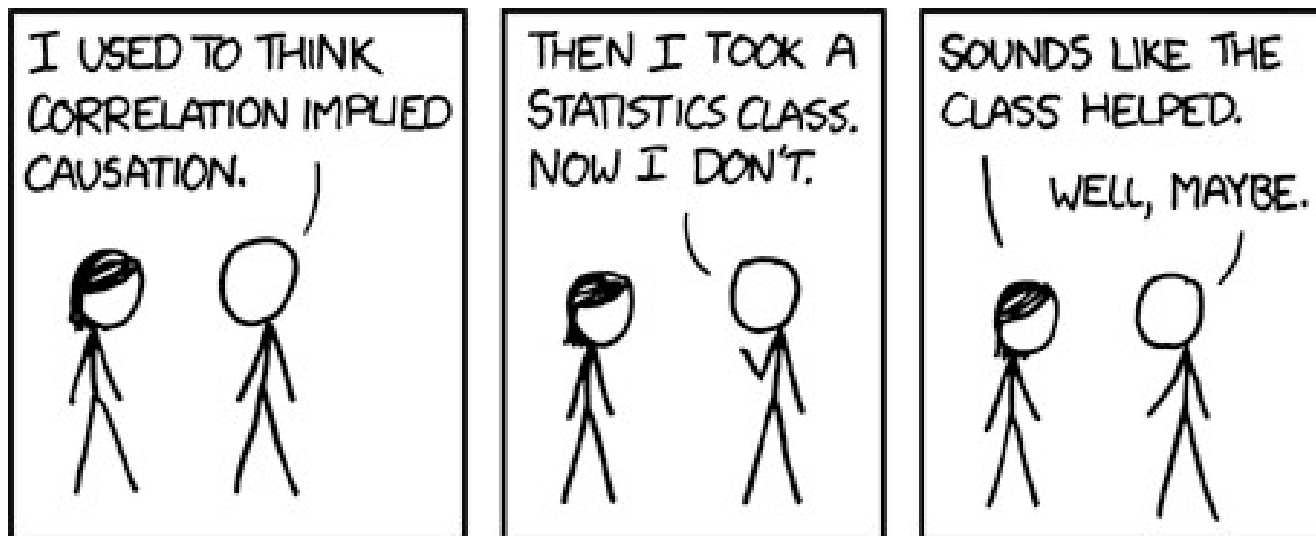


Kausalität in der Datenauswertung

Christoph Euler
13. Januar 2020



Datenauswertung: Statistische Aussagen, Vorhersagen und Kausalität

Datenauswertung

Statistische Aussagen

Aufgaben:

- Informationen über eine Population ermitteln
- Aussagen darüber ableiten, in welchem Verhältnis ein Datenpunkt zu allen anderen steht

Werkzeug:
Statistik

Vorhersagen

Aufgaben:

- Information über einen einzelnen (unbekannten) Datenpunkt
- Lernen über Daten, die nicht vorliegen, aber vorliegen könnten

Werkzeug:
Korrelation/Machine Learning

Kausalität

Aufgaben:

- Ursache-Wirkung-Beziehungen zwischen Attributen ermitteln
- Für Variablen X und Y ermitteln, ob eine kausale Wirkung von X auf Y existiert

Werkzeug: ?



