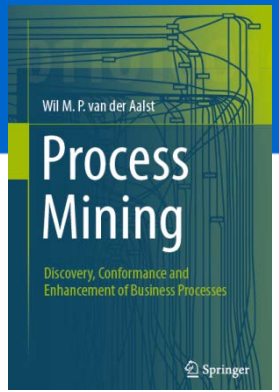
Process Mining: Data Science in Action

Learning Decision Trees

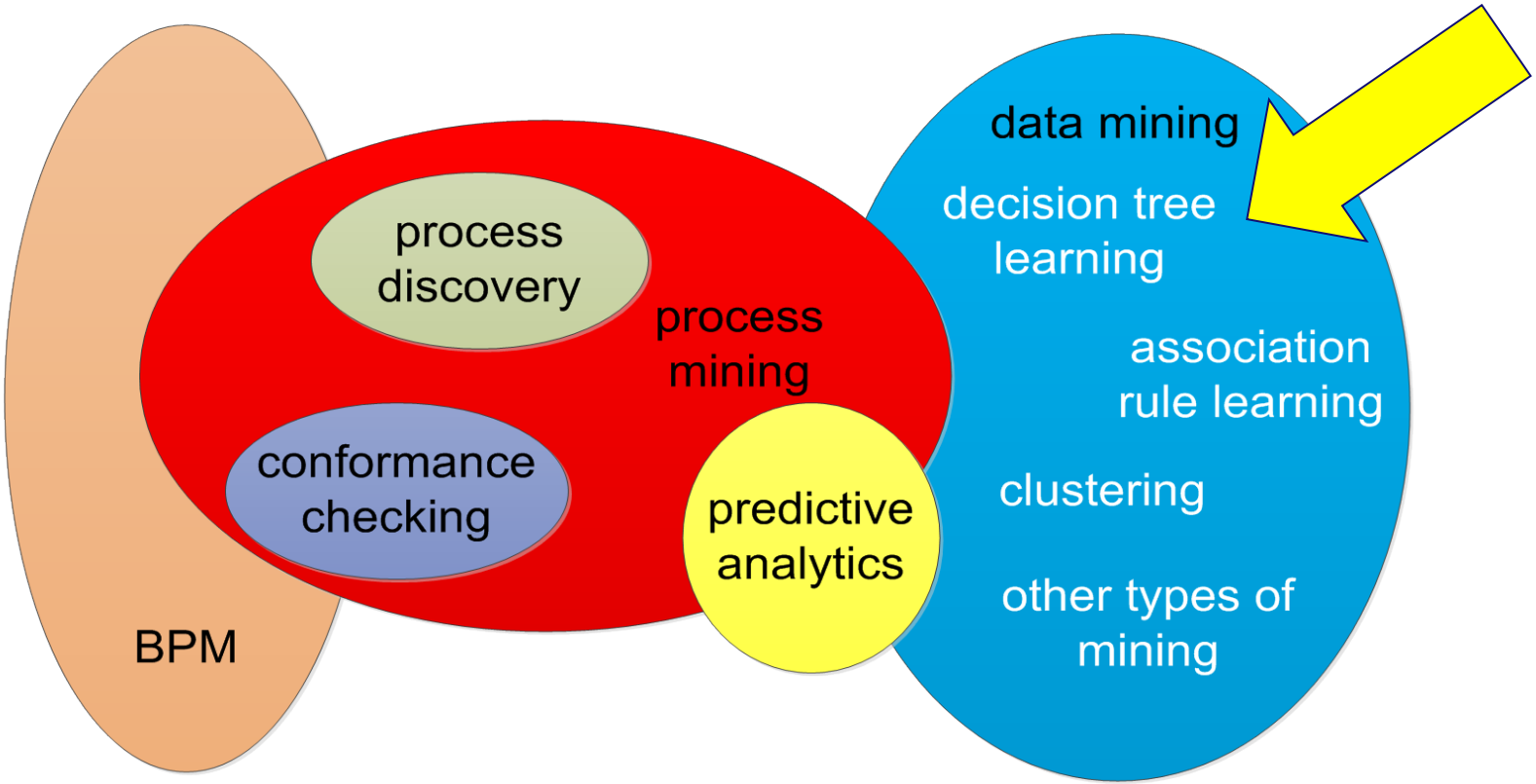


prof.dr.ir. Wil van der Aalst
www.processmining.org

Technische Universiteit Eindhoven University of Technology

Where innovation starts

Positioning decision tree learning

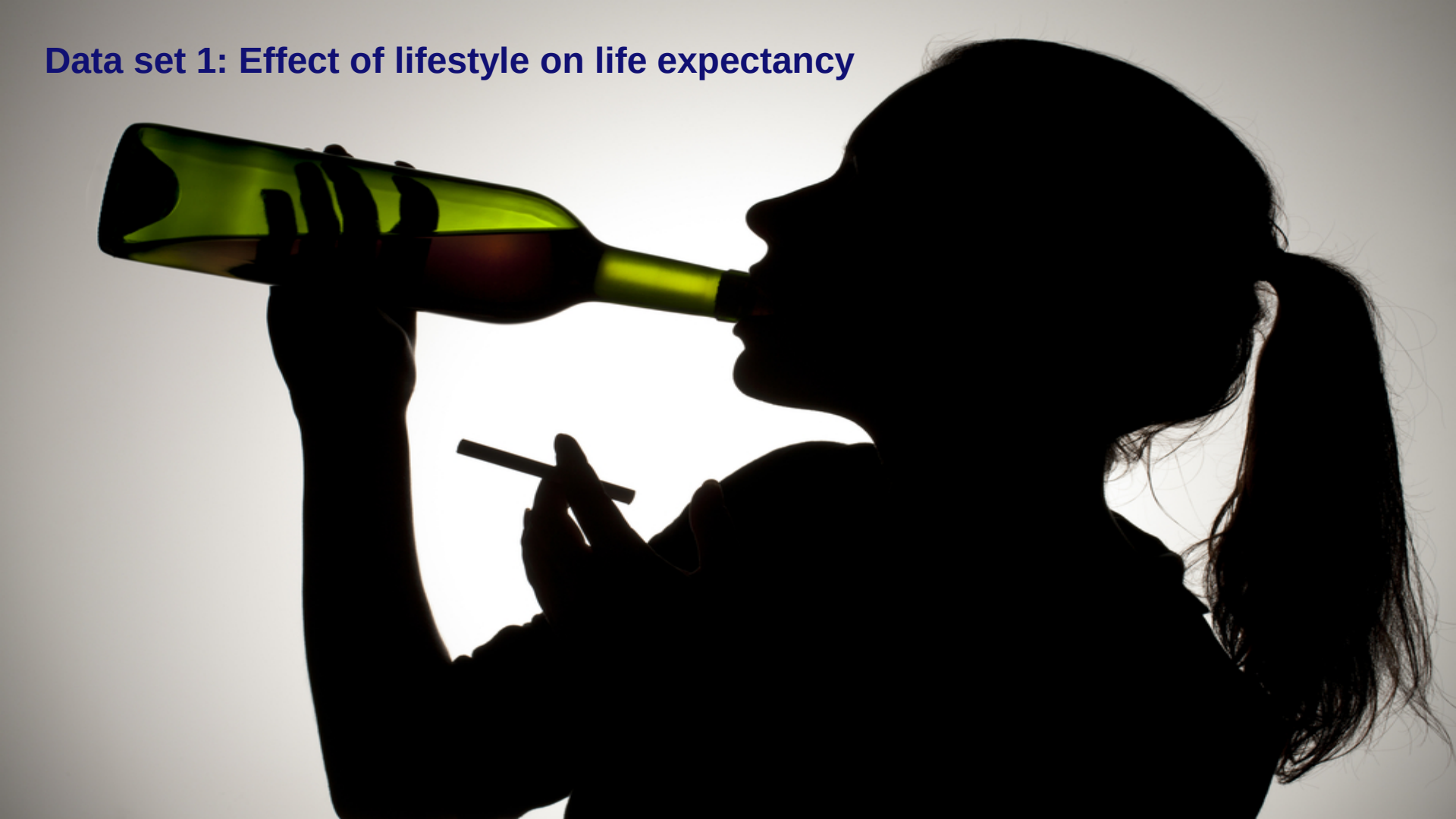


TU/e

Decision tree learning

supervised learning

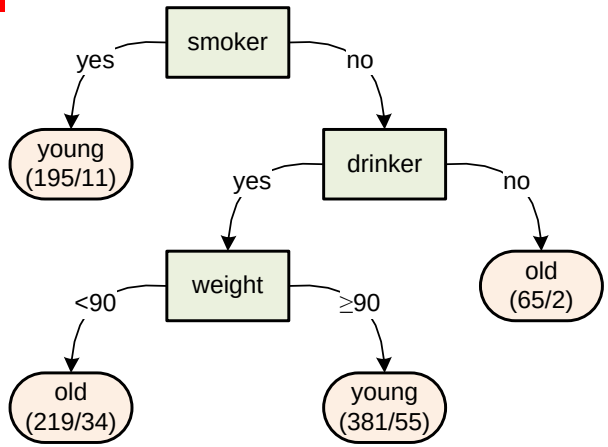
Data set 1: Effect of lifestyle on life expectancy



drinker	smoker	weight	age
yes	yes	120	44 young
no	no	70	96 old
yes	no	72	88 old
yes	yes	55	52 young
no	yes	94	56 young
no	no	62	93 old
	...	...	

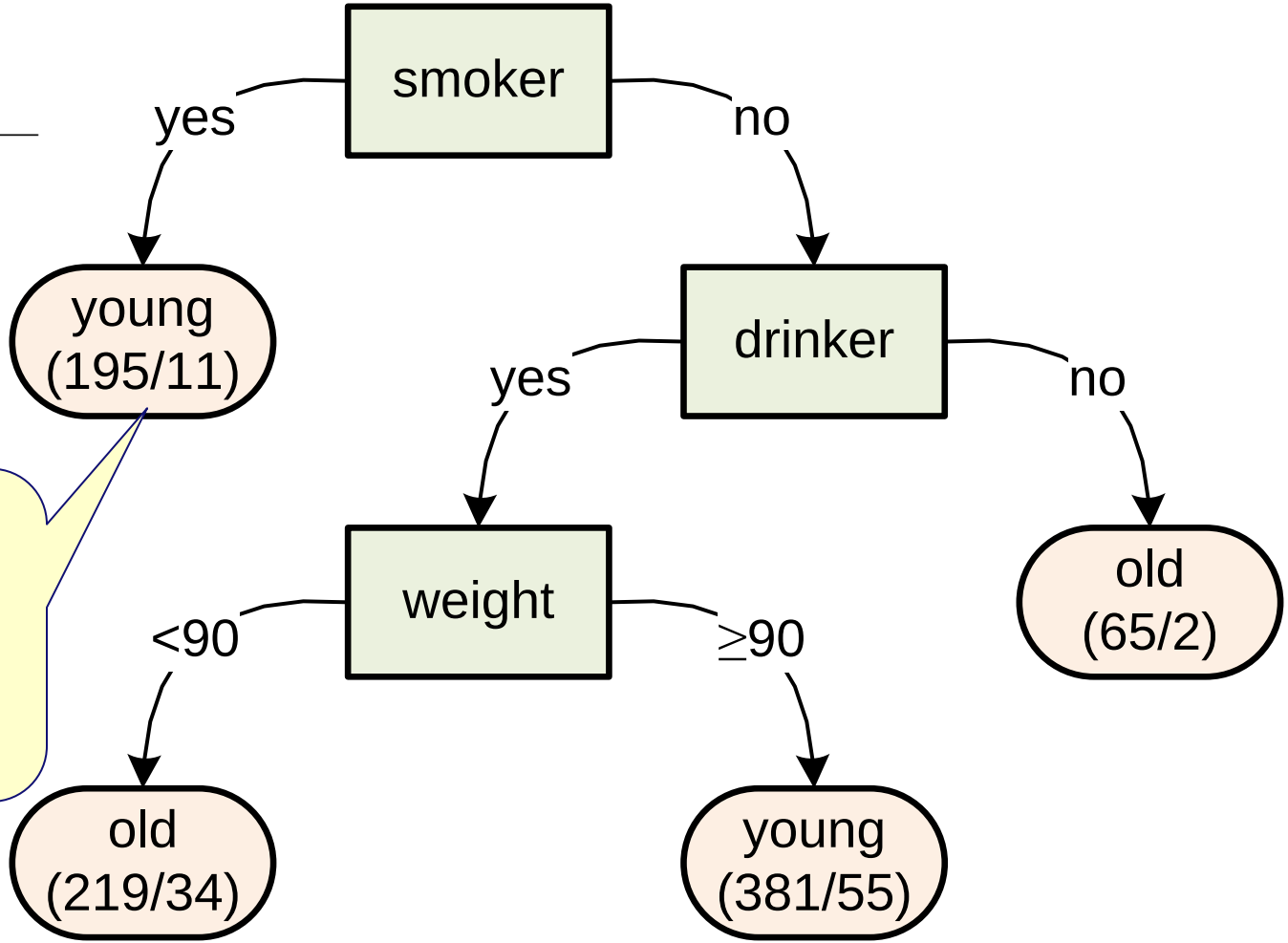
predictor variables

response variable



⩾70 = old
<70 = young

drinker	smoker	weight	age
yes	yes	120	44
no	no	70	96
yes	no	72	88
yes	yes	55	52
no	yes	94	56
no	no	62	93
…	…	…	…

