07
Clustering & Classification

Dominic Schmitz & Janina Esser

Clustering & Classification

Clustering

- das **unüberwachte Finden/Bestimmen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ Multidimensional Scaling, Hierarchical Cluster Analysis

Classification

- das **überwachte Überprüfen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ Classification Trees

Clustering & Classification

Clustering

- das **unüberwachte Finden/Bestimmen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Multidimensional Scaling**, Hierarchical Cluster Analysis

Classification

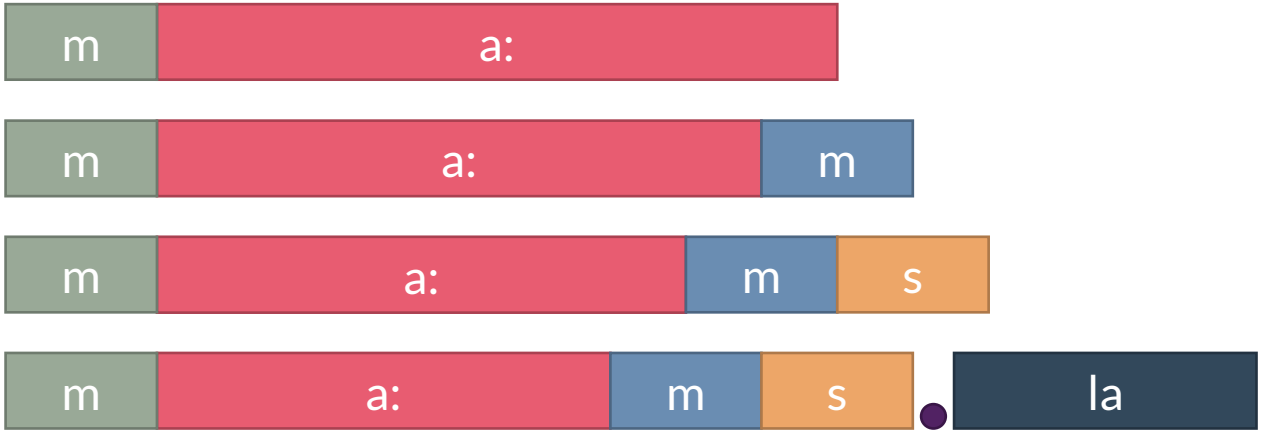
- das **überwachte Überprüfen** von Strukturen/Mustern in Daten anhand von Gruppierungen

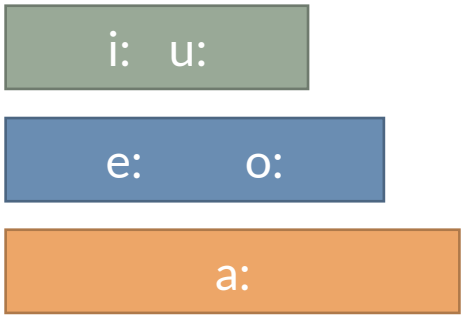
→ Classification Trees

Multidimensional Scaling

- Methode zur Veranschaulichung von (Un-)Ähnlichkeiten zwischen Datenpunkten
- Hierzu werden die Daten auf 2 oder 3 Dimensionen reduziert
- Für die folgenden Beispiele werden wir Daten folgender Studie nutzen:

Compensatory Vowel Shortening in German[1]





[1] Schmitz, D., Cho, H.-E., & Niemann, H. (2018). Vowel shortening in German as a function of syllable structure. Proceedings 13. Phonetik Und Phonologie Tagung (P&P13), 181–184.

Multidimensional Scaling

Frage

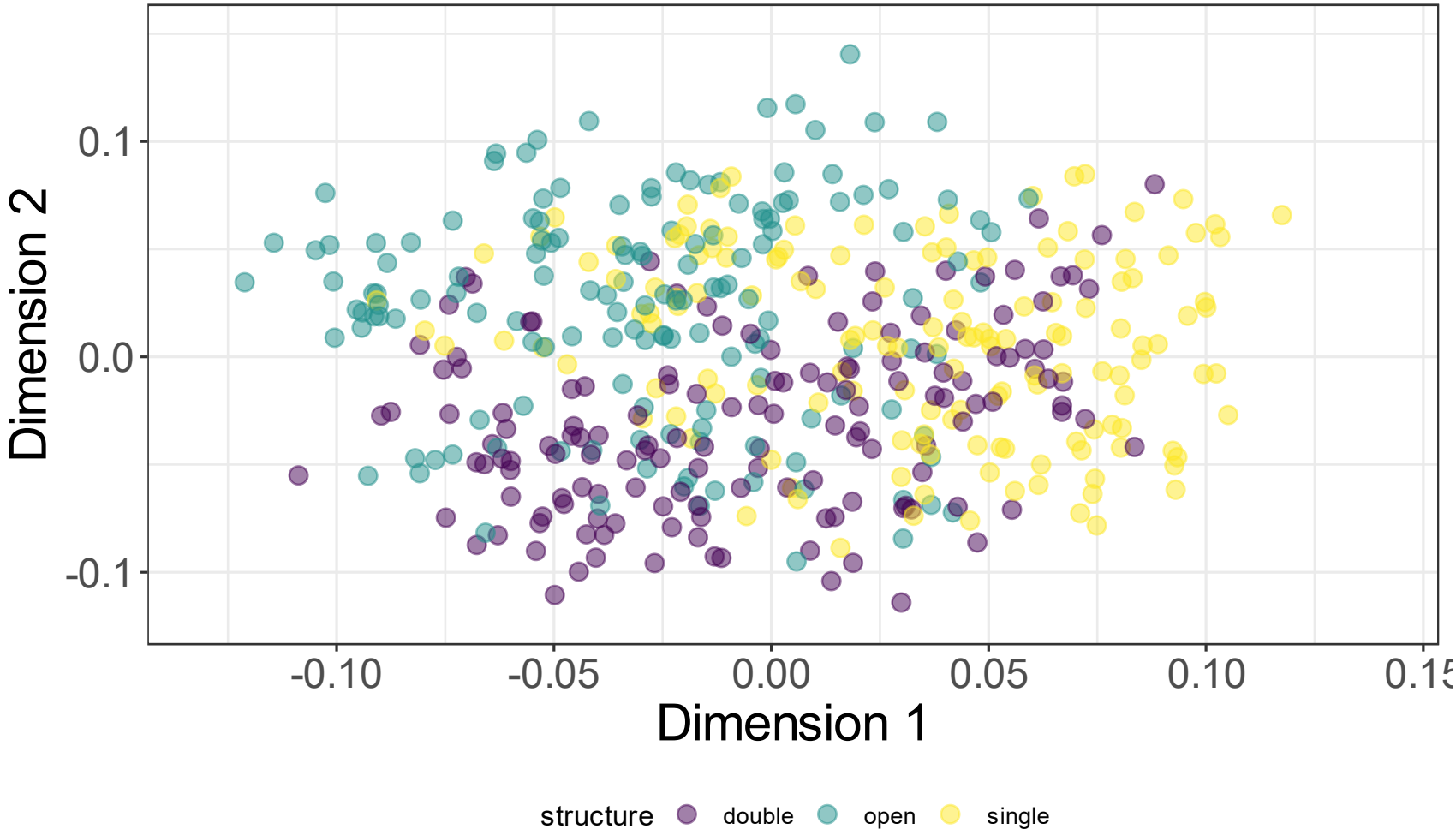
Finden wir in der abhängigen Variable dieser Studie, der Vokaldauer, Muster anhand anderer Variablen, etwa Silbenstruktur und Vokalqualität?

