HAW HAMBURG

MASTERTHESIS
Maria Lüdemann

Quantified Self - eine explorative Selbststudie

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE
WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Maria Lüdemann

Quantified Self - eine explorative Selbststudie

Masterarbeitarbeit eingereicht im Rahmen der Masterarbeitprüfung
im Studiengang Master Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Dr. Susanne Draheim

Abgegeben am 19. Mai 2020

Maria Lüdemann

Data Mining auf Consumer Sensor Daten für Quantified Self

Quantified Self, Companion Technologie, Datenzentralisierung, Data Minung

Kurzzusammenfassung

Daten aus Consumer-Sensoren wie Fitnessarmbändern, Blutdruckmessgeräten etc. lädt jeder Anbieter separat in seine Cloud hoch. Welchen Nutzen können diese Daten haben, wenn sie dem Anwender zentralisiert zur Verfügung stehen und aus allen Bereichen gesammelt und mit manuellen Daten angereichert analysiert werden? Diese Arbeit betrachtet, ob eine derartige Zentralisierung möglich ist und somit ein Grundstein für eine Plattform gelegt werden kann, auf der Companion-Systeme aufsetzten können um Nutzer zu unterstützen, ihre persönlichen Daten sinnvoll zu nutzen.

Maria Lüdemann

Data Mining with consumer sensor data for Quantified Self

Quantified Self, Companion Technology, data centralization, Data Minung

Abstract

Every manufacturer of consumer sensors such as activity trackers, blood presure monitors etc. uploads the data into their own cloud. What benefit can this data provide, if centralised from all domains and enhanced by manualy entered data. This bachelor thesis shows, if such centralisation is feasible and therefore lays the groundwork for a platform that provides companion systems which support the user in making effective use of their personal data.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Technik und Sensorik wird seit Jahren immer kleiner. Computer haben sich von raumfüllenden Monstren auf schlanke Platinen verkleinert. Sie sind allgegenwärtig und aus dieser Allgegenwärtigkeit, dieser Miniaturisierung entsteht die Möglichkeit immer mehr messbar zu machen. Die Welt, die den Menschen umgibt wird datafiziert. Selbst Alltagsgegenstände bis hin zu Textilien oder Pflanzen können und werden mit Sensoren versehen. Die daraus entstehenden Datenmassen müssen verarbeitet werden. Ein sehr wichtiger Aspekt dieser Verarbeitung ist die Kontextualisierung und Interpretation der Daten. Nur dadurch können die Daten nutzbar gemacht werden. Diese Schritte sind so wichtig und gleichzeitig so komplex, dass sich daraus eine eigene Disziplin innerhalb der Informatik gebildet hat, die Data Science.

Diese Entwicklung stoppt nicht bei unserer Umwelt, sie greift auch auf den Menschen und seinen Körper und sein Leben über. In den letzten Jahren entwickelte sich das Themengebiet Self Tracking und Quantified Self von einer Randerscheinung für Technikbegeisterte und aus medizinischen Gründen dazu gezwungenen Personen, hin zu einer weit verbreiteten Erscheinung. Neben einem immer weiter wachsenden Markt für technische Geräte zum Erfassen verschiedenster Körperfunktionen und Werte gibt es eine schiere Flut von Apps, die dem Benutzer helfen sollen, die verschiedensten Aspekte ihres Lebens zu erfassen, zu protokollieren, erinnerbar und messbar zu machen. Allein für den englischsprachigen Markt führt die Homepage (quantifiedself.com/guide/tools) über fünfhundert Tools auf, eine andere Quelle Felizi und Varon (2018) beschreibt allein 225 verschiedene Apps zur Erfassung der Menstruation. Dabei sind die Apps aus dem rein deutschsprachigen Raum nicht mit einbezogen.

Selbsterfassung ist an sich nichts Neues, jedoch ist das Ausmaß und die technische Unterstützung, in der Self Tracking mittlerweile passiert, etwas Neues. Es hat den Anschein, als sei dies der Ausdruck des 21. Jahrhunderts. Der Weg zu seinem Selbst führt über Zahlen, was Meditation im letzten Jahrhundert war ist nun Vermessung mit Technik. Zeitgleich entsteht ein Bedürfnis nach Rationalität, Nachvollziehbarkeit, Kontrolle und vielleicht auch Berechenbarkeit. In einer Zeit, in der die Umwelt immer unverständlicher, hektischer und komplexer wird, entsteht vielleicht auch ein Bedürfnis

danach, den eigenen Körper verstehen und berechnen zu können um daraus Sicherheit und Ruhe zu ziehen. Allerdings steckt wohl auch zu einem sehr großen Teil die menschliche Neugierde dahinter, denn zu keinem Zeitpunkt in der Geschichte war es so einfach wie heute, relativ genaue Daten mit einem derart geringen Aufwand zu erfassen, zudem ist es möglich und erschwinglich Lupton (2016, 2013) (1-37).

Immer neue Erkenntnisse revidieren altes Wissen, immer mehr Super Foods erscheinen auf dem Markt und die Möglichkeiten der sportlichen Betätigung bilden eine schiere Flut. Das heißt, ein Mensch kann schnell überfordert sein vom Versuch sich gesund zu ernähren und zu bewegen, was auch immer gesund im aktuellen Jahr bedeutet. Neben dieser äußeren Verwirrung kommt eine innere dazu. Die Welt wird immer schneller, die Anforderungen immer größer und es wird immer weniger auf sich selbst geachtet. Das Vertrauen in die innere Stimme, das Körpergefühl, kann dadurch verloren gehen oder die Wahrnehmung dessen wird verlernt. Wie glaubwürdig ist das eigene Gefühl? Wir wissen, dass unsere Erinnerungen uns trügen können, dass sich unsere Sicht der Dinge im Nachhinein verklären kann. Aber wie ist das mit dem Gefühl? Kann es sein, dass wichtige Signale nicht gehört oder falsch gedeutet werden? Oder dass sie im schlimmsten Fall vom Körper nicht mehr gesendet werden? Es gibt Menschen, die nicht genug trinken, weil sie das Durstgefühl als Hunger interpretieren oder einfach keines haben. Sie merken dann vielleicht erst, dass sie zu wenig trinken wenn sie Kreislaufprobleme haben, falls sie überhaupt den Zusammenhang ziehen. Dies ist ein Punkt, an dem Erfassung helfen kann. Das Erfassen der Menge, die getrunken wird, führt zu einem genauen Wert, der das vage Gefühl ersetzen kann. Doch wie sieht es mit anderen Aspekten des Lebens, der Gesundheit und des Gefühls aus? Dies ist der Punkt, an dem Technik, an dem die Informatik im Rahmen von Datenerfassung und Verarbeitung, also im Rahmen von Data Science, helfen könnte. Ist es vielleicht möglich ein System zu bauen, das die Einsicht in die Black Box des Körpers ermöglicht? Welches das vage Gefühl trainiert oder im Zweifel ersetzen kann?

Es gibt eine Vielzahl an Möglichkeiten, Daten zu erfassen und in Teilen auch auszuwerten, jedoch lassen es die vorhandenen Lösungen oft an Transparenz und Vertrauenswürdigkeit vermissen. Der Nutzer hat zumeist keine Kontrolle über die eigenen Daten und kann sie somit selten sensorübergreifend auswerten. Es ist oft für den Nutzer nicht klar, wer Zugriff auf seine eigenen Daten hat und was diejenigen damit tun. Außerdem kann er schlecht abschätzen wie glaubwürdig die Daten und Sensoren sind, wie sehr er seinem digitalen Spiegelbild vertrauen kann. In dieser Arbeit soll daher näher betrachtet werden, welchen Nutzen Self-Tracking-Daten, die in einer vertrauenswürdigen Umgebung analysiert wurden, haben. Kann die Analyse von Daten, die im Alltag möglichst bequem erfasst werden, Erkenntnisse und Nutzen für den Benutzer bringen, auf welche Art kann der Benutzer diese Daten am besten erleben und welche Aufwände stecken darin, die Daten zu sammeln, aufzubereiten, zu analysieren und am Ende benutzergerecht anzuzeigen. Ein geeignetes Anzeigemedium soll dabei eben-

falls betrachtet werden. Inwieweit kann Technik als vertrauenswürdiger Companion zwischen dem Mensch und seinen Daten vermitteln?

1.2 Ziele und Abgrenzung

Das Ziel dieser Arbeit ist es zu überprüfen, ob mit handelsüblichen Sensoren ein Quantified-Self-System aufgebaut werden kann, das mithilfe von Mitteln der Datenverarbeitung Teilaspekte des Lebens des Nutzers quantifizieren kann und darüber hinaus dem Nutzer selbst Erkenntnisse liefert. Dafür soll eine mögliche Architektur für ein solches System entwickelt werden, deren Teile dann exemplarisch umgesetzt werden, um die generelle Umsetzbarkeit zu zeigen. Darüber hinaus soll überprüft werden, ob dieses System dem Nutzer Erkenntnisse über sich selbst, seinen Körper oder sein Leben ermöglichen kann, die ihm vorher noch nicht bewusst waren. Dabei soll anstatt eines Feldversuches mit vielen Nutzern ein qualitativer Selbstversuch genutzt werden, der über eine längere Zeit läuft. Dabei wird ein zeitlicher Rahmen von etwa drei Jahren angepeilt. Dafür müssen die Daten verarbeitet und analysiert werden. Es ist abzusehen, dass darauf wohl das Hauptaugenmerk dieser Arbeit liegen wird, um Ergebnisse erzielen zu können. Dabei soll ein Überblick gegeben werden über den aktuellen Stand der Wissenschaft im Bereich Quantified Self und Datenanalyse. Um die Daten zu verarbeiten soll der KDD(Knowledge Discovery in Databases) Prozess herangezogen werden.

Abgrenzung
Das Ziel dieser Arbeit wird nicht sein, ein funktionierendes marktreifes System zu entwickeln oder eine medizinische Qualität zu erreichen. Es ist ebenso nicht Ziel die einzelnen Aspekte des Systems bis zu ihrer Perfektion auszuarbeiten, es reicht völlig, ihre Umsetzbarkeit zu testen.

1.3 Gliederung

Die Arbeit teilt sich in 5 Kapitel. Das erste Kapitel gibt eine Einführung in das dieser Arbeit zugrunde liegende Themengebiet. Das Kapitel 2 ist die Analyse des wissenschaftlichen Umfelds, das sich in Quantified Self 2.1, intelligente Spiegel 2.2 und Datenverarbeitung und Analyse im Rahmen eines geeigneten Prozesses 2.3 aufteilt. Darüber hinaus werden im Abschnitt 2.4 die vorliegenden Daten und ihre Qualität analysiert.

Das Kapitel 3 ist das Entwurfs- und Implementationskapitel. Dort wird zunächst auf das geplante System und dessen Architektur eingegangen 3.1. Die folgenden Abschnitte beschreiben dann die Umsetzung und Implementierung des Systems anhand der einzelnen Komponenten beziehungsweise Phasen aus dem KDD Prozess 3.2. Darauf folgt die Beschreibung des technischen Aufbaus des Systems als intelligenter Spiegel im Labor 3.8.

Das Kapitel 4 ist die Evaluation der Ergebnisse und des Systems. In diesem Kapitel wird zunächst auf die Datenerhebung und Bereinigung 4.1 eingegangen. Dabei wird beleuchtet, welche Probleme und Erfahrungen sich daraus ergaben. Folgend werden die Ergebnisse der Datenanalysen sowie deren Interpretation diskutiert 4.2. Danach wird kurz abschließend auf die Datenqualität 4.3 eingegangen sowie eine Systemevaluation 4.4 durchgeführt.

Im Kapitel 5, dem Schluss, wird die Schlussbetrachtung besprochen. Zunächst wird die Arbeit noch einmal zusammengefasst 5.1. Dann wird aus der Arbeit ein Fazit 5.2 gezogen und als letztes ein Ausblick 5.3 geliefert.

2 Analyse

In der Analyse wird das Umfeld der Arbeit betrachtet, dabei wird ein Augenmerk auf Quantified Self wie auch intelligente Spiegel und Datenverarbeitung sowie den KDD Prozess gelegt. Es wird genauer betrachtet, was für Daten in welcher Qualität vorliegen, wie sie beschaffen sind und wie mit ihnen umgegangen werden muss um sie verarbeiten zu können. Dafür wird ein Validitätsmodell erstellt und besprochen sowie Wege aufgezeigt, um die Datenqualität festzustellen und zu erhöhen. Darüber hinaus wird darauf eingegangen welche Möglichkeiten der Datenanalyse zum Erkenntnisgewinn bestehen.

2.1 Quantified Self

Dieser Abschnitt betrachtet Quantified Self (QS) und das wissenschaftliche wie auch das allgemeine Umfeld. Zunächst wird geklärt, was Quantified Self ist und welche Begrifflichkeiten darum herum benutzt werden. Es wird betrachtet, welche Motivation dieser Bewegung zu Grunde liegt und ein Überblick über aktuelle Sensorik und Technik gegeben. Zudem wird betrachtet, welche Probleme und Chancen sich sowohl technisch wie auch moralisch im Rahmen von Quantified Self ergeben. In der Betrachtung des wissenschaftlichen Umfelds wurde ein Augenmerk auf Arbeiten gelegt, die sich mit dem Erfassen der Daten und dessen Umfeld beschäftigen, um von deren Erkenntnissen zu profitieren.

Begriffserklärung und Erläuterung

Quantified Self ist ein Begriff, der stark von der gleichnamigen Bewegung bzw. Gruppe [1] geprägt wurde. Als Quantified Self, sowohl als wissenschaftliches Feld wie auch als Bewegung immer populärer wurden, wurden diverse Begrifflichkeiten geprägt, darunter Personal Informatics, Self Tracking, Lifelogging, Life Hacking und Body Monitoring. Nach und nach setzte sich aber Quantified Self und Self Tracking durch (Lupton, 2016, S.9-16). Die Begrifflichkeiten wurden oft synonym verwendet, beinhalten aber teilweise

[1] http://quantifiedself.com/

leicht unterschiedliche Bedeutungen. Eine genauere Beschreibung der jeweiligen Bedeutung der Begriffe und deren Differenzierung findet sich in Kamenz (2014); Emmert (2013); Lüdemann (2016b)

Quantified Self oder auch die Vermessung seines Selbst definiert sich durch das Erfassen quantitativer und qualitativer Daten über sich selbst oder der direkten Umwelt quantified self institute (2016). Dies zeigt sich auch im Motto der gleichnamigen Bewegung *"Self knowledge trough numbers"* [2].

Dabei kann gesagt werden, dass Quantified Self das Umfeld und das wissenschaftliche Gebiet sowie die bereits erwähnte Bewegung und Self Tracking die Tätigkeit darin ist. Self Tracking ist also das Erfassen der Daten in einem Quantified-Self-Kontext, beispielsweise zum Erkenntnisgewinn. Welche Motivation hinter dem Erfassen von Daten stecken kann, zeigt der Abschnitt 2.1.2. Im Rahmen dieser Arbeit wird von Quantified Self gesprochen, wenn die Bewegung oder das wissenschaftliche Feld gemeint ist. Self Tracking wird verwendet als Bezeichnung der Datenerfassung über sich selbst.

Die Quantified Self Bewegung

Da im Rahmen dieses Abschnitts immer wieder von der Quantified-Self-Bewegung gesprochen wird, soll diese hier beschrieben werden, um ein grundlegendes Verständnis dafür zu schaffen. Im Jahr 2007 gründeten Gary Wolf und Kevin Kelly die Bewegung und verbreiteten den Quantified-Self-Begriff. Es ging ihnen dabei auch darum, eine Plattform zu bieten, auf der zu Themen rund um Quantified Self diskutiert und Wissen ausgetauscht werden kann. Ebenso sollte eine Möglichkeit etabliert werden, um Disskusionen und Austausch über das Erfassen und dessen Erfolge zu ermöglichen. Selbstvermesser konnten somit einfacher mit Gleichgesinnten in Kontakt treten, ihre Erfahrungen teilen und von denen anderer profitieren. Mittlerweile finden neben regelmäßigen Meetups [3] auch Konferenzen in diesem Rahmen und mit dieser Thematik statt[4]. Die Bewegung erfreut sich außerdem in immer mehr Ländern an wachsendem Zuspruch Meetup (2018), so umfasst allein das Quantified-Self-Meetup mittlerweile ca. 90.000 Mitglieder.

Neben den Forschungen und Tests innerhalb der Gruppe gibt es einige Forscher verschiedenster Disziplinen, die sich die Bewegung genauer angeschaut haben. Es wurde betrachtet, was die Quantified Selfler erfassen und wie. Aber auch warum sie dies überhaupt tun. Was bewegt den Menschen, sich teilweise extrem selbst zu überwachen? Wie groß muss ein Nutzen sein, damit ein Mensch sich dazu entscheidet zu einem gläsernen Menschen zu werden?

[2] http://quantifiedself.com/
[3] https://www.meetup.com/de-DE/topics/quantified-self/
[4] http://qs18.quantifiedself.com/

In ihrer Arbeit haben Whooley u. a. (2014) 51 Videos ausgewertet, in denen Quantified Selfler beschreiben, was, wie und warum sie Selbstvermessung betreiben und welche Probleme technischer oder menschlicher Art ihnen dabei begegnen. In ihrer Arbeit wird systematisch ausgewertet, welche Intentionen hinter den unterschiedlichen Ansätzen der Quantified Selfler stecken, wie sehr dabei die Datenerfassung automatisiert wurde und welche Repräsentationen der Daten genutzt werden. Eine andere Arbeit Choe u. a. (2014) beschäftigt sich damit, welche Probleme und Barrieren sich beim erfassen, speichern, analysieren und auswerten der Daten ergeben und wie QSler damit umgehen. Dabei wird auch genauer betrachtet, welche demographischen Daten die QSler mitbringen und ähnlich wie in der Arbeit von Whooley u. a. (2014) wird auch bei Choe u. a. (2014) genauer analysiert, welche Motivation und welches Ziel hinter dem Erfassen steckt, sowie was mit welchen Werkzeugen erfasst und verarbeitet wird. In dieser Arbeit steht die Problembehandlung etwas mehr im Vordergrund.

2.1.1 Selbstvermessung

Selbstvermessung als solches ist keine neue Erscheinung, allein die Ausmaße und die technische Unterstützung sind durch moderne Mittel stark verändert (Lupton, 2016, S.9). Das Erfassen der Daten und die Neugierde über sich selbst ist jedoch dem Menschen praktisch seit jeher gegeben. Self Tracking kann sehr unterschiedlich betrieben werden und ist dabei keineswegs an technische Mittel gebunden.

Neben den offensichtlich erscheinenden Arten wie das Erfassen von Gewicht in Diäten oder das Überwachen von Blutdruck oder Blutzucker auf Anraten des Arztes gibt es weniger offensichtliche Arten im Alltag, Daten zu erheben. Eine der gängigsten Arten ist das Tagebuch schreiben, aber auch die Striche am Türrahmen um die Größe der Kinder zu erfassen und mit der Zeit zu vergleichen ist eine Art der Datenerfassung. Sogar das Aufbewahren der Zeichnungen der eigenen Kinder stellt in einer Weise ihre motorische und psychologische Entwicklung bildlich dar (Lupton, 2016, S.29). Dazu kommen Fotos, Reisetagebücher, Jahresbücher, Briefwechsel, Traumtagebücher, Haushaltsausgaben, eine Liste der gelesenen Bücher oder gesehenen Filme, all dies sind gängige Arten der alltäglichen Datenerfassung.

Neben diesen alltäglichen Arten der Selbsterfassung gab es auch schon früh die wissenschaftliche Neugierde und den Wunsch nach größerem Verständnis für die Funktionsweise des Körpers. Ein relativ altes Beispiel bildet dabei der italienische Mediziner Santorio Santorio aus dem 16. Jhdt. Er wollte die Funktion des menschlichen Stoffwechsels ergründen und erfasste dafür in einem Zeitraum von dreißig Jahren sein Gewicht und seine Ernährung Neuringer (1981) zitiert nach Swan (2013). Heutzutage wird ein Patient von den Ärzten immer häufiger dazu angehalten, längerfristig Daten zu erheben, seien es regelmäßige Bluttests zum richtigen Einstellen einer Medikation oder eine Überwachung des Blutdrucks. Für Diabetiker des Typs 1 ist das Vermessen

längst habitualisiert und unabdingbar, denn das Gefühl reicht nicht, um einschätzen zu können, ob der Körper in Ordnung ist. Erst die erfassten Zahlen bringen eine Einschätzbarkeit und können den Diabetiker einem Selbstgefühl näher bringen, da die Zahlen als Vermittler zwischen Gefühl und Realität eintreten.

