

ANGEWANDTE
COMPUTER- UND
BIOWISSENSCHAFTEN



SOPHIE FRIEDL

PROZESSKETTE – BUCHKLAPPENTEXTE (COARSE)

PROZESSKETTE – BUCHKLAPPENTEXTE (COARSE)

ORGANISATION UND PLANUNG

ORGANISATION & PLANUNG

Festgelegte Entscheidungen:

- Verwaltungssoftware: GitHub
- Projektleiter: Michelle
- Klassifikator: Nearest Zentroid & Random Forrest
 - Erstellung von Wörterbüchern von verschiedenen Genres
 - Falls beide Klassifikatoren verschiedene Ergebnisse liefern, dann im Wörterbuch nachschauen und daran entscheiden

Offene Entscheidungen:

- Rekursive Klassifikation ?

PROZESSKETTE – BUCHKLAPPENTEXTE (COARSE)

DERZEITIGER STAND

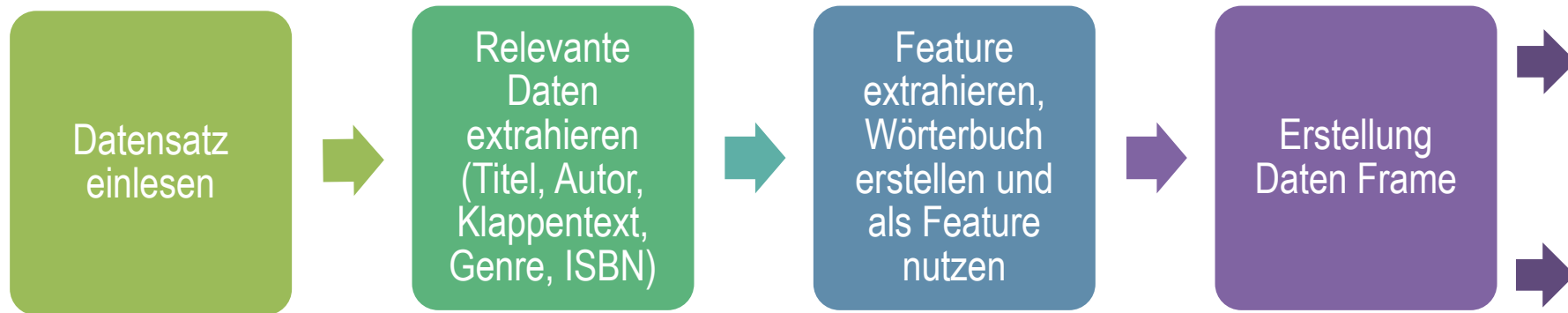
DERZEITIGER STAND

- Erstellung Testdatensatz (ersten 2 Klappentexte)
- Einlesen der Datei
- Klappentexte, Autor, ISBN, Genre herausfiltern
- Erste Features extrahiert:
 - Wie viele Wörter sind in einem Satz ? (Durchschnitt)
 - Wie viele Sätze
 - Relative Häufigkeit von :
 - Nomen
 - Verben
 - Adjektiven
 - Anzahl Kommas und Symbole ! (erste Schwierigkeiten)
- Tools:
 - Textblob.de (Natural Language Toolkit)
 - Pandas
 - SciKit-Learn