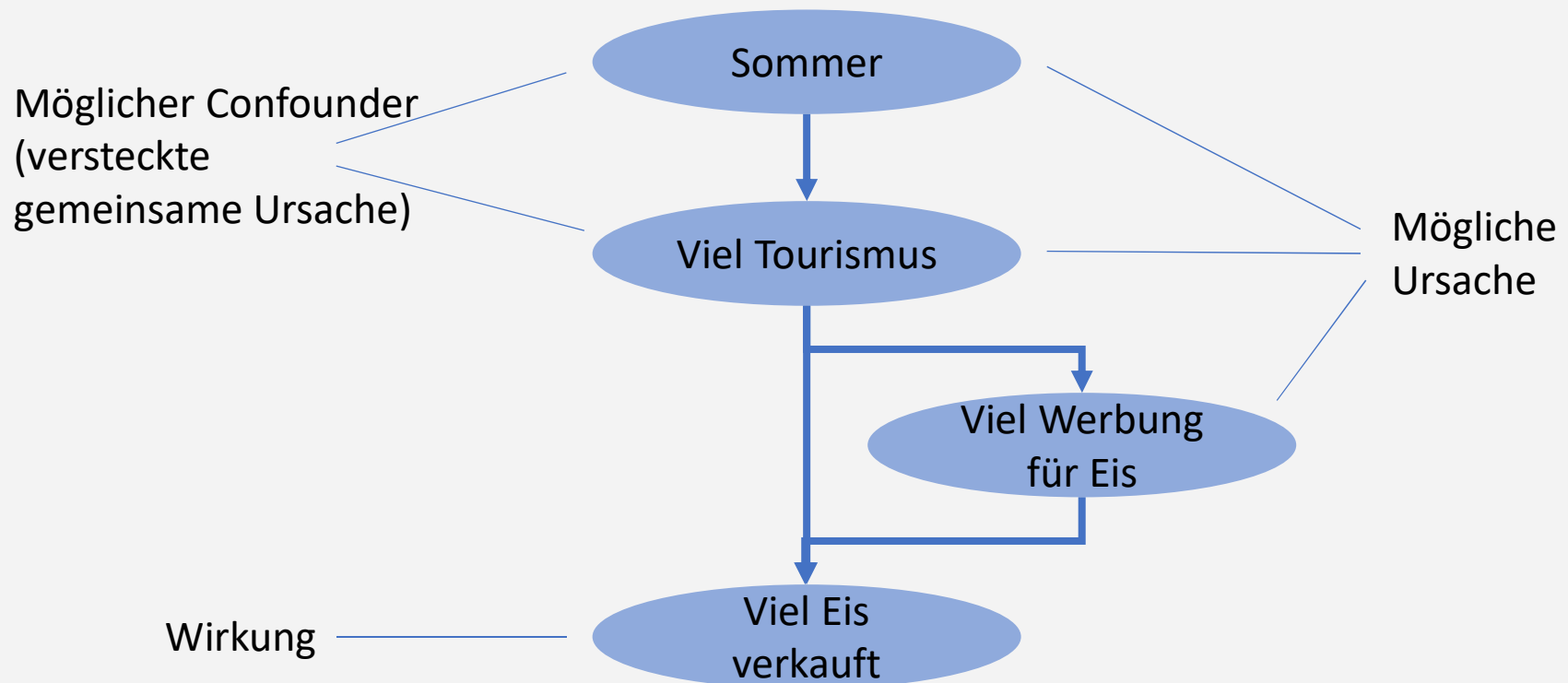
Ziel dieser Vorlesung: Einführung in Möglichkeiten, kausale Wirkungen zu beschreiben
Nicht Ziel der Vorlesung: Experimentelles Design

Spricht Sie Werbung für Eis an?



Quelle: <https://www.scoopon.com.au/deals/70390/-1-scoop-of-ice-cream-new-zealand-natural-bondi>, abgerufen am 12.1.2020 um 14:32.

Ursache und Wirkung grafisch dargestellt



Unterschied zw. Korrelation und Kausalität: Kausalität ist Ergebnis einer Intervention.
→ Wenn die Intervention entfernt wird, entfällt die Wirkung.

Kausalität: Unterschied in Y (Eis-Kauf) durch X (Werbung), der nicht eingetreten wäre, wenn X (Werbung) nicht passiert wäre

→ **Kausalen Effekt messen** als Differenz von Kauf mit Werbung und Kauf ohne Werbung

Idee für ein Experiment: (20.000 Touristen)

Gruppe 1: Sieht keine Werbung

Gruppe 2: Sieht Werbung

ID	Aktion	Kauf ohne Werbung	Kauf mit Werbung	Effekt*
1	Werbung	Nein*	Ja	Ja
2	Werbung	Nein*	Nein	Nein
3	Keine Werbung	Nein	Ja*	Ja
4	Keine Werbung	Ja	Nein*	Ja
5	Werbung	Ja*	Ja	Nein
...




Problem: Wir können nicht jede einzelne Person beobachten.