Methode

Reduktion auf 2 (oder 3) Dimensionen, anschließend **Visualisierung** der Dimensionen sowie **statistische Überprüfung** (Korrelation, t-Test, ANOVA)

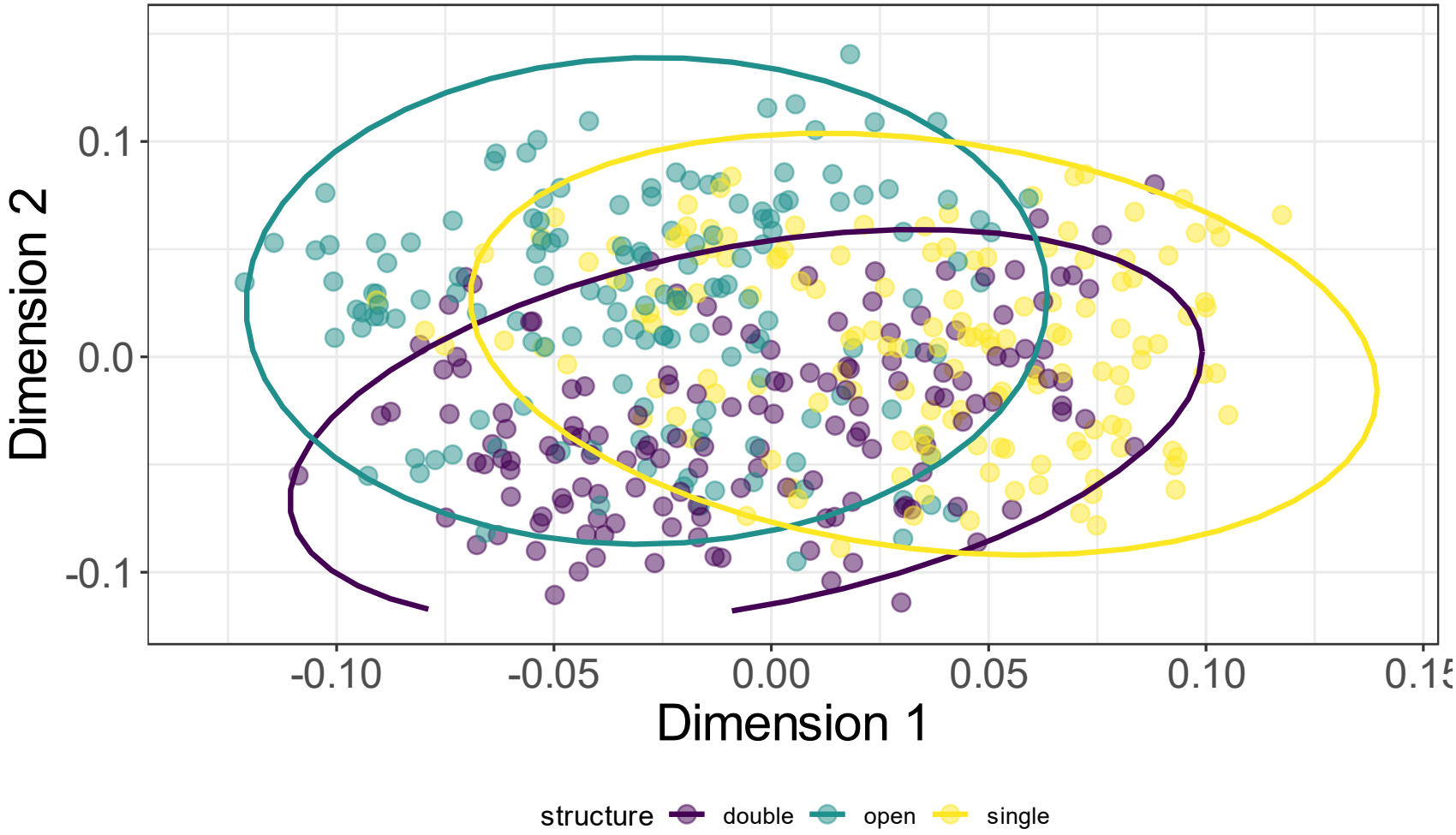
Multidimensional Scaling

Visualisierung - structure



Multidimensional Scaling

Visualisierung - structure



Multidimensional Scaling

Überprüfung - structure

→ via ANOVA, da wir im konkreten Fall eine numerische Variable anhand einer kategorischen Variable mit mehr als 2 Levels modellieren

