

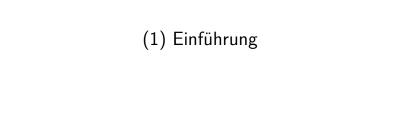
# Programmierung und Deskriptive Statistik

BSc Psychologie WiSe 2023/24

#### Belinda Fleischmann



Aufnahme läuft.



#### Lehrstuhlwebsite

### Homepage der Abteilung für Methodenlehre I



Formalia

Grundbegriffe der Informatik

R und Visual Studio Code

Grundbegriffe der Informatik

R und Visual Studio Code

# Modul C: Einführung in empirisch-wissenschaftliches Arbeiten

#### C1. Programmierung und Deskriptive Statistik

- Einführung in die datenanalytische Programmierung
- Einführung in die Auswertung deskriptiver Statistiken mit R in Visual Studio Code

### C2. Design, Analyse, Dokumentation

- Einführung in das Design quantitativer empirischer Studien
- Einführung in die Analyse und Dokumentation emirischer Daten

# Organisatorisches

- Termine: Mittwochs in Raum G05-307
  - Gruppe 1 um 11-13 Uhr
  - Gruppe 2 um 13-15 Uhr
- Kursmaterialien (Folien, Videos, RMarkdown Code) auf der Kurswebseite
- Code auf Github
- Ankündigungen über die Moodleseite
- Vorherige Iteration des Kurses: PDS (WS 2023/23)
- Empfohlene Vorbereitung: Vorkurs "Mathematische Grundlagen"
- Q&A im Mattermost-Channel
  - Einmalige Registrierung zum Team "bsc-psy-2023"
- Zweiteiliger Leistungsnachweis: Unbenotet, Multiple Choice, digitales Format
  - Teil 1 vor der Weihnachtspause
  - Teil 2 am Semesterende

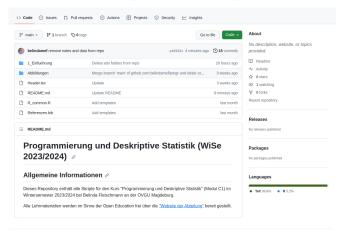
# Termine

Datum	Einheit	Thema	
11.10.23	Einführung	(1) Einführung	
18.10.23	R Grundlagen	(2) R und Visial Studio Code	
25.10.23	R Grundlagen	(2) R und Visial Studio Code	
01.11.23	R Grundlagen	(3) Vektoren	
08.11.23	R Grundlagen	(4) Matrizen und Arrays	
15.11.23	R Grundlagen	(5) Listen und Dataframes	
22.11.23	R Grundlagen	(6) Datenmanagement	
39.11.23	R Grundlagen	(7) Häufigkeitsverteilungen	
06.12.23	R Grundlagen	(8) Verteilungsfunktionen und Quantile	
13.12.23	Deskriptive Statistik (9) Maße der zentralen Tendenz		
20.12.23	Deskriptive Statistik (10) Maße der Datenvariabilität		
20.12.23	Leistunsnachweis Teil 1		
	Weihnachtspause		
10.01.24	Deskriptive Statistik	(11) Anwendungsbeispiel (Deskriptive Statistik)	
17.01.24	Inferenzstatistik (12) Anwendungsbeispiel (Parameterschätzung, Konfidenzintervalle		
24.01.24	Inferenzstatistik	(13) Anwendungsbeispiel (Hypothesentest)	
25.01.24	Leistunsnachweis Teil 2		

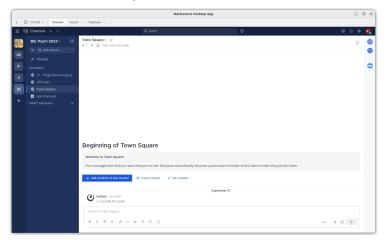
### Webseite des Kurses (Folien, Videos)



### Git-repository des Kurses (Folien, Source Codes)



### Mattermost-Team BSc Psych 2023





# Grundbegriffe der Informatik

R und Visual Studio Code

# Zentrale Komponenten der Datenwissenschaft



## Grundbegriffe der Informatik

- Datenanalyse
- Informatik
- Rechnerarchitektur
- Algorithmen und Programme

R und Visual Studio Code

Grundbegriffe der Informatik

- Datenanalyse
- Informatik
- Rechnerarchitektur
- Algorithmen und Programme

R und Visual Studio Code

### Datenanalyse - Überblick

- Wissenschaftliche Daten liegen heutzutage als digitale Daten vor.
- Digitale Daten werden mit Hilfe eines Computers analysiert.
- Zur Analyse von digitalen Daten schreibt man Computerprogramme.
- Diese Computerprogramme heißen Datenanalyseskripte.