Self Tracking oder Quantified Self in dem heute möglichen Ausmaß ist relativ jung und damit auch die Forschung um dieses Thema, sowohl der Nutzen wie auch die Gefahren, die darin stecken sind noch nicht vollständig erfasst. Weder von der Forschung noch von den eigentlichen Benutzern. Ebenso fehlt es an fundierten Methoden zum erfassen, auswerten, speichern und auch anwenden der gesammelten Daten bzw. den daraus generierten Erkenntnissen. So bemängelt zum Beispiel die Arbeit von Kersten-van Dijk u. a. (2017), dass noch kein allgemeingültiger Weg gefunden wurde, um Menschen eine Optimierung zu ermöglichen.

2.1.2 Motivation der Selbstvermessung

Menschen haben viele Arten, Daten über sich zu erheben und das aus ganz unterschiedlichen Gründen. In ihren Arbeiten über die Quantified-Self-Bewegung haben Whooley u. a. (2014) und Choe u. a. (2014) einige davon herausgearbeitet. Diese waren natürliche Neugierde, entweder an der Technik die benötigt wird um die Daten zu erfassen, oder jene auszuwerten oder Interesse an der Funktionsweise des Körpers selbst. Aber auch der Wunsch nach einer Verbesserung. Diesen Verbesserungswunsch schlüsselte die Arbeit von Choe u. a. (2014) noch einmal in die Verbesserung des Gesundheitszustandes und der Verbesserung eines Aspekts des Lebens auf. Dies kann der Wunsch sein, die Arbeits- oder Lernleistung zu optimieren, sich mehr an gewisse Ereignisse zu erinnern oder generell umsichtiger und achtsamer zu sein. In der Arbeit von Choe u. a. (2014) wurde noch ein Motivationsgrund gefunden, bei dem Daten erfasst werden um neue Erfahrungen zu machen. Zum Beispiel sich zu motivieren neue Orte zu entdecken indem sie getrackt werden oder beim Auswerten der Daten Neues zu lernen. In der Arbeit von (Wiedemann, 2016, S.76) wurde ein Interview mit einem QSler beschrieben, der aus rein präventiven Gründen Daten erfasst. Er begründet dies damit, dass er erst weiß welche Daten er braucht, wenn eine Anomalie merkbar auftritt wie zum Beispiel körperliche Beschwerden. Zu diesem Zeitpunkt ist es jedoch zu spät um die Daten die zu dieser Anomalie geführt haben zu erfassen. Somit sammelt er möglichst kleinteilig alle möglichen Daten, um sie im Zweifel vorliegen zu haben und auswerten zu können.

In Arbeiten, die sich nicht direkt mit den Mitgliedern der Quantified-Self-Bewegung beschäftigen sondern generell mit Menschen die Daten über sich selbst oder ihre Umwelt erfassen, werden noch weitere Motivationen beschrieben. Zum Beispiel, um ein Verhalten anzupassen, dabei kann es um Gesundheitsthemen gehen, wie gesünder Essen oder mehr Bewegung aber auch seine Ausgaben besser im Griff zu haben, besser

zu schlafen oder produktiver zu sein Li u. a. (2011); Kamal u. a. (2010). Ebenso gibt es Menschen, die ihre Daten erfassen müssen oder wollen aufgrund von chronischen Erkrankungen wie Diabetes, Epilepsie oder Migräne MacLeod u. a. (2013); Li u. a. (2011); Rooksby u. a. (2014). Ein sportliches Interesse bezeichnet (Wiedemann, 2016, S.78) als eine der häufigsten Gründe für nicht QSler Daten zu erfassen. Also das Aufzeichnen von sportlichen Aktivitäten, deren Vergleich und Auswertung.

Immer wieder wird Neugierde angeführt MacLeod u. a. (2013), Whooley u. a. (2014), Neugierde auch über sich selbst. Wie beschreibe ich mich in Zahlen, wie sieht mein 'Zahlen-Ich' aus? Das digitale Ich, das so mancher vielleicht einfacher zu verstehen scheint als den eigenen Körper und das Gefühl. Eine Flucht ins Rationale, um die Black Box Körper verständlicher zu machen. In Zeiten, in denen Informationsüberfluss herrscht und uns sowohl falsche wie auch wahre Informationen in einer schier unfilterbaren Flut begegnen, steigt der Wunsch nach Sicherheit, da immer weniger Zeit für Besinnung und Ruhe bleibt Lupton (2014) (Wiedemann, 2016, S.66). Somit steckt im Willen zur Selbsterfassung oft ein Wunsch nach Kontrolle. Das eigene Gefühl scheint zu trügen, es ist in der Regel getrübt von Emotionen oder der Tagesverfassung. Emotionen lassen sich nicht be- oder verrechnen, vergleichen und vermitteln. Es fehlt ihnen das Rationale und Handhabbare, was sie im modernen Sinne handhabbar machen würde. So schrieb auch Gary Wolf

> "Numbering things allows tests, comparisons, experiments. Numbers make problems less resonant emotionally but more tractable intellectually." Wolf (2010)

Zahlen räumt man eher ein nicht zu trügen, greifbarer und in unserer Gesellschaft glaubhafter zu sein als ein vages Gefühl. Des Weiteren lassen sich Zahlen besser vergleichen, verarbeiten und berechnen. Sie scheinen besser in unsere rationale, schnelle Gesellschaft zu passen und vermitteln das Gefühl einer Objektivität, wobei dabei natürlich stark darauf zu achten ist, wie die Daten erhoben wurden. Daten können nur so objektiv sein wie die Umstände ihrer Erhebung.

Jedoch scheinen Zahlen Ausdruck unserer Gesellschaft zu sein und nach und nach alle Lebensbereiche einzuholen. Der Mitgründer der Quantified-Self-Bewegung drückte dieses Gefühl in folgendem Zitat aus:

> "We use numbers when we want to tune up a car, analyse a chemical reaction, predict the outcome of an election. We use numbers to optimize an assembly line. Why not use numbers on ourselves?" Wolf (2010)

Zahlen stellen außerdem ein klares Ziel dar. Sportlicher zu werden ist viel schwerer zu erreichen als drei Kilo abzunehmen, zwei Prozent mehr Muskelmasse zu bekommen oder die fünf Kilometer zwei Minuten schneller zu joggen als bisher. Gefühle lassen sich nicht berechnen, Zahlen schon. Allein zum erreichen gewisser Ziele oder zum optimieren gewisser Bereiche ist es notwendig, ein optimierbares Spektrum zu schaffen.

In der Arbeit von Li u. a. (2011) werden sechs Fragekategorien benannt, in die sich die Fragen der Selftracker einteilen, die sie sich über ihre Daten stellen. Die Kategorien sind: Status, Historie, Ziele, Diskrepanzen, Kontext und Faktoren. Daraus ergeben sich die meisten Motivationen, die Self Tracker haben und die hier benannt wurden.

2.1.3 Technik und Sensorik

Durch die Technik wurde der Neugierde des Menschen ein immer stärker werdendes Werkzeug an die Hand gegeben. Das Erfassen von Körperdaten wird immer einfacher, zum einen dadurch, dass die Sensoren immer kleiner werden und auch in Kleidung verbaut werden können somit immer tragbarer werden. Zum anderen dadurch, dass es Firmen [5] gibt, die Sensorik benutzerfreundlich und praktikabel auf dem Markt anbieten.

Die Akzeptanz der Benutzer steigt stetig, das Leben wird immer digitaler (Lupton, 2016, S.38), Fitnessarmbänder werden besonders von jungen sportlichen Menschen gern getragen und kommen auch in technikferneren Kreisen an Lupton (2016). Es ist nicht mehr nur ein Nerdspielzeug, sondern ein viel benutztes Werkzeug. Smart Watches, die häufig auch gewisse Sensorik beinhalten, werden immer häufiger als Erweiterung zum Smartphone getragen.

Smartphones selbst sind bereits omnipotente Sensoren und in der Lage, weite Teile des Lebens aufzuzeichnen und werden von immer mehr Menschen genutzt Statista. Das Smartphone kann sagen, wann es wo und wie lange war, welche Fotos dort gemacht wurden, mit wem man dort gechattet oder telefoniert hat. Oder auch auf welchen Websites man dort surfte oder was man dort im Internet bestellte. Das Smartphone kennt unsere sozialen Kontakte, in etwa wann wir aufstehen und wann ins Bett gehen, alles anhand der Benutzung. Dienste wie Google werten diese Daten auch immer umfassender aus und verwandeln sie in zusätzliche Dienste. So wertet Gmail zum Beispiel E-Mails aus und setzt auf Google Maps Marker für gebuchte Hotelzimmer oder schreibt Termine für Flüge in den Google Kalender, völlig automatisch wird so das Leben vermessen und digitalisiert.

Neben spezieller tragbarer Sensorik ist das Smartphone ein gutes und sehr weit verbreitetes Gerät, um Daten zu erfassen. Die eingebauten Sensoren können neben Schritten und Standort auch mithilfe der Kamera den Puls erfassen. Es ist ein Gerät, das den Menschen den ganzen Tag begleitet und selbst nachts häufig neben dem Bett liegt. Es gibt kaum ein privateres Gerät, auf dem neben der sozialen Kommunikation und dem E-Mail-Verkehr auch private Fotos und Videos erzeugt, geteilt und konsumiert werden können (Lupton, 2016, S.41). Viele Smartphones werden für Bankverbindungen, Authentifizierung, Zahlvorgänge, Bestellungen und die Bewegung im öffentlichen Raum

[5] Fitbit, Jawbone, Apple etc.

benutzt (DB, Car2Go, MyTaxi, Stadtradt etc.). Hat man Zugang zum Smartphone einer Person, bedeutet dies gleichsam Zugang zu weiten Teilen des Lebens. Das Smartphone speichert, wann wir uns wohin wie lange bewegen und mit wem. In einer Arbeit von Kamal u. a. (2010) wird angemerkt, dass Nutzer auf andere Möglichkeiten für Datenerfassung zurückgreifen, wenn Apps oder Sensorik gewisse Daten nicht erfassen bzw. kein Platz für Notizen lassen. Die Apps selbst befinden sich auf dem Smartphone und auch die Notizen werden unter anderem auf dem Smartphone gespeichert. Es wird zum Tagebuch, zum Protokoll für Gesundheit, Ernährung, Sport und vieles mehr.

Sensorik

In diesem Abschnitt wird darauf eingegangen, was im Augenblick gängige Sensoren sind, sowie welche Sensoren gerade entwickelt werden um ein Gefühl dafür zu vermitteln welchen Stand die Technik hat.

Neben der immer mehr verbreiteten Smartphone-Nutzung wird auch die Nutzung tragbarer Sensorik immer anerkannter. Durch das hohe Interesse der Self Tracker sehen die Firmen auch einen Anreiz daran, immer bessere Technik zu entwickeln. Dies kommt auch einigen chronisch Kranken zugute, da nun durch die höhere Nachfrage mehr Geld im Markt steckt und die auch von chronisch Kranken benötigte Sensorik immer besser und komfortabler wird (Wiedemann, 2016, S.78). Die Sensorik wird dabei nicht nur kleiner und somit tragbarer, sondern auch robuster, preiswerter und genauer. Tragbare Sensorik ist im Endverbrauchermarkt angekommen.

Eine sehr bekannte und weit verbreitete Sensorik ist das Fitnessarmband beziehungsweise die Smart Watch. Dabei gibt es Ausprägungen von sehr auf Sport bezogenen Geräten, zum Beispiel von Fitbit und Garmin über Jawbone, bis zu den Smart Watches, die hauptsächlich das Telefon erweitern und Schritte sowie Bewegungen überwachen und darüber auch z.B den Schlaf. Bekannte Hersteller sind dort unter anderem Apple, Fossil und Fitbit, dabei ist darauf zu achten, dass die Grenzen zwischen Fitness Armband und Smart Watch zunehmend verschwimmen. Die Fitness Armbänder bekommen eine immer bessere Smartphone Integration, sodass das Fitbit Blaze bereits Musik auf dem Smartphone steuern und Nachrichten empfangen kann. Dagegen haben neue Smart Watches bereits GPS, Barometer, Pulssensoren und immer mehr Fitness Features. Neben den Armbändern um Puls zu messen gibt es auch Brustgurte, die direkt am Brustkorb anliegend genauere Informationen über das Herz liefern können [6].

Der Neugierde sind nur wenig Grenzen gegeben, es gibt ebenfalls Analysewaagen für Körpergewicht, Knochen-, Muskel- und Wasseranteil. In den einfachen Ausführungen kann man sich einfach daraufstellen, komplexere Geräte leiten den Messstrom, mit

[6] https://buy.garmin.com/de-DE/DE/p/15490

Handgriffen, auch durch den Oberkörper. Die Ergebnisse lassen sich auf dem Smartphone anschauen. Zum Messen der Schlafgewohnheiten und Qualität können neben diversen Armbändern auch Sensoren genutzt werden, die neben das Bett gestellt werden [7] oder als Matte auf der Matratze liegen [8]. Dabei wird gemessen, wann man einschläft, wann aufwacht, wie oft aufgewacht wird und wann man sich wie lange in welcher Schlafphase befindet. Neben der direkten Schlafqualität gibt es auch Sensoren, die die Qualität der Luft im Schlafzimmer messen sollen. Sauerstoff, CO2 Gehalt, Luftfeuchtigkeit und Temperatur werden gemessen um Faktoren für guten oder schlechten Schlaf besser erkennen zu können.

Nach dem Aufstehen kann das Erfassen gleich weitergehen. Einige Toilettenhersteller planen Analysetoiletten, die den Urin des Benutzers auffangen und ihn analysieren. Dies beschreibt zum Beispiel die Firma Duravit, die einen Prototyp vorgestellt hat, der die Ergebnisse auch direkt an eine App weiterleiten soll [9].

Des Weiteren gibt es unter anderem Blutdruckmessgeräte. Diese gibt es sowohl für das Handgelenk wie auch für den Oberarm zum manuellen Messen. Die Daten werden dann per Bluetooth an das Smartphone übermittelt. Auch Blutzuckermessgeräte sind mittlerweile digital und an das Smartphone angeschlossen. Die Blutzuckermessgeräte sind entweder so erhältlich, dass sie nach dem Piksen in den Finger die Daten an ein Smartphone übertragen [10] oder so, dass sie direkt an das Smartphone angesteckt werden können und die Daten direkt übertragen werden [11]. Neben dem relativ herkömmlichen Weg gibt es mittlerweile Messgeräte, die im Oberarm direkt in die Haut gesteckt werden und dort verbleiben. Spezielle Scanner können dann, an den Sensor gehalten, die Daten auslesen [12].

Ein weiteres Gebiet, in dem immer mehr Sensoren zu finden sind, ist intelligente Kleidung. Dort finden sich unter anderem Fitness Shirts, die vitale Parameter erfassen wie Puls und Stress um daraus einen Fitnesslevel zu errechnen und den Trainingsfortschritt dadurch unterstützen sollen [13]. Neben Fitnessshirts gibt es auch Socken die helfen sollen, das Joggen zu optimieren, diese Socken erfassen wie der Läufer auftritt und den Fuß abrollt. Damit kann ein optimaler Laufstil besser trainiert werden. Neben dem Fitnessgedanken werden die intelligenten Kleidungsstücke auch für andere Zwecke gebaut. So gibt es eine Socke, die den Fuß und seine Druckpunkte vermisst,

[7] https://sleeptrackers.io/sense/
[8] https://health.nokia.com/de/de/sleep
[9] https://www.presseportal.de/pm/106317/3585190
[10] https://www.medisana.de/Gesundheitskontrolle/Blutzuckermessgeraete/MediTouch-2-connect-dual-Blutzuckermessgeraet-inkl-Starterset.html
[11] http://mydario.de
[12] https://www.freestylelibre.de/libre/
[13] https://www.ambiotex.com/,https://www.iis.fraunhofer.de/de/magazin/2017/fitnessshirt.html,https://www.mindtecstore.com/Sensoria-Fitness-T-Shirt

um zum Beispiel orthopädischen Schuhmachern bei der Herstellung optimaler Schuhe zu helfen [14]

Dies waren nur einige Beispiele für die Bandbreite an Sensoren, die es mittlerweile auf dem Markt oder noch in Entwicklung gibt.

Apps

Neben den Sensoren, die zumeist ohnehin auch Apps haben, gibt es einen sehr großen Markt an Apps zum Erfassen von Daten. Auf der Seite der Quantified Self Bewegung Self (2018) werden allein über fünfhundert der gängigeren Apps aufgelistet und beschrieben. Dies sind aber bei weitem nicht alle und es werden immer mehr. In ihrem Artikel beschreiben Felizi und Varon (2018) bei einer Suche im englischsprachigen App Store von Google allein etwas über zweihundert Apps nur für Frauengesundheit und Menstruationserfassung gefunden zu haben. Daraus wird schnell ersichtlich, dass es insgesamt viele hundert Apps geben muss, mit denen Körperdaten aller Art erfasst werden können.

Apps gibt es zum Beispiel, um epileptische Anfälle zu erfassen und mögliche Auslöser zu finden. Das Gleiche gibt es auch für Migräne und allergische Anfälle. Es gibt Diabetes Apps und Traumtagebücher, Stimmungserfassungs-Apps und wie oben schon erwähnt eine große Bandbreite an Apps für Frauengesundheit. Es gibt Apps für die Ernährung, um Kalorien zu zählen und Rezepte zu suchen. Diese können mit Bewegungs-Apps zusammenarbeiten und so die verbrauchten zu sich genommenen Kalorien verrechnen. Es ist ein sehr großer Markt, der von vielen verschiedenen Anbietern bestückt wird.

Ebenfalls ein großes Angebot findet sich unter den Sport Apps, dort gibt es für Laufen, Radfahren oder Fitness Training spezialisierte aber auch generelle um einfach erfassen zu können, dass Sport getrieben wurde. Diese Apps sollen Anleiten, Erfassen und Tipps geben. Durch den Vergleich mit sich selbst und anderen soll eine Motivationshilfe gegeben werden.

Auch für geistige Pflege gibt es Apps, entweder mit Entspannungsübungen oder Meditation. Jeder Bereich des Lebens kann erfasst, bewertet und archiviert werden. Deborah Lupton geht in ihrem Buch Lupton (2016) stark darauf ein und referenziert auch auf weitere ihrer Arbeiten. in denen sie sich ausgiebig mit den Apps beschäftigt.

[14] http://www.alpha-fit.de/produkte/smart-sock.html

2.1.4 Probleme der Sebstvermessung

In der in Abschnitt 2.1.2 beschriebenen Literatur wurde sich damit auseinandergesetzt, wie Selbstvermessung funktioniert und was daran Probleme bereitet oder dazu führt, dass es nicht funktioniert. In den verschiedenen Arbeiten wurden Probleme der Selbstvermessung aufgeführt, die im Rahmen dieser Arbeit verwendet und aufgegriffen werden.

In den Arbeiten von Choe u. a. (2014); Li u. a. (2011); Bentley u. a. (2013) wird beschrieben, dass ein Problem der Datenerfassung ist, dass Daten dezentralisiert erfasst und gespeichert werden. Der Nutzer hat weder einen guten Überblick noch eine Möglichkeit, die Daten untereinander zu vergleichen, wenn sie in unterschiedlichen Apps auf unterschiedliche Art angezeigt werden. Darüber hinaus ist es oft ohnehin schwer, die Daten zu analysieren, der Nutzer bräuchte erweitertes Wissen, sowohl über die Daten und ihre Struktur sowie über das Feld Choe u. a. (2014); West u. a. (2016). Daten müssten für den Nutzer leichter in Plattformen zusammengefasst und aus verschiedenen Quellen zusammengeführt werden können MacLeod u. a. (2013); Li u. a. (2010). Daraus folgt auch, was in den Arbeiten von Whooley u. a. (2014); Choe u. a. (2014); West u. a. (2016) angesprochen wird, dass eine geeignete Anzeige der Daten von enormer Wichtigkeit ist. Der Nutzer kann meist nur so viel aus den Daten ziehen, wie das Gerät oder die Applikation ihm erlaubt. Sind Daten schlecht oder missverständlich angezeigt, kann der Nutzer nur schwerlich einen Mehrwert oder eine Erkenntnis daraus ziehen. Häufig sind Daten für den Nutzer auch nicht ganzheitlich erfahrbar, sie werden akkumuliert oder unvollständig angezeigt und sind oft nicht in ihrer Gänze einsehbar Li u. a. (2011). Darüber hinaus müssen die Anzeigen gut durchdacht sein, was wird dem Nutzer wie angezeigt und wann? Eine Arbeit beschäftigt sich damit, dass es keineswegs optimal ist, dem Nutzer jede negative Information sofort zu geben. Im speziellen geht es darum ob einem gestressten Nutzer angezeigt werden sollte, dass er gestresst ist, dies kann zu einem negativen Trend werden Kersten-van Dijk u. a. (2017). Es ergab sich auch, dass selbst bei guten Anzeigen Nutzer dazu neigen können nur zu sehen was sie sehen wollen Kersten-van Dijk u. a. (2017).