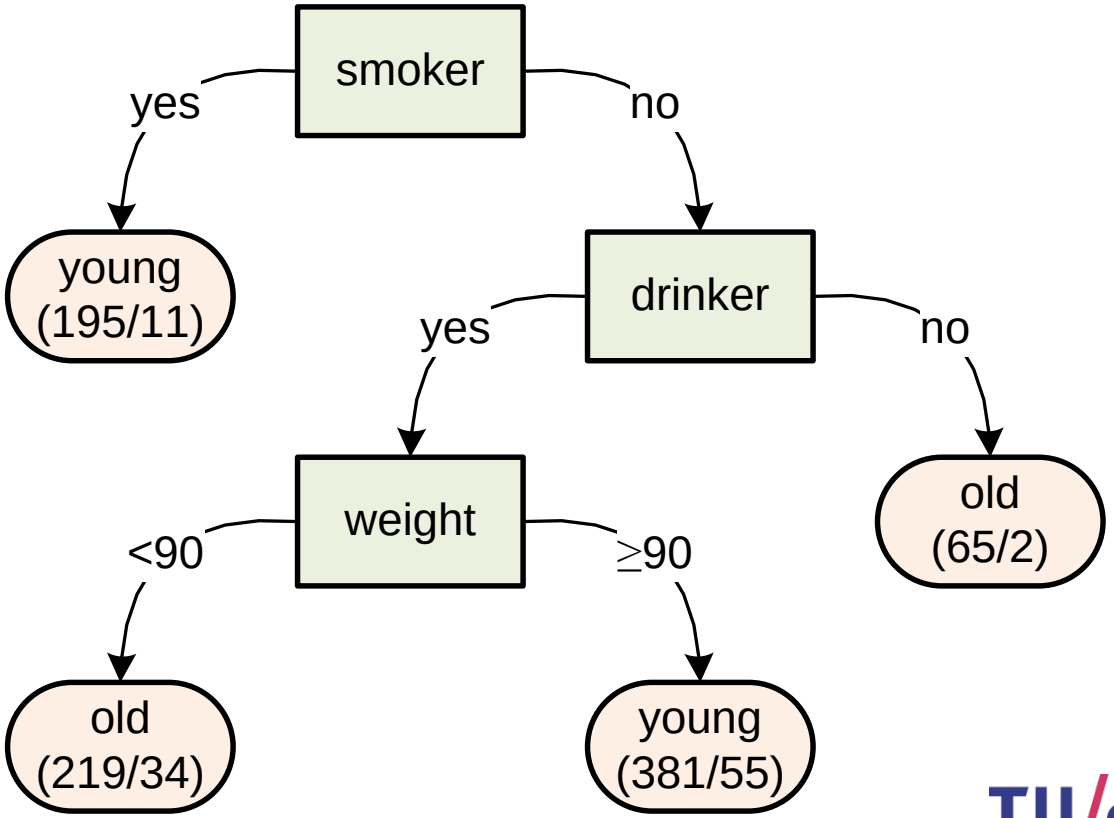


all 195 smokers were classified as "young" but 11 were classified incorrectly

⩾70 = old
<70 = young

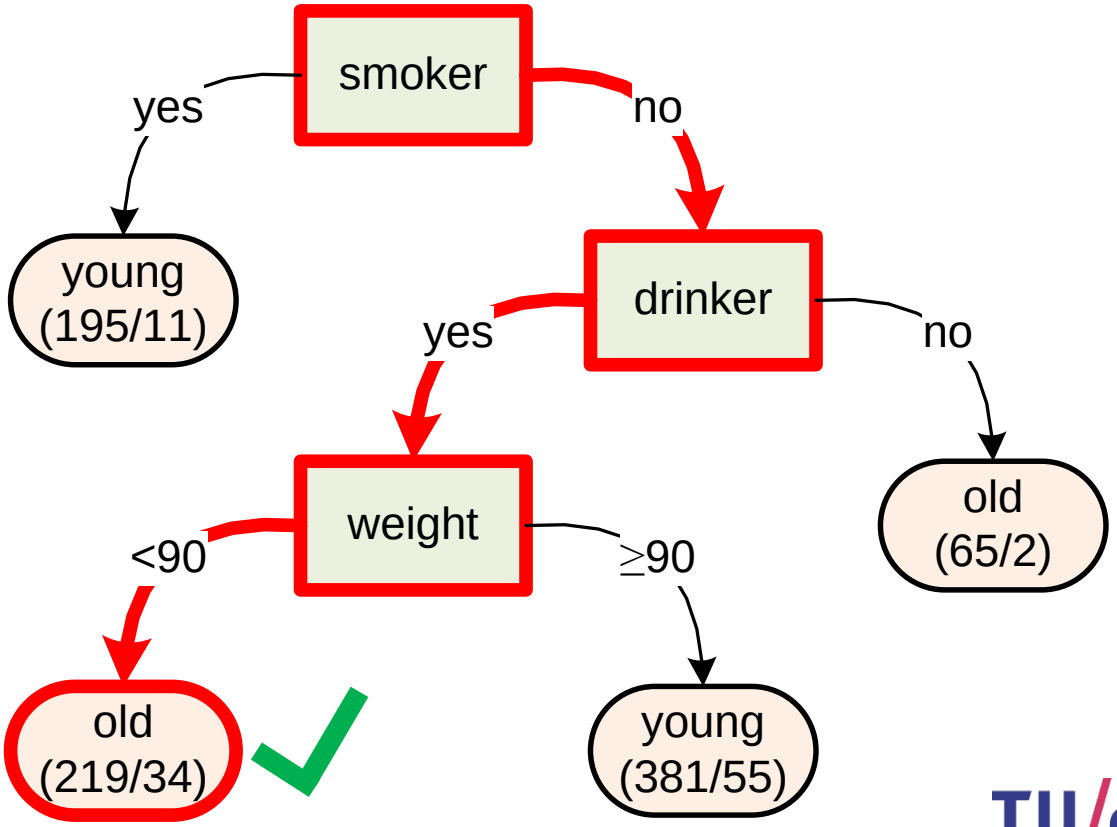
Question: Correctly classified?

Mary Jones	
drinker	yes
smoker	no
weight	70 kg
age	85 year



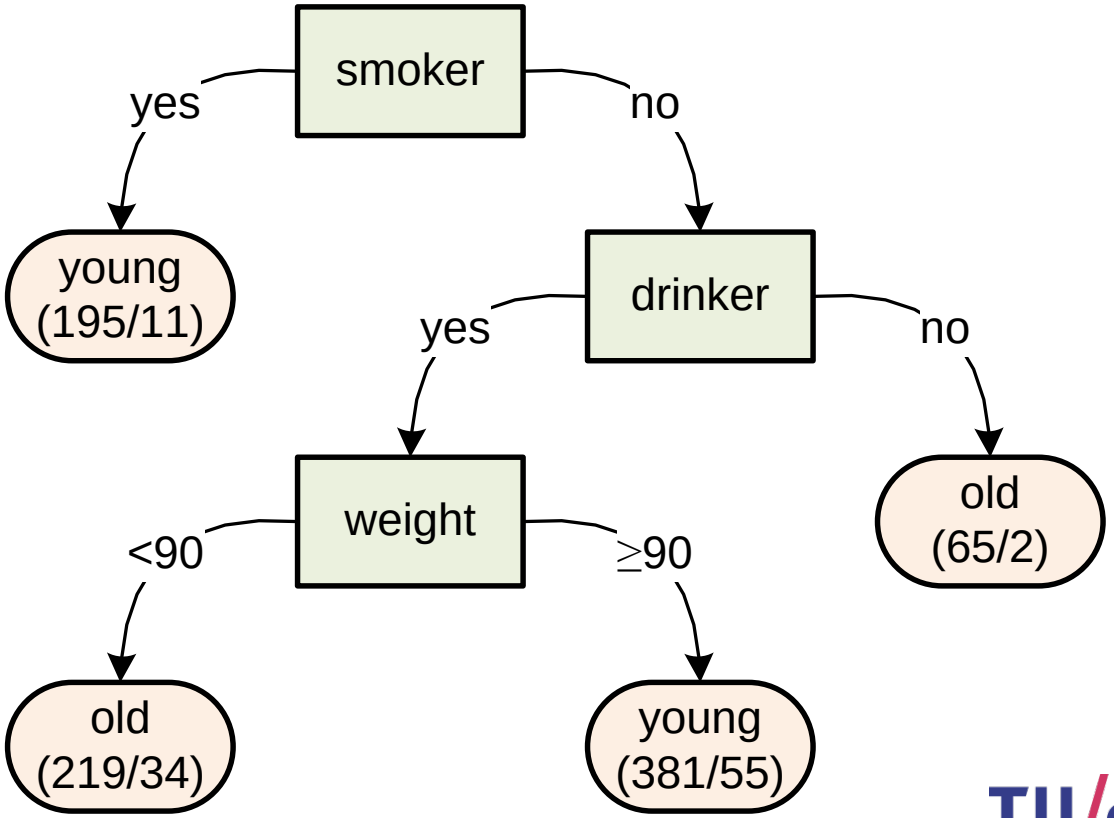
Answer: Yes

Mary Jones	
drinker	yes
smoker	no
weight	70 kg
age	85 year



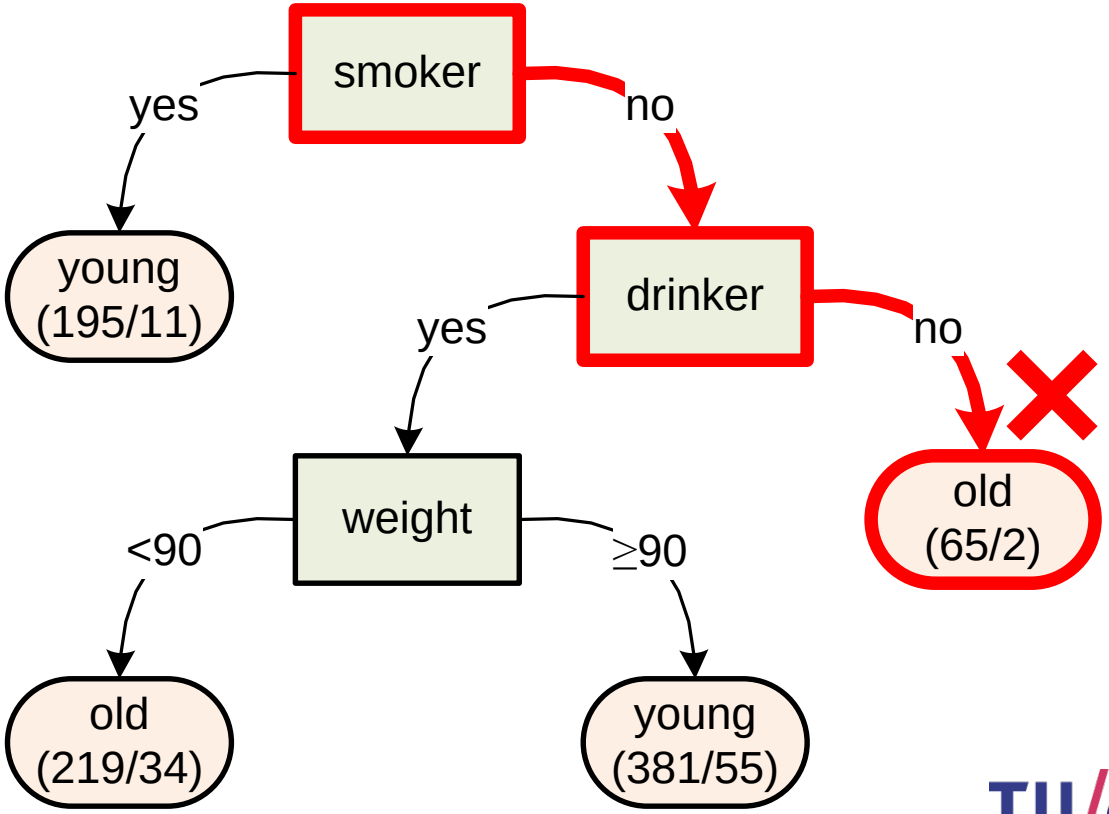
Question: Correctly classified?

