Interpreting Classifiers through Attribute Interactions in datasets

Andreas Henelius, Kai Puolamäki, Antti Ukkonen

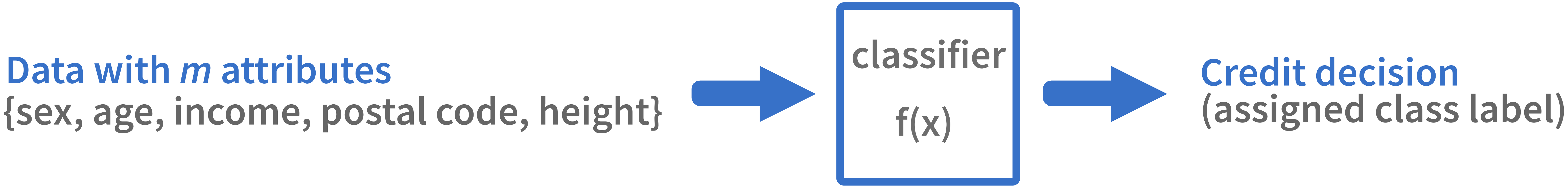
Finnish Institute of Occupational Health
Helsinki, Finland

ICML Workshop on Human Interpretability in Machine Learning 2017

Finnish Institute of
Occupational Health

BACKGROUND

State-of-the-art classifiers are high-performing but essentially **black boxes.**



Data with *m* attributes
{sex, age, income, postal code, height}

classifier

f(x)

Credit decision
(assigned class label)

Good predictions, but why?

BACKGROUND

Need for algorithmic transparency

- decision making (e.g., possible legislative demands)
- to learn about the data (why was the class assigned?)

How can we gain insight into how the algorithms work?

INTERPRETABILITY

through attribute interactions

Interaction

Attributes are conditionally dependent given the class, i.e., attributes are together meaningful when predicting the class.

INTERPRETABILITY

through attribute interactions

Applications

bioinformatics

interactions between SNPs

pharmacogivilance

interactions between simultaneously administered drugs

INTERPRETABILITY

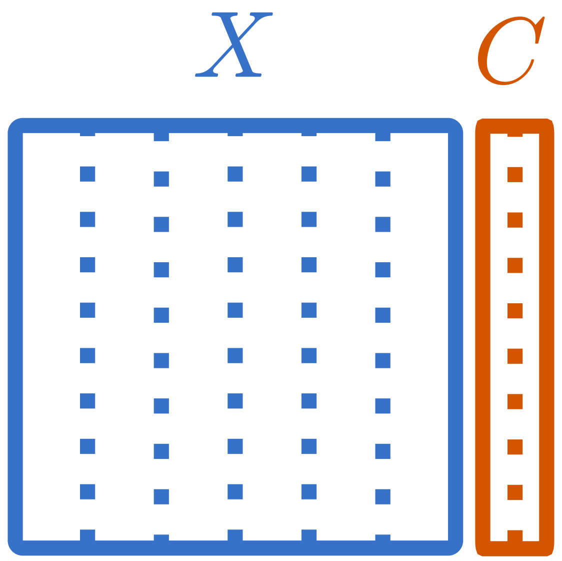
through attribute interactions

Grouping of attributes $\mathcal{S}$

$\mathcal{S}$ = { {sex, age}, {income, postal code}, {height}}

- attributes in the same group interact,
- attributes in different groups are independent