Neben den Problemen der Anzeige und dem Erfahren von Daten bereitet das Erfassen allein oft schon viele Probleme. Zum einen sind Nutzer im Allgemeinen schnell vergesslich, vor allem wenn sie keine hohe intrinsische Motivation haben. Das heißt Nutzer müssen immer wieder erinnert werden die Daten zu erfassen Bentley u. a. (2013). Ebenso müssen Daten möglichst zeitnah erfasst werden. Umso später Daten erfasst werden, umso mehr divergieren sie mit der Realität MacLeod u. a. (2013); Lüdemann (2016a). In einer vorangegangenen Arbeit Lüdemann (2016a) wurde anhand eines Selbsttestes und verschiedener Literatur gezeigt, dass es nicht vorteilhaft ist, Daten manuell zu erheben Whooley u. a. (2014); Li u. a. (2010); Lupton (2014); Kersten-van Dijk u. a. (2017); MacLeod u. a. (2013). Dies kann dazu führen, dass Daten nicht oder un-

genau erhoben werden. Die Ungenauigkeit kann daher rühren, dass Daten erst viel später erfasst als erzeugt werden. Bei der Erfassung der Ernährung kann es sein, dass durch Zeitdruck keine Möglichkeit besteht direkt aufzuschreiben was verzehrt wurde, am Abend ist die Erinnerung schon nicht mehr so genau und am Tag danach sehr schwammig. Ebenfalls kann es sein, dass geschummelt wird, Kleinigkeiten werden nicht erfasst, weil der Aufwand des Erhebens zu groß ist oder sich für das Essen geschämt wird, da man eigentlich gesünder essen wollte. Ebenso kann das manuelle Erfassen zu anderen Nebeneffekten führen, bei der Erfassung der getrunkenen Wassermenge ist zum Beispiel aufgefallen dass sich das Trinkverhalten ändert um weniger Schreibaufwand zu haben. Anstatt viele kleine Mengen über den Tag verteilt wurde einige Male eine größere Menge getrunken. Dies sind im Großen und Ganzen alles unerwünschte Nebenwirkungen vom manuellen Erfassen, die teilweise durch geschicktes Design verbessert werden können.

Das Erfassen muss darüber hinaus einfach sein und möglichst schnell gehen, damit der Nutzer es nicht vor sich herschiebt Choe u. a. (2014). Neben den Erinnerungen sollte der Nutzer motiviert werden, das Erfassen nicht zu vernachlässigen, da besonders nach einer gewissen Zeit, nach dem der Novelty Effekt Koch u. a. (2018) aufhört, der Nutzer das Interesse an dem Erheben verlieren kann Choe u. a. (2014). Ein zu großer Ansporn oder Wettbewerb kann allerdings dazu führen, das Nutzer schummeln um bessere Ergebnisse zu erzielen Rooksby u. a. (2014). Zusammenfassend kommen viele Arbeiten zu dem Schluss, dass bei der Datenerfassung so viel wie möglich vereinfacht und automatisiert werden muss, um dem Nutzer die Arbeit abzunehmen, da manuelles Erfassen zu großen Problemen und Ausfällen führen kann Whooley u. a. (2014); Li u. a. (2010); Lupton (2014); Kersten-van Dijk u. a. (2017); MacLeod u. a. (2013). Was also nicht automatisiert werden kann muss optimal und einfach gestaltet werden, einfache übersichtliche und klein gehaltene Eingabemasken, die am besten dem Nutzer Spaß machen und ihn erinnern, sie auszufüllen. Darüber hinaus sollte dem Nutzer ein Opt Out geboten werden, um den Druck auf ihn gering zu halten, besonders mit negativen Reizen sollte sorgsam umgegangen werden und je nach Art der Daten der Nutzer entsprechend geschützt werden.

Einige Arbeiten bemängelten auch, dass viele Apps und Geräte dem Nutzer nicht ausreichend erlauben Kontexte zu erfassen, Notizen oder Anmerkungen zu machen und so ihre eigenen Daten anzureichern. Dadurch nutzen sie oft zusätzliche Applikationen, was die Daten noch weiter aufteilt oder sie verlieren das Interesse Li u. a. (2011); Kamal u. a. (2010); Mentis u. a. (2017). Ein weiterer wichtiger Punkt ist, dass dem Nutzer Feedback gegeben werden muss und dies möglichst zeitnah, sei es um ihn für das Erfassen zu belohnen und so die Motivation zu steigern, oder auch um ihm direkte Informationen zu seinen Daten zu geben. Das Erfassen sollte nicht nur als Prozess gedacht sein, in dem der Nutzer Daten erhebt, sondern gleichsam Feedback dazu erhält Kersten-van Dijk u. a. (2017). Dazu kann auch gehören, was in den Arbeiten von Li u. a. (2011); Kersten-van Dijk u. a. (2017) angesprochen wird, es wäre für den Nutzer

sinnvoll, ihm Lösungsvorschläge und Tipps auf Grundlage seiner Daten anzubieten, sodass die Technik mehr auf den Nutzer reagiert und eine Kommunikation entstehen kann, die nicht nur in eine Richtung geht.

In den Arbeiten von West u. a. (2016); Rooksby u. a. (2014) wird angesprochen, dass bei der Auswertung oft dadurch Probleme entstehen, dass die Datenqualität nicht so gut ist wie es wünschenswert wäre, besonders manuelles Erfassen ist oft ungenau aber auch Sensoren sind noch nicht so stabil und einfach zu benutzen wie es sinnvoll wäre. So muss jeweils damit gerechnet werden, dass Messwerte mehr Richtwerte sind, deren Tendenzen spannend sind. Dabei muss auch mit Mess- oder Bedienfehlern gerechnet werden. Ebenso sind die Daten nicht immer akkurat. Neben der Datenqualität ist es auch oft so, dass ernsthafte Langzeiterfassung kaum unterstützt wird. Besonders manuelles Erfassen ist in den meisten Fällen zu aufwändig. Beim Erfassen mit Sensoren zielt die Auswertung oft nicht darauf, Daten von mehreren Jahren dem Nutzer näher zu bringen und daraus Wissen zu generieren Li u. a. (2011). Dazu kommt, dass viele Studien nicht auf Langzeit angelegt sind, sie laufen zumeist nur ein paar Wochen, manchmal einige Monate, aber von Langzeit kann dabei nicht gesprochen werden. Ein Teil des Problems liegt darin, dass es schwer ist, die Probanden über längere Zeit zu motivieren Harrison u. a. (2014).

Darüber hinaus gibt es auch technische Probleme. Es wird klar, dass besonders im Rahmen der Sicherheit und des Datenschutzes extreme Defizite vorhanden sind. Die Apps und Geräte sind für einen schnellen Start auf den Markt konzipiert, die wenigsten achten dabei auf das notwendige Maß an Sicherheit im Umgang mit den Daten, sei es unverschlüsselte Übertragung von Gerät zu Smartphone Response (2014) oder in der Speicherung beim Hersteller. Immer wieder kommt es zu Angriffen, denn die Daten sind von enormem Wert. Ein relativ junger Vorfall ist der Angriff auf die MyFitnessPal Datenbank im Februar 2018, bei dem von ca. 150 Millionen Nutzerkonten unter anderem die E-Mail Adressen und Passwörter gestohlen wurden Fipps (2018). Ändern die Nutzer das Passwort nicht, so können darüber auch die empfindlichen Dateninhalte eingesehen werden. Aber nicht nur die Hersteller von Fitness Apps und Geräten sind Ziel diverser Attacken, auch Krankenkassen trifft es Spiegel (2014).

Neben dieser Angreifbarkeit der Daten kommt der Umstand dazu, dass die Firmen, die hinter Geräten und Apps stecken, zum Teil völlig offen in ihre AGBs schreiben, dass sie die Daten zu ihren eigenen Zwecken nutzen bzw. sogar verkaufen. Ein Beispiel dazu sind Apps um Frauen zu ermöglichen, ihre Menstruation besser zu erfassen. Dabei trägt die Benutzerin diverse teilweise extrem persönliche Daten in die App ein, die daraus Verläufe, Symptome, Tipps und Statistiken generiert. Die Bandbreite an Applikationen ist enorm, so auch die Ausrichtung der Ziele, die mit der App verfolgt werden können. Dabei entsteht ein Abbild des Hormonspiegels der Benutzerin. Einige Apps geben dann Ratschläge, beispielsweise wird angegeben, man sei nun in einer Phase, in der man sich besonders attraktiv fühle, in Kaufstimmung gerate oder zu

gesteigertem Hungergefühl neige. Andere Apps verbinden das laut Felizi und Varon (2018) direkt mit passender Werbung und nehmen sich auch das Recht heraus, die Daten weiterzugeben und nach dem Löschen des Accounts die Daten zu speichern. Für Werbefirmen sind dies natürlich wertvolle Informationen.

2.1.5 Moralischer Diskurs

Quantified Self ist ein sehr zwiespältig betrachtetes Feld. Zum einen ist es ein wachsender Markt mit immer neuen Möglichkeiten, zum anderen begegnen viele Menschen der Selbstvermessung mit Argwohn und Zurückhaltung. Schließlich erhebt man empfindliche Daten über sich selbst, die im schlimmsten Fall gegen die eigene Person verwendet werden können. Körper und Gesundheitsdaten gehören zu den sensibelsten Daten. Beim Verlust einer Kredit- oder EC-Karte kann man die Karte sperren oder wenn nötig das Konto auflösen und woanders ein neues erstellen. Werden Gesundheitsdaten gestohlen und veröffentlicht geht das nicht, man kann weder den Körper noch das Leben tauschen. Sobald eine Krankenkasse oder ein Arbeitgeber von einer chronischen Krankheit erfährt, können einem daraus Nachteile erwachsen. Banken könnten einen Kredit verweigern, wenn zum Beispiel herausgekommen ist, dass man an einer tödlichen Krankheit leidet. Vielleicht kommt der Betroffene gar nicht dazu, das Geld zurückzuzahlen oder wird durch die Krankheit vielleicht arbeitsunfähig, vielleicht hat er hohe medizinische Kosten zu stemmen und wird daraufhin als nicht kreditwürdig eingestuft. Ebenso können Menschen mit zerstörerischen Absichten gegebenenfalls die Krankheiten wichtiger Persönlichkeiten ausnutzen um Schaden anzurichten. Seien es starke Allergien, wichtige benötigte Medikamente oder einfach für den Ruf abträgliche Krankheiten, sollten sie öffentlich bekannt werden. Dazu kommt, dass besonders in den Anfängen von Quantified Self schnell Sensoren auf den Markt gebracht wurden, deren Fokus nicht auf Sicherheit sondern auf Funktionalität beruht. Es war am Anfang nicht schwer, die Daten von Sensoren abzugreifen, wenn sie diese an die Smartphones übermittelt haben Institute (2018).

Dazu kommt, dass es nicht ganz geklärt scheint wem die erfassten Daten von Applikationen und Wearables gehören. Wie beim Beispiel der Menstruations-App, aus dem vorherigen Abschnitt ist es nicht gegeben, dass die erhobenen Daten gelöscht werden nachdem man den Account löschte. Ein Beispiel für den Umgang der Firmen mit den persönlichen Daten ist auch Facebook. Die Firma verdient Unsummen an den Daten der Nutzer, ohne davon Gelder an den Nutzer selber weiterzugeben. Alles in allem ist es extrem schwer transparente Datenpolitik zu finden und als Nutzer souverän entscheiden zu können, wer welche Daten speichert, weitergibt und verarbeitet. Ebenso ist es schwierig, die Daten nachhaltig löschen zu lassen.

Häufig wird dies selbst in den Köpfen der Benutzer nicht unbedingt als Problem angesehen. Sie hätten nichts zu verbergen, es interessiere doch eh niemanden wie viele

Schritte sie am Tag machten und wann. Dabei gehen sie davon aus, dass das allein die Daten schon schützt Lupton (2014). Dabei liegen sie falsch, wie allein schon das Menstruations-App-Beispiel gravierend zeigt. Aber auch Angriffe auf Gesundheitsdatenbanken wie in den USA werden immer häufiger Spiegel (2014). Denn Daten sind gefragt und sind für unterschiedliche Abnehmer interessant. Dazu gehören neben Firmen, die ihre Produkte bewerben wollen auch Wissenschaftler, Krankenkassen, Versicherungen oder auch der Staat. Es wird etwas später noch einmal darauf eingegangen.

Darüber hinaus sind Daten ein sehr abstraktes Konstrukt. Man sieht sie nicht direkt, man kann sie nicht anfassen und wenn sie gestohlen wurden merkt man es zumeist nicht mal. Bis vielleicht gehäuft Werbung kommt, Werbeanrufe, eine Versicherung Leistungen versagt, die Einreise in ein Land verwehrt bleibt oder ähnliches. Es gibt immer mehr Informationen, die über uns im Netz stehen und besonders Gesundheitsdaten sind extrem heikel. Eine E-Mail-Adresse oder ein Bankkonto kann man sperren und wechseln. Eine Krankenakte, einen Körper und ein Leben nicht. Vielleicht möchte man nicht, dass der zukünftige Arbeitgeber von der eigenen Krankengeschichte weiß, besonders in Amerika, wo die Arbeitnehmer vom Arbeitgeber versichert werden, ist dies eine sehr unangenehme Vorstellung. Aber auch in Deutschland könnte ein Verweis auf eine chronische Erkrankung die ggf. zu häufigen Arztbesuchen, Behandlungen oder Operationen und somit zu gehäuften Fehltagen führt, sehr unangenehm sein und den gewünschten Arbeitsplatz verwehren. Genauso wie der Verweis auf ein frühes Stadium einer Schwangerschaft der eigenen Person oder des Partners. Schließlich geht auch der Vater immer häufiger in Elternzeit.

Dieses Wissen und auch der Umgang mit der Ressource Daten müssen die meisten Nutzer erst lernen. Es muss ein Umgang ermöglicht werden, der ein Studium der Informatik nicht erforderlich macht, um technische Geräte, die mit dem Internet verbunden sind, zu benutzen. Es muss ein Verständnis dafür entwickelt werden, dass Daten wertvoll sind, auch wenn sie noch so banal wirken, dass die Daten einer Person dieser Person gehören und nicht irgendwelchen Firmen. Bis dahin sollte der Benutzer durch Gesetze und im besten Fall von den Firmen, die solche Werkzeuge zur Verfügung stellen, selbst geschützt werden.

In der Arbeit von Wiedemann (2016) wird anhand der Arbeiten von Deborah Lupton und Melanie Swan ein recht bedrohliches Bild von Quantified Self gemalt. Selbstvermessung als Mittel zur Selbstkontrolle, Gesundheit als zu erlangendes Gut, das man aufrechtzuerhalten habe und sich auf dem Gesundheitsmarkt als brauchbares Gut zu bewahren hat. In einer Arbeit von Rooksby u. a. (2014) zeigt sich aber, dass Selbstvermessung nichts Dauerhaftes sein muss. Häufig nutzen die Befragten in jener Arbeit eine Art der Selbsterfassung nur solange, bis sie die Antwort auf ihre gestellte Frage gefunden hatten oder aber die Lust verloren. Entweder wurden dann andere Sensoren mit einer anderen Fragestellung genutzt oder aber das Erfassen ganz beendet. Selbstvermessung ist ein Werkzeug, das die Menschen nutzen wie es ihnen dienlich ist. Dabei

werden verschiedenste Tools verwendet, um den eigenen Bedürfnissen gerecht zu werden, verschiedenste Apps oder Sensoren werden zusammen genutzt und wenn diese nicht bieten was der Nutzer benötigt, wird auf trivialere Werkzeuge zurückgegriffen, Notiz-Apps, Tabellen oder auch Zettel und Stift Kamal u. a. (2010).

Wichtig ist die Unterscheidung, ob Self Tracking ein Werkzeug ist um dem Leben dienlich zu sein oder ob es im Vordergrund steht und das Leben dahingehend angepasst wird. Quantified Self sollte niemals ein Dogma sein nach dem man lebt, es sollte als Werkzeug für Fragestellung und Neugierde dienen und im Leben mitlaufen und es dabei möglichst wenig beeinflussen, außer in den selbstgewählten Momenten des Nachfragens. Aus dem kritischen Blick Wiedemanns ergibt sich, dass Technik sanft unterstützen und verfügbar sein, sich aber nicht aufdrängen und nicht im Mittelpunkt stehen sollte. In der Arbeit von Kersten-van Dijk u. a. (2017) wird herausgearbeitet, dass es für den Menschen nicht immer von Vorteil ist, in gestressten Situationen auch vorgehalten zu bekommen, dass er gestresst ist. Oder dass seine Körperdaten aufgrund der stressigen Phase schlecht sind. Jedoch ist es auch so, dass die Technik helfen kann. In einer immer hektischeren Welt erlaubt die Technik sich mit sich selbst zu beschäftigen, man ist nicht als unproduktiv verpönt wenn man sich hinsetzt und überlegt wie man das aktuelle Empfinden, den Schlaf, die Laune oder den Schmerz auf einer Skala einordnet. Man kann dem Effizienzdruck durch die eigenen Mittel einen Augenblick entfliehen und sich durch die technisch validierte Notwendigkeit erlauben, auch mal nichts zu tun um sich zu entspannen.

Dazu kommt, dass Ärzte oft schlicht überfordert sind und zu wenig Zeit haben oder sie sich nehmen. Besonders in Großstädten kennen sie die Patienten nicht mehr persönlich. Ein Gefühl dafür zu haben was für einen persönlich 'normal' ist, kann dabei helfen den Arzt zu unterstützen und sich selbst vor Fehlentscheidungen des Arztes abzusichern. Vielleicht schaut man selbst auf die Blutwerte und geht ggf. zu einem anderen Hausarzt, um im Zweifel eine Zweitmeinung zu haben. Dies steht aber dem Problem gegenüber, den Arzt übervorteilen zu wollen, da man es aufgrund der Daten 'besser weiß' oder Probleme aufgrund der Daten sieht, wo keine sind. Zum Beispiel kann es so sein, dass das eigene Befinden gut ist, das Blutdruckmessgerät aber einen zu niedrigen Blutdruck anzeigt. Als Reaktion darauf geht man zum Arzt, obwohl ohne die Technik alles in Ordnung wäre. Auch Messfehler des Pulssensors können zu Ängsten führen, zum Beispiel davor eine Herzerkrankung zu haben. Wichtig ist in jedem Fall ein souveräner Umgang mit der Technik, den Daten und dem Körper bzw. Körpergefühl. Sensoren sind nicht perfekt, messen nicht hoch akkurat und können falsch benutzt werden, das erfordert, dass dies dem Nutzer auch bewusst ist. Hinzu kommt die Befürchtung, die Verwendung von Technik könnte zu einem schlechteren Körperempfinden führen, dass die Menschen mehr auf die Technik und ihre vermeintlich rationalen Zahlen hören als auf das eigene Gefühl. Dies führt über kurz oder lang zu einer regelrechten Abhängigkeit von der Technik. Tatsächlich kann das, was der Nutzer sieht, sein Gefühl beeinflussen, so schreiben es Rogerson u. a. (2016) in ihrer

Arbeit. Fühlt sich der Nutzer gut, sieht jedoch, dass seine Daten etwas anderes sagen, passt sich das Gefühl an. Ebenso kann sich das Gefühl verbessern, allein dadurch dass die Daten exzellent sind. Ob dies ein kurzfristiger Trend ist, der nach einiger Zeit abebbt, wenn die Daten nicht mehr als so stark präsent wahrgenommen werden, ist in der Arbeit nicht geklärt.

