Computational and Mathematical Modeling Program

RWTH AACHEN UNIVERSITY

Diese Arbeit wurde vorgelegt am
Lehrstuhl für Mathematik (MathCCES)

Wie sicher ist meine Privatsphäre in Online Netzwerken? ...und was hat das mit Mathe zu tun?

How private is my data in online networks? ...and what does this have to do with maths?

Masterarbeit
zur Erlangung des akademischen Grads Master of Education

8. Juli 2016

Vorgelegt von Presented by	Maike Sube [redacted]
Erstprüfer First examiner	Prof. Dr. Martin Frank Lehrstuhl für Mathematik (MathCCES) RWTH Aachen University
Zweitprüfer Second examiner	Prof. Dr. Sebastian Walcher Lehrstuhl A für Mathematik RWTH Aachen University
Koreferenten Co-supervisor	Dr. Christina Roeckerath Lehrstuhl für Mathematik (MathCCES) RWTH Aachen University
 Co-supervisor	Thomas Camminady Lehrstuhl für Mathematik (MathCCES) RWTH Aachen University

Eigenständigkeitserklärung

Hiermit versichere ich eidesstattlich, dass ich diese Masterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht. Dasselbe gilt sinngemäß für Tabellen und Abbildungen. Diese Arbeit hat in dieser oder einer ähnlichen Form noch nicht im Rahmen einer anderen Prüfung vorgelegen.

Aachen, im Juli 2016

Maike Sube

Inhaltsverzeichnis

1. Warum das Thema für Schülerinnen und Schüler relevant ist - Einleitung 1
1.1. Soziale Online Netzwerke und aktuelle Forschung - Motivation und Ziele dieser Arbeit . . . 1
1.2. Von der Forschung zum CAMMP day . . . 3
1.3. Was Sie in dieser Masterarbeit erwartet . . . 5

2. ...und das steckt hinter der Konzeption des Lernmoduls - Fachdidaktischer Hintergrund 6
2.1. Was ist CAMMP? . . . 6
2.2. Was bedeutet mathematisches Modellieren und wie soll das gehen? . . . 8
2.2.1. Begrifflichkeit und Arten von Modellen . . . 8
2.2.2. Modellierungsprozess . . . 9
2.2.3. Wichtigkeit und Vorteile von Modellen und Modellierungsaufgaben 13
2.2.4. Modellieren mit digitalen Werkzeugen . . . 14
2.3. Was bedeutet Problemlösen? . . . 14
2.3.1. Begrifflichkeit und Arten von Problemen . . . 14
2.3.2. Problemlöseprozess . . . 15

3. ...das hat das mit Mathe zu tun! - Fachlicher Hintergrund 16
3.1. Problemstellung . . . 16
3.2. Das soziale Netzwerk als ungerichteter Graph . . . 18
3.3. Daten eines sozialen Netzwerks . . . 19
3.3.1. Friendster . . . 19
3.3.2. Grundlegende Annahme . . . 20
3.3.3. Daten aufbereiten . . . 20
3.3.4. Die Nutzer und ihre Eigenschaften . . . 22
3.4. Untersuchung der Fragestellung . . . 29
3.4.1. Vorgehen . . . 29
3.4.2. Konkrete Untersuchungen . . . 31
3.5. Verbesserung von Kappa . . . 43
3.5.1. Optimierung in Anlehnung an die Monte-Carlo-Methode . . . 44
3.5.2. Weitere angedachte Optimierungsverfahren . . . 49
3.6. Gültigkeit der Ergebnisse: Hypothesentests . . . 52
3.6.1. Idee dahinter . . . 53
3.6.2. Aber wie kommen wir auf die Entscheidungsvorschrift? . . . 55
3.6.3. Signifikanzniveau . . . 57
3.7. Veränderungen und -verbesserungen . . . 58

4. ...und das ist das Lernmodul - Der CAMMP day und ein didaktischer Blick darauf 61
4.1. Methodisches Konzept . . . 61
4.1.1. Ziele des CAMMP days . . . 61

4.1.2. Bezug zur Schulmathematik . . . 61
4.1.3. Der Modellierungskreislauf liegt der Struktur des CAMMP days zu Grunde . . . 64
4.1.4. Ablauf und Struktur des CAMMP days . . . 66
4.2. Didaktische Analyse der Materialien . . . 67
4.2.1. Beschreibung und didaktische Zielsetzung der Materialien und der Arbeit mit diesen . . . 67
4.2.2. Das Problemlösen im CAMMP day . . . 73
4.2.3. Die Schulmathematik im CAMMP day . . . 73
4.2.4. Die Nutzung von Vorteilen von Modellierung im CAMMP day . 74
4.2.5. MATLAB als digitales Werkzeug im CAMMP day . . . 74
4.2.6. Antizipierte Lernhürden und Unterstützungsmaßnahmen . . . 75
4.2.7. Erreichung der gesteckten Ziele des CAMMP days . . . 77

5. ...und so verlief die erste Durchführung - Evaluation 78
5.1. Organisatorische Anmerkungen . . . 78
5.2. Beschreibung der Schülerinnen und Schüler . . . 79
5.3. Leitende Gesichtspunkte zur Evaluation und Evaluationsmethoden . . . 79
5.4. Evaluation . . . 79

6. ...und das kann verbessert werden - Ausblick zur Weiterentwicklung des CAMMP days 85
6.1. Organisatorische Aspekte . . . 85
6.2. Arbeitsmaterial . . . 86
6.3. Weiterentwicklungen in der Konzeption . . . 87

7. Abschließende Worte 89

Abbildungsverzeichnis 90

Literatur 91

A. Anhang 95
A.1. Methodisches Konzept und Ablaufübersicht . . . 95
A.2. Präsentation zur Problemstellung . . . 99
A.2.1. Abbildungsverzeichnis . . . 99
A.3. Arbeitsblätter mit Antwortblättern, Besprechungsfolien und Lösungen . 110
A.4. Codes . . . 144
A.4.1. sozialenetzwerke.m . . . 144
A.4.2. ZusatzGuetebewerten.m . . . 148
A.4.3. ZusatzHypothesentests.m . . . 149
A.5. Abschlussbesprechung . . . 153
A.5.1. Abbildungsverzeichnis . . . 153
A.6. Evaluationsbogen und Auswertung . . . 157

1. Warum das Thema für Schülerinnen und Schüler relevant ist - Einleitung

Not on a social network? You've still got a privacy problem

Diese Überschrift eines Artikels von Robert McMillan aus dem Jahr 2014 [21] beschreibt treffend das Thema dieser Masterarbeit. Es geht um soziale Online Netzwerke und die Privatsphäre in diesen. Schauen wir uns das etwas genauer an:

1.1. Soziale Online Netzwerke und aktuelle Forschung - Motivation und Ziele dieser Arbeit

In unserem Leben und im Leben unserer Schülerinnen und Schüler spielen sogenannte „soziale Netzwerke" eine große Rolle. Unter einem sozialen Netzwerk oder in englischer Sprache einem „social network" verstehen wir laut Duden [5]:

> **Was ist ein *soziales Netzwerk*?**
> Portal im Internet, das Kontakte zwischen Menschen vermittelt und die Pflege von persönlichen Beziehungen über ein entsprechendes Netzwerk ermöglicht.

Schaue ich mich in meinem Freundeskreis um, so sind Personen, die keine Mitglieder von Facebook oder WhatsApp sind, die große Ausnahme. Solche Netzwerke sind abhängig vom Internet. Wir nennen sie **soziale Online Netzwerke**. Auch die heutigen Schülerinnen und Schüler sind in solchen Netzwerken aktiv. Sie sind Teil ihrer Lebenswelt. Diese Behauptungen werden gestützt von einer im Jahr 2011 veröffentlichten Studie vom BITKOM, dem Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. [23]. Zentrale Ergebnisse waren, dass 76% der befragten Internetnutzer in einem sozialen Netzwerk angemeldet sind. Bei den unter 30 Jährigen betrug der Prozentsatz sogar 96%. Facebook ist hierbei das am meisten genutzte Netzwerk mit der größten Nutzergruppe von $14 - 29$ Jährigen.
Die Relevanz von sozialen Netzwerken wird somit schnell deutlich. Mit deren Nutzung werden persönliche Daten preisgegeben. So sagt die Studie vom BITKOM [23], dass vor allem Vor- und Nachnamen, das Alter, ein persönliches Foto und der Beziehungsstatus angegeben werden. Nicht nur diese Daten werden von den Betreibern der Netzwerke gespeichert.
In den letzten Jahren wurde das Thema der Privatsphäre in sozialen Online Netzwer-

ken intensiv erforscht. Eine Auswahl dieser Arbeiten ist in [9, S. 2-3] zu finden. Die grundsätzliche, übergreifende Frage hinter diesen Studien ist:

Inwieweit ist es möglich, mehr über Personen durch Daten eines sozialen Netzwerks herauszufinden, als diese angegeben haben?

Diese Fragestellung betrifft sowohl Nutzer als auch Nicht-Nutzer eines sozialen Netzwerks, denn durch die Zustimmung der Nutzungsbedingungen von Facebook for Mobile wird den Betreibern der Zugriff auf die Kontaktliste erlaubt [9, S. 2]. In dieser Kontaktliste befinden sich bei vielen Personen auch Nicht-Nutzer des sozialen Netzwerks, wodurch Daten über diese preisgegeben werden. Dies ermöglicht (theoretisch) für Nutzer und Nicht-Nutzer die Erstellung von *Schattenprofilen* (*Shadow profiles*). Darunter wird gemäß [9, S. 2] Folgendes verstanden:

> ***Schattenprofile (Shadow profiles)***
> Ein Schattenprofil ist ein Profil von Personen, das Daten enthält, die Freunde der Personen freiwillig preisgegeben haben.

Obwohl noch nicht klar ist, ob die Erstellung solcher Schattenprofile tatsächlich stattfindet [21], ist die Möglichkeit der Erstellung doch besorgniserregend. Die Frage, die sich nun aufdrängt ist:

Wie gut ist es möglich, mehr über Personen durch Daten eines sozialen Netzwerks herauszufinden, als diese angegeben haben?

Dieser Fragestellung sind insbesondere drei Forscher der ETH Zurich in ihrer Veröffentlichung „Online Privacy as a Collective Phenomenon“ [9] nachgegangen. Ihre wissenschaftliche Vorgehensweise (das Modellieren) und die zentralen Erkenntnisse sollen durch diese Masterarbeit Schülerinnen und Schülern zugänglich gemacht werden. Denn: Soziale Netzwerke sind Teil der Lebenswelt von unseren Schülerinnen und Schülern. Der reflektierte Umgang mit ihnen sollte demnach geschult werden. Ein solcher ist jedoch nur möglich, wenn ein Bewusstsein für soziale Netzwerke - und das nicht nur durch ihre Nutzung sondern auch durch eine wissenschaftliche Betrachtung - geschaffen wird. Diese Masterarbeit soll zur Entstehung dieses Bewusstseins beitragen. Überdies trägt diese Masterarbeit zur Allgemeinbildung unserer Schülerinnen und Schüler bei. Ein Grund dafür sind die von Heinrich Winter entwickelten drei Grunderfahrungen, die Schülerinnen und Schüler im Mathematikunterricht machen sollen. Gemäß [40] lauten diese:

1. Mathematik soll als *nützliche, brauchbare Disziplin* verstanden werden, um Erscheinungen der Welt um uns in einer spezifischen Art wahrzunehmen und zu verstehen (**Anwendungen der Mathematik**). Bei diesem Punkt betont Winter die Bedeutung von Modellen.

2. Mathematik soll als *deduktiv geordnete eigene Welt*, als geistige Schöpfung kennengelernt und begriffen werden.

3. Mathematik soll als *Schule des Denkens* verstanden werden (**heuristische Fähigkeiten**). Winter spricht hier die Förderung von Problemlösefähigkeiten und die vielen Beziehungen zur außermathematischen Realität an.

Durch die Fragestellung

Wie sicher ist meine Privatsphäre in Online Netzwerken? ...und was hat das mit Mathe zu tun?

wird die Mathematik im Bezug auf die Lebenswelt der Schülerinnen und Schüler als Werkzeug, alltägliche Erscheinungen wahrzunehmen und zu verstehen, sowie als Lernanlass für hinausgehende Fähigkeiten genutzt werden können.

1.2. Von der Forschung zum CAMMP day

Das zentrale Ziel dieser Arbeit ist die **Entwicklung eines CAMMP Moduls zum Thema soziale Online Netzwerke** für Schülerinnen und Schüler. Die Einrichtung CAMMP, das „Computergestützte Mathematische Modellierungsprogramm “ [36], der RWTH Aachen ermöglicht Schülerinnen und Schülern, (aktuelle) Forschungsthemen zu erarbeiten. Genaueres zum CAMMP finden Sie im nächsten Kapitel.
Diese Masterarbeit ist der ersten Schritt zur Aufbereitung des Themas „soziale Online Netzwerke“. Der Ausgangspunkt hierzu ist die bereits erwähnte Arbeit „Online Privacy as a Collective Phenomenon“ [9] von Garcia, Sarigol und Schweitzer. Neben dieser lagen keine anderen Materialien vor. Um einen CAMMP day zu entwickeln, war zunächst eine Auseinandersetzung mit der Forschung von Garcia, Sarigol und Schweitzer notwendig. Ihre Forschungsarbeit wurde in einem zweimonatigen Forschungsprozess durch einfachere, schülergerechtere Methoden nachgestellt, ausgearbeitet und angepasst. Dabei geschah unter Betreuung von Thomas Camminady Folgendes:
Um den Forschungsprozess zu verstehen, sollen hier (sehr) kurz die Forschungsmethoden und die zentralen Ergebnisse von Garcia, Sarigol und Schweitzer anhand von [9]

skizziert werden. Genauere Erläuterungen der fachlichen Begriffe finden Sie im fachlichen Hintergrund.
Geleitet von der Frage „Wie gut ist es möglich, mehr über Personen durch Daten eines sozialen Netzwerks herauszufinden, als diese angegeben haben?“ schauten sich die drei Forscher als spezielle Information über Nutzer deren sexuelle Orientierung an. Sie nutzten Daten des Netzwerks *Friendster*. Mit Hilfe von maschinellem Lernen (einem sogenannten *Random Forest Classifier*) machten sie Vorhersagen über die sexuelle Orientierung und ermittelten, wann diese Vorhersagen besser sind als der Zufall. Dazu nutzen sie verschiedene Maße (Cohen's Kappa, Trefferquote und Genauigkeit) und untersuchten diverse Abhängigkeiten. Schließlich führten sie zu den ermittelten Abhängigkeiten statistische Tests (Hypothesentests) durch, um die Belastbarkeit der Ergebnisse zu prüfen. Sie fanden zusammenfassend heraus:

- Die Gefährdung der Privatsphäre ist je nach sexueller Orientierung unterschiedlich.
- Ist die Gruppe der sexuellen Orientierung groß oder klein, so ist die Gefährdung der Privatsphäre abhängig von der Netzwerkgröße (je größer desto gefährdeter) und von dem Enthüllungsverhalten der Nutzer (je mehr preisgegeben wird desto gefährdeter).
- Die Gefährdung der Privatsphäre ist von der Vernetzung einer Person innerhalb des Netzwerks und aller Personen im Netzwerk abhängig. Dabei gilt: Je besser die Personen vernetzt sind, desto gefährdeter ist die Privatsphäre.

Ausgehend von diesen Ergebnissen wurde in der Forschung der Masterarbeit ein Teil des Datensatzes von Garcia, Sarigol und Schweitzer erneut aufbereitet, da es nicht möglich war, an die von den drei Forschern genutzten Daten zu gelangen. Dies wurde vor allem von Thomas Camminady getätigt. Mit diesen Daten wurden ebenfalls Vorhersagen über die sexuelle Orientierung der Nutzer macht. Dazu wurde das mathematische Modell genauer aufgearbeitet, da dies aus der Forschungsarbeit nicht detailliert ersichtlich war. Die Vorhersagen wurden mit Hilfe von einfachen Vorhersageregeln und nicht wie bei Garcia, Sarigol und Schweitzer mit gängigen Klassifizierungen getätigt. Bei der Entwicklung der Vorhersageregeln wurden die genannten Ergebnisse der drei Forscher beachtet, was vor allem für den letzten Punkt zur Vernetzung der Personen im Netzwerk gilt. Hierzu wurde eine Vielzahl von Regeln entwickelt und getestet, von denen sich nur eine Auswahl im CAMMP day wiederfindet. Beim Testen der Regeln

wurden die gleichen Maße wie bei Garcia, Sarigol und Schweitzer genutzt. Zur Verbesserung der Regeln wurden zudem verschiedene Optimierungsverfahren getestet und diskutiert. Die Gültigkeit der Ergebnisse wurde - wie bei Garcia, Sarigol und Schweitzer - mittels Hypothesentests untersucht, wobei hier keine Abhängigkeiten sondern Trefferwahrscheinlichkeiten auf ihre Gültigkeit getestet wurden. Es wurde also ein anderes stochastisches Modell genutzt.
Schließlich erfolgte die didaktische Aufbereitung der Thematik unter der Betreuung von Dr. Christina Roeckerath. Ein erster Ansatz dieser Aufbereitung, der auch mit Thomas Camminady diskutiert wurde, ist mit Schülerpraktikanten an der RWTH Aachen getestet worden. Die Erfahrungen hieraus wurden in der endgültigen Aufbereitung genutzt. Diese erfolgte in zweifacher Hinsicht: zum einen für die Durchführenden Hilfswissenschaftler/innen des Moduls, in der Regel Lehramtsstudierende der Mathematik und Computational Engineering Science Studierende ab dem vierten Fachsemester, und zum anderen für Schülerinnen und Schüler der gymnasialen Oberstufe. Das Ergebnis dieser Aufbereitung ist ein erster Entwurf des Moduls. Es wurden Arbeitsblätter, Lösungsvorschläge, Hilfekarten, Präsentationen und Programme entwickelt. Hierzu musste der Umgang mit der Software MATLAB erlernt werden. Die Materialien wurden durch einen Leistungskurs Mathematik der Qualifikationsphase I des Gymnasiums am Wirteltor (Düren) getestet. Dieser erste Durchlauf des Moduls wurde evaluiert und anhand dessen wurden Verbesserungsvorschläge benannt, um eine weitere Entwicklung des Lernmoduls zu ermöglichen.

1.3. Was Sie in dieser Masterarbeit erwartet

In den nächsten Kapiteln erwarten Sie die Antworten auf die folgenden Fragen:

- *Was steckt hinter der Konzeption des Lernmoduls?*
  Hier werden die fachdidaktischen Hintergründe erläutert.

- *Was hat das mit Mathe zu tun?*
  In diesem Kapitel erfolgt eine Darlegung der fachlichen Hintergründe.

- *Wie sieht das Lernmodul aus?*
  Es folgen eine Vorstellung und didaktische Analyse der Konzeption und der Materialien des Moduls.

- *Wie lief die erste Durchführung?*
  Die erste Durchführung wird beschrieben und evaluiert.

- *Was kann verbessert werden?*
  Die Masterarbeit schließt mit ersten Vorschlägen zur Verbesserung des Moduls ab.

Lassen Sie uns diese Fragen gemeinsam beantworten.

2. ...und das steckt hinter der Konzeption des Lernmoduls - Fachdidaktischer Hintergrund

Die Vermittlung der Forschungsmethodik und -ergebnisse von Garcia, Sorigol und Schweitzer [9] an Schülerinnen und Schüler findet in Form eines *CAMMP days* statt.

2.1. Was ist CAMMP?

Die Zusammenstellung der folgenden Informationen erfolgt anhand von [36]. Abweichende Quellen sind gekennzeichnet.
Die Abkürzung **CAMMP** steht für „**Computergestütztes Mathematisches Modellierungsprogramm**“. Der Lehrstuhl Mathematik CCES bietet hierbei Schülerinnen und Schülern die Möglichkeit, sich mit realen wissenschaftlichen Problemen auseinanderzusetzen und diese zu lösen. Gemäß des Namens soll es also um drei Dinge gehen:

- um die Nutzung von Computern,
- um Mathematik und
- um Modellierung.

Gehen wir die Punkte gemeinsam durch: Die *Nutzung von Computern* ist letztendlich die Nutzung von digitalen Werkzeugen zur Lösung eines Problems. Gearbeitet wird mit der Software *MATLAB*. Diese wurde von dem Unternehmen MathWorks [20] konzipiert und bietet durch vielfältige Möglichkeiten den Schülerinnen und Schülern eine Unterstützung bei der Lösung des Problems. Der Einsatz und die genauen Vorteile der Nutzung dieses digitalen Werkzeugs in dem hier konzipierten Modul werden in den folgenden Abschnitten bzw. im folgenden Kapitel dargelegt. Der zweite Punkt lautet *Mathematik*. Die Mathematik bietet Lösungsmöglichkeiten für die betrachteten Probleme. Dabei geht es darum, die mathematischen Werkzeuge zu nutzen und zum Teil

auch neue kennenzulernen. Das Ganze soll in Form von *Modellierung* geschehen. Die Modellierung ist eine Methode der Erkenntnisgewinnung, die im folgenden Abschnitt genauer erläutert wird.
Ziel eines CAMMP Moduls ist es, dass die Schülerinnen und Schüler ein Problem tatsächlich lösen. Dabei wird an den aktuellen Lehrplänen angeknüpft und neue Mathematik, die notwendig wird, vermittelt. Das kann in zwei Organisationsformen geschehen: in der CAMMP week, in der die Schülerinnen und Schüler eine Woche lang ein Problem erarbeiten, oder in einem CAMMP day, in dem die Lösung innerhalb eines Tages erfolgen soll.
Die Organisationsform dieses Moduls ist der CAMMP day. Die Schülerinnen und Schüler kommen hierbei einen Tag lang zur RWTH Aachen und befassen sich mit einem Problem. Aktuell stehen dabei folgende Themen zur Auswahl:

- Wie funktioniert eigentlich GPS und was hat das mit Mathe zu tun?
- Wie funktioniert eigentlich Google und was hat das mit Mathe zu tun?
- Spiegelaufstellung in einem Solarkraftwerk

In der Entwicklung befinden sich Module zum Thema „Brückenbau“, „Shazam“, „mp3“, „Animationsfilme“ und eben „soziale Online Netzwerke“.
Hinter CAMMP steht nicht nur ein organisatorischer Rahmen, sondern auch ein spezielles Verständnis von der Vermittlung von Mathematik, das in [35] erläutert wird. Dabei hat CAMMP das Ziel der Berufs- und Studienorientierung für Schülerinnen und Schüler und möchte ein Studium von CES (Computational Engineering Science) oder Mathematik anregen. Des Weiteren soll die gesellschaftliche Bedeutung von Mathematik und Simulationswissenschaften anhand von ausgewählten Beispielen deutlich werden. Hier kann im Bezug zu den genannten drei Grunderfahrung nach Winter ein Beitrag zur Allgemeinbildung gesehen werden. Das Ganze soll nicht in Form von üblichem Unterricht geschehen. Es werden computergestützt neue forschungsgeleitete didaktische Ideen ausprobiert und evaluiert. Hierbei gilt es, problemorientiert Lösungswege für Schülerinnen und Schüler mit passiven als auch mit aktiven Phasen zu gestalten. Das bedeutet auch, dass die Lösung des Problems die inhaltlichen mathematischen Aspekte des CAMMP Moduls bestimmt, die so eingeführt werden, dass sie direkt zur Lösung beitragen.
Wie bereits deutlich wurde, findet die Lösung von realen Problemen durch die Methode des mathematischen Modellierens statt. In [14, S. 378-379] wird der beschriebene

Organisationsrahmen der Modellierungswochen bzw. -tage als Fördermöglichkeit der Modellierungskompetenz für Schülerinnen und Schüler angeführt. Das mathematische Modellieren wollen wir uns nun genauer anschauen.

2.2. Was bedeutet mathematisches Modellieren und wie soll das gehen?

2.2.1. Begrifflichkeit und Arten von Modellen

Beim mathematischen Modellieren geht es nach Gerd Hinrichs [13, S. 1] um einen Bezug zwischen Realität und Mathematik. Dabei werden reale Situationen mit der Mathematik beschrieben und Probleme möglicherweise gelöst [13, S. 2]. Neben dem Modellieren, bei dem die Beziehung zwischen der Realität und der Modellebene bei komplexen Problemlöseprozessen bewusst betrachtet wird, gibt es nach Kaiser noch drei weitere Arten von Realitätsbezügen im Mathematikunterricht: Eingekleidete mathematische Probleme (Mathematik wird in Anwendungen „eingekleidet“), Veranschaulichungen (z.B. Temperaturskalen mit negativen Zahlen) und Anwendungen (Lösung realer Probleme mit Standardverfahren) [13, S. 4].
Um den Unterschied zwischen dem Modellieren und den anderen Realitätsbezügen besser zu verstehen, wollen wir uns dem Begriff des „Modells“ widmen. Den Begriff beschreibt Hinrichs [13, S. 8] wie folgt:

Was ist ein *Modell*?
Ein Modell ist eine vereinfachende Darstellung der Realität.

Vereinfachend ist die Darstellung deshalb, weil sie nur bestimmte Aspekte der Realität berücksichtigt [13, S. 8]. Je nach Zielsetzung eines Modells werden Arten unterschieden: Man spricht von *deskriptiven Modellen*, wenn ein Aspekt der Realität möglichst gut abgebildet werden soll (Modelle, die vorhersagen, erklären oder beschreiben). Von *normativen Modellen* wird gesprochen, wenn Modelle etwas vorschreiben [13, S. 8-9]. Je nach Zielsetzung gibt es damit zu einer realen Situation viele verschiedene Modelle [13, S. 9].
Der Begriff „Modellieren“ beschreibt einen Prozess. In Hinrichs Worten [13, S. 18] mit angepasster Formulierung:

Was ist *Modellieren*?
Modellieren bezeichnet den gesamten Problemlösungsprozess, für den das Modell verwendet wird.

Dieser Prozess wird im Abschnitt 2.2.2 näher behandelt. Wie es Arten von Modellen gibt, existieren auch Arten des Modellierens. Für uns sind zwei Arten interessant, weshalb diese anhand von [14, S. 362-363] kurz dargestellt werden sollen:
Unter einem *realistischen oder angewandten Modellieren* wird das Anwenden der Mathematik zum Verständnis der Realität verstanden. Diese Modellierung passt zum Vorgehen in der Wissenschaft. Unter einem *pädagogischen Modellieren* versteht man die Förderung von mathematischen Kompetenzen, wobei hier insbesondere die Modellierungskompetenz neben Kommunikations- und Argumentationskompetenzen sowie sozialer Kompetenzen gemeint ist.

2.2.2. Modellierungsprozess

Die Überschrift des Kapitels beinhaltet die Frage „Wie soll das gehen?“. Diese beantworten wir jetzt:
Das Modellieren als Problemlöseprozess wird von Fachdidaktikern in Phasen eingeteilt und häufig durch einen Kreislauf beschrieben [13, S. 18]. Eine solche Beschreibung kann auf viele Arten erfolgen und zahlreiche Fachkundige haben sich bereits Gedanken über Möglichkeiten gemacht. Eine Einteilung und Übersicht versucht Rita Borromeo Ferri in [6]. In der Mathematikdidaktik ist der Modellierungskreislauf von Blum und Leiss gängig [13, S. 18-19]. Seine Vorteile liegen laut Ferri in der Analyse des tatsächlichen Modellierungsprozesses und der Aufschlüsselung von Textaufgaben [6, S. 20-21]. Hinrichs betont, dass dieser Kreislauf die Prozesse fokussiert, die den Schülerinnen und Schülern kognitiv schwer fallen [13, S. 19]. Kommen wir zum Kreislauf nach Blum und Leiss:

Modellierungskreislauf nach Blum und Leiss
In Abbildung 1 ist der Kreislauf nach Blum und Leiss gemäß [13, S. 19] dargestellt. Der Prozess vollzieht sich in sieben Schritten, die nun durch Bezug auf [13, S. 19-31] beleuchtet werden. Abweichende Quellen sind gekennzeichnet:

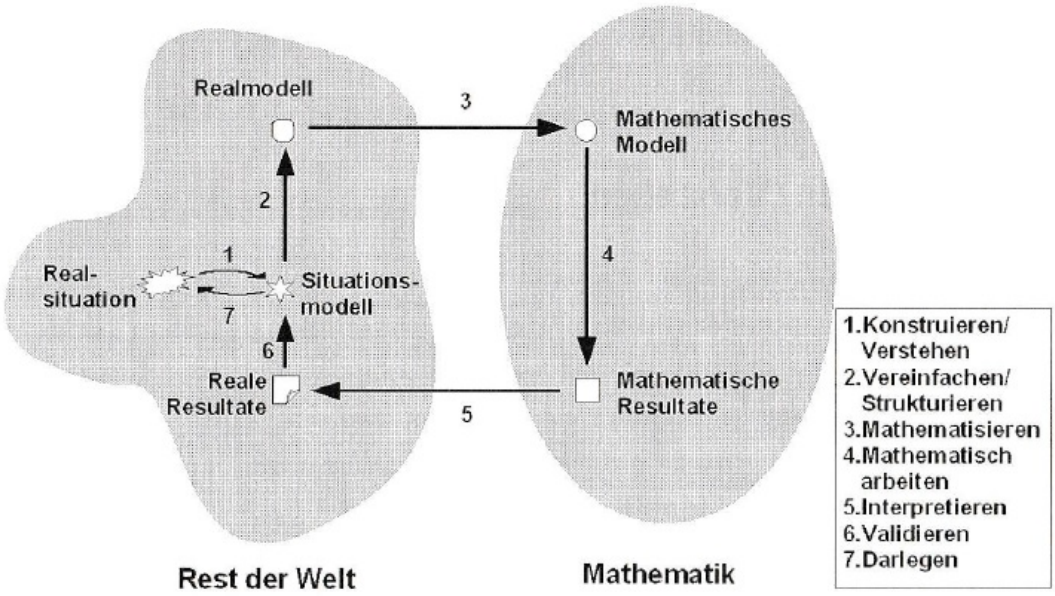


Abbildung 1: **Modellierungskreislauf nach Blum und Leiss.** [13, S. 19]

1. Der erste Schritt ist von der **Realsituation** zu einem Situationsmodell. Genannt wird er *konstruieren/verstehen*. Hierbei muss die Situation, die Aufgabe oder die Fragestellung verstanden werden. Gegebenenfalls ist hier auch eine Datenbeschaffung von Nöten. Ziel ist eine Vorstellung von der Situation, ein **Situationsmodell**.

2. Nun folgt der Schritt vom **Situationsmodell** zum **Realmodell**. Er ist gekennzeichnet durch *vereinfachen/strukturieren*. Wir wollen eine Mathematisierung vornehmen und dabei mit Hilfe der Leitfrage für die Modellierung entscheiden, welche Informationen wichtig und welche nicht wichtig sind. Die Vereinfachungen durch *Annahmen* können häufig eher intuitiv begründet werden [13, S. 11]. Ein Kompromiss zwischen Einfachheit und Angemessenheit der Darstellung des Problems muss gemacht werden. Ein erster Schritt zum mathematischen Modell ist getan.

3. Nun folgt der letztendliche Schritt zum **mathematischen Modell**: das *Mathematisieren*. Ganz trennscharf sind die Schritte 2 und 3 jedoch nicht. In diesem dritten Schritt erfolgt die Überführung des Realmodells in die mathematische Sprache, häufig durch Gleichungen oder Funktionen.

4. Mit dem mathematischen Modell kann nun *mathematisch gearbeitet werden*. Wir

lösen das Problem mit mathematischen Werkzeugen. Gelingt dies, so steht am Ende ein **mathematisches Resultat**.

5. Das Resultat muss nun wieder in die reale Welt (Rest der Welt) überführt werden: Wir *interpretieren* es und sprechen dann von einem **realen Resultat**.

6. Eine Überprüfung des realen Resultats wird nun im Schritt der *Validierung* vollzogen. Dabei können zwei Perspektiven unterschieden werden: Zum einen kann man eine Reflexion der Modellannahmen bezüglich des Modellierungsziels und zum anderen einen Vergleich und eine Bewertung verschiedener Modellierungen vornehmen. Die Validierung kann laut Hinrichs durch Kontrolle der Dimension (Sind die Maßeinheiten sinnig?), der Größenordnung (Ist das Ergebnis genau genug?), des Charakters der Abhängigkeiten (Sind die Zusammenhänge plausibel?), durch Kontrolle extremer Situationen oder Randbedingungen (Wie verhält sich das Modell an besonderen Werten?) oder durch Kontrolle der mathematischen Abgeschlossenheit und Widerspruchsfreiheit (Sind die Resultate bei verschiedenen Modellen deckungsgleich?) sowie durch Kontrolle der Stabilität des Modells (Wie reagiert das Modell auf Veränderungen?) erfolgen.

7. Der letzte Schritt nennt sich *darlegen/erklären*. Die Schülerinnen und Schüler müssen ihre Ergebnisse angemessen kommunizieren und dokumentieren.

Kommt man zu dem Schluss, dass die Modellierung auf eine andere - vermutlich bessere - Art und Weise geschehen kann, so werden die beschriebenen Schritte erneut durchlaufen. Ein deutlich anschaulicheres Bild dieses erneuten Durchlaufes der Schritte ist in einer *Modellierungsspirale* enthalten. Die Begründung dazu gibt Vorhölter, indem sie betont, dass das Modell bei jedem Durchlauf komplexer wird [38, S. 44]. Ein Kreislauf hingegen würde suggerieren, dass man erneut am „Nullpunkt“ ankommt und von Neuem beginnen muss.
Bei jedem Modellierungsschritt können natürlich auch Fehler geschehen und unterschiedliche Kompetenzen werden gefordert. Hinrichs listet eine ganze Reihe solcher möglicher Fehler auf [13, S. 64-75] und beschreibt allgemeine Teilkompetenzen [13, S. 83-92]. Kaiser u. a. [14, S. 369-370] beschreiben ebenfalls verschiedene Teilkompetenzen. An dieser Stelle soll keine Wiederholung dieser sehr allgemeinen Listen geschehen. Bei der Analyse des erstellten Lernmoduls wird stattdessen konkret auf Unterstützungsmaßnahmen und Anforderungen für die Lernenden eingegangen.

Modellierungskreislauf im CAMMP

Der beschriebene Kreislauf nach Blum und Leiss ist laut Hinrichs für Schülerinnen und Schüler häufig zu komplex, weshalb er didaktisch reduzierte Schemata bevorzugt [13, S. 19]. Im CAMMP day wird der folgende vereinfachte Kreislauf in Abbildung 2 behandelt.

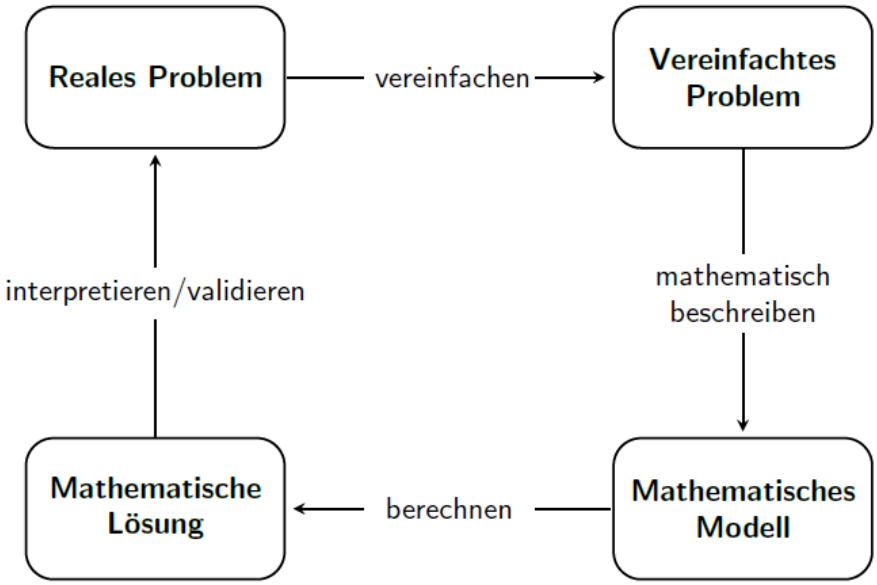


Abbildung 2: **Modellierungskreislauf im CAMMP.**

Mit Blick auf die Abbildung stellt man fest, dass hier nur vier Schritte unterschieden werden:

1. Am Anfang steht das **reale Problem**. Mit diesem beschäftigen sich die Schülerinnen und Schüler und vollziehen hier den Schritt „konstruieren/verstehen“ von Blum und Leiss. Dieses Problem wird *vereinfacht* und so zu einem **vereinfachten Problem**. Der Schritt „vereinfachen/strukturieren“ zum Realmodell ist hier zu finden.

2. Nun kommt die Mathematik ins Spiel und führt uns durch *mathematisches Beschreiben* zum **mathematischen Modell**. Wir haben den Schritt „mathematisieren“ getätigt.

3. Das *Berechnen* liefert uns eine **mathematische Lösung**. Blum und Leiss nennen dies „mathematisch arbeiten“ und „mathematische Resultate erzeugen“.

4. Die Lösung wird schließlich *interpretiert bzw. validiert* und auf das reale Problem

bezogen. Die Schritte „interpretieren“ und „validieren“ sind hier zusammengefasst.

5. Das „Darlegen“ von Blum und Leiss ist nicht explizit erwähnt, wird aber, wie später deutlich wird, im CAMMP an vielen Stellen benötigt.

Eine deutliche Reduzierung des Kreislaufs von Blum und Leiss ist damit vor allem auf der Seite „Rest der Welt“ geschehen.

2.2.3. Wichtigkeit und Vorteile von Modellen und Modellierungsaufgaben

Modelle und Modellierung sind laut Hinrichs besonders interessant, da mit Hilfe dieser *Simulationen* von Situationen getätigt werden können, die nicht eintreten können oder sollen [13, S. 11]. Er nennt als Beispiele ökologische Aspekte, wie Waldbrände, gesellschaftliche Themen, wie Rentenversicherung, und planerische Aspekte, wie die Planung von Projekten [13, S. 11-12]. Modelle helfen somit, objektive Entscheidungen, zumindest mit einer bestimmten Näherung an die Realität, treffen zu können [13, S. 12-13].
Der Bezug zur Realität ist hierbei ein großer Vorteil von Modellen und Modellierungsaufgaben. Hinrichs formuliert einige dieser Vorteile als Anforderungen an „echte“ Modellierungsaufgaben [13, S. 76-81]: Er nennt hier die *Authentizität*. Es handelt sich um wirkliche Probleme, die gelöst werden sollen. Bei Modellierungsaufgaben gibt es einen bestimmten Anteil an *Offenheit*, denn bei der Erstellung eines Modells gibt es nicht eine Lösung sondern verschiedene Möglichkeiten. Auch in der Lösung ist eine gewisse Offenheit vorhanden, z.B durch Probieren, Experimentieren oder Messen. Zudem haben solche Aufgaben *Differenzierungsvermögen*. Unterschiedlich schwere Modelle können erarbeitet werden. Durch Schwerpunktsetzung auf bestimmte Teilschritte des Modellierens kann außerdem auf unterschiedliche *Anforderungsniveaus* und bessere *Verständlichkeit* eingegangen werden.
Mit Hilfe von Modellierungen können auch fächerübergreifende Aspekte behandelt werden (wie im ersten Absatz bereits angesprochen). Hierbei bilden die *Verantwortung und Ethik* einen möglichen Schwerpunkt. Herget und Maaß erläutern dies in [12]. Beim Modellieren kommt laut dieses Artikels auch die individuelle und gesellschaftliche Verantwortung von Wissenschaft zum Tragen, denn laut den beiden Autoren „haben die Entscheidungen, was mit welcher Absicht modelliert und ausgerechnet wird, mögliche Folgen - für andere Menschen und uns selbst.“ [12, S. 6]

2.2.4. Modellieren mit digitalen Werkzeugen

Das Modellieren mit Hilfe von digitalen Werkzeugen, in unserem Fall der Computer mit dem Programm MATLAB, hat vielerlei Vorteile für den Modellierungsprozess. Hinrichs nennt hierzu [13, S. 93-95] unter anderem, dass Rechenfehler vermieden oder Rechnungen von den digitalen Werkzeugen übernommen werden können und der Fokus auf den Modellierungsprozess an sich gerichtet werden kann. Zudem erwähnt er, dass ein Ausprobieren mit Hilfe von digitalen Werkzeugen in einer angemessenen Zeit vollzogen werden kann und dass realistisches Datenmaterial verwendet werden kann. Kaiser u. a. betonen zudem in [14, S. 371-373] das Potential der digitalen Werkzeuge auch beim Experimentieren und Entdecken. Die bereits genannten Simulationen können ebenfalls nach [14, S. 371-373] mit diesen Werkzeugen auch im stochastischen Bereich durchgeführt werden. Der Einsatz ist gemäß [14, S. 371-373] in allen Schritten des Modellierungsprozesses sinnvoll und nützlich.

2.3. Was bedeutet Problemlösen?

Im Zuge der Erklärungen zum Modellieren wurde immer wieder von „Problemen“ und von einem „Problemlöseprozess“ gesprochen. Das Problemlösen ist ein Fachbegriff der Mathematikdidaktik. Da das Modellieren nicht ohne Problemlösen gedacht werden kann, wie auch in [11, S. 282] geschrieben ist, soll diese Thematik in Grundzügen dargelegt werden.

2.3.1. Begrifflichkeit und Arten von Problemen

Jeder von uns hatte schon ein „Problem“. Unsere alltäglichen Probleme sind sehr vielfältig. Unter einem mathematischen Problem verstehen wir nach [11, S. 279-280]:

Was ist ein *Problem*?
Von einem Problem oder einer Problemaufgabe wird gesprochen, wenn ein Individuum ein Ziel hat, derzeit jedoch nicht bzw. nicht genau weiß, wie es dieses (momentan) erreichen kann.

Soll das Ziel erreicht werden, so wird das Problem gelöst. In der Fachdidaktik heißt das gemäß [11, S. 280]:

Was ist *Problemlösen*?
Problemlösen ist der Prozess der Überführung eines Anfangszustandes in einen Zielzustand, bei dem gewisse Schwierigkeiten bzw. Barrieren überwunden werden müssen.

Mathematisches Problemlösen kann gelernt werden - mehr noch: Das Lösen von Problemen kann als grundlegendes Motiv, Mathematik zu betreiben, gesehen werden [11, S. 280].

Wie unsere alltäglichen Probleme sind auch die mathematischen Probleme vielfältig. Sie können aber typisiert werden. Polya und Kranz vollziehen dies nach [11, S. 281-282] so:

- Bestimmungsaufgaben, z.B. berechnen, konstruieren, bestimmen
- Entscheidungsaufgaben, z.B. beweisen, überprüfen
- Entdeckungsaufgaben, z.B. Vermutungen aufstellen, entdecken, auffinden

Dies soll uns zu diesem Abschnitt genügen.

2.3.2. Problemlöseprozess

Wie beim Modellierungsprozess gibt es auch beim Problemlösen eine Vielfalt von Verlaufsmodellen. Ein Modell, auf das viele andere aufbauen, geht auf Polya zurück [11, S.286-287]. Es beschreibt die Schritte wie folgt [11, S.286-287]:

1. Verstehen einer Aufgabe
2. Ausdenken eines Plans
3. Ausführen des Plans
4. Rückschau und Ausblick

Hierbei kann es jedoch auch zu Sprüngen zwischen den Schritten und zu Zyklen innerhalb der Schritte kommen.

Nach dieser Darlegung des fachdidaktischen Hintergrundes folgt nun die Erläuterung des fachlichen Hintergrundes.

3. ...das hat das mit Mathe zu tun! - Fachlicher Hintergrund

Die vorliegende Arbeit basiert auf dem Artikel „Online Privacy as a Collective Phenomenon“ von Sarigol, Garcia und Schweitzer aus dem Jahr 2014 [9]. Von diesem Ausgangspunkt aus erfolgte eine weitere Erforschung der Thematik mit leicht abgewandelten, einfacheren Methoden. Der fachliche Hintergrund wird gemäß der CAMMP Philosophie dargelegt, indem der Forschungsprozess (in Teilen) durchlaufen und an den entscheidenden Stellen die notwendige Mathematik eingeführt und erläutert wird. Konkrete Forschungsergebnisse werden hier nicht dargestellt. Sie sind in den Lösungsvorschlägen zu den Materialien des CAMMP days dargelegt.

3.1. Problemstellung

Sarigol, Garcia und Schweitzer führten eine Studie durch mit dem Ziel, zu zeigen, in welchem Maße Informationen über Nutzer eines sozialen Netzwerks gebraucht werden können, um Wissen über Personen der Gesellschaft zu quantifizieren [9, S. 2]. Konkret gingen sie zwei verschiedenen Möglichkeiten nach. Zum einen untersuchten sie die Erstellung sogenannter *Partial Shadow Profiles*:

Definition *Partial Shadow Profile* [9, S. 4]:
Nutzer des sozialen Netzwerks haben ein gespeichertes Profil. Dieses beinhaltet bestimmte Informationen. Werden zusätzliche Informationen über einen Nutzer generiert, die er aber nicht selber angegeben hat, so sprechen wir von einem Partial Shadow Profile.

Eine solche Situation betrifft alle Nutzer eines sozialen Netzwerks. Zum anderen beschäftigten sie sich mit *Full Shadow Profiles*:

Definition *Full Shadow Profiles* [9, S. 6]:
Full Shadow Profiles sind erstellte Profile von Personen, die nicht Mitglied des Netzwerks sind.

Hierbei werden die Informationen über Nutzer des Netzwerks, zum Beispiel ihre Kontaktlisten, die auch Nicht-Nutzer des Netzwerks enthalten, gebraucht [9, S. 6]. Stimmt man den Nutzungsbedingungen von WhatsApp [39] zu, so billigt man: „In order to provide the WhatsApp Service, WhatsApp will periodically access **your**

address book or contact list on your mobile phone to locate the mobile phone numbers of other WhatsApp users ("in-network" numbers), or otherwise categorize other mobile phone numbers as "out-network" numbers, which are stored as one-way irreversibly hashed values."

In beiden Situationen geht es also um die Generierung von Informationen über Personen. In der Studie wurde hierzu eine spezielle Information fokussiert. Es handelt sich um die *sexuelle Orientierung* der Nutzer, die als sensible und relevante Information angesehen wird [9, S. 2]. Zudem gehen die Forscher davon aus, dass die sexuelle Orientierung in Kombination mit dem Geschlecht inhomogen verteilt ist, wie viele andere Merkmale einer Person (z.B. ihre politische Gesinnung) auch [9, S. 2]. Dazu wurde vereinfachend die sexuelle Orientierung gemäß folgender Annahme festgelegt:

Definition *Sexuelle Orientierung*: [9, S. 4]
Das Geschlecht eines Nutzers ist entweder weiblich oder männlich. Die sexuelle Orientierung wird in Abhängigkeit seines Interesses an einer Beziehung oder an einem Treffen (hier im Sinne eines „Dates") von Nutzern eines bestimmten Geschlechts bestimmt:

- Ist ein Nutzer an einer Beziehung oder an einem Treffen mit einer Person gleichen Geschlechts interessiert, so wird er als **homosexuell** eingestuft.
- Ist umgekehrt ein Nutzer an einer Beziehung oder an einem Treffen mit einer Person des anderen Geschlechts interessiert, so wird er als **heterosexuell** eingestuft.

Die Nutzer müssen, um ihnen eine eindeutige sexuelle Orientierung zuordnen zu können, ein eindeutiges Geschlecht und eine der beschriebenen Interessen in ihrem Profil angegeben haben. Bisexuelle Nutzer sowie eine Differenzierung der Geschlechter werden im Rahmen des CAMMP days als weitere Vereinfachung nicht berücksichtigt. Für Partial Shadow Profiles (und für Full Shadow Profiles) stellen wir uns im CAMMP day die zentrale Frage:

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Dazu betrachten wir jeden Nutzer mit seinen Freundschaften und modellieren ein Netz-

werk als einen Graphen.

3.2. Das soziale Netzwerk als ungerichteter Graph

Graphen werden als Modelle für netzartige Strukturen genutzt [37, S. 11]. Eine Illustration ist in Abbildung 3 zu sehen. Wie zu erkennen ist, gibt es in einem Graphen zwei verschiedene Objekte [37, S. 11]. Wir sehen Punkte und Striche. Die Punkte werden als **Knoten** und die Striche als **Kanten** bezeichnet [37, S. 11]. Zusätzlich bezeichnet man einen Graphen als **ungerichtet** (oder gerichtet) [37, S. 12]. In der Sprache der Mathematik bedeutet das:

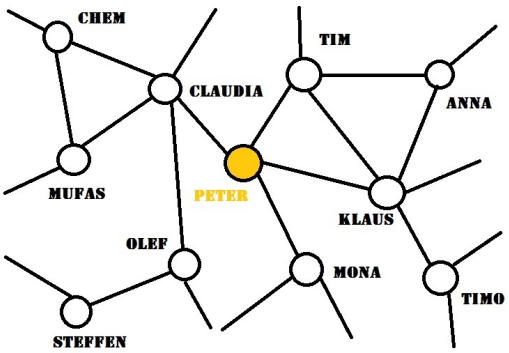


Abbildung 3: **Das Netzwerk als ungerichteter Graph.** Eine Veranschaulichung.

Definition *Ungerichteter Graph* [37, S. 12]:
Ein ungerichteter Graph $G = (V, E)$ besteht aus einer (nichtleeren) Knotenmenge V und einer (nichtleeren) Kantenmenge E, wobei jeder Kante $e \in E$ von G zwei (nicht notwendig verschiedene) Knoten aus V zugeordnet sind.

Bemerkung [37, S. 12]:
Eine Kante $e \in E$ wird in der Form $e = \{u, v\}$ beschrieben, wobei $u \in V$ und $v \in V$ dann Endknoten der Kante e heißen.

Wir übertragen diese Bezeichnungen auf unser soziales Netzwerk: Schnell ist anhand des Bildes klar, dass die Knoten unsere Nutzer und die Kanten die Freundschaften zwischen den Nutzern modellieren. Besteht eine Freundschaft zwischen Nutzer A und Nutzer B, so bezeichnen wir die Nutzer A und B als Endknoten der Freundschaft zwischen A und B. In unserem Fall sind die Knoten, die von einer Kante verbunden werden, verschieden. Eine Freundschaft mit sich selbst ergibt hier keinen Sinn. Kümmern wir uns etwas genauer um den Begriff „ungerichtet“: Laut Definition sollen jeder Kante zwei Knoten zugeordnet werden. Das ist in sozialen Netzwerken der Fall, denn wenn Nutzer A mit Nutzer B befreundet ist, so ist auch Nutzer B mit Nutzer A befreundet. Eine Freundschaft ist sozusagen kommutativ und wird laut der

Bemerkung durch die Personen, zwischen denen sie besteht, beschrieben.

Nebenbei bemerkt ist ein Graph *gerichtet*, wenn Kanten auftreten, die einen Richtungssinn haben [37, S. 12]. Eine solche Situation tritt bei uns jedoch nicht auf. Wir wollen noch weitere Begriffe aus der Graphentheorie einführen und diskutieren:

Definition *Ungerichteter endlicher Graph* [37, S. 12]:
Ein ungerichteter Graph $G = (V, E)$ heißt endlich, wenn die Mengen V und E endlich sind.

In unserem Fall trifft diese Bezeichnung zu, denn die Anzahl der Nutzer (Mächtigkeit der Menge V) und die der Freundschaften (Mächtigkeit der Menge E) sind natürliche Zahlen[1]. Wenn wir die Freundschaft zwischen zwei Nutzern betrachten, so kann ein weiterer Fachbegriff aus der Graphentheorie helfen, den wir auch im Folgenden nutzen wollen:

Definition *Adjazente Knoten* [37, S. 13]:
Zwei Knoten $u, v \in V$, die durch eine Kante $e = \{u, v\}$ verbunden sind, heißen adjazent (benachbart) in G.

Wir können also befreundete Nutzer als adjazent bezeichnen. Diese Ausführungen sollen uns zunächst genügen. Um die Fragestellung untersuchen zu können, sind zunächst konkrete Daten eines sozialen Netzwerks notwendig.