INTUITION OF CLASSIFIERS



Dataset: $D = (X, C)$

Grouping of attributes: $\mathcal{S}$

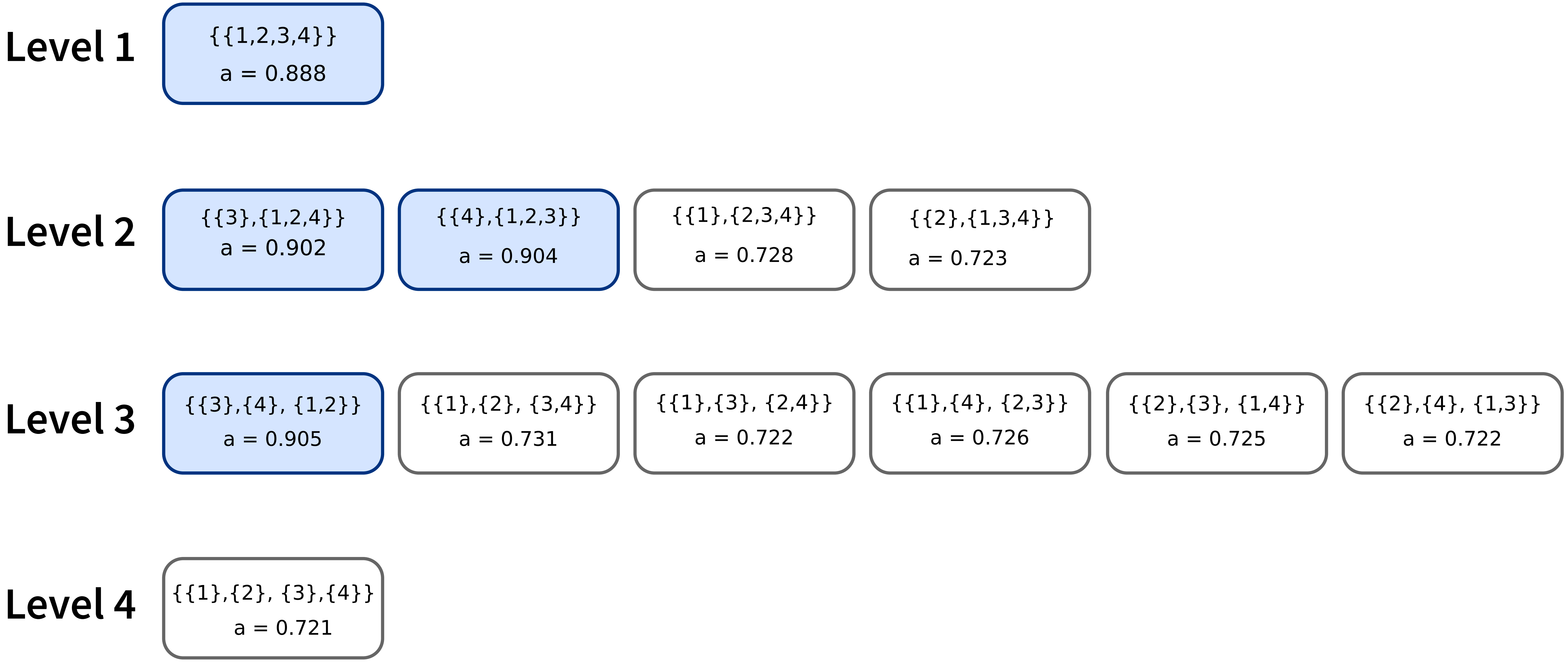
The classifier models $P(C|X) \propto \underbrace{P(X|C)}_{\text{class-conditional distribution}} P(C)$

Factorise the class-conditional joint data distribution (parametrised by $\mathcal{S}$)

$$P(X \mid C; \mathcal{S}) = \prod_{S \in \mathcal{S}} P(X(\cdot, S) \mid C)$$

INTUITION OF CLASSIFIERS

Assumption: classification accuracy decreases monotonically
Known grouping: {{1, 2}, {3}, {4}}