Dimension 1 ~ structure

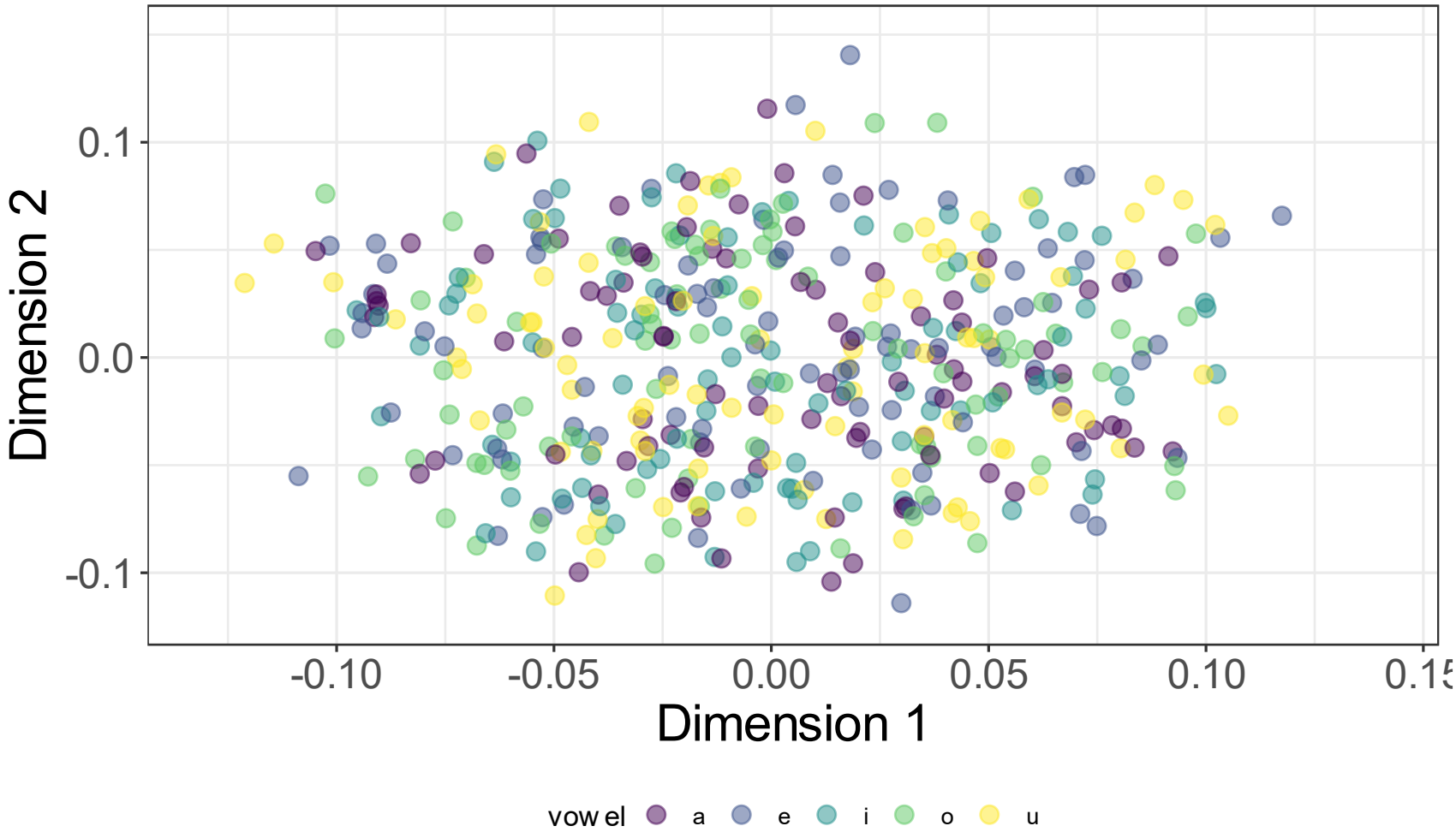
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
structure	2	0.284	0.142	70.990	0.000	***
Residuals	445	0.890	0.002			

Dimension 2 ~ structure

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
structure	2	0.226	0.113	55.410	0.000	***
Residuals	445	0.907	0.002			

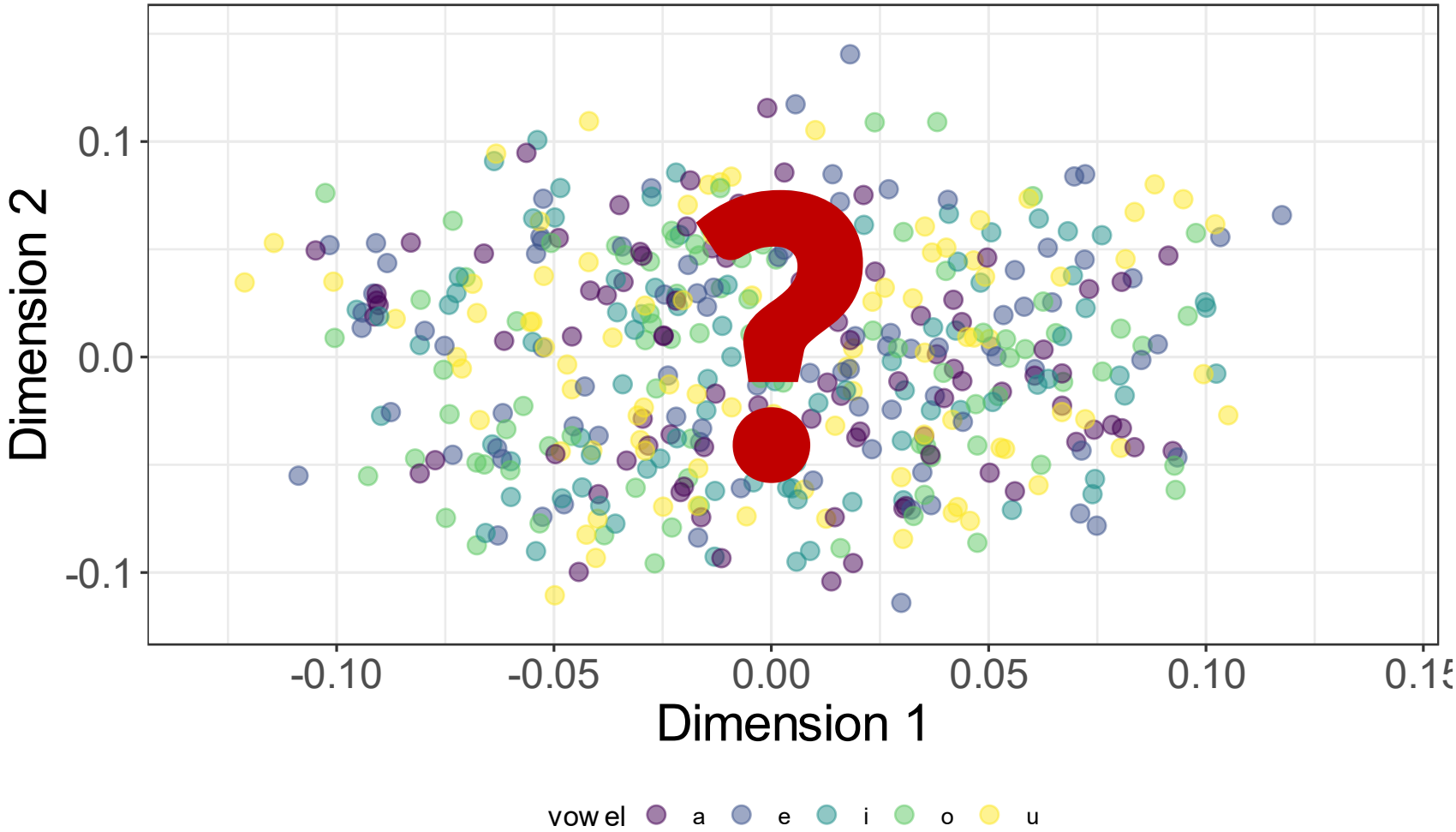
Multidimensional Scaling

Visualisierung - vowel



Multidimensional Scaling

Visualisierung - vowel



Multidimensional Scaling

Überprüfung - vowel

→ via ANOVA, da wir im konkreten Fall eine numerische Variable anhand einer kategorischen Variable mit mehr als 2 Levels modellieren

Dimension 1 ~ vowel

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vowel	4	0.002	0.000	0.148	0.964
Residuals	443	1.173	0.003		

Dimension 2 ~ vowel

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vowel	4	0.005	0.001	0.447	0.775
Residuals	443	1.129	0.003		