3.3. Daten eines sozialen Netzwerks

Wir nutzen Daten des Netzwerks *Friendster*. Zu diesem Netzwerk ein paar erläuternde Worte:

3.3.1. Friendster

Das Netzwerk Friendster wurde im Jahr 2002 gegründet und hatte bis zu 350 Millionen Nutzer [15]. Es gilt als Vorbild für andere Netzwerke, wie zum Beispiel *Facebook* [22]. Aktuell ist das Netzwerk nicht mehr zu erreichen. Auf seiner Internetseite [8] wird eine Pause bis 2015 verkündet, die anscheinend bis heute anhält. Doch die öffentlichen

[1] Eine Menge M heißt *endlich*, wenn $|M| \in \mathbb{N}_0$. $|M|$ ist die Anzahl der paarweise verschiedenen Elemente der Menge M. [16, S. 161]

Daten der Nutzer des Netzwerks sind gespeichert worden und können im Internetarchiv [7] heruntergeladen werden.

3.3.2. Grundlegende Annahme

An dieser Stelle kommen wir zu einer weiteren vereinfachenden Annahme bei der Modellierung:

Annahme: Wir gehen von der Richtigkeit der Angaben der Nutzer aus.

3.3.3. Daten aufbereiten

Wir werden die ersten 1 Millionen Profile analysieren, also die Profile der Personen, die sich in ihrer Beitrittsreihenfolge unter den ersten 1 Millionen befinden. Hierbei handelt es sich vor allem um Nutzer aus den USA [9, S. 3]. Diese Profile sind von unterschiedlichem Informationsgehalt, denn es sind nur die Informationen über die Nutzer vorhanden, die diese angegeben und öffentlich gemacht haben. Ein Profil kann aus den folgenden Informationen bestehen [9, S. 3]:

- Nutzer ID (gibt die Mitgliedsnummer an, Beitrittsreihenfolge ersichtlich)
- Name als Wort
- Geburtsdatum
- Geschlecht (männlich, weiblich oder unspezifisch)
- Interessen (Freunde, Fans, Partner für Aktivitäten, nur herumschauen, Frauen treffen, Männer treffen, Beziehung mit Frauen, Beziehung mit Männern, Beziehung mit Frauen und Männern, Frauen und Männer treffen)[2]
- Beziehungsstatus (allein stehend, verheiratet, in einer Beziehung, in Gemeinschaft lebend, es ist kompliziert)

Für die Studie sind die Profile interessant, die ihre Informationen veröffentlicht haben und deren Freundschaften sichtbar sind. Das bedeutet, dass nicht alle Profile genutzt werden können. Hier setzt die Filterung ein:

[2]Hier können mehrere Aspekte angegeben werden.

- Herausgenommen wurden im ersten Schritt alle Nutzer, deren sexuelle Orientierung (hetero- oder homosexuell) aufgrund fehlender oder nicht eindeutiger Daten nicht bestimmt werden kann.

- Im zweiten Schritt wurden nur die Nutzer erhalten, die mindestens eine Freundschaft zu Nutzern nach dem ersten Filterungsschritt haben.

Konkret erhalten wir einen Datensatz bestehend aus 56284 Nutzern. Diesen wurden die Nummern $1 - 56284$ in ihrer Beitrittsreihenfolge neu zugeordnet. Innerhalb von diesen Nutzern gibt es 183581 ungerichtete Freundschaften. Jede Person hat somit im Durchschnitt ca. 3.26 Freundschaften.
Die Freundschaften in unserem Datensatz können aufgrund der Tatsache, dass befreundete Nutzer in unserem Modell adjazent sind, durch eine Adjazenzmatrix dargestellt werden:

Definition *Adjazenzmatrix* [37, S. 29]:
Es sei $G = (V, E)$ ein ungerichteter Graph mit $n \in \mathbb{N}$ Knoten. Die Adjazenzmatrix $A(G)$ (oder kurz A) von G ist eine $n \times n$ Matrix mit den Eintragungen a_{ij} = Anzahl der Kanten zwischen i und j $(i, j \in \{1, \dots n\})$.

In unserem Fall besteht maximal eine Kante zwischen zwei Nutzern (entspricht maximal einer Freundschaft). Die Adjazenzmatrix für die (neu nummerierten) ersten 1000 Nutzer ist in Abbildung 4 dargestellt. Die blauen Punkte symbolisieren eine Freundschaft. In der Matrix sind viele Einträge nicht blau, d.h. man kann vermuten, dass ein Nutzer hier im Allgemeinen nicht viele Freundschaften hat. Dies wird im nächsten Abschnitt in Tabelle 1 bestätigt. Die Matrix ist außerdem symmetrisch[3]. Diese Eigenschaft ist logisch erschließbar, da zuvor schon aus der Modellierung deutlich wurde, dass Freundschaften stets symmetrisch sind. Des Weiteren fällt eine Häufung der Punkte auf der Diagonalen auf. Nutzer, deren Nummern nahe beieinander liegen, sind demnach eher miteinander befreundet.

[3]Eine reelle Matrix A heißt *symmetrisch*, wenn $A^{tr} = A$, ihr Transponiertes gleich der Matrix selbst ist [27, S. 73].

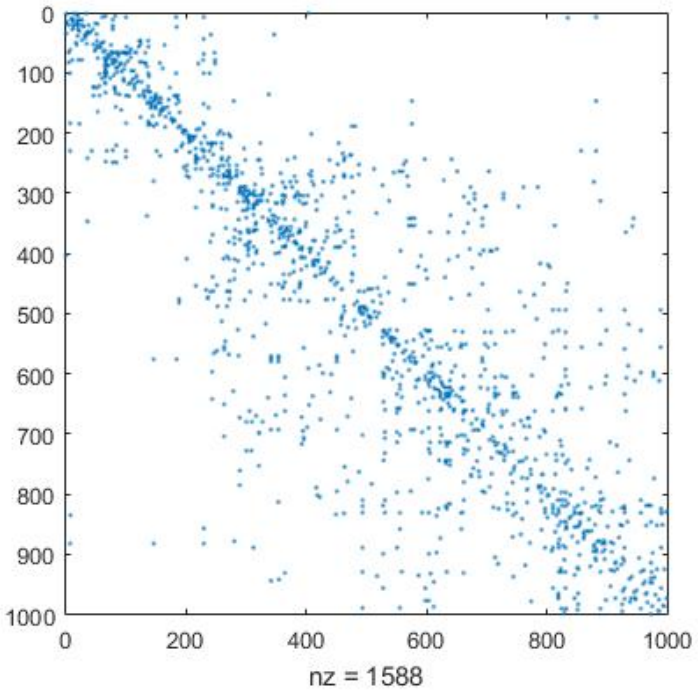


Abbildung 4: **Adjazenzmatrix des Datensatzes der ersten 1000 Nutzer.** Blaue Punkte symbolisieren eine Kante.

3.3.4. Die Nutzer und ihre Eigenschaften

Wir schauen uns die nach der Filterung verbliebenen 56284 hetero- und homosexuellen Nutzer und ihre Eigenschaften genauer an. Dabei spielen sogenannte Freunde mit Abstand oder Ordnung 1 oder 2 eine Rolle. Um diesen Ausdruck zu verstehen, sind einige neue Bezeichnungen einzuführen:

Definition *Kantenfolge* [37, S. 15]:
Sei $k \in \mathbb{N}$. Eine Kantenfolge ist eine Folge $v_1, e_1, v_2, e_2, v_3, \ldots, v_{k-1}, e_{k-1}, v_k$ von Knoten $v_i \in V, i \in \{1, \ldots k\}$, und Kanten $e_i \in E, i \in \{1, \ldots k\}$, eines Graphen G, sodass die Kante e_i jeweils die Endknoten v_i und v_{i+1} besitzt.

Eine solche Kantenfolge wäre in Abbildung 3 zum Beispiel die Folge

- *Chem, Freundschaft zwischen Chem und Claudia, Claudia, Freundschaft zwischen Claudia und Peter, Peter, Freundschaft zwischen Klaus und Peter, Klaus* oder
- *Chem, Freundschaft zwischen Chem und Claudia, Claudia, Freundschaft zwi-*

schen Claudia und Peter, Peter, Freundschaft zwischen Klaus und Peter, Klaus, Freundschaft zwischen Klaus und Peter, Peter.

Interessant ist für uns auch die Bezeichnung einer Länge einer Kantenfolge:

Definition *Länge der Kantenfolge* [37, S. 15]:
Die Länge der Kantenfolge ist die Anzahl der Kanten dieser Folge.

In den betrachteten Beispielen haben wir Längen von:

- *Chem, Freundschaft zwischen Chem und Claudia, Claudia, Freundschaft zwischen Claudia und Peter, Peter, Freundschaft zwischen Klaus und Peter, Klaus* mit Länge 3.
- *Chem, Freundschaft zwischen Chem und Claudia, Claudia, Freundschaft zwischen Claudia und Peter, Peter, Freundschaft zwischen Klaus und Peter, Klaus, Freundschaft zwischen Klaus und Peter, Peter* mit Länge 4.

Formal gesehen kann auch ein einzelner Knoten eine Kantenfolge sein, wobei die Länge dann als Null bestimmt wird [37, S. 15]. Es gibt einen für uns bedeutenden Spezialfall einer Kantenfolge, es ist:

Definition *Weg* [37, S. 15]:
Eine Kantenfolge in einem Graphen $G = (V, E)$ ist ein Weg, wenn jeder Knoten aus V höchstens einmal in dieser Folge auftritt.

Betrachten wir noch einmal unsere Beispiele:

- *Chem, Freundschaft zwischen Chem und Claudia, Claudia, Freundschaft zwischen Claudia und Peter, Peter, Freundschaft zwischen Klaus und Peter, Klaus*
  Es treten die Personen (also die Knoten) Chem, Claudia, Peter, Klaus auf. Diese werden genau einmal genannt. Andere Personen werden nicht genannt. Diese Kantenfolge ist demnach ein Weg.
- *Chem, Freundschaft zwischen Chem und Claudia, Claudia, Freundschaft zwischen Claudia und Peter, Peter, Freundschaft zwischen Klaus und Peter, Klaus, Freundschaft zwischen Klaus und Peter, Peter*
  Es treten Chem, Claudia, Peter, Klaus auf. Chem, Claudia und Klaus werden genau einmal genannt. Peter wird jedoch genau zweimal genannt. Diese Kantenfolge erfüllt somit die Bedingungen eines Weges nicht.

Mit der Länge einer Kantenfolge haben wir bereits die Länge eines Weges, als spezielle Kantenfolge, definiert und können zur nächsten Bezeichnung übergehen:

Definition *Abstand von zwei Knoten* [37, S. 33]:
Der Abstand $d(u, v)$ von zwei Knoten $u \in V$ und $v \in V$ eines Graphen $G = (V, E)$ ist die Länge eines kürzesten Weges von u nach v. Wenn kein solcher Weg existiert, so setzen wir $d(u, v) = \infty$. Der Abstand $d(v, v)$ eines Knoten zu sich selbst ist stets gleich Null.

Aus der Definition geht hervor, dass der kürzeste Weg nicht eindeutig sein muss. Im Bild 3 haben wir beispielsweise: Der Abstand von Chem zu Klaus ist 3 (ein kürzester Weg wäre: *Chem, Freundschaft zwischen Chem und Claudia, Claudia, Freundschaft zwischen Claudia und Peter, Peter, Freundschaft zwischen Klaus und Peter, Klaus*). Mit diesem theoretischen Hintergrund können wir nun die Begriffe der Freunde 1. oder 2. Ordnung festlegen:

Definition *Freunde 1. oder 2. Ordnung*:
Wir betrachten einen bestimmten Nutzer und bezeichnen alle Personen als *Freunde 1. Ordnung des Nutzers*, deren Knoten den Abstand 1 zum Knoten des betrachteten Nutzers haben. Entsprechend bezeichnen wir alle Personen als *Freunde 2. Ordnung des Nutzers*, deren Knoten den Abstand 2 zum Knoten des betrachteten Nutzers haben.

Anschaulich gesprochen sind Freunde 1. Ordnung die direkten Freunde eines Nutzers und Freunde 2. Ordnung die Freunde der Freunde des Nutzers, wobei immer die kleinere Ordnung gezählt wird. Zur Erläuterung ein Beispiel:
In Abbildung 5 sind die Freunde 1. Ordnung von Peter orange und die Freunde 2. Ordnung blau gekennzeichnet. Peters Freunde 1. Ordnung sind Tim, Klaus, Mona und Claudia. Anna, Timo, Olef, Chem und Mufas sind Freunde 2. Ordnung. Obwohl

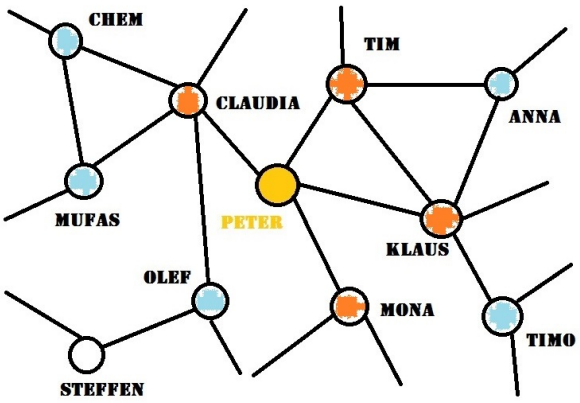


Abbildung 5: Peter und seine Freunde.
Peters Freunde 1. Ordnung
Peters Freunde 2. Ordnung

beispielsweise Tim und Klaus als Freunde 1. Ordnung miteinander befreundet sind und damit zu den Freunden von Peters Freunden gehören, sind sie nicht Freunde 2. Ordnung. Der kürzeste Weg wird gezählt.
Eine Übersicht über die Eigenschaften der Nutzer unseres Datensatzes gibt Tabelle 1.

Tabelle 1: Informationen über den Datensatz

- Anzahl der Nutzer: 56284
- Anzahl der homosexuellen Nutzer: 11174 (Anteil: 0.1985)

Merkmal eines Nutzers	**Homosexuelle**	**Heterosexuelle**
Anzahl	11174	45110
Mittlere Anzahl Freunde 1. Ordnung	3.8799	3.1086
Mittlere Anzahl homosex. Freunde 1. O. (Anteil)	3.018 (0.7778)	0.2136 (0.0687)
Mittlere Anzahl heterosex. Freunde 1. O. (Anteil)	0.862 (0.2222)	2.895 (0.9313)
Mittlere Anzahl Freunde 2. Ordnung	29.3882	18.4629
Mittlere Anzahl homosex. Freunde 2. O. (Anteil)	21.8854 (0.7447)	1.8585 (0.1007)
Mittlere Anzahl heterosex. Freunde 2. O. (Anteil)	7.5028 (0.2553)	16.6044 (0.8993)
Interessen in Anteilen		
Freunde	0.9565	0.9607
Partner für Aktivitäten	0.8135	0.8310
Männer treffen	0.8764	0.3182
Frauen treffen	0.0985	0.6581
Beziehung mit Männern	0.7181	0.1956
Beziehung mit Frauen	0.0541	0.4524
Beziehungsstatus in Anteilen		
unbekannt	0.0311	0.0345
es ist kompliziert	0.0121	0.0074
allein stehend	0.8666	0.9027
in einer Beziehung	0.0797	0.0468
verheiratet	0.0106	0.0086

Diese Daten bilden die Grundlage, um Vorhersagestrategien über die Nutzer zu entwickeln. Dabei stellt man fest, dass viele Informationen nicht zur Vorhersage der sexuellen Orientierung geeignet sind. Zur Definition der sexuellen Orientierung nutzen wir die Interessen „Männer/Frauen treffen“ und „Beziehung mit Männern/Frauen“. Sie kön-

nen daher nicht als Basis einer Vorhersage dienen. Außerdem geben Interessen, wie „Freunde“ oder „Partner für Aktivitäten“ sowie ein Beziehungsstatus der Form von „unbekannt“, „es ist kompliziert“, „allein stehend“ und „in einer Beziehung“ keine sinnvollen Hinweise. Lediglich der Status „verheiratet“ kann genutzt werden mit der Idee, dass die Hochzeit von homosexuellen Paaren in den USA erst seit 2015 landesweit legalisiert wurde [28]. Demnach bietet sich vor allem eine Arbeit mit den Freunden der Nutzer an, um Vorhersagen zur sexuellen Orientierung zu machen.
Dabei stellen die folgenden beiden Abbildungen die Verteilungen der Freundschaften mit homosexuellen Nutzern dar:

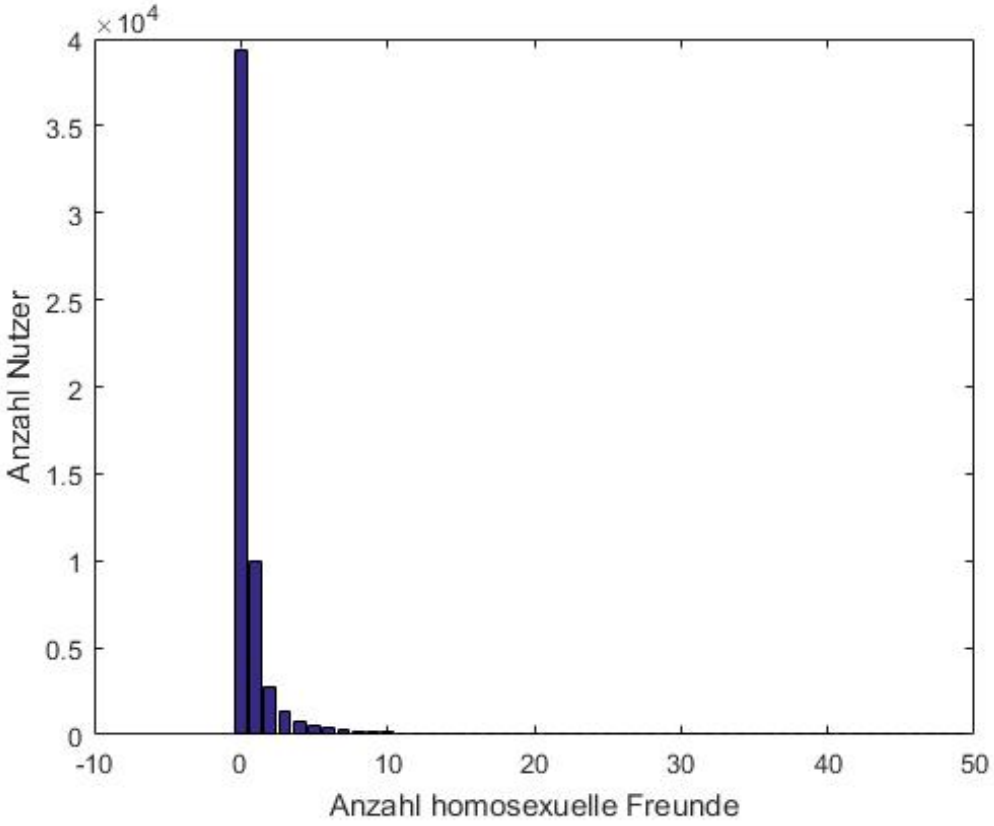


Abbildung 6: **Säulendiagramm zur Anzahl der Nutzer pro Anzahl der homosexuellen Freunde bis 50.** Die Säule zu 50 homosexuellen Freunden stellt die Anzahl der Nutzer mit 50 oder mehr homosexuellen Freunden dar.

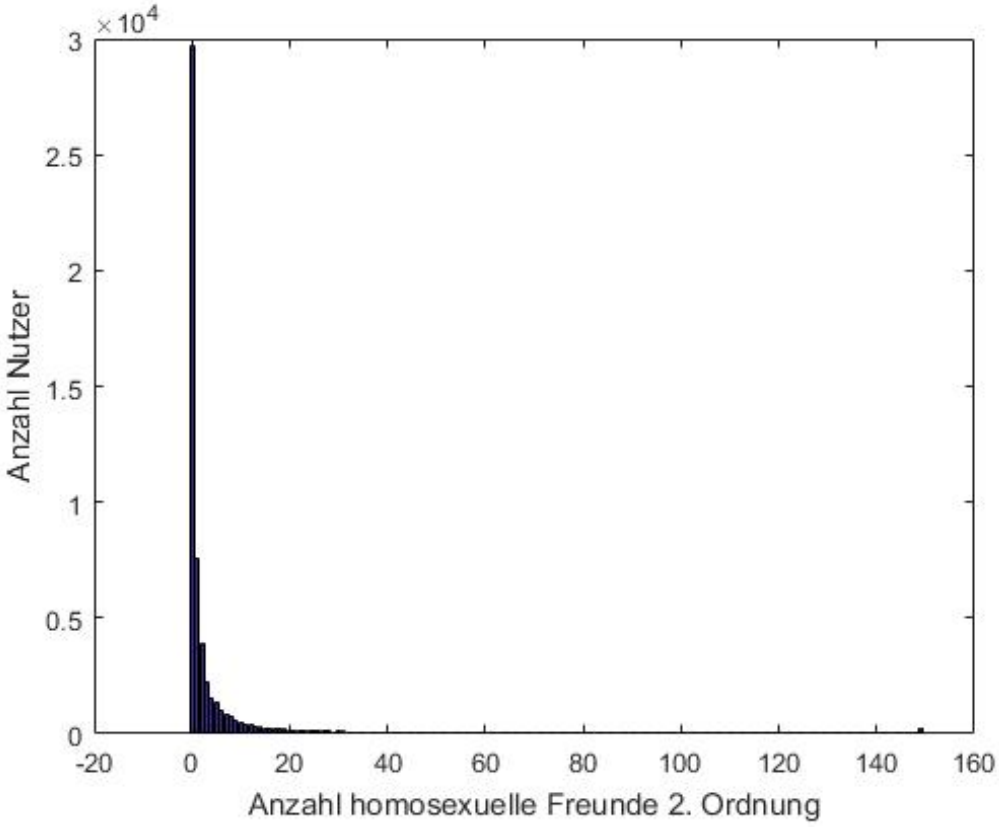


Abbildung 7: **Säulendiagramm zur Anzahl der Nutzer pro Anzahl der homosexuellen Freunde 2. Ordnung bis 150.** Die Säule zu 150 homosexuellen Freunden 2. Ordnung stellt die Anzahl der Nutzer mit 150 oder mehr homosexuellen Freunden 2. Ordnung dar.

Abbildung 6 zeigt die Anzahl der Nutzer pro Anzahl der homosexuellen Freunde bis 50. Die ersten Balken sind deutlich zu sehen, was bedeutet, dass die meisten Nutzer zwischen 0 und 15 homosexuelle Freunde haben. In der Tat sind es 55981, also ca. $99,46\%$, die diese Aussage betrifft. Abbildung 7 verdeutlicht die Anzahl der Nutzer pro Anzahl der homosexuellen Freunde 2. Ordnung bis 150. Hier sind die Säulen ab 35 kaum noch zu erkennen, wobei an der letzten Stelle mit mehr als 150 homosexuellen Freunden 2. Ordnung nochmal eine kleine Säule zu sehen ist. Es sind 53984 Nutzer, die zwischen 0 und 35 homosexuelle Freunde 2. Ordnung haben. Das ist ein Prozentsatz von ca. $95,91\%$. Dieses Wissen wird im späteren Verlauf von Nutzen sein.

```
                          Valid: 1
                             ID: 100
                            Age: 41
                         Gender: 2
               RelationshipStatus: 5
                        Interests: [4 6 2 3]
                   NumberFriends: 1
         NumberFriendsSecondOrder: 3
                        Friends: 66
                         Hetero: 1
                      OriginalID: 932
                NumberGayFriends: 1
             NumberHeteroFriends: 0
       NumberGayFriendsSecondOrder: 1
    NumberHeteroFriendsSecondOrder: 2
```

Abbildung 8: **Gespeicherte Daten über Nutzer 100.**

Zu jedem der 56284 Nutzer wurden seine Eigenschaften gespeichert. Abbildung 8 zeigt die hinterlegten Daten des Nutzers mit der (neuen) ID 100. Die alte ID war die Nummer 932. Die Codierung der Eigenschaften erfolgte in englischer Sprache nach dem folgenden Schema, wobei die Zahlen zu Eigenschaften, die kein Nutzer hat, hier nicht aufgeführt sind:

Codierung:

- Gender: 1 - Male; 2 - Female
- Relationship Status: 0 - Unknown; 1 - It's complicated; 2 - Single; 4 - In Relationship; 5 - Married
- Interest: 2 - Friends; 3 - Activity Partners; 4 - Dating Men; 5 - Dating Women; 6 - Relationship with Men; 7 - Relationship with Women
- Hetero: 0 - gay; 1 - hetero

Wir können also beispielsweise ablesen, dass der Nutzer mit der Nummer 100 weiblich und verheiratet sowie heterosexuell ist und Freunde für Aktivitäten sucht.

Nachdem die Daten in dieser Weise aufbereitet und somit vereinfacht wurden, erfolgte die Untersuchung der Fragestellung.

3.4. Untersuchung der Fragestellung

3.4.1. Vorgehen

Ziel der Arbeit ist es, zu zeigen, dass mit einfachen Methoden gute Vorhersagen der sexuellen Orientierung eines Nutzers getätigt werden können. Wir orientieren uns am Vorgehen, maschinelles Lernen einzusetzen, und wenden uns diesem kurz zu:

Maschinelles Lernen
Ziel des maschinellen Lernens ist es unter anderem, Regeln automatisch zu erwerben [24, S. 2]. Es geht also darum, „Regeln aus Daten zu gewinnen oder eine gegebene Regelmenge zu verbessern“ [24, S. 12]. Im Allgemeinen nutzt man eine bestimmte Teilmenge von Daten, das sogenannte Lernset oder bei uns das **Trainingsset**, um eine Menge von Regeln zu generieren. Die erhaltenen Ergebnisse werden dann durch Anwendung auf eine andere Teilmenge, das **Testset**, geprüft, wobei das Testset in einer bestimmten Weise klassifiziert wird und diese Klassifikation auf ihre Richtigkeit geprüft wird [24, S. 12].

Die Regeln geben wir durch sogenannte **Heuristiken** vor. Dies ist eine sehr einfache Form maschinellen Lernens. Was es damit genauer auf sich hat, wird im nächsten Abschnitt erläutert.

Heuristiken - Was ist das?
Befragt man Duden online, so werden Heuristiken beschrieben als „Lehre, Wissenschaft von den Verfahren, Probleme zu lösen; methodische Anleitung, Anweisung zur Gewinnung neuer Erkenntnisse“ [4]. Und in der Tat: Wir nutzen Heuristiken in Form von methodischen Anleitungen, um Vorhersagen zu treffen. Das Wort Heuristik kommt aus dem Griechischen („heuriskein“). Es bedeutet „finden oder entdecken“ [19, S. 17]. In der Mathematik werden Heuristiken im Allgemeinen als Verfahren beschrieben, die für ein Problem eine Lösung liefern, die aber nicht notwendigerweise optimal ist [19, S. 17]. Für unsere Zwecke reichen diese Verfahren, da wir nur zeigen wollen, dass gute Vorhersagen machbar sind. Die Heuristiken wurden von uns für den CAMMP day als Funktionen in MATLAB realisiert. Sie beinhalten Regeln, wie die sexuelle Orientierung eines Nutzers bestimmt werden könnte.
Zusammenfassend wird in folgenden Schritten gearbeitet:

Vorgehen:

1. Wähle die Größe der Stichprobe N. Teile die Nutzer des Datensatzes zufällig in zwei Teile ein. Dabei wird durch „Ziehen ohne Zurücklegen“ eine Stichprobe aus N Nutzern erstellt, deren sexuelle Orientierung wir zunächst ausblenden.
2. Wende eine Heuristik auf die Stichprobe an und sage damit die sexuelle Orientierung der Nutzer dieser Stichprobe vorher.
3. Überprüfe die Vorhersagen auf ihre Richtigkeit, indem die sexuelle Orientierung wieder eingeblendet und die Vorhersage mit der Realität verglichen wird.

Im letzten Schritt können genau zwei Fälle eintreten:

- „richtig eingeschätzt“: Vorhersage entspricht der tatsächlichen sexuellen Orientierung.
- „falsch eingeschätzt“: Vorhersage entspricht nicht der tatsächlichen sexuellen Orientierung.

Diese Vorgehensweise spiegelt ein stochastisches Experiment wieder, das wir uns in Anlehnung an [3, S. 202] genauer anschauen wollen:

N-fache unabhängige Wiederholung des Zufallsexperiments:
Wir haben N Nutzer. Jeder Nutzer wird unabhängig von den andern mit Hilfe einer Heuristik eingeschätzt. Dies ist das **Zufallsexperiment**. Für jeden Nutzer erhalten wir entweder das Ergebnis „richtig eingeschätzt“ oder aber das Ergebnis „falsch eingeschätzt“, wobei eine richtige Einschätzung mit Wahrscheinlichkeit $p \in [0, 1]$ eintritt. Wir nennen sie Trefferwahrscheinlichkeit. Diese Wahrscheinlichkeit ist uns nicht bekannt. Wir schreiben nun für eine Einschätzung X und codieren eine richtige Einschätzung mit 1 und eine falsche mit 0. Da wir N-mal einschätzen, nennen wir die i-te Einschätzung X_i (mit $i \in \{1, \dots N\}$). Somit ist für jede Einschätzung die Wahrscheinlichkeit, richtig zu liegen, $P(X_i = 1) = p$.

Da wir in diesem Fall nur von zwei Ereignissen ausgehen („richtig eingeschätzt“ oder „falsch eingeschätzt“) liegt ein sogenanntes **Bernoulli-Experiment** vor [3, S. 202]; [1, S. 356-357]. Wir können für einen solchen Fall die Wahrscheinlichkeit für k „Treffer“, also richtige Einschätzungen, durch die sogenannte *Binomialverteilung* [3, S. 165]

bestimmen.

Binomialverteilung ($bin(N, p)$):
Liegt ein Bernoulli-Experiment der zuvor beschriebenen Art vor, so gilt für die Wahrscheinlichkeit P für k ($k \in \{1, \dots N\}$) richtige Einschätzungen:

$$P(\text{„genau k richtige Einschätzungen“}) = \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k},$$

wobei p die Wahrscheinlichkeit einer richtigen Einschätzung pro Wiederholung und N die Anzahl der Wiederholungen angibt.

Diese Formel komplettiert das zugrundeliegende stochastische Modell.
Wird der Stichprobenumfang „groß genug“ gewählt, so ist die Vorgehensweise wissenschaftlich zulässig aufgrund des *Gesetzes großer Zahlen*. Wir können uns der unbekannten Trefferwahrscheinlichkeit annähern, denn es gilt für unsere Zwecke in Anlehnung an [3, S. 229-230]:

Gesetz großer Zahlen (schwache Version)
Die relative Häufigkeit von richtigen Einschätzungen $\frac{1}{N} \sum_{i=1}^{N} X_i$ (nur bei richtiger Einschätzung gilt $X_i = 1$, sonst $X_i = 0$) bei N Versuchen konvergiert stochastisch gegen die uns unbekannte Trefferwahrscheinlichkeit p.

Stochastische Konvergenz bedeutet: Die Wahrscheinlichkeit des Abstandes zwischen dem arithmetischen Mittel und dem eigentlichen Wert von p konvergiert im Grenzwert gegen Unendlich gegen Null [3, S. 230]. Die Ergebnisse bezogen auf eine Stichprobe sind also für die Gesamtheit in einem gewissen Sinne „repräsentativ“.

3.4.2. Konkrete Untersuchungen

In Anlehnung an die Betrachtung von Partial und Full Shadow Profiles aus der Studie werden wir fünf Vereinfachungen unterscheiden und untersuchen:

- Wir betrachten nur die *Anzahl der homosexuellen Freunde* (1. Ordnung) eines Nutzers und arbeiten mit dem Zufall.
- Wir betrachten nur die *Anzahl der homosexuellen Freunde der Freunde* (Freunde 2. Ordnung) eines Nutzers und arbeiten mit dem Zufall.
- Wir betrachten nur den *Beziehungsstatus* eines Nutzers.

- Wir betrachten nur den *Anteil der homosexuellen Freunde* (1. Ordnung) eines Nutzers und arbeiten mit dem Zufall.
- Wir betrachten nur den *Anteil der homosexuellen Freunde der Freunde* (Freunde 2. Ordnung) eines Nutzers und arbeiten mit dem Zufall.

Die erste, zweite, vierte und fünfte Vereinfachung passt zu Partial und Full Shadow Profiles, da in beiden Fällen ein Freundschaftskreis betrachtet werden kann. Die dritte Annahme entspricht einer Untersuchung von Partial Shadow Profiles, da hier eine Angabe des Beziehungsstatus notwendig ist.

Unsere Heuristiken

Die Heuristiken begründen sich durch verschiedene Vereinfachungen in der Modellierung. Bei den ersten und letzten beiden wird unter gewissen Bedingungen eine Vorhersage der sexuellen Orientierung mit Hilfe des Zufalls getätigt. Wir schauen uns eine **Bestimmung der sexuellen Orientierung mit Hilfe des Zufalls** an:

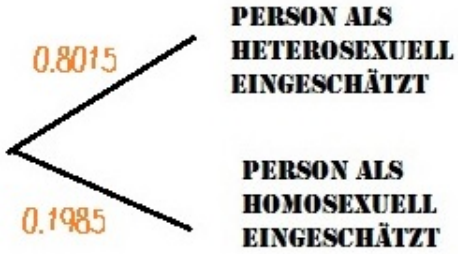


Abbildung 9: **Baumdiagramm zur Bestimmung der sexuellen Orientierung mit Hilfe des Zufalls.**

Stellen wir uns eine Urne mit Kugeln vor. Wir haben rote Kugeln, die die homosexuellen Nutzer darstellen, und blaue Kugeln, die die heterosexuellen Nutzer symbolisieren. Aus unserer Datenanalyse wissen wir, dass wir 19.85% homosexuelle Nutzer in dem Datensatz haben. Das bedeutet, dass 19.85% unserer Kugeln in der Urne rot sind. Greifen wir blind in unsere Urne, so haben wir (wenn wir von einer gleichmäßigen Verteilung der Kugeln ausgehen) eine Chance von 19.85%, eine rote Kugel zu ziehen. Die Chance, eine blaue Kugel zu finden, ist demnach $100\% - 19.85\% = 80.15\%$. Für die Einschätzungen können wir von einem einstufigen Zufallsexperiment in Anlehnung an [1, S. 338] sprechen und das Baumdiagramm in Abbildung 9 zeichnen.

Nun zu den Heuristiken: Die Zahlen, die in den Heuristiken gewählt wurden, sind nicht näher begründbar. Vielmehr sind sie ein erster Versuch der Untersuchung.

Vereinfachung 1: Wir betrachten nur die Anzahl der homosexuellen Freunde (1. Ordnung) des Nutzers und arbeiten mit dem Zufall.
Hintergrund ist die Idee, dass jemand, der viele homosexuellen Freunde hat, auch homosexuell sein könnte. Die Heuristik lautet in Worten:

Wenn der Nutzer 5 oder mehr homosexuelle Freunde hat, dann ist er homosexuell.
Wenn der Nutzer 2 oder weniger homosexuelle Freunde hat, dann ist er heterosexuell.
Sonst lasse den Zufall entscheiden.

Sie ist realisiert im Code:

```
% Name der Funktion:
function hetero = Heuristic_Regel1params(Users,i,params)

% Wenn keine anderen Parameter vorgegeben sind,
% dann nehme die Parameter 5 und 2:
if nargin==2
    params(1) = 5;
    params(2) = 2;
end

% Wenn der erste Parameter nicht größer als der zweite ist,
% so gebe eine Fehlermeldung aus:
if (params(2)>params(1) | params(2)==params(1))
        disp('Ihr habt Zahlen gewählt,
        die für die Regel keinen Sinn machen.')
end

% Lasse den Zufall entscheiden:
p = 0.1985;
hetero = (rand()>p);

% Wenn aber der Nutzer genau so viele wie
% oder mehr homosexuelle Freunde
% als der erste Parameter hat, so sage, dass er homosexuell ist:
if Users(i).NumberGayFriends > (params(1)-1)
    hetero = false;
end
```

```
% Wenn aber der Nutzer genau so viele wie
% oder weniger homosexuelle Freunde
% als der zweite Parameter hat, so sage, dass er heterosexuell
% ist:
if Users(i).NumberGayFriends < (params(2)+1)
    hetero = true;
end
end
```

Vereinfachung 2: Wir betrachten nur die Anzahl der homosexuellen Freunde der Freunde (Freunde 2. Ordnung) des Nutzers und arbeiten mit dem Zufall.

Die Idee dahinter ist, dass jemand, der viele homosexuellen Freundes Freunde hat, auch homosexuell sein könnte. Wir nutzen die Heuristik:

Wenn der Nutzer 5 oder mehr homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer 3 oder weniger homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.

Mit dem Code:

```
% Name der Funktion:
function hetero = Heuristic_Regel2params(Users,i,params)

% Wenn keine anderen Parameter vorgegeben sind,
% dann nehme die Parameter 5 und 3:
if nargin==2
    params(1) = 5;
    params(2) = 3;
end

% Wenn der erste Parameter nicht größer als der zweite ist,
% so gebe eine Fehlermeldung aus:
if (params(2)>params(1) |params(2)==params(1))
    disp('Ihr habt Zahlen gewählt,
    die für die Regel keinen Sinn machen.')
end

% Lasse den Zufall entscheiden:
p = 0.1985;
```

```
hetero = (rand()>p);

% Wenn aber der Nutzer genau so viele wie
% oder mehr homosexuelle Freunde 2. Ordnung
% als der erste Parameter hat, so sage, dass er homosexuell ist:
if Users(i).NumberGayFriendsSecondOrder > (params(1)-1)
    hetero = false;
end

% Wenn aber der Nutzer genau so viele wie
% oder weniger homosexuelle Freunde 2. Ordnung
% als der zweite Parameter hat, so sage, dass er heterosexuell ist:
if Users(i).NumberGayFriendsSecondOrder < (params(2)+1)
    hetero = true;
end
end
```

Vereinfachung 3: Wir betrachten nur den Beziehungsstatus des Nutzers. Hier kann die Idee verfolgt werden, dass verheiratete Nutzer vermutlich häufiger heterosexuell sind. Dies ist realisiert durch die Heuristik:

Wenn der Nutzer verheiratet ist, so ist er heterosexuell. Sonst ist er homosexuell.

Mit dem Code:

```
%% Name der Funktion:
function  hetero = Heuristic_Regel3(Users,i)

%% Wenn Beziehungsstatus verheiratet lautet,
%% dann sage: Nutzer ist heterosexuell
if Users(i).RelationshipStatus==5
    hetero = true;

%% Sonst sage: Nutzer ist homosexuell
else
   hetero = false;
end
end
```

Vereinfachung 4: Wir betrachten nur den Anteil der homosexuellen Freunde (Freunde 1. Ordnung) des Nutzers und arbeiten mit dem Zufall.
Die Idee dahinter ist, dass Homosexuelle einen höheren Anteil an homosexuellen Freunden haben könnten. Wir nutzen die Heuristik:

Wenn der Nutzer mehr als 20% *homosexuelle Freunde hat, dann ist er homosexuell.* *Wenn der Nutzer weniger als* 10% *homosexuelle Freunde hat, dann ist er heterosexuell.* *Sonst lasse den Zufall entscheiden.*

Mit dem Code:

```
% Name der Funktion:
function hetero = Heuristic_Regel4params(Users,i,params)

% Wenn keine anderen Parameter vorgegeben sind,
% dann nehme die Parameter 0.2 und 0.1:
if nargin==2
    params(1) = 0.2;
    params(2) = 0.1;
end

% Wenn der zweite Parameter größer als der erste ist
% oder ein Parameter nicht zwischen 0 und 1 liegt,
% so gebe eine Fehlermeldung aus:
if (params(2)>params(1))
    disp('Ihr habt Zahlen gewählt,
    die für die Regel keinen Sinn machen.')
end
if (params(1)< 0 | params(2)< 0 | params(1) > 1 | params(2) > 1)
    disp('Ihr habt Zahlen gewählt,
    die für die Regel keinen Sinn machen.')
end

% Lasse den Zufall entscheiden:
p = 0.1985;
hetero = (rand()>p);

% Wenn aber der Nutzer mehr homosexuelle Freunde
% als der erste Parameter hat, so sage, dass er homosexuell ist:
N = Users(i).NumberFriends;
if Users(i).NumberGayFriends/N > params(1)
hetero = false;
```

```
end

% Wenn aber der Nutzer weniger homosexuelle Freunde
% als der zweite Parameter hat, so sage, dass er heterosexuell
% ist:
if Users(i).NumberGayFriends/N < params(2)
hetero = true;
end
end
```

Nun kommen wir zu unserer letzten Vereinfachung:

Vereinfachung 5: Wir betrachten nur den Anteil der homosexuellen Freunde der Freunde (Freunde 2. Ordnung) des Nutzers und arbeiten mit dem Zufall.

Die Idee dahinter ist, dass Homosexuelle einen höheren Anteil an homosexuellen Freunden 2. Ordnung haben könnten. Dies ist verwirklicht in der Heuristik:

Wenn der Nutzer mehr als 30% *homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer weniger als* 10% *homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

Mit dem Code:

```
% Name der Funktion:
function hetero = Heuristic_Regel5params(Users,i,params)

% Wenn keine anderen Parameter vorgegeben sind,
% dann nehme die Parameter 0.3 und 0.1:
if nargin==2
    params(1) = 0.3;
    params(2) = 0.1;
end

% Wenn der zweite Parameter größer als der erste ist
% oder ein Parameter nicht zwischen 0 und 1 liegt,
% so gebe eine Fehlermeldung aus:
if (params(2)>params(1))
    disp('Ihr habt Zahlen gewählt,
    die für die Regel keinen Sinn machen.')
end
if (params(1)< 0 | params(2)< 0 | params(1) > 1 | params(2) > 1)
```

```
    disp('Ihr habt Zahlen gewählt,
    die für die Regel keinen Sinn machen.')
end

% Lasse den Zufall entscheiden:
p = 0.1985;
hetero = (rand()>p);

% Wenn aber der Nutzer mehr homosexuelle Freunde 2. Ordnung
% als der erste Parameter hat, so sage, dass er homosexuell ist:
N = Users(i).NumberFriendsSecondOrder;
if (Users(i).NumberGayFriendsSecondOrder/N > params(1))
hetero = false;
end

% Wenn aber der Nutzer weniger homosexuelle Freunde 2. Ordnung
% als der zweite Parameter hat, so sage, dass er heterosexuell
% ist:
if (Users(i).NumberGayFriendsSecondOrder/N < params(2) )
hetero = true;
end
end
```

Soweit zu unseren Heuristiken. Aber wann ist eine Heuristik „gut“?

Gütekriterien

Wir haben für jede Vorhersage die beiden Szenarien:

- Klassifikation ist falsch: vorhergesagte sexuelle Orientierung ist nicht gleich der tatsächlichen,
- Klassifikation ist richtig: vorhergesagte sexuelle Orientierung ist gleich der tatsächlichen.

Ziel 1:

Unsere Arbeit wäre unnütz, wenn die Vorhersagen nicht besser wären als der Zufall. Aber wie gut ist eine Klassifikation durch den Zufall? Wenn wir noch einmal einen Blick in unsere Daten werfen, so stellen wir fest, dass wir 0.1985 homosexuelle Nutzer haben. Die Vorhersage und ihre Überprüfung entspricht einem zweistufigem Zufallsexperiment [1, S. 338] wie im Bild 10.
Eine richtige Klassifikation entspricht in Abbildung 10 also den Pfaden:

- Person ist heterosexuell → Person als heterosexuell eingeschätzt und
- Person ist homosexuell → Person als homosexuell eingeschätzt.

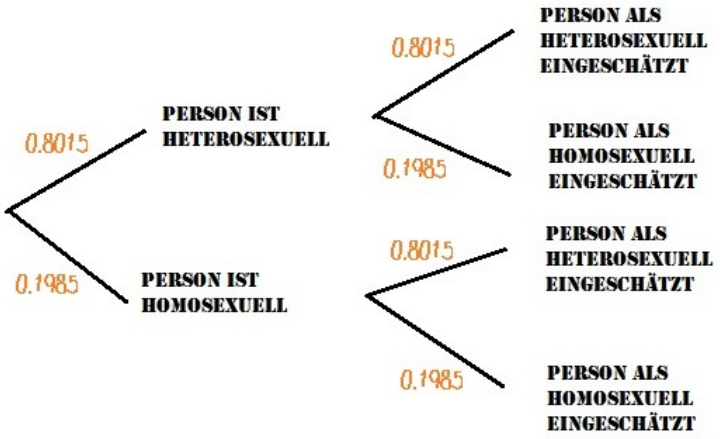


Abbildung 10: **Klassifikation durch Zufall mit Überprüfung im Baumdiagramm.**

Mit Hilfe der Pfadregel (Multipliziere die Wahrscheinlichkeiten entlang des Pfades [1, S. 338].) ergeben sich die Wahrscheinlichkeiten

- Person ist heterosexuell → Person als heterosexuell eingeschätzt: $0.8015 \cdot 0.8015 \approx 0.6424$
- Person ist homosexuell → Person als homosexuell eingeschätzt: $0.1985 \cdot 0.1985 \approx 0.0394$

Da beide Pfade eine richtige Vorhersage liefern, erhalten wir die Wahrscheinlichkeit für eine richtige Vorhersage durch Summation (Summenregel [1, S. 338]) der Wahrscheinlichkeiten zu

$$0.6424 + 0.0394 = 0.6818.$$

Diese Quote gilt es also zu schlagen! Um zu messen, ob das gelingt, wird eine Größe namens *Cohen's Kappa*[4] verwendet. Hier werden die Kategorisierungen aller sexuellen Orientierungen gemeinsam betrachtet:

[4]Jacob Cohen war Psychologe an der New York University [31].

Definition *Cohen's Kappa* [29, S. 85]:
Cohen's Kappa ist definiert als

$$\kappa = \frac{p_H - p_Z}{1 - p_Z},$$

wobei p_H die relative Häufigkeit der richtigen Kategorisierungen und p_Z die Häufigkeit an Übereinstimmungen wiedergibt, die allein durch den Zufall auftreten.

Cohen's Kappa ist ein Übereinstimmungskoeffizient [29, S. 85]. Beziehen wir die vorherigen Überlegungen ein, so lautet Kappa

$$\kappa = \frac{p_H - 0.6818}{1 - 0.6818} = \frac{p_H - 0.6818}{0.3182}.$$

Wenn wir die Größe etwas genauer unter die Lupe nehmen, so stellen wir fest, dass

- κ nahe Null ist für Vorhersagen, deren relative Häufigkeit der richtigen Kategorisierungen nahe der Häufigkeit an Übereinstimmungen durch den Zufall sind. Hier ist $p_H \approx p_Z$; die Heuristik ist so gut wie der Zufall.
- κ nahe 1 ist für Vorhersagen, deren relative Häufigkeit der richtigen Kategorisierungen nahe 1 sind. Hier ist $p_H \approx 1$; die Heuristik bestimmt nahezu alle sexuellen Orientierungen richtig.
- κ negativ ist für Vorhersagen, deren relative Häufigkeit der richtigen Kategorisierungen kleiner als die Häufigkeit an Übereinstimmungen durch den Zufall sind. Hier ist $p_H < p_Z$; die Heuristik ist schlechter als der Zufall.
- κ positiv ist für Vorhersagen, deren relative Häufigkeit der richtigen Kategorisierungen größer als die Häufigkeit an Übereinstimmungen durch den Zufall sind. Hier ist $p_H > p_Z$; die Heuristik ist besser als der Zufall.

Mit diesen Beobachtungen kann κ die Güte der Vorhersagen messen. Dies ist unabhängig von der Anzahl der getesteten Nutzer, da hier immer relative Anteile betrachtet werden [9, S. 5].

Ziel 2:
Ein weiteres Ziel ist natürlich, möglichst viele richtige Klassifikationen mit unseren Heuristiken zu erreichen. Um das zu messen, können zwei mathematische Größen betrachtet werden. Es sind die Trefferquote und die Genauigkeit. Zur Darlegung dieser beiden Größen wird Bezug auf [30, S. 90-91] genommen.

Das ersten Gütekriterium für die Vorhersagen ist die *Trefferquote*. Sie wird für jede Gruppe einer sexuellen Orientierung bestimmt und als *Richtig-Positiv-Rate* bezeichnet. Dabei betrachten wir den Anteil der richtig Klassifizierten in Bezug auf die relevanten Personen. Wir definieren die Trefferquote zu unseren Zwecken:

Definition *Trefferquote*:
Die Trefferquote für eine Gruppe ist gleich dem Anteil der richtig Klassifizierten einer Gruppe. Sei S die Menge der Nutzer einer Gruppe einer sexuellen Orientierung, S_k die Menge der Nutzer, die durch die Klassifikation der Gruppe zugeordnet wurden. Die richtig Klassifizierten bilden dann die Schnittmenge $S_k \cap S$. Dann ist die Trefferquote T_S der Gruppe

$$T_S = \frac{|S_k \cap S|}{|S|},$$

wobei $|\cdot|$ für die Mächtigkeit der Mengen steht.

Das zweite Gütekriterium für die Vorhersagen ist die *Genauigkeit*. Auch sie wird für jede Gruppe einer sexuellen Orientierung bestimmt. Hierbei sprechen wir auch von dem *positiven Vorhersagewert* bzw. dem *positiv prädikativen Wert* und bestimmen den Anteil der richtig Klassifizierten im Bezug auf alle so Klassifizierten. Wir formulieren für uns:

Definition *Genauigkeit*:
Die Genauigkeit für eine Gruppe ist gleich dem Anteil der richtig Klassifizierten einer Gruppe unter allen so klassifizierten. Wir nutzen die eingeführten Bezeichnungen aus der vorherigen Definition. Dann ist die Genauigkeit G_S der Gruppe

$$G_S = \frac{|S_k \cap S|}{|S_k|},$$

wobei $|\cdot|$ für die Mächtigkeit der Mengen steht.

Zur Erläuterung betrachten wir ein Beispiel, das in Abbildung 11 illustriert wird: Wir betrachten die heterosexuellen Nutzer in unserem Datensatz.

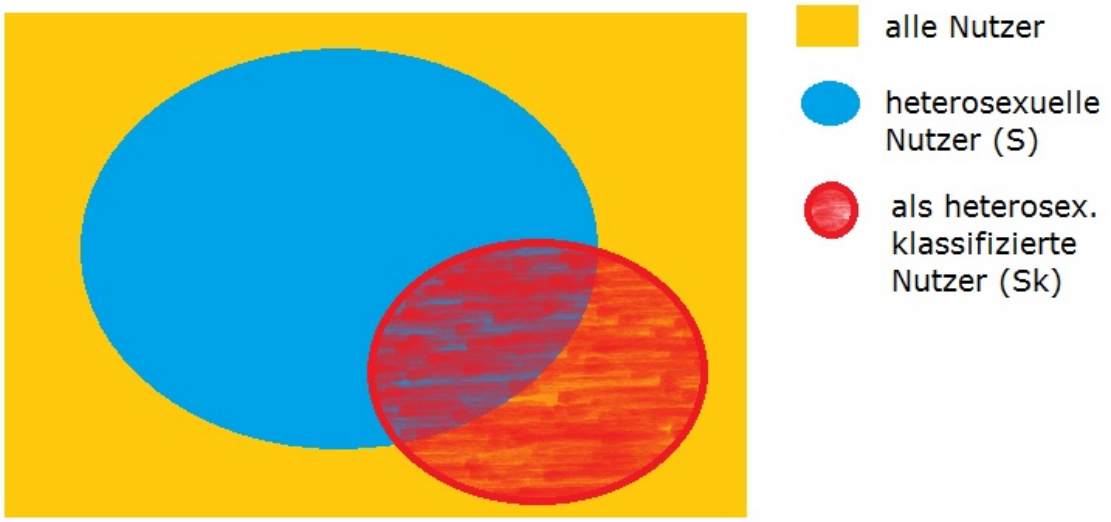


Abbildung 11: **Venn-Diagramm zur Trefferquote und Genauigkeit der Heterosexuellen.**

In der Abbildung 11 sehen wir in orange alle Nutzer des Datensatzes. Die Menge der heterosexuellen Nutzer ist in blau gezeichnet. Sie entsprechen der Menge S in den Definitionen. Wir bestimmen durch unsere Heuristik die rote Menge an Nutzern als heterosexuell (Menge S_k), wobei davon längst nicht alle tatsächlich heterosexuell sind. Die Nutzer, die heterosexuell sind und als solche klassifiziert wurden, befinden sich in der Schnittmenge von blau und rot. Dies ist die Menge $S \cap S_k$.
Die Menge S beinhaltet, gemäß unserer Angaben über den Datensatz, $56284 - 11174 = 45110$ Nutzer. Umfasst nun beispielsweise S_k 2000 Nutzer und $S \cap S_k$ 1000 Nutzer, so erhalten wir folgende Trefferquote T_S und Genauigkeit G_S:

$$T_S = \frac{|S_k \cap S|}{|S|} = \frac{1000}{45110} \approx 0.0222$$

und

$$G_S = \frac{|S_k \cap S|}{|S_k|} = \frac{1000}{2000} = 0.5.$$

Doch was sagen uns diese Zahlen? Beide Größen sind relative Anteile, sie liegen also immer zwischen 0 und 1. Bei beiden gilt:

- Sind die Größen nahe 0, so befinden sich in der Menge $S_k \cap S$ wenige Nutzer. $|S_k \cap S|$ ist nahe 0. Die Klassifikation hat damit kaum Nutzer richtig klassifiziert.
- Sind die Größen nahe 1, so befinden sich in der Menge $S_k \cap S$

– bei der Trefferquote nahezu so viele Nutzer, wie in der Menge S.