Level 1

{{1,2,3,4}}
a = 0.888

Level 2

{{3},{1,2,4}}
a = 0.902

{{4},{1,2,3}}
a = 0.904

{{1},{2,3,4}}
a = 0.728

{{2},{1,3,4}}
a = 0.723

Level 3

{{3},{4}, {1,2}}
a = 0.905

{{1},{2}, {3,4}}
a = 0.731

{{1},{3}, {2,4}}
a = 0.722

{{1},{4}, {2,3}}
a = 0.726

{{2},{3}, {1,4}}
a = 0.725

{{2},{4}, {1,3}}
a = 0.722

Level 4

{{1},{2}, {3},{4}}
a = 0.721

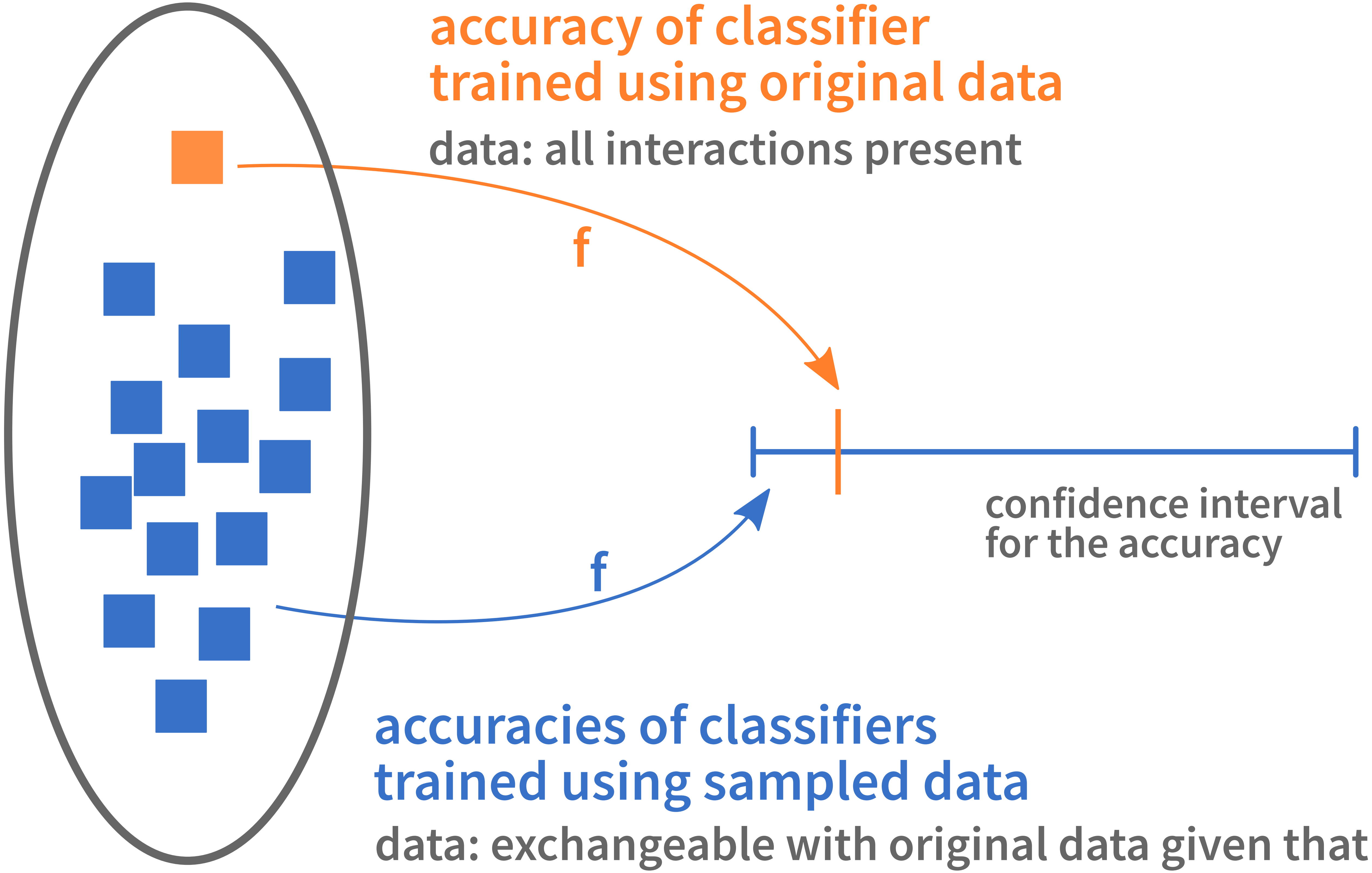
GOALS

$\mathcal{S}$ = { {sex, age}, {income, postal code}, {height}}