Clustering & Classification

Clustering

- das **unüberwachte Finden/Bestimmen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Multidimensional Scaling**, Hierarchical Cluster Analysis

Classification

- das **überwachte Überprüfen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ Classification Trees

Clustering & Classification

Clustering

- das **unüberwachte Finden/Bestimmen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Multidimensional Scaling**, **Hierarchical Cluster Analysis**

Classification

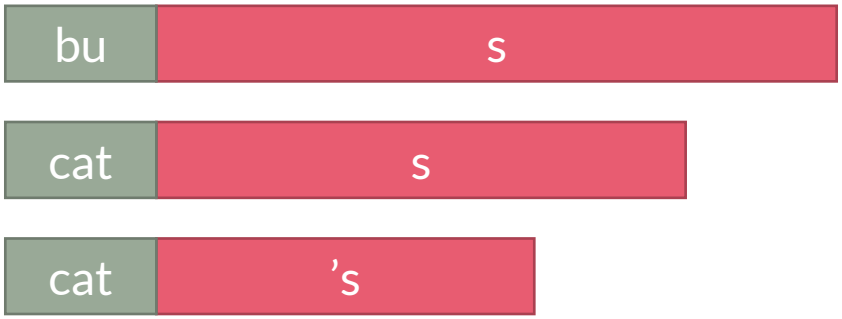
- das **überwachte Überprüfen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ Classification Trees

Hierarchical Cluster Analysis

- Methode zur Veranschaulichung von (Un-)Ähnlichkeiten zwischen Variablen
- Hierzu werden die Variablen schrittweise geclustert
- Für die folgenden Beispiele werden wir Daten folgender Studie nutzen:

The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords[1]



[1] Schmitz, D., Baer-Henney, D., & Plag, I. (2021). The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords. Phonetica, 78(5-6), 571-616. doi: 10.1515/phon-2021-2013

Hierarchical Cluster Analysis

Frage

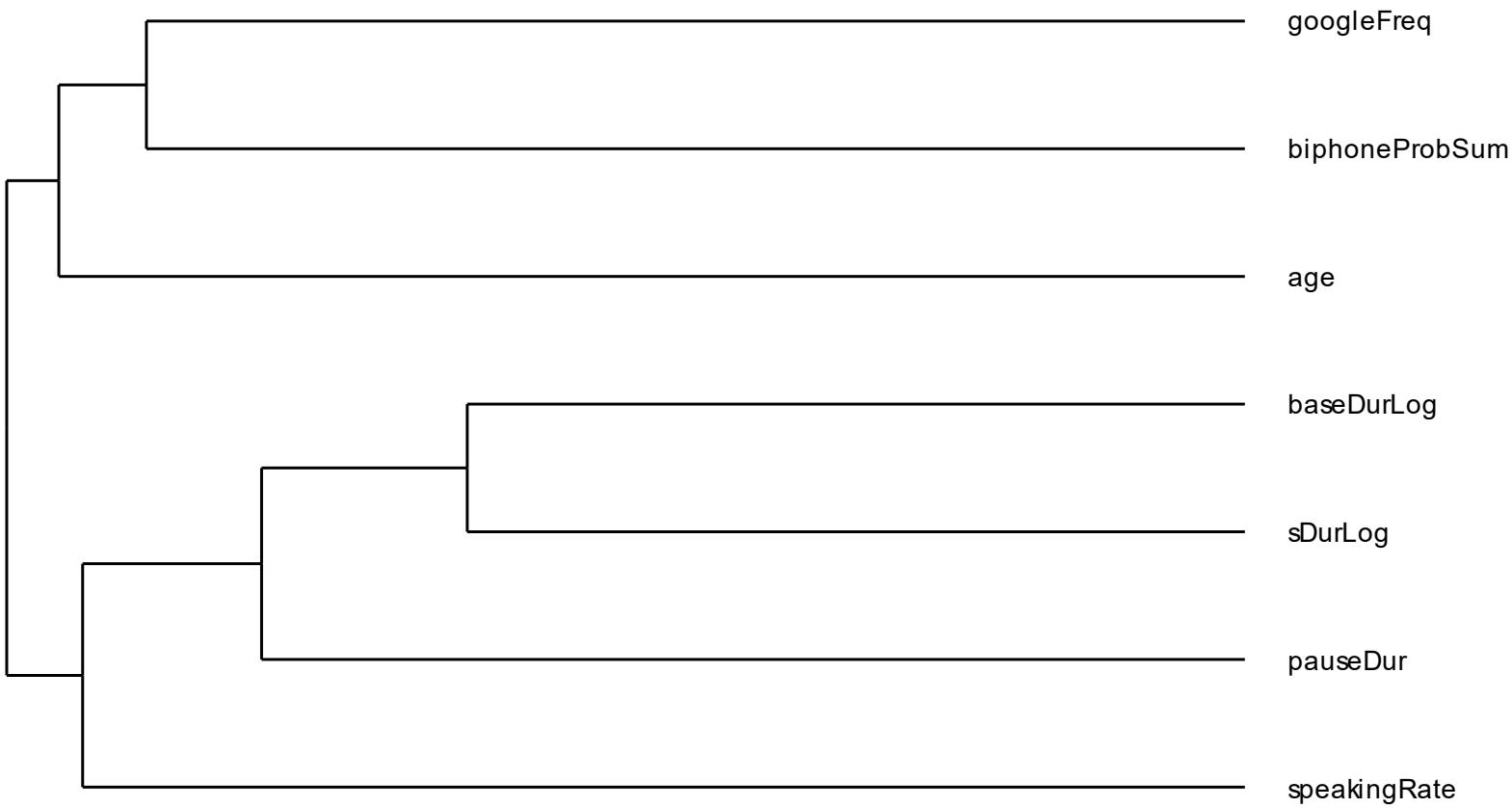
Finden wir in der Menge unserer Variablen solche, die einander ähnlich sind?

Methode

Jeder Datenpunkt wird anfangs als Cluster identifiziert. Dann werden wiederholt die zwei Cluster, die am nächsten zueinander sind zu einem Cluster zusammengefasst. Beendet ist der Vorgang dann, wenn nur noch ein Cluster vorhanden ist.