Sue Smith	
drinker	no
smoker	no
weight	60 kg
age	35 year



Answer: No

Sue Smith	
drinker	no
smoker	no
weight	60 kg
age	35 year



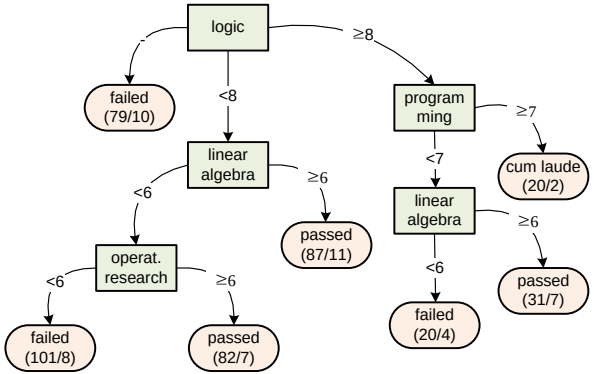
TU/e

Data set 2: Effect of individual course results on graduation

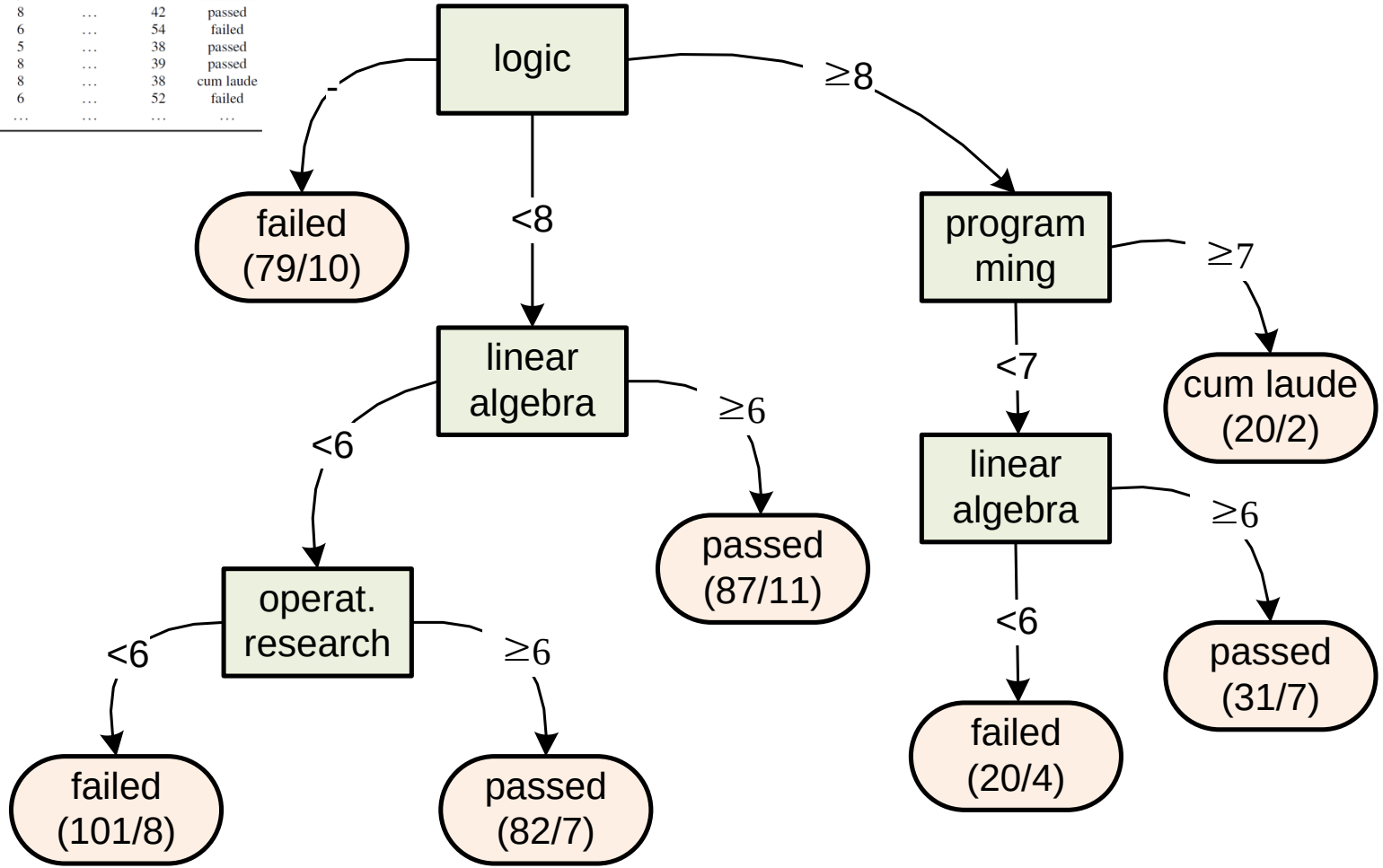
linear algebra	logic	program-ming	operations research	workflow systems	...	duration	result
9	8	8	9	9	..	36	cum laude
7	6	-	8	8	..	42	passed
-	-	5	4	6	..	54	failed
8	6	6	6	5	..	38	passed
6	7	6	-	8	..	39	passed
9	9	9	9	8	..	38	cum laude
5	5	-	6	6	..	52	failed
	..	...	...	...	...	...	...

predictor variables

response variable

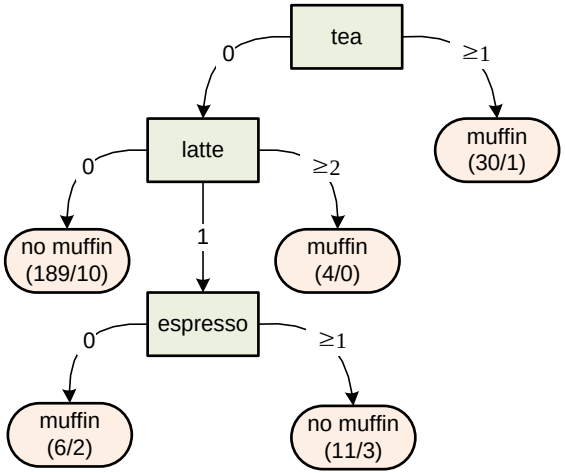


linear algebra	logic	program-ming	operations research	workflow systems	…	duration	result
9	8	8	9	9	…	36	cum laude
7	6	-	8	8	…	42	passed
-	-	5	4	6	…	54	failed
8	6	6	6	5	…	38	passed
6	7	6	-	8	…	39	passed
9	9	9	9	8	…	38	cum laude
5	5	-	6	6	…	52	failed
…	…	…	…	…	…	…	…



Data set 3: Muffin or no muffin

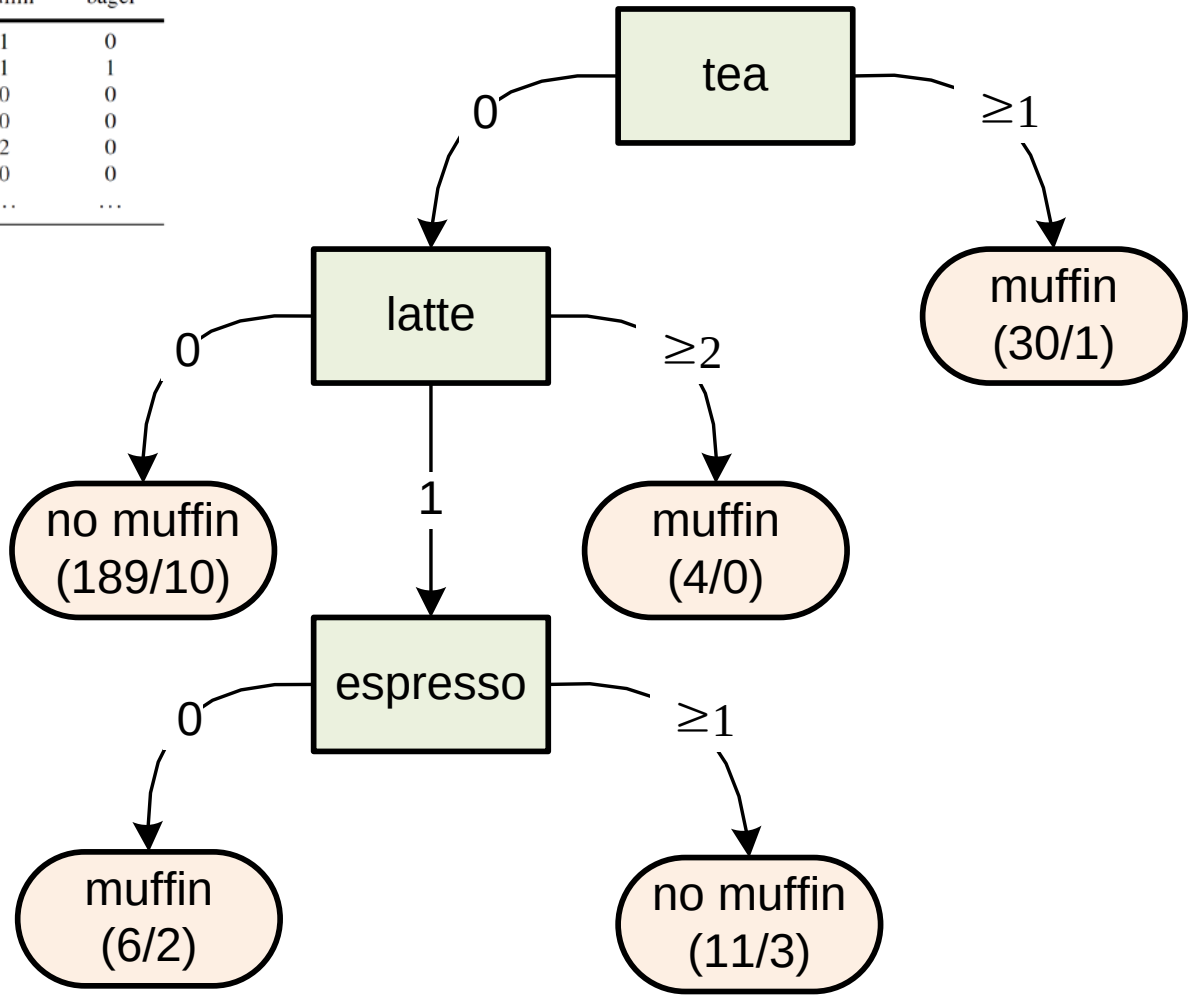
cappuccino	latte	espresso	americano	ristretto	tea	muffin	bagel
1	0	0	0	0	0	1	0
0	2	0	0	0	0	1	1
0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	1	2	0
0	0	0	1	1	0	0	0
...	...	...	...	...	...	...	...



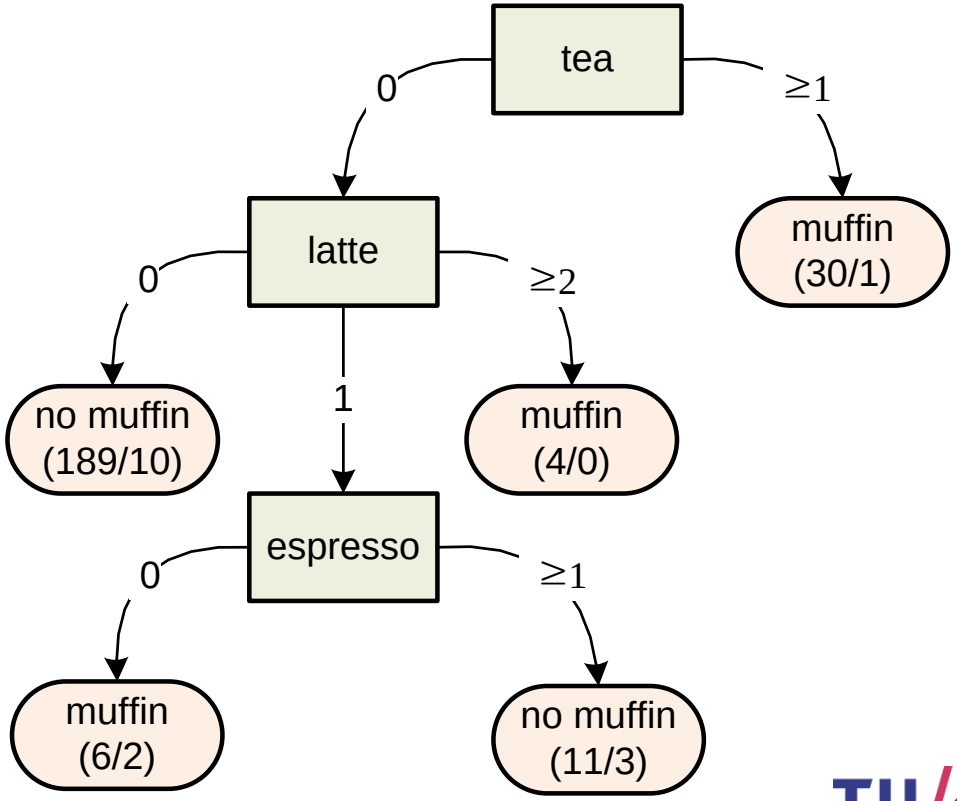
response variable

⩾1 = "muffin"
0 = "no muffin"