* Nicht beobachtbar („Counterfactual“) → Daten können prinzipiell nicht erhoben werden

Lösung für Counterfactuals-Problem: Betrachte Effekt der Werbung nur „im Mittel“ der Gruppen

Problem: Counterfactuals (niemand kann Daten erheben, die es nicht geben kann, da nicht jede einzelne Person beobachtbar ist)

Frage: Wie können Daten trotzdem genutzt werden?

 **Lösung:** Betrachte kausalen Effekt nur im Mittel der Stichprobe („durchschnittlicher Effekt“)

	Ohne Werbung	Mit Werbung	Differenz (M-O)
Verkauftes Eis	500/10.000 (5%)	540/10.000 (5,4%)	0,4%

Ist Werbung erfolgreich? – Ja!

Aufteilung der Stichprobe in Untergruppen kann problematisch sein

Aufgabe: Weiterreichende Analyse nach geringem / hohem Tourismusaufkommen

Frage: Hängt der Effekt der Werbung davon ab, wie viele Touristen in der Stadt sind?

Tourismus	Ohne Werbung	Mit Werbung	Differenz (M-O)
Gesamt	500/10.000 (5%)	540/10.000 (5,4%)	0,4%
Gering	100/4000 (2,5%)	40/2000 (2%)	-0,5%
Hoch	400/6000 (6,6%)	500/8000 (6,3%)	-0,3%

Verhältnis
"ohne" : "mit" = 1:1

Verhältnis
"ohne" : "mit" = 2:1

Verhältnis
"ohne" : "mit" = 3:4

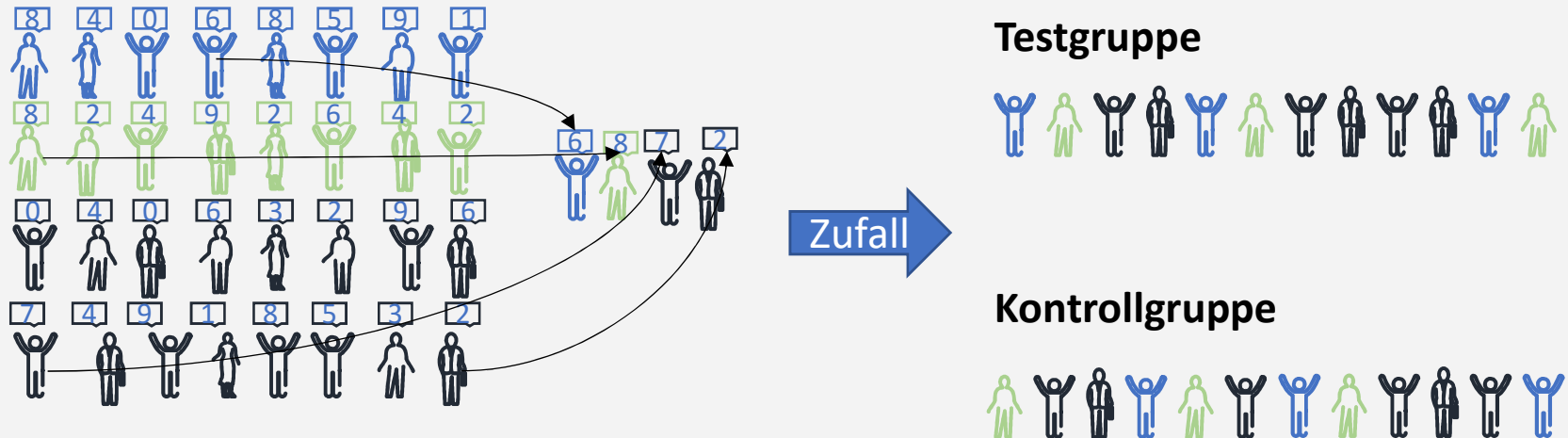
→ Mit diesen Daten keine Aussage zu Untergruppen möglich, da Verhältnisse verzerrt

→ Im Experiment müssen die zu untersuchenden (Unter-) Gruppen gleich behandelt werden: Verzerrung der (Unter-) Stichproben („Sampling Bias“) vermeiden



Verzernte Stichproben verfälschen kausale Aussagen.