Jedoch ist Quantified Self und die Vermessung des Körpers auch dazu geeignet, ein eigenes Gefühl zu entwickeln oder zu verbessern. Es gibt Erkrankungen, deren gefährliche Zustände nicht erfühlt werden können, wie zum Beispiel Diabetes. Hier kann die Technik als Vermittler zwischen Gefühl und Realität treten. Andererseits haben Menschen ein sehr unterschiedlich stark ausgeprägtes Körpergefühl. Zum einen gibt es Menschen, mit einem sehr guten Körpergefühl, die kaum bis keine technische Unterstützung bedürfen um zu wissen, wie es ihrem Körper geht. Dagegen gibt es aber auch Menschen, die schon Probleme damit haben, ausreichend Flüssigkeit über den Tag aufzunehmen. Besonders für diese kann die Technik eine Chance sein sich ihrem Körpergefühl wieder anzunähern.

Quantified Self hat oft auch eine soziale Komponente, die meisten Sensoren und Apps bieten an, die Daten in einem gewissen Kreis zu teilen. Dies soll motivieren und anspornen. Man kann eigene Freunde zu Wettbewerben herausfordern und stolz auf sozialen Plattformen neueste Ergebnisse posten. Dies kann durchaus positiv sein, den Einzelnen motivieren und ihm das Gefühl geben, Teil einer großen Gemeinschaft zu sein. Es kann aber auch anders herum zu großem Druck und Stress führen. Wenn die Ziele nicht erreicht werden, wenn an Wettbewerben nicht teilgenommen wird oder dergleichen. Dann muss der Nutzer mit sozialen Sanktionen rechnen, mit Ausschluss oder Missgunst.

In der Arbeit von Wiedemann (2016) wird eine sehr dystopische Zukunft gemalt, in der Gesundheit zur Selbstverständlichkeit wird, dass man sich nicht mehr glücklich schätzen kann gesund zu sein, sondern vor der Gesellschaft dazu verpflichtet ist, einen Zustand der Fitness und Gesundheit aufrechtzuerhalten. Ein Gesundheitszwang, der nicht mehr viel mit persönlicher Freiheit zu tun hat. In der Arbeit von Whitson (2013) sowie im Buch von Deborah Lupton 'The Quantified Self' Lupton (2016) wird darauf hingewiesen, dass in immer mehr Bereichen des Lebens Tracking verwendet wird. Zum Beispiel auf der Arbeit, um als Gruppenmotivation zu dienen. Dies kann aber schnell dazu führen, dass man sozial ausgeschlossen wird oder gar schlechtere Arbeitsleistungen erbringt. Ein Beispiel dafür ist, dass einige Firmen Wettbewerbe anbieten, in denen Gruppen aus Mitarbeitern über die Zeit Schritte sammeln, Joggen oder Radfahren tracken. Diese Daten werden dann mit dem Arbeitgeber geteilt und entsprechend belohnt. Will man dem Arbeitgeber keine Informationen über die sportlichen Aktivitäten geben, die man durchführt, sieht man sich im Nachteil.

Wer nicht trackt, an Wettbewerben teilnimmt und sich dem allgemeinen Gesundheitstrend unterwirft steht schnell am Rande der sozialen Gesellschaft. Darüber hinaus

könnte es aber auch ganz weltlich zum Zwang werden, wenn der Staat, wie es in China geschieht, Druck ausübt und großflächig Daten auswertet, um sich einen nach seinem Ermessen guten Bürger zu schaffen. Der Staat hat dort den Bürger zum gläsernen Menschen gemacht der danach bewertet wird, wie nah er dem staatlichen Bild eines perfekten Bürgers kommt. Es werden unter anderem Kaufverhalten, politische Einstellung und deren Äußerungen im Netz sowie der Bankverkehr bewertet. Umso mehr Punkte ein Bürger hat, umso wahrscheinlicher ist es, dass er an einen Kredit oder ein Visum kommt. Die Datenbasis, auf der die Bewertung getroffen wird, soll weiter ausgebaut werden Plass-Flessenkämpfer (2015); Hanfeld (2015). Da sind Körper- und Fitnessdaten nur ein weiterer Schritt.

Krankenkassen und Firmen zeigen ebenfalls vor allem monetäres Interesse an Gesundheitsdaten. Sie lassen sich gut verkaufen, da Forschungsinstitute, Werbefirmen und ähnliche Einrichtungen großes Interesse daran haben. Zuden können sie aber auch zur Individualisierung von Krankenkassentarifen beitragen oder zur Risikoanalyse über die jeweiligen Personen. Getarnt als ein Motivtionsgrund für die Versicherten sich mehr zu bewegen kann eine Krankenkasse so Gesundheitsdaten über ihre Nutzer sammeln. Für andere Versicherungen können diese Daten aber genauso interessant sein. Lohnt es sich für eine Versicherung diesen Kunden zu versichern? Kann man demjenigen einen Job oder einen Kredit geben? Aber auch die Forschung ist interessiert, um Medizin und Medikamente zu verbessern. So können allgemeine Produkte zum Beispiel im sportlichen Rahmen damit optimiert und gezielter an den Markt gebracht werden. Es ist schwer abzuwägen, welche Folgen es haben kann, wenn Gesundheitsdaten für alle frei zugänglich wären.

In ihrer Arbeit Lupton (2018) geht Lupton darauf ein, dass einer der nächsten Schritte das Erfassen der Daten direkt im Körper sein könnte, dies würde das Erheben stark automatisieren und dadurch für den Nutzer einfacher gestalten. Gleichzeitig macht es das aber auch invasiver und für den Nutzer viel schwerer auszusteigen, sowie Daten nur kurzfristig zu erfassen und seinen eigenen Zielen damit zu folgen. Sie beschreibt den Mensch als Cyborg, ein Schritt, der deutlich näher ist als man zuerst annehmen mag. Erste Schritte in diese Richtung sind bereits getan, angefangen mit steuerbaren Herzschrittmachern, Insulinpumpen, Prothesen und Blutzuckermessgeräten, die im Arm stecken, sowie intelligente Tattoos, die die Farben je nach Blutwerten verändern Focus. Dadurch scheint diese Vision bereits in nahe Zukunft gerückt zu sein.

Auf der anderen Seite liegt ein vermeintlich großer Nutzen im Erfassen der Daten, wenn sie richtig behandelt und eingesetzt werden. Das Individuum an sich könnte durch Antworten auf seine Fragen, verbesserte Unterstützung einer Zielerreichung und verbesserte medizinische Versorgung profitieren. Ebenso könnte daraus ein besseres Selbstverständnis erwachsen, sowohl für Vorgänge und Veränderungen im Körper als auch für Zusammenhänge. Nicht ohne Grund gibt es eine Vielzahl von Apps, die dafür gedacht sind, die Ursachen für Migräne, Kopfschmerzen, allergische Reaktionen oder

epileptische Anfälle zu erkennen und dadurch vermeiden zu können. Daraus ergibt sich ein verbessertes Wissen, welches dem Individuum aber auch der Gemeinschaft nutzen kann. Denn einmal erkannte Auslöser könnten der breiten Masse bekannt werden und in Zukunft helfen, viel schneller die Auslöser solcher Leiden zu finden. Das Verständnis des eigenen Körpers kann eng mit dem eigenen Wohlbefinden oder der Gesundheit verknüpft sein. Aber auch die schiere Neugierde treibt die Menschen an und kann so zu mehr Wissen und besserer Erkenntnis führen oder einfach zu einem gesteigerten Wohlbefinden des Einzelnen. Besonders die Langzeiterfassung kann dazu führen, dass Krankheiten entdeckt werden können bevor sie entstehen. Dazu gehört auch, dass frühe Anzeichen im Nachhinein aus dem Datenstrom identifiziert werden, um sie so in Zukunft als frühe Anzeichen zu erkennen. Gerade wenn Langzeitdaten auch mit den Daten vieler anderer Menschen verglichen werden, können so Risikogruppen gebildet, verbesserte Behandlungen entdeckt und Krankheitsverläufe untersucht werden. Neben verbesserten Studien zu Sport, Ernährung und Medizin kann der Einzelne davon profitieren indem ihm viel stärker individualisierte Informationen gegeben werden können.

Neben dem Individuum könnte auch die Gesellschaft an sich ihr Wissen über Krankheiten, Verläufe sowie gute und schlechte Behandlungsmethoden erweitern. Ebenfalls möglich wäre das Erkennen noch unbekannter Muster, Krankheiten und Methoden. Vielleicht könnten Menschen mehr über sich und den Mensch an sich erfahren, wenn sie mit den Daten richtig umgehen. Somit ist das Potential des Self Tracking noch keineswegs völlig aufgedeckt. Klar ist, dass sorgsam mit den Daten und den Methoden umzugehen ist, um Schäden zu vermeiden.

2.2 Der intelligente Spiegel

In diesem Abschnitt soll näher darauf eingegangen werden, warum sich ein Spiegel als Ort der Begegnung mit den eigenen Daten eignet, was es bereits für intelligente Spiegel gibt und wie sich in wissenschaftlichen Arbeiten damit auseinandergesetzt wurde. Daraus werden Informationen und Anforderungen für diese Arbeit entwickelt.

Der Spiegel ist seit jeher ein Ort der Selbsterfahrung und Erkenntnis, ein Ort, an dem man seiner äußeren Erscheinung, seinem Spiegelbild begegnet. Ein Ort, um zu validieren, ob das Erscheinungsbild den eigenen oder fremden Normen entspricht und sich gegebenenfalls auch diesen Normen anzunähern, sich zu optimieren. Darüber hinaus taucht der Spiegel in vielen Geschichten als mehr oder weniger magisches Objekt auf, als allwissender Ratgeber oder als Portal zu anderen Welten, es gibt viele Geschichten in denen Spiegel eine besondere Rolle spielen. Somit scheint der Spiegel für die Menschen schon immer ein ganz besonderes Objekt zu sein, es erscheint somit naheliegend, dem Spiegel im Rahmen des Internet of Things und den sich immer weiter

ausbreitenden Trends der Smart Object und dem Smart Home die Fähigkeit zu geben, einen noch tieferen Einblick in den Menschen zu ermöglichen. Ein Ort, in dem nicht nur dem äußeren Bild, sondern auch dem virtuellen Spiegelbild aus Daten begegnet werden kann. Ein Ort, an dem neben der äußeren Widerspiegelung seines Selbst ein Blick auf Vitalwerte geworfen werden kann. So wird der Spiegel zu einem Ort der erweiterten Selbsterkenntnis.

Im Smart Home werden immer mehr Alltagsgegenstände technisiert und 'smart' gemacht. Kühlschränke, Kaffeemaschinen, Vorhänge, Licht, selbst Toiletten und Türschlösser werden im Wohnen der nächsten Generation intelligent, vernetzt und mit zusätzlichen Funktionen versehen. Somit ist es naheliegend, einen Spiegel zu vernetzen und mit Technologie anzureichern, um die ursprüngliche Funktion eines Spiegels zu verbessern und zu erweitern. Dafür finden sich viele Ansätze auf dem freien Markt wie in wissenschaftlichen Arbeiten.

2.2.1 Wissenschaftliches Umfeld

Die meisten Ansätze reichern den Spiegel mit Funktionalitäten eines Smartphones an, Kalenderinformationen, Nachrichten, Wetter, Uhrzeit und To-do-Listen. Darunter die Arbeiten von Yu u. a. (2012); Hossain u. a. (2007); Njaka u. a. (2018); Gold u. a. (2016); Yusri u. a. (2017); Athira u. a. (2016); Cvetkoska u. a. (2017); Johri u. a. (2018). Einige dieser Arbeiten bringen zusätzlich Funktionalitäten für die Bedienung von anderen Smart-Home-Elementen dazu, wie die Steuerung von Licht, Musik oder den Vorhängen Athira u. a. (2016); Yusri u. a. (2017). Darüber hinaus gibt es eine Arbeit, die dem Nutzer ein System bieten möchte, aufgemuntert zu werden Yu u. a. (2012), dazu werden unter anderem Pulsdaten verwendet. Mehr in Richtung Gesundheit und Körperdaten geht die Arbeit von Besserer u. a. (2016), in der ein Spiegel vorgestellt wird, der die Morgenroutine des Nutzers verbessern soll und dieser somit glücklicher und aktiver in den Tag startet. Dies soll geschehen, indem er sich vor dem Spiegel ein wenig sportlich betätigt. Dabei werden Gewicht, Schritte und verbrannte Kalorien angezeigt. In der Arbeit von Fujinami (2010) werden die täglich gegangenen Schritte angezeigt und bei Cvetkoska u. a. (2017) Daten aus diversen Sensoren gesammelt, die als Frühwarnsystem die klinische Überwachung außerhalb der Klinik ermöglichen sollen.

Warum ein Spiegel
In der Arbeit von Fujinami (2010) wird angeführt, dass die Verbindung zwischen einem Spiegelbild und körperbezogenen Daten besonders praktisch ist, da Nutzer dazu neigen, stärkere Verknüpfungen zwischen Daten und Objekten zu ziehen, die eng beieinander angezeigt werden. So können auch abstraktere Werte enger an den Körper geknüpft werden, da sie daneben angezeigt werden. Somit bietet sich ein Spiegel für die Anzeige solcher Daten regelrecht an.

Spiegelaufbau
Die Arbeiten nutzen unterschiedliche Methoden um Spiegel zu bauen, die meisten verwenden halb durchlässige Spiegel, hinter denen möglichst helle Monitore angebracht werden Fujinami (2010); Njaka u. a. (2018); Athira u. a. (2016); Johri u. a. (2018); Gold u. a. (2016). Eine der gesichteten Arbeiten verwendet Spiegelfolie vor einem Monitor Besserer u. a. (2016). Ein anderer Ansatz ist, einen Monitor zu nehmen, der einen Spiegel mimt, indem ein Kamerabild darauf projiziert wird Hossain u. a. (2007); Yusri u. a. (2017).

Bedienung
Die Arten der Bedienung sind dabei ebenfalls sehr unterschiedlich, es gibt Arbeiten in denen der Spiegel direkt über Touch bedient wird Besserer u. a. (2016); Hossain u. a. (2007); Fujinami (2010) oder per Touch auf einem nebengelegten Touchdisplay Gold u. a. (2016). In der Arbeit von Cvetkoska u. a. (2017) wird unter anderem Gestensteuerung verwendet, wohingegen der Großteil der Arbeiten Sprachsteuerung benutzt Besserer u. a. (2016); Johri u. a. (2018); Cvetkoska u. a. (2017); Athira u. a. (2016); Yusri u. a. (2017); Njaka u. a. (2018); Yu u. a. (2012). Darüber hinaus verwendet auch eine Arbeit eine Bedienmethode über ein weiteres Gerät, das Zugriff auf eine Webapplikation hat Gold u. a. (2016).

Nutzererkennung und Authentifizierung
Da die meisten Spiegel personalisierte Dienste anbieten oder mit empfindlichen Daten arbeiten, ist zumeist eine Nutzerauthentifizierung nötig. Auch hier werden unterschiedliche Ansätze verwendet. Zum einen arbeiten manche Arbeiten mit der Gesichtserkennung Cvetkoska u. a. (2017); Njaka u. a. (2018); Hossain u. a. (2007). Andere Arbeiten nutzen Spracherkennung Athira u. a. (2016); Njaka u. a. (2018) oder auch weitere Gegenstände wie das Smartphone Besserer u. a. (2016) oder andere persönliche Objekte wie die Zahnbürste Fujinami u. a. (2005). Manche der Arbeiten verwenden auch gleich mehrere Authentifizierungen um die Sicherheit zu steigern oder bieten darüber hinaus an, sich mit Nutzername und Passwort anzumelden, falls die anderen Möglichkeiten aus irgendwelchen Gründen fehlschlagen. Einige Projekte sehen einen Ruhezustand der Spiegel vor, sodass das System nicht die ganze Zeit aktiv ist, sondern sich erst aktiviert oder den Zustand wechselt, wenn ein Nutzer in die Nähe kommt oder sich direkt davor stellt. Dabei werden ebenfalls unterschiedliche Methoden verwendet, den Nutzer auszumachen. Die Arbeit von Besserer u. a. (2016) sieht vor, dass sich der Nutzer auf ein Balance Board stellt, dies erkennt den Nutzer und schaltet den Spiegel ein. Auch Bewegungssensoren die erkennen, wenn der Nutzer näher kommt Yu u. a. (2012) und pyroelektrische Sensoren(PIR)[15] die erkennen ob jemand in der direkten Nähe des Spiegels ist Athira u. a. (2016) werden verwendet. Eine Arbeit verwendet unter anderem einen direkten Knopfdruck am Spiegel Cvetkoska u. a. (2017). Bei der

[15] Ein Sensor der mithilfe von piezoelektrischen Halbleiterkristallen Temperaturunterschiede messen kann.

Sprachsteuerung gibt es eine Funktionalität, die das System durch definierte Wake Words aufweckt, diese wird ebenfalls von einigen Arbeiten verwendet Cvetkoska u. a. (2017); Njaka u. a. (2018).

Bei den vorgestellten Möglichkeiten ist immer zu beachten, dass der Spiegel und seine Funktionen sowie Technologien an die Situation und den Verwendungszweck angepasst werden müssen. In einem Badezimmer ist es fraglich, ob eine Gesichtserkennung sinnvoll ist, da dafür Kameras notwendig sind, die aus Gründen der Privatsphäre dort eher gemieden werden. Dazu kommt, dass Gesichtserkennung an schlecht beleuchteten Orten nicht zuverlässig funktioniert, ebenso wenn Oberflächen beschlagen wie zum Beispiel im Bad. In lauten Umgebungen eignet sich sowohl Spracherkennung wie -steuerung eher weniger. Bei der Bedienung ist es sehr ähnlich, Touch eignet sich wenig für Orte, an denen der Nutzer oft nasse oder schmutzige Finger hat, wie im Bad oder in der Küche. Ebenso ist von Touch eher abzuraten, wenn mit einem Ganzkörperspiegel gearbeitet wird, da der Nutzer dort meistens zu weit von entfernt steht, er müsste für jede Bedienung viel zu dicht an den Spiegel treten.

Oberflächen müssen ansprechend und übersichtlich gestaltet werden Fujinami (2010). In der Arbeit von Fujinami (2010) ist herausgearbeitet worden, dass das Erfassen und Betrachten der Daten ansprechend sein muss, um den Nutzer zu fesseln. Es sollten interessante datenbezogene Visualisierungen und ggf. Interaktionsmöglichkeiten, die dem Nutzer vielleicht auch Freude bereiten, um ihn immer wieder in die Betrachtung und Interaktion mit dem Spiegel zu locken, genutzt werden. Oberflächen und Daten müssen somit ansprechend und übersichtlich gestaltet werden und den Nutzer zum Entdecken und Mitdenken animieren. In der Arbeit von Fujinami (2010) wurden Schritte, die die Testpersonen gegangen sind, in abstrakten Formen auf einem Spiegel dargestellt, die sich je nach Schrittzahl, erreichten Zielen und Schritten am Vortag verändern. So mussten sich die Probanden damit auseinandersetzen, was die Figuren aussagen und konnten mit steigendem Verständnis die Figuren sogar beeinflussen, indem sie mehr oder weniger gegangen sind.

2.2.2 Intelligente Spiegel auf dem freien Markt

Neben den wissenschaftlichen Arbeiten, die sich mit intelligenten Spiegeln befassen, gibt es auf dem Markt bereits eine recht große Bandbreite ganz unterschiedlicher Produkte, auf die hier kurz eingegangen werden soll um zu zeigen, wie stark die Idee aufgenommen wird sowie um unterschiedlichste Herangehensweisen aufzuzeigen und daraus nützliche Informationen für diese Arbeit zu ziehen. Dabei wird hier nur ein kurzer Überblick gegeben, da die meisten in der vorangegangenen Arbeit Lüdemann (2017b) ausführlicher beschrieben wurden.

Es lassen sich für das eigene Heim intelligente Spiegel erstehen, die ähnliche Funktionalitäten haben wie ein Smartphone, sie bieten dabei eine weitere Anzeigefläche für Nachrichten, Wetter, Kalender oder Reiseinformationen. Ein Hersteller bzw. Vertreiber solcher Spiegel, die bereits auch auf Amazon [16] zu finden sind, ist MySmartMirror. Funktional noch breiter aufgestellt ist der MagicMirror, der für öffentliche Orte vor allem für Kaufhäuser, Boutiquen und Museen, aber auch für große Veranstaltungen und Feiern gedacht ist. Er soll Informationen über Produkte, Werbung oder auch interaktive Medien anzeigen können. Über ein Touchdisplay ist er bedienbar und soll auch als Fotobox oder als Gerät um Frisuren, Accessoires und Outfits auszuprobieren, aufgestellt werden können. Dabei soll er Emotionen und Aufmerksamkeitspunkte der Nutzer erfassen, was besonders zum Auswerten für Werbung und Produktplatzierung interessant ist. Ein privates Projekt ist der AppleMirror, ein Spiegel, der im Prinzip dieselben Funktionalitäten bietet wie ein IPad, da er mit iOS 10 läuft und somit die Bandbreite der Apps eines IPads bieten kann. Neben Kalender, Wetter und Nachrichten kann man also auch Videos konsumieren, mit Freunden chatten oder im Internet surfen. Ein eher auf Fitness basiertes Projekt ist Naked, der Spiegel arbeitet mit einer Waage und wird durch eine Smartphone-App erweitert. Der Spiegel erstellt ein 3D-Scan des Nutzers und ermöglicht so das Körperbild besser zu erfassen und Trainingsfortschritte direkt am 3D-Bild des eigenen Körpers zu sehen. Dabei werden Informationen aus dem Scan und der Waage kombiniert.