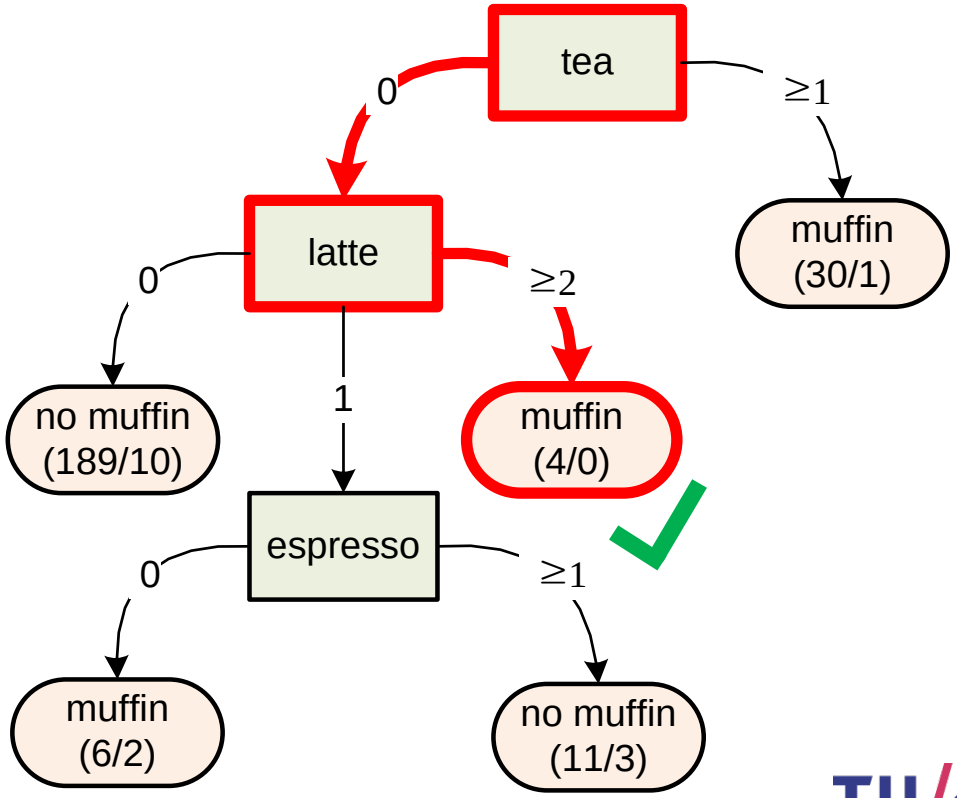
cappuccino	latte	espresso	americano	ristretto	tea	muffin	bagel
1	0	0	0	0	0	1	0
0	2	0	0	0	0	1	1
0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	1	2	0
0	0	0	1	1	0	0	0
…	…	…	…	…	…	…	…



Question: Correctly classified?

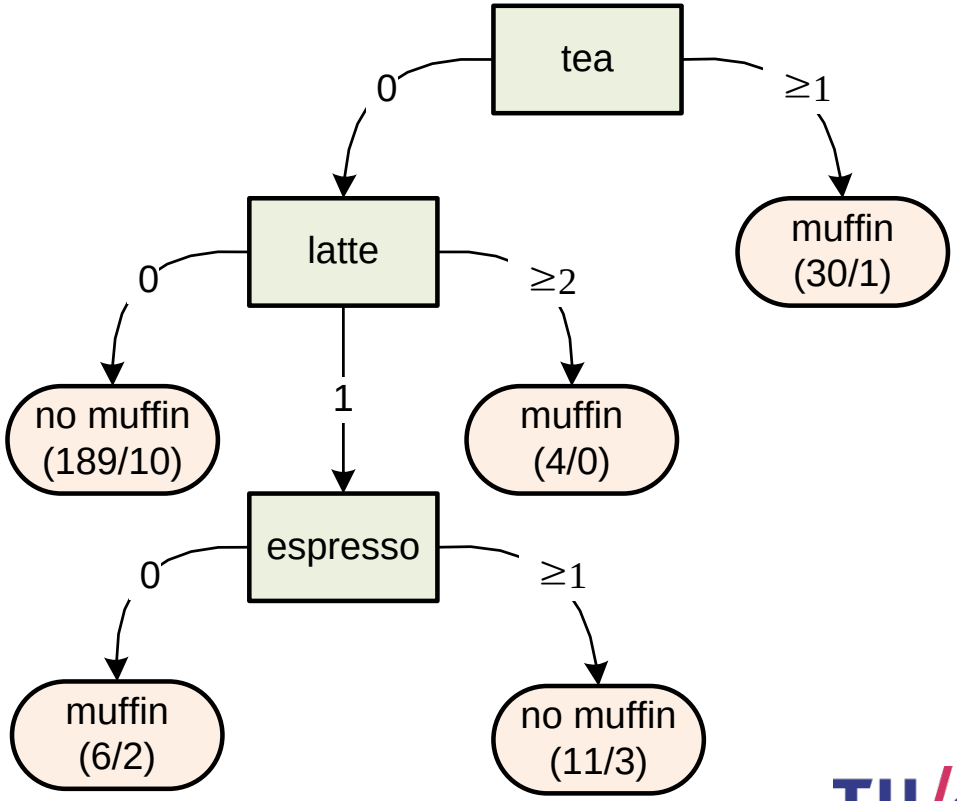


Answer: Yes

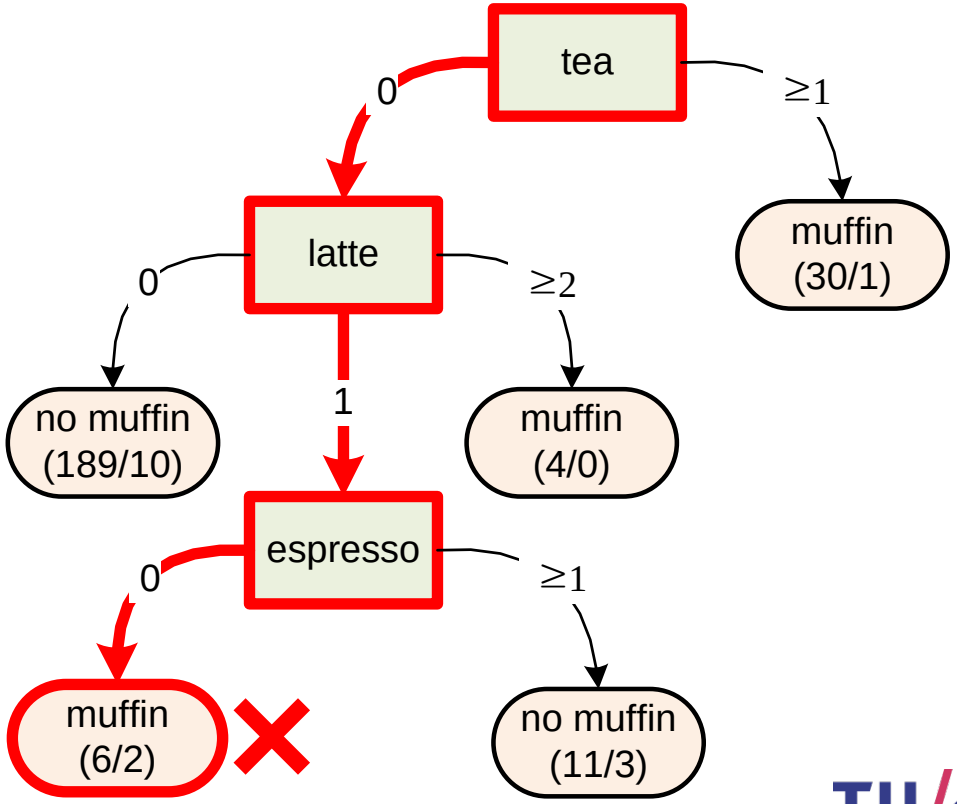


TU/e

Question: Correctly classified?

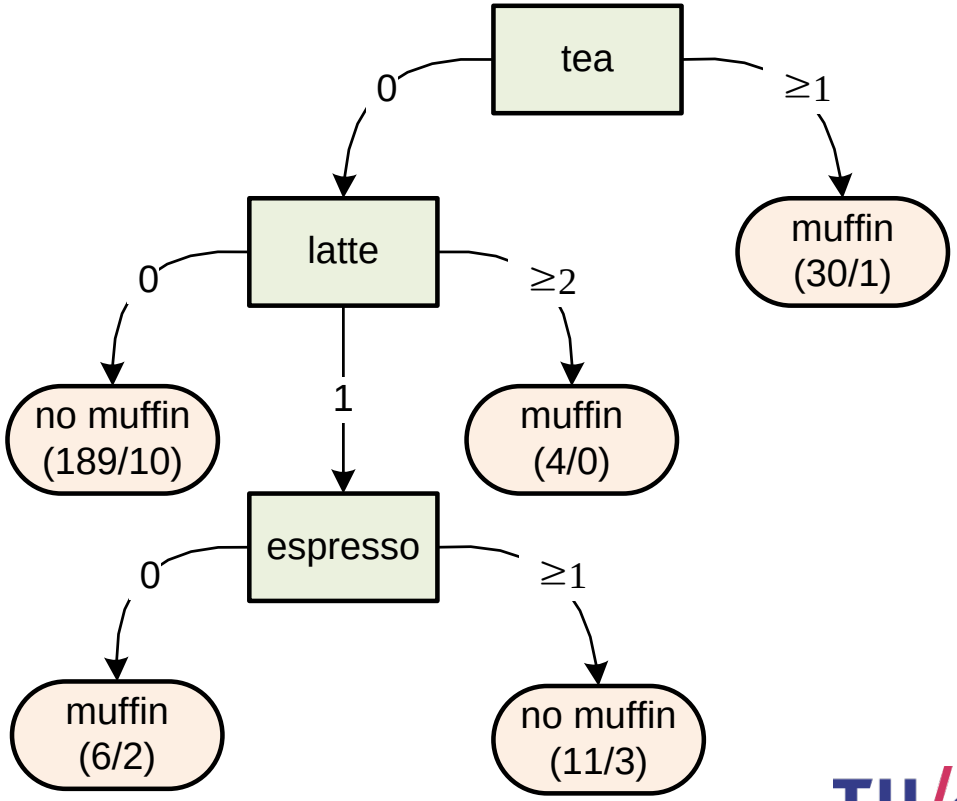


Answer: No



TU/e

Question: Did this person eat a muffin?



TU/e

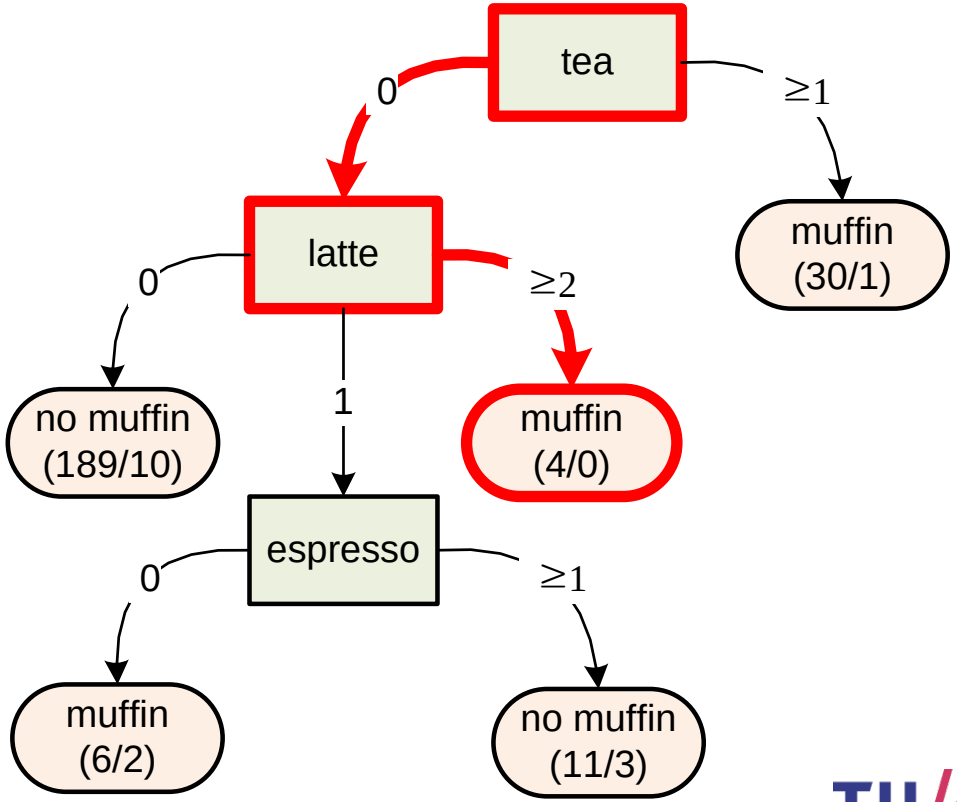
Answer: Yes!