Kausalität: Unterschied in Y durch X, der nicht eingetreten wäre, wenn X nicht geändert worden wäre **und alle anderen Faktoren gleich sind**



Problem: Verzerrte Stichprobe

Lösung: Der Mechanismus der Zuteilung von Werbung und der Effekt müssen unkorreliert sein!

! → Teile die Population in **zufällige Stichproben** ein, um den Sampling Bias zu vermeiden („**stratified sampling**“ mit gleichen Verhältnissen aller Untergruppen)

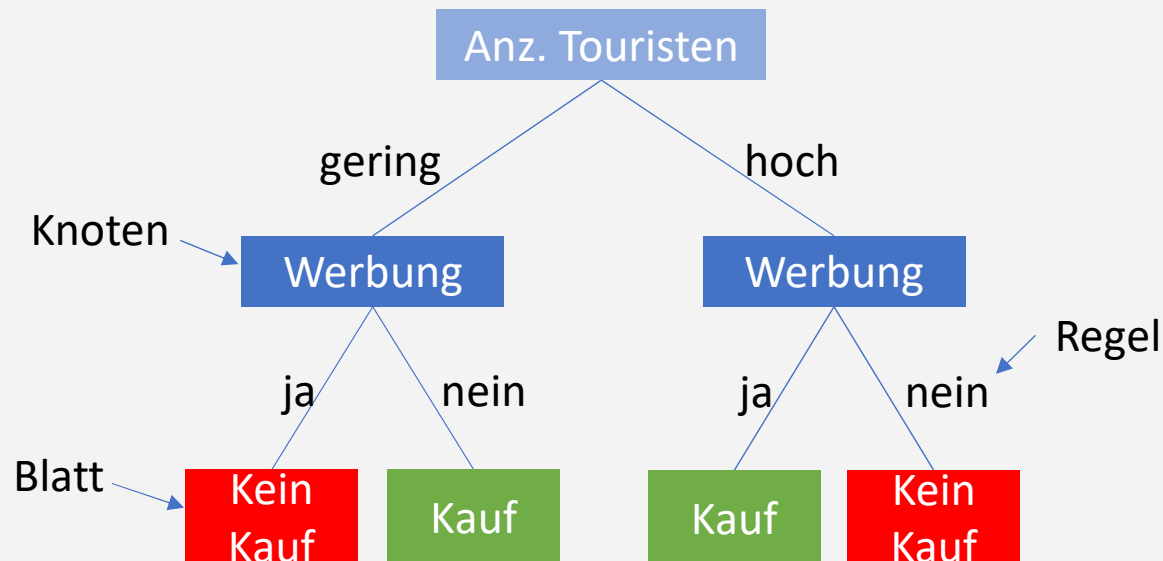
! In randomisierten kontrollierten Studien impliziert Korrelation Kausalität!
(Annahme: Keine Confounder, die nicht kontrolliert werden)

Entscheidungsbaum als Methode für Analysen bereits existierender Daten ohne randomisierte Studie