## Struktur computergestützter Datenanalyse

- 1. Einlesen und Bereinigen eines digitalen Datensatzes.
- 2. Berechnung und Visualisierung deskriptiver Statistiken.
- 3. Probabilistische Modellierung und Inferenz.
- 4. Dokumentation und Präsentation der Ergebnisse.

## Typische Werkzeuge zur Analyse psychologischer Daten

- R (frei, Datenwissenschaft, Statistik, Psychologie)
- Python (frei, Datenwissenschaft, Anwendung)
- Matlab (kommerziell, Engineering, Neuroimaging)

#### Altmodisch

- SPSS (kommerziell, Sozialwissenschaften, Psychologie)
- JMP (kommerziell, Biologie, Psychologie)
- STATA (kommerziell, Wirtschaftswissenschaften)

# Programmiersprachen Trends

### PYPL Index (Stand: September 2023)

Worldwide, Sept 2023 :				
Rank	Change	Language	Share	1-year trend
1		Python	27.99 %	+0.1 %
2		Java	15.9 %	-1.1 %
3		JavaScript	9.36 %	-0.1 %
4		C#	6.67 %	-0.4 %
5		C/C++	6.54 %	+0.3 %
6		PHP	4.91 %	-0.4 %
7		R	4.4 %	+0.2 %
8		TypeScript	3.04 %	+0.2 %
9	ተተ	Swift	2.64 %	+0.6 %
10		Objective-C	2.15 %	+0.1 %
11	ተተ	Rust	2.12 %	+0.5 %
12	$\downarrow \downarrow \downarrow \downarrow$	Go	2.0 %	-0.1 %
13	<b>4</b>	Kotlin	1.78 %	-0.0 %
14		Matlab	1.58 %	+0.1 %
15		Ruby	1.05 %	-0.1 %

- PopularitY of Programming Language
- Basierend auf Googlesuchanfragen zu Programmiersprachentutorials

## Datenanalyseskripte

- Dokumentation aller Schritte von Rohdaten bis zur Datenvisualisierung.
- Reproduktion wissenschaftlicher Ergebnisse durch Dritte.
- Essentieller Teil wissenschaftlicher Publikationen.
- Essentieller Teil täglicher wissenschaftlicher Arbeit.

# Datenanalyse - Zusammenfassung

- Die Digitalisierung betrifft insbesondere auch die Wissenschaft.
- Forschungsdatenmanagement ist eine akute Herausforderung.
- Programmierung als zentrales Handwerkszeug wissenschaftlicher Arbeit.
- Informatikkenntnisse sind in der Arbeitswelt unverzichtbar.
- Dies gilt auch für Psychotherapeut:innen (z.B. Online-Intervention).

Grundbegriffe der Informatik

- Datenanalyse
- Informatik
- Rechnerarchitektur
- Algorithmen und Programme

R und Visual Studio Code

# Informatik (engl. Computer Science)

Bei der Informatik handelt es sich um die Wissenschaft von der systematischen Darstellung, Speicherung, Verarbeitung und Übertragung von Informationen, wobei besonders die automatische Verarbeitung mit Computern betrachtet wird. Sie ist zugleich Grundlagen- und Formalwissenschaft als auch Ingenieurdisziplin.

Wikipedia

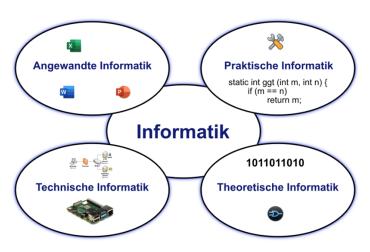
## Zentrale Komponenten der Informatik

## Computer

- Maschinen zum Datenspeichern und Ausführen einfacher Datenoperationen.
- Einfache Operationen mit extrem hoher Geschwindigkeit.
- Universalität durch Speicherung von Daten und Programmen.