GUEST CHECK

DATE	SERVER	TABLE	QUESTS	CHECK NUMBER
				689561

bagel	1
latte	2
capp.	1
muffin	≥1
TAX	
TOTAL	FREE

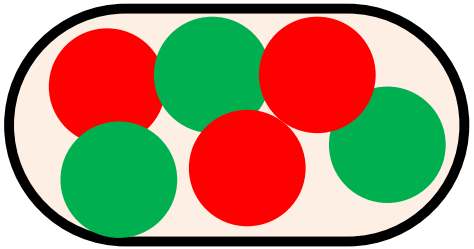
Thank You - Please Come Again



TU/e

How does it work? - Basic idea

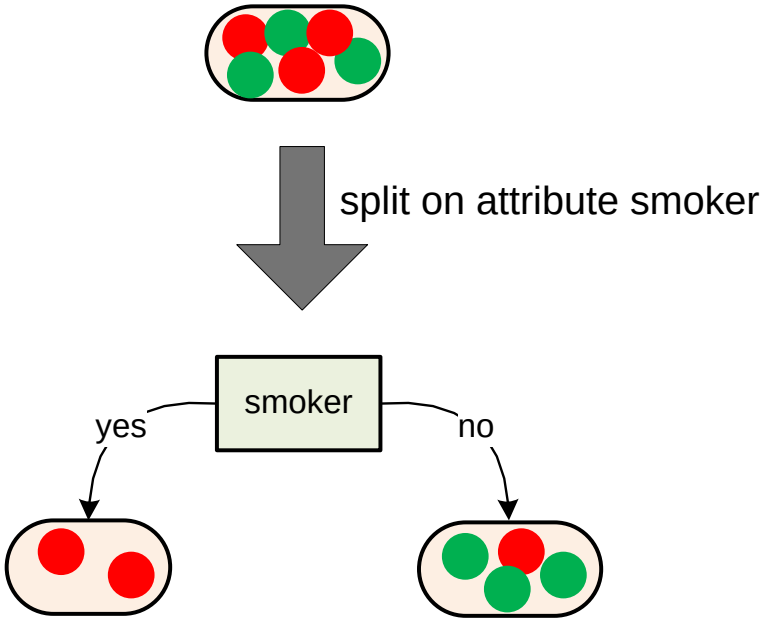
- **Split the set of instances in subsets such that the variation within each subset becomes smaller.**



high entropy

How does it work? - Basic idea

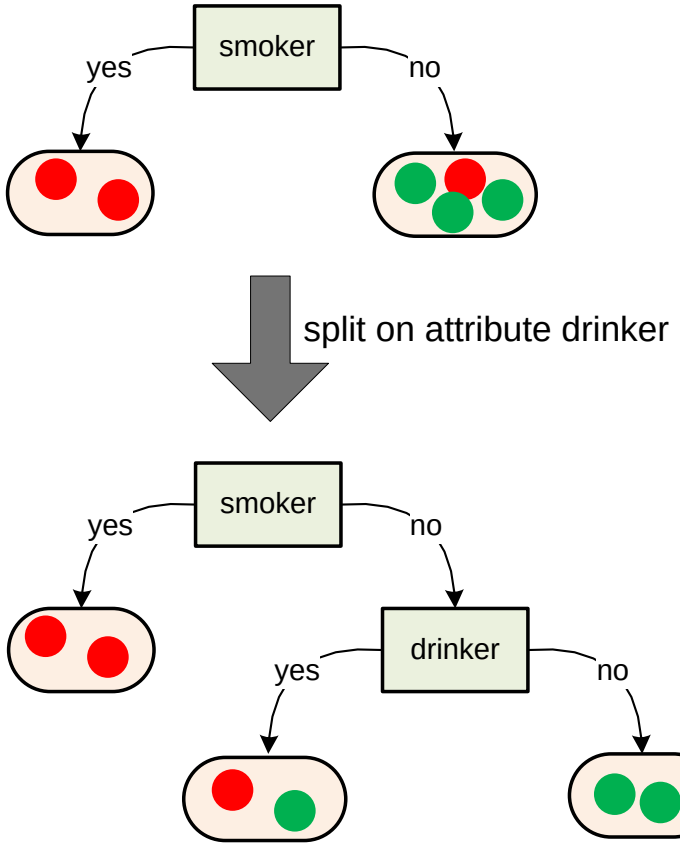
- **Split the set of instances in subsets such that the variation within each subset becomes smaller.**



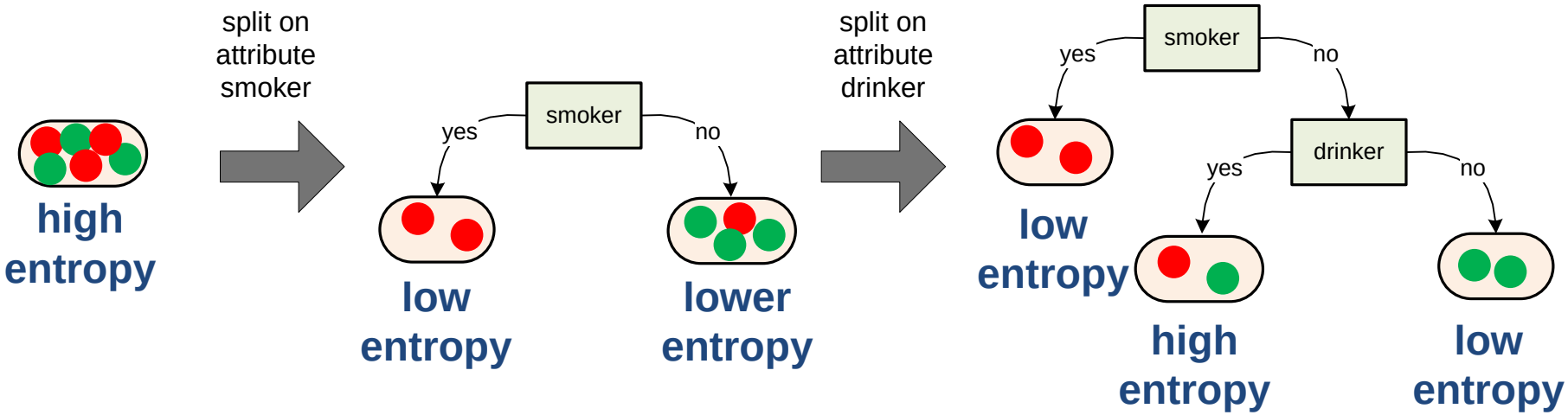
TU/e

How does it work? - Basic idea

- **Split the set of instances in subsets such that the variation within each subset becomes smaller.**



Decreasing entropy



High entropy

- Degree of uncertainty.
- Inverse of "compressibility" ("zippability") .
- Goal: reduce entropy in leaves of tree to improve predictability.

TU/e

Intermezzo: Logarithms

(needed for computing entropy)

$$\log_2(x) = y \Leftrightarrow 2^y = x$$

$$\log_2(2^n) = n$$

$$\log_2(\frac{1}{2^n}) = -n$$

$$\log_2(1) = 0$$

$$\log_2(2) = 1$$

$$\log_2(8) = 3$$

$$\log_2(0.125) = -3$$

$$\log_2(1024) = 10$$

$$\log_2(0.75) = -0.415$$

TU/e

Definition entropy

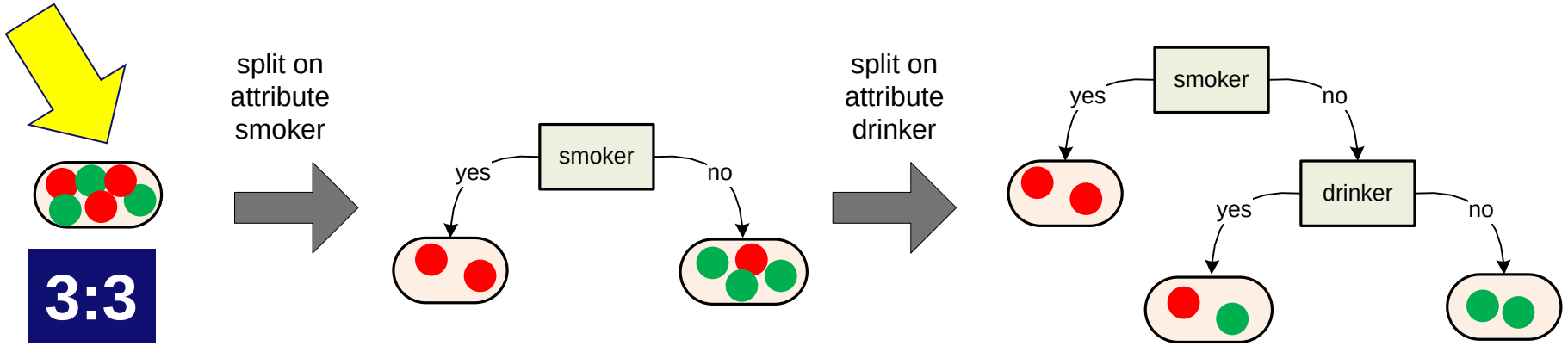
$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

k possible values enumerated $1, 2, \ldots, k$

$p_i = \frac{c_i}{n}$ is the fraction of elements having value i

with $c_i \geq 1$ the number of i values and $n = \sum_{i=1}^{k} c_i$

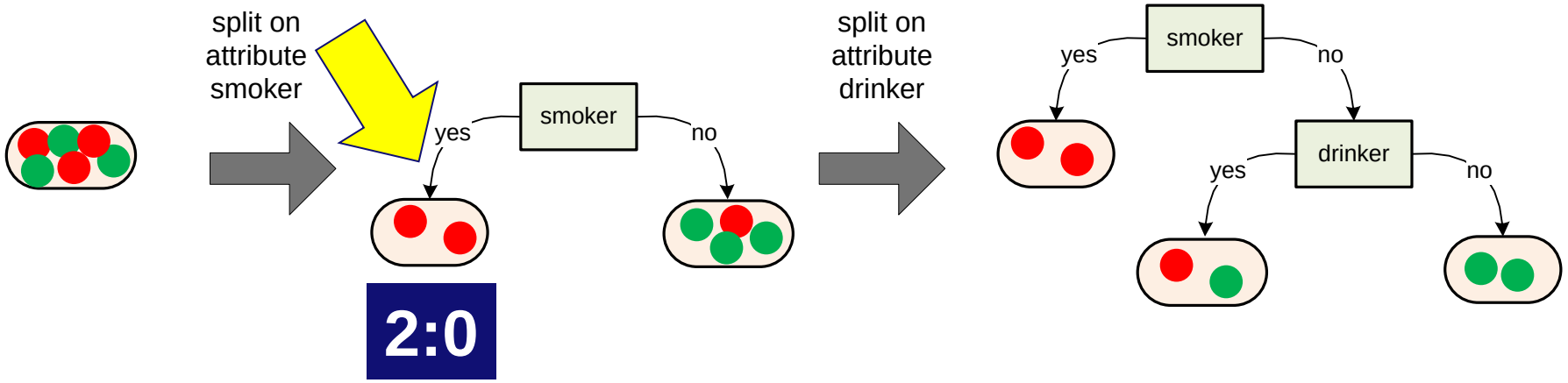
Example $E = 1$ (three red, three green)



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i) = -(\frac{3}{6}\log_2(\frac{3}{6}) + \frac{3}{6}\log_2(\frac{3}{6})) = -(\frac{1}{2} \times -1 + \frac{1}{2} \times -1) = 1$$