Aufgabe: Nutze historische Daten als „natürlich vorkommendes Experiment“

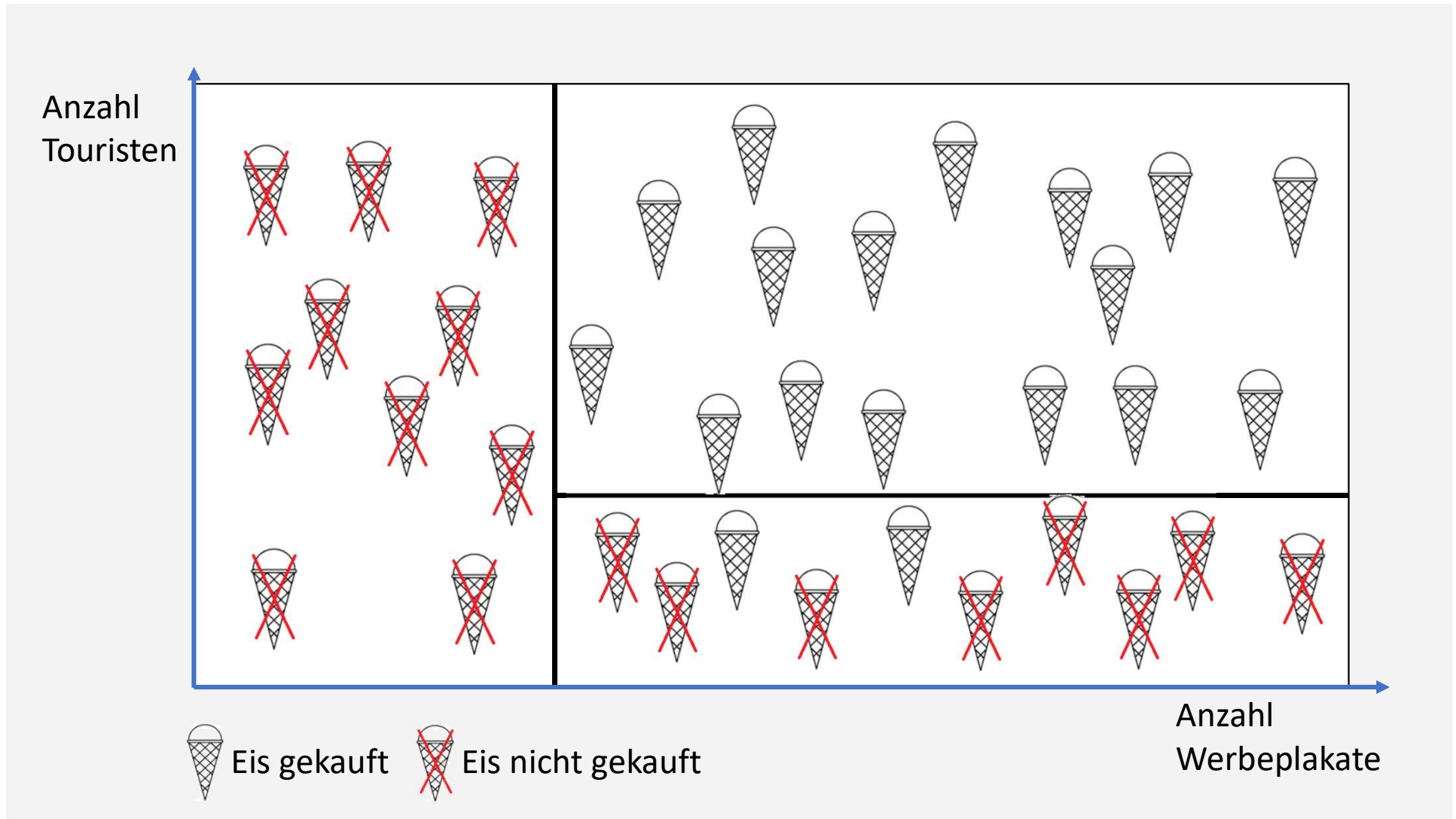
Frage: Gibt es einen kausalen Effekt zwischen Werbung und Kauf?