– bei der Genauigkeit nahezu so viele Nutzer, wie in der Menge S_k.

Die Klassifikation hat damit nahezu alle Nutzer (in zwei Sichtweisen) richtig klassifiziert.

Im Beispiel spricht ein Wert von 0.0222 für eine niedrige Trefferquote der Heterosexuellen. Die Genauigkeit mit einem Wert von 0.5 besagt, dass 50% der Klassifikationen auch der Realität entsprechen. Diese beiden Größen helfen, Heuristiken miteinander zu vergleichen.

Bemerkung am Rande:
Die beiden Größen können auch als bedingte Wahrscheinlichkeiten eingeführt werden. Da wir in unserem Fall jedoch immer die gleiche Grundgesamtheit (Nutzer des Datensatzes) betrachten und eine Laplace-Verteilung (Gleichverteilung) vorliegt, sind die Definitionen einander entsprechend. Zur Erläuterung wird demnach die anschaulichere Version genutzt.

3.5. Verbesserung von Kappa

Vier der zuvor vorgestellten Heuristiken sind von Parametern abhängig. Unter Parametern verstehen wir die Zahlen in den Heuristiken, die vorgegeben werden. In den Regeln sind es die blau markierten Zahlen. Zur Erinnerung:

Vereinfachung 1: Wir betrachten nur die homosexuellen Freunde (1. Ordnung) eines Nutzers und nutzen den Zufall.

Wenn der Nutzer 5 oder mehr homosexuelle Freunde hat, dann ist er homosexuell. Wenn der Nutzer 2 oder weniger homosexuelle Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.

Vereinfachung 2: Wir betrachten nur die homosexuellen Freunde der Freunde (Freunde 2. Ordnung) eines Nutzers und nutzen den Zufall.

Wenn der Nutzer 5 oder mehr homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer 3 oder weniger homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.

Vereinfachung 4: Wir betrachten nur den Anteil der homosexuellen Freunde (Freunde 1. Ordnung) des Nutzers und nutzen den Zufall.

Wenn der Nutzer mehr als 20% homosexuelle Freunde hat, dann ist er homosexuell. Wenn der Nutzer weniger als 10% homosexuelle Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.

Vereinfachung 5: Wir betrachten nur den Anteil der homosexuellen Freunde der Freunde (Freunde 2. Ordnung) des Nutzers und nutzen den Zufall.

Wenn der Nutzer mehr als 30% homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer weniger als 10% homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.

Die Wahl dieser blauen Zahlen hat Einfluss auf Kappa. Wir wollen also die Parameter finden, die uns einen möglichst hohen Kappawert liefern. Wir wollen mit dem Wert Kappa möglichst nahe an die Zahl 1 gelangen, da hierfür die Vorhersagen am besten sind. Natürlich kann man dazu zunächst einige Parameter ausprobieren, dieses Vorgehen ist jedoch sehr mühsam. Hier stellt sich die Frage, ob es bessere Verfahren gibt, an geeignete Parameter zu kommen. Im CAMMP day wird ein Optimierungsverfahren behandelt werden können. Es ist angelehnt an den Monte-Carlo-Algorithmus.

3.5.1. Optimierung in Anlehnung an die Monte-Carlo-Methode

Bei einer Monte-Carlo-Methode handelt es sich um einen sogenannten randomisierten Algorithmus [25, S. 15]. Grundsätzlich werden hier Zufallszahlen genutzt, die auf eine bestimmte Art und Weise erzeugt werden [25, S. 1]. Wir suchen in unserem Fall nach zwei Parametern in den folgenden Schritten:

Optimierung in Anlehnung an die Monte-Carlo-Methode

1. Gebe für jeden Parameter ein (abgeschlossenes) Intervall für die Suche an.
2. Erzeuge eine bestimmte Anzahl von Zahlen mit gleichem Abstand aus diesen Intervallen.
3. Gebe für jede Kombination dieser Zahlen, bei der der erste Parameter größer ist als der zweite, den Wert Kappa aus.
4. Vergleiche die Werte für Kappa und wähle die Parameter mit größtem Kappa.
5. Gebe das größte Kappa (kOpt) und die zugehörigen Parameter (zOpt) aus.

Für das Erzeugen des Wertes Kappa wird zudem zu Beginn die Größe N der Stichprobe festgelegt. Außerdem ist am Anfang eine Eingabe für mögliche Parameter von Nöten. Diese hat jedoch keinen Einfluss auf die Optimierung an sich, die Zahlen werden überschrieben. In den ersten beiden Heuristiken werden natürliche Zahlen zur Regelbildung verwendet, in den letzten beiden jedoch rationale. Die Optimierung wird, um diesen unterschiedlichen Zahlbereichen gerecht zu werden, für diese beiden Arten an Heuristiken etwas angepasst:

Optimierung für die Heuristiken 1 und 2

Die im zweiten Schritt genannte Anzahl von Zahlen ist bei Heuristiken (1 und 2) mit absoluten Parametern die Anzahl der natürlichen Zahlen aus den Intervallen. Bei der Optimierung werden also alle natürlichen Zahlen in den Intervallen erzeugt. In diesem Fall haben wir keine zufällige Auswahl. Dieses geänderte Vorgehen ergibt für unsere Zwecke jedoch Sinn, denn wir haben zuvor erfahren, dass 99% unserer Nutzer zwischen 0 und 15 homosexuelle Freunde bzw. über 95% der Nutzer zwischen 0 und 35 homosexuelle Freundes Freunde haben. Bei der Optimierung ist es also sinnig, genau in diesen Zahlbereichen nach optimalen, natürlichen Parametern zu suchen, sodass sich die Anzahl an Kombinationen im (zeitlichen) Rahmen hält und daher alle getestet werden können. Wir schauen uns hierzu zwei Szenarien an: Bei gleicher Wahl der Intervalle würden wir alle Kombinationen im Intervall von 0 bis zu einer natürlichen Zahl n testen, so sind dies insgesamt $\sum_{k=0}^{n} k = \sum_{k=1}^{n} k = \frac{n^2+n}{2}$ Möglichkeiten gemäß der Gaußschen Summenformel [16, S. 3]. Im Fall $n = 35$ entspricht dies 630 Kombinationen. Sind die Intervalle jedoch disjunkt und beinhalten n bzw. m natürliche

Zahlen mit $n \geq m$, so sind im aufwändigsten Fall (mit $m = n$) n^2 Kombinationen zu testen. Mit $n = 35$ wären das 1225 Kombinationen. Die Optimierung für Heuristik 1 ist implementiert durch die folgenden Befehle, die zur Erklärung kommentiert sind. Der Code für Heuristik 2 ist analog.

```
%%% Von den Schülerinnen und Schülern sind einzustellen:

%% Größe der Stichprobe:
N = 1000;

%% Intervalle vorgeben für beide Parameter;
%% vor dem Semikolon Intervall für ersten Parameter,
%% hinter dem Semikolon für den zweiten Parameter:
range = [0,5;0,5];

%%% Durchführung der Optimierung:

%% Name der Funktion:
function [kOpt,zOpt] = Optimization3_Regel1(Users,N,range)

%% Generiere Stichprobe:
testSet = randi(length(Users),]N,1);

%% Generiere Zahlen mit gleichem Abstand aus den Intervallen:
n = range(1,2)-range(1,1)+1;
range1 = linspace(range(1,1),range(1,2),n);
m = range(2,2)-range(2,1)+1;
range2 = linspace(range(2,1),range(2,2),m);

%% Zur Speicherung der Werte:
hist =[];

%% Erzeuge für alle Kombinationen den Wert Kappa,
%% die die Bedingung aus Schritt 3 erfüllen:
for a = 1:length(range1)
    for b = 1:length(range2)
        if (range2(b)<range1(a))
            params = [range1(a),range2(b)]
            Kappa = heuristicTester(@Heuristic_Regel1params,...
                Users,N,testSet, params,false)
            fitnessTable(a,b) = Kappa;
```

```
                hist(a,b,1) = range1(a);
                hist(a,b,2) = range2(b);
                end
    end
end

%% Bestimme maximales Kappa und die zugehörigen Parameter:
[idxi,idxj] = find(max(max(max(max(fitnessTable))))==fitnessTable);
kOpt = [];
zOpt = [];
for i=1:length(idxi)
   kOpt = [kOpt,  fitnessTable(idxi(i),idxj(i))];
   zOpt = [zOpt;range1(idxi(i)), range2(idxj(i))];
end

%% Gebe die Werte kommentiert aus:
for i=1:length(kOpt)
    disp('-----------------------------------------')
str = sprintf('Der optimale Kappawert ist %6.2f\n',kOpt(i));
disp(str)
str = sprintf('mit den zugehörigen Zahlen
[ %i, %i]\n',zOpt(i,1),zOpt(i,2));
disp(str)
 disp('-----------------------------------------')
end
end
```

Optimierung für die Heuristiken 4 und 5

Bei den Heuristiken mit relativen Parametern werden 20 Pseudo-Zufallszahlen, die von der deterministischen Funktion „rand“ erzeugt werden [25, S. 2], generiert. Diese Anzahl ist empirisch motiviert. Die Optimierung soll pro Durchlauf nicht zu lange dauern. Das Wort „Pseudo“ motiviert sich dadurch, dass es mit dem Computer nicht möglich ist, echte Zufallszahlen zu generieren, aber dennoch Zahlenfolgen angegeben werden können, die hinreichend gleich verteilt sind [26, S. 139]. Nun wird die Implementierung der Optimierung für Heuristik 4 vorgestellt. Die für Heuristik 5 ist analog.

```
%%% Von den Schülerinnen und Schülern sind einzustellen:

%% Größe der Stichprobe:
```

```
N = 1000;

%% Intervalle vorgeben für beide Parameter;
%% vor dem Semikolon Intervall für ersten Parameter,
%% hinter dem Semikolon für den zweiten Parameter:
range = [0,1;0,1];

%%% Durchführung der Optimierung:

%% Name der Funktion:
function [kOpt,zOpt] = Optimization3_Regel4(Users,N,range)

%% Generiere Stichprobe:
testSet = randi(length(Users),N,1);

%% Generiere 20 Zahlen mit gleichem Abstand aus den Intervallen:
n = 20;
range1 = linspace(range(1,1),range(1,2),n);
range2 = linspace(range(2,1),range(2,2),n);

%% Zur Speicherung der Werte:
hist =[];

%% Erzeuge für alle Kombinationen den Wert Kappa,
%% die die Bedingung aus Schritt 3 erfüllen:
for a = 1:length(range1)
    for b = 1:length(range2)
        if (range2(b)<range1(a))
            params = [range1(a),range2(b)]
            Kappa = heuristicTester(@Heuristic_Regel4params,...
                Users,N,testSet, params,false)
            fitnessTable(a,b) = Kappa;
            hist(a,b,1) = range1(a);
            hist(a,b,2) = range2(b);

        end
    end
end

%% Bestimme maximales Kappa und die zugehörigen Parameter:
[idxi,idxj] = find(max(max(max(max(fitnessTable))))==fitnessTable);
kOpt = [];
```

```
zOpt = [];
for i=1:length(idxi)
    kOpt = [kOpt,  fitnessTable(idxi(i),idxj(i))];
    zOpt = [zOpt;range1(idxi(i)), range2(idxj(i))];
end

%% Gebe die Ergebnisse kommentiert aus:
for i=1:length(kOpt)
     disp('------------------------------------------')
str = sprintf('Der optimale Kappawert ist %6.2f\n',kOpt(i));
disp(str)
str = sprintf('mit den zugehörigen Zahlen
[ %6.2f, %6.2f]\n',zOpt(i,1),zOpt(i,2));
disp(str)
 disp('------------------------------------------')
end
end
```

Nimmt die Größe der Stichprobe zu, dauert die Optimierung natürlich länger, da bei der Generierung von Kappa mehr Nutzer betrachtet werden. Eine solche Optimierung ergibt zudem nur bei wenigen Parametern Sinn, denn das Durchlaufen aller möglichen Kombinationen braucht viel Zeit. Damit das Verfahren gewinnbringend ist, können die Intervalle für die relativen Parameter immer weiter eingeschränkt werden [2, S. 160]. Hier wird die Idee verfolgt, dass in der Nähe von guten Parametern noch bessere liegen könnten. Bei der vorgestellten Optimierung für relative Parameter ist der Fall, dass der erste und der zweite Parameter gleich sein können, nicht beachtet worden. Die Einarbeitung dieses Falls kann eine Verbesserungsmöglichkeit sein.

3.5.2. Weitere angedachte Optimierungsverfahren

Im Laufe der Forschungsarbeit wurden weitere Optimierungsverfahren getestet. In diesem Abschnitt sollen sie der Vollständigkeit halber in ihrer Idee in Anlehnung an [2, S. 160-175] dargelegt und erläutert werden, warum ihre Umsetzung im CAMMP day nicht stattgefunden hat.

Optimierung mit Gradientenverfahren

Die zweite Optimierungsmöglichkeit ist eine Optimierung im Sinne des Gradientenverfahrens. Wir suchen in unserem Fall den größtmöglichen Kappawert. Hierbei stellen wir uns die Werte für Kappa als Funktion der Parameter vor. In der Literatur wird

das Gradientenverfahren häufig zur Suche nach Minima beschrieben. Hier spricht man von der „Methode des steilsten Abstiegs“ [2, S. 165]. Wir suchen jedoch ein Maximum, sodass hier die Methode angepasst wird zu einer *Methode des steilsten Anstiegs*. Sie ist in Abbildung 12 illustriert. Wir möchten, wie zuvor, das maximale Kappa bestimmen, also suchen wir mit Hilfe des Gradienten (der Steigung) dieses Kappa durch folgende Schritte:

Gradientenverfahren zur Suche nach einem Maximum

1. Gebe Parameter p_0 und Schrittlänge β vor.
2. Bestimme Kappa(p_0).
3. Bestimme neue Parameter p_1 durch $p_1 = p_0 + 0.1$.
4. Bestimme Kappa(p_1).
5. Bestimme die Steigung, den Gradienten durch
$$\text{grad}_{0,1} = \frac{\text{Kappa}(p_1) - \text{Kappa}(p_0)}{p_1 - p_0}$$
6. Bestimme neue Parameter p_2 durch $p_2 = p_0 + \beta \cdot \text{grad}_{0,1}$.
7. Bestimme Kappa(p_2).
8. Bestimme neue Parameter p_3 durch $p_3 = p_2 + 0.1$.
9. Bestimme Kappa(p_3). Weiter mit 5.

Das Verfahren ergibt Sinn, denn bei der Berechnung des Gradienten können drei Situationen auftreten:

- Gradient ist Null: Das bedeutet, dass das neue Kappa nicht besser ist als das alte. Wir behalten die Parameter.
- Gradient ist positiv: Das neue Kappa ist besser als das alte. Deshalb bewegen wir uns in die Richtung der neuen Parameter mit der Berechnung $p_2 = p_0 + \beta \cdot \text{grad}_{0,1}$.
- Gradient ist negativ: Das neue Kappa ist nicht besser als das alte. Wir bewegen uns in die entgegengesetzte Richtung (vgl. Abb. 12).

Ziel ist hierbei immer eine Bewegung in Richtung des Maximums.
Es ist leicht ersichtlich, dass der Erfolg dieses Verfahrens von vielen Faktoren abhängig ist. Sind die Anfangsparameter oder die Schrittlänge schlecht gewählt, so kann es vorkommen, dass das Verfahren nicht konvergiert.

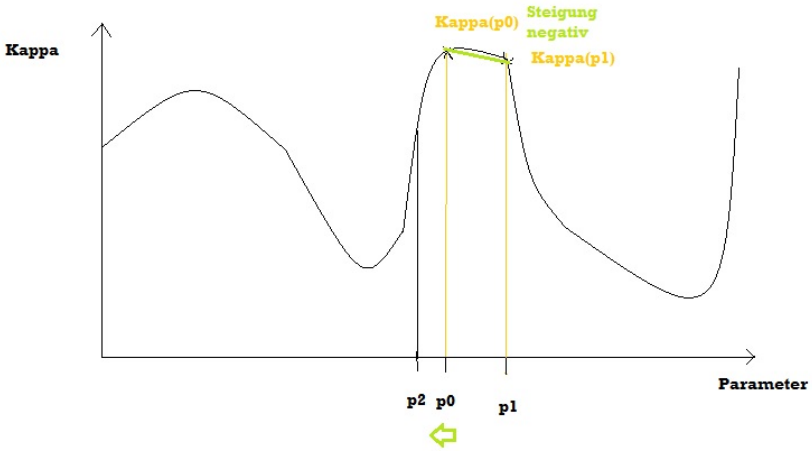


Abbildung 12: **Gradientenverfahren vereinfacht dargestellt.**

Optimierung mit Gradientenverfahren mit Liniensuche
Die dritte Möglichkeit, die Heuristiken zu optimieren, ist das Gradientenverfahren mit Liniensuche. Es erfolgt im Wesentlichen wie das zuvor vorgestellte Verfahren, versucht aber die Willkürlichkeit in der Wahl der Schrittlänge zu beheben.
Die Suche erfolgt in unserem Fall im Allgemeinen entlang der Linie

$$p(\beta) = p_0 + \beta \cdot \text{grad}\,.$$

Hier ist, wie zuvor, β die Schrittlänge. Sie kann nun zum Beispiel mit numerischen Methoden so bestimmt werden, dass eine Verbesserung von Kappa für $p(\beta)$ eintritt. Für unsere Zwecke probieren wir verschiedene Werte für eine Schrittlänge aus und vergleichen sie. Die Liniensuche erfolgt bei uns entlang der Geraden $y = 0.5 \cdot x$. Diese ist aus empirischen Überlegungen motiviert. Durch das Testen der verschiedenen Schrittlängen benötigt das Verfahren mehr Zeit als das ursprüngliche Gradientenverfahren, ist jedoch auch erfolgversprechender. Da auch hier Anfangsparameter vorgegeben werden müssen, ist diese mögliche Fehlerquelle immer noch vorhanden.

Zur Nichtumsetzung dieser beiden Verfahren
Die beiden Optimierungen eignen sich für die Heuristiken mit absoluten Parametern

(Heuristik 1 und 2) in dieser Form nicht. Bei der Generierung neuer Parameter müssten immer natürliche Zahlen als Ergebnisse herauskommen. Bei den Heuristiken mit relativen Parametern ergibt es nur Sinn, im Intervall $[0, 1]$ nach optimalen Parametern zu suchen. Das kann durch die Optimierungen aber in dieser Form nicht gewährleistet werden. Ein weiterer ausschlaggebender Punkt, die Optimierungen nicht in den CAMMP day einzubauen, ist, dass nicht klar ist, ob die Funktion der Kappawerte stetig oder sogar differenzierbar ist. Anschaulich wird daher deutlich: Kommen Sprünge in der Funktion vor, so ist eine systematische Bewegung durch das Gradientenverfahren zum Maximum schwer möglich. Um einen Eindruck von der Optimierung zu erhalten, sind die erzeugten Kappawerte der Optimierung der Heuristik 4 in der Abbildung 13 dargestellt. Das Gradientenverfahren wurde hier angewandt, wobei die Kappawerte durch Linien verbunden worden sind.

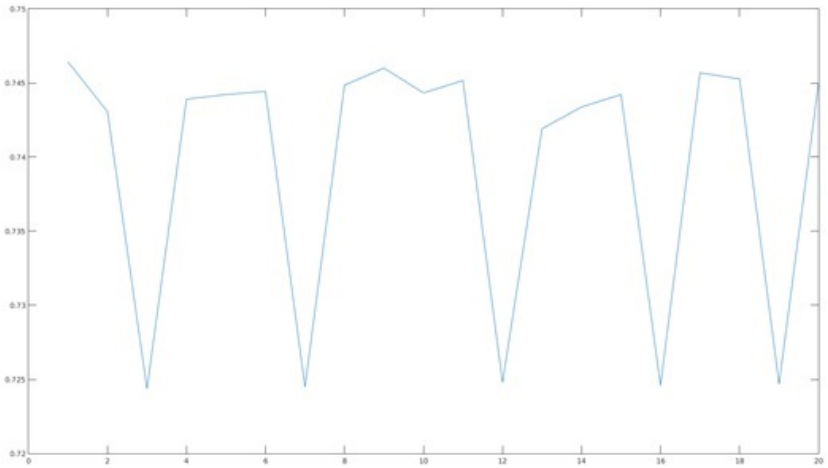


Abbildung 13: **Mit dem Gradientenverfahren erzeugte Kappawerte für Heuristik 4.** Stichprobengröße: 30000, x-Achse: Anzahl der Iterationen; y-Achse: Kappawert; erstellt von T. Camminady.

Die genannten Aspekte führten zwar zur Nichtumsetzung in dieser ersten CAMMP day Version, andere passendere Optimierungsmethoden können jedoch in der Weiterentwicklung thematisiert werden.

3.6. Gültigkeit der Ergebnisse: Hypothesentests

Die Heuristiken wurden nun getestet und zum Teil auch optimiert. Wir haben als Ergebnisse für jede Heuristik die Übereinstimmungen der Kategorisierungen. Doch

diese Ergebnisse wurden (meist[5]) anhand von Stichproben erzeugt. Wir wollen aber verlässliche Ergebnisse für alle Nutzer erhalten. Deshalb führen wir im letzten Schritt Hypothesentests durch. Die folgenden Erläuterungen basieren auf den Ausführungen von [3, S. 266-273, 294-298].

3.6.1. Idee dahinter

Wie der Name schon sagt, geht es bei Hypothesentests darum, eine Hypothese zu prüfen. Die Idee ist, dass die Gültigkeit einer **Hypothese**, also einer Aussage, für eine Grundgesamtheit[6] behauptet wird und dann mit Hilfe einer Stichprobe überprüft wird. Ziel ist, eine Entscheidung zu treffen, ob die Hypothese stimmt oder nicht. Ein solches Vorgehen ergibt Sinn, wenn wir die Richtigkeit der Aussage für die Grundgesamtheit nicht nachprüfen können oder wollen. Wir erinnern uns, dass wir die Trefferwahrscheinlichkeit einer Heuristik nicht kennen (vgl. Abschnitt 3.4.1). Deshalb formulieren wir eine Behauptung:

„Die Wahrscheinlichkeit einer richtigen Vorhersage der Heuristik ist besser als die Wahrscheinlichkeit einer richtigen Vorhersage durch den Zufall."

Diese nachzuweisende Hypothese wird als **Alternativhypothese** oder Alternative (in Zeichen: H_1) bezeichnet. Die gegenteilige Aussage wird ebenfalls formuliert und als **Nullhypothese** (in Zeichen: H_0) bezeichnet. Wir erhalten also das Entscheidungsproblem $H_0 \leftrightarrow H_1$. Genauer gesagt:

Wir müssen uns als **Entscheidungsproblem** zwischen der *Alternativhypothese* (H_1) und der gegenteiligen Aussage, der *Nullhypothese*(H_0), entscheiden:

H_0 : „Die Wahrscheinlichkeit einer richtigen Vorhersage der Heuristik ist *gleich der oder kleiner als* die Wahrscheinlichkeit einer richtigen Vorhersage durch den Zufall."

$\leftrightarrow$

H_1 : „Die Wahrscheinlichkeit einer richtigen Vorhersage der Heuristik ist *größer* als die Wahrscheinlichkeit einer richtigen Vorhersage durch den Zufall."

Wichtig ist bei der Formulierung eines Problems, dass sich die Hypothesen gegenseitig

[5] Grundsätzlich ist es auch möglich, alle Nutzer zu testen.
[6] Auch Population genannt.

ausschließen. Zu unseren Zwecken betrachten wir eine Situation, in der es darum geht, nachzuweisen, dass ein reeller Parameter größer ist als ein fester, reeller Wert.

Die Wahrscheinlichkeit einer richtigen Vorhersage durch den Zufall ist bereits berechnet (vgl. Abschnitt 3.4.2). Wir nennen sie p_Z. Die Wahrscheinlichkeit einer richtigen Vorhersage der Heuristik ist uns unbekannt (vgl. Abschnitt 3.4.1). Wir nennen sie p_H. Wir haben somit ein Problem der Form:

$$H_0 : p_H \leq p_Z \leftrightarrow H_1 : p_H > p_Z.$$

Damit haben wir eine disjunkte Teilung der reellen Zahlen für unsere beiden Hypothesen. Trifft H_0 zu, so ist der Parameter p_H kleiner oder gleich dem festen, reellen Wert p_Z; trifft H_1 zu, so ist der Parameter p_H größer als der feste, reelle Wert p_Z. Wir wollen daher die genannten Intervalle ebenfalls mit H_0 bzw. H_1 bezeichnen:

Bemerkung: Sei $p_Z \in \mathbb{R}$ eine feste reelle Zahl. Dann teilen die beiden Hypothesen $\mathbb{R}$ in zwei disjunkte Intervalle, nämlich

$$H_0 = \{p_H \in \mathbb{R}; \quad p_H \leq p_Z\} \quad \text{und} \quad H_1 = \{p_H \in \mathbb{R}; \quad p_H > p_Z\}.$$

Nachdem nun das Entscheidungsproblem benannt ist, gilt es, eine Entscheidung zu treffen. Dazu gibt es eine Entscheidungsregel. Diese hängt von der Art des statistischen Tests ab. Wir haben hier einen sogenannten *approximativen Binomialtest*, der sich durch das zugrundeliegende Modell (vgl. 3.4.1) begründen lässt. Anhand der betrachteten Stichprobe wird nach bestimmten Regeln ein Wert berechnet, der **Teststatistik** (in Zeichen: V) genannt wird. Dieser Wert wird dann mit einem weiteren Wert, einer festgelegten **kritischen Schranke** (in Zeichen: u), verglichen. Das Ergebnis des Vergleichs ergibt dann die Entscheidung. Die Entscheidungsvorschrift kann wie folgt formuliert werden:

Entscheidungsvorschrift

$$\text{Verwerfe} \quad H_0, \text{falls} \quad V > u.$$

Das bedeutet nichts anderes als:

Die H_0-**Hypothese wird abgelehnt**, wenn

$$V > u.$$

Das bedeutet, dass dann die Richtigkeit der H_1-Hypothese angenommen wird und wir unsere Behauptung statistisch belegen können.
Die Richtigkeit der H_0-**Hypothese wird angenommen**, wenn

$$V \leq u.$$

Das bedeutet, dass dann die Richtigkeit der H_1-Hypothese abgelehnt wird und wir unsere Behauptung statistisch *nicht* belegen können.

3.6.2. Aber wie kommen wir auf die Entscheidungsvorschrift?

Die genaue Herleitung erfordert einiges an stochastischen und statistischen Grundlagen, die an dieser Stelle zu weit führen würden. Daher beschränken wir uns auf ein grundlegendes Verständnis:
Die **Teststatistik** ist häufig ein *Schätzer* für den gesuchten Parameter. Möchte man zum Beispiel eine Trefferwahrscheinlichkeit nachweisen, so liegt es nahe, als Schätzung die durchschnittliche Trefferzahl in der Stichprobe zu nehmen. Solche Schätzer werden mit Hilfe statistischer Verfahren für Hypothesentests bestimmt.
Nun zur **kritischen Schranke**: Bei einer Entscheidung kann es, wie auch im richtigen Leben, zu einer Fehlentscheidung kommen. Es können vier Entscheidungen auftreten, die in Abbildung 14 aus [3, S. 270] dargestellt sind.

		In der Grundgesamtheit ist richtig:	
		Nullhypothese H_0	Alternative H_1
Das Testverfahren liefert aufgrund der Stichprobe die Entscheidung für	H_0	korrekte Entscheidung	Fehler 2. Art (β-Fehler)
	H_1	Fehler 1. Art (α-Fehler)	korrekte Entscheidung

Abbildung 14: **Entscheidungsmöglichkeiten.** [3, S. 270]

Fehlentscheidungen treffen wir dann, wenn wir uns für H_1 entscheiden, aber H_0 richtig ist (Fehler 1. Art, α-Fehler), oder wenn wir uns für H_0 entscheiden, aber H_1 richtig ist (Fehler 2. Art, β-Fehler). Solche Fehler sind zu vermeiden bzw. die Möglichkeit einer Fehlentscheidung ist möglichst gering zu halten. Hierbei ist, wenn man an medizinische Tests denkt, eine Fehlentscheidung erster Art gravierender als die zweiter Art. Möchte man nachweisen, dass eine neue Heilmethode besser ist als die alte (H_1-Hypothese) und entscheidet sich fälschlicherweise für diese Hypothese, so kann den Patienten weniger gut geholfen werden als zuvor. Begeht man jedoch bei diesem Beispiel einen Fehler zweiter Art, so entscheidet man, die alte Methode beizubehalten und erhält den Erfolg der alten Heilmethode. Die Minimierung beider Fehlertypen ist jedoch im Allgemeinen nicht möglich. Deshalb hilft man sich aus, indem man den Fehler 1. Art nach oben begrenzt und den Fehler 2. Art möglichst klein macht.
Betrachten wir die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen. Wir entscheiden uns für H_1, wenn die Entscheidungsvorschrift erfüllt ist, also dann, wenn $V > u$. Das bedeutet, die Wahrscheinlichkeit, eine solche Entscheidung zu treffen, ist:

Definition *Fehlerwahrscheinlichkeit 1. Art*:
Es seien die Teststatistik V und die kritische Schranke u gegeben für das Testproblem $H_0 \leftrightarrow H_1$ mit Vorschrift „Verwerfe H_0, falls $V > u$“. Wir wählen die Bezeichnung P für die Wahrscheinlichkeit, dann gilt: Die Wahrscheinlichkeit für die Ablehnung von H_0, also die Fehlerwahrscheinlichkeit 1. Art, ist

$$P(\text{„}H_1 \text{ entscheiden, obwohl } p_H \in H_0 \text{ richtig“}) = \sup_{p_H \in H_0} P(V > u)$$

Die Abkürzung *sup* steht hierbei (und bei der nächsten Definition) für das Supremum, das gemäß [16, S. 16] als kleinste obere Schranke definiert ist.
Die andere Art, einen Fehler zu begehen, ist, wenn wir uns für H_0 entscheiden, aber H_1 richtig ist. Diesen Fehler 2. Art machen wir aufgrund der Entscheidungsregel, wenn $V \leq u$ gilt. Es ist also:

Definition *Fehlerwahrscheinlichkeit 2. Art*:
Wir wählen die Voraussetzungen wie zuvor, dann gilt: Die Wahrscheinlichkeit für die Akzeptanz von H_0, also die Fehlerwahrscheinlichkeit 2. Art, ist

$$P(\text{„}H_0 \text{ entscheiden, obwohl } p_H \in H_1 \text{ richtig“}) = \sup_{p_H \in H_1} P(V \leq u)$$

Bei solchen Hypothesentests wird als Begrenzung die Fehlerwahrscheinlichkeit 1. Art vorgegeben durch ein $\alpha \in (0, 1)$. Der Wert α ist dann die Wahrscheinlichkeit, die für eine solche Fehlentscheidung noch akzeptabel ist. Üblich sind Werte $\alpha \in \{0.1, 0.05, 0.01\}$. Mit Hilfe dieser Vorgabe wird dann die Fehlerwahrscheinlichkeit 2. Art möglichst klein gehalten. Verfahren dazu werden in der mathematischen Statistik entwickelt und sollen hier nicht weiter diskutiert werden.
Die kritische Schranke wird nun aus der Forderung hergeleitet, dass die Fehlerwahrscheinlichkeit 1. Art maximal α betragen darf. Hierzu werden Überlegungen zu Wahrscheinlichkeitsverteilungen genutzt. In unserem Fall gilt:

Die **Teststatistik** ist

$$V = \sqrt{N} \cdot \frac{\overline{p} - p_Z}{\sqrt{\overline{p} \cdot (1 - \overline{p})}},$$

wobei $\overline{p}$ der relative Anteil der richtigen Kategorisierungen der Heuristik, N die Größe der Stichprobe und p_Z die Wahrscheinlichkeit von Übereinstimmungen durch den Zufall ist. Sie wird mit der **Schranke** $u_{1-\alpha}$ verglichen.

Die Schranke ist abhängig von der Fehlerwahrscheinlichkeit 1. Art α. Ihre Berechnung würde hier zu weit führen und weitaus größere statistische und stochastische Grundlagen erfordern. Die Werte der Schranke können in Tabellen nachgeschaut werden und sind für den CAMMP day in MATLAB bereits durch eine Funktion eingegeben worden.

Diese Ausführungen sollen uns genügen, um die Idee hinter den Werten zu verstehen. Sie dienen uns auch zur Beleuchtung eines weiteren wichtigen Begriffs:

3.6.3. Signifikanzniveau

Ein entscheidender Begriff im Umgang mit Hypothesentests ist das Signifikanzniveau.

Definition *Signifikanzniveau*:
Für ein $\alpha \in (0,1)$ wird ein Test der Form „Verwerfe H_0, falls $V > u$“ durch eine geeignete Festlegung der kritischen Schranke u definiert. Die Bedingung an diesen Wert lautet:

$$P(\text{„}H_1 \text{ entscheiden, obwohl } p_H \in H_0 \text{ richtig “}) = \sup_{p_H \in H_0} P(V > u) \leq \alpha.$$

α heißt **Signifikanzniveau**. Lehnt ein Test die Nullhypothese ab, so wird der Inhalt der Alternative als *signifikant zum Niveau* α bezeichnet.

Mit Hilfe dieses Begriffs kann die „Verlässlichkeit“ eines Ergebnisses beurteilt werden. Je niedriger das Signifikanzniveau ist, desto niedriger ist auch die Wahrscheinlichkeit für einen Fehler 1. Art. Aus der Definition wird außerdem deutlich:

Sind $\alpha_1, \alpha_2 \in (0,1)$ mit $\alpha_1 < \alpha_2$ und gilt für einen Test der Form „Verwerfe H_0, falls $V > u$“,

$$\sup_{p \in H_0} P(V > u) \leq \alpha_1,$$

so gilt auch

$$\sup_{p \in H_0} P(V > u) \leq \alpha_2,$$

Das bedeutet: Ist der Inhalt der Alternative signifikant zum Niveau α_1, so ist er dies auch für jedes höhere Signifikanzniveau α_2.

Diese Eigenschaft werden wir uns bei der Durchführung der Hypothesentests zu Nutze machen.

3.7. Veränderungen und -verbesserungen

Entscheidend für ein Modell sind die Vereinfachungen bzw. die Annahmen, die getroffen werden. Mögliche Modellveränderungen und -verbesserungen können daher durch Änderung der Vereinfachungen und Annahmen erzielt werden. In diesem Kapitel werden Vorschläge dazu vorgestellt. In der ersten CAMMP day Version sind sie noch nicht umgesetzt worden, können jedoch in der weiteren Entwicklung diskutiert werden. Die Vorschläge lauten:

- **Modellierung als gerichteter Graph:** Die Modellierung erfolgte, wie dargelegt wurde, durch einen ungerichteten Graphen. Betrachtet man reale soziale Netzwerke, so findet man nicht nur Informationen über die Freundschaften der

Nutzer, sondern auch über angeklickte Seiten, Kommunikationen und vieles weitere. Gerade bei Netzwerken wie beispielsweise Twitter ist es möglich, anderen Nutzern „zu folgen“, sodass die Kanten des Graphen einen Richtungssinn erhalten können. Dies bedingt jedoch auch die Nutzung eines anderen Datensatzes.

- **Nutzung eines anderen Datensatzes:** Der hier verwendete Datensatz des Netzwerks Friendster ist aufgrund der Anlehnung an die Forschung von Sarigol, Garcia und Schweitzer verwendet worden. Die Abrufbarkeit der Daten war ein weiteres Argument hierfür. Dennoch muss beachtet werden, dass aktuelle Netzwerke mehr Informationen beinhalten als hier durch die Daten bereitgestellt werden konnte. Die Zugänglichkeit von diesen Daten kann jedoch problematisch sein. Des Weiteren ist die Größe des Datensatzes (hier: die ersten 1 Millionen Nutzer) längst nicht repräsentativ für die aktuell genutzten Netzwerke. Dazu beachte man, dass Facebook über 1,3 Milliarden Nutzer hat [15]. Hier kann eine weitere Filterung der Daten erfolgen, wobei die Nutzeranzahl von Friendster insgesamt deutlich kleiner als die Größe der aktuellen Netzwerke ist.

- **Modellierung mit gewichteten Kanten:** In der Forschung von Sarigol, Garcia und Schweitzer erfolgte die Modellierung mit gewichteten Freundschaften. Hier wurde eine Freundschaft anhand der gemeinsamen Freunde gewichtet [9, S. 4]. Diese Erweiterung der Modellierung ist auch hier durchführbar. Hat man zudem Informationen über das Anschauen einer Seite der Nutzer innerhalb des sozialen Netzwerks, so kann eine Gewichtung mit einem ähnlichen Vorgehen wie beim *Page Rank Algorithmus* aus dem CAMMP day zum Thema Google, erläutert in der Bachelorarbeit von Sarah Schönbrodt [32], genutzt werden.

- **Verringerung der Vereinfachung durch Kombination der Vereinfachungen 1 bis 5 und weitere Heuristiken:** Im Rahmen der Forschungsarbeit wurden zudem Kombinationen der Vereinfachungen 1 bis 5 als Heuristiken getestet. Außerdem wurden Heuristiken genutzt, die beispielsweise nur die Freunde betrachten, die im Altersspektrum von ± 5 Jahren der Nutzer lagen. Eine gewinnbringendere Variante als die Heuristik 4 wurde dabei jedoch nicht gefunden. In diesem Gebiet ist durchaus eine Weiterentwicklung - auch für Schülerinnen und Schüler, die etwas programmieren können - möglich.

- **Beachtung von falschen Angaben:** Laut der Studie „Soziale Netzwerke. Eine repräsentative Untersuchung zur Nutzung sozialer Netzwerke im Internet“ des

BITKOM [23] haben ca. 25% der Netzwerkmitglieder schon einmal bewusst falsche Angaben gemacht. Da wir zu Beginn der Modellierung die Annahme der Richtigkeit der Angaben gemacht haben, sollte diese überdacht werden. Eine erneute Untersuchung unter Beachtung der Studienergebnisse des BITKOM könnte zu einer Verbesserung oder zumindest zu einer Näherung an die Realität führen.

- **Weitere Optimierungen:** Ein weiterer Punkt zur Verbesserung wäre, eine andere Optimierungsmethode einzuarbeiten. Ein erster Schritt ist die Überarbeitung der Optimierung für relative Parameter durch Beachtung der Tatsache, dass die Parameter gleich sein können. Zudem kann in der Weiterentwicklung die Stetigkeit bzw. die Differenzierbarkeit der Funktion der Kappawerte diskutiert bzw. erforscht werden. Sicherlich sind auch viele weitere Verfahren zur Optimierung geeignet.

- **Weiterreichende Forschungsfragen:**
  - Auch bei unseren Daten kann, wie bei Sarigol, Garcia und Schweitzer [9], beispielsweise die Abhängigkeit von Kappa von der Anzahl der Nutzer, die ihre sexuelle Orientierung angegeben haben, untersucht werden. In der vorliegenden Arbeit sind wir davon ausgegangen, das wir die sexuelle Orientierung aller Freunde eines Nutzers kennen. Hier kann man verschieden **Grade der Enthüllung** der sexuellen Orientierungen untersuchen.
  - Das **maschinelle Lernen** kann noch erweitert werden, indem die Nutzer des Datensatzes in ein Trainingsset und in ein Testset unterschiedlicher Größen eingeteilt werden. Die Heuristiken werden dann auf dem Trainingsset in ihrer Vorhersagekraft untersucht. Bei den Nutzern im Testset werden die sexuellen Orientierungen ausgeblendet und schließlich die Heuristiken, die gute Vorhersagen für das Trainingsset geliefert haben, hier angewandt.

  Diese Vorschläge bringen die hier betriebene Forschung näher an die Arbeit der drei Forscher Sarigol, Garcia und Schweitzer.

4. ...und das ist das Lernmodul - Der CAMMP day und ein didaktischer Blick darauf

In diesem Kapitel wird der CAMMP day vorgestellt und (teilweise zugleich) die konkreten didaktischen Gedanken erläutert.

4.1. Methodisches Konzept

Anknüpfend an die fachdidaktischen und fachlichen Hintergründe verfolgt der CAMMP day konkret verschiedene Ziele:

4.1.1. Ziele des CAMMP days

Die Schülerinnen und Schüler sollen im Verlauf des CAMMP days ein für sie *relevantes Thema* bearbeiten und hierbei den *Modellierungskreislauf durchlaufen und näher begreifen*. Dabei sollen sie am Ende die gestellte *Problemfrage lösen* können, was mit einer *Reflexion über ihre Verantwortung in sozialen Netzwerken* einhergeht. Die Erarbeitung soll weitestgehend *eigenständig* erfolgen. Dazu sollen die Lernenden *die Software MATLAB als digitales Werkzeug nutzen*. Das Material soll *an das Vorwissen der Schülerinnen und Schüler anknüpfen*. Es soll ihrer Heterogenität (im Sinne von Verschiedenheit) in *inhaltlichen Interessen*, *Lerntempo* und *Leistungsniveau* begegnen. Schließlich soll eine *Vermittlung der (neuen) mathematischen Inhalte* (Graph, Heuristiken, Gütekriterien, Optimierung, Hypothesentests) erfolgen. Das Ganze sollte den Schülerinnen und Schülern zudem *Spaß machen*.

Um die gesteckten Ziele zu erreichen, wird zunächst der Bezug zur Schulmathematik, bei dem unter anderen das vorausgesetzte Wissen der Schülerinnen und Schüler erläutert wird, beschrieben:

4.1.2. Bezug zur Schulmathematik

Das Lernmodul, in Form des CAMMP days, richtet sich an Schülerinnen und Schüler in der Qualifikationsphase der Sekundarstufe II eines Gymnasiums oder einer Gesamtschule. Dabei kann ein Beitrag zur Allgemeinbildung sowie ein deutlicher Bezug zum Lehrplan in den Bereichen „Modellieren“, „Problemlösen“ und „Stochastik“ festgestellt werden. Im Folgenden werden die den CAMMP day betreffenden Aspekte dargestellt:

- **Beitrag zur Allgemeinbildung**
  In den Beschlüssen der Kultusministerkonferenz wurde 2003 für die Sekundarstufe I und 2012 für die Sekundarstufe II festgelegt, welche Bildungsziele länderübergreifend gelten sollen [17], [18]. Das Modellieren und das Problemlösen sind hierbei bei den *allgemeinen mathematischen Kompetenzen* sowohl für den Mittleren Schulabschluss [17, S. 8] als auch für die Allgemeine Hochschulreife [18, S. 10] aufgeführt. Bezug wird dabei auf die von Heinrich Winter entwickelten drei Grunderfahrungen im Mathematikunterricht genommen. Hier ist somit auch ein Bezug zur Einleitung dieser Masterarbeit und zur CAMMP Philosophie zu erkennen: Die beschriebenen Prozesse des Modellierens und des Problemlösens, die fachdidaktische Grundlage das CAMMP days sind, sind unsere Werkzeuge, um alltägliche Erscheinungen wahrzunehmen und zu verstehen. Sie sind zudem unsere Mittel zum Erwerb von über die Mathematik hinausgehenden Fähigkeiten, da die in den Prozessen vermittelten Denkweisen auch in anderen Lebensbereichen eine Rolle spielen.

- **Bezug zur Schulmathematik**
  Prozessbezogene Aspekte: Modellieren und Problemlösen
  Modellieren und Problemlösen sind ebenfalls feste Bestandteile der Kernlehrpläne des Landes Nordrhein-Westfalen [33], [34]:
  Am Ende der Sekundarstufe I sollen die Schülerinnen und Schüler gemäß [33, S. 14] beim *Modellieren* die Mathematik als Werkzeug zum Erfassen von Phänomenen der realen Welt nutzen, indem sie

  – Realsituationen in mathematische Modelle übersetzen.

  – gewonnene Lösungen in den realen Situationen überprüfen und interpretieren und ihren Lösungsweg ggf. verändern.

  In der Sekundarstufe II erfolgt dann gemäß [34, S. 18-19] das

  – Strukturieren: Mit konkreten Fragestellungen Sachsituationen erfassen und strukturieren sowie begründet Annahmen und Vereinfachungen treffen.

  – Mathematisieren: Sachsituationen in mathematische Modelle übersetzen und eine Lösung im mathematischen Modell erarbeiten.

  – Validieren: Die Lösung auf die Sachsituation beziehen und die Angemessenheit beurteilen sowie die Modelle verbessern und die Abhängigkeit der Lösung von den Annahmen reflektieren.

Hier sind deutliche Bezüge, gerade in der Benennung der Schritte, zu den Modellierungskreisläufen zu erkennen.

Im Bereich des *Problemlösens* sollen die Schülerinnen und Schüler gemäß [33, S. 14] am Ende der Sekundarstufe I Probleme erfassen, erkunden und lösen, indem sie

- inner- und außermathematische Problemstellungen mit eigenen Worten wiedergeben, sie erkunden, Vermutungen aufstellen und Probleme in Teilprobleme zerlegen.
- verschiedene Darstellungsformen, mathematische Verfahren und Problemlösestrategien, wie systematisches Probieren und Schlussfolgern, nutzen.
- Lösungswege und Ergebnisse, auch die Möglichkeit mehrerer Lösungen, überprüfen und bewerten.

Erlangt man die Allgemeine Hochschulreife, so sollte man gemäß [34, S. 19-20] unter anderem fähig sein,

- zu erkunden, indem Informationen recherchiert werden und Muster und Beziehungen erkannt werden.
- zu lösen, indem Ideen für Lösungswege entwickelt werden und Strategien und Prinzipien (z.B. Schätzen und Überschlagen, systematisches Probieren) genutzt werden.
- zu reflektieren, indem die Plausibilität von Ergebnissen geprüft, Ergebnisse bezüglich einer Fragestellung interpretiert werden oder verschiedene Lösungswege verglichen werden.

Diese genannten prozessbezogenen Kompetenzen werden in der Analyse der Konzeption des Lernmoduls eine Rolle spielen.

Inhaltsbezogener Aspekt: Stochastik

Im Lernmodul werden vor allem inhaltliche Kompetenzen im Bereich „Stochastik“ vorausgesetzt. Die Schülerinnen und Schüler sollten

- relative Häufigkeiten kennen und interpretieren können.
- das arithmetische Mittel kennen und interpretieren können.

- ein- und zweistufige Zufallsexperimente kennen, nachvollziehen und erstellen können.
- die Darstellung des Baumdiagramms deuten und erstellen können.
- die Summen- und die Pfadregel anwenden können.
- Chancen und Risiken beurteilen können.

Diese Kompetenzen werden in der Sekundarstufe I laut [33, S. 33-34] vermittelt. Dennoch bekommen die Lernenden Gelegenheit, ihr Wissen im Lernmodul aufzufrischen.
Außerdem werden inhaltliche Aspekte des Kernlehrplans der Sekundarstufe II [34, S. 23-34] aufgegriffen:

- Im Rahmen von mehrstufigen Zufallsexperimenten werden Situationen als solche gedeutet, Pfadregeln angewandt und Baumdiagramme genutzt.
- Die Binomialverteilung spielt implizit eine Rolle.
- Entscheidungsregeln werden in Hypothesentests verwendet und die Ergebnisse interpretiert.

Neben diesen Aspekten geht der fachliche Inhalt des Moduls über die Lehrpläne hinaus.

Als nächstes wollen wir uns den CAMMP day in seiner Konzeption näher anschauen.

4.1.3. Der Modellierungskreislauf liegt der Struktur des CAMMP days zu Grunde

Der Modellierungskreislauf im CAMMP day
Im CAMMP day wird die zentrale Fragestellung bzw. Problemstellung

> Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

durch Modellierung gelöst. Dabei wird der Modellierungskreislauf wie folgt durchlaufen:

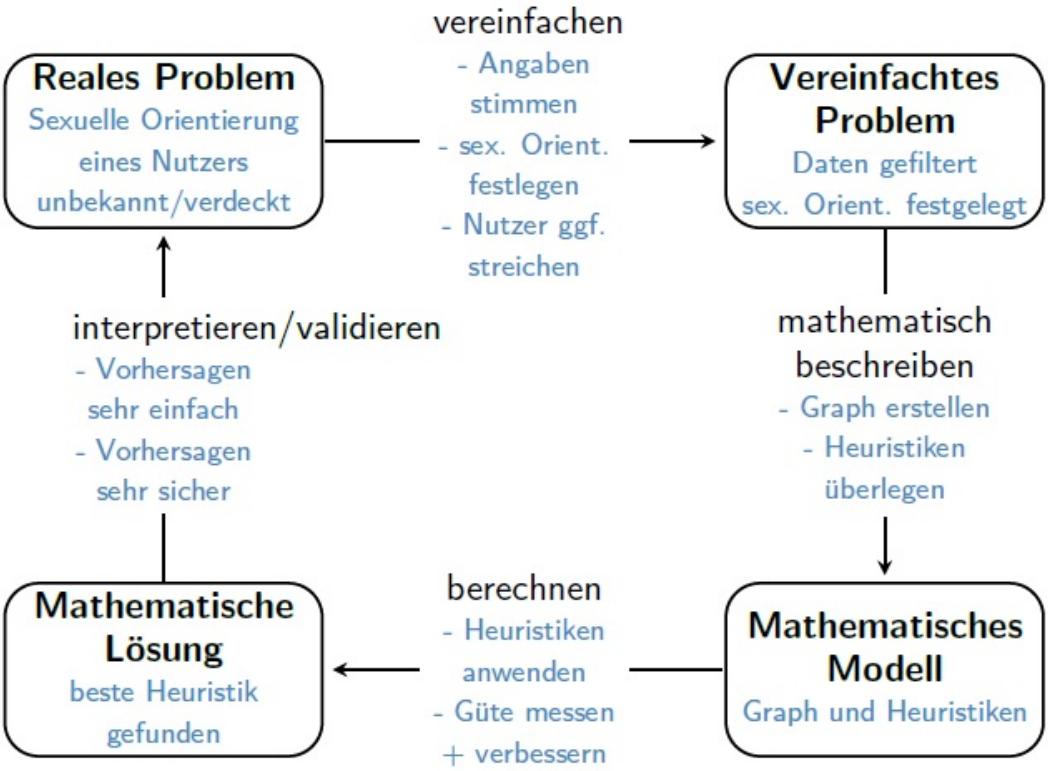


Abbildung 15: **Modellierungskreislauf im CAMMP day.**

Wir wollen uns Abbildung 15 genauer anschauen:
Das *reale Problem* im CAMMP day besteht darin, dass wir die sexuelle Orientierung eines Nutzers nicht kennen, aber kennen wollen. Wir möchten sie also vorhersagen und machen dazu einige *Vereinfachungen*. Zunächst gehen wir davon aus, dass die Angaben der Nutzer stimmen, wir legen, wenn möglich, vereinfachend ihre sexuelle Orientierung fest, betrachten nur Nutzer mit mindestens einem Freund und streichen Nutzer aus unserem anfänglichen Datensatz der ersten 1 Millionen, die unsere Kriterien nicht erfüllen. Im *vereinfachten Problem* haben wir dann die Daten gefiltert und jedem Nutzer eine sexuelle Orientierung zugeordnet. Nun gilt es, das Problem *mathematisch zu beschreiben*. Dazu werden ein erstellter Graph betrachtet und Heuristiken entwickelt. Das *mathematische Modell* besteht dann aus dem Graphen und den Heuristiken. Beim Schritt *berechnen* werden die Heuristiken angewendet, ihre Güte gemessen und ggf. verbessert. Um schließlich als *mathematische Lösung* die beste Heuristik zu benennen, wird der Schritt des Berechnens mehrfach durchlaufen. Zunächst werden die Heuristiken auf eine Person und dann auf mehrere Personen angewendet. Danach wird die Güte gemessen und im nächsten Schritt verbessert. Zusätzlich kann in einem

weiteren Durchlauf die Gültigkeit mittels Hypothesentests geprüft werden. Die *mathematische Lösung* besteht schließlich aus der besten Heuristik. Diese Lösung wird im letzten Schritt *interpretiert und validiert*, wobei wir zu dem Ergebnis kommen, dass die Vorhersagen sehr einfach (mit „Wenn-dann-Aussagen“) und sehr sicher (mit ca. 90%-iger Sicherheit) getätigt werden können. Im Anschluss daran bietet sich noch eine Diskussion von Verbesserungsmöglichkeiten bei den einzelnen Schritten an.
Werfen wir noch einmal einen Blick zurück in die fachdidaktische Theorie:
Es wurde zwischen *deskriptiven Modellen* (Modelle, die vorhersagen, erklären oder beschreiben) und *normativen Modellen* (Modelle, die etwas vorschreiben) unterschieden. Betrachten wir unser mathematisches Modell im CAMMP day, den Graphen und die Heuristiken, so stellen wir fest, dass sich hier beide Modelltypen wiederfinden lassen. Der Graph beschreibt das Netzwerk in seiner Struktur als deskriptives Modell und die Heuristiken sagen die sexuelle Orientierung durch eine bestimmte Vorschrift als normatives Modell vorher.
Die Modellierung wurde in der Theorie in Form von *realistischem oder angewandtem Modellieren* und von *pädagogischem Modellieren* beschrieben. Auch diese beiden Typen sind im CAMMP day zu finden: Das beim realistischen oder angewandten Modellieren verfolgte Ziel, die Realität zu verstehen, ist auch hier im CAMMP unser Ziel. Wir wollen verstehen, wie gut Schattenprofile erstellt werden können. Das pädagogische Modellieren wird durch Vermittlung des Modellierungsprozesses vor allem in den Präsentationen, durch Teamarbeit sowie durch Diskussionen und Präsentationen der Schülerinnen und Schüler bei den Besprechungen geschult.