### Algorithmen und Programme

- Programme sind in einer Programmiersprache verfasste Algorithmen.
- Algorithmen sind Folgen von Anweisungen durchzuführender Operationen.
- Bei Algorithmen unterscheidet man
  - Beschreibung (Kochrezept, IKEA Bauanleitung, R Skript)
  - Anweisungen ("Mehl und Wasser vermengen", o - -, x = c(1,2,3))
  - Durchführung (Kochvorgang, Zusammenbau, R Skript laufen lassen)



Hattenhauer (2020) Informatik

# Teilgebiete der Informatik mit Relevanz für die Psychologie

#### Angewandte Informatik

• Anwendungssoftware, Human-Computer-Interaction, Informatik und Gesellschaft

#### Technische Informatik

Mikroprozessortechnik, Rechnerarchitektur, Netzwerktechnik

#### Praktische Informatik

• Programmierung, Algorithmen, Datenbanken

#### Theoretische Informatik

• Automatentheorie, Berechenbarkeitstheorie, Komplexitätstheorie

# Spezialgebiete der Informatik mit Relevanz für die Psychologie

#### Maschinelles Lernen und Künstliche Intelligenz

Datenanalyse aus Sicht der Informatik

#### Computervisualistik

• Bilderkennung und Bildsynthese, Virtuelle Realität, Augmented Reality

#### Computerlinguistik

Spracherkennung und Sprachsynthese

#### Bioinformatik

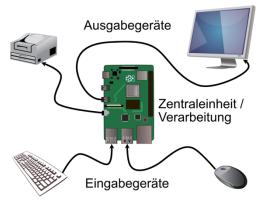
• Lebenswissenschaften, Genomik, Bildgebende Verfahren der Medizin

Grundbegriffe der Informatik

- Datenanalyse
- Informatik
- Rechnerarchitektur
- Algorithmen und Programme

R und Visual Studio Code

# $\textbf{EVA-Prinzip} : \mathsf{Eingabe} \to \mathsf{Verarbeitung} \to \mathsf{Ausgabe}$



Hattenhauer (2020) Informatik

## Zentraleinheit eines Computers

Auch Hauptplatine, Motherboard oder Mainboard genannt



Figure 1: Hattenhauer (2020) Informatik

Hattenhauer (2020) Informatik

## Zentraleinheit eines Computers

### CPU (Central Processing Unit/Mikroprozessor)

- Rechenwerk, Steuerwerk, und Leitwerk des Systems
- Cache (flüchtiger schneller Speicher)
- Intel(R) Core(TM) i5-7300HQ CPU @ 2.50GHz

### RAM (Random Access Memory)

- Temporärer, flüchtiger Arbeitsspeicher des Systems
- Begrenzt, z.B. 16 GB

#### Massenspeicher

- Stationärer Speicher des Systems
- SSD (Solid State Drive), Cloudspeicher

### GPU (Graphical Processing Unit)

- Leistungsstarke, speziell für Visualisierung optimierte Prozessoren
- Unterstützung der CPU in manchen Anwendungen, z.b. Neuronale Netze

#### Von Neumann-Architektur

Abstraktion eines Rechensystems mit wohldefinierten Komponenten und Datenflüsse.

Rechner := Steuerwerk, Rechenwerk, Speicher, Eingabewerk, Ausgabewerk.

#### Zentrale Eigenschaften

- Struktur des Rechners unabhängig von dem zu bearbeitenden Problem
- Daten, Programme, Zwischen- und Endergebnisse liegen im gleichen Speicher.
- Speicher ist in gleichgroße nummerierte (addressierte) Zellen unterteilt.
- Über die Adresse einer Speicherzelle kann deren Inhalt abgerufen verändert werden.
- Ein Programm ist eine Reihe von Befehlen.
- Aufeinanderfolgende Befehle eines Programms liegen in benachbarten Speicherzellen und werden entsprechend nacheinander aufgerufen.
- ightarrow Die Architektur eines Rechners impliziert das Grundprinzip der Programmierung: Befehle werden streng sequentiell abgearbeitet.