Annahme: Alle Confounder sind im Datensatz enthalten

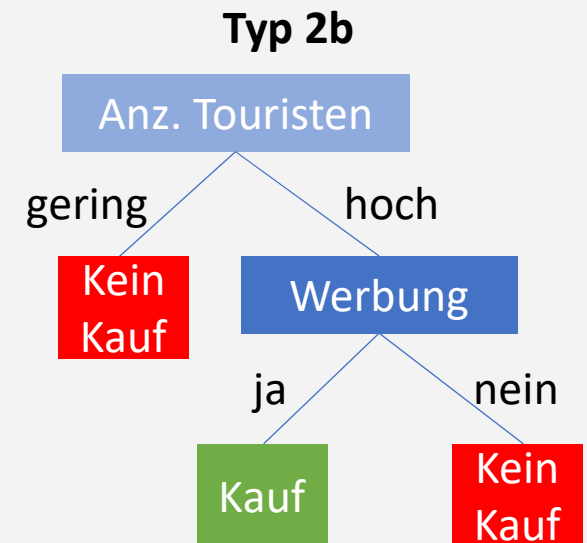
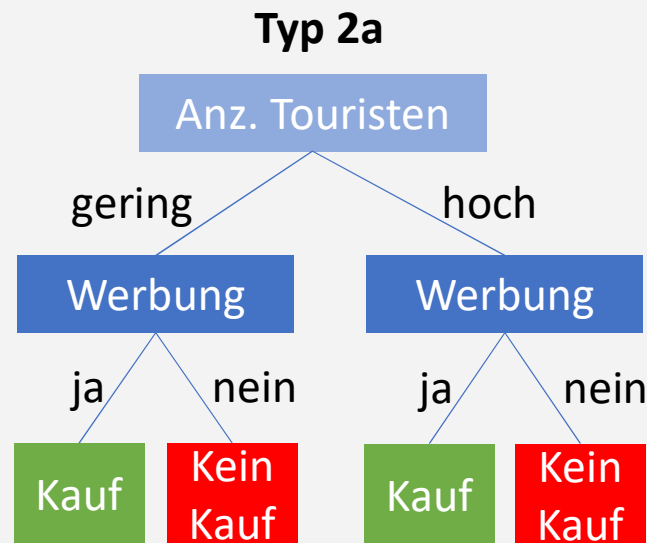
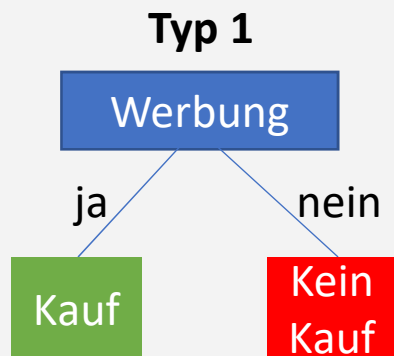


Ein Entscheidungsbaum macht keine Aussage zur Größe des kausalen Effekts.

Strategie eines Entscheidungsbaums: Schneide den Datensatz, um möglichst „sortenreine“ Bereiche zu erhalten



Entscheidungsbaum zur Untersuchung des Einflusses von Werbung (1/2)



Kausale Beziehung: ja

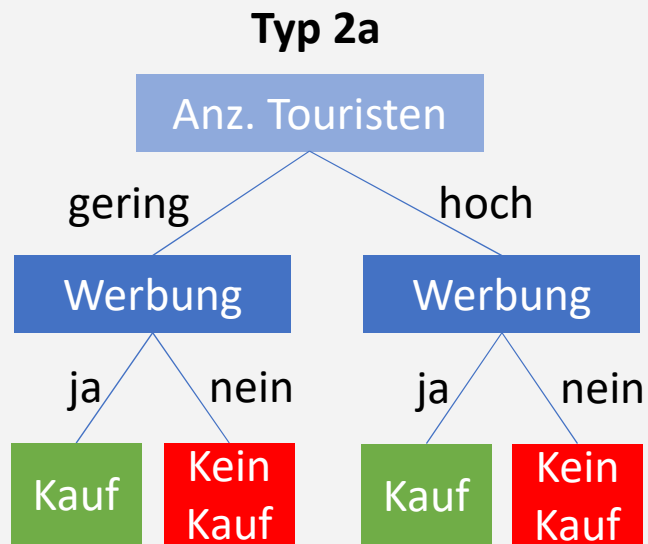
**Kausale Beziehung: ja,
sogar in Untergruppen:
Anz. Touristen hoch & gering**