(1) **Verify** if a given grouping of attributes represents the attribute interaction structure

(2) **Automatically find** the maximum cardinality grouping of the attributes in the dataset

PROBLEM 1: VERIFY A GIVEN GROUPING



accuracy of classifier
trained using original data

data: all interactions present

accuracies of classifiers
trained using sampled data

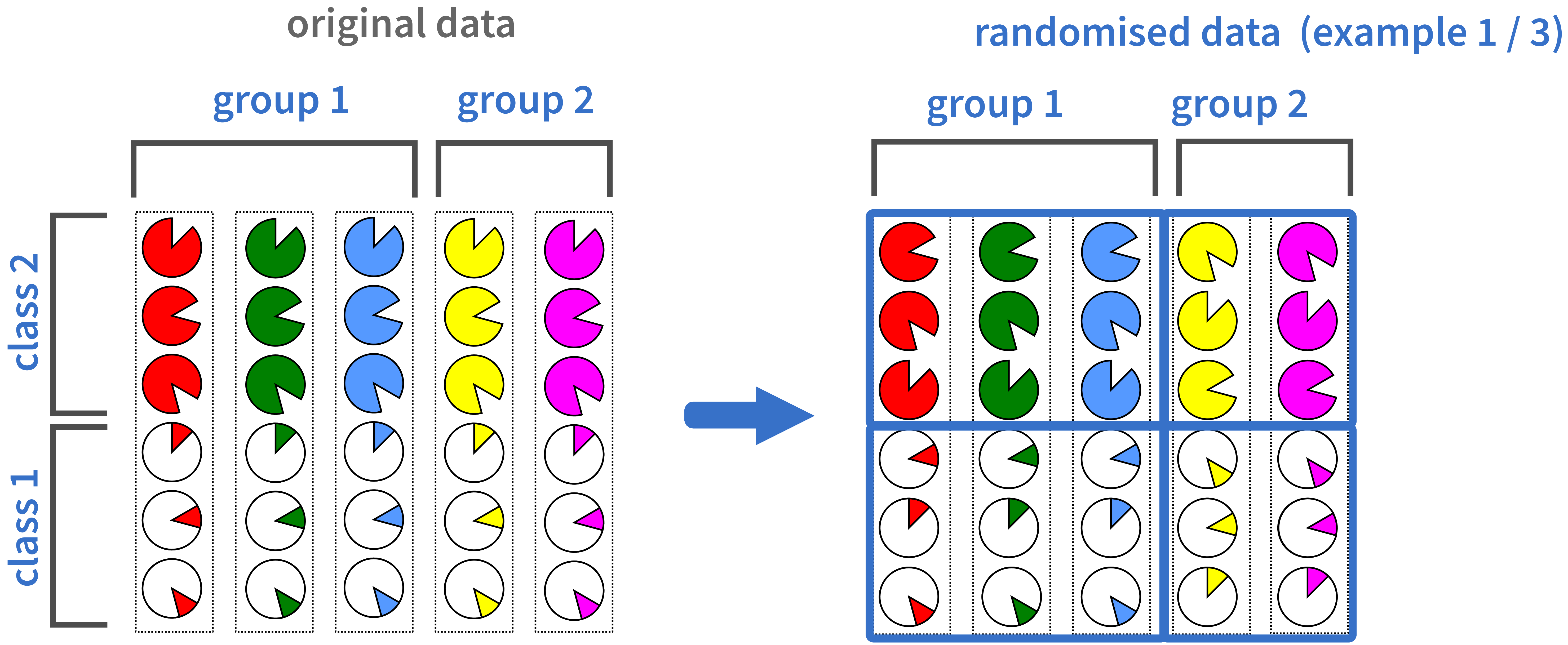
data: exchangeable with original data given that
the factorisation holds