4.1.4. Ablauf und Struktur des CAMMP days

Die Durchführung des CAMMP days erfolgt durch Professoren, Assistenten/-innen und Hilfswissenschaftler/-innen des Instituts. Die Schülerinnen und Schüler werden zu Beginn des CAMMP days mit einer *Begrüßungspräsentation* willkommen geheißen, wobei CAMMP und die dahinterstehenden Personen vorgestellt und der Tagesablauf erläutert wird. In dem folgenden Modellierungsvortrag, der von Professoren oder Assistenten des Instituts gehalten wird, wird der Prozess des Modellierens den Schülerinnen und Schülern an einem ausgewählten Beispiel (z.B. der Strahlentherapie) erläutert. Die nächsten Schritte werden vor allem von Hilfswissenschaftlern durchgeführt. Sie halten die Vorträge, unterstützen die Lernenden in ihren Arbeitsphasen und leiten die Besprechungen. Nach dem Modellierungsvortrag werden die Schülerinnen und Schüler durch eine *Präsentation zur Problemstellung* in die Thematik eingeführt und

die Schritte des Modellierungskreislaufs bis zum mathematischen Modell mit ihnen erarbeitet. Die Schritte „berechnen“ und „mathematische Lösung“ werden übersichtlich skizziert und das „Interpretieren und Validieren“ auf die Abschlussbesprechung verschoben. Nun sollen die Schülerinnen und Schüler den Schritt des Berechnens selbst durchführen. In der *ersten Partnerarbeitsphase* beschäftigen sich die Schülerinnen und Schüler mit Hilfe von Arbeitsblatt 1 genauer mit den Heuristiken und wenden sie auf einzelne Personen sowie auf viele Personen an. Die Ergebnisse werden in einer *ersten Besprechung* diskutiert. Nach einer kurzen Pause steht die Frage der Güte einer Heuristik im Mittelpunkt. Die Schülerinnen und Schüler bearbeiten in der *zweiten Partnerarbeitsphase* das Arbeitsblatt 2, bei dem sie Cohen's Kappa kennen- und nutzenlernen. Es folgt die *zweite Besprechung* der Ergebnisse. Während der zweiten Arbeitsphase haben die Lernenden zudem die Möglichkeit, mit Hilfe eines Zusatzblattes die Trefferquote und die Genauigkeit der Heuristiken zu untersuchen. Es folgt eine etwas längere *Mittagspause*. Nach dieser arbeiten die Schülerinnen und Schüler in der *dritten Arbeitsphase* mit Arbeitsblatt 3 an der Verbesserung der Heuristiken durch Optimierung. Die Ergebnisse werden in der *dritten Besprechung* diskutiert. Ist noch Zeit vorhanden, so ermitteln die Lernenden die Gültigkeit der Ergebnisse mittels Hypothesentests in der vierten oder als Zusatz in der dritten Partnerarbeitsphase mit Arbeitsblatt 4, dessen Besprechung ebenfalls stattfindet. Nach einer kurzen Pause erfolgt das Interpretieren und Validieren der Ergebnisse in der *Abschlusspräsentation*, in der die letztendliche Beantwortung der Problemstellung und eine Reflexion stattfinden. Nach dieser werden den Schülerinnen und Schülern noch weitere Angebote von CAMMP und verbundene Studiengänge vorgestellt und der CAMMP day wird von ihnen evaluiert.

Nach dieser intensiven Betrachtung des Modellierens im CAMMP day wollen wir nun die für den CAMMP day entwickelten Materialien genauer durch Betrachtung der weiteren fachdidaktischen Aspekte des fachdidaktischen Hintergrunds analysieren.

4.2. Didaktische Analyse der Materialien

4.2.1. Beschreibung und didaktische Zielsetzung der Materialien und der Arbeit mit diesen

Vorliegende Materialien

Zur Durchführung des CAMMP days liegen die Begrüßungs-, die Modellierungs-

und die abschließende Informationspräsentation zum CAMMP aus dem Archiv vor. Bezogen auf den konkreten CAMMP day wurde zur fachlichen Klärung für die Durchführenden ein sogenanntes basic paper erstellt, was dem fachlichen Hintergrund (Kapitel 3) entspricht. Sie werden so über die fachlichen Hintergründe informiert. Den organisatorischen Ablauf finden die Durchführenden im stichpunktartig zusammengefassten methodischen Konzept. Es sind Präsentationen zur Problemstellung und zur abschließenden Beantwortung der Fragestellung erstellt worden. Hierbei dienen Hinweise den Dozenten zur Durchführung. Für die Schülerinnen und Schüler sind eine Übersicht über den Ablauf des Tages sowie Arbeitsblätter mit Antwortblättern erstellt worden. Hierzu erhalten die Dozenten Lösungsvorschläge und Folien zur Besprechung. Zudem sind Hilfekarten verfügbar. Ein arbeitsprozessbegleitender Code wurde als MATLAB-Programm entwickelt, zur Arbeit mit diesem liegen den Dozenten ebenfalls Vorschläge in den Lösungsvorschlägen der Arbeitsblätter vor. Schließlich bewerten die Schülerinnen und Schüler den CAMMP day mit einem Evaluationsbogen.
Die genannten Materialien zum konkreten CAMMP day sind, bis auf das basic paper, im Anhang zu finden. Im Folgenden wird das für die Durchführung des CAMMP days für Schülerinnen und Schüler entscheidende Material didaktisch analysiert:

Modellierungsvortrag
Der Modellierungsvortrag wird von den Assistenten bzw. den Professoren gehalten und vorbereitet. Er dient **didaktisch** den Schülerinnen und Schülern zur Information über den Vorgang des Modellierens und zur Darstellung einer möglichen (aktuellen) Anwendung. Die einzelnen Schritte des vierstufigen CAMMP-Kreislaufs werden mit Inhalt gefüllt und erläutert.

Präsentation zur Eröffnung der Problemstellung
Bei der Präsentation zur Eröffnung der Problemstellung werden den Schülerinnen und Schülern zunächst einige Fragen zum Thema gestellt. Hierbei soll eine Identifikation mit der Thematik erfolgen. Es wird nach ihrer Mitgliedschaft in sozialen Netzwerken, nach ihrer Vorstellung und Einschätzung von Sicherheit in sozialen Netzwerken und nach preisgegebenen Informationen gefragt. Zudem wird ein Szenario diskutiert, indem ein Betreiber eines sozialen Netzwerks Informationen über Nutzer herausfinden möchte. Es werden mögliche Interessen und Gründe des Betreibers besprochen und eine Einschätzung des Erfolgs gefordert. Danach wird die Erstellung von Schattenprofilen thematisiert und ein Schattenprofil erklärt. Die leitende Fragestellung wird

anschließend vorgestellt und der Einstieg in den Modellierungskreislauf vollzogen. Dieser wird dann Schritt für Schritt mit den Schülerinnen und Schülern durchlaufen, indem die Datenfilterung und das Modell zum Teil interaktiv erarbeitet werden. Am Schluss der Präsentation wird das organisatorische Vorgehen im CAMMP erläutert.
Didaktisches Ziel der Präsentation ist, dass die Schülerinnen und Schüler die Problematik verstehen, sich mit ihr identifizieren und auseinandersetzen. Die Diskussionen werden **im Plenum** durchgeführt. Diese sehr gelenkte Art und Weise der Vermittlung erschien als notwendig, da zur Lösung des Problems das Verständnis der Problemstellung absolute Grundlage ist und so der Dozent die Erläuterung vornehmen kann.

Lösung des Problems:
Die folgenden Arbeitsphasen führen die Schülerinnen und Schüler zur Lösung des Problems. In jedem Schritt werden die Lernenden mit der zentralen Problemstellung konfrontiert, die sie nach und nach immer besser beantworten können. Den Phasen ist gemein, dass sie sich in **Partner-/Gruppenarbeit und anschließender Besprechung** gliedern. Diese Organisation wurde vorgenommen, da so die Schülerinnen und Schüler weitestgehend eigenständig arbeiten können, zudem eine Reflexion des Lösungsweges und eine Zusammenführung der (zum Teil arbeitsteiligen) Ergebnisse möglich ist. Die Besprechungen werden durch Folien und ggf. durch Nutzung der Tafel begleitet und die Ergebnisse so fixiert und visualisiert, wobei die Schülerinnen und Schüler ihre Ergebnisse angemessen darstellen und präsentieren sollen. Schauen wir uns die Phasen im Einzelnen an:

Arbeitsphase 1: Heuristiken - Einstieg
Das erste Arbeitsblatt beginnt **inhaltlich** mit einer genaueren Betrachtung der Heuristiken. Die Regeln sind dargestellt und die Phrase „Der Zufall entscheidet“ wird mit Hilfe eines Baumdiagramms erläutert. Die Schülerinnen und Schüler werden aufgefordert, Vermutungen zur Güte der Regeln aufzustellen. Im zweiten Schritt arbeiten die Schülerinnen und Schüler zum ersten Mal mit MATLAB. Sie wenden die Regeln auf einzelne Nutzer an, die so ausgewählt wurden, dass für jede der parameterabhängigen Heuristiken eine Person dabei ist, die in die Zufallsbestimmung fällt. Auch hier soll die Güte der Regeln eingeschätzt werden. Im dritten Schritt werden viele Nutzer gleichzeitig untersucht. Die Schülerinnen und Schüler dürfen sich hier eine Regel aussuchen, die sie behandeln wollen, wobei jede Regel von mindestens

einer Gruppe betrachtet werden soll. Hierbei untersuchen die Schülerinnen und Schüler zunächst verschiedene Stichprobengrößen und schauen, welche sexuelle Orientierung gut und welche weniger gut eingeschätzt werden kann. Bei den parameterabhängigen Heuristiken werden zudem verschiedene Parameter ausprobiert. Das Gesetz großer Zahlen wird indirekt behandelt, da die Schülerinnen und Schüler begründen müssen, warum eine „große“ Stichprobe gewählt wird bzw. welche Stichprobengröße am aussagekräftigsten ist. Am Ende steht erneut die Frage nach der Güte der Heuristiken. Ist eine Gruppe schnell fertig, so kann diese eine andere Regel bearbeiten. Die Ergebnisse der Schritte werden in einer anschließenden Besprechung verglichen und besprochen.
Didaktisches Ziel dieser Arbeitsphase ist es, anhand der konkreten Betrachtung der Regeln, anhand der Anwendung auf einzelne Personen und anhand der Anwendung auf viele Personen erneut die Güte zu beurteilen. Im zweiten Schritt soll zudem das Konzept einer Zufallsentscheidung deutlich werden durch die gezielte Vorgabe der Nutzer. Stärken und Schwächen der Heuristiken für Hetero- oder Homosexuelle werden im Schritt 3 deutlich. Hierbei wird auch die Parameterabhängigkeit der Heuristiken erforscht und der Inhalt des Gesetzes großer Zahlen erkannt. Das Aussuchen einer Regel erzeugt ein **arbeitsteiliges Vorgehen**. So können die Interessen der Schülerinnen und Schüler beachtet werden und am Ende alle Lernenden von einem gemeinsamen Ergebnis profitieren. Ein weiterer Vorteil dieser Arbeitsteilung liegt im Sparen von Arbeitszeit, sodass hier der Fokus nicht auf die Behandlung vieler Regeln, sondern auf das exemplarische Behandeln und auf ein tieferes Verständnis gelegt wird. Der Vergleich der Heuristiken soll den Schülerinnen und Schülern schwer bzw. umständlich fallen, sodass eine einheitliche Größe zur Bewertung im nächsten Schritt sinnvoll wird.

Arbeitsphase 2: Heuristiken - Güte bewerten
In der zweiten Arbeitsphase wird **inhaltlich** Cohen's Kappa eingeführt und genutzt. Das Übereinstimmungsmaß wird zunächst definiert und diese Definition erarbeitet. Zudem bestimmen die Schülerinnen und Schüler die Wahrscheinlichkeit einer richtigen Vorhersage mit Hilfe des Zufalls, wobei sie hier durch Hilfekarten Unterstützung finden. Ihre Antwort überprüfen sie durch das Programm in MATLAB. Die Deutung eines Kappawertes wird in der nächsten Aufgabe behandelt. Mit Hilfe eines Lückentextes und weiterer Hilfekarten werden die Schülerinnen und Schüler zur Deutung geführt, die sie erneut mit MATLAB kontrollieren können. Im letzten Teilschritt suchen sich die Lernenden erneut eine Regel zur Erforschung aus und bestimmen bei festgelegter

„großer“ Stichprobengröße den Kappawert. Bei den parameterabhängigen Heuristiken erforschen sie verschiedene Parameter. Es folgt ein Vergleich der Güte der Heuristiken mit Hilfe von Kappa. Zusätzlich haben die Schülerinnen und Schüler die Möglichkeit, weitere Regeln zu testen und/oder ein zusätzliches Arbeitsblatt zu bearbeiten.
Didaktisches Ziel dieser Arbeitsphase ist, dass die Lernenden ein wissenschaftliches Werkzeug kennen- und nutzenlernen, das den Vergleich der Güte vereinfacht. Hierbei steht jedoch immer noch das Entdecken von guten Heuristiken im Vordergrund.

Zusatz: Heuristiken - Güte bewerten
In dieser zusätzlichen Arbeitsphase bekommen die Schülerinnen und Schüler **inhaltlich** die Möglichkeit, zwei weitere Gütekriterien, nämlich die Trefferquote und die Genauigkeit, kennenzulernen und die Heuristiken mit Hilfe dieser einzuschätzen. Die Gütekriterien werden zunächst definiert und somit eingeführt. Anschließend füllen die Lernenden, analog zur Vorgehensweise bei Cohen's Kappa, einen Lückentext zu den Eigenschaften der Gütemaße aus. Hierzu können sie auch zwei Hilfekarten nutzen. Die Lösung kann mit MATLAB kontrolliert werden. Schließlich dürfen sich die Schülerinnen und Schüler eine Regel zur Untersuchung aussuchen, hier ggf. verschiedene Parameter ausprobieren und die Güte einschätzen.
Didaktisches Ziel ist hierbei zum einen das Arbeiten mit weiteren Gütemaßen zum anderen aber auch die Einsicht, dass die Heuristiken zum Teil anders eingeschätzt werden und Regeln unter bestimmen Fragestellungen genutzt werden sollten (z.B. Regel 3 mit der besten Trefferquote für Homosexuelle). Die Stärken und Schwächen der Regeln kommen deutlich zum Ausdruck.

Arbeitsphase 3: Kappa verbessern
Das dritte und letzte Arbeitsblatt im Basisprogramm dieses CAMMP days wird **inhaltlich** in der Motivation bearbeitet, in einfacherer Form „gute“ Parameter für die Heuristiken 1, 2, 4 und 5 zu finden. Die Schülerinnen und Schüler arbeiten **in Teams** mit der Monte-Carlo-Methode. So kann der zeitliche Aufwand für die Optimierungen reduziert werden. Zunächst wird erneut eine Regel gewählt und sich auf ein gemeinsames Vorgehen durch Festlegung der Stichprobengröße und der Intervalle geeinigt. Die Optimierungen werden nachfolgend mehrfach durchgeführt. Um den Einfluss des Zufalls zu verringern, wird dann innerhalb der Gruppe der Mittelwert der Kappawerte der am häufigsten auftretenden Parameter bestimmt und als Ergebnis angesehen. Eine Reflexion und ggf. eine Verfeinerung der Optimierung

wird am Ende der Arbeitsphase getätigt.
Didaktisches Ziel dieser dritten Arbeitsphase ist, das Kennenlernen und Nutzen der Optimierung als wissenschaftliche Vorgehensweise mit ihren Vor- und Nachteilen.

Zusatz: Hypothesentests
Zusätzlich zu den beschriebenen Materialien kann der Lerngruppe das Arbeiten mit Hypothesentests angeboten werden. **Inhaltliche** Motivation hierzu ist, dass die Gültigkeit der (besten) Ergebnisse geprüft werden kann. Mit Hilfe des vierten Arbeitsblattes werden die Schülerinnen und Schüler mit der Theorie hinter der Thematik bekannt gemacht. Sie durchlaufen, teils aktiv, den Schritt der Hypothesenbildung. Hier werden sie durch Kontrollinstanzen in MATLAB unterstützt. Die Tests, zunächst bezogen auf eine Regel, werden durchgeführt und das Ergebnis in Form des Vergleichs von der Teststatistik mit der Schranke in MATLAB angezeigt. Eine Interpretation wird von den Schülerinnen und Schülern vollzogen. Zusätzlich können Tests mit weiteren Regeln durchgeführt werden.
Didaktisches Ziel ist hierbei, dass den Schülerinnen und Schülern die Bedeutung einer Stichprobe deutlich wird. Hypothesentest sind zudem ein gängiges Mittel in der statistischen Forschung, weshalb sich ihre Kenntnis lohnt.

Präsentation zum Abschluss der Problembearbeitung
Die Abschlusspräsentation rundet den CAMMP day ab. Als erstes wird hier die zentrale Fragestellung, der Modellierungskreislauf mit Fokus auf den Schritt des Berechnens und der aktuelle Schritt des Interpretierens der Lösung vom Dozenten wiederholt. Gemeinsam mit den Schülerinnen und Schülern wird **im Plenum** die Fragestellung beantwortet, indem das beste Vorgehen zur Bestimmung der sexuellen Orientierung und durch Umformung von Kappa die Treffsicherheit der besten Heuristik erläutert wird. Beispielhaft wird den Lernenden anhand eines Graphen die Bedeutung des Ergebnisses visualisiert. Danach werden die Schülerinnen und Schüler zur persönlichen Reflexion des Ergebnisses aufgefordert. Dies kann durch Bezug auf aktuelle Netzwerke (beispielsweise Facebook) geschehen. Optionalerweise erfolgt eine Diskussion von möglichen Verbesserungen im Modellierungsprozess. Zusammenfassend liegt das **didaktische Ziel** in der Beantwortung der Fragestellung und der (persönlichen) Reflexion.

Evaluation
Anhand des im Anhang befindlichen Evaluationsbogens erhalten die Lernenden die Möglichkeit, den CAMMP day zu bewerten. Die zuvor formulierten Ziele des Lernmoduls sind hier durch Aussagen wiederzufinden. Genauer wird auf diesen Aspekt des CAMMP days im nächsten Kapitel eingegangen.

4.2.2. Das Problemlösen im CAMMP day

Die ersten Schritte „Verstehen einer Aufgabe“ und „Ausdenken eines Plans“ des Problemlöseprozesses nach Polya werden in der Präsentation zur Eröffnung der Problemstellung gemeinsam mit den Lernenden gemacht. Außerdem wurde eine erste Einschätzung zur Güte von Vorhersagen getätigt und auf einer Folie fixiert. Hier begegnen die Schülerinnen und Schüler der ersten Entdeckungsaufgabe nach Polya und Kranz.
Alle folgenden Phasen zur Problemlösung finden im Problemlöseprozess ihren Platz im Schritt „Ausführen des Plans“. Im Bezug auf das Problemlösen stellen sich die Schülerinnen und Schüler bei der Bearbeitung der Arbeitsblätter Entdeckungsaufgaben. Sie bekommen an vielen Stellen die Chance, Möglichkeiten auszuprobieren und stellen am Ende anhand ihrer Entdeckungen erneut (immer konkretere) Vermutungen auf. Diese Vermutungen werden anhand von Entscheidungen für oder gegen eine Regel getätigt. Hier handelt es sich also um Entscheidungsaufgaben. Zum Kennen- und Nutzenlernen von Cohen's Kappa werden ihnen zudem Bestimmungsaufgaben gestellt.
Die Abschlussbesprechung beinhaltet den letzten Schritt des Problemlösens „Rückschau und Ausblick“ durch Reflexion des Lösungsprozesses, Beantwortung der Fragestellung und Diskussion von Verbesserungen im Modellierungsprozess.

4.2.3. Die Schulmathematik im CAMMP day

Beziehen wir die Phase der Einführung in die Problemstellung auf die Schulmathematik, so kann hier in eher passivem Sinne das Übersetzen einer Realsituation in mathematische Modelle, der Schritt des Strukturierens und ein Teil des Mathematisierens getätigt werden. Inhaltlich müssen hier relative Häufigkeiten bekannt sein. Bei den Arbeitsphasen zur Lösung des Problems wird im Bereich des Modellierens das Mathematisieren, durch Erarbeitung einer Lösung, und im Bereich des Problemlösens vor allem das Erkunden und das Aufstellen von Vermutungen sowie die Nutzung von Problemlösestrategien und das Prüfen verschiedener Lösungen geschult. Inhaltlich werden nun zu den bereits genutzten relativen Häufigkeiten die Kenntnis des arithmetischen Mittels, von ein- und zweistufigen Zufallsexperimenten, von Baumdiagrammen (und

hierbei die Pfad- und Summenregel) und schließlich das Beurteilen von Chancen und Risiken vorausgesetzt. Im vierten und zusätzlichen Arbeitsauftrag werden durch das zugrundeliegende Modell die Binomialverteilung genutzt und Entscheidungen mit Hypothesentests getroffen. In der Abschlussbesprechung/-präsentation werden die Kompetenzen, gewonnene Lösungen zu überprüfen, zu interpretieren und den Lösungsweg ggf. zu verändern (Validieren und Reflektieren) genutzt bzw. geschult.

4.2.4. Die Nutzung von Vorteilen von Modellierung im CAMMP day

Ein Vorteil von Modellierung wird bereits bei der Problemstellung genutzt: die Authentizität. Im CAMMP day dürfen sich die Schülerinnen und Schüler mit einem echten Problem auseinandersetzen. Bei der Lösung des Problem stoßen die Lernenden immer wieder auf Phasen, in denen sie sich einen Lösungsweg aussuchen und viele Ideen ausprobieren können. Hier ist der Vorteil der Offenheit zu finden. Das Differenzierungsvermögen der Aufgaben liegt hier im Aussuchen des (unterschiedlich komplexen) Inhalts aber auch im Auffangen von verschiedenen Lerntempi, da Zusatzaufgaben verfügbar sind. Der Schwerpunkt des CAMMP days liegt auf der Bestimmung der mathematischen Lösung. So sind die Schülerinnen und Schüler nicht direkt mit der kompletten Durchführung der Modellierung konfrontiert. Das Anforderungsniveau ist reduziert und ein besseres Verständnis für die Teilschritte der Lösungsbestimmung kann erreicht werden. Des Weiteren ist der Vorteil von Simulationen zur Entscheidung für die beste Regel genutzt, da verschiedene Regeln als verschiedene Szenarien zur Bestimmung der sexuellen Orientierung mit Hilfe von MATLAB ausprobiert werden. Gerade in der Optimierung wird dieser Vorteil intensiv genutzt. Der Aspekt der Verantwortung und Ethik ist in der Abschlussbesprechung verortet. Eine individuelle und gesellschaftliche Reflexion des Ergebnisses findet hier statt.

4.2.5. MATLAB als digitales Werkzeug im CAMMP day

Ein MATLAB-Programm begleitet die Schülerinnen und Schüler durch den CAMMP day. Nachdem die realistischen Daten durch die Dozenten bereits geladen wurden, finden die Schülerinnen und Schüler zu jedem Lösungsschritt einen Abschnitt im MATLAB-Programm. Hierbei wird MATLAB genutzt, um die Heuristiken auf eine Person und auf viele Personen anzuwenden. MATLAB dient als Kontrollinstanz bei der Berechnung der Trefferwahrscheinlichkeit durch den Zufall sowie bei der Lösung der Lückentexte und der Formulierungen der Hypothesen. Des Weiteren bestimmt MATLAB die Gütekriterien zu den Heuristiken und führt die Optimierungen sowie die

Hypothesentests durch. Zudem kann MATLAB als Taschenrechner eingesetzt werden. Zusammengefasst wird MATLAB auf vielerlei Weisen eingesetzt: zur Vermeidung von Rechenfehlern, zur Bewältigung der Bearbeitung von großen Datenmengen, zum Experimentieren und Entdecken, zum Ausprobieren (Optimieren) in einer angemessenen Zeit und zur Bestimmung von statistischen Größen. Der Einsatz ist hier sehr nützlich und sinnvoll.

4.2.6. Antizipierte Lernhürden und Unterstützungsmaßnahmen

An bestimmten Stellen des CAMMP days wurden Lernhürden der Schülerinnen und Schüler antizipiert. Auf diese wollen wir nun genauer eingehen:
Im Allgemeinen sind zunächst Antwortblätter entwickelt worden, um die Lernenden in einer Strukturierung und Dokumentation der Ergebnisse zu leiten und eine gemeinsame Besprechung zu ermöglichen. Außerdem sind inhaltliche Hinweise, wie auf dem ersten Arbeitsblatt zu den einzelnen Regeln, zur Deutung der Ergebnisse gegeben. Des Weiteren finden sich methodische Hinweise, wie z.B. auf Arbeitsblatt 2 zur Nutzung von MATLAB als Taschenrechner, im Material. Durch den wiederholt auftretenden Hinweis zur Besprechung der Ergebnisse am Ende jedes Arbeitsblattes werden der CAMMP day und die Schülerinnen und Schüler in der organisatorischen Struktur unterstützt. Zum Umgang mit MATLAB ist die Vorgehensweise für die Schülerinnen und Schüler, die dem Programm vermutlich zum ersten Mal begegnen, sehr detailliert unter anderem durch Angabe von Zeilennummern beschrieben.
An einzelnen Stellen begegnen Hilfekarten antizipierten Lernhürden. Bei der Bestimmung der Trefferwahrscheinlichkeit durch den Zufall können die Schülerinnen und Schüler drei Hilfekarten nutzen. Das Konzept der Wahrscheinlichkeit wurde hier als schwer zu verstehen eingestuft. Auf der ersten Karte ist das zugehörige Baumdiagramm gezeichnet und die Vorgehensweise zur Lösung beschrieben. Die nächste Karte ergänzt die erste durch eingetragene Wahrscheinlichkeiten. Die dritte und letzte Karte enthält zusätzlich den Ansatz zur Anwendung der Pfad- und der Summenregel. Eine Unterstützung findet hier in drei Stufen statt.
Die für die Schülerinnen und Schüler neue Größe Cohen's Kappa kann ebenfalls mit Hilfekarten (4 und 5) erarbeitet werden. Auch hier sind zwei Stufen der Unterstützung zu finden: Zunächst werden die auftretenden Möglichkeiten beim Vergleich der Heuristik mit dem Zufall benannt und die Anteile richtiger Vorhersagen konkret verglichen. Konkrete Fragen zu den Zahlen, die Zähler und Nenner annehmen, und zum gesamten Bruch unterstützen die Schülerinnen und Schüler auf dem Weg zur Lösung. Die nächs-

te Stufe der Hilfe beinhaltet auf der fünften Hilfekarte die einzusetzenden Zahlen. Die Aufgabe wurde so zu einer Zuordnungsaufgabe reduziert.
Vergleichbar ist die Unterstützung durch Hilfekarten beim Kennenlernen der Trefferquote und der Genauigkeit. Zunächst sollen auf Hilfekarte 6 Leitfragen die Gedanken der Schülerinnen und Schüler unterstützen und danach wird auf Hilfekarte 7 die Aufgabe zu einer Zuordnungsaufgabe reduziert.
Neben diesen analogen Unterstützungsmaßnahmen erhalten die Schülerinnen und Schüler bei der Arbeit mit dem Code ein differenziertes Feedback. Dieses unterstützt sie bei der Arbeit mit dem Code, beispielsweise bei der Wahl der Parameter:

```
%%      Heuristiken auf viele Personen anwenden

%% mit Regel 1:
clc
N = 1000;
params = [4  5];

% Personen einschätzen:
heuristicTester_EinstiegRegel12(@Heuristic_Regel1params,
Users, N, params)
```

Die Parameter sind hier nicht sinnvoll. Es erscheint das Feedback:

```
Ihr habt Zahlen gewählt, die hier keinen Sinn machen.
>>
```

An den Stellen, an denen MATLAB als Kontrollinstanz auftritt, werden den Schülerinnen und Schülern Hinweise zum Weiterarbeiten gegeben. Zur Erläuterung dient ein Beispiel zur Erkundung der Eigenschaften von Cohen's Kappa:

```
%%      Kappa betrachten
clc

a1 = 0;
a2 = 1;
a3 = 0;
a4 = 0;

% Prüfe die Eingabe:
PruefeLueckentext(a1,a2,a3,a4);
```

```
Gut! Eure Antwort ist richtig
```

Bei Fehler:

```
%%      Kappa betrachten
clc

a1 = 1;
a2 = 1;
a3 = 0;
a4 = 0;

% Prüfe die Eingabe:
PruefeLueckentext(a1,a2,a3,a4);
```

```
Die Antwort ist fast richtig.
Ihr seid auf dem richtigen Weg!
Schaut euch nochmal die Antworten an.
a2 ist richtig.
a3 ist richtig.
a4 ist richtig.
```

Richtige und falsche Antworten werden zurückgemeldet.
Zusätzlich sind die Durchführenden des CAMMP days vor Ort und beantworten den Schülerinnen und Schülerinnen konkret auftretende Fragen, wobei das selbstständige Arbeiten gefördert werden soll.

4.2.7. Erreichung der gesteckten Ziele des CAMMP days

Zu Beginn dieses Kapitels wurden Ziele des CAMMP days formuliert. Nun wird zusammenfassend beschrieben, wodurch diese erreicht werden sollen:

Das *relevante Thema* ist hier die Privatsphäre in sozialen Online Netzwerken. Dieses wird bearbeitet durch Modellierung. Der *Modellierungskreislauf* wird mit den Schülerinnen und Schülern zum Teil gemeinsam, zum Teil nahezu eigenständig durchlaufen und durch eine einführende Präsentation und konkrete Auslegung auf die Problemstellung näher dargestellt. Die *Problemfrage* wird Stück für Stück durch die Plenumsgespräche und die Arbeitsphasen gelöst und abschließend gemeinsam beantwortet. Im Abschlussgepräch erfolgt eine *Reflexion über die Verantwortung in sozialen Netzwerken*. Viele Unterstützungsmaßnahmen fördern die *eigenständige Erarbeitung*. Durch Nutzung vieler Vorteile von digitalen Werkzeugen in Form der *Software MATLAB* erfolgt die Lösung des Problems. Das Material knüpft *an das*

Vorwissen der Schülerinnen und Schüler durch Beachtung des Kernlehrplans bei der Konzeption des CAMMP days und durch die genannten Unterstützungsmaßnahmen an. Den Schülerinnen und Schülern wird in ihrer *Heterogenität* in *inhaltlichen Interessen* durch (zum Teil gegebene) Offenheit des Lösungsweges und im *Lerntempo* durch Zusatzaufgaben begegnet. Das *Schwierigkeitsniveau* der Aufgaben wird durch Unterstützungsmaßnahmen unterschiedlichen *Leistungsniveaus* zugänglich gemacht. Die *Vermittlung der (neuen) mathematischen Inhalte* erfolgt durch geleitete Erkundungen. Diese genannten Maßnahmen dienen dazu, dass die Schülerinnen und Schüler *Spaß* im CAMMP day haben.

Die Vermittlung von Mathematik ist hier also besonders durch Entdecken und Ausprobieren in Bezug auf eine konkrete Fragestellung gekennzeichnet.

5. ...und so verlief die erste Durchführung - Evaluation

In diesem Kapitel wird die Evaluation der ersten Durchführung des CAMMP days vorgenommen.

5.1. Organisatorische Anmerkungen

Am 17. Juni 2016 fand die erste Durchführung des CAMMP days „Soziale Netzwerke“ statt. Ein Mathematikleistungskurs des Gymnasiums am Wirteltor aus Düren führte den CAMMP day unter Begleitung der Lehrerin, Frau Sigrid Schneider-Haag, und der Referendarin, Frau Nina Wolf, durch. Zusätzlich nahmen zwei Praktikantinnen des Instituts an dem CAMMP day teil. Der Modellierungsvortrag wurde an diesem Tag von Thomas Camminady gehalten. Die Arbeit mit dem konzipierten Material wurde hauptsächlich von mir geleitet, wobei Benjamin Terschanski und Sarah Schönbrodt unterstützend tätig waren. Der zeitliche Rahmen des CAMMP days wurde zunächst auf 6.5 Zeitstunden von 8.30 Uhr bis 15.00 Uhr festgelegt. Letztendlich wurde nur ein Zeitraum von ca. 6 Stunden genutzt, da es aus didaktischer Sicht aufgrund der nachlassenden Arbeitskraft der Schülerinnen und Schüler nicht mehr sinnvoll war, das vierte, zusätzliche Arbeitsblatt bearbeiten zu lassen.

5.2. Beschreibung der Schülerinnen und Schüler

Die Beschreibung der Schülerinnen und Schüler findet anhand der Evaluation der Lernenden (s. Auswertung im Anhang) und anhand eines Gesprächs mit den Lehrkräften statt:
Der Leistungskurs der Qualifikationsphase 1 besteht aus 15 Schülerinnen und Schüler, wobei nur 13 davon am CAMMP day teilnahmen. Unter Einbezug der beiden Praktikantinnen bestand die Gruppe der Lernenden aus zehn männlichen und fünf weiblichen Personen. Neun Lernende belegen einen Leistungskurs in den Fächern Mathematik, Informatik und Physik und drei Schülerinnen und Schüler sogar zwei. Bezogen auf den Kernlehrplan waren, nach Auskunft der Lehrkräfte, den Leistungskursschülerinnen und -schülern lediglich Hypothesentests unbekannt.

5.3. Leitende Gesichtspunkte zur Evaluation und Evaluationsmethoden

Unter Berücksichtigung der Ziele des CAMMP days wurden leitende Gesichtspunkte zur Evaluation formuliert. Diese sind in den Überschriften der folgenden Abschnitte zu finden. Die Evaluation fand zum einen durch einen Evaluatuionsbogen (s. Anhang) statt. Er wurde von den Schülerinnen und Schüler, den Praktikantinnen aber auch von den Lehrkräften ausgefüllt. Zum anderen wird durch auf Beobachtungen beruhende Gespräche mit den Lehrkräften und mit Thomas Camminady, Sarah Schönbrodt und Benjamin Terschanski evaluiert. Kommen wir zur konkreten Evaluation:

5.4. Evaluation

Die Anzahlen der Personen, die einen entsprechenden Aspekt evaluiert haben, schwankt. Demnach werden hier absolute Zahlen und die relativen Anteile angegeben.

*Wurde der **Modellierungsprozess** durchlaufen? Konnte dies im zeitlichen Rahmen geschehen? Haben die Schülerinnen und Schüler mathematische Modellierung besser begreifen können?*
Im tatsächlichen zeitlichen Rahmen konnte der Modellierungsprozess mit den Schülerinnen und Schülern durchlaufen werden. In der Abschlussbesprechung fand sogar eine Diskussion über Verbesserungsmöglichkeiten statt. Betrachtet man die Auswertung der Evaluation, so kann festgestellt werden, dass nur eine Person (ca. 6.67%) angab,

mathematisches Modellieren nicht besser verstanden zu haben, wohingegen acht (ca. $53,3\%$) angaben, Modellieren teilweise besser verstanden zu haben. Sechs (40%) Personen gaben sogar an, das Modellieren besser verstanden zu haben. Ähnlich sieht es bei der Bewertung des Hilfepotentials des Modellierungsvortrags aus: Eine Person (ca. 6.67%) fand ihn weniger hilfreich, sechs Personen (40%) teilweise hilfreich und acht Personen (ca. $53,3\%$) hilfreich. Eine Person gab zudem im Kommentar an, Modellierung besser verstanden zu haben.

*Konnte das **gestellte Problem** von den Schülerinnen und Schülern gelöst werden?*
Bezogen auf die Beobachtungen im CAMMP day waren die Schülerinnen und Schüler fähig, die Problemstellung in der Abschlussbesprechung zu beantworten. Eine Hürde stellte jedoch die Deutung von Cohen's Kappa dar. Hier waren die Lernenden zunächst nicht fähig, auf die Trefferwahrscheinlichkeit der Heuristik zu schließen. Eine gemeinsame, durch die Durchführenden unterstützte Diskussion und Umrechnung waren von Nöten. Konsistent mit der ersten Einschätzung sind die Angaben der Teilnehmenden: Acht Evaluierende (50%) gaben an, dass die Schülerinnen und Schüler das Problem teilweise gelöst haben, und ebenfalls acht Evaluierende (50%) gaben an, dass die Schülerinnen und Schüler das Problem gelöst haben.

*Konnten die Schülerinnen und Schüler **eigenständig** arbeiten? Inwieweit war Unterstützung durch die Durchführenden notwendig?*
Der Aspekt der Eigenständigkeit wurde von den Teilnehmenden des CAMMP days sehr unterschiedlich bewertet: Eine Person (6.25%) gab an, dass gar keine Eigenständigkeit vorhanden war, zwei Personen (12.5%), dass kaum Eigenständigkeit vorhanden war, fünf Personen (31.25%), dass teilweise Eigenständigkeit möglich war, und acht Personen (50%), dass Eigenständigkeit möglich war. Der größte Teil empfand somit, dass die Schülerinnen und Schüler eigenständig arbeiten konnten. Bei der Durchführung des CAMMP days fiel auf, dass sich eine allein arbeitende Schülerin sehr schwer in der Bearbeitung der Materialien tat. Sie wurde von Durchführenden unterstützt. Zudem waren an einigen Stellen Unterstützungsmaßnahmen durch Hilfekarten notwendig, wobei die Arbeit mit diesen eigenständig erfolgte. Die Förderung des eigenständigen Arbeitens durch die Betreuer des CAMMP days sah eine Person (ca. 8.33%) als nicht vorhanden an. Fünf Personen (ca. 41.67%) sahen dies jedoch anders und gaben an, dass es teilweise der Fall war, und sechs Personen (50%) gaben an, dass eine Förderung der Selbstständigkeit vorhanden war.

*Inwieweit begegneten die Materialien der **Heterogenität** der Schülerinnen und Schüler in*

- *ihren Interessen, bezogen auf den Lösungsweg?*
  Den Schülerinnen und Schülern wurden an den Stellen, an denen sie sich separate Arbeitsblätter zu den Heuristiken holen sollten, eine bestimmte Anzahl von jedem Arbeitsblatt angeboten, sodass jede Heuristik bearbeitet wurde. An den Stellen, an denen sie die Auswahl einer Regel treffen sollten, ohne ein zusätzliches Blatt bearbeiten zu müssen, wurden die Heuristiken ohne äußeren Einfluss so gewählt, dass hier ebenfalls jede Heuristik behandelt wurde. Diesen Aspekt evaluierten 15 Personen. Eine Person (ca. 6.67%) gab an, dass der Lösungsweg nicht ihren Interessen entsprach, vier Personen (ca. 26.67%) empfanden, dass der Lösungsweg teilweise nicht ihren Interessen entsprach, neun Personen (60%) konnten den Lösungsweg teilweise nach ihren Interessen wählen und eine Person (ca. 6.67%) komplett nach ihren Interessen.

- *ihrem Lerntempo, bezogen auf Zusatzaufgaben?*
  Das „Basisprogramm“ (Material ohne Zusatzaufgaben) konnte von allen Schülerinnen und Schüler bearbeitet werden. Einzelne Gruppen testeten mehr als eine Regel. Eine Gruppe bearbeitete das Material zur Trefferquote und zur Genauigkeit. Das vierte Arbeitsblatt wurde nicht ausgehändigt.

- *ihrem Leistungsniveau, bezogen auf die Unterstützungsmaßnahmen bei schwierigeren Aufgaben?*
  Bei der Durchführung des CAMMP days wurde die Hilfekarte 1 an sechs Gruppen, die Hilfekarte 2 an eine Gruppe und die Hilfekarte 4 an zwei Gruppen ausgehändigt. Die Unterstützungsmaßnahmen wurden also vor allem bei der Herleitung der Trefferwahrscheinlichkeit durch den Zufall und bei der Behandlung von Cohen's Kappa genutzt. Zu den Hilfekarten nahmen 14 Teilnehmende Stellung: Zwei (ca. 14.29%) gaben an, dass sie eher nicht hilfreich waren, sieben (50%) schrieben, dass sie teilweise halfen und fünf (ca. 35.17%) gaben an, dass sie hilfreich waren.
  Die Beobachtungen ergaben ebenfalls, dass die Unterstützungsmaßnahmen an den antizipierten Stellen angemessen waren und keine weiteren kritischen Stellen sichtbar waren.

*Inwieweit knüpft der CAMMP day am **Vorwissen** der Schülerinnen und Schüler an? Waren die vorausgesetzten Inhalte realistisch?*

Zum Schwierigkeitsgrad der Aufgaben äußerten sich 16 Teilnehmende: Vier Personen (25%) empfanden die Aufgaben als eher einfach, zehn Personen (62.5%) empfanden sie als eher nicht zu einfach und zwei (12.5%) als nicht zu einfach. Die umgekehrte Aussage wurde ebenfalls beurteilt: Eine (6.25%) Person empfand die Aufgaben als zu schwierig, zwei (12.5%) als eher zu schwierig, zehn (62.5%) als eher nicht zu schwierig und drei (18.75%) als nicht zu schwierig. Die Schwierigkeit der Aufgaben war somit für die meisten angemessen.

Als weiterer Punkt wurde die Verständlichkeit der Anleitungen zu den Aufgaben evaluiert, zu der sich elf Personen äußerten: Eine Person (ca. 9.09%) gab an, dass die Aufgaben eher nicht zufriedenstellend gestellt und eher nicht verständlich waren. Sechs Personen (ca. 54.55%) empfanden, dass das eher der Fall war, und fünf Personen (ca. 45.45%) empfanden, dass die Anleitungen zu den Aufgaben verständlich waren.

Auch die Lehrkräfte empfanden, dass das Vorausgesetzte durchaus realistisch war.

*Inwieweit wurde das **Bewusstsein** der Schülerinnen und Schüler bezogen auf ihre **Verantwortung in sozialen Online Netzwerken** geschult?*

Beobachtbar war, dass in der Abschlussdiskussion eine Reflexion stattfand. Die Schülerinnen und Schüler stellten die Bedeutung der Ergebnisse dar und erschlossen sich ihre eigenen Rollen in Netzwerken. Diese Beobachtung ist konsistent mit den Kommentaren in der Evaluation, in der Kommentare wie „soziale Netzwerke sind unheimlich“ oder „Ich habe zu viel gelernt.“ zu finden sind. Benannt wurde auch die Thematik und im Kommentar der Lehrerinnen, dass der CAMMP day eine „Sensibilisierung“ und eine „effektive Medienkritik“ ermöglicht.

*Konnte **MATLAB** durch die Anleitungen genutzt werden? Welche Probleme zeigten sich?*

Der Umgang mit MATLAB verlief weitestgehend unproblematisch. Bei einigen Gruppen kam es zur Verwechslung von „Run“ und „Run Section“, die direkt behoben werden konnte. Zudem nutzten die Schülerinnen und Schüler MATLAB nicht als Taschenrechner, sondern arbeiteten mit den von ihnen mitgebrachten. Zum Umgang mit MATLAB gaben 15 Personen ihre Meinung an: Eine Person (ca. 6.67%) fand den Umgang sehr schwer, drei (20%) eher schwer, fünf (ca. 33.33%) eher weniger schwer und sechs (40%) nicht schwer. Dabei gaben zwei Personen (ca. 28.57%) an,

dass das Hilfesystem von MATLAB hilfreich war, und fünf (ca. 71.43%), dass es eher hilfreich war. Der Einsatz technischer Hilfsmittel wurde in den Kommentaren als sehr motivierend dargestellt.

*War das Thema für die Lernenden **interessant**?*
Diese Frage kann aufgrund der Beobachtungen mit „ja“ beantwortet werden. Die Schülerinnen und Schüler stellten viele weitergehende Fragen zur Thematik und zeigten in den Arbeitsphasen Durchhaltevermögen und Interesse. Gerade die Entwicklung von Ideen für neue Heuristiken unterstützt diese Ansicht. Zudem fragten die Schülerinnen und Schüler ohne Aufforderung nach Zusatzaufgaben.
Ein Interesse an Naturwissenschaften und Technik hat der CAMMP day bei vielen gesteigert: Von elf Personen gaben drei (ca. 27.27%) an, dass ihr Interesse eher nicht gesteigert wurde, fünf (ca. 45.45%) gaben eher eine Steigerung des Interesses an und drei (ca. 27.27%) eine Steigerung des Interesses.

*Wurden die neuen **mathematischen Inhalte** verständlich vermittelt? Haben die Schülerinnen und Schüler die Inhalte verstanden?*
Zu der ersten dieser beiden Fragen äußerten sich elf Personen. Eine Person (ca. 9.09%) gab an, dass die Inhalte von den Dozenten nicht verständlich vermittelt wurden, zwei (ca. 18.18%) gaben an, dass sie eher verständlich vermittelt wurden und acht (ca. 72.72%), dass sie verständlich vermittelt wurden.
Die subjektive Einschätzung der Schülerinnen und Schüler, ob sie die Inhalte verstanden haben (2. Frage), sah folgendermaßen aus: elf von 16 Personen (68.75%) gaben an, dass sie die Inhalte teils verstanden haben und fünf von 16 (31.25%), dass sie die Inhalte verstanden haben.
Auch in den Besprechungen entstand der Eindruck, dass die Schülerinnen und Schüler die Inhalte verstanden haben. Die Äußerungen in den Diskussionsphasen zeigten, dass die Schülerinnen und Schüler bei der Entwicklung und Lösung der Problemstellung mitarbeiten konnten, indem sie beispielsweise die Heuristiken antizipieren und einschätzen konnten.

*Hat der CAMMP day den Schülerinnen und Schülern **Spaß** gemacht?*
Hierzu gab eine Person von 15 Personen (ca. 6.67%) an, eher keinen Spaß gehabt zu haben. Zehn Personen (ca. 66.67%) hatten teilweise Spaß und vier (ca. 26.67%) hatten Spaß. Der Spaß am CAMMP day zeigte sich in der Beobachtung vor allem an

der Kreativität beim Ausprobieren verschiedener Heuristiken und im Arbeitseifer der meisten Schülerinnen und Schüler.

Hat der CAMMP day das Interesse der Schülerinnen und Schüler an einem ***Studium oder einer Ausbildung im Bereich Technik oder Naturwissenschaften*** *gefördert?*
Zu diesem Punkt wurde die Frage gestellt, ob interessante Berufe vorgestellt wurden. Eine Person von elf (ca. 9.09%) gab an, dass das nicht der Fall war. Sieben Personen (ca. 63.64%) fanden, dass dies teils der Fall war und drei Personen (ca. 27.27%) fanden, dass interessante Berufe vorgestellt wurden.
Es kann sich eine Person von zehn Personen (10%) kein Studium oder eine Ausbildung in den genannten Bereichen vorstellen und eine weitere (10%) eher nicht. Fünf Personen (50%) können es sich teilweise vorstellen und drei (30%) können es sich vorstellen.

Wie viele Personen würden einen solchen Workshop ***erneut besuchen****?*
Hier gaben vier von elf Personen (ca. 36.36%) an, dies eher nicht noch einmal tun zu wollen. Vier (ca. 36.36%) gaben an, dass sie dies teilweise tun möchten und weitere drei (ca. 27.27%), dass sie dies tun möchten.

Insgesamt ist festzuhalten...
... , dass die erste Durchführung des CAMMP days sehr erfolgreich war. Der CAMMP day ist nun durchführbar. Der Modellierungprozess wurde durchlaufen und von den meisten Lernenden besser begriffen, was unter anderem am Modellierungsvortrag lag. Die Lösung des Problems gelang den Schülerinnen und Schülern (teilweise mit Unterstützung). Die Mehrheit der Lernenden arbeitete eigenständig, was auch durch die Betreuerinnen und Betreuer gefördert wurde. Der Heterogenität konnte man bei den meisten Schülerinnen und Schülern in ihren inhaltlichen Interessen, ihrem Lerntempo und ihrem Leistungsniveau durch Unterstützungsmaßnahmen und zusätzlichen Aufgaben begegnen. Der Schwierigkeitsgrad der Aufgaben wurde zusammenfassend als angemessen empfunden, wobei auch die Aufgabenstellung für die meisten zufriedenstellend und gut verständlich war. Die Anknüpfung an das Vorwissen ist somit gelungen. Das ethische Ziel des CAMMP days wurde erreicht. Die Reflexion über die Verantwortung in sozialen Netzwerken fand statt. Das technische Hilfsmittel MATLAB konnte von den meisten Schülerinnen und Schülern ohne Probleme genutzt werden, was auch durch

das Hilfesystem unterstützt wurde. Die Thematik war für die meisten Lernenden interessant und motivierend, was sich auch in einer Steigerung des Interesses an Technik und Naturwissenschaften ausdrückt. Die Vermittlung der mathematischen Inhalte ist ebenfalls gelungen, da ein Großteil die Vermittlung verständlich fanden und angaben, die Inhalte auch tatsächlich verstanden zu haben. Der Spaß im CAMMP day blieb nicht aus. Die meisten Teilnehmer hatten Spaß im CAMMP day und gaben dies auch an. Das Interesse der Schülerinnen und Schüler an einem Studium oder einer Ausbildung im Bereich Technik oder Naturwissenschaften und an einem weiteren Workshop besteht bei den meisten von ihnen.
Abschließend ist zu betonen, dass die grundlegende didaktische Idee, vor allem durch Ausprobieren, in den Kommentaren positiv hervorgehoben wurde. Des Weiteren ist einiges an Lob zu lesen (s. Anhang). Dieses Lob wurde auch von Frau Sigrid Schneider-Haag ausgesprochen. Ihr Bericht soll bald auf der Homepage des Gymnasiums am Wirteltor [10] zu lesen sein.

6. ...und das kann verbessert werden - Ausblick zur Weiterentwicklung des CAMMP days

Bei der ersten Durchführung sind neben vielen positiven Aspekten auch Änderungsmöglichkeiten aufgefallen. Diese sollen nun in Form eines Ausblicks anhand der Evaluation und nun vorhandener Erfahrungswerte dargestellt werden.

6.1. Organisatorische Aspekte

- Sehr hilfreich bei den Besprechungen und allgemein bei der Arbeit mit den Schülerinnen und Schülern war ein von den Lehrkräften erstellter **Sitzplan**. Die Arbeit wird so persönlicher. Dies ist auch für weitere CAMMP days zu empfehlen.

- Der ursprünglich angedachte **zeitliche Ablauf** ist im Anhang dargestellt. Der tatsächliche wird nun dargelegt:
  Der Einstieg in die Problemstellung durch den zum CAMMP day gehörenden Vortrag dauerte 20 Minuten. Die anschließende erste Arbeitsphase war 70 Minuten lang mit anschließender 20-minütiger Besprechung. Die zweite Arbeitsphase fand innerhalb von 45 Minuten statt, wobei sich eine Gruppe nach 35 Minuten das zusätzliche Arbeitsmaterial holte. Die dazugehörige Besprechung dauerte 6 Minuten. Das dritte Arbeitsblatt wurde in 40 Minuten bearbeitet. Die Besprechung

mit anschließender Abschlussdiskussion und Evaluation brauchte 30 Minuten. Die erste Besprechung war sehr lang. Eine mögliche Abkürzung kann durch vorheriges Austeilen der Besprechungsfolien erreicht werden oder sie kann durch zwei kürzere Besprechungen ersetzt werden.