**Kausale Beziehung: ja, aber
nur in der Untergruppe
Anz. Touristen = hoch**

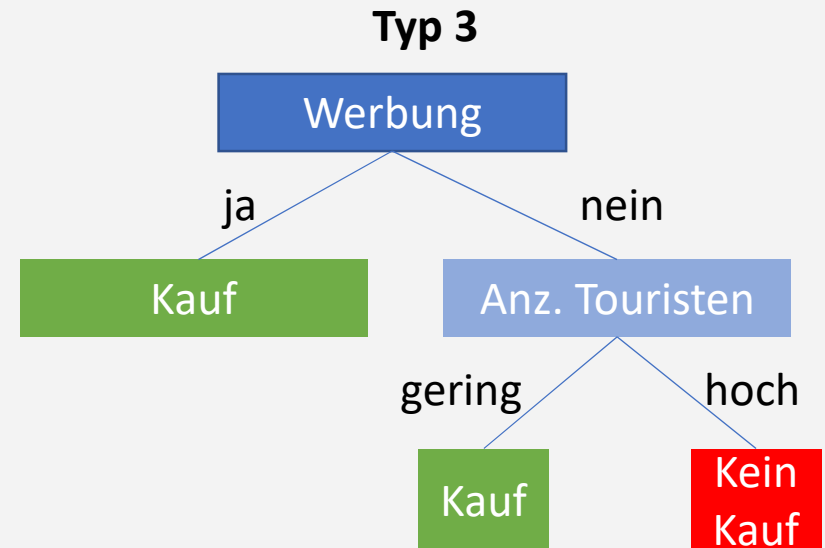


„Werbung“ ist letzter Knoten → Kausalität liegt vor

Entscheidungsbaum zur Untersuchung des Einflusses von Werbung (2/2)



**Kausale Beziehung: ja,
sogar in Untergruppen:
Anz. Touristen hoch & gering**




Kausale Beziehung: unklar



„Werbung“ nicht letzter Knoten → Kausale Beziehung unklar
Aber dass eine Variable nicht im Baum auftritt, schließt Kausalität nicht aus!

Praktische Umsetzung eines Entscheidungsbaums für kausale Interpretation: Conditional Inference Tree

Eigenschaft	Klassischer Entscheidungsbaum	Conditional Inference Tree
Algorithmus	CART, ID3, ...	
Funktionsweise	<ol style="list-style-type: none"> 1. Kriterium im Datensatz berechnen 2. Datensatz an einer Variable aufteilen, sodass Krit. reduziert ist 3. Wiederholen, bis Konvergenz-Kriterium erfüllt ist 	
Knoten einfügen, wenn...	...Unterschied in der Varianz (oder anderem Kriterium) ausreichend groß ist	...die Mengen in den beiden erzeugten Blättern signifikant unterschiedlich sind
Umsetzung in R (Beispiel) 	<pre>library(rpart) mdl <- rpart(class~., data=df)</pre>	<pre>library(partykit) mdl <- ctree(class~., data=df)</pre>
Hyperparameter (Auswahl)	cp (Pruning-Parameter)	Signifikanzniveau (z.B. 0,95) für Permutations-Test



Ein Conditional Inference Tree erzeugt einen Baum auf Basis eines Signifikanztests