2.2.3 Anforderungen an einen intelligenten Spiegel

In der Betrachtung von intelligenten Spiegeln in wissenschaftlichen Arbeiten und auf dem freien Markt hat sich ergeben, dass ein Spiegel sehr gut dazu geeignet ist, körperbezogene Daten anzuzeigen. Dabei war zu sehen, dass der Gedanke eines intelligenten Spiegels sowohl im öffentlichen wie auch im Wohnraum immer wieder und verstärkt aufgegriffen wird. Dabei gibt es unterschiedlichste Herangehensweisen und Methoden, jedoch sehen die meisten Arbeiten in einem Spiegel die Möglichkeit, Medien anzuzeigen, die sonst einen Blick auf das Smartphone benötigen. Einige Arbeiten widmen sich darüber hinaus spezielleren Zielen, jedoch scheinen sie selten einen tieferen Blick in den Menschen selbst ermöglichen zu wollen. Dadurch unterscheidet sich die hier beschriebene Arbeit maßgeblich von den vorgestellten wissenschaftlichen Arbeiten. Es sollen keine Home Controll Möglichkeiten oder Smartphone Funktionalitäten auf den Spiegel übertragen werden, sondern der Spiegel durch Körperdaten des Benutzers in seiner ursprünglichen Funktion erweitert und dem Nutzer eine erweiterte Selbsterkenntnis geboten werden. Darüber hinaus wurden aus den vorgestellten Arbeiten Informationen für Methoden und Vorgehensweisen sowie verschiedene Techniken für Hardware und Software eines Spiegelsystems gezogen.

[16] https://www.amazon.de/

Aus den wissenschaftlichen Arbeiten und den darin beschriebenen Herangehensweisen, Problemen und Erfolgen wurden einige Punkte herausgearbeitet, die als Anforderungen an einen intelligenten Spiegel im Rahmen dieser Arbeit sinnvoll erscheinen.

Im Vordergrund stehen dabei die Aspekte der Ambient Intelligence Qualität, Bequemlichkeit, Effizienz, Security(Schutz) und Safety(Gefahrlosigkeit) Raisinghani u. a. (2006).

Der Nutzer muss sich im System anmelden und authentifizieren können, um ein qualitatives (individualisiertes) und geschütztes System nutzen zu können. Da es sich bei den hier geplanten Nutzerdaten um sensible Körperdaten handelt ist es sinnvoll, diesen Prozess möglichst sicher zu gestalten, also durch eine zweite oder gar dritte Anmeldemöglichkeit. Neben der Gesichtserkennung, die für den Nutzer angenehm und einfach funktioniert, kann durch Spracherkennung eine weitere Instanz eingebaut werden, auch dies kann für den Nutzer schnell und einfach durch Passphrasen ermöglicht werden.

Für den Fall, dass die Technik versagt oder die Raumumstände (Dunkelheit, Lärm oder ähnliches) es nicht zulassen, wäre ein Anmelden über Nutzername und Passwort sinnvoll. Jedoch muss dabei darauf geachtet werden, ein sicheres Login zu ermöglichen, ggf. auch über eine Zwei-Wege-Authentifizierung. Ebenso muss sichergestellt werden, dass der Nutzer wieder ausgeloggt wird, wenn er sich entfernt.

Bei der Verwendung einer Kamera für Gesichtserkennung und weitere Datenverarbeitung ist es nötig, die Verzögerung zu minimieren und den Winkel der Kamera zu optimieren, damit der Nutzer ein möglichst verzögerungsfreies und perspektivisch richtiges Spiegelbild erhält. Wenn ein Monitor verwendet wird, um einen Spiegel zu mimen, ist es im ersten Schritt gut, den Kamera-Livestream nicht abzuspeichern, um den Datenschutz zu erhöhen. Es sollten nur Daten abgespeichert werden, die zur Auswertung dienen.

Die Übersicht und Anzeige der Daten muss möglichst einfach und übersichtlich gestaltet werden, dabei aber auch spannend und nicht zu langweilig sein. Inwieweit dies mit dem Wunsch, möglichst detaillierte und unkontextualisierte Daten anzuzeigen, einhergehen kann ist noch nicht abzuschätzen. Dazu kommt, dass die Anzeige schnell und zuverlässig arbeiten soll. Der Nutzer sollte nicht lange auf Daten und Visualisierungen warten müssen.

Für die Bedienung bietet sich Sprachkontrolle oder Bedienung über ein weiteres Gerät an, da ein Ganzkörperspiegel sich nicht zwingend für Touch eignet, denn der Nutzer steht in der Regel zu weit entfernt.

2.3 Knowledge Discovery in Databases (KDD)

In diesem Abschnitt soll erklärt werden, worum es sich beim KDD-Prozess handelt und gleichsam mit dem Model von Li u. a. (2010) verglichen werden. Dabei wird der KDD-Prozess nicht in seiner umfassenden Allgemeinheit behandelt, sondern in Betrachtung der vorliegenden Thematik.

Der KDD-Prozess wurde in der Arbeit von Fayyad u. a. (1996) beschrieben und definiert. Das Ziel des KDD-Prozesses ist es, aus Datenbeständen bisher unbekannte fachliche Zusammenhänge zu erkennen. Der Prozess und seine einzelnen Schritte sind in der untenstehenden Abbildung 2.1 dargestellt, entnommen aus der Arbeit von Fayyad u. a. (1996).

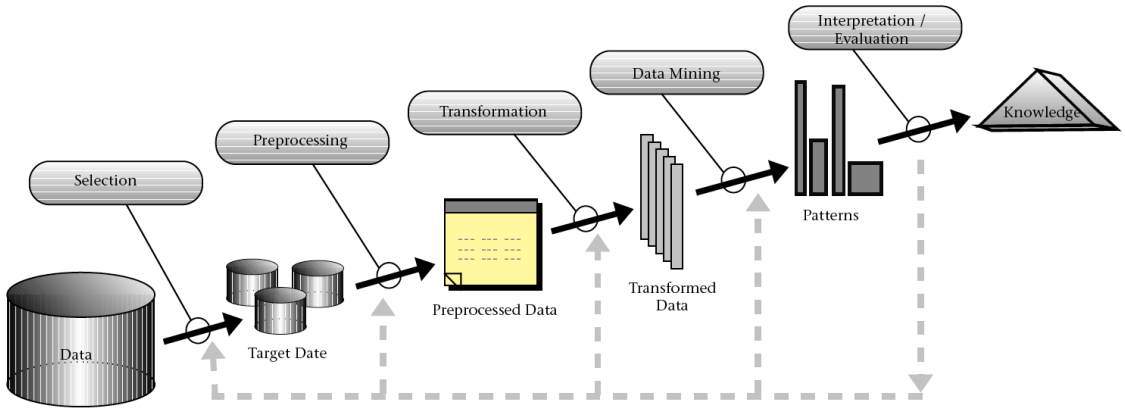

Abbildung 2.1: Die Phasen des KDD Prozesses nach Fayyad u. a. (1996). Entnommen aus selbigem.

Die neun Schritte des Prozesses nach Fayyad u. a. (1996) sind im Folgenden beschrieben. Der Erste ist, das Erlangen des Hintergundwissens und Domänenverständnisses sowie eine Zielsetzung für den Prozess zu finden, die zweite Phase ist die Datenselektion und -integration, dort werden die Daten ausgewählt die für eine Fragestellung relevant erscheinen und ggf. aus verschiedenen Quellen integriert. Die dritte Phase ist die Vorverarbeitung und Bereinigung der Daten, da die Daten selten die Qualität haben, die zur Analyse benötigt wird. In der vierten Phase, der Datentransformation, wird ein von der Analyse benötigtes Format erzeugt. Die fünfte Phase ist die Zuweisung des Ziels des KDD-Prozesses zu einer Data Mining Methode. Die sechste Phase

ist die Wahl der Data Mining Verfahren. Die siebte Phase ist die Analyse, also das Data Mining. In der achten Phase werden die Ergebnisse der Analyse interpretiert und im besten Fall ergeben sich neue Erkenntnisse über fachliche Zusammenhänge. Die neunte Phase ist das Handeln mit und durch die gewonnenen Erkenntnisse. In der Arbeit von (Beierle und Isberner, 2006, S.145-146) wird diese Phase als dem Prozess anschließend beschrieben, dort ist sie also kein inhärenter Teil.

Die einzelnen Phasen des Prozesses werden in den folgenden Abschnitten genauer untersucht und beschrieben, um auf typische Probleme einzugehen, die im späteren Verlauf der Arbeit gelöst werden müssen sowie um zu beschreiben, wie mit ihnen verfahren werden soll. Dabei wird vermehrt auf die Beschreibungen aus der Arbeit von Cleve und Lämmel (2014) zurückgegriffen. Sowohl die fünfte wie auch die sechste Phase werden im hier implizit durchgeführt und finden sich nicht explizit als einzelne Abschnitte wieder.

Zunächst soll ein weiteres Modell betrachtet werden, das dem KDD ähnelt aber an Fragestellungen aus dem Bereich Quantified Self angepasst wurde. Dieser Prozess wird in der Arbeit von Li u. a. (2010) beschrieben. Es besteht hingegen nur aus fünf Phasen, Vorbereitung, Sammeln, Integrieren, Reflektieren, Handeln. Dabei ist das Grundprinzip dem KDD-Prozess sehr ähnlich, aber das Augenmerk liegt mehr auf kurzfristigeren Fragestellungen, die sich nach dem Umlauf häufig nicht als beantwortet, sondern nur modifiziert zeigen. Gerade in der Quantified Self Bewegung sind Fragestellungen veränderlich beziehungsweise kurzlebig. Sie benötigen oft mehrere Durchläufe von erfassen, verarbeiten, auswerten und analysieren, bis eine zufriedenstellende Antwort gegeben werden kann, dabei müssen oft die erfassten Daten oder verwendeten Analysen überdacht und justiert werden. So können sich entweder immer speziellere Fragen entwickeln von 'Warum habe ich Kopfschmerzen?' über 'Wann genau habe ich Kopfschmerzen und was habe ich unmittelbar davor getan?' zu 'Was esse ich, das Kopfschmerzen auslöst?'. Es können sich aber auch bei gleichbleibender Fragestellung 'Warum schlafe ich schlecht?' veränderliche Daten und Analysen ergeben, so könnte es sein, dass das alleinige Überwachen der Ernährung, Alkohol, Koffein und dem Schlafrhythmus keine Antwort ergibt, dann kann noch ein Schallpegel-Sensor, ein Temperatur- und CO2-Sensor im Schlafzimmer angebracht werden. Selbst durch die neuen Daten könnte es sein, dass sich kein Ergebnis ergibt, weil die falsche Analyse gewählt wurde. Es ist also vielmehr ein iterativer Prozess als durch eine einmalige Abfolge abbildbar. Aus dieser Erkenntnis wird für diese Arbeit gezogen, dass der KDD-Prozess iterativ und nicht singulär gesehen werden sollte. Zudem ist es durchaus möglich, dass nach der Analyse noch einmal bei einem früheren Schritt eingesetzt werden muss, weil die Verarbeitung der Daten oder die nachfolgende Transformation oder Analyse nicht stimmig war.

2.3.1 Hintergrund- und Domänenwissen

In dieser Phase wird Wissen über den Datenbestand und die dazugehörigen Domänen geschaffen. Dies ist von Bedeutung, um im späteren Verlauf die Daten verarbeiten zu können. Um entscheiden zu können, ob Daten fehlerhaft sind oder fehlen, ist ein Verständnis für die Natur und Herkunft der Daten notwendig. Mögliche Fragestellungen dieser Phase sind: Was sind semantische Zusammenhänge? Wo überschneiden sich Datensätze? Was ist semantisch äquivalent, was nicht? Welche Zusammenhänge oder Korrelationen lassen sich vermuten? Was wäre eine zu erwartende aber irrelevante Korrelation und weitere. Ebenso ist das Interpretieren von Ergebnissen schwierig wenn nicht klar ist, was sie überhaupt auf die Domäne bezogen bedeuten. Somit ist es essentiell, zu Beginn ein Gefühl und eine Wissensgrundlage für die Daten selbst und ihre Domäne zu schaffen, um im späteren Verlauf qualitative Entscheidungen treffen zu können. Des Weiteren sollte überlegt werden, welche Fragestellung verfolgt werden soll. Geht es um die Antwort auf eine spezielle Frage, die Vorhersage eines bestimmten Wertes oder die Klärung der Herkunft einer Anomalie? Die spätere Wahl der Daten, Verfahren und Techniken kann von der Fragestellung abhängig sein.

2.3.2 Datenselektion und -integration

Die Datenselektion und -integration ist die Phase, in der Daten ausgewählt werden, die betrachtet werden sollen. Sie werden entweder aus einer bestehenden Datenbasis oder als Vorbereitung um sie zu erfassen ausgewählt.(Bezug auf das Modell von Li u. a. (2010)). Es muss entschieden werden, welche Daten für den Prozess relevant und/oder interessant sind. Diese Selektion wurde für die Arbeit bereits in der vorangegangenen Bachelorarbeit getroffen Lüdemann (2016a). Die dort ausgewählten Geräte wurden seither weiter verwendet, sodass sich eine vergleichsweise große Datenbasis ergeben hat. Im Rahmen dieser Arbeit muss entschieden werden, ob diese Daten ausreichen oder ob aus anderen Quellen weitere Daten mit einbezogen werden sollen. Neben der Selektion der Daten gehört auch deren Integration zu dieser Phase. Bei der Integration werden Daten aus verschiedenen Quellen zu einem zentralen Datenbestand zusammengeführt. Dies kann Probleme hervorrufen, auf die hier kurz eingegangen werden soll.

Mögliche Probleme:

- Entitätenidentifikationsproblem (Das Datumsattribut ist in jedem Datensatz anders benannt)
- verschiedene Zeitschritte (stündliche, tägliche, minütliche Daten)
- gleich benannte Attribute die etwas anderes beinhalten (Name für den kompletten Namen oder nur den Nachnamen)

- Datenwertkonflikt (Zeitstempel oder Maßeinheiten in anderen Formen (kg vs Gramm oder Zeitstempel in hh:mm gegen Millisekunden)
- Redundanzen (Für einen täglichen oder stündlichen Wert mehrere Werte in schneller Folge oder mit demselben Zeitstempel)

2.3.3 Datenvorverarbeitung und -bereinigung

Die nächste Phase des KDD-Prozesses ist die Datenvorverarbeitung und -bereinigung. Da die Daten zumeist nicht in der Qualität vorliegen, in der sie analysiert werden, können muss diese Qualität erzeugt werden. Dies geschieht durch eine geeignete Verarbeitung und Bereinigung der Daten. Die Datenqualität ist eine maßgebliche Grundlage für den Erfolg der Analyse und die Qualität der Ergebnisse. Daten aus fremden Quellen sind oft nicht bereinigt oder auf fremde Szenarien angepasst. Dadurch können sie Dimensionen enthalten, die für die eigene Verarbeitung nicht relevant oder im eigenen Szenario zu ungenau sind. Des Weiteren müssen fehlerhafte, unsinnige oder nutzlose Daten entfernt oder angepasst, Ausreißer und Rauscheffekte betrachtet und ggf. behandelt werden.

Fehlende Daten
Fehlende Daten sind ein Problem, da Analysen und Verfahren oft nicht damit umgehen können, wenn ein Wert fehlt. Das heißt, im Allgemeinen müssen sie irgendwie ersetzt werden. Zuallererst muss erkannt werden, wann ein Datum fehlt und was dieses Fehlen bedeutet. Denn auch das Fehlen einer Information kann eine Information sein. Zum Beispiel, dass ein Wert nicht erhoben wurde. Bei manueller Erhebung wäre dies ein Hinweis darauf, dass er entweder nicht in die vorgegebene Skala passt, dass etwas dazwischen kam oder, dass das System dem Benutzer zu umständlich ist etc. Bei maschineller Erhebung kann das Fehlen ein Hinweis darauf sein, dass ein Sensor kaputt, nicht getragen oder ohne Strom ist. Es kann aber auch sein, dass ein Wert in einem größeren Zeitabstand gemessen wird als alle anderen, dann gäbe es zu diesem Zeitpunkt einfach keinen Messwert. Soll der fehlende Wert ersetzt werden, gibt es verschiedene Möglichkeiten, fehlende Daten zu ersetzen. Zum einen können sie händisch nachgetragen werden, sofern die Anzahl es zulässt und ein Wert realistisch vermutet werden kann. Es könnte aber auch ein Platzhalter eingefügt werden. Dies kann entweder ein Wert sein, der außerhalb des Wertebereichs liegt, bei Puls z.B 999 um zu erkennen, dass dieser Wert fehlt, aber nicht ausgewertet werden kann oder unknown bei nominalen Werten. Dies gewährleistet, dass die Daten ausgewertet werden können aber dennoch als ersetzt zu erkennen sind. Für numerische Werte kann dabei aber das Problem entstehen, dass mathematische Auswertungen verfälscht werden, da zum Beispiel Durchschnitt und Steigung verfälscht sind. Somit können Werte auch interpoliert werden, also von einem Algorithmus nach Betrachtung der vorangegangenen und folgenden Werte eingefügt werden.

Falsche Daten
Um zu erkennen, ob Daten falsch sind, wird ein Validitätsmodell benötigt, in dem beschrieben ist, wann Daten valide sind. Daten können besonders aus Sensoren auch falsche Daten beinhalten. Sie entstehen bei Messfehlern, Ungenauigkeit oder weil die Sensoren eine Software haben, die fehlende Daten extrapoliert. Um einschätzen zu können, ob ein Wert falsch ist, muss ein Datenmodell herangezogen werden. Dies kann sich unter anderem auf Grenzwerte und Steigung von einem Wert auf den nächsten beziehen. Bei Puls wäre dies zum Beispiel, dass der Wert zwischen 0 und 230 liegen muss. Dies wäre erstmal eine grobe erste Einschätzung, da die Randbedingungen je nach Person schwanken, da sie sich auf diverse körperliche Merkmale beziehen. Dadurch ist aber jeder Wert, der außerhalb dieser Grenzen liegt, als invalide einzustufen. Es können auch Platzhalter aus den APIs der Daten kommen, sodass beim Puls ein numerischen, Wert, alphanumerische Platzhalter eingefügt werden. Invalide Daten können ähnlich zu fehlenden Daten behandelt werden, durch Ersetzen mit geschätzten, konstanten oder interpolierten Daten.

2.3.4 Datentransformation

Die Phase der Datentransformation ist notwendig, um die vorliegenden Daten in eine Form zu bringen, in der sie von den gewählten Data Mining Verfahren verarbeitet werden können. Dies umfasst unter anderem die Anpassung des Datums, der Datengranularität und den Datentypen. Einige Verfahren können nur mit numerischen Daten arbeiten, andere nur mit nominalen bzw. ordinalen Daten. Ebenfalls kann es notwendig sein, Attribute zu selektieren, um ein besseres oder schnelleres Ergebnis gemäß der folgenden Verfahren zu gewährleisten. Bei sehr großen Datendimensionen kann es sein, dass einige Attribute die Datenmenge nur unnötig vergrößern, ohne dass von ihnen ein zu erwartender Mehrwert ausgeht.