PROBLEM 1: VERIFY A GIVEN GROUPING

How do we get the data needed to construct the confidence intervals?

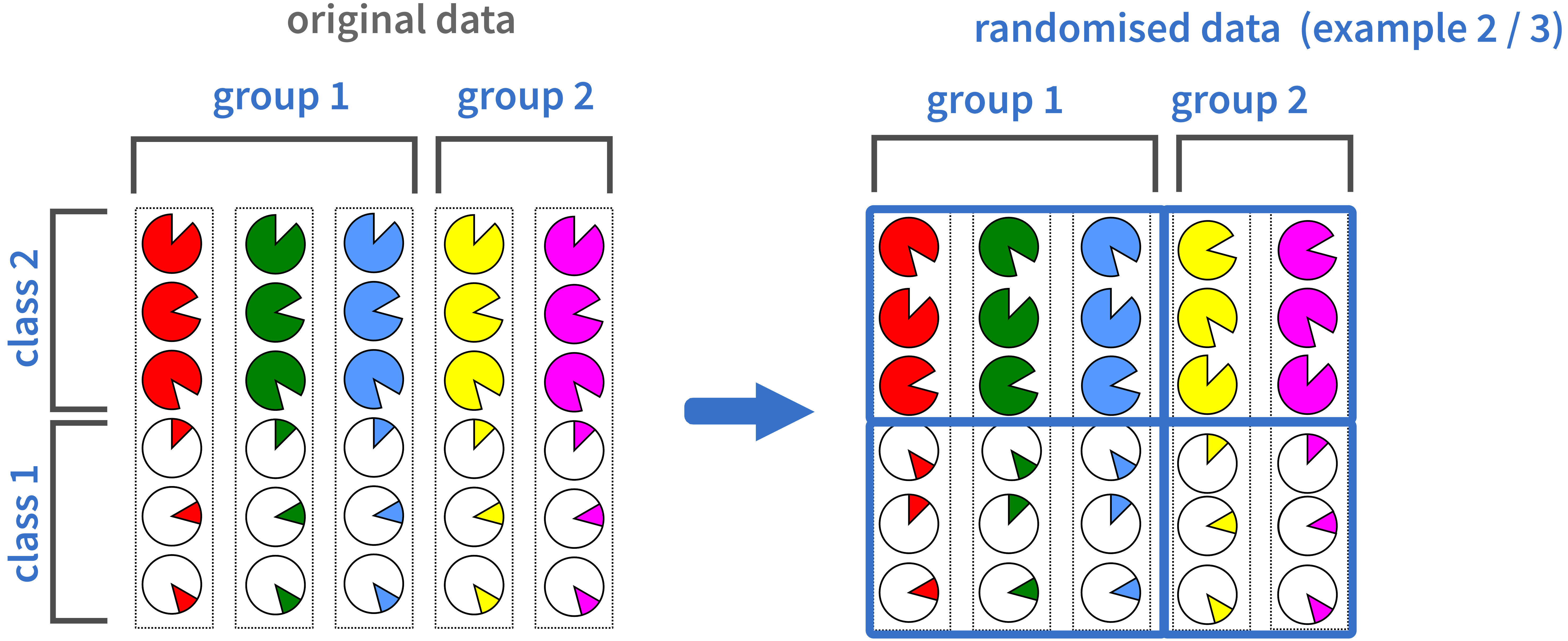
GENERATING EXCHANGEABLE DATASETS

Constrained randomisation, parametrised by the attribute grouping $\mathcal{S}$



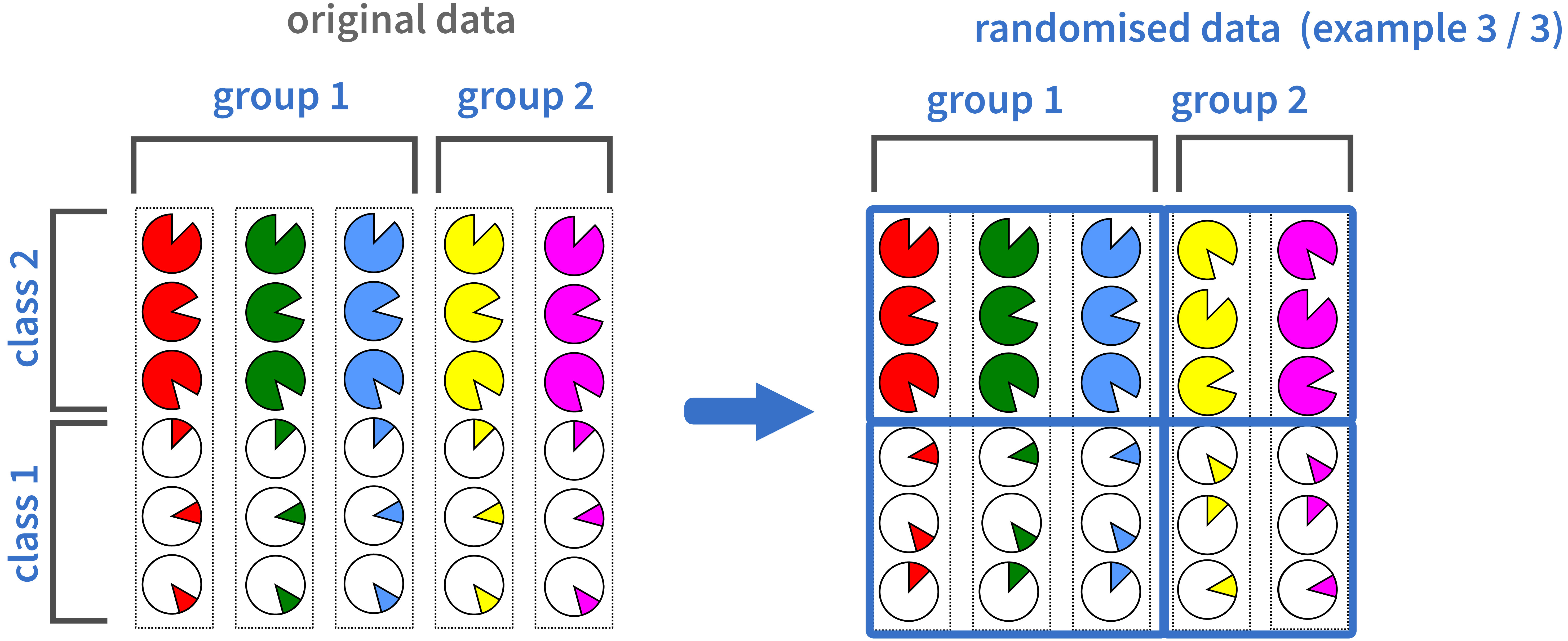