Praktische Umsetzung in R: Conditional Inference Tree



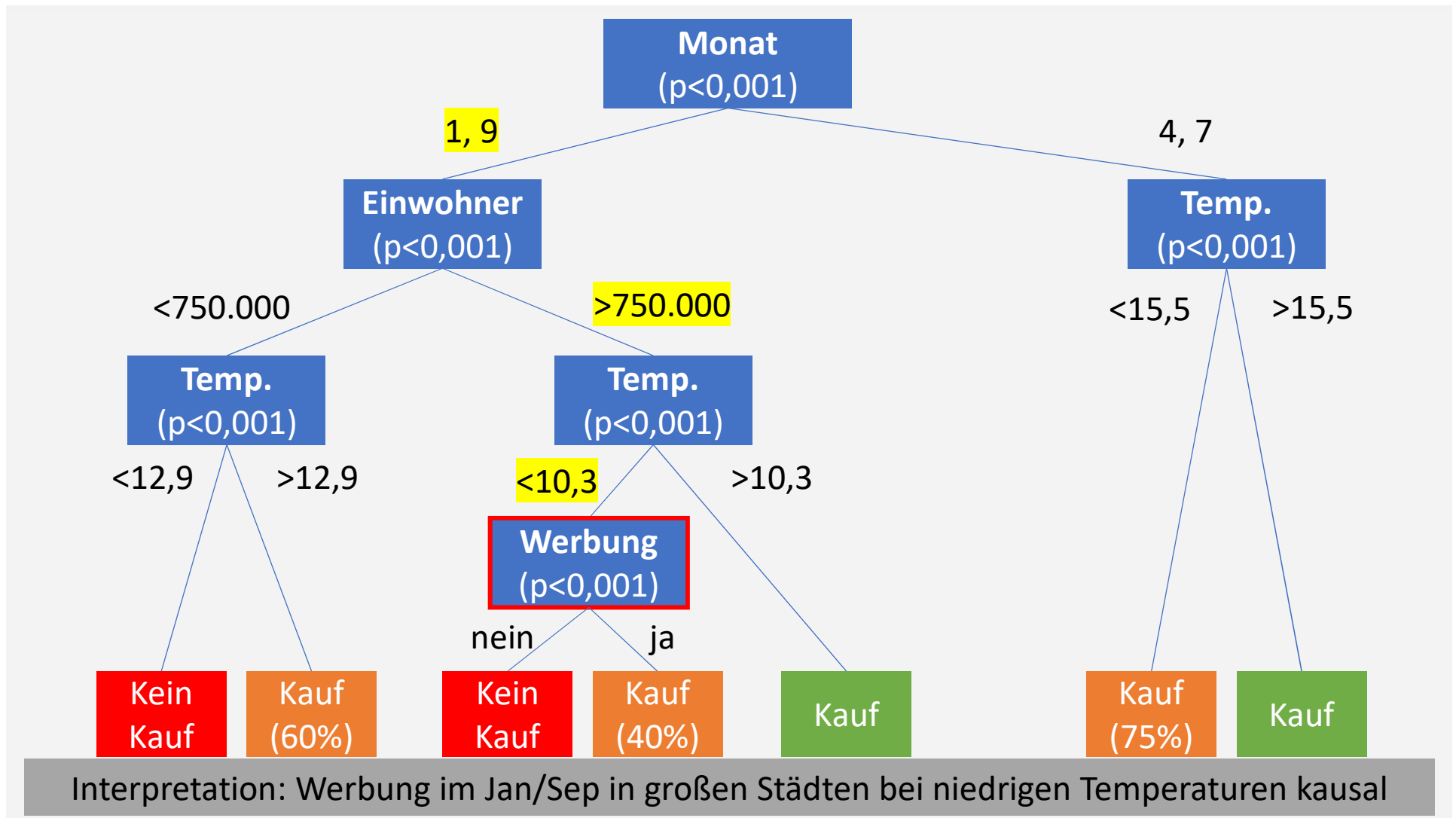
```
library(partykit)
df <- read.csv("kausalitaet/data.csv")

# Daten vorbereiten
# Variable: Kauf, Werbung, Temperatur, Monat, Einwohner
# kauf und werbung als Faktor-Variable (binär, nicht 1>0)
df$kauf <- as.factor(df$kauf)
df$werbung <- as.factor(df$werbung)

# Modell trainieren
# (Random Seed für Reproduzierbarkeit, Signifikanzniveau 95%)
set.seed(-753)
mdl <- ctree(kauf~., data=df,
             control = ctree_control(mincriterion = 0.95))

# Baum ausgeben
plot(mdl)
```

Visualisierung eines Conditional Inference Tree



Zusammenfassung: Kausalität in der Datenauswertung

Definition: Unterschied in Y durch X, der nicht eingetreten wäre, wenn X nicht geändert worden wäre und alle anderen Faktoren gleich sind (→ Intervention)

Randomized Controlled Trial

Eigenschaften:

- Zufällige Aufteilung in Test- und Kontrollgruppe → kein Sampling Bias
- Ausgleich von Counterfactuals
- Korrelation impliziert Kausalität!

Conditional Inference Tree (ctree in R)

Eigenschaften:

- ML-erzeugtes grafisches Modell
- Zu untersuchende Größe: letzter Knoten
- Nichtauftreten einer Variable schließt Kausalität nicht aus
- Keine Aussage über Größe des Effekts

Folien, Daten und Code: <https://github.com/ChristophEuler/FrankfurtUAS>



1. Kausale Zusammenhänge sind nur eine Interpretation von Korrelation.
2. Bedingungen dazu: Kontrollierte Randomisierung und Abwesenheit von Confoundern.