Diese Erfahrungen sollten bei den nächsten Durchführungen im organisatorischen Bereich beachtet werden.

6.2. Arbeitsmaterial

- Zu Beginn scheint es in der **Präsentation zur Problemstellung** hilfreich, den **Begriff Heuristik mit dem Begriff Vorhersageregel häufiger zu verknüpfen**. Es entstanden zwar keine konkreten Verständnisprobleme bei der betrachteten Lerngruppe, dennoch kann so ein besseres Verständnis gefördert werden.

- Bei der Einführung der parameterabhängigen Heuristiken mit dem **Arbeitsblatt 1** wurde in einem Gespräch mit den Lehrkräften deutlich, dass der Begriff „Zahlen“ zu Verwirrung führen kann. Hier sollte der **Begriff „Parameter“** durchaus verwendet werden.
  Ein weiterer Aspekt ist bei den Arbeitsaufträgen auf diesem Blatt zu verbessern: Bei den einzelnen Regeln in Schritt 3 war in Teil c) nicht ganz klar, dass die Schülerinnen und Schüler die Summe der richtigen bzw. falschen Vorhersagen betrachten sollten.

- Bei **Arbeitsblatt 2** sollte in Teil a) deutlicher werden, dass man aufgrund der konkreten Daten auf die **Trefferwahrscheinlichkeit des Zufalls** schließt. Der Ursprung der Prozentangaben war einigen nicht klar. Des Weiteren sollte deutlicher werden, was genau in dieser Aufgabe das „Ergebnis“ ist. Viele Lernende gaben eine falsche Zahl in MATLAB ein, obwohl sie die Trefferwahrscheinlichkeit richtig bestimmt hatten.

- Die **Gruppengröße der Teamarbeit** auf **Arbeitsblatt 3** sollte variabel von den Dozenten vorgegeben werden, damit der CAMMP day für jede Lerngruppengröße durchführbar ist.

- In der **Abschlussbesprechung** wurde eine **Hürde, Cohen's Kappa zu deuten**, festgestellt. Zum besseren Verständnis kann für die Schülerinnen und Schü-

ler ein weiterer Arbeitsauftrag vor der Abschlussbesprechung gestellt werden. Die selbstständige Umrechnung kann so erfolgen.

- Der **Code** des MATLAB-Programms kann weiter ergänzt werden. Das **Feedback kann differenzierter** gestaltet werden, sodass die Lernenden zu jedem Schritt sowohl konkretere Fehlermeldungen (z.B. wenn Stichprobengrößen größer als 56284 gewählt werden) als auch konkreteres Feedback erhalten. Das Feedback kann zudem an einigen Stellen **formatiert** werden, damit es komplett im Command Window erscheint. Des Weiteren sollten durch veränderte Abspeicherung **die mit dem Code zusammenhängenden Programme nicht zugänglich** sein. Genutzte Begriffe im Code, wie „params“ oder „range“ könnten zudem übersetzt und näher kommentiert werden. Die Bezeichnungen „v2adj“ und „v3friendsteruser“ sowie „pe“ könnten näher erklärt werden, indem sie in „Freundschaften“ und „Nutzer“ sowie „Trefferwahrscheinlichkeit Zufall“ umbenannt werden.

- Grundsätzlich müssen für den CAMMP day **viele Blätter gedruckt** werden. Im Sinne des Umweltschutzes sollte über eine Reduzierung nachgedacht werden. Beispielsweise könnten Daten in Tabellen in Matlab eingetragen werden oder kurze Arbeitsaufträge, wie auf Arbeitsblatt 3, mittels Beamer visualisiert werden.

6.3. Weiterentwicklungen in der Konzeption

- Schon bei der Entwicklung dieses ersten Entwurfes des CAMMP days wurde über eine mögliche **Visualisierung der Güte der Heuristiken** nachgedacht. Eine Visualisierung des Ergebnisses ist bereits im CAMMP day zum Thema GPS umgesetzt. Möglich wäre hier, einen Graphen zu zeigen, bei dem die Knoten unterschiedlich gefärbt werden. Diese Idee ist auch nach dem CAMMP day diskutiert worden.

- Nach der Mittagspause wurde deutlich, dass die ständige Arbeit mit Arbeitsblättern und Codes nur noch mäßig motivierend ist. Hier kann über einen **methodischen Wechsel** nachgedacht werden. Möglich wäre beispielsweise eine Rechercheaufgabe zur Frage „Was kann ich alles in meinem Facebook-Profil angeben?“ oder „Wie lauten die Nutzungsbedingungen von WhatsApp?“. Ein arbeitsteiliges Vorgehen ist hier zu diskutieren.

- Eine andere Möglichkeit, den Interessen der Schülerinnen und Schülern zu begegnen, wäre, dass sie sich zu Beginn des Workshops **auf eine Regel festlegen** und alle weiteren Schritte mit dieser (soweit es möglich ist) durchlaufen. Ein Vorteil hierbei wäre, dass so der Papierverbrauch reduziert werden könnte. Vor- oder Nachteile eines solchen Vorgehens, bezogen auf das Lernen der Schülerinnen und Schüler, können erst nach einer Durchführung benannt werden.

- Eine weitere Ergänzung in der Konzeption wäre, die **Arbeit mit den Heuristiken zu öffnen**. Damit ist gemeint, dass die Schülerinnen und Schüler die Möglichkeit bekommen, ihre eigenen Heuristiken zu entwickeln. Außerdem könnten andere Merkmale der Nutzer, neben der sexuellen Orientierung, untersucht werden. Über Hürden im Bereich des Programmierens ist hier jedoch zu diskutieren. Dieser Aspekt wurde auch in der Evaluation als Verbesserungsvorschlag angeführt. Konkret kann man die folgenden Vorschläge umsetzen:
  - Im Verlauf der weiteren CAMMP days können die Vorschläge der Schülerinnen und Schüler gesammelt und für die nächsten CAMMP days programmiert werden, damit sie damit für weitere Lerngruppen verfügbar werden.
  - Als Ergänzung können die im Verlaufe dieser Masterarbeit entwickelten Heuristiken für die Schülerinnen und Schüler zur Verfügung gestellt werden. Hier sind bereits weitere Untersuchungsideen verwirklicht.
  - Es kann ein Programm geschrieben werden, das die vorhandenen Heuristiken kombinieren kann. So können die Schülerinnen und Schüler mit „Metaheuristiken“ arbeiten.
  - In kleinen Gruppen und mit Unterstützung der Durchführenden ist es, falls die Durchführenden programmierfähig sind, möglich, direkt im CAMMP day die Ideen der Schülerinnen und Schüler umzusetzen.
  - Mit Hilfe eine „Bausteinsystems“, was noch entwickelt werden muss, kann den Schülerinnen und Schülern das Programmieren erleichtert werden. Gerade die Nutzung von „Wenn-Dann-Aussagen“ ist mit wenigen Programmierbausteinen zu bewerkstelligen.

- Zur Ergänzung und zum besseren Verständnis der Arbeitsweise des Zufalls kann diese als **weitere Heuristik** in den CAMMP day eingebaut und mit den anderen Heuristiken verglichen werden. Grundsätzlich können auch andere Heuristiken, basierend auf anderen Ideen der Untersuchung, eingearbeitet werden.

- Neben diesen Aspekten ist über eine **Eröffnung der Thematik für Schülerinnen und Schüler vor der gymnasialen Oberstufe** nachzudenken. Grundsätzlich kann durch das Auslassen bzw. das weitere Ausarbeiten bestimmter Schritte eine Anknüpfung an ein weniger differenziertes Vorwissen geschehen.

Diese Aspekte dienen der Weiterentwicklung des CAMMP days.

7. Abschließende Worte

Wir sind am Ende der Masterarbeit angekommen. Der CAMMP day zum Thema

Wie sicher ist meine Privatsphäre in sozialen Online Netzwerken? ...und was hat das mit Mathe zu tun?

ist entwickelt, durchgeführt und evaluiert worden. Das Ziel meiner Masterarbeit ist erreicht: Der CAMMP day kann nun für Schülerinnen und Schüler angeboten werden. Die Vorschläge für die fachliche und fachdidaktische Entwicklung können weiter verfolgt werden.

Mit diesen Worten beende ich meine Masterarbeit. Ich freue mich, einen Beitrag zum CAMMP geleistet zu haben, und bin gespannt auf die Entwicklung, die der CAMMP day in Zukunft machen wird.

Abbildungsverzeichnis

1. **Modellierungskreislauf nach Blum und Leiss.** [13, S. 19] 10
2. **Modellierungskreislauf im CAMMP.** 12
3. **Das Netzwerk als ungerichteter Graph.** Eine Veranschaulichung. . 18
4. **Adjazenzmatrix des Datensatzes der ersten 1000 Nutzer.** Blaue Punkte symbolisieren eine Kante. 22
5. Peter und seine Freunde. Peters Freunde 1. Ordnung Peters Freunde 2. Ordnung 24
6. **Säulendiagramm zur Anzahl der Nutzer pro Anzahl der homosexuellen Freunde bis 50.** Die Säule zu 50 homosexuellen Freunden stellt die Anzahl der Nutzer mit 50 oder mehr homosexuellen Freunden dar. 26
7. **Säulendiagramm zur Anzahl der Nutzer pro Anzahl der homosexuellen Freunde 2. Ordnung bis 150.** Die Säule zu 150 homosexuellen Freunden 2. Ordnung stellt die Anzahl der Nutzer mit 150 oder mehr homosexuellen Freunden 2. Ordnung dar. 27
8. **Gespeicherte Daten über Nutzer 100.** 28
9. **Baumdiagramm zur Bestimmung der sexuellen Orientierung mit Hilfe des Zufalls.** 32
10. **Klassifikation durch Zufall mit Überprüfung im Baumdiagramm.** 39
11. **Venn-Diagramm zur Trefferquote und Genauigkeit der Heterosexuellen.** 42
12. **Gradientenverfahren vereinfacht dargestellt.** 51
13. **Mit dem Gradientenverfahren erzeugte Kappawerte für Heuristik 4.** Stichprobengröße: 30000, x-Achse: Anzahl der Iterationen; y-Achse: Kappawert; erstellt von T. Camminady. 52
14. **Entscheidungsmöglichkeiten.** [3, S. 270] 55
15. **Modellierungskreislauf im CAMMP day.** 65

Literatur

[1] Manfred Baum u. a. *Lambacher Schweizer. Mathematik für Gymnasien. Qualifikationsphase Leistungskurs. Nordrhein-Westfalen.* Klett Verlag: Leipzig; Stuttgart, 2011.

[2] Dieter Bestle. *Analyse und Optimierung von Mehrkörpersystemen. Grundlagen und rechnergestützte Methoden.* Springer Verlag: Berlin; Heidelberg, 1994.

[3] Erhard Cramer und Udo Kamps. *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik. Ein Skript für Studierende der Informatik, der Ingenieur- und Wirtschaftswissenschaften.* 3., überarb. Aufl., Springer Verlag: Berlin; Heidelberg, 2014.

[4] Duden. *Heuristik, die.* URL: `http://www.duden.de/rechtschreibung/Heuristik(26.5.2016)`.

[5] Duden. *Social Network, das.* URL: `http://www.duden.de/rechtschreibung/Social_Network(3.6.2016)`.

[6] Rita Borromeo Ferri. *Wege zur Innenwelt des mathematischen Modellierens. Kognitive Analyse zu Modellierungsprozessen im Mathematikunterricht.* Vieweg+Teubner und Springer Fachmedien: Wiesbaden, 2011.

[7] *Friendster: Archive.* URL: `https://archive.org/details/archive-team-friendster(2.4.2016)`.

[8] *Friendster: Website.* URL: `www.friendster.com(2.6.2016)`.

[9] David Garcia, Emre Sarigol und Frank Schweitzer. *Online Privacy as a Collective Phenomenon.* 2014. URL: `http://arxiv.org/pdf/1409.6197v1.pdf(6.6.2016)`.

[10] *Gymnasium am Wirteltor, Düren: Homepage.* URL: `http://www.wirteltor-gymnasium.de/`.

[11] Frank Heinrich, Regina Bruder und Christina Bauer. *Handbuch der Mathematikdidaktik. Problemlösen lernen.* Hrsg. von Regina Bruder u. a. Springer Verlag: Berlin; Heidelberg, 2015.

[12] Wilfried Herget und Jürgen Maaß. *Mathematik nutzen mit Verantwortung. In: mathematik lehren.* 194. 2016, S. 2–6.

[13] Gerd Hinrichs. *Modellierung im Mathematikunterricht.* Spektum Akademischer Verlag: Heidelberg, 2008.

[14] Gabriele Kaiser u. a. *Handbuch der Mathematikdidaktik. Anwendungen und Modellieren.* Hrsg. von Regina Bruder u. a. Springer Verlag: Berlin; Heidelberg, 2015.

[15] Jürgen Klanert. *Das Online-Verhalten der User wird immer komplexer.* 2014. URL: http://www.focus.de/digital/experten/klanert/wandel-durch-das-internet-das-online-verhalten-der-user-wird-immer-komplexer_id_4078023.html(2.6.2016).

[16] Aloys Krieg. *Analysis I. Skript zur Vorlesung.* Lehrstuhl A für Mathematik, RWTH Aachen, 2007.

[17] Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland. *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss.* Wolters Kluver: München, 2004.

[18] Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland. *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife.* Wolters Kluver: München, 2015.

[19] Rafael Marti und Gerhard Reinelt. *The Linear Ordering Problem. Exact and Heuristic Methods in Combinatorial Optimization.* Springer Verlag: Berlin; Heidelberg, 2011.

[20] *MathWorks: Website.* URL: http://de.mathworks.com/products/matlab/(4.6.2016).

[21] Robert McMillan. *Not on a social network? You 've still got a privacy problem.* 2014. URL: http://www.wired.com/2014/10/privacy-friendster/(2.6.2016).

[22] Robert McMillan. *The Friendster Autopsy: How a Social Network Dies.* 2013. URL: http://www.wired.com/2013/02/friendster-autopsy/(2.6.2016).

[23] Bundesverband Informationswirtschaft Telekommunikation und neue Medien e. V (BITKOM). *Soziale Netzwerke. Eine repräsentative Untersuchung zur Nutzung sozialer Netzwerke im Internet.* 2011. URL: https://www.bitkom.org/Publikationen/2011/Studie/Studie-Soziale-Netzwerke/BITKOM-Publikation-Soziale-Netzwerke.pdf(3.6.2016).

[24] Katharina Morik. *Maschinelles Lernen.* Lehrstuhl VIII, Informatik, Universität Dortmund, 1993. URL: https://eldorado.tu-dortmund.de/bitstream/2003/2574/1/morik%5C_93c.pdf(26.5.2016).

[25] Thomas Müller-Gronbach, Erich Novak und Klaus Ritter. *Monte Carlo-Algorithmen.* Springer Verlag: Berlin; Heidelberg, 2012.

[26] Harald Nahrstedt. *Algorithmen für Ingenieure. Technische Realisierung mit Excel und VBA.* 2. überarb. Aufl. Vieweg+Teubner und Springer Fachmedien: Wiesbaden, 2012.

[27] Wilhelm Plesken. *Lineare Algebra I. Skript zur Vorlesung.* Lehrstuhl B für Mathematik, RWTH Aachen, 2008.

[28] Andreas Ross. *Oberster Gerichtshof legalisiert Homoehe in allen Bundesstaaten.* 2015. URL: http://www.faz.net/aktuell/politik/ausland/amerika/vereinigte-staaten-oberster-gerichtshof-legalisiert-homoehe-in-allen-bundesstaaten-13670061.html(2.6.2016).

[29] Jürgen Rost. *Lehrbuch Testtheorie - Testkonstruktion.* Huber Verlag: Bern; Göttingen; Toronto; Seattle, 2004.

[30] Thomas Runkler. *Data Mining: Modelle und Algorithmen intelligenter Datenanalyse.* 2. Aufl. Springer Fachmedien Verlag: Wiesbaden, 2015.

[31] Wolfgang Saxon. *Jacob Cohen, 74, Psychologist And Pioneer in Statistical Studies.* 1998. URL: http://www.nytimes.com/1998/02/07/nyregion/jacob-cohen-74-psychologist-and-pioneer-in-statistical-studies.html(2.6.2016).

[32] Sarah Schönbrodt. *Didaktisch-methodische Ausarbeitung eines Lernmoduls zum Thema Google.* 2015. URL: http://www.cammp.rwth-aachen.de/wp-content/uploads/2013/03/BA%5C_thesis%5C_schoenbrodt.pdf(26.5.2016).

[33] Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. *Kernlehrplan für das Gymnasium - Sekundarstufe I (G8) in Nordrhein-Westfalen. Mathematik.* Ritterbach Verlag, Frechen, 2007.

[34] Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. *Kernlehrplan für das Gymnasium - Sekundarstufe II Gymnasium/Gesamtschule in Nordrhein-Westfalen. Mathematik.* Ritterbach Verlag, Frechen, 2013.

[35] CAMMP Team. *CAMMP: Agenda 2015.* 2015.

[36] CAMMP Team. *CAMMP: Website.* URL: http://www.cammp.rwth-aachen.de/(3.6.2016).

[37] Peter Tittmann. *Graphentheorie. Eine anwendungsorientierte Einführung.* 2. akt. Aufl. Carl Hanser Verlag: München., 2011.

[38] Katrin Vorhölter. *Sinn im Mathematikunterricht. Zur Rolle von mathematischen Modellierungsaufgaben bei der Sinnkonstruktion von Schülerinnen und Schülern. Studien zur Bildungsgangforschung. Band 27.* Barabara Budrich Verlag: Warschau, 2009.

[39] *WhatsApp: Nutzungsbedingungen.* URL: `https://www.whatsapp.com/legal/#Privacy(6.6.2016)`.

[40] Heinrich Winter. *Mathematikunterricht und Allgemeinbildung. In: tteilungen der Gesellschaft für Didaktik der Mathematik.* Hrsg. von Mitteilungen der Gesellschaft für Didaktik der Mathematik. 61. 1995, S. 37–46. URL: `http://geosoft.ch/fachdidaktik/fd3/Winter.pdf(6.6.2016)`.

A. Anhang

A.1. Methodisches Konzept und Ablaufübersicht

CAMMP Day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Methodisches Konzept CAMMP day Soziale Netzwerke

für die Dozenten

Material

Arbeitsblätter für die Schülerinnen und Schüler:
- ☐ Ablauf.pdf
- ☐ AB 1_Heuristiken Einstieg.pdf
- ☐ AB 1_Heuristiken Einstieg_Antwortblatt.pdf
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 1.pdf
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 2.pdf
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 3.pdf
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 4.pdf
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 5.pdf
- ☐ AB 2_Heuristiken Güte bewerten.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Antwortblatt.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz_Antwortblatt.pdf
- ☐ AB 3_Kappa verbessern.pdf
- ☐ AB 3_Kappa verbessern_Antwortblatt.pdf
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 1.pdf
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 2.pdf
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 4.pdf
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 5.pdf
- ☐ AB 4_Zusatz Hypothesentests.pdf
- ☐ AB 4_Zusatz Hypothesentests_Antwortblatt.pdf
- ☐ Evaluation.pdf

Hilfekarten:
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 1.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 2.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 3.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 4.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 5.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz_Hilfekarte 6.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz_Hilfekarte 7.pdf

Folien zur Besprechung:
- ☐ AB 0_Heuristiken Übersicht.pdf
- ☐ AB 1_Heuristiken Einstieg _Besprechung.pdf
- ☐ AB 2_Heuristiken Güte bewerten _Besprechung.pdf
- ☐ AB 3_Kappa verbessern _Besprechung.pdf
- ☐ AB 4_Zusatz Hypothesentests _Besprechung.pdf

Präsentationen:
- ☐ Begrüßung aus Bestand
- ☐ Präsentation Eröffnung.pdf
- ☐ Präsentation Abschluss.pdf
- ☐ Werbung aus Bestand

Hinweise zu den Präsentationen:
- ☐ Präsentation Eröffnung Notizen.pdf
- ☐ Präsentation Abschluss Notizen.pdf

Dateien:
- ☐ Ordner in SVN modules -> -> code

Sonstige Materialien:
- ☐ Folienstifte
- ☐ saubere Tafel
- ☐ OHP

Lösungen für Dozenten:
- ☐ AB 1_Heuristiken Einstieg _Lösungen.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Lösungen.pdf
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz _Lösungen.pdf
- ☐ AB 3_Kappa verbessern _Lösungen.pdf
- ☐ AB 4_Zusatz Hypothesentests _Lösungen.pdf

Modellierungsvortrag:
wird von den Doktoranten organisiert

Vorbereitungen

vor 8:30 Vorbereitungen
- ☐ Datei sozialenetzwerke auf Laptops öffnen und ersten Abschnitt „Lade Daten“ durch *Run Section* ausführen Command Window sollte rechts von oben bis unten offen sein
- ☐ Ablauf.pdf auf jedem Schülerplatz auslegen
- ☐ AB 1_Heuristiken Einstieg.pdf bereitlegen
- ☐ AB 1_Heuristiken Einstieg_Antwortblatt.pdf bereitlegen
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 1.pdf auf separatem Tisch bereitlegen
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 2.pdf auf separatem Tisch bereitlegen
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 3.pdf auf separatem Tisch bereitlegen
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 4.pdf auf separatem Tisch bereitlegen
- ☐ AB 1_Heuristiken Einstieg_Schritt 3 Regel 5.pdf auf separatem Tisch bereitlegen
- ☐ Präsentation Begrüßung auf Hiwi-Laptop öffnen
- ☐ Präsentation Eröffnung.pdf auf Hiwi-Laptop öffnen
- ☐ Modellierungsvortrag auf Hiwi-Laptop öffnen
- ☐ OHP vorbereiten
- ☐ AB 1_Heuristiken Einstieg_Besprechung.pdf auf OHP bereitlegen (bis Schritt 1 aufdecken)
- ☐ www.friendster.com öffnen

Ablauf

8:30 Begrüßung
- ☐ Präsentation Begrüßung.pdf halten

Modellierungsvortrag
- ☐ wird gehalten

8:45 Einstieg in das Projekt
- ☐ Präsentation Eröffnung.pdf halten mit Präsentation Eröffnung Notizen.pdf
- ☐ Einschätzungen der SuS auf AB 1_Heuristiken Einstieg_Besprechung.pdf festhalten

9:15 Partnerarbeit (Teil I)
- ☐ AB 1_Heuristiken Einstieg.pdf austeilen
- ☐ AB 1_Heuristiken Einstieg_Antwortblatt.pdf austeilen
- ☐ Hinweis auf Arbeitsblätter vorne
- ☐ AB 1_Heuristiken Einstieg.pdf bearbeiten

10:30
- ☐ Besprechung mit AB 1_Heuristiken Einstieg_Besprechung.pdf mit AB 1_Heuristiken Einstieg_Lösungen.pdf Antworten (stichpunktartig) auf Folie notieren

10:45 Pause
- ☐ unbenutzte Arbeitsblätter von separatem Tisch einsammeln
- ☐ AB 2_Heuristiken Güte bewerten.pdf austeilen
- ☐ AB 2_Heuristiken Güte bewerten_Antwortblatt.pdf austeilen
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 1.pdf auf separatem Tisch bereitlegen
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 2.pdf auf separatem Tisch bereitlegen
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 3.pdf auf separatem Tisch bereitlegen
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 4.pdf auf separatem Tisch bereitlegen
- ☐ AB 2_Heuristiken Güte bewerten_Hilfekarte 5.pdf auf separatem Tisch bereitlegen

10:45 immer noch Pause
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz.pdf auf separatem Tisch bereitlegen
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz_Antwortblatt.pdf auf separatem Tisch bereitlegen
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz_Hilfekarte 6.pdf auf separatem Tisch bereitlegen
- ☐ AB 2_Heuristiken Güte bewerten_Zusatz_Hilfekarte 7.pdf auf separatem Tisch bereitlegen

11:05 Partnerarbeit (Teil II)
- ☐ AB 2_Heuristiken Güte bewerten.pdf bearbeiten
- ☐ Hinweis auf Hilfekarten und Zusatzblatt vorne
- ☐ Zusatzblatt bei Bedarf bearbeiten lassen und individuell mit Dozentenhilfe kontrollieren mit AB 2_Heuristiken Güte bewerten_Zusatz_Lösungen.pdf

11:45
- ☐ Besprechung mit AB 2_Heuristiken Güte bewerten_Besprechung.pdf mit AB 2_Heuristiken Güte bewerten_Lösungen.pdf Antworten (stichpunktartig) auf Folie festhalten Tafel für Erläuterungen nutzen

12:10 Mittagspause
- ☐ unbenutzte Arbeitsblätter von separatem Tisch einsammeln
- ☐ AB 3_Kappa verbessern.pdf austeilen
- ☐ AB 3_Kappa verbessern_Antwortblatt.pdf austeilen
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 1.pdf auf separatem Tisch bereitlegen
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 2.pdf auf separatem Tisch bereitlegen
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 4.pdf auf separatem Tisch bereitlegen
- ☐ AB 3_Kappa verbessern_Schritt 5 Regel 5.pdf auf separatem Tisch bereitlegen
- ☐ AB 4_Zusatz Hypothesentests.pdf auf separatem Tisch bereitlegen
- ☐ AB 4_Zusatz Hypothesentests_Antwortblatt.pdf auf separatem Tisch bereitlegen

13:10 Partnerarbeit (Teil III)
- ☐ AB 3_Kappa verbessern.pdf bearbeiten
- ☐ Hinweis auf Arbeitsblätter und Zusatzblatt vorne
- ☐ bei Bedarf AB 4_Zusatz Hypothesentests.pdf mit AB 4_Zusatz Hypothesentests_Antwortblatt.pdf austeilen und bearbeiten lassen

14:00
- ☐ Besprechung mit AB 3_Kappa verbessern_Besprechung.pdf mit AB 3_Kappa verbessern_Lösungen.pdf Antworten (stichpunktartig) auf Folie festhalten
- ☐ bei Bedarf Besprechung mit AB 4_Zusatz Hypothesentests_Besprechung.pdf mit AB 4_Zusatz Hypothesentests_Lösungen.pdf Antworten (stichpunktartig) auf Folie festhalten kann auch individuell mit Dozentenhilfe besprochen werden

14:20 Pause
- ☐ Präsentation Abschluss.pdf und Präsentation Werbung öffnen
- ☐ Facebookprofil öffnen (siehe Präsentation Abschluss Notizen.pdf)
- ☐ Evaluationsbogen.pdf bereitlegen

14:30 Reflexion
- ☐ Präsentation Abschluss.pdf halten mit Präsentation Abschluss Notizen.pdf Diskussion mit Schülerinnen und Schülern durchführen AB 3_Kappa verbessern_Besprechung.pdf als Unterstützung nutzen

14:45 Werbeblock
- ☐ Präsentation Werbung halten

14:50 Evaluation
- ☐ Evaluationsbogen.pdf austeilen und ausfüllen lassen

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

CAMMP day Soziale Netzwerke

Was euch heute erwartet!

ungefähre (!) Zeit		Programm
8:30	Begrüßung Modellierungsvortrag	Hallo und herzlich willkommen! Was heißt eigentlich „Modellieren“?
8:45	Unsere Fragestellung	Was machen wir heute?
9:15	Partnerarbeit (Teil I)	Schritt 1 zur Lösung
10:45	Pause	
11:05	Partnerarbeit (Teil II)	Schritt 2 zur Lösung
12:10	Mittagspause	
13:10	Partnerarbeit (Teil III)	Schritt 3 zur Lösung
14:20	Pause	
14:30	Abschlussgespräch	Was wir heute gelernt haben!
14:45	Ein bisschen Werbung	Das gibt es sonst noch!
14:50	Evaluation	Wie hat es euch gefallen?

A.2. Präsentation zur Problemstellung

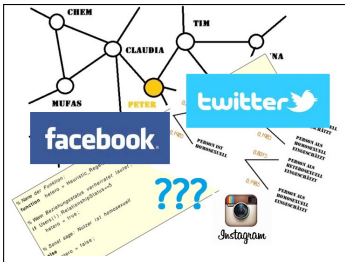
A.2.1. Abbildungsverzeichnis

- Titelbild: Logos
  facebook:
  `http://teenspeak.org/2011/11/23/the-latest-facebook-virus/` (Stand: 24.5.2016)
  twitter:
  `http://ictville.com/2013/02/forget-hate-speech-and-make-twitter-a-job-search-tool/` (Stand: 24.5.2016)
  Instagram:
  `http://www.mysmn.com/how-to-the-basics-of-instagram/instagram-logo-2/` (Stand: 24.5.2016)

- Männchen mit Fragezeichen:
  `https://tse3.mm.bing.net/th?id=OIP.M89deedb40f72e5f5787e1d4f25ca4967o0&pid=15.1&P=0&w=300&h=300` (Stand: 24.5.2016)

- Friendsterprofil:
  `https://tse4.mm.bing.net/th?id=OIP.M3f7238763e4e6185391d8e272b13f985o0&pid=15.1&P=0&w=204&h=151` (Stand: 24.5.2016)

- Organisationsfolie: Fischbild:
  `https://tse4.mm.bing.net/th?id=OIP.Ma926f92217a82c0377eb64bd472e7f29o0&pid=15.1&P=0&w=300&h=300` (Stand: 24.5.2016)



Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Sarah Schönbrodt, Benjamin Terschanski, Maike Sube
Lehrstuhl für Mathematik
Center for Computational Engineering Science

Aachen, 17. Juni 2016

Was denkt ihr?

Wie sicher ist meine Privatsphäre in sozialen Netzwerken?

Was denkt ihr?

Wie sicher ist meine Privatsphäre in **sozialen Netzwerken**?

Wer von euch ist Mitglied von sozialen Netzwerken?

Was denkt ihr?

Wie **sicher** ist **meine Privatsphäre** in sozialen Netzwerken?

Was bedeutet für euch Sicherheit in sozialen Netzwerken?

Was denkt ihr?

Wie **sicher** ist **meine Privatsphäre** in sozialen Netzwerken?

Wie sicher schätzt ihr eure Privatsphäre in sozialen Netzwerken ein?

Was denkt ihr?

Wie sicher ist meine Privatsphäre in sozialen Netzwerken?

Welche Informationen gebt ihr in einem sozialen Netzwerk über euch preis? Welche nicht?

Stellt euch vor...ein Betreiber eines sozialen Netzwerks möchte etwas über euch herausfinden, das ihr nicht angegeben habt.

Welche Informationen über euch könnten das sein?

Stellt euch vor...ein Betreiber eines sozialen Netzwerks möchte etwas über euch herausfinden, das ihr nicht angegeben habt.

Warum sollte er ein Interesse an zusätzlichen Informationen über euch haben?

Stellt euch vor...ein Betreiber eines sozialen Netzwerks möchte etwas über euch herausfinden, das ihr nicht angegeben habt.

Schätzt: Wie gut kann ihm das gelingen? Zu wie viel Prozent wird er richtig liegen?

Betreiber eines sozialen Netzwerks können **Schattenprofile** erstellen

Was ist denn ein Schattenprofil?

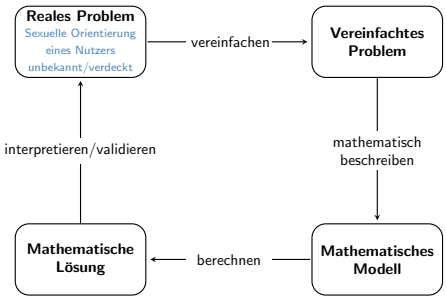
Das ist ein Profil für Nutzer des Netzwerks, das Informationen über den Nutzer enthält, die nicht angegeben wurden.

Unser heutiges Thema:

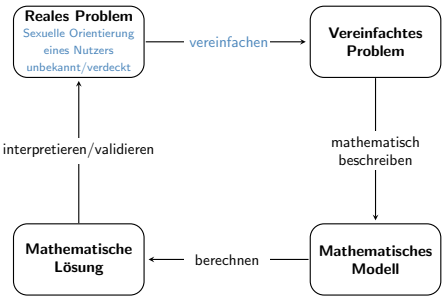
Unser heutiges Thema:

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Modellierungskreislauf



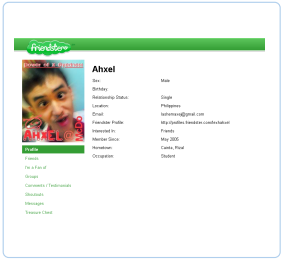
Modellierungskreislauf



Vereinfachen | Profile von Nutzern des Netzwerks *Friendster*

1. Vereinfachung: Angaben stimmen

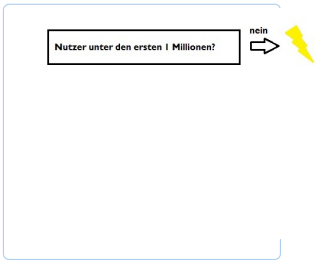
2. Vereinfachung: Daten filtern



Vereinfachen | Daten aufarbeiten

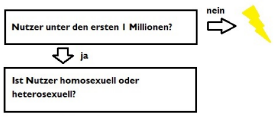
1. Vereinfachung: Angaben stimmen

2. Vereinfachung: Daten filtern



Vereinfachen | Daten aufarbeiten

1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern



Vereinfachen | Sexuelle Orientierung

1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern

- Nutzer ID (gibt die Mitgliedsnummer an, Beitrittsreihenfolge ersichtlich)
- Name als Wort
- Geburtsdatum
- Geschlecht (männlich, weiblich oder unspezifisch)
- Interessen (Freunde, Fans, Partner für Aktivitäten, Nur herumschauen, Frauen treffen, Männer treffen, Beziehung mit Frauen, Beziehung mit Männern, Beziehung mit Frauen und Männern, Frauen und Männer treffen)
- Beziehungsstatus (allein stehend, verheiratet, in einer Beziehung, in Gemeinschaft lebend, es ist kompliziert)

Vereinfachen | Sexuelle Orientierung

1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern

- Nutzer ID (gibt die Mitgliedsnummer an, Beitrittsreihenfolge ersichtlich)
- Name als Wort
- Geburtsdatum
- Geschlecht (männlich, weiblich oder unspezifisch)
- Interessen (Freunde, Fans, Partner für Aktivitäten, Nur herumschauen, Frauen treffen, Männer treffen, Beziehung mit Frauen, Beziehung mit Männern, Beziehung mit Frauen und Männern, Frauen und Männer treffen)
- Beziehungsstatus (allein stehend, verheiratet, in einer Beziehung, in Gemeinschaft lebend, es ist kompliziert)

Vereinfachen | Sexuelle Orientierung

1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern

Vereinfachen | Sexuelle Orientierung

1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern

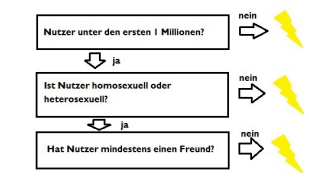
Vereinfachen | Daten aufarbeiten

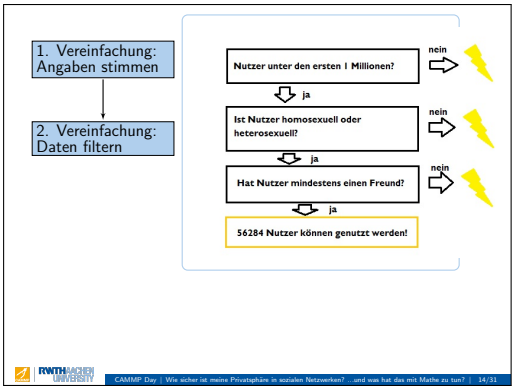
1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern

Vereinfachen | Daten aufarbeiten

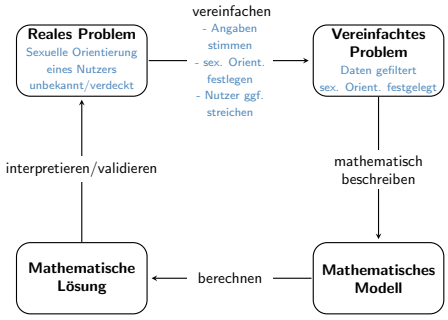
1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern

1. Vereinfachung: Angaben stimmen
2. Vereinfachung: Daten filtern

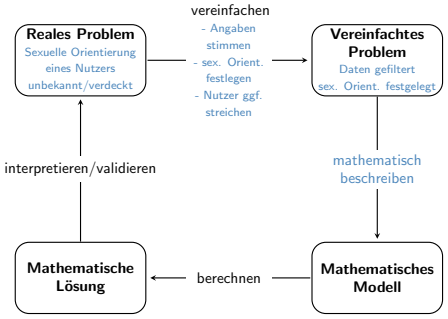




Modellierungskreislauf

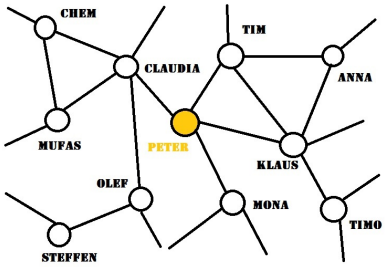


Modellierungskreislauf

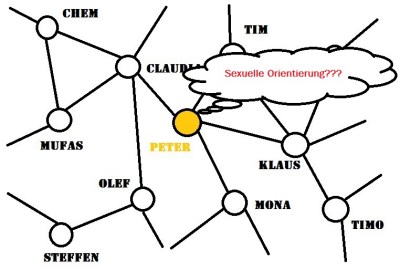


Mathematisch beschreiben | als **Graph**

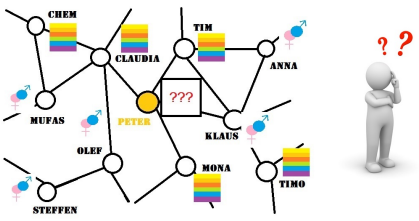
Mathematisch beschreiben | als **Graph**



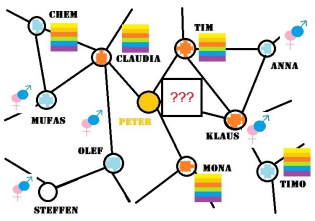
Mathematisch beschreiben | Vorhersagen machen



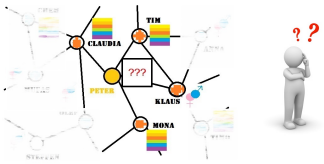
Mathematisch beschreiben | Vorhersagen machen



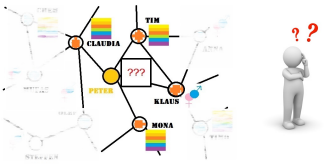
Mathematisch beschreiben | Vorhersagen machen



Mathematisch beschreiben | Vorhersagen machen

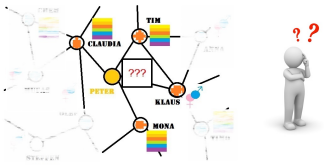


Mathematisch beschreiben | Vorhersagen machen



...anhand der **Anzahl an homosexuellen Freunden**

Mathematisch beschreiben | Vorhersagen machen



...anhand der **Anzahl an homosexuellen Freunden**

...anhand des **Anteils an homosexuellen Freunden**

Mathematisch beschreiben | Vorhersagen machen

Mathematisch beschreiben | Vorhersagen machen

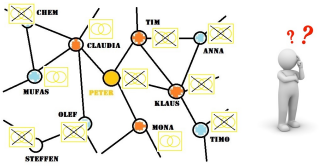
...anhand der **Anzahl an homosexuellen Freunden der Freunde**

Mathematisch beschreiben | Vorhersagen machen

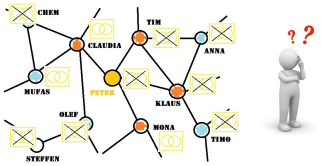
...anhand der **Anzahl an homosexuellen Freunden der Freunde**

...anhand des **Anteils an homosexuellen Freunden der Freunde**

Mathematisch beschreiben | Vorhersagen machen

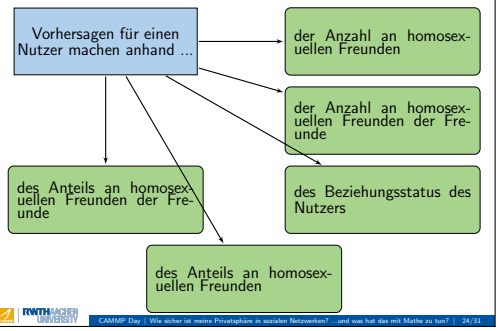


Mathematisch beschreiben | Vorhersagen machen

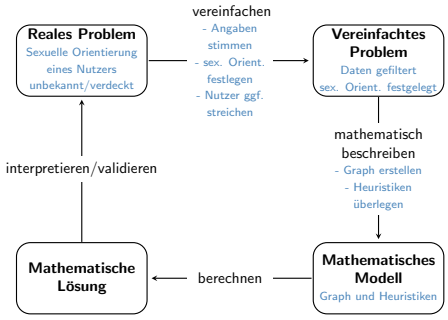


...anhand des **Beziehungsstatus des Nutzers**

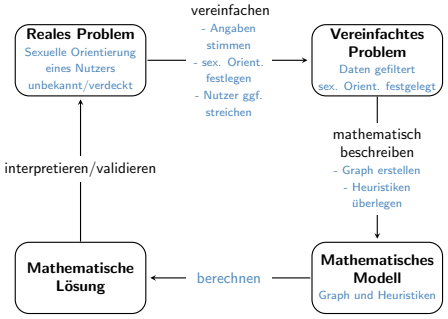
Mathematisch beschreiben | Heuristiken (Vorhersageregeln)



Modellierungskreislauf



Modellierungskreislauf



Berechnen | Heuristiken anwenden

1. Wir kennen die sexuelle Orientierung aller Nutzer, die nach der Filterung übrig geblieben sind.

Berechnen | Heuristiken anwenden

1. Wir kennen die sexuelle Orientierung aller Nutzer, die nach der Filterung übrig geblieben sind.
2. Wähle eine bestimmte Person aus.

Berechnen | Heuristiken anwenden

1. Wir kennen die sexuelle Orientierung aller Nutzer, die nach der Filterung übrig geblieben sind.
2. Wähle eine bestimmte Person aus.
3. Blende ihre sexuelle Orientierung aus.

Berechnen | Heuristiken anwenden

1. Wir kennen die sexuelle Orientierung aller Nutzer, die nach der Filterung übrig geblieben sind.
2. Wähle eine bestimmte Person aus.
3. Blende ihre sexuelle Orientierung aus.
4. Mache eine Vorhersage mit Hilfe der Regeln.

Berechnen | Heuristiken anwenden

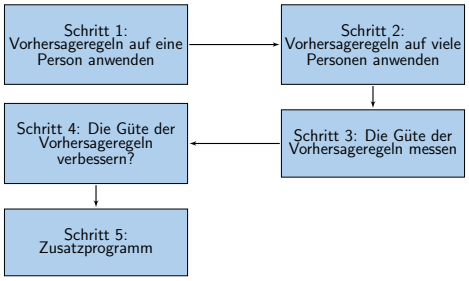
1. Wir kennen die sexuelle Orientierung aller Nutzer, die nach der Filterung übrig geblieben sind.
2. Wähle eine bestimmte Person aus.
3. Blende ihre sexuelle Orientierung aus.
4. Mache eine Vorhersage mit Hilfe der Regeln.
5. Blende die sexuelle Orientierung wieder ein und vergleiche die Vorhersage mit der tatsächlichen sexuellen Orientierung.

Berechnen | Heuristiken anwenden

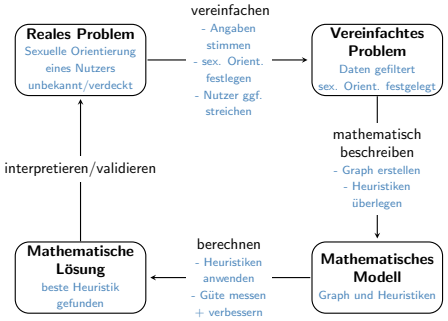
1. Wir kennen die sexuelle Orientierung aller Nutzer, die nach der Filterung übrig geblieben sind.
2. Wähle eine bestimmte Person aus.
3. Blende ihre sexuelle Orientierung aus.
4. Mache eine Vorhersage mit Hilfe der Regeln.
5. Blende die sexuelle Orientierung wieder ein und vergleiche die Vorhersage mit der tatsächlichen sexuellen Orientierung.
6. ⇒ Werden viele Nutzer richtig bestimmt? Dann funktioniert das Verfahren/die Heuristik auch bei anderen Daten!

Werkzeug zur Bestimmung der sexuellen Orientierung von Nutzern eines sozialen Netzwerks gefunden!

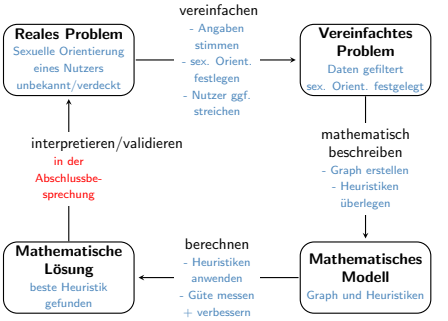
Berechnen | in mehreren Schritten



Modellierungskreislauf



Modellierungskreislauf



Jetzt seid ihr dran...

- Bearbeitet die Arbeitsblätter!
- Nutzt MATLAB!
- Arbeitet im Team!
- Nutzt die Hilfekarten!
- Fragt die Betreuer!
- Nutzt das Internet!
- Präsentiert Ergebnisse anderen Mitschülern!

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Hinweise zum Eröffnungsvortrag

Folie 1/2 | Zentrale Frage Das ist unsere heutige Fragestellung. Zuerst interessiert uns, was ihr dazu denkt!

Die folgenden Fragen werden im Plenum kurz besprochen.

Folie 3 | Frage: Soziale Netzwerke

- Wer von euch ist Mitglied von sozialen Netzwerken (z.B. facebook)? (Dazu gehört auch WhatsApp!)
- Wer ist KEIN Mitglied?
- Diese beiden Fragen werden gestellt, um die persönliche Relevanz hervorzuheben.
- Erwartet wird, dass die meisten Schülerinnen und Schüler Mitglieder sind.
- Optimal wäre es, wenn mindestens eine Person, Lehrer oder vielleicht zur Not ein Durchführender kein Mitglied ist.

Folie 4/5 | Frage: Sicherheit und Privatsphäre

- Was bedeutet für euch Sicherheit in sozialen Netzwerken?
- Hier wird der zweite Teil des Titels betrachtet.
- Erwartet werden Antworten, wie „meine Daten sollen nicht verkauft werden“ oder „es sollen keine Fakeprofile erstellt werden“.
- Die Schülerinnen und Schüler werden hier informiert: In diesem Workshop wird die Sicherheit bezüglich Daten betrachtet. Soll heißen: Ich gebe nur das preis, was ich über mich preisgeben möchte. Der Rest bleibt geheim. Das bedeutet, dass meine Privatsphäre gewahrt, also sicher, bleibt.
- Wie sicher schätzt ihr die Privatsphäre in sozialen Netzwerken (z.B. Facebook) ein?
- Diese Frage dient der späteren Reflexion.
- Eine Einschätzung, wie „sehr hoch, da ich das ja einstellen kann“ oder „eher nicht so hoch, da ich schon einmal gehört habe, dass Facebook nicht sicher ist“ kann erwartet werden.

Folie 6 | Frage: preisgegebene Informationen

- Welche Informationen gebt ihr preis? Welche nicht?
- mögliche Antwort: „ich gebe mein Alter, aber nicht meine Adresse preis“

Folie 7/8/9 | Frage: Einschätzung der Situation

- An welchen Informationen könnte ein Betreiber interessiert sein?
- mögliche Antwort: Interessen wie Sport, Beruf, **sexuelle Orientierung?! (als Dozent nennen)**...
- Warum sollte ein Betreiber eines sozialen Netzwerks Vorhersagen über Eigenschaften von dir treffen wollen? Welchen Nutzen könnte er davon haben?
- mögliche Antwort: Werbung
- Schätzt: Wie gut kann ein Betreiber eines sozialen Netzwerks Vorhersagen über Eigenschaften von dir treffen? Zu wie viel Prozent liegt er richtig?
- **auf Folie Einschätzungen notieren!!**

Folie 10/11 | Problemstellung *Die Diskussionsergebnisse werden gebündelt und zur Problemstellung überführt:*

- Wie ihr merkt, ist das Thema sehr aktuell und betrifft euch.
- Es betrifft diejenigen, die Mitglied eines Netzwerks sind
- - aber, und das ist vielleicht etwas überraschend, auch die, die KEIN Mitglied eines Netzwerks sind. Dazu kommen wir aber am Abschluss des Tages.
- Studien haben nämlich herausgefunden, dass es möglich ist, **Schattenprofile** von euch anzufertigen.
- **Definition Schattenprofil:**
  - für Nutzer des Netzwerks: Profil, das Informationen über euch enthält, die ihr nicht angegeben habt
  - (für Nicht-Nutzer: Profil mit Informationen wird erstellt.)
- Mit der Möglichkeit der Erstellung von solchen Schattenprofilen wollen wir uns heute beschäftigen.
- Dazu betrachten wir eine Information, die manche Nutzer vielleicht nicht preisgeben wollen. Es ist die **sexuelle Orientierung**.
- *Hier kann je nach Zeitfortschritt noch diskutiert werden, warum die sexuelle Orientierung nicht unbedingt preisgegeben werden möchte. Mögliche Ideen: Homosexuelle werden diskriminiert, Fußballspieler z.B.*

Wir stellen uns heute die Frage:

> Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Nun folgt der Übergang zum Modellieren:

Um eine solche Problemstellung zu beantworten, kann man die Methode des Modellierens nutzen.

Folie 12 | Reales Problem Die Problemstellung, dass die sexuelle Orientierung eines Nutzers unbekannt ist, ist unser reales Problem. Wir wollen sie vorhersagen und machen nun einige Vereinfachungen:

Folie 13/14 | Vereinfachen

Das Netzwerk Friendster

- echtes Netzwerk: Friendster
- Hier Link zeigen: http://www.friendster.com/
- diente als Vorbild für z.B. facebook.
- gegründet 2002
- bis zu 350 Millionen Nutzer (im Vergleich: facebook 1,3 Milliarden Nutzer)
- die meisten Nutzer stammten aus den USA und aus Asien
- Netzwerk nicht mehr aktiv
- Daten der Nutzer sind gespeichert worden, sodass wir mit ihnen arbeiten können

1. Vereinfachung: Angaben der Nutzer stimmen

2. Vereinfachung: Daten filtern

- Wir betrachten die ersten 1 Millionen Profile
- Wir wollen, damit wir die Vorhersagen prüfen können, nur Nutzer haben, deren sexuelle Orientierung wir kennen.
- Ein Profil kann in der originalen Form aus folgenden Informationen bestehen:
  - Nutzer ID (gibt die Mitgliedsnummer an, Beitrittsreihenfolge ersichtlich)
  - Name als Wort
  - Geburtsdatum
  - Geschlecht (männlich, weiblich oder unspezifisch)
  - Interessen (Freunde, Fans, Partner für Aktivitäten, nur herumschauen, Frauen treffen, Männer treffen, Beziehung mit Frauen, Beziehung mit Männern, Beziehung mit Frauen und Männern, Frauen und Männer treffen)[1]
  - Beziehungsstatus (allein stehend, verheiratet, in einer Beziehung, in Gemeinschaft lebend, es ist kompliziert)
- aber für unsere Frage sind nur die Nutzer interessant,
  1. die ihr Profil veröffentlicht haben,
  2. denen wir eine sexuelle Orientierung zuordnen können (wir wollen unsere Vorhersagen ja kontrollieren)
- Wir legen vereinfachend die sexuelle Orientierung eines Nutzers so fest:

> **Definition Sexuelle Orientierung:**
> Das Geschlecht eines Nutzers ist entweder weiblich oder männlich. Die sexuelle Orientierung wird in Abhängigkeit seines Interesses an einer Beziehung oder an einem Treffen (hier im Sinne eines „Dates“) von Nutzern eines bestimmten Geschlechts bestimmt:
> - Ist ein Nutzer an einer Beziehung oder an einem Treffen mit einer Person gleichen Geschlechts interessiert, so ist er **homosexuell**.
> - Ist umgekehrt ein Nutzer an einer Beziehung oder an einem Treffen mit einer Person des anderen Geschlechts interessiert, so ist er **heterosexuell**.

[1] Hier können mehrere Aspekte angegeben werden.