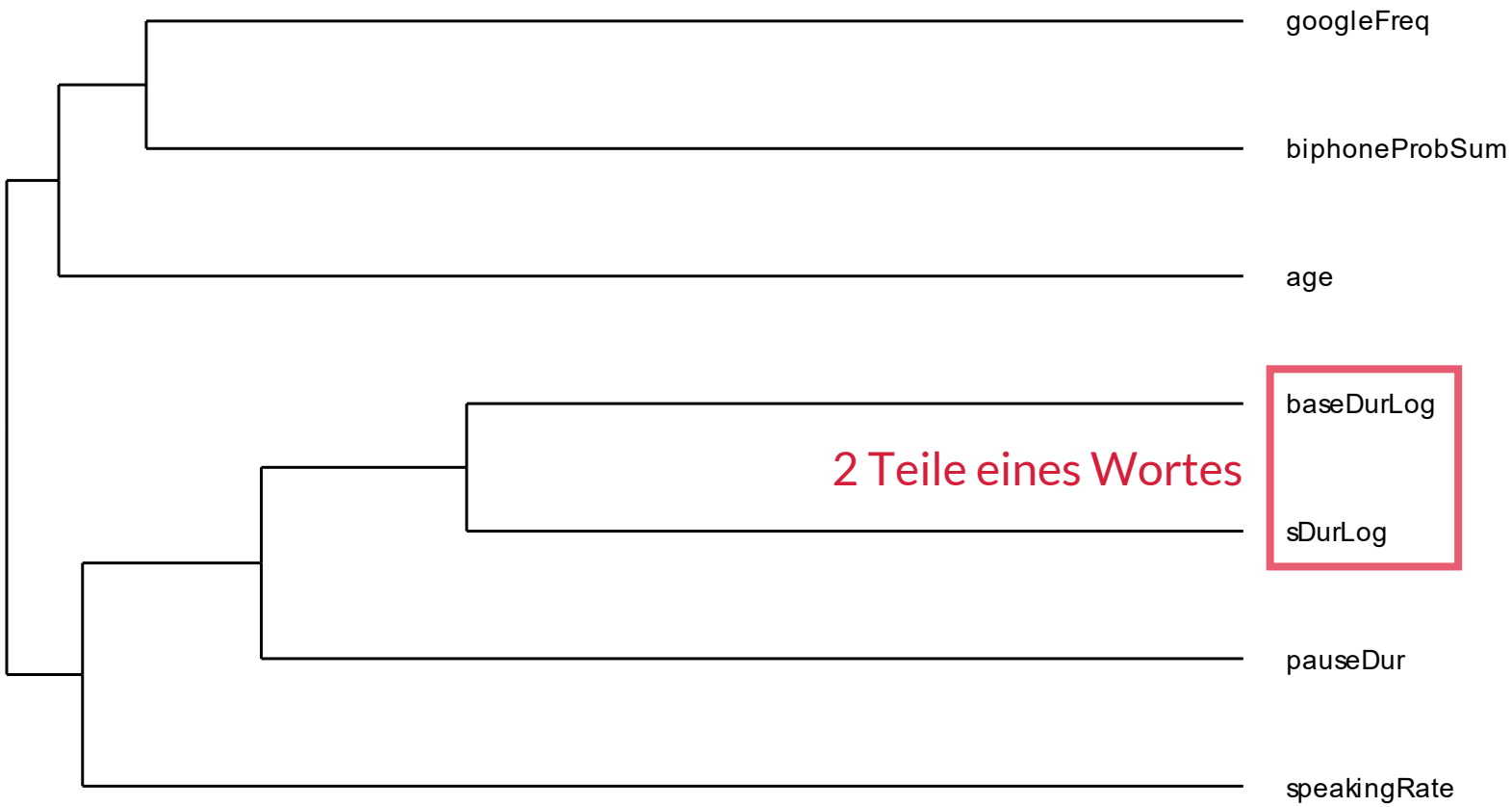
Hierarchical Cluster Analysis

Visualisierung



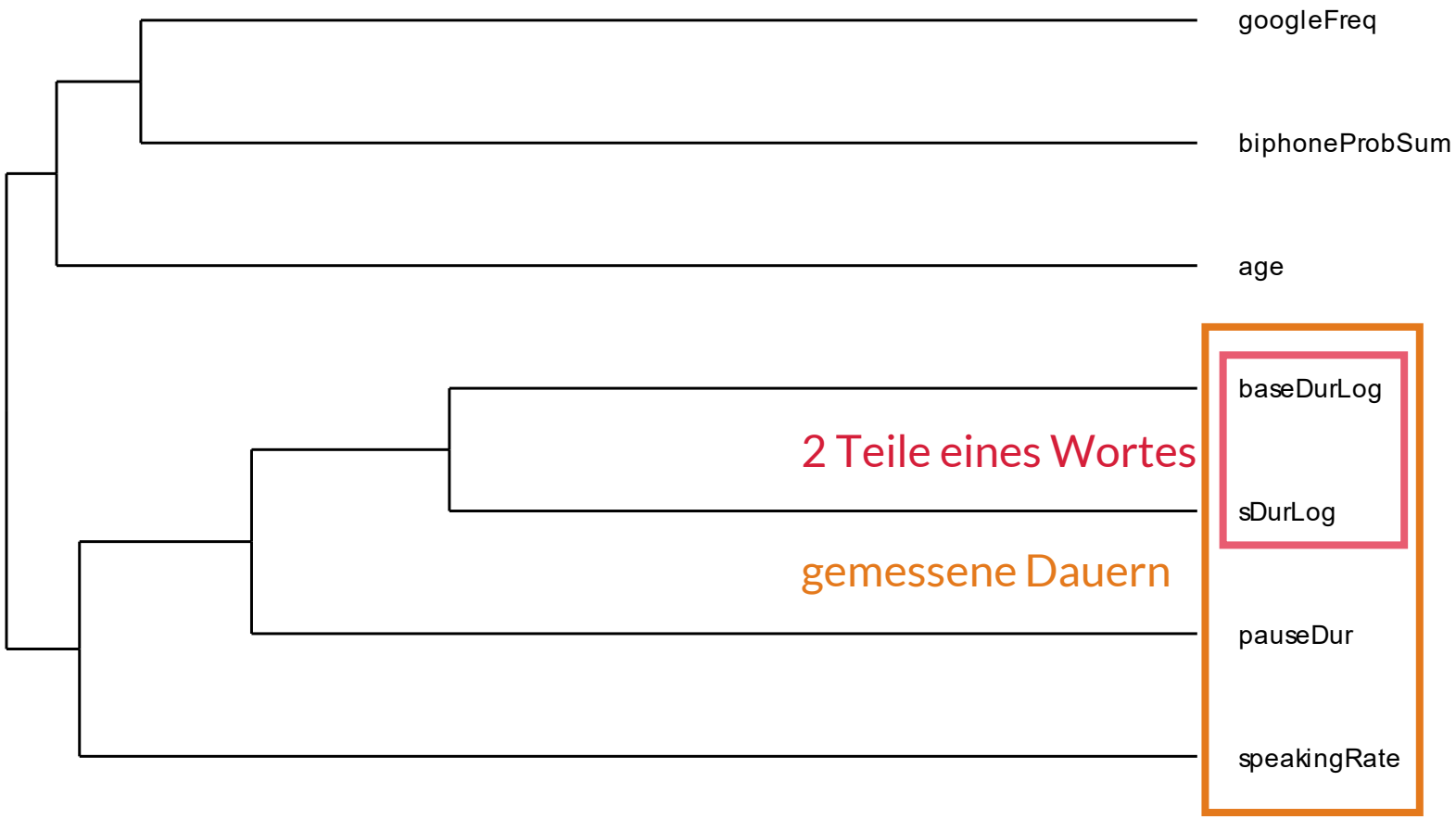
Hierarchical Cluster Analysis

Visualisierung



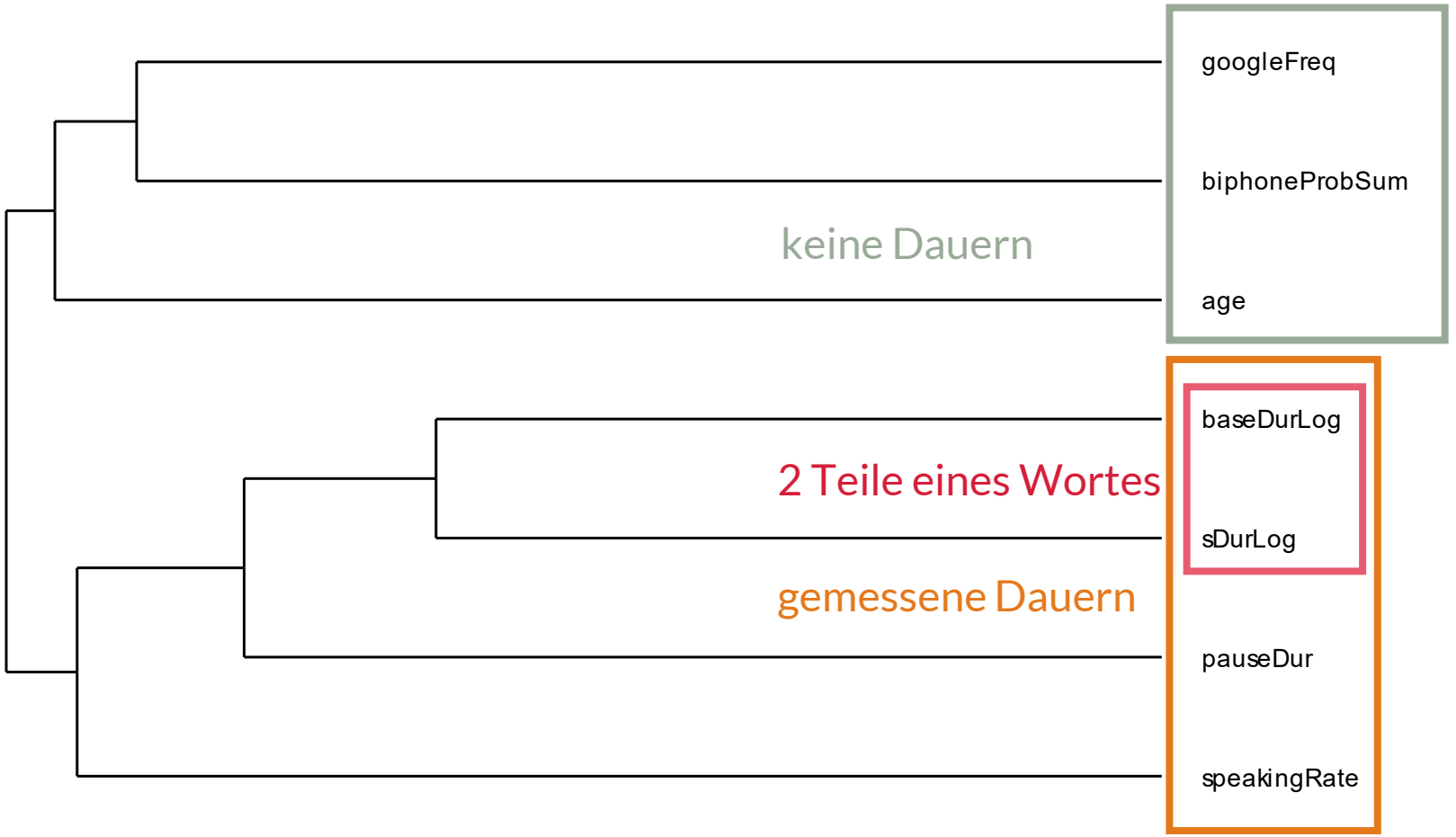
Hierarchical Cluster Analysis

Visualisierung



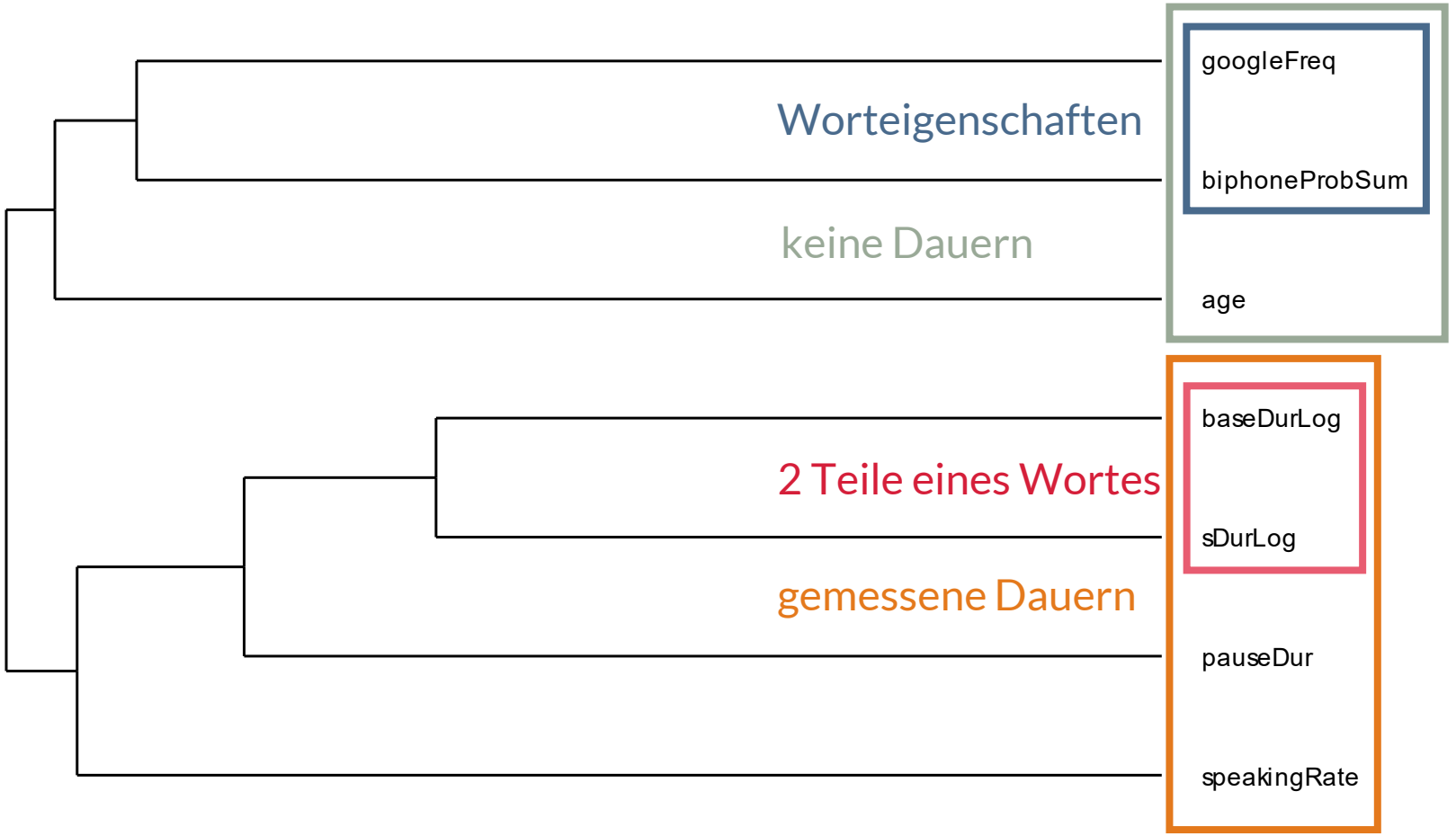
Hierarchical Cluster Analysis

Visualisierung



Hierarchical Cluster Analysis

Visualisierung



Clustering & Classification

Clustering

- das **unüberwachte Finden/Bestimmen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Multidimensional Scaling**, **Hierarchical Cluster Analysis**

