

1 Was ist Computational Communication Science?	1
1.1 Daten, Datafizierung und Big Data	4
1.2 Software, Algorithmen und Modelle	6
1.3 CCS im empirischen Forschungsprozess.	9
1.3.1 Befragung, Beobachtung, Inhaltsanalyse oder Experiment?	12
1.3.2 Qualitativ oder quantitativ?	13
1.3.3 Deskriptiv oder explanativ?	14
1.3.4 Deterministisch oder probabilistisch?	15
1.3.5 Offen oder proprietär?	16
1.3.6 Ethisch oder unethisch?	17
1.3.7 Legal oder illegal?	17
1.4 Zwischenfazit und Literaturhinweise.	18
2 Daten sichten	21
2.1 Empirische Perspektive	23
2.1.1 Merkmale, Merkmalsträger und Merkmalsausprägungen.	23
2.1.2 Mess- und Skalenniveaus.	24
2.2 Technische Perspektive	26
2.2.1 Tabellarische Datensätze	26
2.2.2 Hierarchische Datensätze.	28
2.3 Informatische Perspektive	30
2.3.1 Zahlen	32
2.3.2 Fehlende Werte.	33
2.3.3 Texte (und ihre Probleme).	34
2.3.4 Listen, Arrays und Tabellen	35
2.3.5 Objekte und binäre Daten	36
2.4 Statistische Perspektive	36
2.4.1 Häufigkeiten und Histogramme	37
2.4.2 Lagemaße, Streuungsmaße und Box-Plots	38
2.4.3 Fehlende Werte.	41
2.5 Zwischenfazit und Literaturhinweise.	42

3	Daten bewerten	45
3.1	Nach ihrer Aussagekraft	46
3.2	Nach ihrer Darstellungsweise	49
3.2.1	Querschnittsdaten	50
3.2.2	Längsschnittsdaten und Zeitreihen	53
3.2.3	Relationale Daten	58
3.3	Nach ihrer rechtlichen Grundlage	61
3.3.1	Eigentumsrechte	61
3.3.2	Personenbezogene Rechte	64
3.3.3	Lizenzen	65
3.4	Nach ihren ethischen Prinzipien	65
3.4.1	Menschen stehen im Mittelpunkt	68
3.4.2	Daten und Software sind sicher, robust und nachvollziehbar	69
3.4.3	Entscheidungen verlaufen fair und transparent	71
3.5	Nach ihrer Genese	72
3.5.1	Tracking-Daten	72
3.5.2	Social-Media-Daten	74
3.5.3	Input-Output-Daten	75
3.5.4	Datenspenden	76
3.5.5	Medieninhalte	77
3.5.6	Modellierte und simulierte Daten	78
3.6	Zwischenfazit und Literaturhinweise	79
4	Forschungssoftware entwickeln	81
4.1	R und/oder Python	82
4.2	Programmierung	83
4.2.1	Hallo, Welt!	85
4.2.2	Variablen	87
4.2.3	Bedingte Verzweigungen	88
4.2.4	Schleifen	89
4.2.5	Funktionen und objektorientierte Programmierung	90
4.2.6	Modularisierung	92
4.2.7	Bibliotheken und Pakete	93
4.3	Dokumentation	94
4.4	Tests	95
4.5	Deployment, Pipelines und Versionierung	96
4.6	Zwischenfazit und Literaturhinweise	97
5	Daten beziehen	99
5.1	Daten von anderen	101
5.1.1	Nutzungsrechte	102
5.1.2	Zugang	103
5.1.3	Verzerrungen	103

5.2	Datenbanken	107
5.2.1	Integrität und Normalisierung	109
5.2.2	Effizienz und Leistung	111
5.2.3	Structured Query Language (SQL)	112
5.3	Application Programming Interfaces (APIs)	113
5.3.1	Schnittstellenprogrammierung	114
5.3.2	Anwendungssoftware	117
5.3.3	GET und POST	118
5.3.4	Streaming	118
5.4	Die „APIcalypse“ und das „post-API“-Zeitalter	119
5.5	Zwischenfazit und Literaturhinweise	122
6	Fremde Daten sammeln	125
6.1	Web Scraping	126
6.1.1	Rechtliche Grundlagen	128
6.1.2	Ethische Prinzipien	129
6.1.3	Umsetzung	129
6.2	Agentenbasierte Tests	130
6.2.1	Rechtliche Grundlagen	131
6.2.2	Ethische Prinzipien	132
6.2.3	Umsetzung	133
6.3	Web Tracking	133
6.3.1	Rechtliche Grundlagen	134
6.3.2	Ethische Prinzipien	136
6.3.3	Umsetzung bei Webseitenbetreibenden	137
6.3.4	Umsetzung während der Nutzung	138
6.3.5	Umsetzung bei Nutzenden	139
6.4	Datenspenden	140
6.4.1	Rechtliche Grundlagen	140
6.4.2	Ethische Prinzipien	141
6.4.3	Umsetzung	142
6.5	Zwischenfazit und Literaturhinweise	143
7	Eigene Daten generieren	145
7.1	Rechtliche Grundlagen	147
7.2	Ethische Prinzipien (und CCS-Ressourcenverbrauch)	147
7.3	Monte-Carlo-Simulation	149
7.3.1	Zufallszahlen	151
7.3.2	Simulation	154
7.3.3	Mehrwert und Limitationen	156
7.4	Bootstrapping	157
7.4.1	Zufallsstichproben	158
7.4.2	Simulation	159
7.4.3	Mehrwert und Limitationen	160