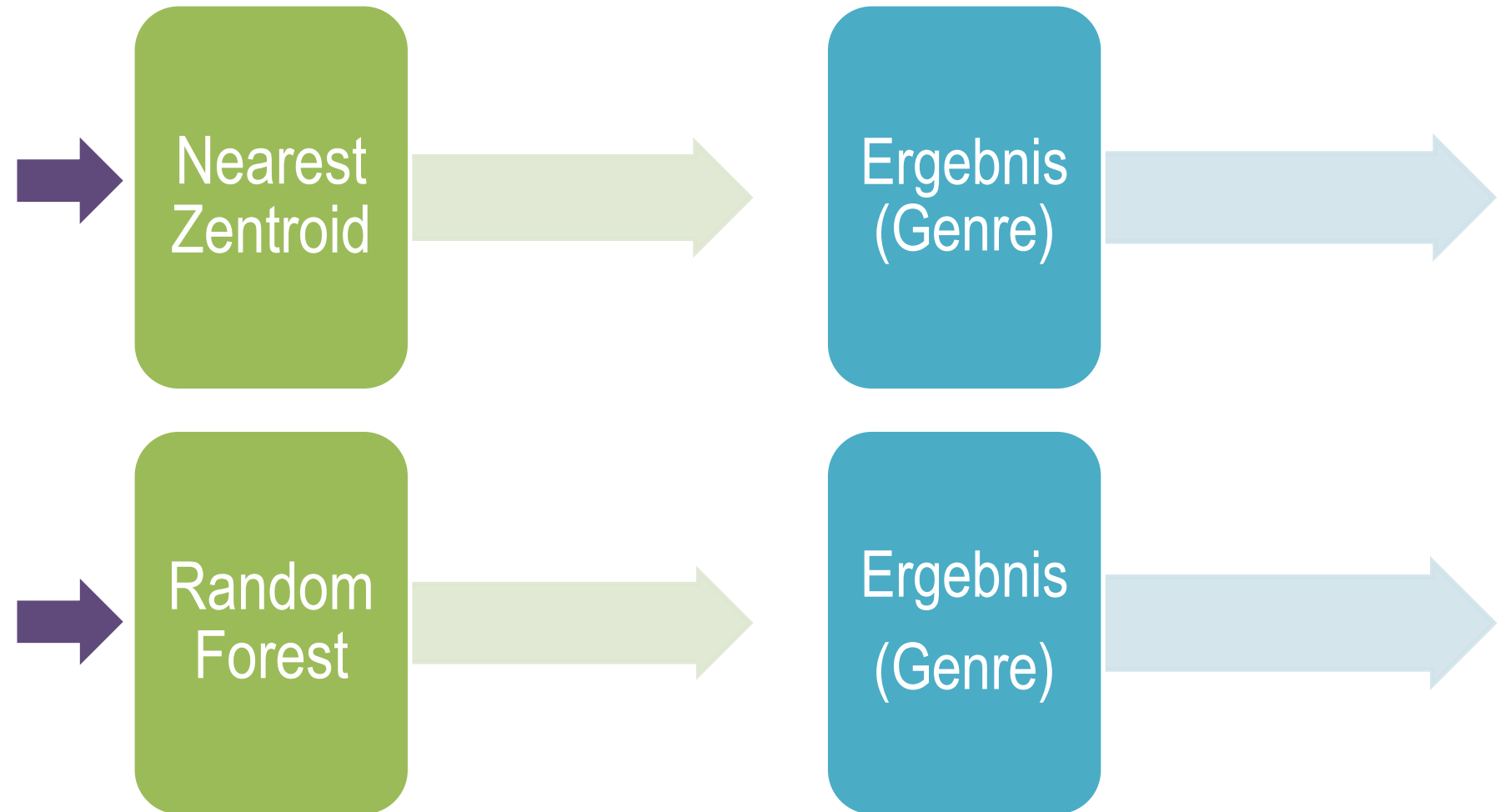
PROZESSKETTE – BUCHKLAPPENTEXTE (COARSE)