Grundbegriffe der Informatik

- Datenanalyse
- Informatik
- Rechnerarchitektur
- Algorithmen und Programme

R und Visual Studio Code

## Vom Realweltproblem zum Programm

#### Realweltproblem

- Das Problem, das mithilfe eines Computers gelöst werden soll.
- z.B. Auswertung von Fragebogendaten einer psychologischen Studie.

#### Problemspezifikation

- Genaue sprachliche Fassung des Realweltproblems.
- z.B. Methodenteil einer wissenschaftlichen Publikation.

#### Algorithmus

- Folge von Anweisungen zur Lösung des Problems.
- z.B. Dateneinlesen, deskriptive Statistiken berechnen, T-Test durchführen.

### Programm

- Ein Algorithmus, der von einem Computer ausgeführt werden kann.
- Eine in einer Programmiersprache verfasste Textdatei.

### Algorithmus

# Definition (Algorithmus)

Ein Algorithmus ist eine Folge von Anweisungen, um aus gewissen Eingabedaten bestimmte Ausgabedaten herzuleiten, wobei folgende Bedingungen erfüllt sein müssen

- Finitheit. Die Anweisungsfolge muss in einem endlichen Text vollständig beschrieben sein.
- Effektivität. Jede Anweisung muss tatsächlich ausführbar sein.
- Terminierung. Der Algorithmus endet nach endlich vielen Anweisungen.
- Determiniertheit. Der Ablauf des Algorithmus ist zu jedem Punkt fest vorgeschrieben.

Wenn E die Menge der zulässigen Eingabedaten und A die Menge der zulässigen Ausgabedaten bezeichnet, dann ist ein Algorithmus eine Funktion

$$f: E \to A, e \mapsto f(e)$$
 (1)

Umgekehrt heißen Funktionen, die durch einen Algorithmus beschrieben werden können, berechenbare Funktionen.

#### Bemerkung

Effektivität sollte nicht mit Effizienz verwechselt werden.

#### Eine Programmiersprache

- ... bestimmt die Regeln, denen ein Programm gehorchen muss.
- ... definiert eine Syntax, also Vokabular und Programmaufbau.
- ... definiert Semantik, also die Bedeutung der erlaubten Anweisungen.

```
#if [ -z "$USER NAME" -o -z "$USER TYPE" -o -z "$GROUP" ]
if [ -z "$USER NAME" -o -z "$USER TYPE" ]
thon
       echo "Please set the user name, type and group"
        echo "Please set the user name and type"
        exit 1
fi
# generate a random password
# -v: include special characters
# -n: include numbers
# -1: one generated passwords per Line
#pwgen -y 15 -n 5 -1
echo "Propositions for random passwords to use in next step:"
pwgen -s -n -1 15 5
# add the user
# requires password to be given via input
adduser --firstuid 1000 --lastuid 9999 --no-create-home ${USER NAME}
```

### Programmiersprachen

### Maschinensprache

- Elementare Operationsbefehle (z.B. Speichern, Vergleichen, Addieren)
- Elementare Operationsbefehle werden als Binärzahlen kodiert

```
Addiere Inhalt R1 zu Inhalt R2
                                        \Rightarrow 1001\ 0010
Erhöhe Inhalt R1 um 1
                                        \Rightarrow 1001 0110
```

- Übertrage Inhalt R1 nach R3 • Programme in Maschinensprache heißen Maschinenprogramme.
- De facto führt ein Computer nur Maschinenprogramme aus.
- Für Menschen ist die Programmierung in Maschinensprache mühselig.

#### Höhere Programmiersprache

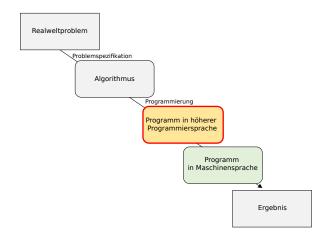
- An die menschliche Sprache angelehnte Wörter und Sätze
- Interpreter oder Compiler übersetzen Programme in Maschinensprache
- R. Python, Matlab, C++, Java, FORTRAN, COBOL,...