Datumsangaben
Daten können mit verschiedenen Datumsangaben vorliegen. Je nachdem, ob die Systemsprache Deutsch oder Englisch ist sowie der Handhabung der Zeitstempel. Diese können in das Datum kodiert sein oder als weiteres Attribut gespeichert werden, auch dessen Form kann variieren, je nach Systemsprache und Zeitzone. Dies muss validiert werden, um dann einen gemeinsamen Nenner für alle Daten zu finden. Die gewählte Form sollte eine sein, die in den weiteren Verfahren nutzbar ist. Ebenfalls ist es möglich, dass die Daten nicht mit einem *dd.MM.YYYY* Stempel vorliegen, sondern sogar mit drei einzelnen Attributen. In diesem Fall müssen die Attribute zusammengeführt werden. Zu erwartende Datumsangaben wären *MM.dd.YY, MM.dd.YYYY, MM.dd.YY HH:mm, YYYY.MM.dd* oder anstatt der Punkte mit Bindestrichen getrennt, ein Beispiel wäre *YYYY-MM-dd*. Ebenfalls ist es möglich, dass es keinerlei Trennzeichen

gibt, dann sähe ein Zeitstempel so aus: *YYYYMMdd*. Weiterhin kann es sein, dass der Monat nicht als Zahl sondern als Kürzel angegeben ist, zum Beispiel *Sep* anstatt *09*.

Einheiten und Schreibweisen
Daten können mit unterschiedlichen Einheiten vorliegen. Bei den Körperdaten, mit denen hier gearbeitet werden soll, ist es nicht zu erwarten, dass sie unterschiedliche Attribute verschiedene Maßeinheiten haben, wie z.B. Messwerte in Gramm und Kilogramm. Aber es ist durchaus möglich, dass der Ort der Einheitsangabe variiert oder sie nur den Metadaten zu entnehmen ist. Die Einheitsangabe kann im Attributnamen oder im Attribut selbst zu finden sein. Sollte die Angabe im Attribut stehen, verändert sich dadurch der Datentyp und verhindert eine numerische Auswertung. Die untenstehende Tabelle 2.1 zeigt, wie Maßeinheiten angegeben sein könnten. Dabei ist zu beachten, dass die Systole vom Datentyp String wäre, das verhindert, die Verarbeitung mit metrischen Analysen. Hier müsste also das Einheitsmaß entfernt werden.

Gewicht (kg)	**Systole**	**Ruhepuls**
74,6	112 mmHg	61
74,4	110 mmHg	60
74,7	111 mmHg	60

Tabelle 2.1: Beispiel der zu erwartenden Einheitenbeschreibung, kg im Attributs Namen, mmhG im Wert und Ruhepuls nur in den Metadaten.

Bei den anderen beiden Möglichkeiten muss eingeschätzt werden, wie notwendig oder störend die Maßeinheit im Attributs Namen ist und dieser gegebenenfalls angepasst werden.

Datengranularität
Daten, die von Sensoren stammen, können durchaus in unterschiedlichen Intervallen erhoben werden. Neben Daten, die einmal am Tag erhoben werden, gibt es welche, die stündlich, minütig oder nur alle paar Tage erhoben werden. Zum Beispiel gibt es Werte, die automatisch einmal am Tag oder pro Minute erfasst werden. Des Weiteren gibt es Werte, die manuell einmal am Tag erfasst werden müssen, bei diesen kann es vorkommen, dass aus verschiedenen Gründen kein Wert erfasst wird oder erfasst werden kann. Dadurch entstehen Werte, die seltener als einmal am Tag erfasst werden. Um hier eine Analyse und auch Visualisierung zu ermöglichen, sollten die Werte in gleichbleibenden Zeitabständen vorliegen. Dafür müssen sie im Falle von fehlenden oder seltener erfassten Werten interpoliert werden. Im Falle von Minutendaten, die mit Tagesdaten verglichen werden sollen, müssen Intervalle gebildet werden, die die Daten eines Zeitraums zusammenfassen. Zum Beispiel mithilfe des Mittelwertes oder Medians.

Datentypen
Die Daten können in unterschiedlichen Datentypen vorliegen. Allerdings gibt es Analyseverfahren, die einen bestimmten Datentypen vorschreiben. In diesen Fällen muss der Datentyp der Daten ggf. angepasst werden. Dies kann unter anderem durch die Kodierung von Integer (metrisch) in Kategorien bzw. Klassen (ordinal) erreicht werden. Dabei wird einem Wert oder einer Werteskala eine ordinale Kategorie zugeordnet. Ein Beispiel dafür sei am Blutdruck gegeben und in folgender Tabelle dargestellt 2.2. Dies ist ein Beispiel, wie der Blutdruck in ordinale Kategorien eingeteilt werden könnte. Dadurch würde der Wert der Systole 110 durch optimal und der der Diastole 60 ebenfalls als optimal ersetzt.

Systole mmHG	**Diastole mmHg**	**Kategorie**
< 120	< 80	optimal
129-129	80-84	normal
130-139	85-89	erhöht
140-159	85-89	hoch
160-179	90-99	sehr hoch
>= 180	>= 110	extrem hoch

Tabelle 2.2: Beispiel für die Abbildung metrischer auf ordinale Werte

Für metrische Verfahren kann es notwendig sein, nominale oder ordinale Werte in numerische zu kodieren. Nominale Eigenschaften *ja/nein* können in boolesche Werte kodiert werden *1/0*. Bei ordinalen Werten ist es notwendig darauf zu achten, dass die natürliche Ordnung erhalten bleibt und nicht ein Zahlenwert beliebig gewählt wird. So könnte aus optimal 0 werden, aus normal 1, aus erhöht 2, und so weiter. Optional kann eine andere Werteskala mit unterschiedlicher Gewichtung gewählt werden.

Attributs-Selektion
Manche Mining Verfahren sind sehr aufwändig und können bei fehlenden Ressourcen sehr lange dauern oder zu keinem bzw. schlechteren Ergebnis kommen. Um derart aufwändige Verfahren nutzen zu können, muss im Voraus ermessen werden, welche Attribute in welchen Ausmaßen relevant sind und eingesetzt werden können. Es ist zu überlegen, welche Attribute in welchem Ausmaß in einem solchen Verfahren einen Mehrwert liefern können. Es kann sein, dass nur ein gewisser Datenausschnitt notwendig ist. So können die Daten nicht als Gesamtes betrachtet werden, sondern zum Beispiel in Ein-Jahres-Abschnitten. Ebenfalls ist es möglich, durch genügend Wissen über die Domäne die für die Analyse wichtigen Attribute manuell zu selektieren. Darüber hinaus gibt es diverse Verfahren der Datenaggregation und -selektion. Diese werden hier aber nicht näher betrachtet, da eine derart große Datendimension hier nicht abzusehen ist.

2.3.5 Data Mining

In der Phase des Data Minings werden verschiedene Analyseverfahren angewendet, um die Daten auszuwerten und Erkenntnisse zu gewinnen. Dabei können sowohl automatische wie manuelle Verfahren verwendet werden. Die Wahl der Verfahren hängt dabei von der Beschaffenheit der Datengrundlage wie auch dem Ziel des Data Minings ab. Manche Verfahren haben gewisse Ansprüche an die Daten was Datentypen, Dimension und Qualität angeht. Darüber hinaus sind einige Verfahren für manche Fragestellungen besser geeignet als für andere. Hier wird kurz auf für Körperdaten möglicherweise interessante Verfahren eingegangen.

2.3.5.1 Cluster-Analyse

Die Cluster-Analyse ist unter anderem gut, um Daten in Gruppen mit großer Ähnlichkeit zu gruppieren. Zum Beispiel kann dadurch eine homogene Kundengruppe erstellt werden, die aufgrund ihrer ähnlichen Einkäufe gezielte Produktvorschläge bekommt Cleve und Lämmel (2014). Dabei ist es notwendig, die Daten mithilfe einer Distanz- oder Abstandsfunkion auf Cluster zu verteilen, sodass sich die Daten innerhalb des Clusters möglichst ähnlich sind und die Cluster untereinander möglichst unähnlich. Im Folgenden soll kurz darauf eingegangen werden, welche Algorithmen als potentiell nützlich angesehen werden. Die Definition und Erklärung der Algorithmen wird dabei dem Buch von Cleve und Lämmel (2014) entnommen.

Liste der möglichen Algorithmen

- k-Means
- Fuzzy c-Means
- k-Medoids
- DBSCAN

Partitionierende Algorithmen
K-Means ist ein effizienter Algorithmus für numerische Daten, der nach einer angegebenen Clusteranzahl die Centroiden jedes Clusters rät und sich dann mit der Distanzfunktion über Iterationen hinweg einer optimalen Clusterverteilung nähert. Das heißt auch, dass durch die Iteration die partitionierenden Cluster-Algorithmen den Punkten erlauben, die Cluster während der Berechnung zu wechseln. Bei ihm muss jedoch die Anzahl der Cluster k im Vorhinein festgelegt werden. Dies kann durchaus zu Problemen führen, da der Nutzer entscheiden muss, was die optimale Clusteranzahl ist.

Der Fuzzy c-Means Algorithmus arbeitet ebenfalls mit numerischen Daten und kann genutzt werden wenn sich abzeichnet, dass eine totale Zuordnung zu einem einzigen Cluster für die Datenpunkte nicht optimal ist. Im Fuzzy c-Means Algorithmus können Daten zu mehreren Clustern gleichzeitig gehören, jeweils zu verschiedenen prozentualen Anteilen.

Der k-Medoids Algorithmus ähnelt dem k-Means sehr, nur, dass bei ihm nicht der Centroid den Mittelpunkt eines Clusters ist, sondern der Medoid. Der Centroit ist ein errechneter Mittelpunk, der Medoid ist ein möglichst zentraler Punkt der Eingabemenge. Der k-Medoids Algorithmus kann im Prinzip auch mit nicht numerischen Daten arbeiten, dann muss er allerdings so implementiert sein, dass der Medoid nicht berechnet, sondern im Vorhinein bestimmt wird, berechnet wird der dichteste Punkt am Centroid. Sollte der k-Means Algorithmus keine zufriedenstellenden Ergebnisse liefern, sollte dieser Algorithmus ausprobiert werden.

Dichtebasierende Algorithmen
DBSCAN gehört anders als die drei bereits beschriebenen Algorithmen zu den dichtebasierten und nicht zu den partitionierenden Cluster-Algorithmen. Diese haben den Vorteil, dass sie nicht auf konvexe Clusterformen angewiesen sind, sondern auch Cluster erkennen können, die sich wie ein Schlauch durch den n-dimensionalen Raum ziehen. Beim DBSCAN Algorithmus werden Cluster daran bestimmt, wie dicht die Datenpunkte beisammenliegen. Bereiche mit großer Dichte werden zu Clustern, die durch Bereiche mit geringer Dichte getrennt sind. Dabei ergibt sich die Zahl der Cluster automatisch, es ist jedoch wichtig, die Parameter zur Dichtebestimmung gut zu wählen, da sie einen großen Einfluss auf das Ergebnis haben.

2.3.5.2 Klassifikation

Die Klassifikation teilt Datensätze in vorher festgelegte und durch Kriterien beschriebene Klassen ein. Diese Klassen können zum Beispiel die Kreditwürdigkeit eines Kunden anhand seines Alters, Wohnsitzes, Bestelleigenschaften und Grundeinkommens klassifizieren Cleve und Lämmel (2014). Wurden Testdaten klassifiziert, können Algorithmen diese Klassifizierung auf große Datensätze anwenden, deren Klassenzugehörigkeit noch unbekannt ist. Dabei ist es notwendig, eine genaue Vorstellung davon zu haben, welche Klassen es geben soll und wie sie sich definieren, da von diesen Kriterien abhängt, wie gut das Ergebnis sein kann. Im Folgenden werden einige Algorithmen kurz behandelt, die potentiell nützlich sein können, wenn die hier vorliegenden Daten klassifiziert werden sollen. Die Definition und Erklärung der Algorithmen wird dabei dem Buch von Cleve und Lämmel (2014) entnommen.

Liste der möglichen Algorithmen

- k-Nearest-Neighbour

- ID3 bzw. C4.5
- Naive-Bayes

Instanzenbasierte Verfahren
Der k-Nearest-Neighbour-Algorithmus kann mit metrischen sowie mit nominalen und ordinalen Daten arbeiten, indem die Berechnung des Abstandsmaßes angepasst wird. Wichtig dabei ist, dass dem Algorithmus eine Möglichkeit gegeben ist, die Ähnlichkeit der Attribute festzustellen. Die Klasse eines unklassifizierten Objekts wird anhand der Klassen seiner k-ähnlichsten Objekte bestimmt. Dieser Algorithmus bezieht alle Attribute eines Datensatzes mit ein und läuft dadurch Gefahr, unpräzise zu werden, wenn die herangezogenen Attribute für die Klassifizierung irrelevant sind. Dem muss entweder durch Gewichtung oder Selektion der Attribute vorgebeugt werden. Ebenso ist es notwendig, die Trainingsdaten gut zu wählen, damit er den Lösungsraum möglichst gleichmäßig aufspannt.

Entscheidungsbäume und wahrscheinlichkeitsbasierte Verfahren
Entscheidungsbäume arbeiten in der Regel auf diskreten Werten, nur durch weitere Arbeitsschritte wie Wertgruppierungen und Definition von Schwellwerten die den Werteraum in zwei Teile teilen, können nicht diskrete Werteräume verarbeitet werden. Dabei arbeiten sie auf metrischen Daten. Bei Entscheidungsbäumen wird ein Attribut als Startattribut manuell oder zufällig gewählt oder berechnet. Davon ausgehend wird ein Baum aufgebaut, der entscheiden kann, in welche Klasse ein unklassifiziertes Objekt gehört. Bei diskreten Werten kann der ID3 Algorithmus angewendet werden, dieser berechnet automatisch das optimale Startattribut und den optimalen Baum zur Bestimmung. Für numerische Daten ist der Nachfolger des ID3 der C4.5 geeignet. Bei diesem Algorithmus werden numerische Daten in Intervalle unterteilt, die dadurch zu ordinalen Werten werden. Entscheidungsbäume haben manchmal das Problem des Overfittings auf Testdaten. Dabei lernen sie sozusagen die Testmenge auswendig und bilden dabei kein gutes Modell zur Bestimmung. Um dies zu verhindern, können die Bäume künstlich verkürzt werden, dies wird als Prunning bezeichnet.

Um dieses Problem zu umgehen kann der Naive-Bayes Algorithmus verwendet werden. Er ist ein wahrscheinlichkeitsbasiertes Verfahren und stellt somit fest, welches die wahrscheinlichste Klasse eines Objektes ist. Der Vorteil dabei ist, dass vorher kein Modell trainiert wird, dahingegen geht dieser Algorithmus aber leider auch von der Unabhängigkeit der Attribute aus und diese ist oft nicht gegeben.

2.3.5.3 Assoziationsanalyse

Die Assoziationsanalyse soll Zusammenhänge zwischen Datensätzen erkennen. Ein beliebtes Beispiel hierbei ist die Warenkorbanalyse. Sie wird genutzt um zu erkennen,

welche Waren häufig zusammen gekauft werden um dann gezielt Werbung oder Angebote zu geben oder um die Produktplatzierung zu optimieren Cleve und Lämmel (2014). Die Analyse dient dazu, Regelmäßigkeiten in den Daten nicht nur zu erkennen, sondern auch das Verhalten neuer Datensätze vorherzusagen. Dies geschieht, indem Regeln aufgestellt werden, die den Klassifikationsregeln zwar ähneln aber auch auf Zusammenhänge zwischen Werten anwendbar sind. Solche Regeln haben oft die Form: Wenn A, dann auch B. Die Algorithmus Beschreibung und Definition wird dem Buch von Cleve und Lämmel (2014) entnommen.

A-Priori
Der A-Priori Algorithmus ist ein iteratives Verfahren dessen Ziel es ist, Assoziationsregeln zu erstellen. Dafür werden aus den Objekten *Frequent Item Sets* gesucht, also Objektmengen, deren relative Häufigkeit einen Schwellwert übersteigt. Diese Assoziationsregeln beschreiben dann die den Daten inhärente Struktur und können genutzt werden, um das Verhalten neuer Daten vorauszusagen.

2.3.5.4 Explorative Datenanalyse

Neben den algorithmenunterstützten Möglichkeit können die Daten auch explorativ analysiert werden. Das Ziel der explorativen Datenanalyse ist es unter anderem, einen Überblick über die Daten und ihre Struktur zu bekommen. Es handelt sich um ein deskriptives, statistisch unterstütztes Verfahren, bei dem Abhängigkeiten, Regelmäßigkeiten und die Datenstruktur beobachtet und aufgezeigt werden sollen. Außerdem sollen Zusammenhänge deutlicher werden und die Daten besser einzuschätzen und zu bewerten sein. Ebenso sollen Ausreißer und fehlende Daten erkannt und analysiert werden, um die Daten dann ggf. in einen größeren Kontext setzen zu können und Fragestellungen zu beantworten. Geprägt wurde diese Art der Analyse von John W. Tukey, erstmalig in einem 1962 veröffentlichten Artikel 'The Future of Data Analysis' Tukey (1962) und dem daraufhin erschienenen Buch Tukey (1977).

Das Untersuchen der Daten geschieht mit verschiedenen Werkzeugen. Neben statistischen Auswertungen (Median, arithmetisches Mittel etc.) kommen tabellarische Ansichten, Diagramme und Visualisierungen zum Einsatz. Dies können sein: Box Plot, Histogramme, Linien -Kreis - Balkendiagramme, Mosaik Plots, Streudiagramme und viele mehr. Dabei werden neben den Daten als Ganzes auch einzelne Aspekte betrachtet.

Der Vorteil der explorativen Datenanalyse besteht unter anderem darin, dass die Daten nicht nur maschinell, sondern auch manuell untersucht werden. Dabei kann besser auf den Kontext der Daten eingegangen und Hypothesen genau untersucht werden. Zum anderen wird dem Analysten dabei ein gutes Gefühl für die Daten und ihre Eigenschaften gegeben, sodass er ggf. besser die richtigen Algorithmen bestimmen kann

um mit den Daten zu arbeiten. Durch die visuelle Art der Datenuntersuchung entstehen gleichzeitig mit der Analyse Möglichkeiten, die Daten transparent zu machen. Es ist mit dieser Methode viel einfacher, an die Daten neugierig heranzugehen, um ein Gefühl für die möglichen Erkenntnisse zu bekommen, die die Daten liefern könnten. Die meisten Algorithmen, die hier auch schon vorgestellt wurden, benötigen eine präzise Fragestellung, die noch dazu oft thematisch eingegrenzt ist.

Der Analytiker benötigt ein gewisses Maß an Anwendungswissen, um die erkannten Auffälligkeiten in einen sinnvollen Kontext setzen zu können. Ebenso hängen die möglichen Ergebnisse an den Erfahrungen und Möglichkeiten des Analysten. Dies beinhaltet gegebenenfalls auch selektive Wahrnehmung. Es besteht die Gefahr, dass der Analyst im Voraus Annahmen über die Daten trifft, die er dann nur zu bestätigen sucht. Dies kann dazu führen, dass Annahmen mit einer dünnen Datenlage untermauert werden und gegenteilige Beweise oder Erkenntnisse ganz anderer Art übergangen werden.

2.3.5.5 Zeitreihen

Zeitreihen sind Daten, die durch einen Zeitstempel zeitlich geordnet sind. Dies ermöglicht, dass Datensätze nicht nur für sich allein betrachtet werden, sondern in einem Kontext mit den Datensätzen deren Zeitstempel davor und danach datieren. Eine Zeitreihe lässt sich mithilfe statistischer Kennzahlen beschreiben, dazu gehören der Trend, gleitende Durchschnitte, Differenzen, Extreme und Wachstumsraten. Des Weiteren können Zeitreihen in ihre Komponenten geteilt werden. Dies sind zum einen die Trendkomponente, also ein Abschnitt in einer Zeitreihe, die eine langfristige systematische Veränderung zeigt. Zum Anderen die zyklische Komponente, sie umfasst gleichbleibende Veränderungen, die sich in regelmäßigen Abschnitten wiederholen. Die dritte Komponente ist die zufällige oder Restkomponente, diese beschreibt all jene Abschnitte, die nicht mit zur Trend- oder Zyklischen-Komponente zu zählen sind.

Wenn es möglich ist ein Modell zu erstellen, das die Veränderungen über Zeit beschreibt, ist es möglich Prognosen zu geben, wie sich die Zeitreihe in der Zukunft verändern wird. Außerdem ist es durch die zeitliche Verortung einfacher, die Daten in einen Kontext mit anderen Daten zu setzen, die am gleichen Datum erhoben wurden, sei es Wetter, Kalendereinträge, Nachrichten oder weitere Sensoren. Zeitreihen lassen sich sowohl deskriptiv wie auch explorativ analysieren.