- Hierzu zwei Beispiele mit den Schülerinnen und Schülern besprechen: **Symbole erklären!**
  a) Peter ist an Frauen interessiert, er ist also heterosexuell.
  b) Peter ist an Männern interessiert, er ist also homosexuell.
- Die Nutzer müssen, um ihnen eine sexuelle Orientierung zuordnen zu können, ein beschriebenes Interesse in ihrem Profil angegeben haben.
- Bisexuelle Nutzer werden vereinfachend im Rahmen des CAMMP days nicht berücksichtigt.
- Außerdem sollten die Nutzer noch mindestens einen Freund im Netzwerk haben, damit wir mit ihnen arbeiten können.
- nur 56284 Nutzer erfüllen unsere Bedingungen

Folie 15 | Zusammenfassung der letzten Schritte

Die letzten Schritte werden kurz zusammengefasst.

Folie 16 | Mathematisch beschreiben

Nun wollen wir das vereinfachte Problem aus dem mathematischen Blickwinkel betrachten.

Folie 17 | Mathematisch beschreiben

Das Netzwerk als Graph

- Wie der Name „Netzwerk“ schon sagt, können wir uns das Netzwerk als „Netz“ vorstellen. (Habt ihr in Ansätzen bereits gesehen!)
- Eine solche Modellierung nennt man in der Mathematik **Graph**.
- Er besteht aus **Knoten und Kanten**. *(Hier Bild nutzen!)* Dabei modellieren der Graph das Netzwerk, die Knoten die Nutzer und die Kanten die Freundschaften.

Folie 18 | Vorhersagen machen

Heuristiken werden nun interaktiv erarbeitet.

- Stellen wir uns vor, wir kennen die sexuelle Orientierung von Peter nicht, wollen sie aber vorhersagen.

Folie 19 | Vorhersagen machen

- Nehmen wir weiter an, wir würden die sexuellen Orientierungen der Freunde kennen.
- Welche sexuelle Orientierung könnte Peter haben?
- Begründet eure Vermutung

Folie 20 | Vorhersagen machen

- Peter hat „enge“ Freunde und „weniger enge“ Freunde.
- Die engen Freunde, seine direkten Freunde sind orange markiert.
- Die weniger engen Freunde, die Freunde von Peters Freunden, sind blau markiert.
- Ist ein Freund schon ein direkter Freund von Peter, so nennen wir ihn nicht mehr Freundes Freund.

Folie 21 | Vorhersagen machen

- Nehmen wir an, wir betrachten nur Peters Freunde.
- Welche sexuelle Orientierung könnte Peter haben?
- Begründet eure Vermutung.

Folie 22 | Vorhersagen machen

- Nehmen wir an, wir betrachten nur Peters Freundes Freunde.
- Welche sexuelle Orientierung könnte Peter haben?
- Begründet eure Vermutung.

Folie 23 | Vorhersagen machen

- Nehmen wir an, wir wissen, ob die Nutzer verheiratet sind oder nicht.
- Welche sexuelle Orientierung könnte Peter haben?
- Begründet eure Vermutung.

Folie 24 | Heuristiken *Eure Vermutungen kann man als Vorhersageregeln, in der Fachsprache Heuristiken genannt, formulieren. Wir schauen uns heute die folgende Aspekte an:*

- Wir betrachten die Anzahl der homosexuellen Freunde eines Nutzers.
- Wir betrachten die Anzahl der homosexuellen Freunde der Freunde eines Nutzers.
- Wir betrachten den Beziehungsstatus eines Nutzers.
- Wir betrachten den Anteil an homosexuellen Freunden eines Nutzers.
- Wir betrachten den Anteil an homosexuellen Freunden der Freunde eines Nutzers.

Folie 25 | Zusammenfassung der letzten Schritte *Hier kommt eine kurze Zusammenfassung der letzten Schritte.*

Folie 26 | Berechnen *Nun wollen wir mit dem Modell arbeiten.*

Folie 27 | Berechnen - Heuristiken anwenden *Wenn wir eine Heuristik/Vorhersageregel auf einen Nutzer anwenden, dann machen wir das in den folgenden Schritten.*

Folie 28 | Berechnen - in mehreren Schritten *Dieses Prozedere werden wir mehrfach in 4/5 Schritten durchlaufen, um die beste Heuristik zu finden.*

Folie 29 | Zusammenfassung der letzten Schritte *Hier kurze Zusammenfassung der letzten Schritte.*

Folie 30 | validieren/interpretieren *...findet in der Abschlussbesprechung statt.*

Folie 31 | Erläuterung zum Vorgehen im CAMMP day Alles Notwendige auf der Folie.

A.3. Arbeitsblätter mit Antwortblättern, Besprechungsfolien und Lösungen

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Heuristiken

1: Anzahl der homosexuellen Freunde des Nutzers

Regel 1: *Wenn der Nutzer 5 oder mehr homosexuelle Freunde hat, dann ist er homosexuell. Wenn der Nutzer 2 oder weniger homosexuelle Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

2: Anzahl der homosexuellen Freunde der Freunde des Nutzers

Regel 2: *Wenn der Nutzer 5 oder mehr homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer 3 oder weniger homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

3: Beziehungsstatus des Nutzers

Regel 3: *Wenn der Nutzer verheiratet ist, so ist er heterosexuell. Sonst ist er homosexuell.*

4: Anteil der homosexuellen Freunde des Nutzers

Regel 4: *Wenn der Nutzer mehr als 20%(= 0.2) homosexuelle Freunde hat, dann ist er homosexuell. Wenn der Nutzer weniger als 10%(= 0.1) homosexuelle Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

5: Anteil der homosexuellen Freunde der Freunde des Nutzers

Regel 5: *Wenn der Nutzer mehr als 30%(= 0.3) homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer weniger als 10%(= 0.1) homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

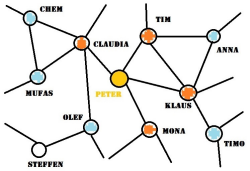


Abbildung 1: Peter und seine Freunde. Peters Freunde; Freunde von Peters Freunden

„Der Zufall entscheidet" bedeutet: 19.85% unserer Nutzer sind homosexuell. Wir sagen mit einer Wahrscheinlichkeit von 19.85% „der Nutzer ist homosexuell" und mit einer Wahrscheinlichkeit von 80.15% „der Nutzer ist heterosexuell".

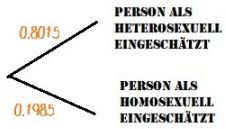


Abbildung 2: Zufallsentscheidung im Baumdiagramm

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Arbeitsblatt 1: Heuristiken Einstieg

Schritt 1 - Heuristiken kennenlernen

Wir werden mit Vorhersageregeln, den sogenannten *Heuristiken*, arbeiten. Heuristiken sind Regeln, die bestimmen, wann ein Nutzer als homo- oder heterosexuell eingestuft wird. Wir nutzen nun konkret die folgenden Regeln:

Aufgabe 1: *Lest euch die konkreten Regeln und die Erklärung zur Zufallsentscheidung (nächste Seite) genau durch.*

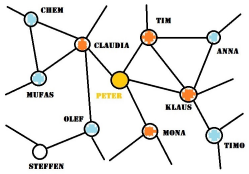


Abbildung 1: Peter und seine Freunde. Peters Freunde; Freunde von Peters Freunden

1: Anzahl der homosexuellen Freunde des Nutzers

Regel 1: *Wenn der Nutzer 5 oder mehr homosexuelle Freunde hat, dann ist er homosexuell. Wenn der Nutzer 2 oder weniger homosexuelle Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

2: Anzahl der homosexuellen Freunde der Freunde des Nutzers

Regel 2: *Wenn der Nutzer 5 oder mehr homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer 3 oder weniger homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

3: Beziehungsstatus des Nutzers

Regel 3: *Wenn der Nutzer verheiratet ist, so ist er heterosexuell. Sonst ist er homosexuell.*

4: Anteil der homosexuellen Freunde des Nutzers

Regel 4: *Wenn der Nutzer mehr als 20%(= 0.2) homosexuelle Freunde hat, dann ist er homosexuell. Wenn der Nutzer weniger als 10%(= 0.1) homosexuelle Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

5: Anteil der homosexuellen Freunde der Freunde des Nutzers

Regel 5: *Wenn der Nutzer mehr als 30%(= 0.3) homosexuelle Freundes Freunde hat, dann ist er homosexuell. Wenn der Nutzer weniger als 10%(= 0.1) homosexuelle Freundes Freunde hat, dann ist er heterosexuell. Sonst lasse den Zufall entscheiden.*

„Der Zufall entscheidet“ bedeutet: 19.85% unserer Nutzer sind homosexuell. Wir sagen mit einer Wahrscheinlichkeit von 19.85% „der Nutzer ist homosexuell“ und mit einer Wahrscheinlichkeit von 80.15% „der Nutzer ist heterosexuell“.

Aufgabe 2: *Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden? Stellt Vermutungen auf und notiert sie auf dem Antwortblatt.*

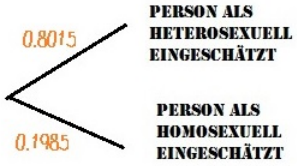


Abbildung 2: Zufallsentscheidung im Baumdiagramm

Schritt 2 - Heuristiken auf eine Person anwenden

Wir wollen die Regeln an einem Nutzer testen:

a) Arbeitet mit Matlab. Die Datei „sozialenetzwerke“ ist bereits geöffnet. Geht zum Abschnitt „Heuristiken auf eine Person anwenden“ (Z. 13-19).

b) Überlegt euch eine Nummer zwischen 1 und 56284 (Nummer eines Nutzers). Ersetzt den Ausdruck NaN in Zeile 15 durch diese Nummer. So wählt ihr den Nutzer mit dieser Nummer aus.

c) Drückt nun auf *Run Section*. Es erscheinen die tatsächliche sexuelle Orientierung des Nutzers und die Vorhersagen der Heuristiken im Command Window.

d) *Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden?*

Wendet zur Beantwortung diese Frage das beschriebene Vorgehen auf die Nutzer mit den Nummern **66, 6811, 55975 und 16651** an. Notiert auf eurem Antwortzettel, ob die Regeln richtig oder falsch lagen. Wiederholt die Bestimmung mehrfach für die Nutzer 6811, 55975 und 16651. Notiert auf eurem Antwortzettel, ob die Regeln richtig oder falsch lagen.

Welche Vorhersageregeln funktionieren gut, welche nicht? Notiert eure Einschätzungen auf dem Antwortblatt.

Schritt 3 - Heuristiken auf viele Personen anwenden

Ein Betreiber eines Netzwerks möchte nicht nur Vorhersagen über eine Person machen sondern über viele Nutzer und dabei möglichst viele richtige Vorhersagen treffen. Schauen wir uns das an:

Jetzt könnt ihr euch aussuchen, welche Regel ihr testen möchtet. Holt euch dazu das entsprechende Arbeitsblatt.

Hinweis zu Matlab: Matlab kann auch als Taschenrechner verwendet werden. Gebt in das Command Window hinter den zwei Pfeilen (») eine Rechnung (z.B. 100/(30+50)) ein und drückt die Entertaste. Das Ergebnis erscheint als *ans =....*

Zusatzaufgabe - weitere Heuristiken testen

Arbeitet mit einer anderen Regel, bearbeitet dazu das entsprechende Arbeitsblatt.

Wir besprechen nun eure Ergebnisse!

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 1 - Schritt 3 - Regel 1

a) Arbeitet mit dem Abschnitt „Heuristiken auf viele Personen anwenden“ (Z. 20-60). Geht in den Abschnitt „mit Regel 1“ (Z. 22-29).

b) **Verschiedene Anzahlen von Nutzern testen:**

- Überlegt euch, wie viele Personen ihr einschätzen möchtet (Zahl zwischen 1 und 56284). Diese Zahl ist die Größe unserer *Stichprobe*. Gebt die Zahl in Zeile 24 anstelle von NaN ein.
- Führt die Einschätzung der Stichprobe durch, indem ihr auf *Run Section* drückt.
- Notiert auf eurem Antwortblatt: die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- Testet noch weitere Anzahlen von Nutzern, indem ihr die Zahl N (Z. 24) neu eingebt und den Abschnitt „mit Regel 1“ ausführt (*Run Section*). Notiert auch hier die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- *Welche sexuellen Orientierungen können mit Regel 1 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort auf dem Antwortblatt.*

c) **Verschiedene Zahlen in der Regel testen:**

In der Regel 1 seht ihr im Kasten bei Schritt 1 rote Zahlen markiert. Diese sind entscheidend für die Arbeitsweise der Heuristik. Aber sind die roten Zahlen schlau gewählt?

- Legt eine „große“ Zahl für N (Z. 24) fest (mindestens 10000, höchstens 56284). Notiert sie auf dem Antwortblatt.
- Testet andere Zahlen, indem ihr in Zeile 25 die Zahlen 5 und 2 durch Zahlen eurer Wahl ersetzt (die Zahlen stehen in der gleichen Reihenfolge wie in der Regel, ein Leerzeichen muss zwischen den Zahlen stehen).
- Führt den Abschnitt aus (*Run Section*).
- Notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- Testet weitere Zahlen und notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- *Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?* Begründet eure Antwort auf dem Antwortblatt.
- *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 1 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein. Begründet eure Antwort auf dem Antwortblatt.*

Hinweis: Homosexuelle Nutzer haben im Durchschnitt 3.018 homosexuelle Freunde und heterosexuelle Nutzer haben im Durchschnitt 0.2136 homosexuelle Freunde.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 1 - Schritt 3 - Regel 2

a) Arbeitet mit dem Abschnitt „Heuristiken auf viele Personen anwenden“ (Z. 20-60). Geht in den Abschnitt „mit Regel 2“ (Z. 30-37).

b) **Verschiedene Anzahlen von Nutzern testen:**

- Überlegt euch, wie viele Personen ihr einschätzen möchtet (Zahl zwischen 1 und 56284). Diese Zahl ist die Größe unserer *Stichprobe*. Gebt die Zahl in Z. 32 anstelle von NaN ein.
- Führt die Einschätzung der Stichprobe durch, indem ihr auf *Run Section* drückt.
- Notiert auf eurem Antwortblatt: die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- Testet noch weitere Anzahlen von Nutzern, indem ihr die Zahl N (Z. 32) neu eingebt und den Abschnitt „mit Regel 2“ ausführt (*Run Section*). Notiert auch hier die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- *Welche sexuellen Orientierungen können mit Regel 2 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort auf dem Antwortblatt.*

c) **Verschiedene Zahlen in der Regel testen:**

In der Regel 2 seht ihr im Kasten bei Schritt 1 rote Zahlen markiert. Diese sind entscheidend für die Arbeitsweise der Heuristik. Aber sind die roten Zahlen schlau gewählt?

- Legt eine „große“ Zahl für N (Z. 32) fest (mindestens 10000, höchstens 56284). Notiert sie auf dem Antwortblatt.
- Testet andere Zahlen, indem ihr in Zeile 33 die Zahlen 5 und 3 durch Zahlen eurer Wahl ersetzt (die Zahlen stehen in der gleichen Reihenfolge wie in der Regel, ein Leerzeichen muss zwischen den Zahlen stehen).
- Führt den Abschnitt aus (*Run Section*).
- Notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- Testet weitere Zahlen und notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- *Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?* Begründet eure Antwort auf dem Antwortblatt.
- *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 2 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein. Begründet eure Antwort auf dem Antwortblatt.*

Hinweis: Homosexuelle Nutzer haben im Durchschnitt 21.8854 homosexuelle Freundes Freunde und heterosexuelle Nutzer haben im Durchschnitt 1.8585 homosexuelle Freundes Freunde.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 1 - Schritt 3 - Regel 3

a) Arbeitet mit dem Abschnitt „Heuristiken auf viele Personen anwenden“ (Z. 20-60). Geht in den Abschnitt „mit Regel 3“ (Z. 38-44).

b) **Verschiedene Anzahlen von Nutzern testen:**

- Überlegt euch, wie viele Personen ihr einschätzen möchtet (Zahl zwischen 1 und 56284). Diese Zahl ist die Größe unserer *Stichprobe*. Gebt diese Zahl in Z. 40 anstelle von NaN ein.
- Drückt auf *Run Section*.
- Notiert auf eurem Antwortblatt: die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- Testest noch weitere Anzahlen von Nutzern, indem ihr die Zahl N (Z. 40) neu eingebt und den Abschnitt ausführt (*Run Section*). Notiert auch hier die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- *Welche sexuellen Orientierungen können mit Regel 3 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort auf dem Antwortblatt.*
- *Welche Stichprobengröße liefert die aussagekräftigsten Ergebnisse? Begründet eure Antwort auf dem Antwortblatt.*
- *Wie gut kann die sexuelle Orientierung mit Regel 3 vorhergesagt werden? Schätzt die Güte der Heuristik ein. Begründet eure Antwort auf dem Antwortblatt.*

Hinweis: Es sind 1.06% der homosexuellen und 0.86% der heterosexuellen Nutzer verheiratet.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 1 - Schritt 3 - Regel 4

a) Arbeitet mit dem Abschnitt „Heuristiken auf viele Personen anwenden“ (Z. 20-60). Geht in den Abschnitt „mit Regel 4“ (Z. 45-52).

b) **Verschiedene Anzahlen von Nutzern testen:**

- Überlegt euch, wie viele Personen ihr einschätzen möchtet (Zahl zwischen 1 und 56284). Diese Zahl ist die Größe unserer *Stichprobe*. Gebt die Zahl in Zeile 47 anstelle von NaN ein.
- Führt die Einschätzung der Stichprobe durch, indem ihr auf *Run Section* drückt.
- Notiert auf eurem Antwortblatt: die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- Testet noch weitere Anzahlen von Nutzern, indem ihr die Zahl N (Z. 47) neu eingebt und den Abschnitt „mit Regel 4“ ausführt (*Run Section*). Notiert auch hier die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- *Welche sexuellen Orientierungen können mit Regel 4 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort auf dem Antwortblatt.*

c) **Verschiedene Zahlen in der Regel testen:**

In der Regel 4 seht ihr im Kasten bei Schritt 1 rote Zahlen markiert. Diese sind entscheidend für die Arbeitsweise der Heuristik. Aber sind die roten Zahlen schlau gewählt?

- Legt eine „große“ Zahl für N (Z. 47) fest (mindestens 10000, höchstens 56284). Notiert sie auf dem Antwortblatt.
- Testet andere Zahlen, indem ihr in Zeile 48 die Zahlen 0.2 und 0.1 durch Zahlen eurer Wahl ersetzt (die Zahlen stehen in der gleichen Reihenfolge wie in der Regel, Achtung! Komma als Punkt eingeben!, ein Leerzeichen muss zwischen den Zahlen stehen).
- Führt den Abschnitt aus (*Run Section*).
- Notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- Testet weitere Zahlen und notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- *Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?* Begründet eure Antwort auf dem Antwortblatt.
- *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 4 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein. Begründet eure Antwort auf dem Antwortblatt.*

Hinweis: Homosexuelle Nutzer haben im Durchschnitt einen Anteil von 0.7778 homosexuellen Freunden und heterosexuelle Nutzer haben im Durchschnitt einen Anteil von 0.0687 homosexuellen Freunden.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 1 - Schritt 3 - Regel 5

a) Arbeitet mit dem Abschnitt „Heuristiken auf viele Personen anwenden“ (Z. 20-60). Geht in den Abschnitt „mit Regel 5“ (Z. 53-60).

b) **Verschiedene Anzahlen von Nutzern testen:**

- Überlegt euch, wie viele Personen ihr einschätzen möchtet (Zahl zwischen 1 und 56284). Diese Zahl ist die Größe unserer *Stichprobe*. Gebt die Zahl in Z. 55 anstelle von NaN ein.
- Führt die Einschätzung der Stichprobe durch, indem ihr auf *Run Section* drückt.
- Notiert auf eurem Antwortblatt: die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- Testet noch weitere Anzahlen von Nutzern, indem ihr die Zahl N (Z. 55) neu eingebt und den Abschnitt „mit Regel 5“ ausführt (*Run Section*). Notiert auch hier die Größe der Stichprobe und die Ergebnisse der Einschätzung.
- *Welche sexuellen Orientierungen können mit Regel 5 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort auf dem Antwortblatt.*

c) **Verschiedene Zahlen in der Regel testen:**

In der Regel 5 seht ihr im Kasten bei Schritt 1 rote Zahlen markiert. Diese sind entscheidend für die Arbeitsweise der Heuristik. Aber sind die roten Zahlen schlau gewählt?

- Legt eine „große“ Zahl für N (Z. 55) fest (mindestens 10000, höchstens 56284). Notiert sie auf dem Antwortblatt.
- Testet andere Zahlen, indem ihr in Zeile 56 die Zahlen 0.3 und 0.1 durch Zahlen eurer Wahl ersetzt (die Zahlen stehen in der gleichen Reihenfolge wie in der Regel, Achtung! Komma als Punkt eingeben!, ein Leerzeichen muss zwischen den Zahlen stehen).
- Führt den Abschnitt aus (*Run Section*).
- Notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- Testet weitere Zahlen und notiert die Zahlen, die ihr eingegeben habt, und die Anzahlen der richtigen und der falschen Einschätzungen auf dem Antwortblatt.
- *Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?* Begründet eure Antwort auf dem Antwortblatt.
- *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 5 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein. Begründet eure Antwort auf dem Antwortblatt.*

Hinweis: Homosexuelle Nutzer haben im Durchschnitt einen Anteil von 0.7447 homosexuellen Freundes Freunden und heterosexuelle Nutzer haben im Durchschnitt einen Anteil von 0.1007 homosexuellen Freundes Freunden.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 1: Antwortblatt

Schritt 1 - Heuristiken kennenlernen

Aufgabe 2: *Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden?*

Wir vermuten **anhand der Betrachtung der konkreten Regeln**, dass...

Schritt 2 - Heuristiken auf eine Person anwenden

Aufgabe d) *Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden?*

Notiert in der Tabelle, ob die Regeln die sexuelle Orientierung richtig ($\surd$) oder falsch ($\emptyset$) bestimmt haben.

Regel	Nutzer 66	Nutzer 6811 | Wiederh.	Nutzer 55975 | Wiederh.	Nutzer 16651 | Wiederh.
1				
2				
3				
4				
5				

Wir vermuten **anhand der Anwendung der konkreten Regeln auf einzelne Nutzer**, dass...

Schritt 3 - Heuristiken auf viele Personen anwenden

mit Regel:

Aufgabe b) Verschiedene Anzahlen von Nutzern testen

Größe der Stichprobe:				
richtige Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Hetero:				
Geschätzt Homo und ist Homo:				
falsche Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Homo:				
Geschätzt Homo und ist Hetero:				

Welche sexuellen Orientierungen können mit der Regel gut, welche schlecht eingeschätzt werden?

- Für Regel 3:
  Welche Stichprobengröße liefert die aussagekräftigsten Ergebnisse?

  Wie gut kann die sexuelle Orientierung mit der Regel vorhergesagt werden? Schätzt die Güte der Heuristik ein.

- Für Regel 1, 2, 4, 5:

Aufgabe c) Verschiedene Zahlen in der Regel testen

Größe der Stichprobe:

mit den Zahlen:					
richtige Vorhersagen:					
falsche Vorhersagen:					

Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?

Wie gut kann die sexuelle Orientierung mit Hilfe der Regel vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Einstieg und Arbeitsblatt 1: Besprechung

Einschätzung zu Beginn

Wie gut kann ein Betreiber eines sozialen Netzwerks Informationen über euch herausfinden?

Wir vermuten, dass er zu ..... % richtig liegen wird.

Schritt 1 - Heuristiken kennenlernen

Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden?

Wir vermuten **anhand der konkreten Regeln**, dass ..... die beste Regel ist, da....

Schritt 2 - Heuristiken auf eine Person anwenden

Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden?

Regel	Nutzer 66	Nutzer 6811 Wiederh.	Nutzer 55975 Wiederh.	Nutzer 16651 Wiederh.
1				
2				
3				
4				
5				

Wir vermuten **anhand der Anwendung der Regeln auf einzelne Nutzer**, dass

..... die beste Regel ist, da....

Schritt 3 - Heuristiken auf viele Personen anwenden

Regel	schätzt am besten .... ein	schätzt insgesamt ... % der Nutzer richtig ein	mit den Zahlen
1			
2			
3			•
4			
5			

Welche Stichprobengröße liefert die aussagekräftigsten Ergebnisse?
Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 1: Lösungsvorschlag

Anmerkung: Wird im Folgenden von einer „Lösung“ gesprochen, so ist dies die einzige richtige Lösungsvariante. Wird jedoch von einem „Lösungsvorschlag“ gesprochen, so ist hier nur eine Variante zur Lösung dargelegt. In diesem CAMMP day spielt der Zufall auch in den Ergebnissen eine große Rolle, sodass durchaus mehrere Ergebnisse möglich sind und längst nicht alle aufführbar. Wichtig ist hier, dass das Vorgehen annähernd wissenschaftlich (z.B. durch genügend Wiederholungen) ist. Die Lösungsvorschläge dienen in jedem Fall zur Orientierung, in welchen Bereichen gute Werte liegen können.

Schritt 1 - Aufgabe 2

- **Aufgabe:** Anhand der Betrachtung der Regeln die Frage beantworten: *Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden? Stellt Vermutungen auf und notiert sie.*
- **Lösungsvorschlag:** Regel 1 und 4 sind am besten, da die engeren Freunde vermutlich am meisten über einen aussagen. Regel 2 und 5 werden am zweitbesten sein, da hier der weitere Freundeskreis betrachtet wird und der auch etwas über einen aussagen kann. Regel 3 ist vermutlich am schlechtesten, da ich denke, dass sich viele Singles im Netzwerk befinden, die heterosexuell sind. Genaue Trefferzahlen sind schwer anzugeben.
- **Hinweis zur Besprechung:** Grundsätzlich sind hier auch andere Lösungen möglich, solange die Schülerinnen und Schüler ihre Antworten begründen können. Diese Frage dient einem ersten Einschätzungsversuch der Heuristiken aus dem reinen Vorwissen bzw. aus dem reinen Gefühl.

Schritt 2 - Aufgabe d)

- **Aufgabe:** *Wie gut kann die sexuelle Orientierung mit Hilfe dieser Regeln vorhergesagt werden? Welche Heuristik scheint die sexuelle Orientierung gut zu bestimmen, welche eher nicht?*
- **Anmerkung zur Deutung der Ergebnisse:** Fehleinschätzungen der sexuellen Orientierungen durch Regel 1 bzw. 2 und 4 bzw. 5 können an einer unzureichenden Festlegung der Freundeszahlen (5 und 2; 5 und 3; 0.2 und 0.1; 0.3 und 0.1) oder an einer Fehleinschätzung durch den Zufall, falls der Nutzer 3 oder 4 bzw. zwischen 10% und 20% homosexuelle Freunde oder 4 bzw. zwischen 10% und 30% homosexuelle Freundes Freunde hat, liegen. Fehleinschätzungen mit Regel 3 sind immer auf eine falsche Deutung aus dem Verheiratetsein zurückzuführen. Auch hier kann an einer sinnvollen Formulierung der Regel gezweifelt werden.
- **Lösungsvorschlag:**

Regel	66	6811	6811 (Wdh.)	55975	55975 (Wdh.)	16651	16651 (Wdh.)
1	∅	√	√	∅	√	∅	∅
2	∅	∅	∅	√	√	∅	√
3	√	∅	∅	∅	∅	√	√
4	√	√	∅	√	√	√	√
5	∅	√	∅	√	√	∅	√

  - Nutzer **66** ist homosexuell und
    - hat 1 homosex. Freund 1. Ordnung (Regel 1 ist immer hetero)
    - hat 0 homosex. Freunde 2. Ordnung (Regel 2 ist immer hetero)
    - ist nicht verheiratet (Regel 3 ist immer homo)

    - hat 25 % homosex. Freunde 1. Ordnung (Regel 4 ist homo)
    - hat 0 % homosex. Freunde 2. Ordnung (Regel 2 ist hetero)
  - **Wiederholte Anwendung der Bestimmung bei Nutzer 6811 (heterosex.):** *Das Ergebnis bei Regel 4 und 5 hat sich geändert. Der Nutzer muss zwischen 10% und 20% homosexuelle Freunde und zwischen 10% und 30% homosexuelle Freundes Freunde haben. In der Tat hat der Nutzer ca. 18,2% homosexuelle Freunde und ca. 16% homosexuelle Freundes Freunde. Damit fällt er bei beiden Regeln in die Regel „bestimme mit Zufall“. Regel 1 und 2 werden sich nicht ändern, da der Nutzer 2 homosexuelle Freunde und 13 homosexuelle Freundes Freunde hat.*
  - **Wiederholte Anwendung der Bestimmung bei Nutzer 55975 (homosex.):** *Das Ergebnis bei Regel 1 hat sich geändert. Der Nutzer muss also 3 oder 4 homosexuelle Freunde haben und fällt damit in die Regel „bestimme mit Zufall“. Also sind die Ergebnisse bei Regel 1 nicht immer gleich! In der Tat hat der Nutzer 3 homosexuelle Freunde. Die anderen Ergebnisse werden sich nicht ändern, da der Nutzer 8 homosexuelle Freundes Freunde hat sowie 50% homosexuelle Freunde und ca. 35% homosexuelle Freundes Freunde.*
  - **Wiederholte Anwendung der Bestimmung bei Nutzer 16651 (homosex.):** *Das Ergebnis bei Regel 2 und 5 hat sich geändert. Der Nutzer muss also 4 homosexuelle Freundes Freunde haben und fällt damit in die Regel „bestimme mit Zufall“. Also sind die Ergebnisse bei Regel 2 nicht immer gleich! In der Tat hat der Nutzer 4 homosexuelle Freundes Freunde. Regel 1 wird sich nicht ändern, da der Nutzer 2 homosexuelle Freunde hat. Der Nutzer muss außerdem zwischen 10% und 30% homosexuelle Freundes Freunde haben, da er bei Regel 5 in „bestimme mit Zufall“ fällt. In der Tat hat der Nutzer ca. 28% homosexuelle Freundes Freunde. Regel 4 wird sich nicht ändern, da der Nutzer ca. 33% homosexuelle Freunde hat.*
- **Insgesamt ist festzuhalten:**
  - Fällt ein Nutzer mit den Freundeszahlen in die Zufallsbestimmung von Regel 1, 2, 4 oder 5, so muss die Bestimmung bei wiederholter Durchführung nicht konsistent sein.
  - Regel 1, 2, 3 haben bei insgesamt 3 von 7 Bestimmungen richtig gelegen, Regel 4 bei 6 von 7 und Regel 5 bei 4 von 7.
  - Mit diesen Ergebnissen war Regel 4 die beste Heuristik und Regel 5 die zweitbeste. Die Regeln 1, 2 und 3 war die schlechtesten.
- **Hinweise zur Besprechung:** Sind die Ergebnisse anders oder wurden mehr Durchläufe gemacht, so kann die Deutung anders ausfallen. Tendenziell ist die Regel 3 am schlechtesten, Regel 2 am zweitschlechtesten, Regel 1 in der Mitte, Regel 5 am zweitbesten und die Regel 4 am besten. Grundsätzlich kann aber an dieser Stelle noch nicht viel über die Güte gesagt werden. Wir müssen also weiter machen!
- **Code und Ausgabe:**

```
%       Wie sicher ist meine Privatsphäre in sozialen Netzwerken
%               ... und was hat das mit Mathe zu tun?

%%                      Heuristiken – Einstieg
%%      Lade Daten

clc
close all
load v2Adj
load v3FriendsterUser

%%      Heuristiken auf eine Person anwenden
clc
i = 66;

% Heurostiken werden angewendet:
```

```
AnwendenaufeinePerson(Users, i);
```

Der Nutzer mit Nummer
66

ist homosexuell
_____

Die Vorhersagen lauten:
Regel 1: Person ist heterosexuell
Regel 2: Person ist heterosexuell
Regel 3: Person ist homosexuell
Regel 4: Person ist homosexuell
Regel 5: Person ist heterosexuell

Schritt 3 - Heuristiken auf viele Personen anwenden

- **Regel 1**:
  - **b)** *Welche sexuellen Orientierungen können mit Regel 1 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort.*

Lösungsvorschlag:

Größe der Stichprobe:	100	1000	10000	30000
richtige Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Hetero:	73	786	7954	24087
Geschätzt Homo und ist Homo:	6	39	457	1302
falsche Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Homo:	21	174	1556	4508
Geschätzt Homo und ist Hetero:	0	1	33	103

Beantwortung der Frage:

- Besonders gut kann die sexuelle Orientierung bei heterosexuellen Nutzern vorhergesagt werden. Über 99% der heterosexuellen Nutzer werden richtig eingeschätzt.
- Homosexuelle Nutzer können nicht gut eingeschätzt werden. Hier werden nur um die 20% der Nutzer richtig eingeschätzt.

  - **c)** Warum ist es sinnvoll, eine „große“ Zahl zu wählen?
  Lösung: Wir brauchen eine hohe Zahl als Zahl N, da sonst die Ergebnisse nicht belastbar sind (Stichwort: Gesetz großer Zahlen).
  - **c)** *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 1 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein.*
  Lösungsvorschlag: Wir wählen exemplarisch 10000 (das sind ca. 20% der Nutzer) und testen 5 Kombinationen:
  Ideen hinter den Zahlen in Spaltenreihenfolge:
  Zahlen aus Hinweis verwenden - Regel ohne Zufall - Zahlen, die weit voneinander entfernt sind - größere Zahlen verwenden - Einfach etwas ausprobiert

mit den Zahlen:	3 und: 1	3 und: 2	9 und: 2	10 und: 7	1 und: 0
richtige Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Hetero:	7899	7942	7973	7985	6763
Geschätzt Homo und ist Homo:	731	700	263	149	1732
falsche Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Homo:	1259	1269	1747	1863	240
Geschätzt Homo und ist Hetero:	111	89	17	3	1265

Beantwortung der Frage: Mit den Zahlen 3 und 1 können ca. $86,3\%$, mit den Zahlen 3 und 2 können ca. $86,42\%$, mit den Zahlen 9 und 2 können ca. $82,36\%$, mit den Zahlen

10 und 7 können ca. $81,34\%$ und mit den Zahlen 1 und 0 können ca. $84,95\%$ der Nutzer richtig eingeschätzt werden. Hier ist eine Steigerung zu den zuvor genutzten Zahlen 5 und 2 (mit ca. 80%) zu sehen. 3 und 2 sind hier insgesamt am besten. Es scheint, als würde der Zufall hier am besten ausgeschlossen werden.

- **Regel 2**:
  - **b)** *Welche sexuellen Orientierungen können mit Regel 2 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort.*

Lösungsvorschlag:

Größe der Stichprobe:	100	1000	10000	30000
richtige Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Hetero:	80	715	7045	21167
Geschätzt Homo und ist Homo:	8	118	1175	3567
falsche Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Homo:	7	63	858	3567
Geschätzt Homo und ist Hetero:	5	104	922	2755

Beantwortung der Frage:

- Heterosexuelle Nutzer können gut eingeschätzt werden. Hier werden zwischen 87% und 94% der heterosexuellen Nutzer richtig eingeschätzt.
- Homosexuelle Nutzer können weniger gut eingeschätzt werden. Hier werden zwischen 53% und 65% richtig eingeschätzt.

  - **c)** Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?
  Lösung: Wir brauchen eine hohe Zahl als Zahl N, da sonst die Ergebnisse nicht belastbar sind (Stichwort Gesetz großer Zahlen).
  - **c)** *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 2 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein.*
  Lösungsvorschlag: Wir wählen exemplarisch 10000 und testen 5 Kombinationen:
  Ideen hinter den Zahlen in Spaltenreihenfolge:
  Zahlen aus Hinweis verwenden - Regel ohne Zufall - Zahlen, die weit voneinander entfernt sind - größere Zahlen verwenden - Einfach etwas ausprobiert

mit den Zahlen:	22 und: 2	5 und: 4	50 und: 7	45 und: 39	150 und: 9
richtige Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Hetero:	7786	7175	7848	7989	7903
Geschätzt Homo und ist Homo:	660	1135	414	259	192
falsche Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Homo:	1200	851	1618	1725	1823
Geschätzt Homo und ist Hetero:	354	839	120	27	82

Beantwortung der Frage: Mit den Zahlen 22 und 2 können ca. $84,46\%$, mit den Zahlen 5 und 4 können ca. $83,1\%$, mit den Zahlen 50 und 7 können ca. $82,62\%$, mit den Zahlen 45 und 39 können ca. $82,48\%$ und mit den Zahlen 150 und 9 können ca. $80,95\%$ der Nutzer richtig eingeschätzt werden. Hier sind die Zahlen 22 und 2 insgesamt am besten. Eine deutliche Steigerung ist hier im Vergleich zu den Zahlen 5 und 3 (mit zwischen 82% und 88%) nicht zu sehen. Hier sollte der Zufall also in der Regel erhalten bleiben.

- Ihr nutzt **Regel 3**:
  - **b)** *Welche sexuellen Orientierungen können mit Regel 3 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort.*

Lösungsvorschlag:

Größe der Stichprobe:	100	1000	10000	30000
richtige Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Hetero:	0	7	73	202
Geschätzt Homo und ist Homo:	26	210	1948	5902
falsche Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Homo:	0	4	19	63
Geschätzt Homo und ist Hetero:	74	779	7960	23833

Beantwortung der Fragestellung:

* Besonders gut können homosexuelle Nutzer eingeschätzt werden. Hier werden immer über 98% der Nutzer richtig eingeschätzt. Hier sagen auch die Daten, dass wenige homosexuelle Nutzer verheiratet sind.
* Heterosexuelle können äußerst schlecht eingeschätzt werden. Zwischen 0% und 9% werden richtig eingeschätzt. Dies liegt daran, dass nur 0.86% verheiratet sind.

– **b)** *Welche Stichprobengröße liefert die aussagekräftigsten Ergebnisse? Begründet eure Antwort.*
Lösung: Wir brauchen eine hohe Zahl als Zahl N (mehr als 10000), da sonst die Ergebnisse nicht belastbar sind (Stichwort Gesetz großer Zahlen).

– **b)** *Wie gut kann die sexuelle Orientierung mit Regel 3 vorhergesagt werden? Schätzt die Güte der Heuristik ein.*
Lösungsvorschlag: Die Heuristik schätzt $20-26\%$ der Nutzer richtig ein. Das ist schlecht. Die Idee hinter der Regel ist für unseren Datensatz also nicht gewinnbringend.

- Ihr nutzt **Regel 4**:

– **b)** *Welche sexuellen Orientierungen können mit Regel 4 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort.*

Lösungsvorschlag:

Größe der Stichprobe:	100	1000	10000	30000
richtige Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Hetero:	71	723	7075	21253
Geschätzt Homo und ist Homo:	18	164	1732	5148
falsche Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Homo:	3	30	294	795
Geschätzt Homo und ist Hetero:	8	83	899	2804

Beantwortung der Frage:

* Heterosexuelle Nutzer können gut eingeschätzt werden. Hier werden zwischen 88% und 90% der heterosexuellen Nutzer richtig eingeschätzt.
* Homosexuelle Nutzer können auch gut eingeschätzt werden. Hier werden zwischen 84% und 87% richtig eingeschätzt.

– **c)** Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?
Lösung: Wir brauchen eine hohe Zahl als Zahl N, da sonst die Ergebnisse nicht belastbar sind (Stichwort Gesetz großer Zahlen).

– **c)** *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 4 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein.*
Lösungsvorschlag: Wir wählen exemplarisch 10000 und testen 5 Kombinationen :
Ideen hinter den Zahlen in Spaltenreihenfolge:
Zahlen aus Hinweis verwenden - Regel mit wenig Zufall - Zahlen, die weit voneinander entfernt sind - größere Zahlen verwenden - Einfach etwas ausprobiert

mit den Zahlen:	0.7 ; 0.1	0.56 ; 0.55	0.9 ; 0.1	0.9 ; 0.7	0.8 ; 0.3
richtige Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Hetero:	7528	7619	7504	7725	7643
Geschätzt Homo und ist Homo:	1515	1558	1375	1309	1417
falsche Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Homo:	470	439	659	656	541
Geschätzt Homo und ist Hetero:	487	384	462	310	399

Beantwortung der Frage: Mit den Zahlen 0.7 und 0.1 können ca. $90,43\%$, mit den Zahlen 0.56 und 0.55 können ca. $91,77\%$, mit den Zahlen 0.9 und 0.1 können ca. $88,79\%$, mit den Zahlen 0.9 und 0.7 können ca. $90,34\%$ und mit den Zahlen 0.8 und 0.3 können ca. $90,6\%$ der Nutzer richtig eingeschätzt werden. Hier sind die Zahlen 0.56 und 0.55 am besten. Der Zufall sollte also aus der Regel weitestgehend herausgenommen werden. Eine Steigerung ist hier im Vergleich zu den Zahlen 0.2 und 0.1 (mit $88-89\%$) zu sehen.

- Ihr nutzt **Regel 5**:

– **b)** *Welche sexuellen Orientierungen können mit Regel 5 gut, welche schlecht eingeschätzt werden? Begründet eure Antwort.*

Lösungsvorschlag:

Größe der Stichprobe:	100	1000	10000	30000
richtige Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Hetero:	65	689	6786	20353
Geschätzt Homo und ist Homo:	17	153	1548	4642
falsche Vorhersagen:	•	•	•	•
Geschätzt Hetero und ist Homo:	4	35	428	1295
Geschätzt Homo und ist Hetero:	14	123	1238	3710

Beantwortung der Frage:

* Heterosexuelle Nutzer können gut eingeschätzt werden. Hier werden zwischen 82% und 85% der heterosexuellen Nutzer richtig eingeschätzt.
* Homosexuelle Nutzer können auch recht gut eingeschätzt werden. Hier werden zwischen 78% und 82% richtig eingeschätzt.

– **c)** Warum ist es sinnvoll, eine „große“ Zahl für N zu wählen?
Lösung: Wir brauchen eine hohe Zahl als Zahl N, da sonst die Ergebnisse nicht belastbar sind (Stichwort Gesetz großer Zahlen).

– **c)** *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 5 vorhergesagt werden? Welche Zahlen liefern die besten Ergebnisse? Schätzt die Güte der (geänderten) Heuristik ein.*
Lösungsvorschlag: Wir wählen exemplarisch 10000 und testen 5 Kombinationen:
Ideen hinter den Zahlen in Spaltenreihenfolge:
Zahlen aus Hinweis verwenden - Regel mit wenig Zufall - Zahlen, die weit voneinander entfernt sind - größere Zahlen verwenden - Einfach etwas ausprobiert

mit den Zahlen:	0.7 ; 0.1	0.56 ; 0.55	0.9 ; 0.1	0.9 ; 0.7	0.8 ; 0.3
richtige Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Hetero:	7127	7371	7214	7645	7325
Geschätzt Homo und ist Homo:	1361	1371	1027	870	1227
falsche Vorhersagen:	•	•	•	•	•
Geschätzt Hetero und ist Homo:	666	606	988	1007	839
Geschätzt Homo und ist Hetero:	846	652	771	478	609

Beantwortung der Frage: Mit den Zahlen 0.7 und 0.1 können ca. $84,88\%$, mit den Zahlen 0.56 und 0.55 können ca. $87,42\%$, mit den Zahlen 0.9 und 0.1 können ca. $82,41\%$, mit den Zahlen 0.9 und 0.7 können ca. $85,15\%$ und mit den Zahlen 0.8 und 0.3 können ca. $85,52\%$ der Nutzer richtig eingeschätzt werden. Hier sind die Zahlen 0.56 und 0.55 am

besten. Auch bei dieser Regel sollte der Zufall am besten eine geringe Rolle spielen. Eine Steigerung ist hier im Vergleich zu den Zahlen 0.2 und 0.1 (zwischen 82% und 84%) zu sehen.

- **in Besprechung: Zusammenfassender Vergleich der fünf Heuristiken anhand der Anzahl richtiger Einschätzungen:** Regel 3 ist am schlechtesten. Regel 1 und 2 schätzen mit entsprechenden Zahlen um die 85% der Nutzer richtig ein. Dabei scheint Regel 1 im Ganzen betrachtet etwas bessere Ergebnisse zu liefern als Regel 2. Regel 5 erbringt bis zu 87% richtige Einschätzungen. Die beste Regel ist hier Regel 4 mit bis zu 91,77% richtigen Einschätzungen.

- **Hinweise zur Besprechung:** Der Vergleich der Ergebnisse sollte den Schülerinnen und Schülern sehr aufwendig und schwer fallen. Deshalb wird eine gemeinsame, einheitliche Größe benötigt! Zudem kann man feststellen, dass die Anzahl richtiger Vorhersagen bei der Testung weniger Nutzer (z.B. 100) häufig besser ausfällt. Der Grund dafür ist, dass hier der Einfluss der konkreten Stichprobe größer ist als bei höheren Zahlen.
  Wurde eine Regel von den Schülerinnen und Schülern nicht bearbeitet, so werden die Ergebnisse von den Dozenten ergänzt.

- **Code und Ausgabe:**

```
%%      Heuristiken auf viele Personen anwenden

%% mit Regel 1:
clc
N = 10000;
params = [5 2];

% Personen einschätzen:
heuristicTester_EinstiegRegel12(@Heuristic_Regel1params, Users, N, params)


_____________________________
Wir haben diese Heuristik genutzt:

Heuristic_Regel1params

mit den Zahlen:           5
und:      2

Größe der Stichprobe:    10000

Ergebnisse:

richtige Vorhersagen:

Geschätzt Hetero und ist Hetero:          7988
Geschätzt Homo und ist Homo:      426

falsche Vorhersagen:

Geschätzt Hetero und ist Homo:    1563
Geschätzt Homo und ist Hetero:    23
```

```
%% mit Regel 3:
clc
N = 10000;

% Personen einschätzen:
heuristicTester_EinstiegRegel3(@Heuristic_Regel3, Users, N)


_____________________________
Wir haben diese Heuristik genutzt:
```



```
Heuristic_Regel3

Größe der Stichprobe:    10000

Ergebnisse:

richtige Vorhersagen:

Eingeschätzt Hetero und ist Hetero:       71
Eingeschätzt Homo und ist Homo:    1960

falsche Vorhersagen:

Eingeschätzt Hetero und ist Homo:         20
Eingeschätzt Homo und ist Hetero:         7949


_____________________________
```

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Arbeitsblatt 2: Heuristiken Güte bewerten

Schritt 4 - Einheitliche Größe zur Bewertung der Heuristiken

Wir lernen nun eine Größe kennen, die uns die Bewertung der Heuristiken erleichtern wird. Es ist eine Zahl, die **Cohen's Kappa** genannt wird. Sie wird in der Wissenschaft genutzt, um Übereinstimmungen zu messen. Berechnet wird sie so:

$$\kappa = \frac{\text{Anteil richtiger Vorhersagen der Heuristik} - \text{Anteil richtiger Vorhersagen mit Hilfe des Zufalls}}{1 - \text{Anteil richtiger Vorhersagen mit Hilfe des Zufalls}}$$

a) Wie ihr seht, spielt hier der Vergleich mit dem Zufall eine wichtige Rolle. Um zu beurteilen, ob eine Vorhersage besser ist als der Zufall, müssen wir herausfinden, wie hoch die **Wahrscheinlichkeit einer richtigen Vorhersage mit Hilfe des Zufalls** bzw. der *Anteil richtiger Vorhersagen mit Hilfe des Zufalls* ist. Das sind zwei Bezeichnung für dieselbe Zahl.

- Eine *Person wird mit Hilfe des Zufalls richtig vorhergesagt*, wenn sie
  - heterosexuell ist und als heterosexuell klassifiziert wurde oder
  - homosexuell ist und als homosexuell klassifiziert wurde.
- *Bestimmt die Wahrscheinlichkeit, dass eine Person mit Hilfe des Zufalls richtig klassifiziert wird mit der Information, dass* 19.85% *der Nutzer homosexuell sind.* *Es gibt Hilfekarten 1-3! Nutzt Matlab als Taschenrechner.*

Zur Erinnerung: *„Der Zufall entscheidet"* bedeutet: 19.85% unserer Nutzer sind homosexuell. Wir sagen mit einer Wahrscheinlichkeit von 19.85% „der Nutzer ist homosexuell" und mit einer Wahrscheinlichkeit von 80.15% „der Nutzer ist heterosexuell".

- Zur Überprüfung eures Ergebnisses gebt euer Ergebnis im Abschnitt „Übereinstimmung durch Zufall" (Z. 63-69) als Kommazahl (zwischen 0 und 1, z.B. 0.5) oder als Rechnung (Achtung! Komma als Punkt eingeben!) in Z. 65 ein. Um es zu überprüfen, drückt auf *Run Section.*
- Tragt die richtige Antwort auf eurem Antwortzettel ein.

b) Wir schauen uns das Übereinstimmungsmaß **Cohen's Kappa** nun genauer an:

- *Welche Zahlen müssen im Kasten „Eigenschaften von Kappa" an den Stellen a1, a2, a3 und a4 stehen, damit die Aussagen stimmen?* *Es gibt Hilfekarten 4 und 5!*

Eigenschaften von Kappa:

- Ist Kappa nahe **a1**, so ist die Heuristik so gut wie der Zufall. Der Anteil richtiger Vorhersagen der Heuristik ist ungefähr so groß wie der Anteil richtiger Vorhersagen mit Hilfe des Zufalls.
- Ist Kappa nahe **a2**, so bestimmt die Heuristik nahezu alle sexuellen Orientierungen richtig. Der Anteil richtiger Vorhersagen der Heuristik ist ungefähr gleich 1. (**bitte wenden**)

- Ist Kappa kleiner als **a3**, so ist die Heuristik schlechter als der Zufall. Der Anteil richtiger Vorhersagen der Heuristik ist kleiner als der Anteil richtiger Vorhersagen mit Hilfe des Zufalls.
- Ist Kappa größer als **a4**, so ist die Heuristik besser als der Zufall. Der Anteil richtiger Vorhersagen der Heuristik ist größer als der Anteil richtiger Vorhersagen mit Hilfe des Zufalls.

- Gebt eure Lösung in Abschnitt „Kappa betrachten" (Z. 70-80) anstelle von *NaN* ein und überprüft eure Lösung durch Drücken auf *Run Section.*

c) Nun wollen wir die Güte unserer Heuristiken mit **Cohen's Kappa** einschätzen.

- Schaut euch den Abschnitt „Heuristiken vergleichen" (Z. 81-121) an.
- Sucht euch eine Regel aus, die ihr untersuchen wollt, und geht in den entsprechenden Abschnitt. Führt dann die zugehörigen Schritte durch:
  - **Regel 1** (Z. 83-90):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 85) fest (am besten die „große" Zahl aus Schritt 3). Tragt sie auf eurem Antwortzettel ein. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und den Wert Kappa. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 86 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert auch hier die genutzten Zahlen und den Wert Kappa.
  - **Regel 2** (Z. 91-98):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 93) fest (am besten die „große" Zahl aus Schritt 3). Tragt sie auf eurem Antwortzettel ein. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und den Wert Kappa. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 94 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert auch hier die genutzten Zahlen und den Wert Kappa.
  - **Regel 3** (Z. 99-105):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 101) fest (am besten die „große" Zahl aus Schritt 3). Tragt sie auf eurem Antwortzettel ein. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert das Ergebnis.
  - **Regel 4** (Z. 106-113):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 108) fest (am besten die „große" Zahl aus Schritt 3). Tragt sie auf eurem Antwortzettel ein. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und den Wert Kappa. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 109 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert auch hier die genutzten Zahlen und den Wert Kappa.
  - **Regel 5** (Z. 114-121):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 116) fest (am besten die „große" Zahl aus Schritt 3). Tragt sie auf eurem Antwortzettel ein. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und den Wert Kappa. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 117 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert auch hier die genutzten Zahlen und den Wert Kappa.
- *Wie gut kann die sexuelle Orientierung vorhergesagt werden? Vergleicht ggf. die Werte für Kappa und findet die beste Regel. Entsprechen die Ergebnisse euren Vermutungen aus den Aufgaben zuvor? Tragt eure Antwort auf dem Antwortzettel ein.*

Zusatzaufgabe - Weitere Heuristiken bewerten

Testet weitere Heuristiken, indem ihr die Anweisungen zu den Regeln (Abschnitt oben) befolgt.

Zusatzaufgabe - Weitere Größen zur Bewertung der Heuristiken

Fragt nach einer Zusatzaufgabe.

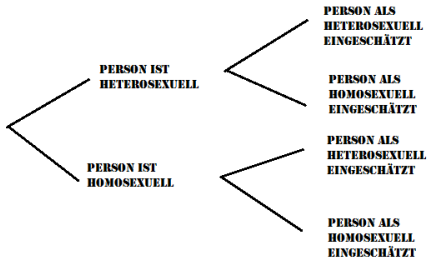
Wir besprechen nun eure Ergebnisse!