7.5	Agentenbasierte Modellierung	160
7.5.1	Agenten, Umwelt, Regeln und Ticks	161
7.5.2	Simulation	163
7.5.3	Mehrwert und Limitationen	165
7.6	Zwischenfazit und Literaturhinweise	166
8	Texte als Daten I	169
8.1	Zählen	171
8.1.1	Sätze, Typen und Tokens	172
8.1.2	Reguläre Ausdrücke	174
8.1.3	Stemming und Lemmatisierung	176
8.1.4	Diktionäre	177
8.2	Kontextualisieren	180
8.2.1	Konkordanzen	181
8.2.2	Kollokationen	182
8.3	Verstehen	184
8.3.1	Kollokationsmatrizen	185
8.3.2	Typen und Tokens als Vektoren	186
8.3.3	Transformer-Modelle	190
8.4	Zwischenfazit und Literaturhinweise	192
9	Texte als Daten II	195
9.1	Rechtliche Grundlagen	195
9.2	Ethische Prinzipien	197
9.3	Datenaufbereitung („Pre-Processing“)	198
9.3.1	Definition eines Dokuments	198
9.3.2	Tokenisierung	199
9.3.3	Stoppwörter	201
9.3.4	Vereinheitlichung	202
9.3.5	Stemming und Lemmatisierung	204
9.4	Annotation	204
9.4.1	Tagging und Parsing	204
9.4.2	Named Entity Recognition	205
9.4.3	Word Embeddings	206
9.5	Document-Feature-Matrizen (oder: Wie passt das alles zusammen?)	207
9.6	Beschreibung und Visualisierung	211
9.6.1	Wichtigkeit und Wortwolken	212
9.6.2	Nähe und Netzwerke	214
9.6.3	Distanz und Distinktion	214
9.6.4	Valenz und Verlauf	217
9.7	Zwischenfazit und Literaturhinweise	219

10	Maschinelles Lernen mit Goldstandard („überwachtes Lernen“)	221
10.1	Daten	225
10.2	Features	230
10.3	Algorithmen	234
10.3.1	Naïve Bayes	235
10.3.2	Decision Tree und Random Forest	238
10.3.3	Support-Vector Machine (SVM)	240
10.3.4	Neuronale Netze	242
10.3.5	Ensemble	245
10.4	Validierung	246
10.4.1	Accuracy	247
10.4.2	Precision, Recall und F_1	248
10.4.3	ROC-Kurve	251
10.5	Klassifizierung	253
10.6	Zwischenfazit und Literaturhinweise	254
11	Maschinelles Lernen ohne Goldstandard („unüberwachtes Lernen“)	257
11.1	Rechtliche Grundlagen	258
11.2	Ethische Prinzipien	259
11.3	Daten, Features und Validierung	262
11.4	Algorithmen	263
11.4.1	k-means	265
11.4.2	Generative Modelle	267
11.4.3	Latent Dirichlet Allocation (LDA)	269
11.4.4	Structural Topic Modeling (STM)	271
11.5	Anwendung und Ressourcen	273
11.6	Zwischenfazit und Literaturhinweise	276
12	Netzwerke als Daten	279
12.1	Knoten und Kanten	281
12.2	Dichte und Distanz	284
12.3	Homophilie und Heterophilie	291
12.4	Beschreibung und Visualisierung	295
12.5	Rechtliche Grundlagen	298
12.6	Ethische Prinzipien	299
12.7	Zwischenfazit und Literaturhinweise	299
13	Gruppen und Sequenzen als Daten	301
13.1	Gruppen als abhängige Größe	302
13.1.1	Grundprinzip Clusteranalyse	302
13.1.2	Beispiel Inhaltliche Muster	305
13.1.3	Mehrwert und Limitationen	305

13.2	Gruppen als unabhängige Größe	306
13.2.1	Grundprinzip Mehrebenenmodellierung	307
13.2.2	Beispiel Mediensystemeinflüsse	310
13.2.3	Mehrwert und Limitationen	311
13.3	Zeitliche Sequenzen	312
13.3.1	Grundprinzip Zeitreihenmodellierung	313
13.3.2	Beispiel Nachrichtendiffusion	316
13.3.3	Mehrwert und Limitationen	317
13.4	Bedingte Sequenzen	317
13.4.1	Grundprinzip Markow-Ketten	318
13.4.2	Beispiel Surfverhalten	320
13.4.3	Mehrwert und Limitationen	321
13.5	Zwischenfazit und Literaturhinweise	321
14	Bilder und multimodale Daten	323
14.1	Visuelle Daten	324
14.1.1	Dienstleistende oder selbst aktiv werden?	326
14.1.2	Beispiel visuelle Texterkennung (OCR)	327
14.2	Audio-Daten	328
14.2.1	Dienstleistende oder selbst aktiv werden?	330
14.2.2	Beispiel Spracherkennung (TTS)	331
14.3	Audiovisuelle Daten	331
14.3.1	Dienstleistende oder selbst aktiv werden?	333
14.3.2	Beispiel „optical flow estimation“	334
14.4	Geo-Daten	335
14.4.1	Dienstleistende oder selbst aktiv werden?	338
14.4.2	Beispiel Visualisierung der 7-Tage-Inzidenz	339
14.5	Rechtliche Grundlagen	341
14.6	Ethische Prinzipien	342
14.7	Zwischenfazit und Literaturhinweise	342
	Literatur	345
	Stichwortverzeichnis	351