Classification

- das **überwachte Überprüfen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ Classification Trees

Clustering & Classification

Clustering

- das **unüberwachte Finden/Bestimmen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Multidimensional Scaling**, **Hierarchical Cluster Analysis**

Classification

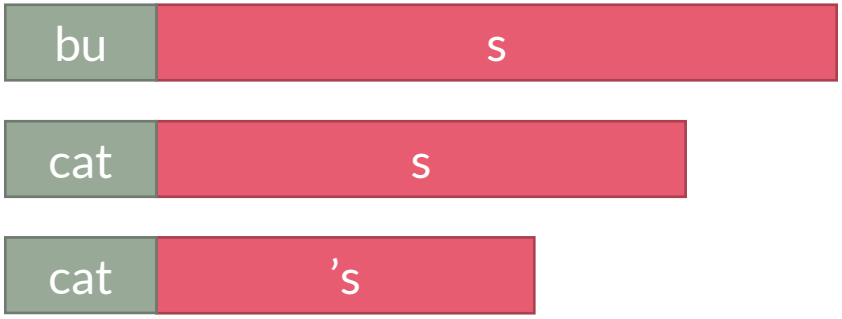
- das **überwachte Überprüfen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Classification Trees**

Classification Trees

- Methode zur Überprüfung von Mustern in einer Variable durch andere Variablen
- Für die folgenden Beispiele werden wir Daten folgender Studie nutzen:

The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords[1]



[1] Schmitz, D., Baer-Henney, D., & Plag, I. (2021). The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords. Phonetica, 78(5-6), 571-616. doi: 10.1515/phon-2021-2013

Classification Trees

Frage

Anhand welcher Variablen lässt sich die abhängige Variable am effizientesten vorhersagen?