PROZESSKETTE

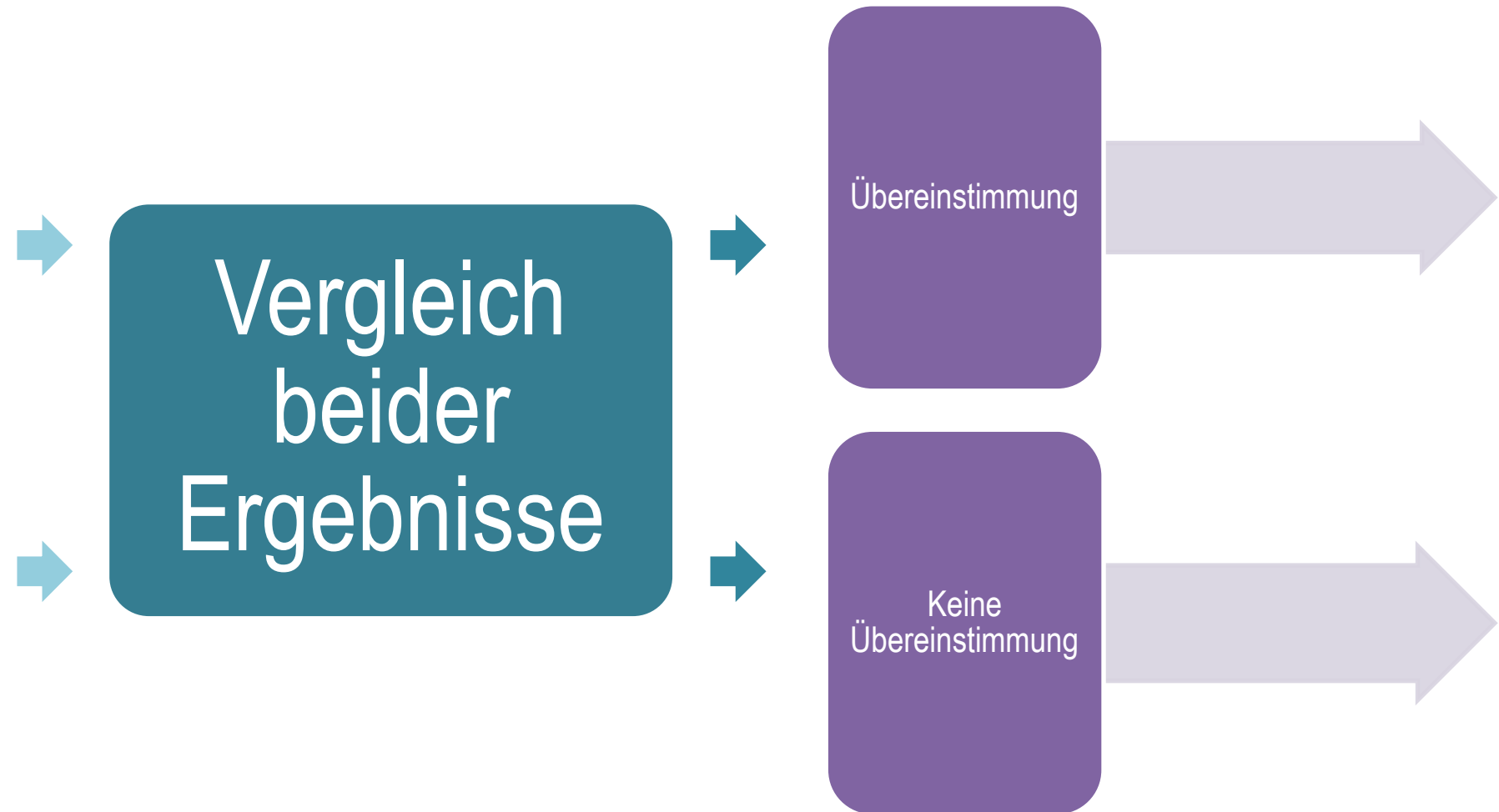
PROZESSKETTE



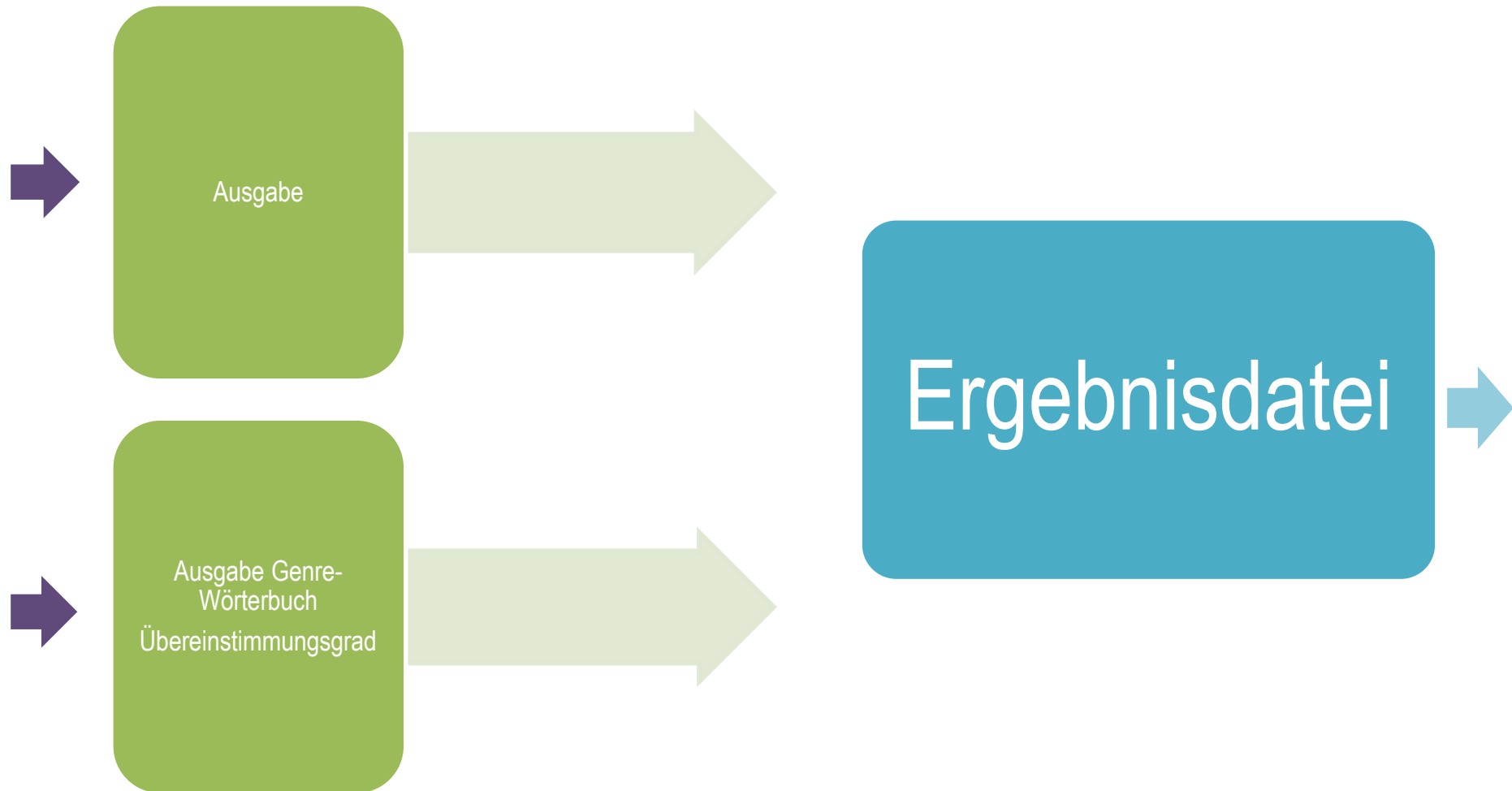
PROZESSKETTE




PROZESSKETTE



PROZESSKETTE



PROZESSKETTE



Evaluierung mit
Evaluierungsdatensatz
(F-Maß und 10-fache
Kreuzvalidierung)



Paper schreiben

PROZESSKETTE – BUCHKLAPPENTEXTE (COARSE)

**VIELEN DANK FÜR IHRE
AUFMERKSAMKEIT !**