GENERATING EXCHANGEABLE DATASETS

Constrained randomisation, parametrised by the attribute grouping $\mathcal{S}$



GENERATING EXCHANGEABLE DATASETS

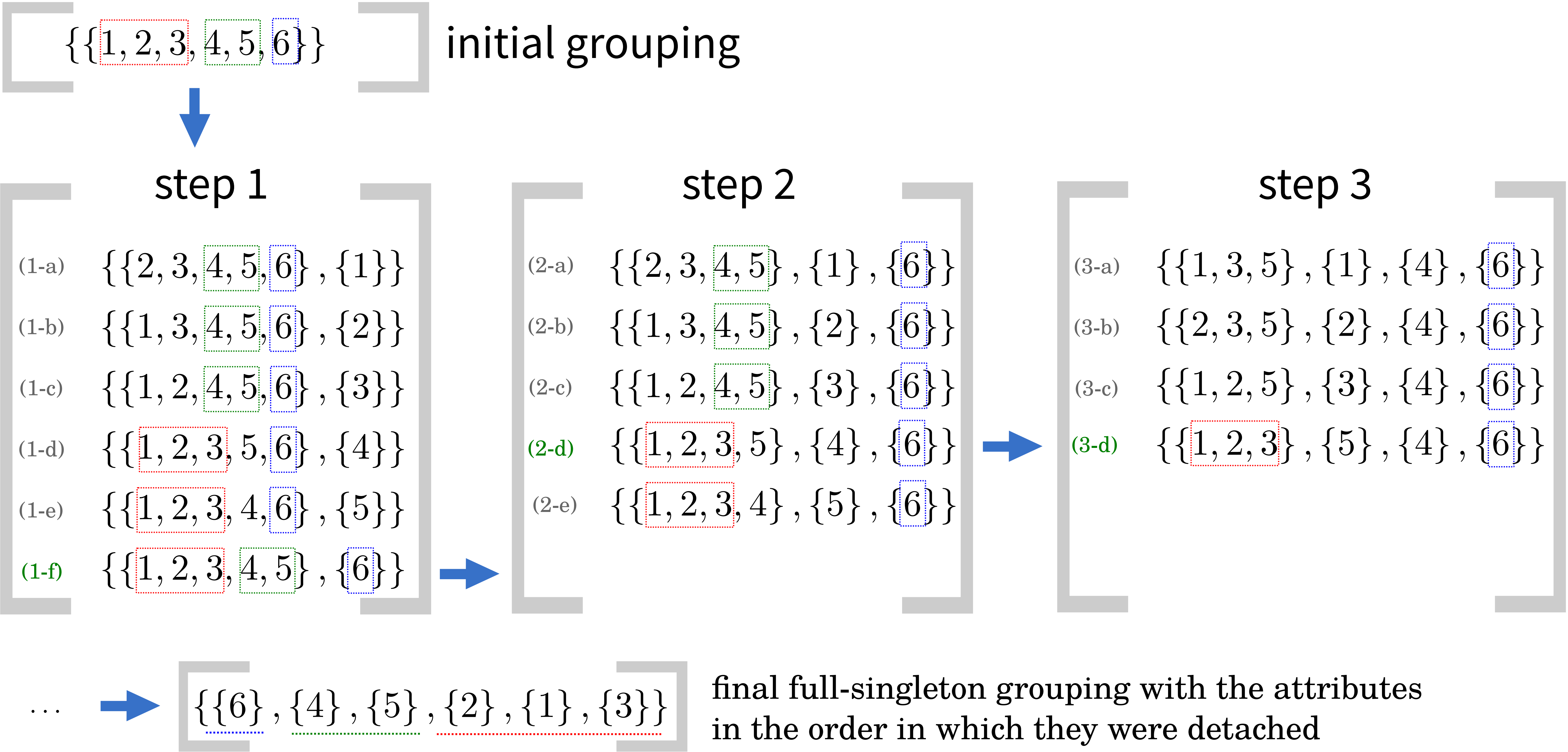
Constrained randomisation, parametrised by the attribute grouping $\mathcal{S}$