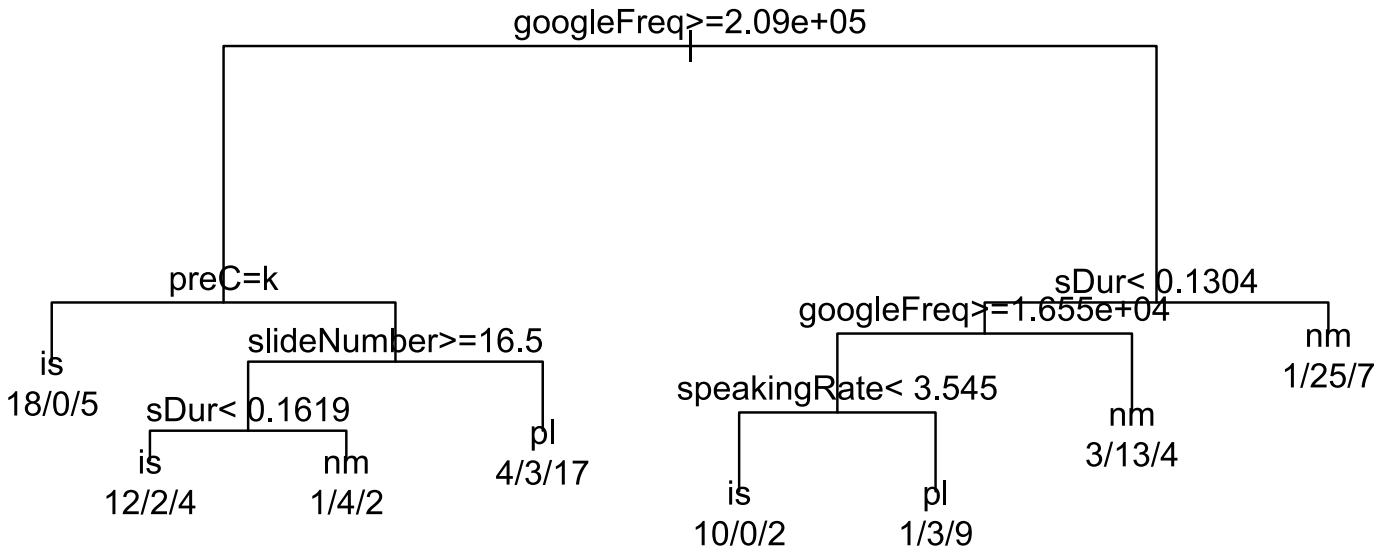
Methode

Kategorische Abhängige: Multiple ANOVAs

Kontinuierliche Abhängige: Lineare Regression

Classification Trees

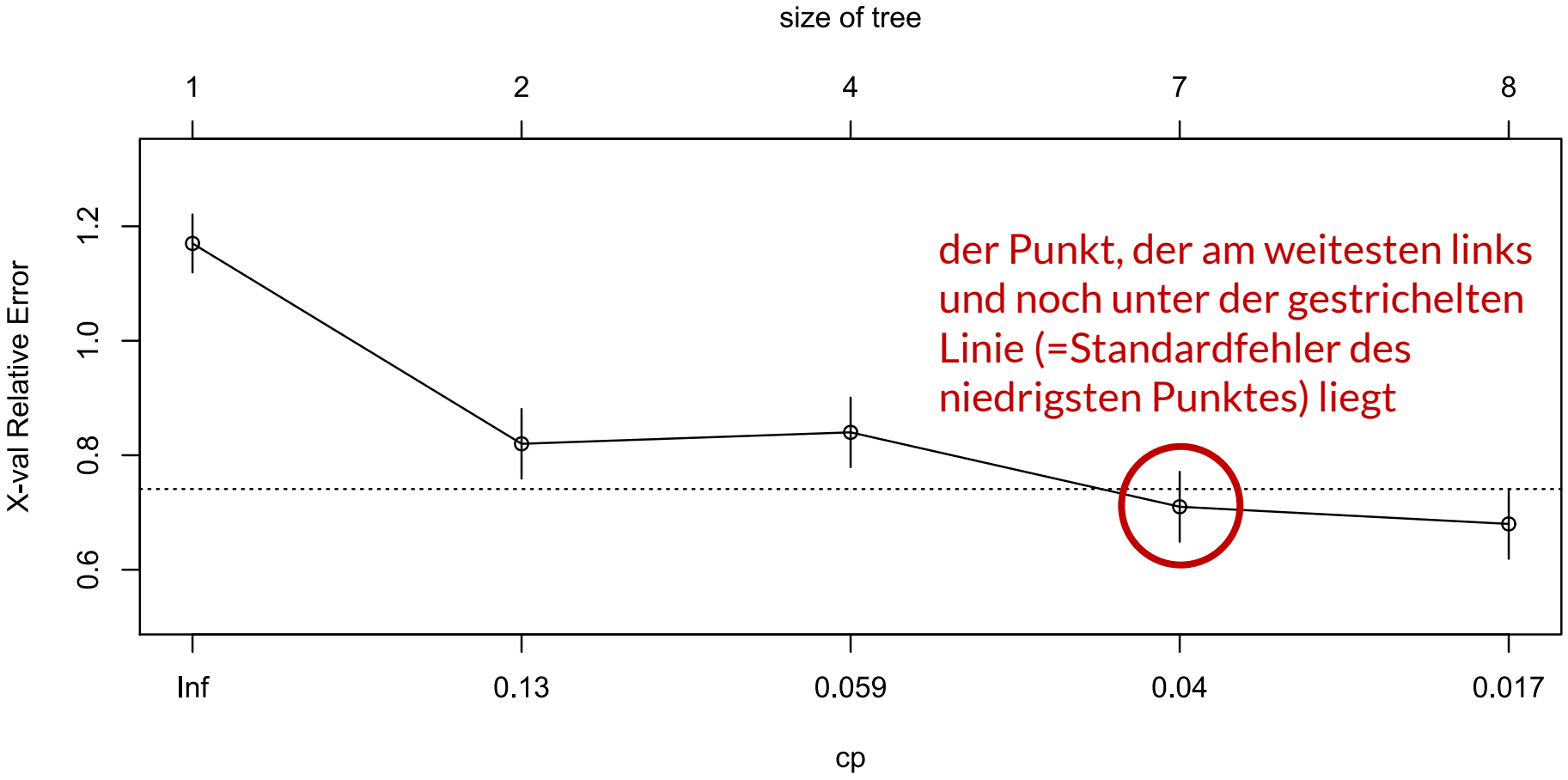
Schritt 1: Originaler Baum



Problem: Overfitting

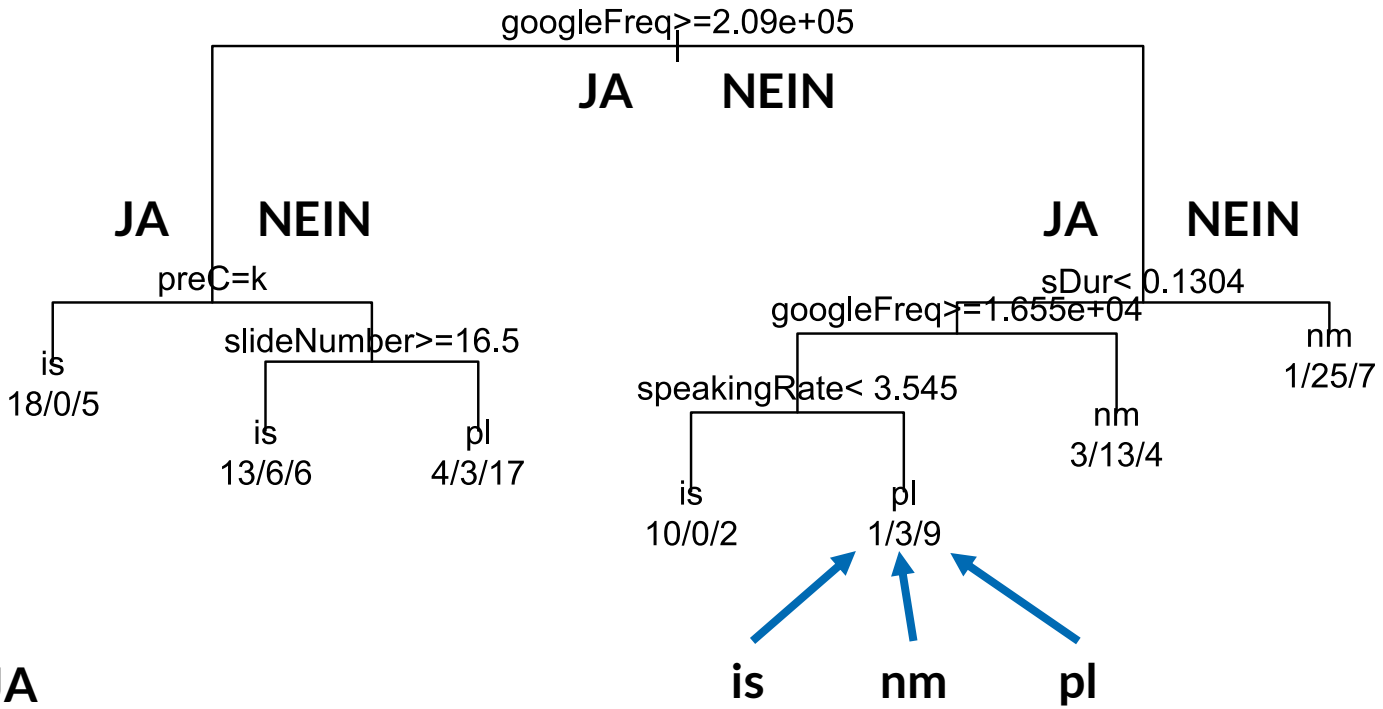
Classification Trees

Schritt 2: Stutzen – aber wo?



Classification Trees

Schritt 3: Gestutzter Baum



links = JA

rechts = NEIN

Classification Trees

Schritt 4: Fehlerraten berechnen

type of /s/	Fehlerrate
is-clitic	18 %
nicht-morphemisch	24 %
plural	48 %
Gesamt	30 %

Clustering & Classification

Clustering

- das **unüberwachte Finden/Bestimmen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Multidimensional Scaling**, **Hierarchical Cluster Analysis**

Classification

- das **überwachte Überprüfen** von Strukturen/Mustern in Daten anhand von Gruppierungen

→ **Classification Trees**