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Hilfekarte 1 (zu Arbeitsblatt 2 Schritt 4)

Ihr könnt euch die Situation wie folgt veranschaulichen:



a) Beschriftet die Striche mit den jeweiligen Wahrscheinlichkeiten.

b) Bestimmt die Wahrscheinlichkeiten der vier Pfade.

c) Welche Pfade beschreiben richtige Vorhersagen?

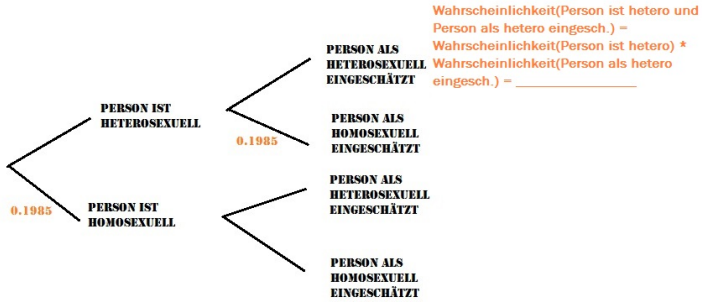
d) Bestimmt nun die gesuchte Wahrscheinlichkeit.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Hilfekarte 2 (zu Arbeitsblatt 2 Schritt 4)

Ihr könnt euch die Situation wie folgt veranschaulichen:



a) Beschriftet die Striche mit den jeweiligen Wahrscheinlichkeiten.

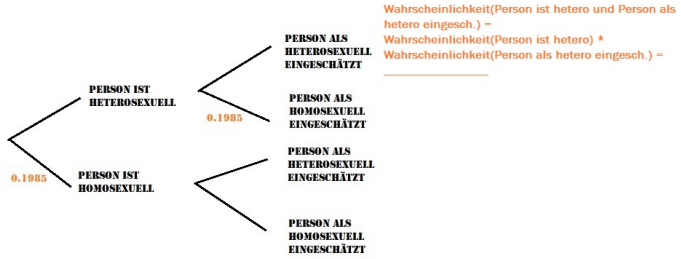
b) Bestimmt die Wahrscheinlichkeiten der vier Pfade.

c) Welche Pfade beschreiben richtige Vorhersagen?

d) Bestimmt nun die gesuchte Wahrscheinlichkeit.

Hilfekarte 3 (zu Arbeitsblatt 2 Schritt 4)

Ihr könnt euch die Situation wie folgt veranschaulichen:



Wahrscheinlichkeit(Person ist hetero und Person als hetero eingesch.) =
Wahrscheinlichkeit(Person ist hetero) *
Wahrscheinlichkeit(Person als hetero eingesch.) =

Wahrscheinlichkeit(richtige Vorhersage) =
Wahrscheinlichkeit(Person ist hetero und Person als hetero eingeschätzt) +
Wahrscheinlichkeit(Person ist homo und Person als homo eingeschätzt) =

a) Beschriftet die Striche mit den jeweiligen Wahrscheinlichkeiten.

b) Bestimmt die Wahrscheinlichkeiten der vier Pfade.

c) Welche Pfade beschreiben richtige Vorhersagen?

d) Bestimmt nun die gesuchte Wahrscheinlichkeit.

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Hilfekarte 4 (zu Arbeitsblatt 2 Schritt 4)

Wir schauen uns die Definition gemeinsam an:

$$\kappa = \frac{\text{Anteil richtiger Vorhersagen der Heuristik} - \text{Anteil richtiger Vorhersagen mit Hilfe des Zufalls}}{1 - \text{Anteil richtiger Vorhersagen mit Hilfe des Zufalls}}$$

- Wenn die Heuristik so gut ist wie der Zufall, dann gilt:

  $$\text{Anteil richtiger Vorhersagen der Heuristik} = \text{Anteil richtiger Vorhersagen mit Hilfe des Zufalls}$$

  Welche Zahl steht dann oben im Bruch (im Zähler)? Welche Zahl nimmt dann der gesamte Bruch an?

- Wenn die Heuristik alle sexuellen Orientierungen richtig bestimmt, dann gilt:

  $$\text{Anteil richtiger Vorhersagen der Heuristik} = 1$$

  Oben und unten im Bruch stehen also die gleichen Zahlen. *Welche Zahl nimmt dann der gesamte Bruch an?*

- Wenn die Heuristik schlechter ist als der Zufall, dann gilt:

  $$\text{Anteil richtiger Vorhersagen der Heuristik} < \text{Anteil richtiger Vorhersagen mit Hilfe des Zufalls}$$

  Ist die Zahl oben im Bruch (im Zähler) negativ oder positiv? Ist dann der gesamte Bruch negativ oder positiv?

- Wenn die Heuristik besser ist als der Zufall, dann gilt:

  $$\text{Anteil richtiger Vorhersagen der Heuristik} > \text{Anteil richtiger Vorhersagen mit Hilfe des Zufalls}$$

  Ist die Zahl oben im Bruch (im Zähler) negativ oder positiv? Ist dann der gesamte Bruch negativ oder positiv?

Hilfekarte 5 (zu Arbeitsblatt 2 Schritt 4)

Die Lücken müssen mit den folgenden Antworten gefüllt werden: 0, 0, 0, 1. Ordnet sie zu.

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 2: Antwortblatt

Schritt 4 - Einheitliche Größe zur Bewertung der Heuristiken

a) Die **Wahrscheinlichkeit einer richtigen Vorhersage mit Hilfe des Zufalls** bzw. der *Anteil richtiger Vorhersagen mit Hilfe des Zufalls* beträgt

.....

c) Vergleich mit Cohen's Kappa:

mit Regel:

Größe der Stichprobe:

- Für die Regeln 1,2,4,5:

mit den Zahlen:					
Kappa:					

Wie gut kann die sexuelle Orientierung vorhergesagt werden? Vergleicht die Werte für Kappa und findet die beste Regel. Entsprechen die Ergebnisse euren Vermutungen aus den Aufgaben zuvor?

- Für Regel 3:

  Kappa =.....

  Wie gut kann die sexuelle Orientierung vorhergesagt werden? Entsprechen die Ergebnisse euren Vermutungen aus den Aufgaben zuvor?

Arbeitsblatt 2: Besprechung

Schritt 4 - Einheitliche Größe zur Bewertung der Heuristiken

a) Die **Wahrscheinlichkeit einer richtigen Vorhersage mit Hilfe des Zufalls** bzw. der *Anteil richtiger Vorhersagen mit Hilfe des Zufalls* beträgt

.....

c) Vergleich mit Cohen's Kappa:

Größe der Stichprobe(n):

Regel	1	2	3	4	5
Kappa					
mit den Zahlen			•		

Wie gut kann die sexuelle Orientierung vorhergesagt werden? Vergleicht die Werte für Kappa und findet die beste Regel. Entsprechen die Ergebnisse euren Vermutungen aus den Aufgaben zuvor?

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 2: Lösungsvorschlag

Schritt 4 - Einheitliche Größe zur Bewertung der Heuristiken

a) Wahrscheinlichkeit einer richtigen Vorhersage mit Hilfe des Zufalls:

Lösung: Die Vorhersage und ihre Überprüfung entspricht einem zweistufigem Zufallsexperiment wie im Bild. Eine richtige Klassifikation entspricht also den Pfaden

- Person ist heterosexuell -> Person als heterosexuell eingeschätzt und
- Person ist homosexuell -> Person als homosexuell eingeschätzt.

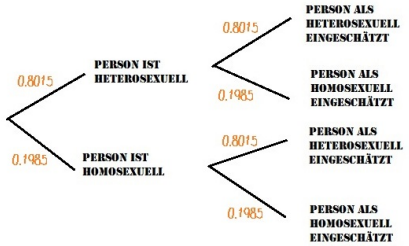


Abbildung 1: **Klassifikation durch Zufall mit Überprüfung im Baumdiagramm.**

Mit Hilfe der **Pfadregel** (Multipliziere die Wahrscheinlichkeiten entlang des Pfades) ergeben sich die Wahrscheinlichkeiten

- Person ist heterosexuell -> Person als heterosexuell eingeschätzt: $0.8015 \cdot 0.8015 \approx 0.6424$
- Person ist homosexuell -> Person als homosexuell eingeschätzt: $0.1985 \cdot 0.1985 \approx 0.0394$

Da beide Pfade eine richtige Vorhersage liefern, erhalten wir die Wahrscheinlichkeit für eine richtige Vorhersage durch Summation (**Summenregel**) der Wahrscheinlichkeiten zu

$$0.6424 + 0.0394 = 0.6818.$$

Hinweis zur Besprechung:

Die Schülerinnen und Schüler sollen hier ihren Lösungsweg z.B. mit Hilfe eines Baumdiagramms oder auf andere nachvollziehbare Art und Weise den Mitschülern präsentieren.

Code und Ausgabe:

```
%%                    Heuristiken – Güte bewerten

%%      Übereinstimmung durch Zufall
clc
pe = 0.6818;

% Prüfe die Eingabe:
Pruefepe(pe);
```

```
Gut! Eure Antwort ist richtig
```

oder als Rechnung

```
%%                      Heuristiken - Güte bewerten

%%      Übereinstimmung durch Zufall
clc
pe = (0.8015*0.8015)+(0.1985*0.1985);

% Prüfe die Eingabe:
Pruefepe(pe);
```

```
0.6818

Gut! Eure Antwort ist richtig
```

mit Feedback bei Fehler

```
%%                      Heuristiken - Güte bewerten

%%      Übereinstimmung durch Zufall
clc
pe = 0.6;

% Prüfe die Eingabe:
Pruefepe(pe);
```

```
Die Antwort ist falsch. Probiert es erneut. Holt euch, wenn ihr möchtet, einen Hilfezettel.
```

oder

```
%%                      Heuristiken - Güte bewerten

%%      Übereinstimmung durch Zufall
clc
pe = 0.68;

% Prüfe die Eingabe:
Pruefepe(pe);
```

```
Die Antwort ist fast richtig.
Vermutlich habt ihr gerundet.
Eine exaktere Antwort ist 0.6818.
```

b) Cohen's Kappa:
Aufgabe: *Welche Zahlen müssen in den Lückentext im Kasten „Eigenschaften von Kappa" an den Stellen a1, a2, a3 und a4 stehen, damit die Aussagen stimmen?*

Lösung:

- Ist Kappa nahe **a1=0**, so ist die Heuristik so gut wie der Zufall. Der Anteil richtiger Vorhersagen der Heuristik ist ungefähr so groß wie der Anteil richtiger Vorhersagen mit Hilfe des Zufalls.
- Ist Kappa nahe **a2=1**, so bestimmt die Heuristik nahezu alle sexuellen Orientierungen richtig. Der Anteil richtiger Vorhersagen der Heuristik ist ungefähr gleich 1.
- Ist Kappa kleiner als **a3=0**, so ist die Heuristik schlechter als der Zufall. Der Anteil richtiger Vorhersagen der Heuristik ist kleiner als der Anteil richtiger Vorhersagen mit Hilfe des Zufalls.
- Ist Kappa größer als **a4=0**, so ist die Heuristik besser als der Zufall. Der Anteil richtiger Vorhersagen der Heuristik ist größer als der Anteil richtiger Vorhersagen mit Hilfe des Zufalls.

Code und Ausgabe:

```
%%      Kappa betrachten
clc

a1 = 0;
a2 = 1;
a3 = 0;
a4 = 0;

% Prüfe die Eingabe:
PruefeLueckentext(a1,a2,a3,a4);
```

```
Gut! Eure Antwort ist richtig
```

Bei Fehler:

```
%%      Kappa betrachten
clc

a1 = 1;
a2 = 1;
a3 = 0;
a4 = 0;

% Prüfe die Eingabe:
PruefeLueckentext(a1,a2,a3,a4);
```

```
Die Antwort ist fast richtig.
Ihr seid auf dem richtigen Weg!
Schaut euch nochmal die Antworten an.
a2 ist richtig.
a3 ist richtig.
a4 ist richtig.
```

c) Vergleich mit Cohen's Kappa:

Aufgabe: ***Wie gut kann die sexuelle Orientierung vorhergesagt werden?*** *Vergleicht ggf. die Werte für Kappa und findet die beste Regel. Entsprechen die Ergebnisse euren Vermutungen aus den Aufgaben zuvor?*

Lösungsvorschlag: Wir betrachten wieder 10000 Nutzer und die Zahlen von Schritt 3.

- **Regel 1:**

mit den Zahlen:	5 und: 2	3 und: 1	3 und: 2	9 und: 2	10 und: 7	1 und: 0
Kappa =	0.49	0.56	0.54	0.44	0.43	0.53

  Vergleich: Die Zahlen 3 und 1 sind hier am besten mit Kappa = 0.56 . Diese gehörten auch im Schritt zuvor zu den besten Heuristiken.

- **Regel 2:**

mit den Zahlen:	5 und: 3	22 und: 2	5 und: 4	50 und: 7	45 und: 39	150 und: 9
Kappa =	0.45	0.48	0.47	0.45	0.46	0.40

  Vergleich: Die Zahlen 22 und 2 sind hier am besten mit Kappa = 0.48. Diese gehörten auch im Schritt zuvor zu den besten Heuristiken.

- **Regel 3:** Kappa = -1.50

- **Regel 4:**

mit den Zahlen:	0.2 ; 0.1	0.7 ; 0.1	0.56 ; 0.55	0.9 ; 0.1	0.9 ; 0.7	0.8 ; 0.3
Kappa =	0.63	0.69	0.75	0.65	0.67	0.70

  Vergleich: Die Zahlen 0.56 und 0.55 sind hier am besten mit Kappa = 0.75. Diese waren auch im Schritt zuvor insgesamt die beste Wahl.

- **Regel 5:**

mit den Zahlen:	0.3; 0.1	0.7 ; 0.1	0.56 ; 0.55	0.9 ; 0.1	0.9 ; 0.7	0.8 ; 0.3
Kappa =	0.47	0.52	0.61	0.42	0.48	0.55

Vergleich: Die Zahlen 0.56 und 0.55 sind hier am besten mit Kappa = 0.61. Diese waren auch im Schritt zuvor insgesamt die beste Wahl.

- **Zusammenfassender Vergleich:** Den besten Kappawert von 0.75 liefert Regel 4 mit den Parametern 0.56 und 0.55. Den zweitbesten Wert erbringt Regel 5 mit 0.61 und den Parametern 0.56 und 0.55. Regel 1 liefert hier den drittgrößten Wert für Kappa (0.56) mit den Parametern 3 und 1. Regel 2 ergibt den Wert 0.48 für Kappa mit den Zahlen 22 und 2. Die Ergebnisse weichen von den vorherigen in der Parameterwahl kaum ab. Bei den relativen Anteilen (Regel 4 und 5) scheinen die guten Parameter gleich zu sein. Hier sind alle vier genannten Regeln besser als der Zufall, da Kappa stets positiv ist. Außerdem sind die Regeln deutlich besser als der Zufall, da die Kappawerte nicht nahe 0, sondern mittig zwischen 0 und 1 liegen bzw. eher größer als 0.5 sind. Regel 3 liefert einen negativen Wert für Kappa (-1.5). Die Heuristik ist also schlechter als der Zufall.

- **Code und Ausgabe:**

```
%%      Heuristiken vergleichen

%% mit Regel 1:
clc
N = 10000;
params=[5 2];

% Kappa bestimmen:
heuristicTester_kappaRegel12(@Heuristic_Regel1params,Users,N, params);
```

```
-----------------------------------
Wir haben diese Heuristik genutzt:

Heuristic_Regel1params

mit den Zahlen:          5
und:     2

Größe der Stichprobe:    10000

Kappa =              0.51
-----------------------------------
```

```
%% mit Regel 3:
clc
N = 10000;

% Kappa bestimmen:
heuristicTester_kappaRegel3(@Heuristic_Regel3,Users,N);
```

```
-----------------------------------
Wir haben diese Heuristik genutzt:

Heuristic_Regel3

Größe der Stichprobe:    10000

Kappa =             -1.49
-----------------------------------
```


CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

CAMMP | RWTH AACHEN UNIVERSITY

Zusatzblatt 1 (zu Arbeitsblatt 2 Schritt 4)

Um die Güte von verschiedenen Heuristiken einfach vergleichen zu können, brauchen wir mathematische Größen. Neben Cohen's Kappa kann man auch die **Trefferquote** und die **Genauigkeit** nutzen. Sie werden für Vorhersagen über Homosexuelle und Heterosexuelle einzeln bestimmt. Aber was bedeuten die Begriffe?

Trefferquote
Die Trefferquote für eine Gruppe (Homosexuelle oder Heterosexuelle) ist gleich dem Anteil der richtig Klassifizierten einer Gruppe. Wir berechnen sie so:

$$\text{Trefferquote der Vorhersage für Homosexuelle} = \frac{\text{Anzahl Nutzer, die homosexuell sind und als homosexuell eingeschätzt wurden}}{\text{Anzahl der Homosexuellen}},$$

$$\text{Trefferquote der Vorhersage für Heterosexuelle} = \frac{\text{Anzahl Nutzer, die heterosexuell sind und als heterosexuell eingeschätzt wurden}}{\text{Anzahl der Heterosexuellen}}.$$

Genauigkeit
Die Genauigkeit für eine Gruppe (Homosexuelle oder Heterosexuelle) ist gleich dem Anteil der richtig Klassifizierten einer Gruppe unter allen so Klassifizierten. Wir berechnen sie so:

$$\text{Genauigkeit der Vorhersage für Homosexuelle} = \frac{\text{Anzahl Nutzer, die homosexuell sind und als homosexuell eingeschätzt wurden}}{\text{Anzahl Nutzer, die als homosexuell eingeschätzt wurden}},$$

$$\text{Genauigkeit der Vorhersage für Heterosexuelle} = \frac{\text{Anzahl Nutzer, die heterosexuell sind und als heterosexuell eingeschätzt wurden}}{\text{Anzahl Nutzer, die als heterosexuell eingeschätzt wurden}}.$$

a) Auch hier schauen wir uns die beiden Größen genauer an:

- *Welche Zahlen müssen im Kasten „Eigenschaften von Trefferquote und Genauigkeit" an den Stellen a1, a2, a3 und a4 stehen, damit die Aussagen stimmen? Es gibt Hilfekarten 6 und 7!*

Eigenschaften von Trefferquote und Genauigkeit:

- Ist die Trefferquote nahe **a1**, so ist die Anzahl der richtig Klassifizierten niedrig. Die Heuristik hat damit kaum Nutzer richtig klassifiziert.
- Ist die Trefferquote nahe **a2**, so ist die Anzahl der richtig Klassifizierten fast so groß wie die Anzahl der Nutzer der Gruppe. Die Heuristik hat damit nahezu alle Nutzer richtig klassifiziert.
- Ist die Genauigkeit nahe **a3**, so ist die Anzahl der richtig Klassifizierten niedrig. Die Heuristik hat damit kaum Nutzer richtig klassifiziert.
- Ist die Genauigkeit nahe **a4**, so ist die Anzahl der richtig Klassifizierten fast so groß wie die Anzahl der so Klassifizierten. Die Heuristik hat damit nahezu alle Nutzer richtig klassifiziert.

- Öffnet das Programm „ZusatzGuetebewerten“. Gebt eure Lösung in Abschnitt „Trefferquote und Genauigkeit betrachten“ (Z. 4-14) anstelle von *NaN* ein und überprüft eure Lösung durch Drücken auf *Run Section*.

b) Nun wollen wir die Güte unserer Heuristiken mit Hilfe der Trefferquote bzw. der Genauigkeit einschätzen.

- Schaut euch den Abschnitt „Trefferquote und Genauigkeit der Heuristiken“ (Z. 15-55) an.
- Sucht euch eine Regel aus, die ihr untersuchen wollt, und geht in den entsprechenden Abschnitt. Führt dann die zugehörigen Schritte durch:
  - **Regel 1** (Z. 17-24):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 19) fest (am besten die „große“ Zahl aus Schritt 3). Notiert sie auf dem Antwortblatt. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 20 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle.
  - **Regel 2** (Z. 25-32):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 27) fest (am besten die „große“ Zahl aus Schritt 3). Notiert sie auf dem Antwortblatt. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 28 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle.
  - **Regel 3** (Z. 33-39):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 35) fest (am besten die „große“ Zahl aus Schritt 3). Notiert sie auf dem Antwortblatt. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die Ergebnisse in der Tabelle.
  - **Regel 4** (Z. 40-47):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 42) fest (am besten die „große“ Zahl aus Schritt 3). Notiert sie auf dem Antwortblatt. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 43 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle.
  - **Regel 5** (Z. 48-55):
    Legt eine Größe für die Stichprobe (Zahl für N, Z. 50) fest (am besten die „große“ Zahl aus Schritt 3). Notiert sie auf dem Antwortblatt. Führt den Abschnitt durch Drücken auf *Run Section* aus. Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle. Testet verschiedene Zahlen, indem ihr die Zahlen in Zeile 51 (params) ändert und den Abschnitt ausführt (*Run Section*). Notiert die genutzten Zahlen und die Ergebnisse in der Tabelle.
- *Wie gut kann die sexuelle Orientierung vorhergesagt werden? Wie genau können Hetero- und Homosexuelle vorhergesagt werden? Wie gut können Hetero- und Homosexuelle getroffen werden?*

Fragt nach den Lösungen, um eure Ergebnisse zu überprüfen!

Hilfekarte 6 zu Zusatzblatt 1 (zu Arbeitsblatt 2 Schritt 4)

Die Trefferquote und die Genauigkeit sind beides Anteile. Anteile nehmen Werte zwischen 0 und 1 an. Wir betrachten beispielhaft Vorhersagen zu Homosexuellen:

Denkt über die folgenden Fragen nach!

Welchen Wert nimmt die Trefferquote der Vorhersage für Homosexuelle an,

- wenn die **Anzahl der Nutzer, die homosexuell sind und als homosexuell eingeschätzt** wurden,
- gleich der **Anzahl der Homosexuellen** ist?

Tritt diese Situation für eine gute oder eine schlechte Heuristik ein?

Welchen Wert nimmt die Genauigkeit der Vorhersage für Homosexuelle an,

- wenn die **Anzahl der Nutzer, die homosexuell sind und als homosexuell eingeschätzt** wurden,
- gleich der **Anzahl der Nutzer, die als heterosexuell eingeschätzt wurden,** ist?

Tritt diese Situation für eine gute oder eine schlechte Heuristik ein?

Hilfekarte 7 zu Zusatzblatt 1 (zu Arbeitsblatt 2 Schritt 4)

Die Lücken müssen mit den folgenden Werten gefüllt werden: 0, 0, 1, 1. Ordnet sie zu.

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Zusatzblatt 1: Antwortblatt

Trefferquote und Genauigkeit der Heuristiken

- mit Regel:
- Größe der Stichprobe:

Für Regel 1,2,4,5:

Zahlen:					
Genauigkeit Hetero					
Genauigkeit Homo					
Trefferquote Hetero					
Trefferquote Homo					

Für Regel 3:

Genauigkeit Hetero					
Genauigkeit Homo					
Trefferquote Hetero					
Trefferquote Homo					

Wie gut kann die sexuelle Orientierung vorhergesagt werden? Wie genau können Hetero- und Homosexuelle vorhergesagt werden? Wie gut können Hetero- und Homosexuelle getroffen werden?

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Zusatzblatt 1: Lösungsvorschlag

a) Trefferquote und Genauigkeit und ihre Eigenschaften

Aufgabe: *Welche Zahlen müssen im Lückentext im Kasten „Eigenschaften von Trefferquote und Genauigkeit" an den Stellen a1, a2, a3 und a4 stehen, damit die Aussagen stimmen?*

Lösung:

Eigenschaften von Trefferquote und Genauigkeit:

- Ist die Trefferquote nahe **a1 = 0**, so ist die Anzahl der richtig Klassifizierten niedrig. Die Heuristik hat damit kaum Nutzer richtig klassifiziert.
- Ist die Trefferquote nahe **a2 = 1**, so ist die Anzahl der richtig Klassifizierten fast so groß wie die Anzahl der Nutzer der Gruppe. Die Heuristik hat damit nahezu alle Nutzer richtig klassifiziert.
- Ist die Genauigkeit nahe **a3 = 0**, so ist die Anzahl der richtig Klassifizierten niedrig. Die Heuristik hat damit kaum Nutzer richtig klassifiziert.
- Ist die Genauigkeit nahe **a4 = 1**, so ist die Anzahl der richtig Klassifizierten fast so groß wie die Anzahl der so Klassifizierten. Die Heuristik hat damit nahezu alle Nutzer richtig klassifiziert.

Code und Ausgabe:

mit richtiger Antwort:

```
%%                          Zusatzblatt 1: Heuristiken Güte bewerten

%%      Trefferquote und Genauigkeit betrachten
clc

a1 = 0;
a2 = 1;
a3 = 0;
a4 = 1;

% Prüfe die Eingabe:
PruefeLueckentext2(a1,a2,a3,a4);
```

```
Gut! Eure Antwort ist richtig
```

bei Fehler:

```
%%                          Zusatzblatt 1: Heuristiken Güte bewerten

%%      Trefferquote und Genauigkeit betrachten
clc

a1 = 1;
a2 = 1;
a3 = 0;
a4 = 1;

% Prüfe die Eingabe:
PruefeLueckentext2(a1,a2,a3,a4);
```

```
Die Antwort ist fast richtig.
Ihr seid auf dem richtigen Weg!
Schaut euch nochmal die Antworten an.
a2 ist richtig.
a3 ist richtig.
a4 ist richtig.
```


b) Heuristiken mit Hilfe der Trefferquote bzw. der Genauigkeit einschätzen:

Aufgabe: *Wie gut kann die sexuelle Orientierung mit den Regeln vorhergesagt werden? Welche Heuristik hat die höchste Genauigkeit und welche die höchste Trefferquote für Hetero- und Homosexuelle?*

Lösungsvorschlag: Wir nutzen wieder 10000 Nutzer und die Zahlen von Schritt 3.

- **Regel 1:**

mit den Zahlen:	5 und: 2	3 und: 1	3 und: 2	9 und: 2	10 und: 7	1 und: 0
Genauigkeit						
Heterosexuelle:	0.84	0.87	0.86	0.83	0.81	0.97
Homosexuelle:	0.92	0.87	0.90	0.92	0.99	0.58
Trefferquote						
Heterosexuelle:	1.00	0.99	0.99	1.00	1.00	0.84
Homosexuelle:	0.23	0.38	0.34	0.13	0.06	0.88

Vergleich:

- Genauigkeit Heterosexuelle:
  Die höchste Genauigkeit hat hier mit einem Wert von 0.97 die Regel mit den Parametern 1 und 0.
- Genauigkeit Homosexuelle:
  Die höchste Genauigkeit hat hier mit einem Wert von 0.99 die Regel mit den Parametern 10 und 7.
- Trefferquote Heterosexuelle:
  Die höchste Trefferquote hat hier mit einem Wert von 1.00 die Regel mit den Parametern 5 und 2; 9 und 2; 10 und 7.
- Trefferquote Homosexuelle:
  Die höchste Trefferquote hat hier mit einem Wert von 0.88 die Regel mit den Parametern 1 und 0.

- **Regel 2:**

mit den Zahlen:	5 und: 3	22 und: 2	5 und: 4	50 und: 7	45 und: 39	150 und: 9
Genauigkeit						
Heterosexuelle:	0.89	0.86	0.89	0.83	0.82	0.81
Homosexuelle:	0.56	0.65	0.59	0.77	0.92	0.78
Trefferquote						
Heterosexuelle:	0.88	0.95	0.90	0.98	1.00	0.99
Homosexuelle:	0.58	0.34	0.56	0.20	0.15	0.10

Vergleich:

- Genauigkeit Heterosexuelle:
  Die höchste Genauigkeit hat hier mit einem Wert von 0.89 die Regel mit den Parametern 5 und 3; 5 und 4.
- Genauigkeit Homosexuelle:
  Die höchste Genauigkeit hat hier mit einem Wert von 0.92 die Regel mit den Parametern 45 und 39.
- Trefferquote Heterosexuelle:
  Die höchste Trefferquote hat hier mit einem Wert von 1.00 die Regel mit den Parametern 45 und 39.

– Trefferquote Homosexuelle:
Die höchste Trefferquote hat hier mit einem Wert von 0.58 die Regel mit den Parametern 5 und 3.

- **Regel 3:**

Genauigkeit	Heterosexuelle:	0.86	Homosexuelle:	0.19
Trefferquote	Heterosexuelle:	0.01	Homosexuelle:	0.99

- **Regel 4:**

mit den Zahlen:	0.2 ; 0.1	0.7 ; 0.1	0.56 ; 0.55	0.9 ; 0.1	0.9 ;0.7	0.8 ;0.3
Genauigkeit						
Heterosexuelle:	0.96	0.94	0.95	0.92	0.92	0.93
Homosexuelle:	0.66	0.75	0.80	0.72	0.79	0.77
Trefferquote						
Heterosexuelle:	0.89	0.94	0.95	0.94	0.96	0.95
Homosexuelle:	0.85	0.74	0.77	0.67	0.67	0.70

Vergleich:

– Genauigkeit Heterosexuelle:
Die höchste Genauigkeit hat hier mit einem Wert von 0.96 die Regel mit den Parametern 0.2 und 0.1.

– Genauigkeit Homosexuelle:
Die höchste Genauigkeit hat hier mit einem Wert von 0.80 die Regel mit den Parametern 0.56 und 0.55.

– Trefferquote Heterosexuelle:
Die höchste Trefferquote hat hier mit einem Wert von 0.96 die Regel mit den Parametern 0.9 und 0.7.

– Trefferquote Homosexuelle:
Die höchste Trefferquote hat hier mit einem Wert von 0.85 die Regel mit den Parametern 0.2 und 0.1.

- **Regel 5:**

mit den Zahlen:	0.3 ; 0.1	0.7 ; 0.1	0.56 ; 0.55	0.9 ; 0.1	0.9 ;0.7	0.8 ;0.3
Genauigkeit						
Heterosexuelle:	0.94	0.92	0.92	0.88	0.87	0.90
Homosexuelle:	0.54	0.60	0.67	0.56	0.64	0.65
Trefferquote						
Heterosexuelle:	0.84	0.89	0.91	0.90	0.94	0.92
Homosexuelle:	0.78	0.68	0.70	0.49	0.45	0.59

Vergleich:

– Genauigkeit Heterosexuelle:
Die höchste Genauigkeit hat hier mit einem Wert von 0.94 die Regel mit den Parametern 0.3 und 0.1.

– Genauigkeit Homosexuelle:
Die höchste Genauigkeit hat hier mit einem Wert von 0.67 die Regel mit den Parametern 0.56 und 0.55.

– Trefferquote Heterosexuelle:
Die höchste Trefferquote hat hier mit einem Wert von 0.94 die Regel mit den Parametern 0.9 und 0.7.

– Trefferquote Homosexuelle:
Die höchste Trefferquote hat hier mit einem Wert von 0.78 die Regel mit den Parametern 0.3 und 0.1.

- **Zusammenfassender Vergleich:**

– Genauigkeit Heterosexuelle:
Die höchste Genauigkeit hat hier mit einem Wert von 0.97 die Regel 1 mit den Parametern 1 und 0. Die zweithöchste liefert Regel 4 mit 0.96 (Parameter: 0.2 und 0.1). Die dritthöchste hat Regel 5 mit 0.94 (Parameter: 0.3 und 0.1). Regel 2 hat die zweitschlechteste mit 0.89 (Parameter: 5 und 3; 5 und 4) und Regel 3 die schlechteste mit 0.86. Insgesamt sind jedoch vor allem die ersten drei Werte nahe 1 und die letzten beiden nahe 0.9. Die Genauigkeit für Heterosexuelle ist damit insgesamt hoch.

– Genauigkeit Homosexuelle:
Die höchste Genauigkeit hat hier mit einem Wert von 0.99 die Regel 1 mit den Parametern 10 und 7. Die zweithöchste liefert Regel 2 mit 0.92 (Parameter: 45 und 39). Die dritthöchste und deutlich niedrigere hat Regel 4 mit 0.8 (Parameter: 0.56 und 0.55). Regel 5 hat die zweitschlechteste und wiederum viel niedrigere mit 0.67 (Parameter: 0.56 und 0.55) und Regel 3 die schlechteste mit 0.19. Insgesamt sind die ersten beiden Werte nahe 1. Regel 4 und 5 liefern deutlich niedrigere Werte, die jedoch immer noch über 0.6 liegen. Der Wert 0.19 von Regel 3 ist nahe 0.

– Trefferquote Heterosexuelle:
Die höchste Trefferquote haben hier mit einem Wert von 1.00 die Regeln 1 bzw. 2 mit den Parametern 10 und 7; 5 und 2; 9 und 2 bzw. 45 und 39. Die zweithöchste liefert Regel 4 mit 0.96 (Parameter: 0.9 und 0.7). Die dritthöchste hat Regel 5 mit 0.94 (Parameter: 0.9 und 0.7). Regel 3 liefert die schlechteste mit 0.01. Insgesamt erreichen Regel 1 und 2 den höchst möglichen Wert und bestimmen damit alle sexuellen Orientierungen richtig. Die Werte von Regel 4 und 5 liegen nahe 1. Der Wert 0.01 von Regel 3 ist nahe 0.

– Trefferquote Homosexuelle:
Die höchste Trefferquote hat hier mit einem Wert von 0.99 die Regel 3. Die zweithöchste aber deutliche niedrigere liefert Regel 1 mit 0.88 (Parameter: 1 und 0). Die dritthöchste hat Regel 4 mit 0.85 (Parameter: 0.2 und 0.1). Regel 5 hat die zweitschlechteste mit 0.78 (Parameter: 0.3 und 0.1) und Regel 2 die schlechteste und deutlich niedrigere mit 0.58 (Parameter: 5 und 3). Insgesamt erreicht Regel 3 fast den Maximalwert. Regel 1 und 4 liefern deutlich niedrigere Werte, die aber immer noch recht hoch (zwischen 0.85 und 0.9) sind. Die restlichen Werte liegen zudem immer noch über 0.5. Die Werte sind damit mittel bis sehr hoch.

– Das bedeutet...
je nach Ziel der Untersuchung kann eine passende Heuristik angegeben werden. Gerade die Heuristik 3, die einen negativen Kappawert erzeugt, ist jedoch hervorragend für eine hohe Trefferquote für Homosexuelle geeignet. Insgesamt sind hier auch die Heuristiken mit absoluten Parametern (1 und 2) besser als die mit den relativen (4 und 5). Demnach ergeben andere Vergleichskriterien auch andere Bevorzugungen der Heuristiken.

- **Code und Ausgabe:**

```
%%      Trefferquote und Genauigkeit der Heuristiken

%% mit Regel 1:
clc
N = 10000;
params=[5 2];

% Kappa bestimmen:
heuristicTester_TrefferquoteGenauigkeitRegel12(@Heuristic_Regel1params,Users,N, params);
```

```
_____________________________
Wir haben diese Heuristik genutzt:

Heuristic_Regel1params

mit den Zahlen:          5
und:     2
```

```
Größe der Stichprobe:    10000

Genauigkeit Heterosexuelle:         0.84
Genauigkeit Homosexuelle:           0.95
Trefferquote Heterosexuelle:        1.00
Trefferquote Homosexuelle:          0.25


_____________________________
```

```
%% mit Regel 3:
clc
N = 10000;

% Kappa bestimmen:
heuristicTester_TrefferquoteGenauigkeitRegel3(@Heuristic_Regel3,Users,N);
```

```
_____________________________
Wir haben diese Heuristik genutzt:

Heuristic_Regel3

Größe der Stichprobe:    10000

Genauigkeit Heterosexuelle:         0.70
Genauigkeit Homosexuelle:           0.20
Trefferquote Heterosexuelle:        0.01
Trefferquote Homosexuelle:          0.99


_____________________________
```


CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Arbeitsblatt 3: Kappa verbessern

Schritt 5 - Kappa systematisch verbessern

Bisher habt ihr verschiedene Heuristiken und bei einigen von ihnen sogar verschiedene Formen ausprobiert, die euch als gewinnbringend erschienen. Doch das war ziemlich mühsam. Eine schnellere Möglichkeit wäre, bei Heuristiken mit Zahlen (Regel 1, 2 und 4, 5) systematisch nach optimalen Zahlen zu suchen. Das Wort „optimal“ ist dabei ein Schlüsselwort. Eine systematische Suche kann nämlich mit Hilfe von einer Optimierung geschehen. Wir schauen uns hier ein „systematisches Ausprobieren“ an.

- Jetzt ist Teamarbeit gefragt! Bildet Teams mit drei oder vier Laptops.
- Sucht euch im Team eine Regel (1, 2, 4 oder 5) aus, die ihr optimieren wollt.
- Holt euch das entsprechende Arbeitsblatt und bearbeitet es.

Zusatzaufgabe - Weitere Heuristiken optimieren

Optimiert weitere Heuristiken, indem ihr die entsprechenden Arbeitsblätter bearbeitet.

Wir besprechen nun eure Ergebnisse!

Arbeitsblatt 3 - Schritt 5 - Regel 1

a) Schaut euch den Abschnitt „Kappa verbessern“ (Z. 122-155) an. Arbeitet mit dem Abschnitt „mit Regel 1“ (Z. 124-131).

b) **Gemeinsames Vorgehen:**

- Legt eine **gemeinsame** Größe für die Stichprobe (Zahl für N, Z. 126) fest (eine „große“ Zahl, mindestens 10000, höchstens 56284). Gebt sie alle im Programm ein und notiert sie auf eurem Antwortblatt!
- Gebt **gemeinsam** für jede Zahl ein Intervall vor, indem gesucht werden soll. Schaut euch dazu das Beispiel unten an und ersetzt in Zeile 127 die Ausdrücke NaN. Notiert die Intervalle auf eurem Antwortblatt.

  Beispiel: $range = [2, 7; 4, 6]$; bedeutet, dass für die erste Zahl im Bereich $[2, 7]$ und für die zweite Zahl im Bereich $[4, 6]$ gesucht wird.

  Hinweis: Mehr als 99% der Nutzer haben zwischen 0 und 15 homosexuelle Freunde.

c) **Optimierung durchführen:**

- Führt den Abschnitt durch Drücken auf *Run Section* aus. *Was tut die Optimierung? Beschreibt das Suchprinzip auf eurem Arbeitsblatt.*
- Die Ergebnisse sind der hier gefundene optimale Kappawert und die zugehörigen Zahlen. Notiert sie in der Tabelle.
- Die Ergebnisse sind hier von der zufällig ausgewählten Stichprobe der Größe N abhängig. Wir wollen diese Abhängigkeit verringern. Bei jeder Durchführung wird die Stichprobe neu ausgewählt. Führt daher die Optimierung an jedem Laptop mindestens 5 mal aus. Notiert jeweils den optimalen Kappawert und die zugehörigen Zahlen.
- Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für die am häufigsten auftretenden Zahlen den durchschnittlichen Kappawert. Tragt eure Ergebnisse auf dem Antwortblatt ein.

d) *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 1 vorhergesagt werden? Welche Zahlen sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Zahlen? Vergleicht sie mit euren vorherigen Ergebnissen. Tragt eure Antwort auf dem Antwortzettel ein.*

e) *Welche Probleme oder Vorteile seht ihr bei dieser Art der Optimierung? Tragt eure Antwort auf dem Antwortzettel ein.*

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 3 - Schritt 5 - Regel 2

a) Schaut euch den Abschnitt „Kappa verbessern“ (Z. 122-155) an. Arbeitet mit dem Abschnitt „mit Regel 2“ (Z. 132-139).

b) **Gemeinsames Vorgehen:**

- Legt eine **gemeinsame** Größe für die Stichprobe (Zahl für N, Z. 134) fest (eine „große“ Zahl, mindestens 10000, höchstens 56284). Gebt sie alle im Programm ein und notiert sie auf eurem Antwortblatt!
- Gebt **gemeinsam** für jede Zahl ein Intervall vor, indem gesucht werden soll. Schaut euch dazu das Beispiel unten an und ersetzt in Zeile 135 die Ausdrücke NaN. Notiert die Intervalle auf eurem Antwortblatt.

  Beispiel: $range = [2, 7; 4, 6]$; bedeutet, dass für die erste Zahl im Bereich $[2, 7]$ und für die zweite Zahl im Bereich $[4, 6]$ gesucht wird.

  Hinweis: Mehr als 95% der Nutzer haben zwischen 0 und 35 homosexuelle Freundes Freunde.

c) **Optimierung durchführen:**

- Führt den Abschnitt durch Drücken auf *Run Section* aus. *Was tut die Optimierung? Beschreibt das Suchprinzip auf eurem Arbeitsblatt.*
- Die Ergebnisse sind der hier gefundene optimale Kappawert und die zugehörigen Zahlen. Notiert sie in der Tabelle.
- Die Ergebnisse sind hier von der zufällig ausgewählten Stichprobe der Größe N abhängig. Wir wollen diese Abhängigkeit verringern. Bei jeder Durchführung wird die Stichprobe neu ausgewählt. Führt daher die Optimierung an jedem Laptop mindestens 5 mal aus. Notiert jeweils den optimalen Kappawert und die zugehörigen Zahlen.
- *Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für die am häufigsten auftretenden Zahlen den durchschnittlichen Kappawert. Tragt eure Ergebnisse auf dem Antwortblatt ein.*

d) *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 2 vorhergesagt werden? Welche Zahlen sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Zahlen? Vergleicht sie mit euren vorherigen Ergebnissen. Tragt eure Antwort auf dem Antwortzettel ein.*

e) *Welche Probleme oder Vorteile seht ihr bei dieser Art der Optimierung? Tragt eure Antwort auf dem Antwortzettel ein.*

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 3 - Schritt 5 - Regel 4

a) Schaut euch den Abschnitt „Kappa verbessern“ (Z. 122-155) an. Arbeitet mit dem Abschnitt „mit Regel 4“ (Z. 140-147).

b) **Gemeinsames Vorgehen:**

- Legt eine **gemeinsame** Größe für die Stichprobe (Zahl für N, Z. 142) fest (eine „große“ Zahl, mindestens 10000, höchstens 56284). Gebt sie alle im Programm ein und notiert sie auf eurem Antwortblatt!
- Gebt **gemeinsam** für jede Zahl ein Intervall vor, indem gesucht werden soll. Schaut euch dazu das Beispiel unten an und ersetzt in Zeile 143 die Ausdrücke NaN. Notiert die Intervalle auf eurem Antwortblatt.

  Beispiel: $range = [0,1;0,0.5]$; bedeutet, dass für die erste Zahl im Bereich $[0,1]$ und für die zweite Zahl im Bereich $[0,0.5]$ gesucht wird.

  Hinweis: Wir arbeiten hier mit Zahlen für Anteile. Diese liegen immer zwischen 0 und 1.

c) **Optimierung durchführen:**

- Führt den Abschnitt durch Drücken auf *Run Section* aus. *Was tut die Optimierung? Beschreibt das Suchprinzip auf eurem Arbeitsblatt.*
- Die Ergebnisse sind der hier gefundene optimale Kappawert und die zugehörigen Zahlen. Notiert sie in der Tabelle.
- Die Ergebnisse sind hier von der zufällig ausgewählten Stichprobe der Größe N abhängig. Wir wollen diese Abhängigkeit verringern. Bei jeder Durchführung wird die Stichprobe neu ausgewählt. Führt daher die Optimierung an jedem Laptop mindestens 5 mal aus. Notiert jeweils den optimalen Kappawert und die zugehörigen Zahlen.
- *Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für die am häufigsten auftretenden Zahlen den durchschnittlichen Kappawert.* Tragt eure Ergebnisse auf dem Antwortblatt ein.

d) *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 4 vorhergesagt werden? Welche Zahlen sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Zahlen? Vergleicht sie mit euren vorherigen Ergebnissen.* Tragt eure Antwort auf dem Antwortzettel ein.

e) *Welche Probleme oder Vorteile seht ihr bei dieser Art der Optimierung?* Tragt eure Antwort auf dem Antwortzettel ein.

f) *Wie könntet ihr die Suche nach guten Zahlen verfeinern?* Beschreibt ein mögliches Vorgehen auf dem Antwortblatt und führt es durch.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 3 - Schritt 5 - Regel 5

a) Schaut euch den Abschnitt „Kappa verbessern“ (Z. 122-155) an. Arbeitet mit dem Abschnitt „mit Regel 5“ (Z. 148-155).

b) **Gemeinsames Vorgehen:**

- Legt eine **gemeinsame** Größe für die Stichprobe (Zahl für N, Z. 150) fest (eine „große“ Zahl, mindestens 10000, höchstens 56284). Gebt sie alle im Programm ein und notiert sie auf eurem Antwortblatt!
- Gebt **gemeinsam** für jede Zahl ein Intervall vor, indem gesucht werden soll. Schaut euch dazu das Beispiel unten an und ersetzt in Zeile 151 die Ausdrücke NaN. Notiert die Intervalle auf eurem Antwortblatt.

  Beispiel: $range = [0,1;0,0.5]$; bedeutet, dass für die erste Zahl im Bereich $[0,1]$ und für die zweite Zahl im Bereich $[0,0.5]$ gesucht wird.

  Hinweis: Wir arbeiten hier mit Zahlen für Anteile. Diese liegen immer zwischen 0 und 1.

c) **Optimierung durchführen:**

- Führt den Abschnitt durch Drücken auf *Run Section* aus. *Was tut die Optimierung? Beschreibt das Suchprinzip auf eurem Arbeitsblatt.*
- Die Ergebnisse sind der hier gefundene optimale Kappawert und die zugehörigen Zahlen. Notiert sie in der Tabelle.
- Die Ergebnisse sind hier von der zufällig ausgewählten Stichprobe der Größe N abhängig. Wir wollen diese Abhängigkeit verringern. Bei jeder Durchführung wird die Stichprobe neu ausgewählt. Führt daher die Optimierung an jedem Laptop mindestens 5 mal aus. Notiert jeweils den optimalen Kappawert und die zugehörigen Zahlen.
- *Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für die am häufigsten auftretenden Zahlen den durchschnittlichen Kappawert.* Tragt eure Ergebnisse auf dem Antwortblatt ein.

d) *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 5 vorhergesagt werden? Welche Zahlen sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Zahlen? Vergleicht sie mit euren vorherigen Ergebnissen.* Tragt eure Antwort auf dem Antwortzettel ein.

e) *Welche Probleme oder Vorteile seht ihr bei dieser Art der Optimierung?* Tragt eure Antwort auf dem Antwortzettel ein.

f) *Wie könntet ihr die Suche nach guten Zahlen verfeinern?* Beschreibt ein mögliches Vorgehen auf dem Antwortblatt und führt es durch.

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 3 - Antwortblatt

Schritt 5 - Kappa systematisch verbessern

mit Regel:

b) **Gemeinsames Vorgehen:**

- Größe für die Stichprobe:
- Intervall für die erste Zahl:
- Intervall für die zweite Zahl:

c) **Optimierung durchführen:**

- *Was tut die Optimierung? Beschreibt das Suchprinzip.*

- Ergebnisse: *Nicht genug Platz? Dann nutzt die Rückseite!*

Durchführung	opt. Kappa	zugeh. Zahlen	Durchführung	opt. Kappa	zugeh. Zahlen
1			9		
2			10		
3			11		
4			12		
5			13		
6			14		
7			15		
8			16		

Die am häufigsten auftretenden Zahlen lauten .....

mit dem durchschnittlichen Kappawert von .....

d) *Wie gut kann die sexuelle Orientierung mit Hilfe der Regel vorhergesagt werden? Welche Zahlen sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Zahlen? Vergleicht sie mit euren vorherigen Ergebnissen.*

e) *Welche Probleme oder Vorteile seht ihr bei dieser Art der Optimierung?*

Für die Regeln 4 und 5: f) *Wie könntet ihr die Suche nach guten Zahlen verfeinern?*

Die Suche kann verfeinert werden durch...

Dieses Vorgehen liefert die folgenden Ergebnisse:

Arbeitsblatt 3 - Besprechung

Schritt 5 - Kappa systematisch verbessern

Was tut die Optimierung? Beschreibt das Suchprinzip.

Ergebnisse:

Regel	durchschnittl. opt. Kappa	zugehörige Zahlen
1		
2		
4		
5		

Welche Probleme oder Vorteile seht ihr bei dieser Art der Optimierung?

Für die Regeln 4 und 5: *Wie könntet ihr die Suche nach guten Zahlen verfeinern?*

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 3: Lösungsvorschlag

- **c)** *Was tut die Optimierung? Beschreibt das Suchprinzip.*
  - **Regel 1 und 2:**
    Es werden alle möglichen (natürlichen) Zahlen im angegebenen Intervall ausprobiert. Bei der Kombination der Zahlen muss die erste Zahl größer sein als die zweite.
  - **Regel 4 und 5:**
    Es werden 20 Zahlen im angegebenen Intervall ausprobiert. Bei der Kombination der Zahlen muss die erste Zahl größer sein als die zweite.
    Hinweis: Hier ist die Zahl 20 nicht wichtig. Wichtig ist, dass nicht alle Zahlen ausprobiert werden.
- **b) c) d)** Stichprobengröße und Intervalle festlegen. Optimierung mehrfach durchführen. Ergebnisse bündeln und deuten.
  - **Regel 1:**
    Lösungsvorschlag: Wir nehmen wie zuvor 10000 Nutzer und betrachten $range = [0, 15; 0, 15]$ (vom Hinweis geleitet) und machen 15 Durchläufe:

Durchführung	optimales Kappa	zugehörige Parameter
1	0.61	[2, 1]
2	0.60	[2, 1]
3	0.60	[2, 1]
4	0.58	[2, 1]
5	0.60	[2, 1]
6	0.60	[2, 1]
7	0.60	[2, 1]
8	0.62	[2, 1]
9	0.61	[2, 1]
10	0.59	[2, 1]
11	0.62	[2, 1]
12	0.59	[2, 1]
13	0.60	[2, 1]
14	0.61	[2, 1]
15	0.58	[2, 1]

    Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für diese Zahlen den durchschnittlichen Kappawert.
    Lösungsvorschlag: In diesem Fall waren immer die Zahlen [2, 1] die optimalen. Der durchschnittliche Kappawert beträgt ca. 0.6007.

    Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 1 vorhergesagt werden? Welche Parameter sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Parameter? Vergleicht sie mit euren vorherigen Ergebnissen.
    Lösungsvorschlag: Der Kappawert von 0.6007 mit den Parametern [2, 1] ist hier optimal. Dieser Kappawert liegt minimal höher als die bisher ermittelten zu Regel 1. Hier ist ein Ausschluss des Zufalls gewinnbringend.
  - **Regel 2:**
    Lösungsvorschlag: Wir nehmen wie zuvor 10000 Nutzer und betrachten $range = [0, 35; 0, 35]$ (vom Hinweis geleitet) und machen 15 Durchläufe:

Durchführung	optimales Kappa	zugehörige Parameter
1	0.53	[13, 9]
2	0.54	[12, 11]
3	0.55	[16, 9]
4	0.52	[14, 7]
5	0.54	[12, 8]
6	0.52	[13, 9]
7	0.57	[11, 10]
8	0.54	[12, 8]; [12, 11]
9	0.52	[15, 7]
10	0.52	[13, 9]
11	0.53	[13, 10]
12	0.55	[12, 8]
13	0.53	[17, 8]
14	0.55	[11, 8]; [15, 11]
15	0.54	[13, 12]

Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für diese Zahlen den durchschnittlichen Kappawert.
Lösungsvorschlag: In diesem Fall waren die Zahlen [12, 8] und [13, 9] die häufigsten optimalen Zahlen. Der durschnittliche Kappawert von [12, 8] beträgt ca. 0.5433 und der von [13, 9] 0.5233.

Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 2 vorhergesagt werden? Welche Parameter sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Parameter? Vergleicht sie mit euren vorherigen Ergebnissen.
Lösungsvorschlag: Der Kappawert von 0.5433 mit den Parametern [12, 8] ist hier am höchsten. Dieser Kappawert liegt etwas höher als die bisher ermittelten zu Regel 2. Hier ist ein Ausschluss des Zufalls nicht gewinnbringend.