 $\Rightarrow 0010\ 0011$ 

### Generationen von Programmiersprachen

- 1. Generation (1GL)
- Maschinensprachen
- 10110000 01100001 (hexadezimaler Darstellung des Ausdrucks "B0 61")
- 2. Generation (2GL)
- Assemblersprachen ab 1950, erste Form der symbolischen Programmierung
- Bspw. "MOV AI, 61H" # Intel-Prozessor-spezifische Sprache
- 3. Generation (3GL)
- Höhere Programmiersprachen ab 1970 wie FORTRAN, C, C++, Java
- Programmierfreundlich, prozessor-unabhängig
- 4. Generation (4GL)
- Höhere Programmiersprachen ab 1980 wie Python, Matlab, R
- Codeoverhead Minimisierung, Automation, Flexibilität, Multiparadigmatisch

# 4GL Programmierung



# Arten der Programmierung

Imperative Programmierung Von Neuamann-Architektur: Die Architektur des Rechners imppliziert bereits die sequentielle Abarbeitung von Befehlen.

- Problemlösungsweg wird als Folge von Anweisungen (Befehlen) vorgegeben.
- Befehle verarbeiten Daten, die mithilfe von Variablen adressiert werden.
  - Prozedurale imperative Programmierung
    - Daten und sie manipulierende Befehle werden separat behandelt.
    - Prozeduren (Funktionen) bilden das zentrale Strukturkonzept.
  - Objektorientierte imperative Programmierung
    - Daten und manipulierende Befehle werden als Objekte zusammengefasst.
    - Objekte bilden das zentrale Strukturkonzept.
- Praktisch liegen oft Mischformen vor.

### Compiler und Interpreter

#### Kompilierte Programmiersprachen

- Gesamter Quellcode wird vor der Ausführung in Maschinensprache übersetzt.
- Das Übersetzungsprogramm heißt Compiler.
- Der übersetzte Maschinencode wird vom Prozessor ausgeführt.
- Das ausführbare Programm wird nicht übersetzt und läuft schnell.
- Bei Änderungen des Quellcodes muss neu kompiliert werden.
- Beispiele für kompilierte Sprachen sind Java, C, C++.

#### Interpretierte Programmiersprachen

- Quellcode wird während der Ausführung in maschinennahe Sprache übersetzt.
- Das Ausführungsprogramm heißt Interpreter.
- Das Programm läuft aufgrund der Interpretation langsamer.
- Bei Änderungen des Quellcodes muss nicht neu interpretiert werden.
- Beispiele für interpretierte Sprachen sind Python und R.

Formalia

Grundbegriffe der Informatik

R und Visual Studio Code

Selbstkontrollfragen

- Eine Programmiersprache und ein Softwarepaket.
- Entwickelt von Ihaka and Gentleman (1996).
- Freier Dialekt der propietären Software S (Becker, Chambers, and Wilks (1988)).
- Weiterentwickelt und gepflegt durch R Core Team und R Foundation
- Interpretierte imperativ-objektorientierte 4GL Sprache.
- Optimiert und populär für statistische Datenanalysen.
- Große Community mit etwa 20.000 beigetragenen R Paketen (Erweiterungen)
- Evolviert und konservativ im Kern, konsistent und progressiv in R Paketen.

### Über cran.r-project.org die geeignete Version herunterladen und installieren.



Documentation

Manuals EAQs



What are R and CRAN? R is "GNUS", a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques; linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information. CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load. Submitting to CRAN To "submit" a package to CRAN, check that your submission meets the CRAN Repository Policy and then use the web form

If this falls, send an email to CRAN-submissions@R-project.org following the policy. Please do not attach submissions to emails, because this will clutter up the mailboxes of half a

### Was kann man mit R machen?

- Datensätze laden, manipulieren, und speichern.
- Eine Vielzahl von Berechnungen an verschiedenen Datenstrukturen durchführen.
- Eine Vielzahl statistischer Analysemethoden auf Daten anwenden.
- Datenanalyseskripte schreiben und Abbildungen generieren.
- Präsentationen RMarkdown und Bücher RBookdown erstellen.
- Wissenschaftliche Berichte mit Quarto erstellen.