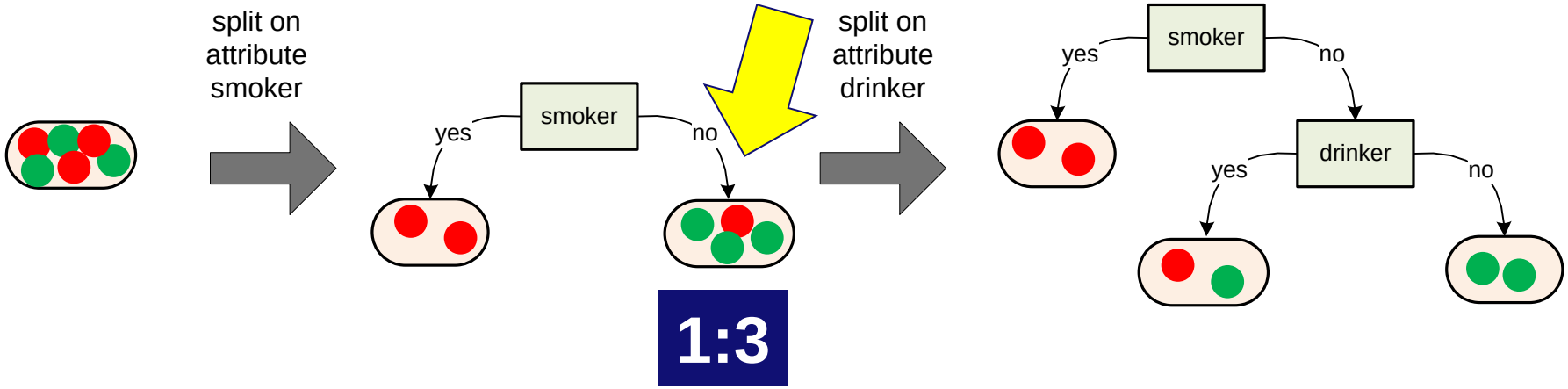
TU/e

Example $E = 0$ (two red, no green)



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i) = -(\frac{2}{2}\log_2(\frac{2}{2})) = -(1 \times 0) = 0$$

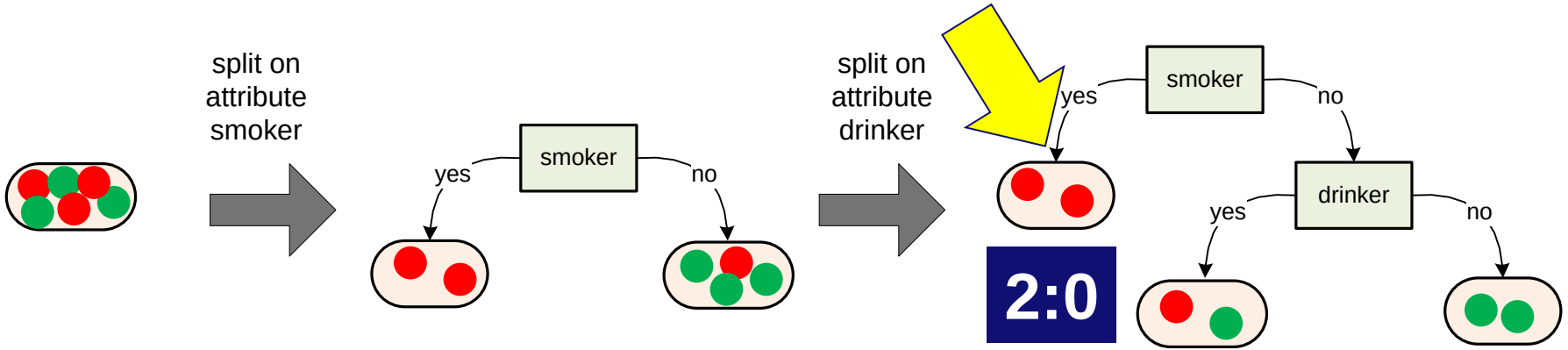
Example $E = 0.811$ (one red, three green)



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i) = -(\frac{1}{4}\log_2(\frac{1}{4})+\frac{3}{4}\log_2(\frac{3}{4})) = -(\frac{1}{4}\times -2+\frac{3}{4}\times -0.415) = 0.811$$

TU/e

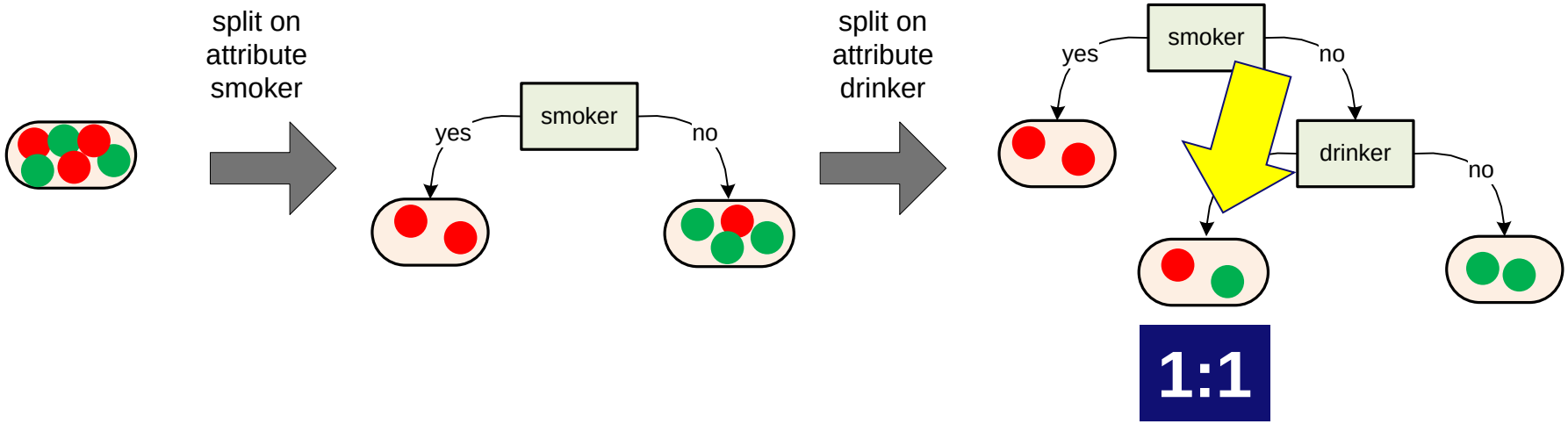
Example $E = 0$ (two red, no green)



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i) = -(\frac{2}{2}\log_2(\frac{2}{2})) = -(1 \times 0) = 0$$

TU/e

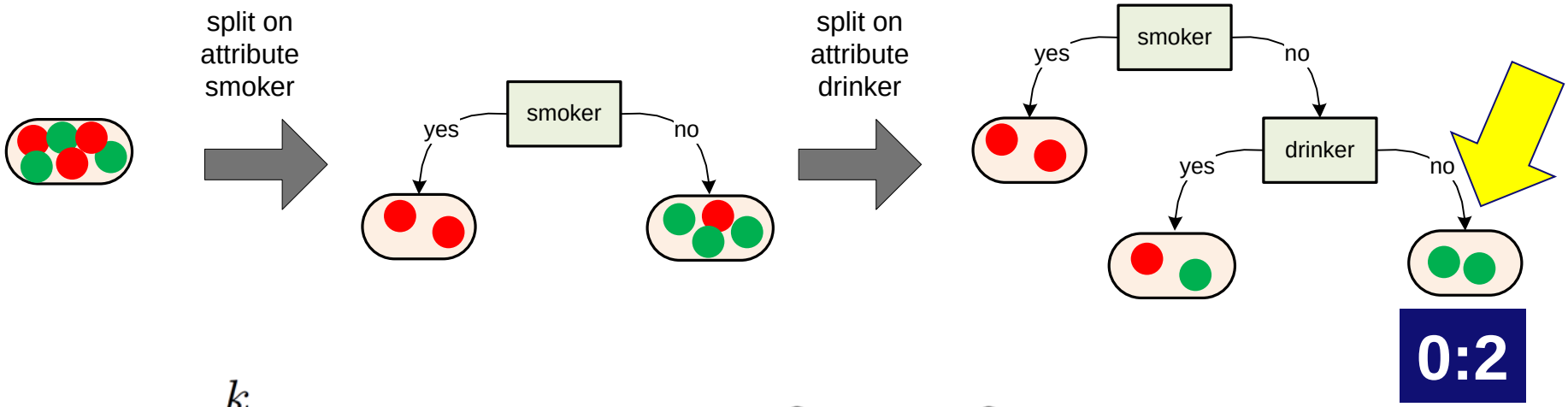
Example $E = 1$ (one red, one green)



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i) = -(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) = -(\frac{1}{2} \times -1 + \frac{1}{2} \times -1) = 1$$

TU/e

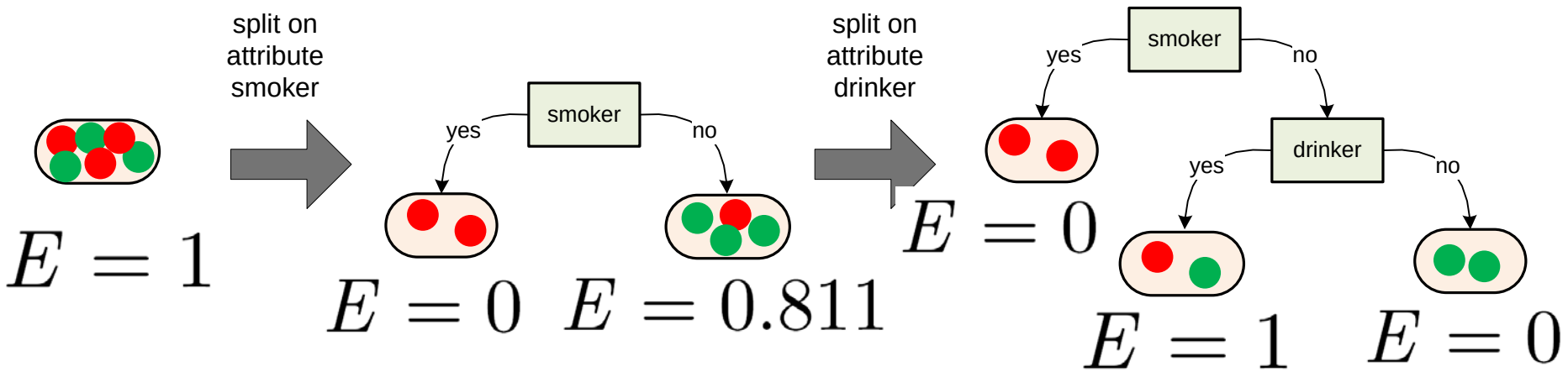
Example $E = 0$ (two red, no green)



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i) = -(\frac{2}{2}\log_2(\frac{2}{2})) = -(1 \times 0) = 0$$