PROBLEM 2: FIND GROUPING

Top-down greedy algorithm: ASTRID
Two steps: sorting and grouping

sorting step



$\{\{1,2,3,4,5,6\}\}$ initial grouping

step 1

(1-a) $\{\{2,3,4,5,6\},\{1\}\}$

(1-b) $\{\{1,3,4,5,6\},\{2\}\}$

(1-c) $\{\{1,2,4,5,6\},\{3\}\}$

(1-d) $\{\{1,2,3,5,6\},\{4\}\}$

(1-e) $\{\{1,2,3,4,6\},\{5\}\}$

(1-f) $\{\{1,2,3,4,5\},\{6\}\}$

step 2

(2-a) $\{\{2,3,4,5\},\{1\},\{6\}\}$

(2-b) $\{\{1,3,4,5\},\{2\},\{6\}\}$

(2-c) $\{\{1,2,4,5\},\{3\},\{6\}\}$

(2-d) $\{\{1,2,3,5\},\{4\},\{6\}\}$

(2-e) $\{\{1,2,3,4\},\{5\},\{6\}\}$

step 3

(3-a) $\{\{1,3,5\},\{1\},\{4\},\{6\}\}$

(3-b) $\{\{2,3,5\},\{2\},\{4\},\{6\}\}$

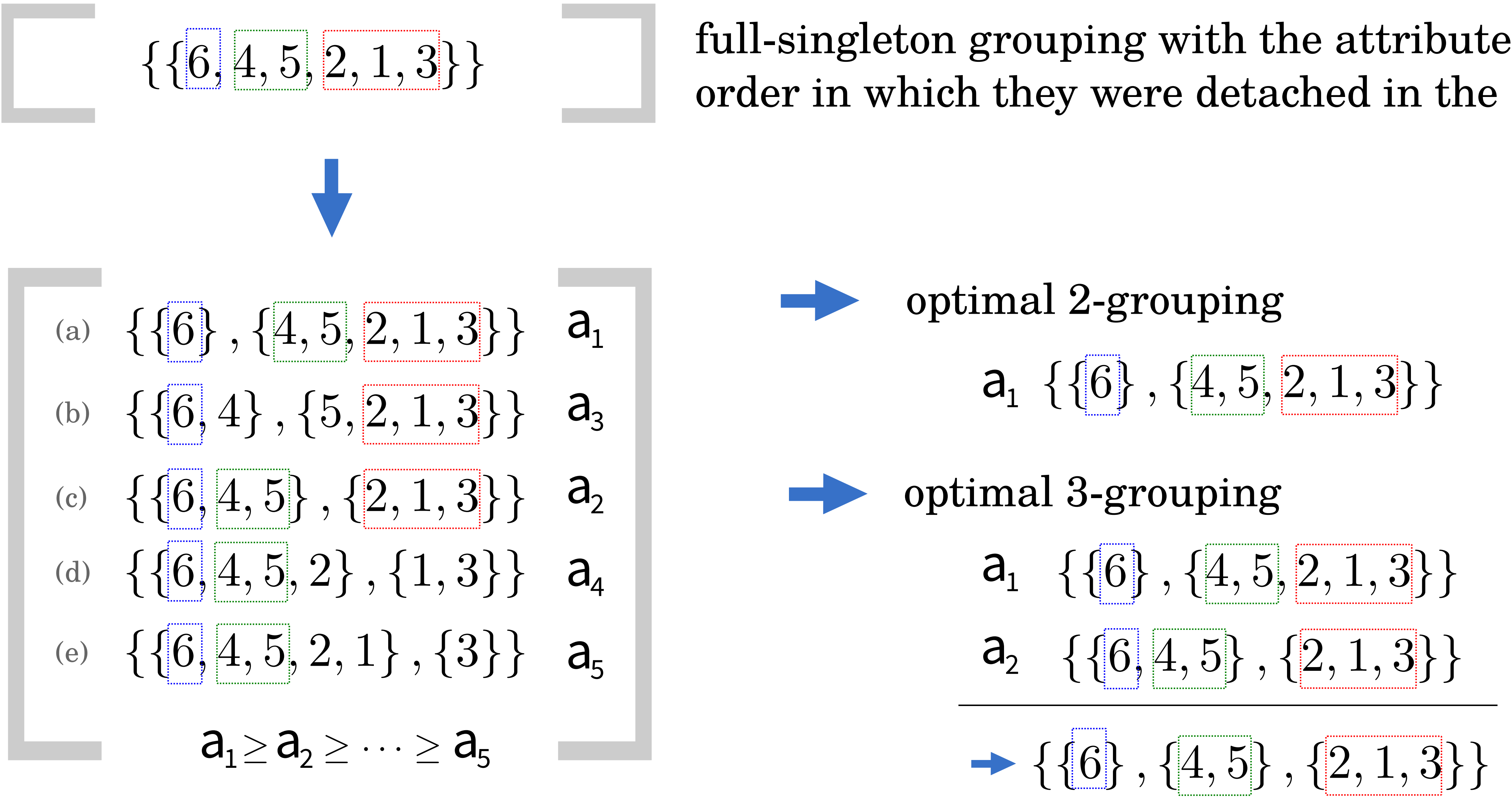
(3-c) $\{\{1,2,5\},\{3\},\{4\},\{6\}\}$

(3-d) $\{\{1,2,3\},\{5\},\{4\},\{6\}\}$

$\ldots$ → $\{\{6\},\{4\},\{5\},\{2\},\{1\},\{3\}\}$ final full-singleton grouping with the attributes in the order in which they were detached

PROBLEM 2: FIND GROUPING

grouping step



$\{\{\boxed{6}, \boxed{4, 5}, \boxed{2, 1, 3}\}\}$

full-singleton grouping with the attributes in the order in which they were detached in the sorting step

(a) $\{\{\boxed{6}\}, \{\boxed{4, 5}, \boxed{2, 1, 3}\}\}$ a_1

(b) $\{\{\boxed{6}, 4\}, \{5, \boxed{2, 1, 3}\}\}$ a_3

(c) $\{\{\boxed{6}, \boxed{4, 5}\}, \{\boxed{2, 1, 3}\}\}$ a_2

(d) $\{\{\boxed{6}, \boxed{4, 5}, 2\}, \{1, 3\}\}$ a_4

(e) $\{\{\boxed{6}, \boxed{4, 5}, 2, 1\}, \{3\}\}$ a_5

$a_1 \geq a_2 \geq \cdots \geq a_5$

optimal 2-grouping

a_1 $\{\{\boxed{6}\}, \{\boxed{4, 5}, \boxed{2, 1, 3}\}\}$

optimal 3-grouping

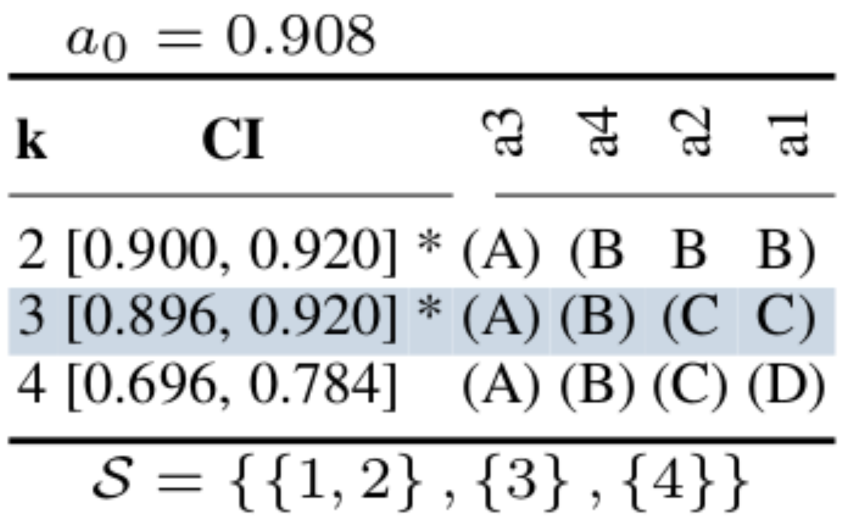
a_1 $\{\{\boxed{6}\}, \{\boxed{4, 5}, \boxed{2, 1, 3}\}\}$