– **Regel 4:**
Lösungsvorschlag: Wir nehmen wie zuvor 10000 Nutzer und betrachten $range = [0, 1; 0, 1]$ (vom Hinweis geleitet) und machen 15 Durchläufe:

Durchführung	optimales Kappa	zugehörige Parameter
1	0.74	[0.53, 0.37]
2	0.75	[0.53, 0.42]
3	0.76	[0.58, 0.53]
4	0.75	[0.58, 0.53]
5	0.76	[0.58, 0.53]
6	0.74	[0.53, 0.47]
7	0.74	[0.58, 0.53]
8	0.75	[0.58, 0.53]
9	0.75	[0.58, 0.53]
10	0.74	[0.58, 0.53]
11	0.75	[0.58, 0.53]
12	0.73	[0.53, 0.42]; [0.58, 0.53]
13	0.74	[0.53, 0.37]; [0.58, 0.37]
14	0.75	[0.58, 0.53]
15	0.76	[0.58, 0.53]

Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für diese Zahlen den durchschnittlichen Kappawert.
Lösungsvorschlag: In diesem Fall waren die Zahlen [0.58, 0.53] die häufigsten optimalen Zahlen. Der durchschnittliche Kappawert beträgt ca. 0.7491.

Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 4 vorhergesagt werden? Welche Parameter sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Parameter? Vergleicht sie mit euren vorherigen Ergebnissen.
Lösungsvorschlag: Die Kappawert von 0.7491 mit den Parametern [0.58, 0.53] ist hier optimal. Dieser Kappawert liegt nicht höher als die bisher ermittelten zu Regel 4. Die getesteten Zahlen [0.56, 0.55] ergaben ähnliche Werte und liegen in der Nähe vom ermittelten Optimum. Hier ist ein Ausschluss des Zufalls nicht direkt gewinnbringend.

– **Regel 5:**
Lösungsvorschlag: Wir nehmen wie zuvor 10000 Nutzer und betrachten $range = [0, 1; 0, 1]$ (vom Hinweis geleitet) und machen 15 Durchläufe:

Durchführung	optimales Kappa	zugehörige Parameter
1	0.61	[0.63, 0.47]
2	0.61	[0.68, 0.53]
3	0.62	[0.63, 0.42]
4	0.63	[0.58, 0.42]; [0.58, 0.47]
5	0.63	[0.53, 0.42]; [0.58, 0.42]
6	0.61	[0.53, 0.47]
7	0.62	[0.58, 0.53]
8	0.62	[0.53, 0.42]
9	0.61	[0.58, 0.53]
10	0.63	[0.63, 0.53]
11	0.61	[0.58, 0.53]
12	0.62	[0.58, 0.53]
13	0. 61	[0.53, 0.42]
14	0.63	[0.53, 0.47]
15	0.63	[0.63, 0.53]

Gibt es in eurem Team gleiche Ergebnisse für die optimalen Zahlen? Wenn nicht, führt die Optimierung so oft durch, bis das der Fall ist. Bestimmt für diese Zahlen den durchschnittlichen Kappawert.
Lösungsvorschlag: In diesem Fall waren die Zahlen [0.58, 0.53] die häufigsten optimalen Zahlen. Der durchschnittliche Kappawert von [0.58, 0.53] beträgt ca. 0.615.

Wie gut kann die sexuelle Orientierung mit Hilfe der Regel 5 vorhergesagt werden? Welche Parameter sind letztendlich optimal? Welche Ergebnisse liefern die optimalen Parameter? Vergleicht sie mit euren vorherigen Ergebnissen.
Lösungsvorschlag: Der Kappawert von 0.615 mit den Parametern [0.58, 0.53] ist hier optimal. Dieser Kappawert liegt nicht höher als die bisher ermittelten zu Regel 5. Die getesteten Zahlen [0.56, 0.55] ergaben ähnliche Werte und liegen in der Nähe vom ermittelten Optimum. Hier ist ein Ausschluss des Zufalls nicht direkt gewinnbringend.

- **e)** *Welche Probleme oder Vorteile seht ihr bei dieser Art der Optimierung?*
Lösungsvorschlag: Vorteilhaft an der Optimierung ist, dass viele Zahlen ausprobiert werden. Die Wahl der Intervalle ist hier jedoch entscheidend. Nachteilig ist, dass durch das komplette Ausprobieren viel Zeit benötigt wird. Bei steigender Stichprobengröße steigt auch die Zeit pro Optimierungsdurchlauf.

- **f)** *Wie könntet ihr die Suche nach guten Zahlen verfeinern?* Beschreibt ein mögliches Vorgehen und führt es durch.

  – **Regel 4:**
  Lösungsvorschlag: Wie festgestellt, liegen „gute“ Zahlen nah beieinander. Wir testen nun ein kleineres Intervall: [0.4,0.6;0.4,0.6] und erhalten als Ergebnis:
  ———————————
  Der optimale Kappawert ist 0.75

mit den zugehörigen Zahlen [0.52, 0.51]
mit den zugehörigen Zahlen [0.53, 0.51]
mit den zugehörigen Zahlen [0.54, 0.51]
mit den zugehörigen Zahlen [0.55, 0.51]
mit den zugehörigen Zahlen [0.56, 0.51]
mit den zugehörigen Zahlen [0.53, 0.52]
mit den zugehörigen Zahlen [0.54, 0.52]
mit den zugehörigen Zahlen [0.55, 0.52]
mit den zugehörigen Zahlen [0.57, 0.52]
mit den zugehörigen Zahlen [0.54, 0.53]
mit den zugehörigen Zahlen [0.55, 0.53]
mit den zugehörigen Zahlen [0.55, 0.54]
———————————

Der Kappawert liegt hier in der bereits ermittelten Größenordnung. Es gibt viele Zahlenkombinationen in diesem Intervall, die diesen Wert hervorbringen. Demnach ist es sinnvoll zur Optimierung dieser Regel, Zahlen knapp über 0.5 zu wählen, die nahe beieinander liegen.

– **Regel 5:**
Lösungsvorschlag: Wie festgestellt, liegen „gute“ Zahlen nah beieinander. Wir testen nun ein kleineres Intervall: [0.4,0.6;0.4,0.6] und erhalten als Ergebnis:
———————————

Der optimale Kappawert ist 0.62
mit den zugehörigen Zahlen [0.59, 0.54]
———————————

Der Kappawert liegt hier in der bereits ermittelten Größenordnung. Es gibt also eine weitere Zahlenkombinationen in diesem Intervall, die einen solchen Wert hervorbringt. Demnach ist es sinnvoll zur Optimierung dieser Regel, Zahlen im ersten Fall um 0.58/0.59 bzw. im zweiten Fall um 0.53/0.54 zu wählen.

- **Hinweise bei der Besprechung:**

Anhand der Ergebnisse soll die Frage nach der besten Heuristik beantwortet werden!

Der relative Anteil an homosexuellen Freunden scheint die beste Aussage über die sexuelle Orientierung zu liefern.
Beim Vergleich der Heuristiken mit absoluten und mit relativen Anteilen ist festzuhalten: Die Heuristik zur Betrachtung der relativen Anteile der Freunde 1. Ordnung liefert die besten Werte für Kappa. Ein Grund dafür ist, dass durch die relativen Anteile mehr Nutzer einbezogen werden. Hat ein Nutzer beispielsweise nur 4 Freunde von denen einer homosexuell ist, so wird er bei alleiniger Betrachtung der Anzahlen mit der hier beschriebenen Regel (Wer mehr als 5 homosexuelle Nutzer hat, ist homosexuell) nicht als homosexuell klassifiziert werden. Bei relativen Anteilen kann das aber der Fall sein. Wir binden also die Nutzer mit wenigen Freunden besser ein. Zur Erinnerung: Im Schnitt haben unsere Nutzer nur 3 oder 4 Freunde erster Ordnung.
Die Heuristik zu den relativen Anteilen der Freunde 2. Ordnung liefert geringfügig bessere Ergebnisse als die der absoluten Anteile. Dies kann daran liegen, dass die Nutzer im Allgemeinen deutlich mehr Freunde zweiter Ordnung haben (Homosexuelle im Schnitt 29.3882 und Heterosexuelle im Schnitt 18.4629) und daher die absoluten Anteile zur Vorhersage schon gute Ergebnisse erbringen. Die Freunde 2. Ordnung scheinen bei der Bestimmung der sexuellen Orientierung weniger aussagekräftig.

- **Code und Ausgabe:**

```
%%      Kappa verbessern

%% mit Regel 1:
clc
N = 10000;
range = [0,15;0,15];

% Optimierung wird durchgeführt
Optimization3_Regel1(Users,N,range);
```

```
....
params =

    15    13


Kappa =

    0.3730


params =

    15    14


Kappa =

    0.3730

_____________________________________
Der optimale Kappawert ist    0.59

mit den zugehörigen Zahlen [ 2, 1]

_____________________________________
```

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

Arbeitsblatt 4: Ergebnisse auf Gültigkeit testen (Zusatz)

Schritt 6 - Ergebnisse auf Gültigkeit testen mit Hypothesentests

Mit Hilfe einer Heuristik können wir die Wahrscheinlichkeit, eine richtige Vorhersagen zu machen, beeinflussen. Wir beeinflussen die *Trefferwahrscheinlichkeit*. Unser Ziel ist es, eine Heuristik zu finden, die eine höhere Trefferwahrscheinlichkeit als der Zufall hat - und zwar für alle Nutzer. Bisher haben wir die Heuristiken auf einen Teil der Nutzer, auf eine Stichprobe, angewendet. Die Ergebnisse sollten aber am besten *verlässlich* für alle Nutzer sein. Man sagt: Die Behauptung, dass die Trefferwahrscheinlichkeit der Heuristik besser ist als die des Zufalls, soll *signifikant* sein. Eine solche Behauptung wird in der Mathematik mit Hilfe eines **Hypothesentests** untersucht. Die Ergebnisse der Heuristiken aus der Anwendung auf eine Stichprobe sind dabei der Grundstein.

a) Der Begriff Hypothese ist euch bestimmt schon begegnet. Eine Hypothese ist (grob gesagt) eine Behauptung, die belegt werden soll. Wir nennen sie „H_1“:

H_1-Hypothese: Die Trefferwahrscheinlichkeit der Heuristik ist größer als die Trefferwahrscheinlichkeit durch den Zufall.

- Formuliert die H_1-Hypothese formaler, indem ihr die Trefferwahrscheinlichkeit der Heuristik als p_H und die Trefferwahrscheinlichkeit durch den Zufall als p_Z bezeichnet und Verhältniszeichen (z.B. $<, >, \leq, \geq, =$) nutzt. Wählt aus den folgenden Möglichkeiten aus:

1	2	3	4	5
$p_H < p_Z$	$p_H > p_Z$	$p_H \leq p_Z$	$p_H \geq p_Z$	$p_H = p_Z$

- Öffnet die Datei „ZusatzHypothesentests“. Überprüft euer Ergebnis, indem ihr im Abschnitt „H1 Hypothese“ in Zeile 6 das NaN durch die eurem Ergebnis entsprechende Zahl ersetzt. Drückt auf *Run Section*. Tragt die H_1-Hypothese in Zeichen in den Kasten auf eurem Antwortblatt ein.

b) Dieser Behauptung gegenüber steht die Hypothese, die wir widerlegen wollen: die „H_0-Hypothese“.

- Formuliert in Worten die Hypothese, die wir widerlegen wollen. *Tipp: Welche Verhältnisse können auftreten, wenn das Verhältnis in der H_1-Hypothese nicht eintritt.*
- Formuliert die H_0-Hypothese formaler. Wählt euer Ergebnis aus der Tabelle oben aus.
- Überprüft euer Ergebnis, indem ihr im Abschnitt „H0 Hypothese“ in Zeile 13 das NaN durch die eurem Ergebnis entsprechende Zahl ersetzt. Drückt auf *Run Section*. Tragt die H_0-Hypothese in Worten und Zeichen in den Kasten auf eurem Antwortblatt ein.

Was heißt „signifikant“? Wir wollen uns zu einem gewissen Maß drauf verlassen können, dass die Heuristik besser ist als der Zufall. Dieses Maß wird durch das „Signifikanzniveau“ (genannt α) vorgegeben. Das Signifikanzniveau ist die Irrtumswahrscheinlichkeit. Die Irrtumswahrscheinlichkeit ist die Wahrscheinlichkeit, dass wir eine Fehlentscheidung treffen, wenn wir uns für eine der Hypothesen entscheiden. Natürlich ist es hier sinnvoll, die Irrtumswahrscheinlichkeit möglichst klein zu halten. Aber sie wird niemals verschwinden. Zu Beginn eines Hypothesentests müssen wir eine Irrtumswahrscheinlichkeit angeben. Wir müssen also angeben, wie sicher wir unsere Aussage treffen möchten. Normalerweise gibt man Irrtumswahrscheinlichkeiten von 0.01, 0.05 oder 0.1 vor.

Wann kommt es zur Annahme bzw. zum Ablehnen von Hypothesen? Wir berechnen dazu eine *Teststatistik* (V), die mit einem festen Wert, der *Schranke* (u), verglichen wird. Die Schranke ist abhängig von der vorgegebenen Irrtumswahrscheinlichkeit α.

Die H_0**-Hypothese wird abgelehnt**, wenn

$$V > u.$$

Das bedeutet, dass dann die Richtigkeit der H_1-Hypothese angenommen wird und wir unsere Behauptung statistisch belegen können.
Die Richtigkeit der H_0**-Hypothese wird angenommen**, wenn

$$V \leq u.$$

Das bedeutet, dass dann die Richtigkeit der H_1-Hypothese abgelehnt wird und wir unsere Behauptung statistisch *nicht* belegen können.

c) Nun führen wir Hypothesentests durch. Betrachtet Abschnitt „Hypothesentests durchführen“ (Z. 18-113). Sucht euch eine Regel aus.

Regel 1, 2, 4 oder 5:
1. Wir wollen die Zuverlässigkeit unserer besten Zahlen untersuchen. Sucht diese aus euren bisherigen Ergebnissen heraus.
2. Jetzt arbeitet ihr mit Abschnitt „mit Regel 1/2/4/5“. Zuerst brauchen wir für unseren Test eine „große“ Stichprobe, also eine Anzahl von Nutzern (N), die wir testen wollen. Gebt eine Anzahl von Nutzern (N, Z. 22/41/78/97) ein. Notiert sie auf eurem Antwortblatt. Gebt zudem die besten Zahlen (params, Z. 23/42/79/98) für eure Regel ein. Notiert sie auf eurem Antwortblatt.
3. Gebt eine Irrtumswahrscheinlichkeit (ein Signifikanzniveau α) an, indem ihr eine der folgenden Zahlen in Zeile 28/47/84/103 statt NaN einsetzt: 0.01, 0.05 oder 0.1. Notiert die eingegebene Zahl auf eurem Antwortblatt.
4. Führt den Hypothesentest mit einem Klick auf *Run Section* durch. Das Ergebnis wird angezeigt. Notiert und interpretiert es!
5. Ändert das Signifikanzniveau α und führt den Test erneut durch. Notiert eure Eingaben, das Ergebnis und eure Interpretation.
6. *Zu welchem Signifikanzniveau sind die Ergebnisse verlässlich?* Beantwortet die Frage auf dem Antwortblatt.

Regel 3: (Z. 58-75)
1. Jetzt arbeitet ihr mit Abschnitt „mit Regel 3“. Zuerst brauchen wir für unseren Test eine „große“ Stichprobe, also eine Anzahl von Nutzern (N), die wir testen wollen. Gebt eine Anzahl von Nutzern (N, Z. 60) ein. Notiert sie auf eurem Antwortblatt.
2. Gebt eine Irrtumswahrscheinlichkeit (ein Signifikanzniveau α) an, indem ihr eine der folgenden Zahlen in Zeile 65 statt NaN einsetzt: 0.01, 0.05 oder 0.1. Notiert die eingegebene Zahl auf eurem Antwortblatt.
3. Führt den Hypothesentest mit einem Klick auf *Run Section* durch. Das Ergebnis wird angezeigt. Notiert und interpretiert es!
4. Ändert das Signifikanzniveau α und führt den Test erneut durch. Notiert eure Eingaben, das Ergebnis und eure Interpretation.
5. *Zu welchem Signifikanzniveau sind die Ergebnisse verlässlich?* Beantwortet die Frage auf dem Antwortblatt.

Zusatzaufgabe - Hypothesentests zu anderen Regeln durchführen

Führt Hypothesentests für andere Regeln durch, indem ihr die entsprechenden Anweisungen befolgt.

Wir besprechen nun eure Ergebnisse!

Arbeitsblatt 4: Antwortblatt

Schritt 6 - Ergebnisse auf Gültigkeit testen mit Hypothesentests

Wir testen die

H_1**-Hypothese:** Die Trefferwahrscheinlichkeit der Heuristik ist größer als die Trefferwahrscheinlichkeit durch den Zufall. (in Zeichen: .....)

gegen die

H_0**-Hypothese:**.....

.....

(in Zeichen: .....)

Ergebnisse der Hypothesentests:

- Regel:
- Größe der Stichprobe:
- Für Regel 1,2,4,5: genutzte Zahlen:

Signifikanzniveau α			
Ergebnis:			
Interpretation:			

Zu welchem Signifikanzniveau sind die Ergebnisse verlässlich?

CAMMP day
Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 4: Besprechung

Schritt 6 - Ergebnisse auf Gültigkeit testen mit Hypothesentests

Wir testen die

H_1**-Hypothese:** Die Trefferwahrscheinlichkeit der Heuristik ist größer als die Trefferwahrscheinlichkeit durch den Zufall. (in Zeichen: .....)

gegen die

H_0**-Hypothese:**.....

.....

.....

(in Zeichen: .....)

Ergebnisse der Hypothesentests:

Die Hypothese „Die Trefferwahrscheinlichkeit der Heuristik ist größer als die Trefferwahrscheinlichkeit durch den Zufall“ wurde für die

Heuristik	1	2	3	4	5
mit den Zahlen			-		
zum Niveau α					
statistisch belegt	ja/nein	ja/nein	ja/nein	ja/nein	ja/nein

CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Arbeitsblatt 4: Lösungsvorschlag

Schritt 6 - Ergebnisse auf Gültigkeit testen mit Hypothesentests

- **a)** Die H_1-Hypothese: "Die Trefferwahrscheinlichkeit der Heuristik ist größer als die Zufallswahrscheinlichkeit" formaler formulieren.
  Lösung: Möglichkeit 2: $p_H > p_Z$ (Trefferwahrscheinlichkeit der Heuristik ist p_H und die Zufallswahrscheinlichkeit ist p_Z)

- **Code und Ausgabe:**
  bei richtigem Ergebnis:

```
%%                  Zusatz: Hypothesentests

%% H1 Hypothese
clc
H1 = 2;

% Prüfe die Eingabe:
PruefeH1(H1);
```

```
Gut! Eure Antwort ist richtig
```

bei falschem Ergebnis:

```
%% H1 Hypothese
clc
H1 = 1;

% Prüfe die Eingabe:
PruefeH1(H1);
```

```
Die Antwort ist fast richtig.
```

oder

```
%% H1 Hypothese
clc
H1 = 5;

% Prüfe die Eingabe:
PruefeH1(H1);
```

```
Die Antwort ist falsch. Probiert es erneut.
```

- **b)** Formuliert in Worten die Hypothese, die wir widerlegen wollen. Formuliert die H_0-Hypothese formaler. Wählt euer Ergebnis aus der Tabelle oben aus.
  Lösung: "Die Trefferwahrscheinlichkeit der Heuristik ist kleiner als die oder gleich der Trefferwahrscheinlichkeit durch den Zufall." Das entspricht der Möglichkeit 3: $p_H \leq p_Z$.

- **Code und Ausgabe:**
  bei richtigem Ergebnis:


```
%% H0 Hypothese
clc
H0 = 3;

% Prüfe die Eingabe:
PruefeH0(H0);
```

```
Gut! Eure Antwort ist richtig
```

bei falschem Ergebnis:

```
%% H0 Hypothese
clc
H0 = 4;

% Prüfe die Eingabe:
PruefeH0(H0);
```

```
Die Antwort ist fast richtig.
```

oder

```
%% H0 Hypothese
clc
H0 = 1;

% Prüfe die Eingabe:
PruefeH0(H0);
```

```
Die Antwort ist falsch. Probiert es erneut.
```

- **c)** Hypothesentest durchführen.
  Hinweis: Ist die H_1-Hypothese schon zum kleinsten Signifikanzniveau (0.01) gültig, so gilt dies auch für alle höheren.(vgl. basic paper)
  Lösungsvorschlag: Wir testen erneut 10000 Nutzer und nehmen bei den parameterabhängigen Heuristiken die aus den Optimierungen als am besten bestimmten Parameter.

Regel	Zahlen	Niveau	Ergebnis
1	2;1	0.01	V > u
2	12;8	0.01	V > u
3	-	0.01	V < u oder V = u
3	-	0.05	V < u oder V = u
3	-	0.1	V < u oder V = u
4	0.58;0.53	0.01	V > u
5	0.58;0.53	0.01	V > u

- **Interpretationen:**
  - **Regel 1:** Die H_0-Hypothese wird zum Signifikanzniveau 0.01 (und auch höher) abgelehnt und wir können statistisch belegen, dass die Trefferwahrscheinlichkeit der Heuristik 1 mit den Zahlen 2 und 1 größer ist als die des Zufalls.
  - **Regel 2:** Die H_0-Hypothese wird zum Signifikanzniveau 0.01 (und auch höher) abgelehnt und wir können statistisch belegen, dass die Trefferwahrscheinlichkeit der Heuristik 2 mit den Zahlen 12 und 8 größer ist als die des Zufalls.

– **Regel 3:** Die H_0-Hypothese wird zu jedem der drei Niveaus angenommen. Wir können also statistisch nicht belegen, dass die Trefferwahrscheinlichkeit der Heuristik 3 größer ist als die des Zufalls.

Hinweis zur Besprechung: Hier kann, falls es die Zeit zulässt, noch diskutiert werden, welche Behauptung sich aus den bisherigen Untersuchungen ergibt, die man statistisch belegen könnte.
Die Heuristik hat bisher schlechtere Ergebnisse erzielt als der Zufall. Daher könnte die Behauptung: "Die Trefferwahrscheinlichkeit der Heuristik ist schlechter als der Zufall" statistisch geprüft werden.

– **Regel 4:** Die H_0-Hypothese wird zum Signifikanzniveau 0.01 (und auch höher) abgelehnt und wir können statistisch belegen, dass die Trefferwahrscheinlichkeit der Heuristik 4 mit den Zahlen 0.58 und 0.53 größer ist als die des Zufalls.

– **Regel 5:** Die H_0-Hypothese wird zum Signifikanzniveau 0.01 (und auch höher) abgelehnt und wir können statistisch belegen, dass die Trefferwahrscheinlichkeit der Heuristik 5 mit den Zahlen 0.58 und 0.53 größer ist als die des Zufalls.

- **Code und Ausgabe:**

```
%%                Hypothesentests durchführen

%% mit Regel 1:
clc
N = 10000;
params=[5 2];

% Heuristik auf Stichprobe anwenden
[TrefferquoteHeuristik,TrefferquoteZufall] =heuristicTester_Test(@Heuristic_Regel1params,Users,
N,params);

alpha = 0.1;

% Teststatistik bestimmen
V = TeststatistikbestimmenN, TrefferquoteHeuristik, TrefferquoteZufall);

%Schranke bestimmen
u = Schrankebestimmen(alpha);

%Test durchführen
HypothesentestdurchfuehrenRegel12(@Heuristic_Regel1params,V,u,alpha,N,params);
```

```
_____________________________
Wir haben diese Heuristik genutzt:

Heuristic_Regel1params

mit den Zahlen:            5
und:       2

Größe der Stichprobe:    10000

_____________________________

Signifikanzniveau:
    0.1000

_____________________________
Ergebnis Hypothesentest: V > u
```

```
%% mit Regel 3:
clc
```



```
N = 10000;

% Heuristik auf Stichprobe anwenden
[TrefferquoteHeuristik,TrefferquoteZufall] =heuristicTester_TestRegel3(@Heuristic_Regel3,Users,N);

alpha = 0.05;

% Teststatistik bestimmen
V = TeststatistikbestimmenN, TrefferquoteHeuristik, TrefferquoteZufall);

%Schranke bestimmen
u = Schrankebestimmen(alpha);

%Test durchführen
HypothesentestdurchfuehrenRegel3(@Heuristic_Regel3,V,u,alpha,N);
```

```
_____________________________
Wir haben diese Heuristik genutzt:

Heuristic_Regel3

Größe der Stichprobe:    10000

_____________________________

Signifikanzniveau:
    0.0500

_____________________________
Ergebnis Hypothesentest: V < u oder V = u
```

A.4. Codes

Hier werden nur die Codes dargestellt, die die Schülerinnen und Schüler im CAMMP day nutzen. Programme, auf die zugegriffen wird, sind nicht dargestellt.

A.4.1. sozialenetzwerke.m

```
%       Wie sicher ist meine Privatsphäre in sozialen Netzwerken
%               ... und was hat das mit Mathe zu tun?

%%                      Heuristiken - Einstieg
%%      Lade Daten

clc
close all
load v2Adj
load v3FriendsterUser

%%      Heuristiken auf eine Person anwenden
clc
i = NaN;

% Heurostiken werden angewendet:
AnwendenaufeinePerson(Users,i);

%%      Heuristiken auf viele Personen anwenden

%% mit Regel 1:
clc
N = 1000;
params = [4  5];

% Personen einschätzen:
heuristicTester_EinstiegRegel12(@Heuristic_Regel1params, Users, N,
params)

%% mit Regel 2:
clc
N = NaN;
params = [5  3];

% Personen einschätzen:
```

```
heuristicTester_EinstiegRegel12(@Heuristic_Regel2params, Users, N,
params)

%% mit Regel 3:
clc
N = NaN;

% Personen einschätzen:
heuristicTester_EinstiegRegel3(@Heuristic_Regel3, Users, N)

%% mit Regel 4:
clc
N = NaN;
params = [0.2 0.1];

% Personen einschätzen:
heuristicTester_EinstiegRegel45(@Heuristic_Regel4params, Users, N,
params)

%% mit Regel 5:
clc
N = NaN;
params = [0.3 0.1];

% Personen einschätzen:
heuristicTester_EinstiegRegel45(@Heuristic_Regel5params, Users, N,
params)

%%                    Heuristiken - Güte bewerten

%%      Übereinstimmung durch Zufall
clc
pe = NaN;

% Prüfe die Eingabe:
Pruefepe(pe);

%%      Kappa betrachten
clc

a1 = NaN;
a2 = NaN;
```

```
a3 = NaN;
a4 = NaN;

% Prüfe die Eingabe:
PruefeLueckentext(a1,a2,a3,a4);

%%      Heuristiken vergleichen

%% mit Regel 1:
clc
N = NaN;
params=[5  2];

% Kappa bestimmen:
heuristicTester_kappaRegel12(@Heuristic_Regel1params,Users,N,
params);

%% mit Regel 2:
clc
N = NaN;
params=[5  3];

% Kappa bestimmen:
heuristicTester_kappaRegel12(@Heuristic_Regel2params,Users,N,
params);

%% mit Regel 3:
clc
N = NaN;

% Kappa bestimmen:
heuristicTester_kappaRegel3(@Heuristic_Regel3,Users,N);

%% mit Regel 4:
clc
N = NaN;
params=[0.2  0.1];

% Kappa bestimmen:
heuristicTester_kappaRegel45(@Heuristic_Regel4params,Users,N,
params);
```

```
%% mit Regel 5:
clc
N = NaN;
params=[0.3  0.1];

% Kappa bestimmen:
heuristicTester_kappaRegel45(@Heuristic_Regel5params,Users,N,
params);

%%      Kappa verbessern

%% mit Regel 1:
clc
N = NaN;
range = [NaN,NaN;NaN,NaN];

% Optimierung wird durchgeführt
Optimization3_Regel1(Users,N,range);

%% mit Regel 2:
clc
N = NaN;
range = [NaN,NaN;NaN,NaN];

% Optimierung wird durchgeführt
Optimization3_Regel2(Users,N,range);

%% mit Regel 4:
clc
N = NaN;
range = [NaN,NaN;NaN,NaN];

% Optimierung wird durchgeführt
Optimization3_Regel4(Users,N,range);

%% mit Regel 5:
clc
N = NaN;
range = [NaN,NaN;NaN,NaN];

% Optimierung wird durchgeführt
Optimization3_Regel5(Users,N,range);
```

```
%% ENDE
```

A.4.2. ZusatzGuetebewerten.m

```
%%                    Zusatzblatt 1: Heuristiken Güte bewerten

%%      Trefferquote und Genauigkeit betrachten
clc

a1 = NaN;
a2 = NaN;
a3 = NaN;
a4 = NaN;

% Prüfe die Eingabe:
PruefeLueckentext2(a1,a2,a3,a4);

%%      Trefferquote und Genauigkeit der Heuristiken

%% mit Regel 1:
clc
N = NaN;
params=[5  2];

% Kappa bestimmen:
heuristicTester_TrefferquoteGenauigkeitRegel12
(@Heuristic_Regel1params,Users,N, params);

%% mit Regel 2:
clc
N = NaN;
params=[5  3];

% Kappa bestimmen:
heuristicTester_TrefferquoteGenauigkeitRegel12
(@Heuristic_Regel2params,Users,N, params);

%% mit Regel 3:
clc
N = NaN;
```

```
% Kappa bestimmen:
heuristicTester_TrefferquoteGenauigkeitRegel3
(@Heuristic_Regel3,Users,N);

%% mit Regel 4:
clc
N = NaN;
params=[0.2  0.1];

% Kappa bestimmen:
heuristicTester_TrefferquoteGenauigkeitRegel45
(@Heuristic_Regel4params,Users,N, params);

%% mit Regel 5:
clc
N = NaN;
params=[0.3  0.1];

% Kappa bestimmen:
heuristicTester_TrefferquoteGenauigkeitRegel45
(@Heuristic_Regel5params,Users,N, params);

%% ENDE
```

A.4.3. ZusatzHypothesentests.m

```
%%                  Zusatz: Hypothesentests

%% H1 Hypothese
clc
H1 = NaN;

% Prüfe die Eingabe:
PruefeH1(H1);

%% H0 Hypothese
clc
H0 = NaN;

% Prüfe die Eingabe:
PruefeH0(H0);
```

```
%%                Hypothesentests durchführen

%% mit Regel 1:
clc
N = NaN;
params=[5  2];

% Heuristik auf Stichprobe anwenden
[TrefferquoteHeuristik, TrefferquoteZufall] =heuristicTester_Test
(@Heuristic_Regel1params, Users,N, params);

alpha = NaN;

% Teststatistik bestimmen
V = Teststatistikbestimmen(N, TrefferquoteHeuristik,
TrefferquoteZufall);

%Schranke bestimmen
u = Schrankebestimmen(alpha);

%Test durchführen
HypothesentestdurchfuehrenRegel12(@Heuristic_Regel1params,V,u,
alpha,N, params);

%% mit Regel 2:
clc
N = NaN;
params=[5  3];

% Heuristik auf Stichprobe anwenden
[TrefferquoteHeuristik, TrefferquoteZufall] =heuristicTester_Test
(@Heuristic_Regel2params, Users,N, params);

alpha = NaN;

% Teststatistik bestimmen
V = Teststatistikbestimmen(N, TrefferquoteHeuristik,
TrefferquoteZufall);

%Schranke bestimmen
u = Schrankebestimmen(alpha);
```

```
%Test durchführen
HypothesentestdurchfuehrenRegel12(@Heuristic_Regel2params,V,u,
alpha,N,params);

%% mit Regel 3:
clc
N = NaN;

% Heuristik auf Stichprobe anwenden
[TrefferquoteHeuristik,TrefferquoteZufall] =
heuristicTester_TestRegel3(@Heuristic_Regel3,Users,N);

alpha = NaN;

% Teststatistik bestimmen
V = Teststatistikbestimmen(N, TrefferquoteHeuristik,
TrefferquoteZufall);

%Schranke bestimmen
u = Schrankebestimmen(alpha);

%Test durchführen
HypothesentestdurchfuehrenRegel3(@Heuristic_Regel3,V,u,alpha,N);

%% mit Regel 4:
clc
N = NaN;
params=[0.2  0.1];

% Heuristik auf Stichprobe anwenden
[TrefferquoteHeuristik,TrefferquoteZufall] =heuristicTester_Test
(@Heuristic_Regel4params,Users,N,params);

alpha = NaN;

% Teststatistik bestimmen
V = Teststatistikbestimmen(N, TrefferquoteHeuristik,
TrefferquoteZufall);

%Schranke bestimmen
u = Schrankebestimmen(alpha);
```

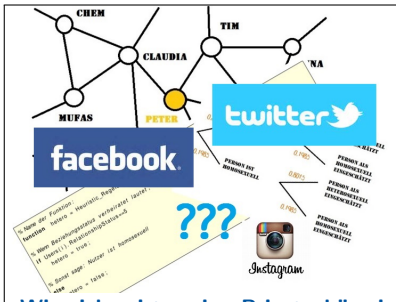
```
%Test durchführen
HypothesentestdurchfuehrenRegel45(@Heuristic_Regel4params,V,u,
alpha,N,params);

%% mit Regel 5:
clc
N = NaN;
params=[0.3  0.1];

% Heuristik auf Stichprobe anwenden
[TrefferquoteHeuristik,TrefferquoteZufall] =heuristicTester_Test
(@Heuristic_Regel5params,Users,N,params);

alpha = NaN;

% Teststatistik bestimmen
V = Teststatistikbestimmen(N, TrefferquoteHeuristik,
TrefferquoteZufall);

%Schranke bestimmen
u = Schrankebestimmen(alpha);

%Test durchführen
HypothesentestdurchfuehrenRegel45(@Heuristic_Regel5params,V,u,
alpha,N,params);

%% ENDE
```

A.5. Abschlussbesprechung

A.5.1. Abbildungsverzeichnis

- Titelbild: Logos
  facebook:
  `http://teenspeak.org/2011/11/23/the-latest-facebook-virus/` (Stand: 24.5.2016)
  twitter:
  `http://ictville.com/2013/02/forget-hate-speech-and-make-twitter-a-job-search-tool/` (Stand: 24.5.2016)
  Instagram:
  `http://www.mysmn.com/how-to-the-basics-of-instagram/instagram-logo-2/` (Stand: 24.5.2016)

- Männchen mit Fragezeichen:
  `https://tse3.mm.bing.net/th?id=OIP.M89deedb40f72e5f5787e1d4f25ca4967o0&pid=15.1&P=0&w=300&h=300` (Stand: 24.5.2016)



Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Sarah Schönbrodt, Benjamin Terschanski, Maike Sube
Lehrstuhl für Mathematik
Center for Computational Engineering Science

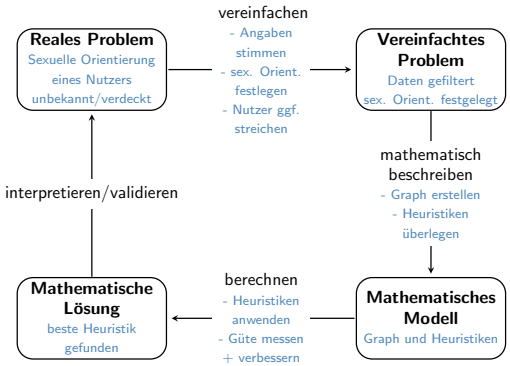
Aachen, 17. Juni 2016

Unser heutiges Thema:

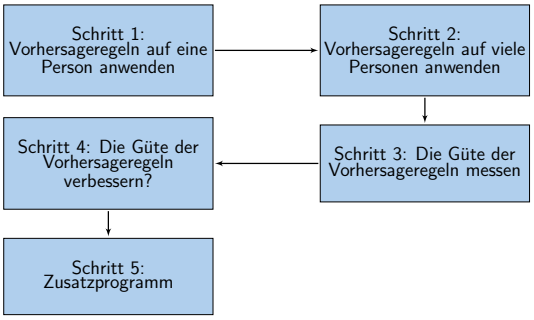
Unser heutiges Thema:

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

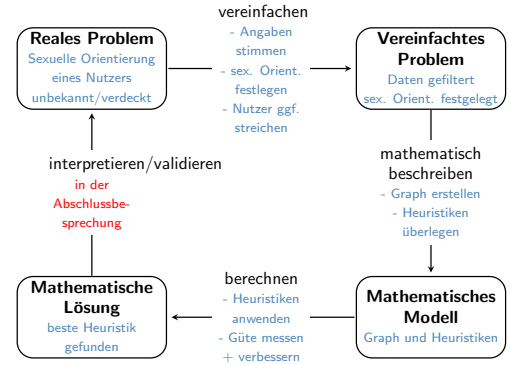
Modellierungskreislauf



Berechnen | in mehreren Schritten



Modellierungskreislauf



Frage beantworten!

Frage beantworten!

Wie gut kann man die sexuelle Orientierung eines Nutzers in einem sozialen Netzwerk vorhersagen, obwohl diese aus den Angaben in seinem Profil nicht ablesbar ist?

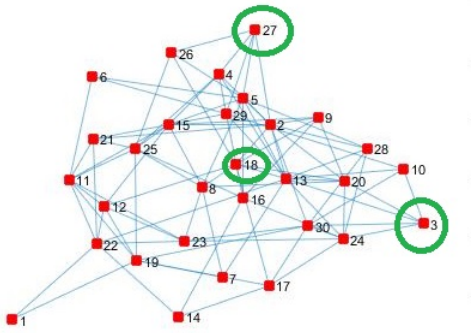
Zusammenfassung der Antworten

- Durch recht **einfache Regeln** (Wenn... dann...-Aussagen) kann man die sexuelle Orientierung richtig (mit bis zu 90% Sicherheit!) vorhersagen.

Zusammenfassung der Antworten

- Durch recht **einfache Regeln** (Wenn... dann...-Aussagen) kann man die sexuelle Orientierung richtig (mit bis zu 90% Sicherheit!) vorhersagen.
- Dabei ist die Nutzung des **Anteils der homosexuellen Freunde eines Nutzers** am besten.

Privatsphäre sicher?



Eure Schlussfolgerungen aus diesem CAMMP day?

- Überrascht es euch, wie einfach eine Vorhersage über euch gemacht werden kann?

Eure Schlussfolgerungen aus diesem CAMMP day?

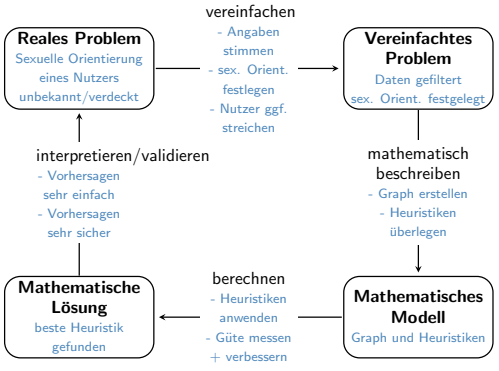
- Überrascht es euch, wie einfach eine Vorhersage über euch gemacht werden kann?
- Wenn die besten Vorhersagen mit Hilfe des Anteils der homosexuellen Freunde eines Nutzers gemacht werden, sind dann Nicht-Nutzer vor den Vorhersagen sicher?

Eure Schlussfolgerungen aus diesem CAMMP day?

- Überrascht es euch, wie einfach eine Vorhersage über euch gemacht werden kann?
- Wenn die besten Vorhersagen mit Hilfe des Anteils der homosexuellen Freunde eines Nutzers gemacht werden, sind dann Nicht-Nutzer vor den Vorhersagen sicher?
- Nehmt Stellung zu der Aussage:

  "Unsere Privatsphäre in sozialen Netzwerken ist von allen Nutzern abhängig."

Modellierung verbessern?



CAMMP day

Wie sicher ist meine Privatsphäre in sozialen Netzwerken? ...und was hat das mit Mathe zu tun?

Hinweise zum Abschlussvortrag

Folie 1/2 | Zentrale Frage

Die zentrale Fragestellung wird wiederholt.

Folie 3 | Modellierungskreislauf

Kurze Wiederholung des Kreislaufs und Orientierung, an welchem Schritt wir standen.

Folie 4 | Der Weg zum Ziel

Kurze Wiederholung des Arbeitsweges.

Folie 5/6 | Ergebnisse interpretieren

- Die zentrale Fragestellung soll von den Schülerinnen und Schülern beantwortet werden.
- Mündliche Sammlung der Antworten.

Folie 7 | Zusammenfassung der Antworten

- Eine Zusammenfassung (recht grob) ist hier visualisiert.
- Anmerkung zu den 90%:
  Diese Zahl erschließt sich aus einer Umrechnung von Kappa:
  Ist Kappa 0.7:

$$\kappa = 0.7 \Leftrightarrow \frac{p - 0.6818}{1 - 0.6818}$$

  also

$$p = 0.7 \cdot 0.3182 + 0.6818 \approx 0.9$$

- Bezug zur anfänglichen Einschätzung (Folie Besprechung 1) kann gezogen werden.
- Zur Verdeutlichung kann ein Graph zur Kassengröße von 30 Schülerinnen und Schülern gezeigt werden.

Folie 8 | Privatsphäre sicher?

Graph zur Klassengröße von 30 Schülerinnen und Schülern wird gezeigt. Bei 90% Richtigkeit sind nur die grün umrandeten Nutzer in ihrer Privatsphäre sicher.

Folie 9 | Schlussfolgerungen

- Die Schülerinnen und Schüler sollen hier Gelegenheit zur Reflexion bekommen.
- Die Fragen werden gestellt und Einschätzungen unkommentiert gesammelt.
- Zur Relevanz/Zur Verdeutlichung, dass die Ergebnisse auch die Schülerinnen und Schüler betreffen, können zudem Einstellungen von facebook gezeigt werden:
  Ein Profil anzeigen lassen -> Bereich Info ->
  a) Familie und Beziehungen -> Beziehungsstatus: verheiratet /in einer eingetragenen Lebenspartnerschaft lebend
  b) Kontaktinformationen und allgemeine Informationen -> interessiert an: Männern / Frauen
- Zu Frage 2: Durch z.B. facebook mobile (WhatsApp) werden Kanten zu Nicht-Nutzern geschaffen, da die Kontaktliste preisgegeben werden muss. Deshalb gelten die Ergebnisse auch für Nicht-Nutzer. Schattenprofile können auch hier erstellt werden.

  Erinnerung:
  Definition Schattenprofil:
  - für Nutzer des Netzwerks: Profil, das Informationen über euch enthält, die ihr nicht angegeben habt
  - (für Nicht-Nutzer: Profil mit Informationen wird erstellt.)
- Zu Frage 3: Erwartete Antwort: Die Aussage ist richtig. Wir alle sind durch das Preisgeben unserer Daten auch für andere verantwortlich.

Folie 10 | Verbesserungen??

- Folie ist optional. Die Dozenten entscheiden je nach Zeitfortschritt, ob sie besprochen wird.
- Hier können mögliche Verbesserungen diskutiert werden.
- Beim Vereinfachten Problem und beim Mathematischen Modell z.B.:
  - Modellierung als gerichteter Graph
  - Nutzung eines anderen Datensatzes
  - Modellierung mit gewichteten Kanten
  - fehlerhafte Angaben berücksichtigen
- Bei der Mathematischen Lösung z.B.:
  - Andere Heuristiken. (Vorschläge können gesammelt werden.)

Nächster Schritt | Werbeblock und Evaluation

A.6. Evaluationsbogen und Auswertung

CAMMP day

**Wie funktioniert eigentlich ...,
und was hat das mit Mathe zu tun?**

Evaluation

Gymnasium am Wirteltor – 17. Juni 2016

Es besteht immer die Möglichkeit, unsere Programme zu verbessern, und wir würden gerne deine Meinung erfahren. Vielen Dank für deine Rückmeldung.

1. Persönliche Angaben:

Jahrgangsstufe: ____ Geschlecht: □ weiblich □ männlich

Leistungskurse: _____

2. Bewertung des Workshops

soziale Netzwerke □ Google □ GPS □ Solarkraftwerk □

	Trifft gar nicht zu (--)	Trifft eher nicht zu (-)	Trifft zum Teil zu (+)	Trifft voll zu (++)
Durch den Workshop habe ich mathematisches Modellieren besser begriffen.	○	○	○	○
Der Vortrag über Modellierung war hilfreich.	○	○	○	○
Der Umgang mit MATLAB fiel mir schwer.	○	○	○	○
Die Aufgaben waren zu einfach.	○	○	○	○
Die Aufgaben waren zu schwierig.	○	○	○	○
Die Hilfekarten waren hilfreich.	○	○	○	○
Das Hilfesystem in MATLAB war hilfreich.	○	○	○	○
Ich konnte das Problem des CAMMP days lösen.	○	○	○	○
Ich konnte eigenständig arbeiten.	○	○	○	○
Ich konnte mir nach meinen Interessen einen Lösungsweg aussuchen.	○	○	○	○
Ich habe die neuen mathematischen Inhalte verstanden.	○	○	○	○
Der CAMMP day hat mir Spaß gemacht.	○	○	○	○

Kommentar:

Was würdest du am Workshop verändern bzw. verbessern wollen?

_____

_____

_____

_____

_____

Bitte wenden. . .

Was hast du für dich persönlich durch die Teilnahme am Workshop gelernt?

_____

_____

_____

_____

_____

3. Weiterführende Fragen:

	Trifft gar nicht zu (--)	Trifft eher nicht zu (-)	Trifft zum Teil zu (+)	Trifft voll zu (++)
1. Der Workshop hat mein Interesse an den Naturwissenschaften und Technik gesteigert.	○	○	○	○
2. Die Anleitungen zu den Aufgaben waren zufriedenstellend und gut verständlich.	○	○	○	○
3. Die Dozenten haben die Inhalte plausibel und klar dargestellt.	○	○	○	○
4. Die Betreuer haben das selbstständige Arbeiten gefördert.	○	○	○	○
5. Es wurden interessante Berufe und Studiengänge vorgestellt.	○	○	○	○
6. Ich kann mir gut vorstellen, später ein Studium oder eine Ausbildung im Bereich Technik oder Naturwissenschaften aufzunehmen.	○	○	○	○
7. Ich würde einen solchen Workshop gerne noch einmal besuchen.	○	○	○	○

Kommentar:

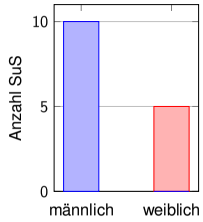
Abschließender persönlicher Kommentar (z.B. Lob, Kritik, Verbesserungsvorschläge):

Evaluation

CAMMP day Gymnasium am Wirteltor
17.06.16

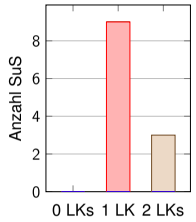
1 Persönliche Angaben

1.1 Geschlecht



m: 10 w: 5

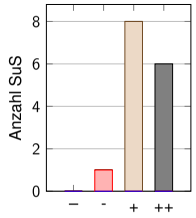
1.2 Anzahl der LKs in Mathematik, Informatik und Physik



0 LK: 0 1 LK: 9 2 LK: 3

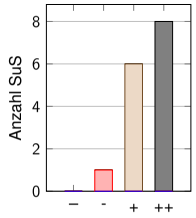
2 Bewertung des Workshops

2.1 Durch den Workshop habe ich mathematisches Modellieren besser begriffen.



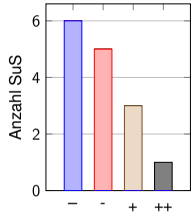
–: 0 -: 1 +: 8 ++: 6

2.2 Der Vortrag über Modellierung war hilfreich.



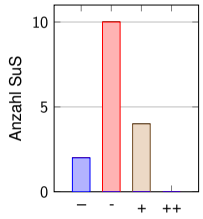
–: 0 -: 1 +: 6 ++: 8

2.3 Der Umgang mit MATLAB fiel mir schwer.



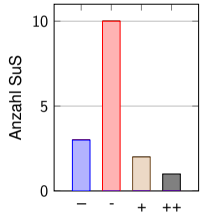
–: 6 -: 5 +: 3 ++: 1

2.4 Die Aufgaben waren zu einfach.



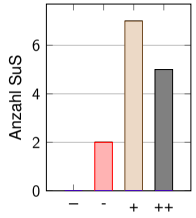
–: 2 -: 10 +: 4 ++: 0

2.5 Die Aufgaben waren zu schwierig.



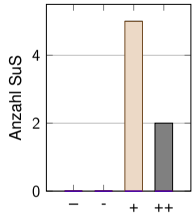
–: 3 -: 10 +: 2 ++: 1

2.6 Die Hilfekarten waren hilfreich.



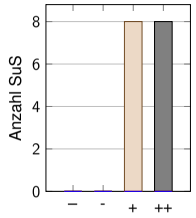
–: 0 -: 2 +: 7 ++: 5

2.7 Das Hilfesystem in MATLAB war hilfreich.



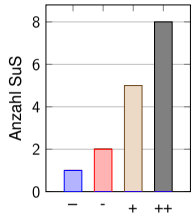
–: 0 -: 0 +: 5 ++: 2

2.8 Ich konnte das Problem des CAMMP days lösen.



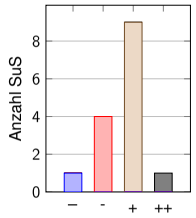
–: 0 -: 0 +: 8 ++: 8

2.9 Ich konnte eigenständig arbeiten.



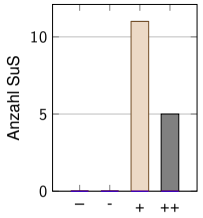
–: 1 -: 2 +: 5 ++: 8

2.10 Ich konnte mir nach meinen Interessen einen Lösungsweg aussuchen.



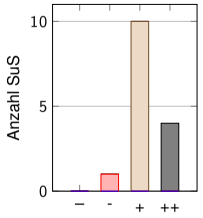
–: 1 -: 4 +: 9 ++: 1

2.11 Ich habe die neuen mathematischen Inhalte verstanden.



–: 0 -: 0 +: 11 ++: 5

2.12 Der CAMMP day hat mir Spaß gemacht.



–: 0 -: 1 +: 10 ++: 4

2.13 Kommentar

2.14 Was würdest du am Workshop verändern bzw. verbessern wollen?

m (Jgst.:): grundsätzlich keine Veränderungen.

m (Jgst.:): individuelles Arbeiten, eigenständig Wege überlegen.

2.15 Was hast du für dich persönlich durch die Teilnahme am Workshop gelernt?

w (Jgst.:): Schüler f. Problematik Datenschutz u. fehlende Privatsphäre in sozialen Netzwerken sensibilisiert $Rightarrow$ effektive Medienkritik.

m (Jgst.:): zu viel.

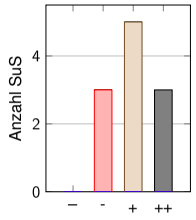
(Jgst.:): soziale Netzwerke sind unheimlich.

w (Jgst.:): Schattenprofile in sozialen Netzwerken.

m (Jgst.:): Modellierung besser verstanden.

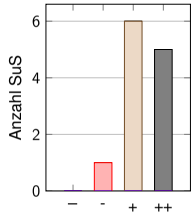
3 Weiterführende Fragen

3.1 Der Workshop hat mein Interesse an den Naturwissenschaften und Technik gesteigert.



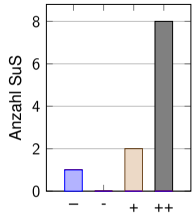
–: 0 -: 3 +: 5 ++: 3

3.2 Die Anleitungen zu den Aufgaben waren zufriedenstellend und gut verständlich.



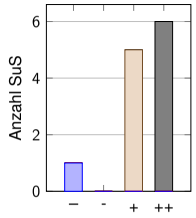
–: 0 -: 1 +: 6 ++: 5

3.3 Die Dozenten haben die Inhalte plausibel und klar dargestellt.



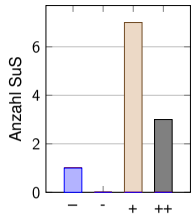
–: 1 -: 0 +: 2 ++: 8

3.4 Die Betreuer haben das selbstständige Arbeiten gefördert.



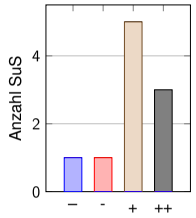
–: 1 -: 0 +: 5 ++: 6

3.5 Es wurden interessante Berufe und Studiengänge vorgestellt.



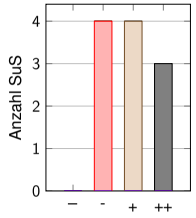
–: 1 -: 0 +: 7 ++: 3

3.6 Ich kann mir gut vorstellen, später ein Studium oder eine Ausbildung im Bereich Technik oder Naturwissenschaften aufzunehmen.



–: 1 -: 1 +: 5 ++: 3

3.7 Ich würde einen solchen Workshop gerne noch einmal besuchen.



–: 0 -: 4 +: 4 ++: 3

3.8 Kommentar

3.9 Abschließender persönlicher Kommentar

w (Jgst.:): Schüler konnten viel selbst ausprobieren, Einsatz techn. Hilfsmittel sehr motivierend.

(Jgst.:): gut gemacht.

w (Jgst.:): gut so.

m (Jgst.:): tip top.