# Was kann man mit R (bisher) nicht so gut machen?

- In einer ansprechenden Umgebung programmieren (⇒ Visual Studio Code).
- Scientific Computing (⇒ Python, Matlab, Julia).
- Psychologische Experimente programmieren (⇒ Python, Matlab)

#### Wie bekomme ich Hilfe zu R?

- Googlen
- ChatGPT
- stackoverflow.com
- r-project.org/help.html
- Während der Programmierung und bei bekanntem Funktionsnamen über die Kommandozeile:

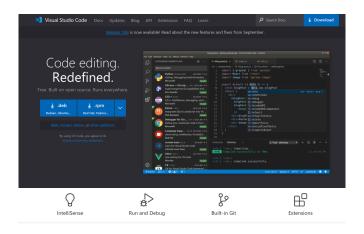
- · rseek.org
- rstudio.com/resources/cheatsheets
- r-bloggers.com

# Was ist Visual Studio Code (VSCode)?

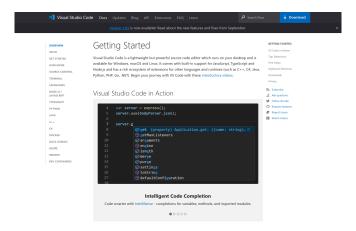
- VSCode ist eine kostenloser Quelltext-Editor von Microsoft.
- VSCode ist eine Softwareentwicklungsumgebung (engl. Integrated Development Environment)
- Seit 2015 für Windows, macOS und Linux verfügbar.
- Seit 2018 ist VSCode der beliebteste Editor laut jährlicher stackoverflow Umfragen.
- Ein Microsoftprodukt ist damit auch der beliebteste Editor der Linuxwelt.
- Über Extensions kann VSCode als IDE für beliebige Sprachen genutzt werden.
- Zum Beispiel funktioniert VSCode als IDE für R, Python, Julia, Shell, Quarto, etc.
- VSCode ist Community-based und sehr konfigurierbar.
- VSCode ist über Microsoft's GitHub über Endgeräte synchronisierbar.

#### Wie bekomme ich VSCode?

Über code.visualstudio.com herunterladen und installieren.



Online Dokumentation: code.visualstudio.com/docs



# Aufgaben bis nächste Woche

- 1. R installieren.
- 2. VSCode installieren.
- 3. VSCode für R startklar machen: (Anleitung hier).

Formalia

 $Grundbegriffe\ der\ Informatik$ 

R und Visual Studio Code

Selbstkontrollfragen

### Selbstkontrollfragen

- 1. Geben Sie die typische Struktur einer computergestützten Datenanalyse wieder.
- 2. Erläutern Sie den Begriff "Datenanalyseskript".
- 3. Definieren Sie den Begriff "Informatik".
- 4. Erläutern Sie die Akronyme CPU, RAM, SSD, und GPU.
- 5. Nennen Sie wesentliche Aspekte der Von-Neumann Rechnerarchitektur.
- 6. Definieren Sie den Begriff Algorithmus.
- 7. Erläutern Sie den Zusammenhang von Algorithmen und Programmen.
- 8. Was bezeichnen die Syntax und Semantik einer Programmiersprache?
- 9. Differenzieren Sie die Begriffe "Maschinensprache" und "höhere Programmiersprache".
- Skizzieren Sie Prinzipien der prozeduralen und objektorientierten imperativen Programmierung.
- 11. Skizzieren Sie die Entwicklung der Programmiersprachen der ersten bis vierten Generation.
- 12. Differenzieren Sie die Begriffe der kompilierten und der interpretierten Programmiersprachen.

#### Referenzen

- Becker, Richard A., John M. Chambers, and Allen Reeve Wilks. 1988. *The New S Language: A Programming Environment for Data Analysis and Graphics*. Reprint. London: Chapman & Hall.
- Ihaka, Ross, and Robert Gentleman. 1996. "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5 (3): 2999–2314.