2.3.5.6 Annotieren/Feature Engineering

Durch die Annotation der Daten können Metadaten, die zur Kontextualisierung dienen, an die Datensätze geschrieben werden, um die Auswertung zu vereinfachen. So

können zum Beispiel subjektive Bemerkungen, Anmerkungen, Kalender oder Ortsinformationen dazugeschrieben werden, die darüber hinaus dazu dienen können, Daten zu klassifizieren.

Das Feature Engineering ist ein Weg, aus den vorliegenden Daten weitere Input Features zu generieren. Dies können von einfachen statistischen Auswertungen bis hin zu komplizierten Berechnungen ganz unterschiedliche Werte sein. Sie können dazu genutzt werden, wichtige Informationen für Algorithmen oder Nutzer hervorzuheben, eigenes oder fremdes Domänenwissen in die Daten zu bringen oder um darauf Algorithmen des maschinellen Lernens laufen zu lassen. Auf maschinelles Lernen soll hier aber nicht weiter eingegangen werden.

2.3.6 Dateninterpretation und -evaluation

Im letzten Schritt des KKD-Prozesses geht es darum, die Daten zu interpretieren und die Ergebnisse zu evaluieren.

Interpretation
Bei der Interpretation geht es darum, die Informationen, entdeckte Muster und Abhängigkeiten aufzubereiten, zu interpretieren und sie, sowie die Ergebnisse, in verständlicher Form darzustellen. Dazu können diverse Visualisierungen dienlich sein. Die Ergebnisse sind genauestens zu betrachten und ggf. in einen Kontext zu setzen. Durch die Anwendung von Domänenwissen können dann aus den Ergebnissen neue Erkenntnisse gewonnen werden. Ebenfalls ist zu entscheiden, ob die vorliegenden Ergebnisse zufriedenstellend sind oder ob zu einem früheren Zeitpunkt des Prozesses zurückgegangen werden sollte, um mit veränderten Parametern, Daten oder Algorithmen zu arbeiten.

Evaluation
Bei der Evaluation werden die Ergebnisse hinsichtlich der definierten Ziele bewertet. Dabei wird auf folgende Kriterien Bezug genommen: *Validität, Neuartigkeit, Nützlichkeit und Verständlichkeit* Cleve und Lämmel (2014). Die Kriterien sollen im Folgenden etwas genauer beschrieben werden.

- *Validität* ist ein objektives Muster dafür, ob ein gefundenes Modell auch für neue Daten gültig ist.
- *Neuartigkeit* beschreibt, ob das Wissen erweitert, ergänzt oder mit Widersprüchen belegt wurde.
- *Nützlichkeit* beschreibt, ob das neue Wissen einen praktischen Nutzen für den Anwender bereithält.
- *Verständlichkeit* beschreibt, wie gut das Wissen nachvollzogen werden kann.

Die Bewertung anhand dieser Kriterien kann je nach verwendetem Data Mining Verfahren schwierig bis unmöglich sein. Bei einer Klassifizierung kann durch die Aufteilung der Daten in Test und Validierungsdaten direkt gezeigt werden, welche Validität ein Modell hat, bei anderen Verfahren ist dies ungleich schwerer.

2.4 Datenanalyse

In diesem Abschnitt soll dargelegt werden, welche Daten im Rahmen dieser Arbeit in Betracht gezogen wurden, welche Probleme dabei jeweils auftraten und welche Daten letztendlich erhoben wurden. Es wird auf die Daten selbst und ihre Struktur eingegangen, darüber hinaus auf die Qualität in der sie vorliegen und das Modell, mit dem unter anderem diese Qualität bestimmt wurde.

2.4.1 Quellen

Bei der Auswahl der Daten ging es vor allem darum, die Daten im Alltag ohne großen Mehraufwand mit möglichst einfachen und bezahlbaren Mitteln zu erheben. Das heißt, verwendete Geräte sollten nicht kompliziert verkabelt oder invasiv in den Körper eingefügt werden müssen, eine Verbindung mit dem Smartphone eingehen und möglichst wenig manuelle Schritte haben. Im Abschnitt 2.1.4 wurde darauf eingegangen, warum manuelles Erfassen ein Problem ist. Als Konsequenz aus diesen Erfahrungen und der Recherche wurde im Rahmen dieser Arbeit auf manuelle Daten verzichtet, da sie über einen größeren Zeitraum nicht sinnvoll zu erfassen sind, wenn es nicht gut designte Lösungen für das Erfassen gibt. Im Folgenden wird auf die Daten eingegangen, die für diese Arbeit in Frage gekommen sind. Dabei wird betrachtet, ob sie erhoben werden oder nicht und was die Gründe dafür sind. Dazu ist zu sagen, dass die Kriterien des Erhebens nicht nur mit dem Nutzen für diese Arbeit zu tun haben, sondern auch sehr weltliche Gründe, wie zu hoher Aufwand und Kosten.

Fitnessdaten/ Bewegungsdaten
Daten darüber, wie viele Schritte gegangen werden, welche Entfernung dabei zurückgelegt wird und wie sportlich aktiv man ist erscheinen, durchaus interessant um einzuschätzen, wie aktiv die Testperson ist, sowohl im sportlichen wie auch generellen Sinne. Sie werden mithilfe des Fitbit-Armbandes erhoben.

Schlafdaten
Daten, die über den Schlaf erhoben werden, können unter anderem sein: die Schlafdauer, der Zeitpunkt des Zubettgehens, die Effektivität des Schlafes und die Anzahl der Unterbrechungen des Schlafes. Dadurch kann der Zusammenhang zwischen dem

Schlaf und anderern Körperdaten betrachtet werden, ggf. sogar im Zusammenhang mit äußeren Umständen.

Die Schlafdaten werden mithilfe eines Fitbit-Armbandes erfasst.

Neben den Körperdaten des Schlafes könnte es interessant sein, die Umstände des Schlafes zu erfassen, Geräuschpegel, Sauerstoff und CO2 Gehalt der Luft. All dies könnte Einfluss auf den Schlaf nehmen. Dafür müssten jedoch Sensoren angeschafft werden und besonders in den ersten Jahren der Datenerfassung war es aufgrund der Umstände nicht möglich, sie zu platzieren.

Puls
Der Puls über den Tag und der Ruhepuls können Aufschluss darüber geben, wie aktiv man gewesen ist, wie schnell der Puls sinkt und steigt, wie anstrengend eine sportliche Aktivität war und auch, wie anstrengend alltägliche Dinge wie Gehen für den Körper sind. Besonders im Zusammenhang mit weiteren Daten kann die Auswertung sehr spannend sein.

Der Puls und Ruhepuls werden mithilfe eines Fitbit-Armbandes erfasst.

Blutdruck
Der Blutdruck steht oft im Zusammenhang mit dem Körperbefinden und verändert sich recht rasch bei Stress und ähnlichen Belastungen, dies wäre eine der Beobachtungen die zu machen sind, wenn der Blutdruck täglich erfasst wird. Darüber hinaus könnte er Hinweise auf Wechselwirkungen mit äußeren und inneren Faktoren geben.

Der Blutdruck wird mit einem Handgelenk-Blutdruck-Messgerät von Medisana erhoben.

Blutzucker
Der Blutzucker ist besonders im Zusammenhang mit der Ernährung und der Bewegung interessant. Darüber hinaus wäre es durchaus spannend, die Wechselwirkungen zwischen diesem und anderen Körperdaten zu erfassen. Allerdings müsste er durch tägliche Blutproben oder ein invasives Gerät erfasst werden. Dieser Wert wird daher nicht erfasst.

Blutwerte
Einen Verlauf der Blutwerte über längere Zeit samt natürlicher Schwankungen zu beobachten könnte eine weitere sehr bereichernde Erhebung sein, gemeinsam mit den Auswirkungen der Ernährung, Sport, Medikation und anderen. Allerdings ist die Erhebung der Blutwerte nur durch die Zusammenarbeit mit einem Arzt und Labor umzusetzen und erzeugt dazu ein recht hohen finanziellen Aufwand. Daher muss darauf verzichtet werden.

Gewicht
Das Gewicht und seine Veränderungen können Aufschluss auf Ernährungs- und Bewegungsgewohnheiten geben. Darüber hinaus ermöglicht dessen Erhebung Ergebnisse im Zusammenhang mit Krankheit, Stress oder Medikamenten und deren Veränderungen zu erfahren. Wo treten hier Regelmäßigkeiten auf, welchen Zusammenhang haben sie? Neben dem Gewicht können auch weitere Daten erfasst werden, Knochenmasse, Körperwasser, Körperfett und Muskelmasse.

Diese Werte werden mithilfe einer Analysewaage von Medisana erfasst.

Wetter
Das Wetter kann unter anderem im Zusammenhang stehen mit Bewegung oder dem Blutdruck. Daher kann es durchaus interessant sein, besonders Temperatur, Sonnenstunden, Luftdruck und Niederschlag zu betrachten. Es kann davon ausgegangen werden, dass die Wetterlage die Bewegung beeinflusst und der Luftdruck vielleicht Einfluss auf den Blutdruck hat.

Die Wetterdaten werden vom Deutschen Wetterdienst bezogen.

Kalender
Der Kalender des Nutzers kann Aufschlüsse darüber geben, ob ein großes Arbeitspensum vorliegt, er sich viel Freizeit gönnt oder Termine im Kalender hat, auf die er sich freuen kann. In erster Linie können diese Informationen zur Kontextualisierung genutzt werden. Darüber hinaus kann aber auch ein Sport-, Arbeits- und Reiseprofil erstellt werden. Die Kalenderdaten können über den Google Kalender bezogen werden.

Diese Anbindung ist jedoch sehr umfangreich, sodass sie im Rahmen dieser Arbeit nur zur händischen Kontextualisierung genutzt wird.

Ort
Der Ort kann dazu genutzt werden, um Daten zu annotieren um Zusammenhänge, die mit der Beschäftigung zu tun haben, zu erkennen. Verändern sich die Körperdaten bei der Arbeit, im Urlaub oder Zuhause? Macht es einen Unterschied wo Urlaub gemacht wird, ländlicher oder urbaner Urlaub? Diese und ähnliche Faktoren erscheinen im Bezug auf den Ort von Interesse.

Die Ortsdaten könnten über Google Timeline bezogen werden, da das Smartphone über Google aufzeichnet, wann es sich wo aufhält. Die Auswertung dieser Daten ist allerdings sehr aufwändig, sodass im Rahmen dieser Arbeit darauf verzichtet wird, sie allumfassend auszuwerten. Stattdessen wird manuell darauf zugegriffen, um einzelne Informationen zu erhalten oder bei der Kontextualisierung behilflich zu sein.

Hydration/Alkohol/Koffein
Sowohl die Zufuhr des Wassers wie auch die Menge an Alkohol und Koffein hat einen Einfluss auf verschiedene Körperdaten, diesen Zusammenhang aufzuzeigen kann durchaus interessant sein, um die genauen Auswirkungen zu sehen. Allerdings müssen

diese Daten manuell erhoben werden und dies ist, wie in den vorangegangenen Arbeiten gezeigt, ein Problem. Daher wird darauf im Rahmen dieser Arbeit verzichtet.

Stimmung
Das Bestimmen der Stimmung kann Aufschluss darüber geben, wie ein Tag oder Tagesabschnitt in der subjektiven Wahrnehmung des Nutzers war. Diese Einschätzung könnte den maschinell erfassten, objektiven Daten gegenübergestellt werden, um dort Zusammenhänge und Besonderheiten zu betrachten. Die Stimmung könnte ebenfalls Aufschluss auf Anomalien in den Daten geben oder Gründe für Veränderungen liefern. Die Stimmung wird nicht erhoben, weil das Erfassen zu manuell ist.

Stress
Der Stress könnte erhoben werden, um in den Daten, Phasen des emotionalen oder arbeitsbedingten Stresses, Phasen ohne Stress gegenüberzustellen. Dadurch könnte aufgezeigt werden, welchen Einfluss verschiedene Stressfaktoren auf die Daten haben.

Der Stress wird nicht erhoben, da das Erfassen zu manuell ist.

Krankheiten
Es könnte Tagebuch darüber geführt werden, wann welche Krankheit auftrat. Daraus könnten die Einflüsse der Krankheit auf die Werte gezeigt werden und vielleicht sogar Vorläufer, Datenanomalien oder Besonderheiten aufgezeigt werden, die die Krankheit ankündigen.

Krankheiten werden nicht erhoben, da das Erfassen zu manuell ist.

Medikation
Wann welche Medikamente eingenommen werden, über welchen Zeitraum und in welcher Stärke, könnte ebenfalls aufzeigen, welchen Einfluss sie auf die Körperdaten haben. Besonders interessant wäre dies für die Veränderung der Dosis eines dauerhaft eingenommenen Medikaments. Die Medikation wird nicht erhoben, da das Erfassen zu manuell ist.

Bilddaten
Über eine Kamera könnten tägliche Bilddaten dazu verwendet werden, diverse Informationen mit Gesichtserkennung aufzunehmen. Neben technisch erhobenen Emotionen und Stress könnten die Bilddaten verglichen werden, um dem Nutzer ein Gefühl dafür zu geben, wie sich die Daten auf das Aussehen niederschlagen. Es könnten saisonale Veränderungen bemerkt und aufgezeigt werden. Diese Daten werden nicht erhoben, da der Aufwand zu groß ist.

Wärmebilder
Neben den Bildern einer gewöhnlichen Kamera könnte eine Wärmebildkamera hinzugezogen werden. Auch diese Kamera könnte der technischen Bestimmung von Emotionen und Stress dienen. Sie könnte aber auch Informationen zur Körpertemperatur liefern, um dort Schwankungen zu beobachten. Ein weiterer interessanter Punkt wäre

die Verwendung um dem Nutzer aufzeigen zu können, welche Muskelpartien nach einem Workout besonders trainiert wurden, da sie wärmer sind als andere. Dadurch könnte sich eine Trainingseffizienz ergeben.

Diese Daten werden nicht erfasst, da der Aufwand im Rahmen dieser Arbeit zu groß wäre.

Überblick
Die untenstehende Tabelle soll abschließend einen kurzen Überblick darüber geben, welche Daten erhoben werden und welche nicht. Ebenso wird gezeigt, ob Literatur vorliegt in der diese Daten verwendet wurden.

Daten	**Erhebung**	**Verwendung in Literatur**
Fitness und Bewegungsdaten	+	Ja
Schlafdaten	o	Nein
Puls	+	Ja
Blutdruck	+	Ja
Blutzucker	-	Nein
Blutwerte	-	Nein
Gewicht	+	Ja
Wetter	+	Nein
Kalender	o	Nein
Ort	o	Ja
Hydration, Alkohol, Koffein	-	Nein
Stimmung	-	Ja
Stress	-	Nein
Krankheiten	-	Nein
Medikation	-	Nein
Bilddaten	-	Ja
Wärmebilder	-	Nein

Tabelle 2.3: Übersicht der möglichen Daten. **Legende:** - Nicht erhoben, o teilweise erhoben, + erhoben.

APIs

In diesem Abschnitt soll auf die APIs der Geräte eingegangen werden, mit denen gearbeitet wurde. Es soll betrachtet werden, wie die APIs aussehen, wie die Daten von dort geholt werden können und in welchen Formaten sie vorliegen. Außerdem soll auf Besonderheiten, Zugriffsrechte und mögliche Beschränkungen eingegangen werden, um ein Bild davon zu bekommen, wie aufwändig das Beziehen der Daten ist. Es wurden

außerdem Wetterdaten vom Deutschen Wetterdienst bezogen, auch diese API wird kurz beschrieben.

Fitbit
Die Fitbit API ist über OAuth2 gesichert und kann über einen Fitbit Account genutzt werden. Jedoch kann durch einen normalen Account nur ein Teil der Daten aus der API geladen werden. Sollen auch Minutendaten für Schlaf und Puls genutzt werden, wird ein spezieller Account benötigt, der von Fitbit genehmigt werden muss. Dieser Account ist unter anderem für Forschungszwecke gedacht und kann somit angefragt werden um, wie im Rahmen dieser Arbeit, für wissenschaftliche Zwecke genutzt zu werden. Dies bedeutet aber auch, dass man über die API als normaler Benutzer nicht im vollen Umfang auf die eigenen Daten zugreifen kann. Die Umsetzung und Nutzung der API Anbindung wurde aus der vorangegangenen Arbeit Lüdemann (2016a) übernommen. Die API wird in der hier genutzten Anbindung über die Konsole via REST Anfragen genutzt. Dabei können je nach Datensatz längere Zeiträume angefragt werden. Wie eine solche Anfrage aussieht zeigt das untenstehende Code Beispiel 2.1

```
1 java -jar target/FitbitBAHeart-0.0.1-SNAPSHOT.jar -call="https://api.
    fitbit.com/1/user/-/activities/activityCalories/date
    /2016-01-20/2016-06-19.json" > activityCalories1.txt
```

Listing 2.1: Code Zeile zum Abfragen der Aktivitätencalorien von Fitbit

Im Rahmen dieser Arbeit wurden die Daten rückwirkend für die letzten dreieinhalb Jahre heruntergeladen, wodurch ein paar Probleme auftreten, die bei regelmäßigeren kleinen Anfragen nicht zu erwarten sind. Das obenstehende Beispiel zeigt ein erstes Problem: selbst bei Daten die nur einmal am Tag erhoben werden, können nicht alle 1300 Tage auf einmal heruntergeladen werden. Die *Aktivitäten Kalorien* mussten in Schritten von ca. 6 Monaten geladen werden. Die *Minuten Sehr aktiv* hingegen konnten in Ein-Jahres-Schritten abgefragt werden. Dies war aus der API und der bereitgestellten Beschreibung allerdings nicht ersichtlich und musste durch ausprobieren und herantasten herausgefunden werden. Bei den Minutendaten wie Schlaf und Puls musste jeder Tag einzeln abgefragt werden. Darüber hinaus gibt es eine API Beschränkung von 150 Anfragen pro Stunde. Daher ist es sehr sinnvoll, die Anfragen so hoch wie möglich zu aggregieren. Was bei den Minutendaten aber dennoch ein hoher Zeitaufwand ist, da jeder Tag eine eigene Anfrage darstellt. Die Schnittstelle liefert als Antwort auf eine wie oben gezeigte Anfrage ein JSON Objekt, das in ein Textdokument gespeichert wird.

Fitbit bietet es auch an, die eigenen Daten über eine Webschnittstelle als CSV Datei herunterzuladen, jedoch war dies nach einem Test als ungeeignet eingestuft, da dort nur Zeiträume von bis zu 31 Tagen gedownloadet werden konnten. Die Datei, die dabei jeweils entsteht, ist darüber hinaus von mäßiger Qualität, die Bezeichnungen der Attribute sind auf Deutsch und die Umlaute erzeugen Codierungsprobleme, sodass so

gut wie jede Überschrift Fehler beinhaltet. Die Daten, die bereitgestellt werden, sind sehr lückenhaft. So gibt es nur die Information, dass an einem Tag Sport getrieben wurde, allerdings nicht wann, welcher und wie der Puls im Verlauf der Anstrengung war. Es fehlen also die Minutendaten für Schlaf, Schritte und Puls.

Neuerdings gibt es eine weitere Möglichkeit, an die eigenen Daten zu kommen. Es ist möglich, das gesamte Kontoarchiv von Fitbit herunterzuladen. Dies ist ein recht aufwändiger Prozess in dem eine Anfrage auf der Homepage gestellt wird, danach muss der Nutzer eine Bestätigungsmail bestätigen und dann je nach Größe des Kontos einige Tage warten, bis er das Archiv auf der Homepage herunterladen darf. Für das hier getestete Konto mit Bestand seit dem 10.06.2015 hat dies jedoch nur dreißig Minuten gedauert. Die Qualität dieses Downloads wird als gut bewertet, da er verständlich und vollständig ist. Dabei handelt es sich um alle Daten, die per Armband und App erhoben werden sowie die Profildaten. Im Rahmen dieser Arbeit musste auf die Web API zurückgegriffen werden, da die Möglichkeit mit dem Kontoarchivexport noch nicht bestand. Für eine fortlaufende Abfrage eines laufenden Systems muss dennoch mit der API gearbeitet werden.

Medisana
Die Medisana API ist ebenfalls über OAuth2 gesichert. Die Implementierung wurde aus der vorangegangenen Arbeit Lüdemann (2016a) übernommen, wobei das Herunterladen der rückwirkenden Daten mithilfe der auf der Homepage angebotenen Download-Funktion durchgeführt wurde. Die Homepage liefert, anders als bei Fitbit, schnell und einfach einen vollständigen Datenbestand in einer gut lesbaren und verarbeitbaren CSV Datei. Es gab dabei keine Probleme oder Schwierigkeiten.