TU/e

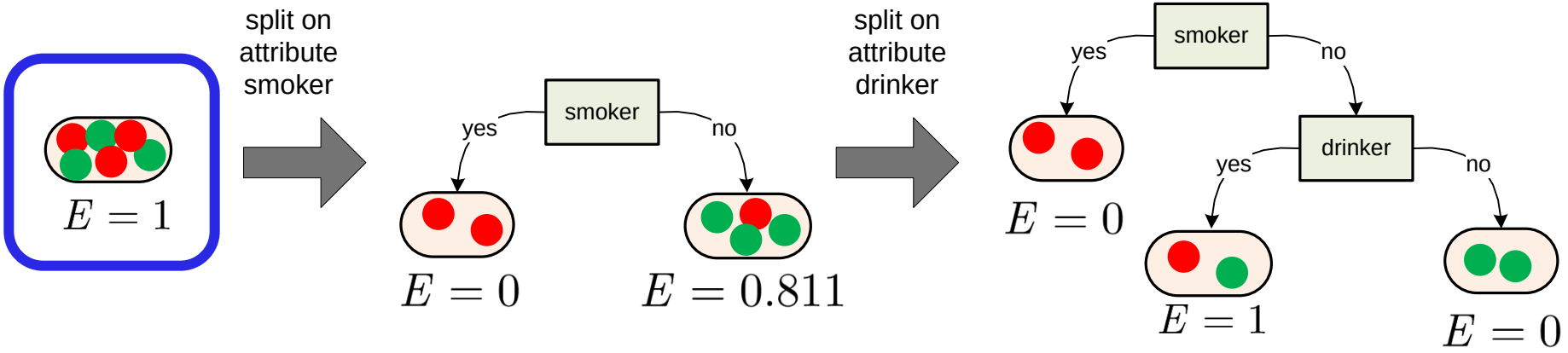
Entropy values



$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

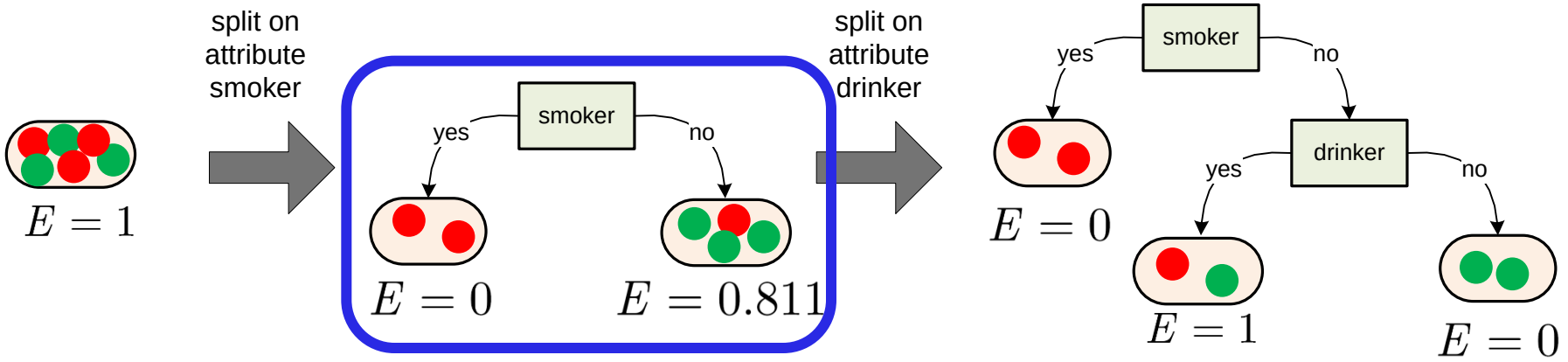
TU/e

Weighted average



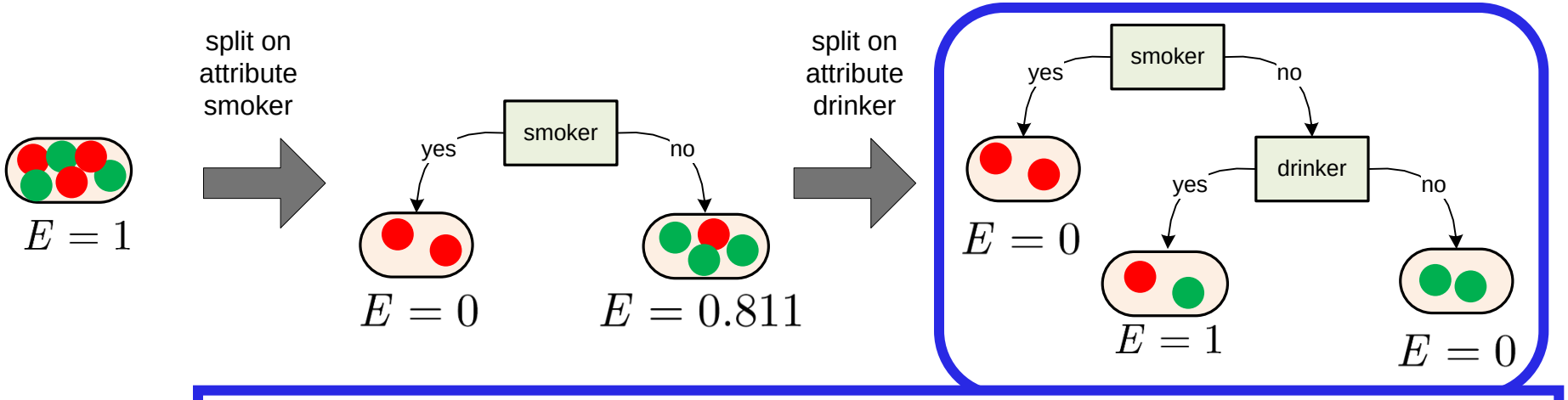
$$E = \frac{6}{6} \times 1 = 1$$

Weighted average



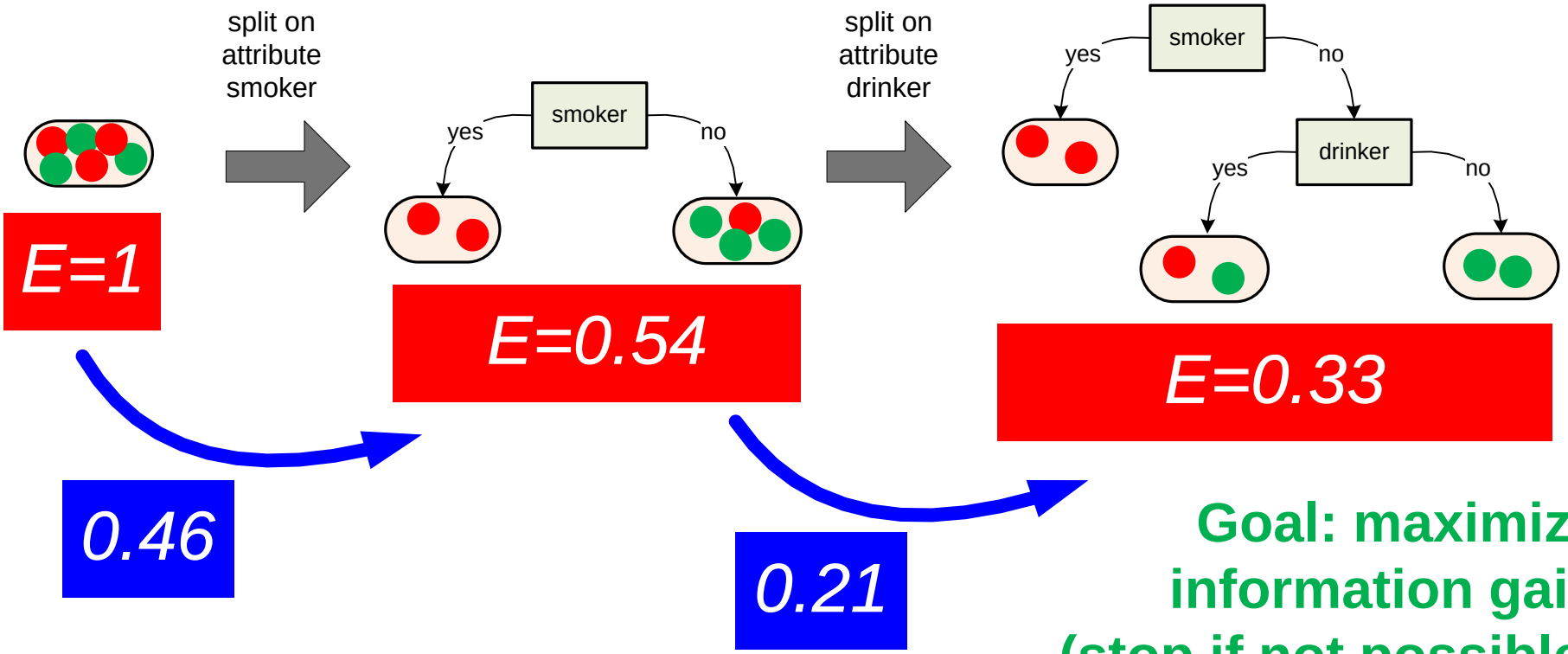
$$E = \frac{2}{6} \times 0 + \frac{4}{6} \times 0.811 = 0.54$$

Weighted average



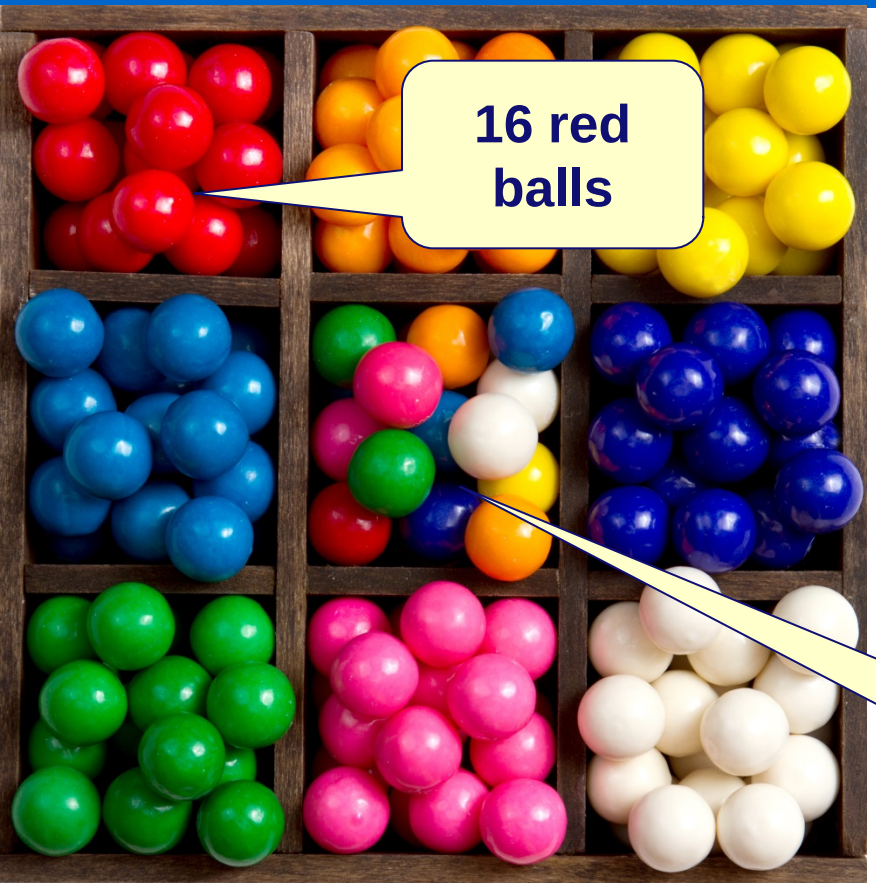
$$E = \frac{2}{6} \times 0 + \frac{2}{6} \times 1 + \frac{2}{6} \times 0 = 0.33$$

Information gain



Goal: maximize information gain (stop if not possible)

Question: Compute entropy



- Compute the entropy of all individual cells.
- What is the overall entropy (weighted average)?
- What is the overall entropy if there is just one cell containing all 144 balls?

TU/e

Answer: *E=3* for cell in middle

- **Cell in the middle: 2+2+2+2+2+2+2+2 balls**

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -\sum_{i=1}^{8} \frac{2}{16} \log_2(\frac{2}{16})$$

$$= -8 \times \frac{1}{8} \times -3 = 3$$

TU/e

Answer: *E=0* for other cells

- **Other cells: 16+0+0+0+0+0+0+0 balls**

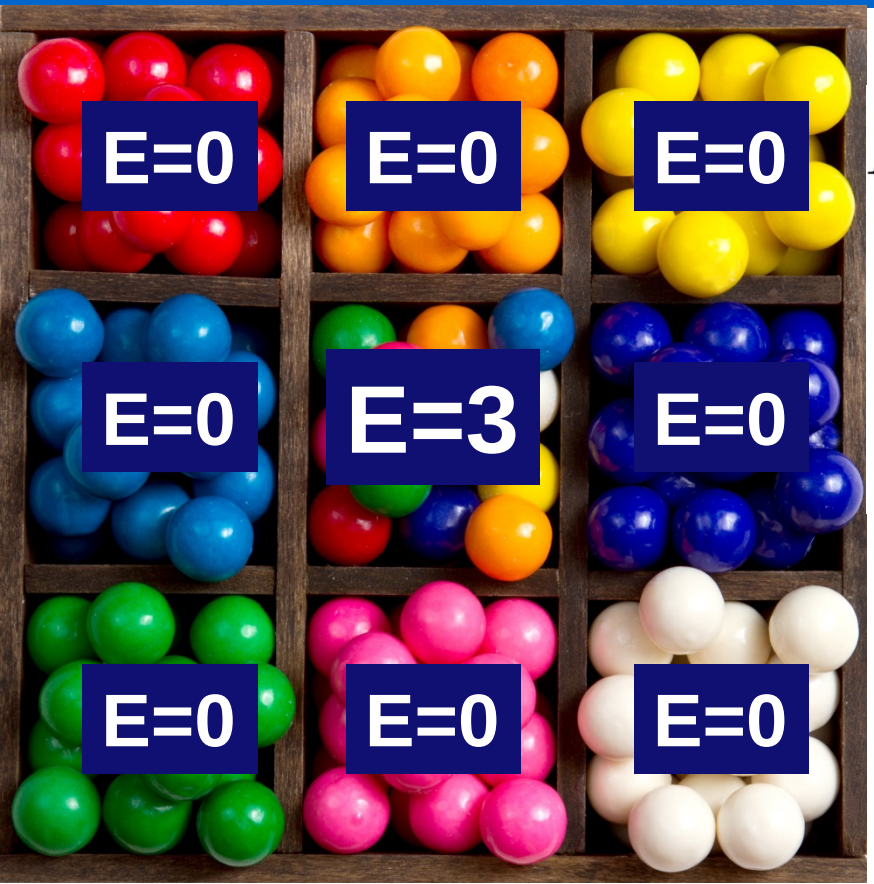
$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -\sum_{i=1}^{1} \frac{16}{16} \log_2(\frac{16}{16})$$

