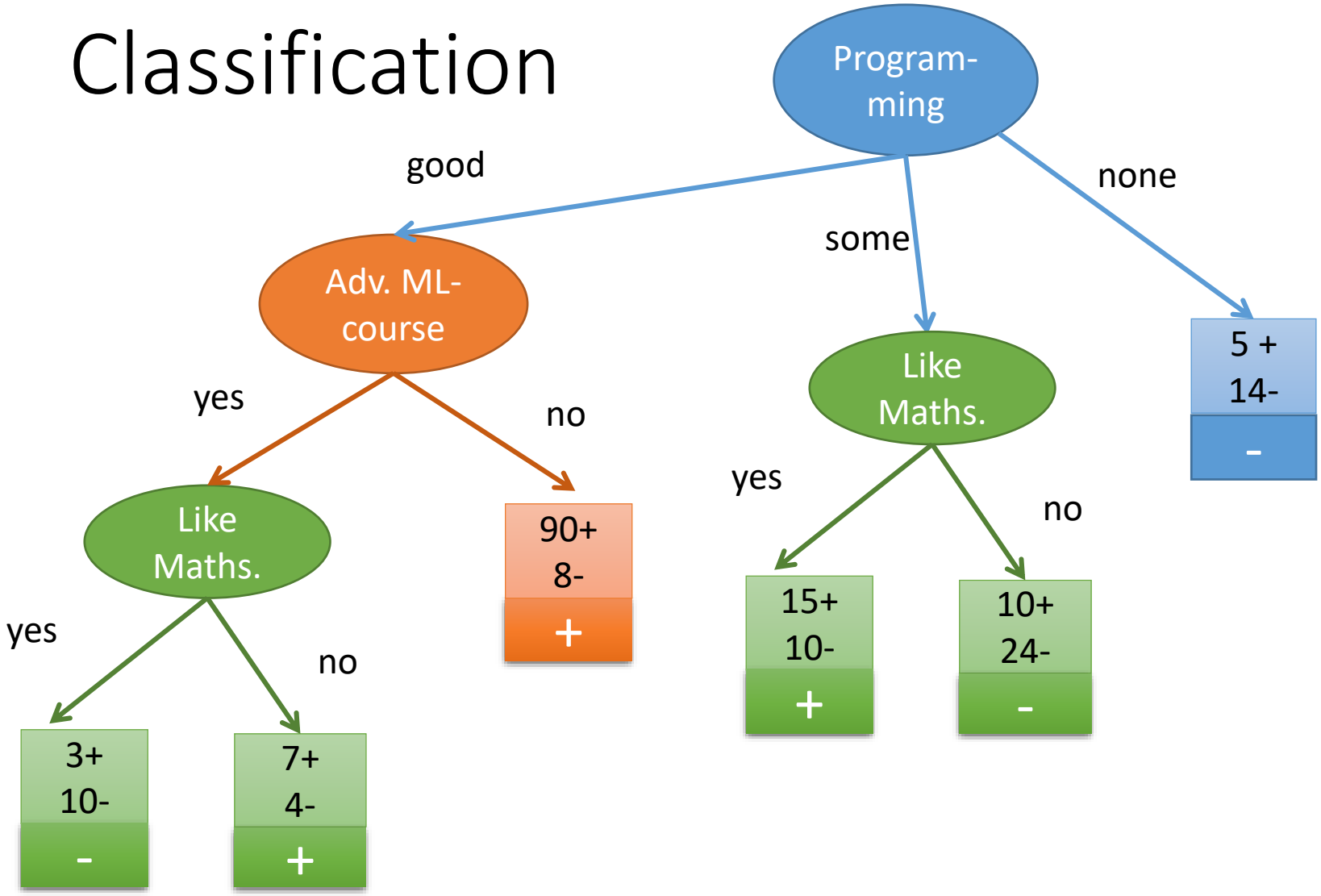
Like maths: 39/59

Final tree



- Stop training when data can't be split further
- Leave nodes may be impure

Classification



- Follow the branch to the leaf
- Choose majority class on the leaf
  - (or return probability distribution of the node)
- Accuracy on the training data (2020):
  - 160/200 = 0.8
  - Up from 0.65
- But results on the test set (2021) may differ

Refinements

- Overfitting:
  - When the model fits the training data very well, but
  - does not generalize well to other sets
- Danger for overfitting with many questions/few instances
- Possible measures:
  - Stop the tree from growing all the way to the leaves, ’’pruning’’
    - Not the details

- In comparing stumps, we simplified using raw counts and fractions
- More advances versions of decision trees use
  - Entropy and information gain (ID3) or
  - Gini impurity
  - (If interested, see Marsland ch. 10)

Properties

Decision trees

- are good at explaining the decisions compared to other ML algorithms.
- are good at categorical data but can also handle numerical data.
- are not the best to handle many features.

Random forests is a generalization of DTs

Some concepts you will meet again

- Supervised learning
  - Classification
- Training set vs. test set
- Accuracy of a classifier
- Baseline classifier
  - Majority class classifier
- Overfitting