Dr. Lars Pelke

Universität Heidelberg, Institut für Politische Wissenschaft

Praxiskurs Datenanalyse und Replikation

Sitzung 3 Auswahl einer geeigneten Studie zur Replikation

Leitfragen und Lernziele

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Suchen nach geeigneten quantifizierenden und bereits publizierten Studien
  - Daten und Softwareskripte zugänglich?
  - Verwendeten Methoden bekannt oder während des Seminars erlernbar?
- Code und Daten herunterladen und übersichtliche Ordnerstruktur erstellen
- Erste Replikationsanalyse: Kommt das raus, was im Paper steht?

Auswahlkriterien Originalstudie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Wählen Sie Paper nur dann aus, wenn die Replikationsmaterialen (Daten, Skripte) über die Zeitschriftenwebsite, Harvard Dataverse, Github etc. öffentlich verfügbar sind.
  - Das sollte aber mittlerweile Standard sein, ist es aber leider nicht immer.
- Um die Wahrscheinlichkeit zu erhöhen, dass Ihre Replikationsstudie ggf. publiziert wird, wählen Sie ein Paper aus, welches in einem *hochrangigen Journal* veröffentlicht worden ist.
- Das Paper sollte nicht älter als 2015 sein.
- Das Paper sollte statistische Methoden verwenden, die Sie bereits erlernt haben oder während der Seminarlaufzeit erlernen können!

Auswahlkriterien Originalstudie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Typische statistische Verfahren, die ggf. kurzfristig erlernbar sind

- Lineare Regressionsverfahren (OLS estimator)
- Generalized linear models (GLS estimator)
- Regressionsverfahren für binäre AVs (logit/probit estimators)
- Panelregressionsmodelle (OLS 2WFE estimators)

Kritisch:

- Multilevel regression models (linear and non-linear)
- Elaborierte Panelregressionsverfahren

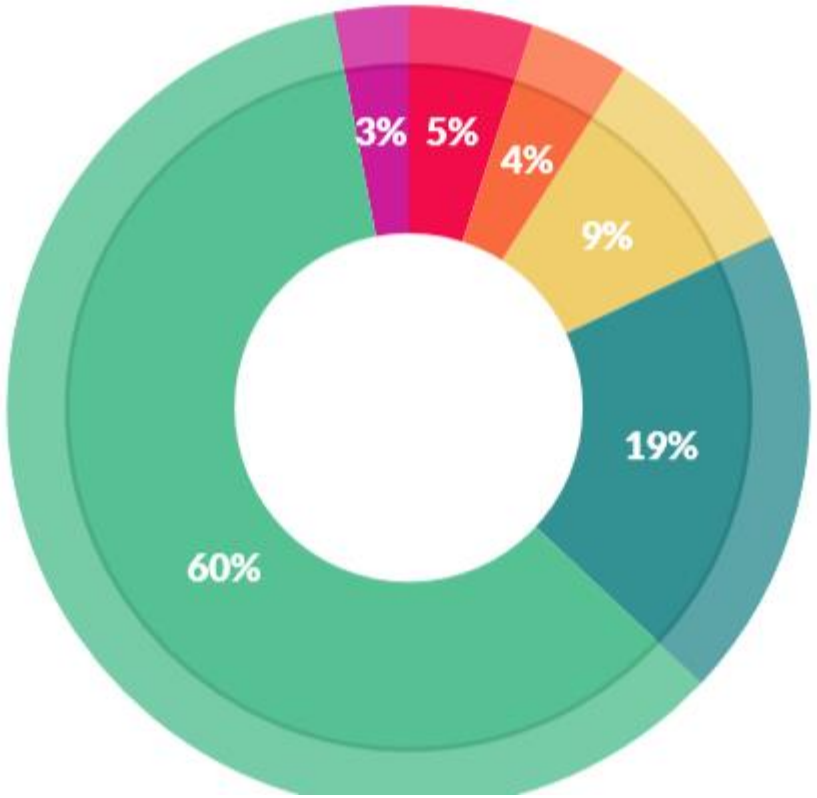
Partnerarbeitsphase

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

ca. 20min

- Erklären Sie Ihrer:m Partner:in, warum Sie die Studie ausgewählt haben
- Fassen Sie im Gespräch die Hauptbefunde zusammen und klären Sie welchen Befund Sie replizieren möchten
- Welche Herausforderungen erwarten Sie bei der Replikation?
- Schätzen Sie, wie viel Zeit die einzelnen Arbeitsschritte benötigen werden.

80/20 rule



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

https://qph.fs.quoracdn.net/main-qimg-c2bea0043c73e556da85f49547338de2

Einzelarbeitsphase

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Originalstudie suchen und Zugang zu Daten und Skripten Prüfen

- Laden Sie nun die Originalstudie herunter und prüfen Sie den Zugang zu den Daten und Softwareskripten
  - Legen Sie die Daten wie in der vorherigen Sitzung besprochen ab und verändern Sie nie die Originaldaten
  - Legen Sie die Skripte wie in der vorherigen Sitzung besprochen ab und verändern Sie diese nicht
- Prüfen Sie nun ob die Skripte laufen? Das wird vermutlich einige Zeit in Anspruch nehmen. Funktioniert das nicht -> folgenden Sie der Checkliste.

Checkliste Skripte zum Laufen bringen

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Alle *library/add-ons* installiert?
- *working directory* definiert und ggf. angepasst?
- Sind alle Ordner vorhanden, die das Skript ansteuert?
- Softwareversionen vergleichen (Eigene versus Originalautor:innen)
- Aufspüren, wo der Fehler passiert! Was sagt die Fehlermeldung?
  - Herausfinden warum ein Fehler passiert:
  - Fehlermeldung bei Stackoverflow suchen
  - Google „rstats“ und Fehlermeldung

Wrap-Up: Wo sind wir?

Kurzfeedback Visualisierung