$$= -1 \times 0 = 0$$

TU/e

Overall entropy (weighted average): *E=0.33*



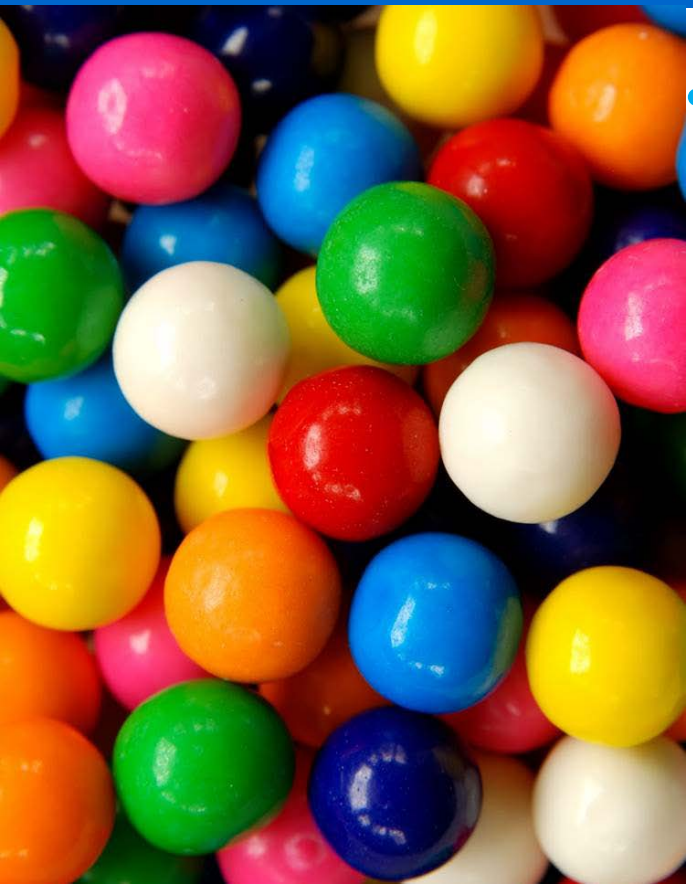
$$E = \frac{16}{144} \times 0 + \frac{16}{144} \times 0 + \ldots + \frac{16}{144} \times 3$$

$$= 8 \times \frac{16}{144} \times 0 + \frac{16}{144} \times 3$$

$$= \frac{1}{9} \times 3 = \frac{1}{3}$$

TU/e

Entropy after mixing the 9 cells: *E=3*



- **144 balls having 8 different colors: 18:18:18:18:18:18:18:18**

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

$$= -\sum_{i=1}^{8} \frac{18}{144} \log_2(\frac{18}{144})$$

$$= -8 \times \frac{1}{8} \times -3 = 3$$

TU/e

E=0.3333

information loss = 2.6666

information gain = 2.6666

E=3

TU/e

$$E = -\frac{546}{860}\log_2\frac{546}{860} + -\frac{314}{860}\log_2\frac{314}{860} = 0.946848$$

#young=546
#old=314
E=0.946848

young
(860/314)

split on attribute smo

information gain is 0.107012

Although the classification did not change there was information gain (we are more certain about the group of smokers).

$$E = -\frac{362}{665}\log_2\frac{362}{665} + -\frac{303}{665}\log_2\frac{303}{665} = 0.994314$$

smoker

yes

no

#young=184
#old=11
E = 0.313027

young
(195/11)

young
(665/303)

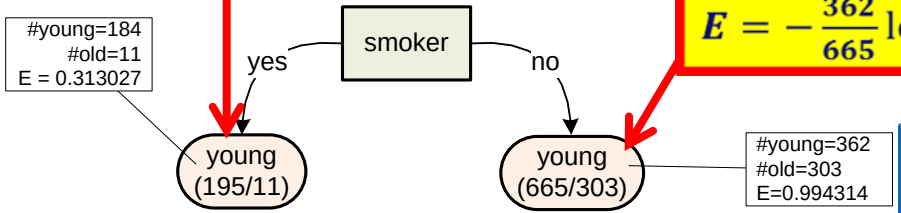
$$\frac{195}{860}0.313027 + \frac{665}{860}0.994314 = 0.839836$$

Overall Entropy

$$E = -\frac{184}{195}\log_2\frac{184}{195} + -\frac{11}{195}\log_2\frac{11}{195} = 0.313027$$

TU/e

$$E = -\frac{184}{195}\log_2\frac{184}{195} + -\frac{11}{195}\log_2\frac{11}{195} = 0.313027$$

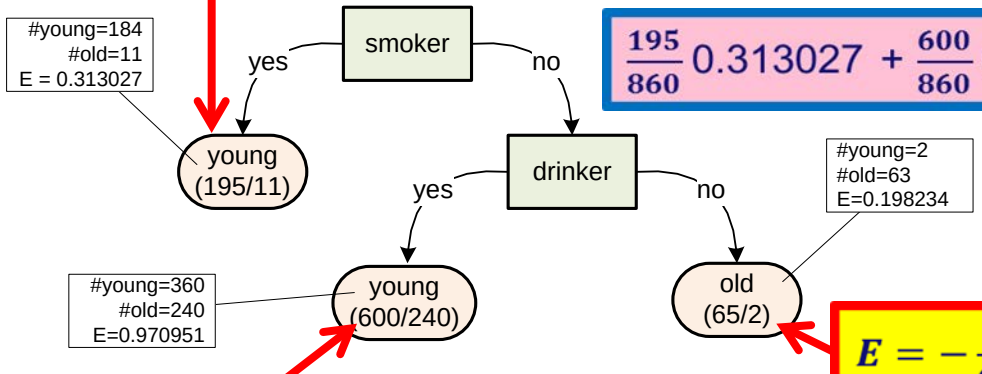


$$E = -\frac{362}{665}\log_2\frac{362}{665} + -\frac{303}{665}\log_2\frac{303}{665} = 0.994314$$

$$\frac{195}{860}\,0.313027 + \frac{665}{860}\,0.994314 = 0.839\ldots6$$

information gain is 0.076468

$$E = -\frac{184}{195}\log_2\frac{184}{195} + -\frac{11}{195}\log_2\frac{11}{195} = 0.313027$$



$$\frac{195}{860}\,0.313027 + \frac{600}{860}\,0.970951 + \frac{65}{860}\,0.198235 = 0.763368$$

$$E = -\frac{2}{65}\log_2\frac{2}{65} + -\frac{63}{65}\log_2\frac{63}{65} = 0.198234$$

$$E = -\frac{360}{600}\log_2\frac{360}{600} + -\frac{240}{600}\log_2\frac{240}{600} = 0.970951$$

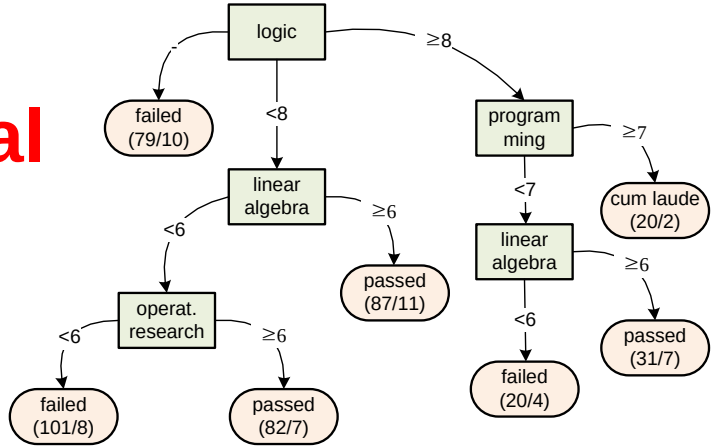
TU/e

Decision tree algorithm (sketch)

- Start with root node corresponding to all instances.
- Iteratively traverse all nodes to see whether "information gain" (i.e., reduction of uncertainty) is possible.
- For each node and for every attribute, check what the effect of splitting the node is in terms of information gain.
- Select the attribute with the biggest information gain above a given threshold.
- Continue until no significant improvement is possible.
- Return the decision tree.

Many parameters/variations are possible

- The **minimal size of a node** before or after splitting.
- **Threshold** setting the **minimal gain** (no split if information gain is too small).
- **Maximal depth** of the tree.
- Allowing the same **label** to appear **multiple** times or not.

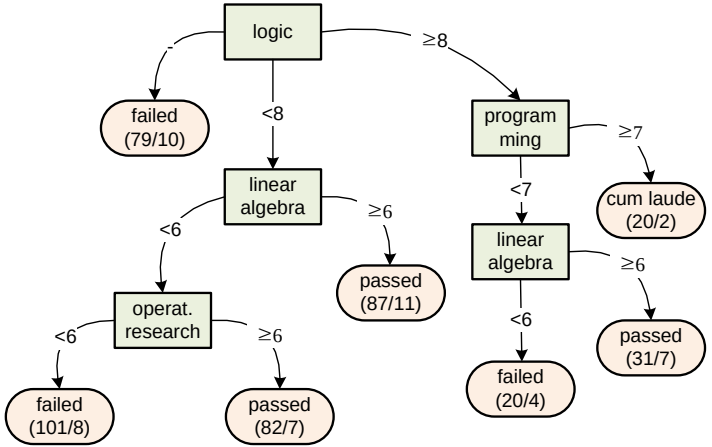


TU/e

Many parameters/variations are possible

- **Alternatives to entropy, e.g., Gini index of diversity.**
- **Splitting the domain of a numerical variable.**
- **Post pruning: removing leaf nodes that do not significantly increase the discriminative power.**

$$G = \sum_{i=1}^{k} p_i(1-p_i) = 1 - \sum_{i=1}^{k} (p_i)^2 \text{ with } p_i = \frac{c_i}{n}$$

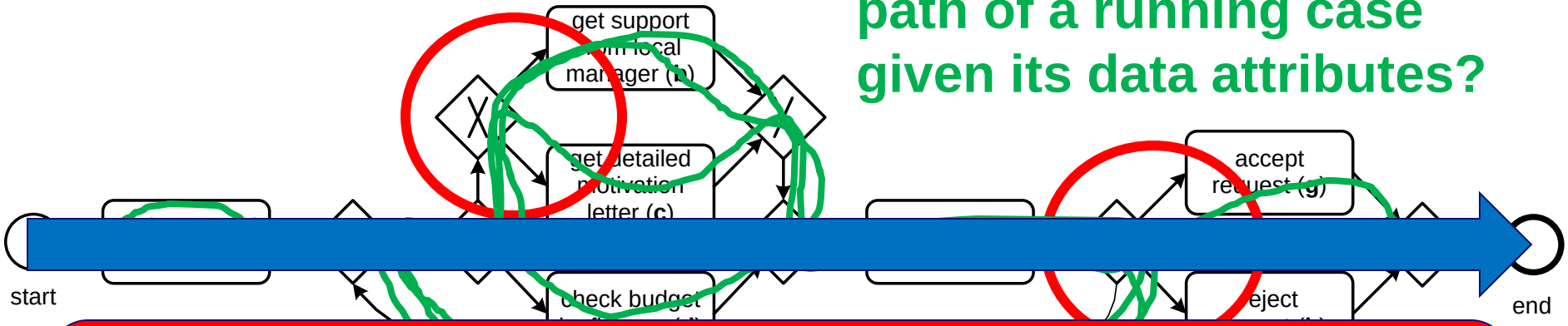


TU/e

Example applications in process mining

What is driving these decisions?

What is the most likely path of a running case given its data attributes?



These questions require a discovered process model on which we can replay the event log!

TU/e

Part I: Preliminaries

- Chapter 1: Introduction
- Chapter 2: Process Modeling and Analysis
- Chapter 3: Data Mining

Part II: From Event Logs to Process Models

- Chapter 4: Getting the Data
- Chapter 5: Process Discovery: An Introduction
- Chapter 6: Advanced Process Discovery Techniques

Part III: Beyond Process Discovery

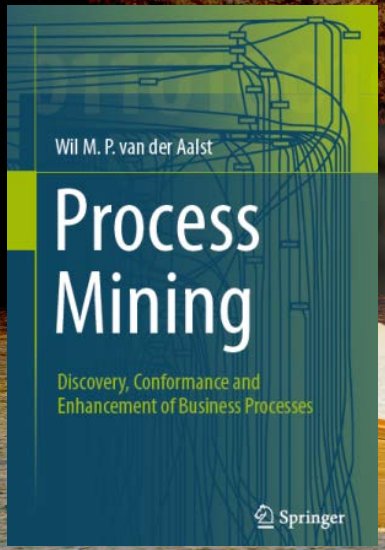
- Chapter 7: Conformance Checking
- Chapter 8: Mining Additional Perspectives
- Chapter 9: Operational Support

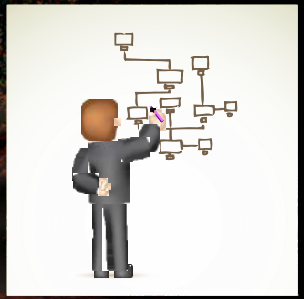
Part IV: Putting Process Mining to Work

- Chapter 10: Tool Support
- Chapter 11: Analyzing “Lasagna Processes”
- Chapter 12: Analyzing “Spaghetti Processes”

Part V: Reflection

- Chapter 13: Cartography and Navigation
- Chapter 14: Epilogue





TU/e