a_2 $\{\{\boxed{6}, \boxed{4, 5}\}, \{\boxed{2, 1, 3}\}\}$

$\{\{\boxed{6}\}, \{\boxed{4, 5}\}, \{\boxed{2, 1, 3}\}\}$

RESULTS

SVM

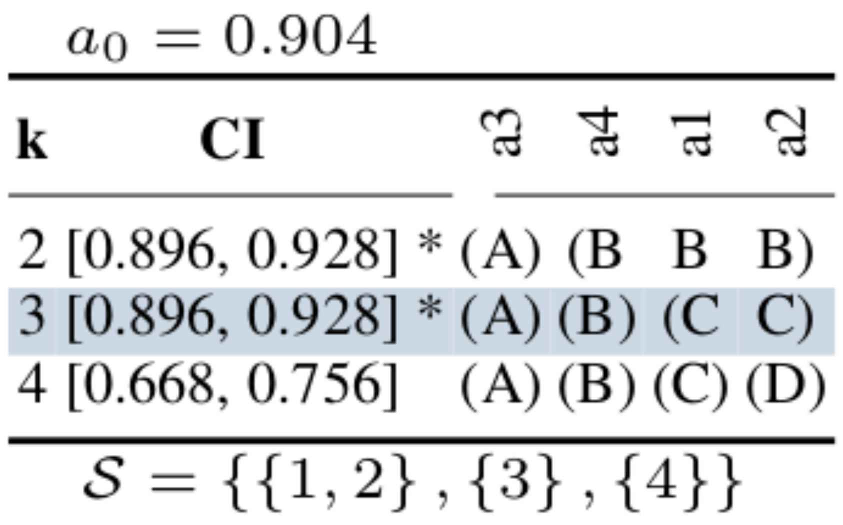


$a_0 = 0.908$

k	CI		a3	a4	a2	a1
2	[0.900, 0.920]	*	(A)	(B	B	B)
3	[0.896, 0.920]	*	(A)	(B)	(C	C)
4	[0.696, 0.784]		(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1,2\},\{3\},\{4\}\}$

Random Forest

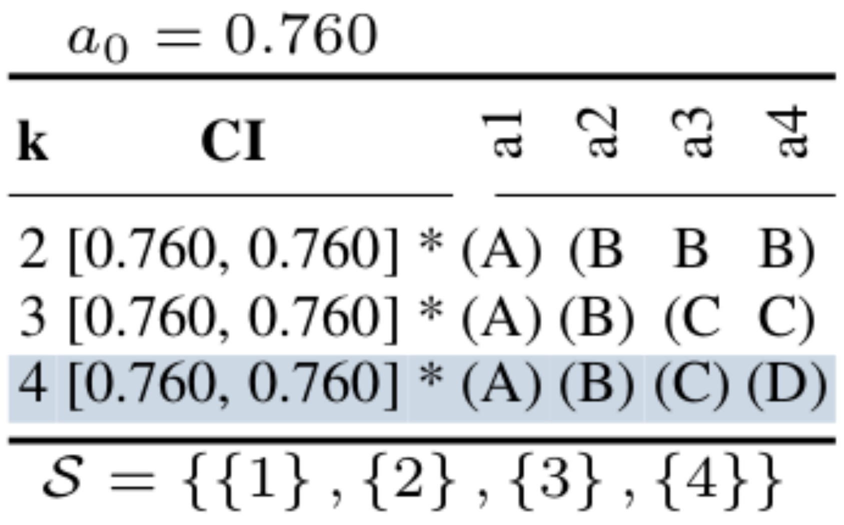


$a_0 = 0.904$

k	CI		a3	a4	a1	a2
2	[0.896, 0.928]	*	(A)	(B	B	B)
3	[0.896, 0.928]	*	(A)	(B)	(C	C)
4	[0.668, 0.756]		(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1,2\},\{3\},\{4\}\}$

Naive Bayes



$a_0 = 0.760$

k	CI		a1	a2	a3	a4
2	[0.760, 0.760]	*	(A)	(B	B	B)
3	[0.760, 0.760]	*	(A)	(B)	(C	C)
4	[0.760, 0.760]	*	(A)	(B)	(C)	(D)

$\mathcal{S} = \{\{1\},\{2\},\{3\},\{4\}\}$

Known grouping: {{1, 2}, {3}, {4}}

SUMMARY

- Investigate attribute interactions used by classifiers
- Find groups of interacting attributes
- Enhances interpretability of opaque classifiers
- R-package: **https://github.com/bwrc/astrid-r/**

Thank you!