Dabei entstehen zwei CSV Dateien, eine für die Daten der Analysewaage und eine für die Daten des Blutdruckmessgerätes. Diese können von Metadaten bereinigt und konvertiert verarbeitet werden. Auf die Schwierigkeiten, die sich dabei aufgrund der inneren Struktur der Daten ergaben, wird später eingegangen.

Deutscher Wetterdienst (DWD)
Die Wetterdaten wurden rückwirkend aus dem Onlinearchiv bezogen, ob und wie sie aktuell gezogen werden können wurde im Rahmen dieser Arbeit nicht betrachtet. Die Daten werden über einen FTP-Server, dem CDC (Climate Data Center) vom DWD angeboten. Dabei ist zu beachten, dass der Deutsche Wetterdienst zwischen validierten Archiven (historisch) und nicht validierten Daten (aktuell) unterscheidet. Diese Unterscheidung variiert allerdings je nach Messstation. Die hier verwendete Messstation ist Hamburg-Fuhlsbüttel, bei der Daten vor dem 31.12.2017 als historisch gelten. Alles danach fällt unter *recent* und ist in einem anderen Verzeichnis abgelegt. Diese Unterscheidung bedeutet, dass die neuen Daten die routinemäßige Qualitätskontrolle noch nicht vollständig durchlaufen haben (dwd). Es ist leider sehr schwer, über die Website des DWDs die Daten und die nötigen Informationen zu finden. Dazu kommt, dass die Daten nach Messstationen IDs sortiert sind, die an einer anderen Stelle der

Website herausgesucht werden müssen, da nicht einfach ersichtlich ist, welche Station wo in Deutschland welche ID hat. Das bedeutet, dass der Aufwand die Daten zu finden und richtig zu identifizieren hoch ist, aber die Daten frei zugänglich sind. Auf die Probleme, die in der Beschriftung der Daten und ihrer Qualität liegen, wird im nachfolgenden Abschnitt eingegangen.

2.4.2 Datenaufbau

In diesem Abschnitt soll genauer darauf eingegangen werden, wie die Daten, die aus den APIs bzw. der Webschnittstelle gezogen werden, aussehen.

Fitbit
Die Daten, die aus der Fitbit API kommen, haben ein JSON Format und sind aufgrund der Anfragestruktur nach Attributen und Zeitpunkten getrennt. Das bedeutet, dass es je nach Attribut unterschiedlich viele Dateien gibt, die zusammengeführt werden müssen. Die meisten Daten liegen in vier Dateien für den Zeitraum vor, dies trifft auf Distanz, Höhe, Kalorien, KalorienGrundumsatz, MinutenAktiv, MinutenLeichtAktiv, MinutenSehrAktiv, Schritte pro Tag und Stockwerke zu. Die Kalorien bei Aktivitäten mussten erneut und mit kleineren Abschnitten gezogen werden, sodass sie in acht Dateien vorliegen. Die Minutendaten für Schlaf, Schritte und Puls mussten für jeden Tag angefragt werden, sodass für jeden der 1300 Tage eine eigene Datei entsteht.

Die meisten Daten sind so aufgebaut, dass das JSON Objekt aus einem Value pro Tag über den Abfragezeitraum besteht. Das untenstehende Beispiel 2.2 veranschaulicht dies noch einmal. Diese Struktur haben alle Daten, die nur einmal am Tag erhoben werden.

```
{"activities-minutesFairlyActive":[
{"dateTime":"2015-07-20","value":"27"},
{"dateTime":"2015-07-21","value":"7"},
{"dateTime":"2015-07-22","value":"0"},
.
.
.
{"dateTime":"2016-07-19","value":"6"}
]}
```

Listing 2.2: Beispiel JSON für MinutenLeichtAktiv

Die Minutendaten beinhalten zusätzlich noch Metadaten sowie die Values für jede gemessene Minute des Tages. Das heißt, wenn der Sensor den Kontakt verliert, nicht getragen wird oder einen Messfehler hat, wird kein Eintrag generiert, die Minute fehlt

dann im JSON. Das untenstehende Beispiel veranschaulicht den Aufbau der Minutendaten anhand eines Ausschnitts aus den Pulsdaten von Fitbit. Diese bringen einen recht großen Satz Metadaten mit, die zumeist aus akkumulierten oder berechneten Daten bestehen und den Rechenaufwand verringern. Dazu gehören das Tagesmaximum und -minimum sowie die Anzahl der Minuten, in denen der Puls innerhalb einer der jeweiligen Pulszonen lag. Des Weiteren wird angegeben, welche Grenzen die Pulszonen haben, dies ist interessant, da sie sich mit dem Alter des Nutzers verschieben. Außerdem wird jeweils angeben, wie viel Kalorien durch die einzelnen Zonen verbraucht wurden.


```
{"activities-heart":[
  {"dateTime":"2015-09-17","value":{
   "customHeartRateZones":[],
     "heartRateZones":[
       {"caloriesOut":1889.36132,
        "max":99,
      "min":30,
      "minutes":1327,
      "name":"Out of Range"},
    {"caloriesOut":233.39041,
     "max":138,
     "min":99,
     "minutes":49,
     "name":"Fat Burn"},
    {"caloriesOut":0,
     "max":168,
     "min":138,
     "minutes":0,
     "name":"Cardio"},
    {"caloriesOut":0,
     "max":220,
     "min":168,
     "minutes":0,
     "name":"Peak"}
    ],
 "restingHeartRate":60}}],
 "activities-heart-intraday":
  {"dataset":[
    {"time":"00:00:00","value":74},
    {"time":"00:01:00",
    .
    .
```

```
33      {"time":"23:59:00","value":64}
34    ],
35   "datasetInterval":1,
36   "datasetType":"minute"}}
```

Listing 2.3: Beispiel JSON für Puls Minutendaten

In dem oben gezeigten Beispiel ist zu sehen, dass die Uhrzeit im Format *hh:mm:ss* und der eigentliche Wert ohne die Maßeinheit angegeben wird.

Medisana

Die Medisana Daten liegen in zwei Dateien vor, die durch die Downloadfunktion auf der Website bezogen wurden. Es liegt jeweils eine *.csv* Datei für den Blutdruck und die Daten der Analysewaage vor. Der Blutdruck enthält, neben den eigentlichen Messwerten, auch einige nutzergenerierte Daten wie Anmerkungen und der Einstellung welchen Aktivitätsgrad der Nutzer hat. Ebenso enthalten die Daten einige Metadaten wie die Optimalwerte nach WHO, wann der letzte Messwert genommen wurde sowie Maximum, Minimum, Durchschnitt und die Anzahl der Messwerte. Die folgende Tabelle 2.4 zeigt einen Ausschnitt der Rohdaten aus der Datei.

Datum - Uhrzeit	Systole	Diastole	Puls	Stimmung	Aktivität
28.07.2015 - 10:26	103 mmHg	63 mmHg	81 bpm	0	2
28.07.2015 - 10:31	105 mmHg	64 mmHg	83 bpm	0	2
28.07.2015 - 10:37	105 mmHg	62 mmHg	80 bpm	0	2
28.07.2015 - 23:44	96 mmHg	58 mmHg	74 bpm	0	2

Tabelle 2.4: Beispiel der Blutdruck Rohdaten

Die obige Tabelle 2.4 zeigt, dass das Datum im Format *dd.MM.yyyy - hh:mm* vorliegt und bei den Messwerten die Maßeinheit mit im Attribut angegeben ist. Dazu kommt die Spalte für die Nutzeranmerkungen, die nicht angezeigt wird, da sie vollständig leer ist. Sie wurde beim Erfassen der Daten nicht genutzt. Die durchgängige 0 der Stimmung deutet ebenfalls darauf hin, dass beim Erfassen des Blutdrucks keine Stimmung in der App angegeben wurde und so dem Defaultwert entspricht. Die Stimmung kann in drei Werten angegeben werden, die Beschreibung ist aus der App entnommen: 0 (super), 1 (so lala) und 2 (nicht so gut). Die Aktivität, die hier ebenfalls nicht verwendet wurde und daher immer auf dem Defaultwert 2 steht, kann vier Ausprägungen annehmen: 0 (aktiv), 1 (ausgeruht), 2 (normal) und 3 (krank). Was an dem oben gezeigten Ausschnitt aus den Rohdaten auffällt ist, dass alle vier Datensätze für denselben Tag sind, es gibt drei Messwerte morgens und einen abends. Dies kann an Messungenauigkeiten oder Mehrfachmessungen liegen. Außerdem kommt es vor, dass

aus einer einzelnen Messung in der App zwei werden, woher dieses Verhalten der App kommt ist nicht nachvollziehbar.

Die Gewichtsdaten bringen ebenfalls einige Metadaten mit, die allerdings nur aus Maximalwert, Minimalwert, Durchschnitt und Gesamtzahl der jeweiligen Messwerte beziehungsweise Attribute bestehen. Die Daten selber werden in der untenstehenden Tabelle 2.5 als kurzes Beispiel angegeben.

Datum - Uhrzeit	**Körpergewicht (kg)**	**Knochenmasse (%)**	**Körperfett (%)**	**Körperwasser (%)**	**Muskelmasse (%)**	**BMI**	**Aktivität**	**Stimmung**
28.07.2015 - 09:33	76,2	3,6	27,2	50,3	34	24,9	2	0
29.07.2015 - 10:09	75,9	3,6	27,1	50,4	34	24,8	2	0
29.07.2015 - 10:10	75,9	3,6	27,1	50,4	34	24,8	2	0
30.07.2015 - 10:39	75,4	3,6	26,8	50,7	34,2	24,6	2	0

Tabelle 2.5: Beispiel der Waagen Rohdaten

Auffällig ist hier, dass anders als in den Blutdruckdaten die Maßeinheiten nicht im Messwert sondern im Attributsnamen angegeben werden, das Datum liegt im selben Format *dd.MM.yyyy - hh:mm* vor. Neben den von der Waage gemessenen Werten Körpergewicht, Knochenmasse, Körperfett, Körperwasser und Muskelmasse wird der BMI bereits beim Erfassen berechnet. Mithilfe der App kann der Nutzer weitere Angaben machen, die sich dann in Form von Aktivität, Stimmung, Mahlzeit und den Anmerkungen in den Daten zeigen. Die Anmerkungen sowie die Mahlzeit-Spalte sind auch hier nicht aufgeführt, da die Anmerkungen leer sind und die Mahlzeit, wie schon bei der Stimmung gezeigt, immer 0 ist. Die Aktivität kann vier Ausprägungen annehmen und ist im Default 2, die Werte stehen für 0 (aktiv), 1 (ausgeruht), 2 (normal), 3 (krank). Die Stimmung wird genauso wie beim Blutdruck angegeben. Der Wert der Mahlzeit beschreibt ob sich 0 (vor) oder 1 (nach) einer Mahlzeit gewogen wurde. Da diese Angaben in der App nie verwendet wurden, stehen sie stets auf ihrem Defaultwert.

Deutscher Wetterdienst

Die Daten des deutschen Wetterdienstes wurden rückwirkend für den gesamten Erfassungszeitraum der Körperdaten heruntergeladen. Dabei wurde der FTP Server des CDC (Climate Data Center) genutzt. Nach einiger Recherche kam heraus, dass die

gewünschte Wetterstation (Hamburg-Fuhlsbüttel) die ID 1975 hat. Mit dieser Information ist es möglich, sowohl historische wie auch nicht historische Werte vom FTP Server zu ziehen. Dabei wird ein Archiv heruntergeladen, in dem eine Vielzahl an Dateien zu finden ist, die zum Großteil Metadaten betreffen und für die Ansprüche dieser Arbeit nur bedingt bis gar nicht relevant sind. Die Datei mit den Messwerten beinhaltet im Falle der historischen Datei die Messwerte vom 1.1.1936 bis zum 31.12.2017. Die Datei mit den aktuellen Daten beinhaltet somit die Daten vom 1.1.2018 bis hin zum Datum der Abfrage, in diesem Fall dem 14.03.2019. Beide Dateien liegen im txt Format vor, die Werte sind mittels Semikolon und Tabstopps getrennt.

Das untenstehende Beispiel 2.2 zeigt einen Ausschnitt aus den aktuellen Wetterdaten.

```
STATIONS_ID;MESS_DATUM;QN_3;  FX;  FM;QN_4; RSK;RSKF; SDK;SHK_TAG;  NM; VPM;  PM; TMK; UPM; TXK; TNK; TGK;eor
      1975;20181222;    1;  15.4;   3.2;    1;   7.6;   6;    0.000;   0;   7.4;   8.8; 1006.01;    6.4;   92.00;    7.4;   5.5;    4.2;eor
      1975;20181223;    1;   9.4;   3.9;    1;   5.3;   6;    0.000;   0;   7.8;   8.9; 1012.50;    6.7;   90.79;    7.5;   4.9;    4.6;eor
      1975;20181224;    1;   8.4;   3.8;    1;   0.0;   6;    3.083;   0;   6.0;   6.4; 1023.92;    2.7;   85.17;    5.0;   0.6;   -0.8;eor
      1975;20181225;    1;  12.5;   5.5;    1;   0.0;   6;    0.000;   0;   7.1;   7.3; 1027.21;    5.9;   78.25;    7.2;   4.3;    2.9;eor
      1975;20181226;    1;  11.1;   4.9;    1;   0.1;   6;    0.000;   0;   8.0;   9.6; 1025.25;    6.9;   95.33;    8.6;   5.1;    4.8;eor
      1975;20181227;    1;   9.8;   4.2;    1;   0.1;   6;    0.000;   0;   8.0;   8.9; 1022.23;    6.0;   95.17;    6.8;   5.3;    5.2;eor
```

Abbildung 2.2: Ausschnitt aus den Wetter Rohdaten.

Die Attributsnamen bestehen zum Großteil aus Abkürzungen, die auf der Website und in anderen Dateien auf dem FTP Server zusammen gesucht werden müssen. Nach ausgiebiger Recherche hat sich folgende Legende ergeben:

Abkürzung	**Erklärung**	**Maßeinheit**
STATIONS_ID	Stationsidentifikationsnummer	Nummer
MESS_DATUM	Das Datum der Messung	yyyymmdd
QN_3	Qualitätsniveau der nachfolgenden Spalten	Code
FX	Tagesmaximum der Windspitze	m/s
FM	Tagesmittel der Windgeschwindigkeit	m/s
QN_4	Qualitätsniveau der nachfolgenden Spalten	Code
RSK	tägl. Niederschlagshöhe	mm
RSKF	tägl. Niederschlagsform	numerischer Code (0,1,4,6,7,8,9)
SDK	tägl. Sonnenscheindauer	h
SHK_TAG	Tageswert Schneehöhe	cm
NM	Tagesmittel des Bedeckungsgrades	1/8
VPM	Tagesmittel des Dampfdruckes	hPa
PM	Tagesmittel des Luftdrucks	hPa
TMK	Tagesmittel der Temperatur	$°C$
UPM	Tagesmittel der Relativen Feuchte	%
TXK	Tagesmaximum der Lufttemperatur in 2m Höhe	$°C$
TNK	Tagesminimum der Lufttemperatur in 2m Höhe	$°C$
TGK	Minimum der Lufttemperatur am Erdboden in 5cm Höhe	$°C$
eor	End data Record, beendet die Zeile	-

Tabelle 2.6: Legende der Wetterdaten des DWD

Anhand der Legende kann ausgewertet werden, welche Daten potentiell interessant sind und welche keinen Mehrwert bieten. Danach wurden sie zusammen mit allen Daten, die außerhalb des Messzeitraums liegen, entfernt. Dadurch wurden die Attribute $STATIONS\_ID$, QN_3, QN_4 und *eor* entfernt. Alle anderen Werte wurden im Datenbestand belassen, da sie keine Performanceprobleme bereiten aber ggf. Mehrwerte liefern könnten. Neben der Erklärung der einzelnen Abkürzungen wurden die Informationen bereit gestellt, dass der Wert -999 in den Daten ein Fehlerwert ist, der vor allem in den ältesten Beständen vorkommt. Des Weiteren wird in diesen Daten ein Tag von sechs Uhr morgens bis sechs Uhr morgens definiert.

2.4.3 Datenqualität

In diesem Abschnitt soll die Datenqualität hinsichtlich der Aspekte, die für die Verarbeitung im Rahmen des KDD relevant sind, betrachtet werden. Die sowohl im Abschnitt 2.3.2 wie auch 2.3.3 heraus gearbeiteten Probleme sollen hier in den Daten identifiziert werden. Darüber hinaus wurden die Daten auf folgende Attribute von Datenqualität untersucht: Sind sie in einem geeigneten Maße verständlich, nützlich, gültig, glaubwürdig, exakt, aktuell und volatil?

Die Verständlichkeit der Daten variiert, wird jedoch durch die Vorverarbeitung so stark erhöht, dass sie kein Problem mehr darstellt, in dem Attribute sprechend benannt werden. Die Nützlichkeit der Daten kann an dieser Stelle noch nicht bestimmt werden, anhand der vorliegenden Daten wird jedoch von einer grundsätzlichen Nützlichkeit ausgegangen. Die Gültigkeit der Daten wird mit dem Modell in Abschnitt 2.4.4 sichergestellt. Die Glaubwürdigkeit der Daten wird durch die Exaktheit der Daten und die Aussage der Sensorenhersteller angenommen. Die Exaktheit wird in den folgenden Abschnitten 'Widersprüchliche Daten' und 'Falsche Daten' analysiert. Die Aktualität ist gegeben, da die Daten eigens für diese Arbeit erhoben wurden. Die Volatilität ist dadurch gegeben, dass der Datensatz über den recht langen Erfassungszeitraum keinen extremen Schwankungen unterlegen ist. Im Folgenden wird auf weitere Aspekte der Datenqualität eingegangen.

Entitätenidentifikationsproblem
In den einzelnen Datensätzen ist das Datumsattribut jeweils anders benannt, bei Fitbit *dateTime*, bei Medisana *Datum - Uhrzeit* und in den Wetter Daten *MESS_DATUM*. Darüber hinaus gibt es keine Überschneidungen in den Datensätzen, die unterschiedlich benannt sind.

Einheitlichkeit der Zeitschritte und Redundanzen
Die Zeitschritte innerhalb der Daten sind unterschiedlich und schwanken in gewissen Datensätzen. Die meisten Daten aus dem Fitbit-Datensatz liegen auf einer täglichen Basis vor, mit ein paar Ausnahmen in denen Schritte, Puls und Schlaf minütlich oder auch stündlich vorliegen. Die Wetterdaten sind einheitlich täglich erhoben. Bei Medisana liegen Daten in veränderlichen Zeitschritten vor, da sie in unterschiedlicher Granularität erhoben wurden und aufgrund der verwendeten Technik teilweise verdoppelte Datensätze beinhalten, die einen leicht abweichenden Zeitstempel aufweisen. Darüber hinaus weisen alle Datensätze außer dem Wetter fehlende Daten auf, sodass sich die Zeitschritte unterscheiden. Während bei den Gewichtsdaten im schlimmsten Fall innerhalb eines Monats anstatt 30 Messungen nur 6 stattfinden, treten in den Blutdruckdaten an einem Tag statt einem, teilweise 6 Messwerte auf. Die Minutendaten von Fitbit weisen selten für jede Minute des Tages einen Messwert auf, somit fehlen oft einige Minuten.

Widersprüchliche Daten
Die befürchteten gleichbenannten Attribute, die etwas Unterschiedliches beinhalten, tauchen in den hier verwendeten Datensätzen nicht auf. Es wurden verschiedene Pulswerte erhoben. Diese werden beim Blutdruckmessgerät *Puls* und beim Fitbit *Ruhepuls* genannt und bezeichnen verschiedene Messwerte.

Datenwertkonflikt Zeitstempel
Alle Datensätze arbeiten mit anderen Zeitstempeln, bei Fitbit wird das Datum unter dem Attributsnamen *dateTime* und dem Format yyyy-MM-dd angegeben. Bei Minutendaten kommt noch eine Uhrzeitangabe als eigenes Attribut unter dem Namen *time* mit dem Format hh:mm:ss hinzu. Medisana nutzt in beiden Datensätzen ein Attribut für Datum und Uhrzeit zusammen *Datum - Uhrzeit* im Format dd.MM.yyyy - hh:ss. Bei den Daten des Deutschen Wetterdienstes wird das Datumsformat yyyyMMdd unter dem Attributsnamen *MESS_DATUM* verwendet. Daraus ergibt sich, dass jeder Datensatz neben einem anderen Attributsnamen auch ein ganz anderes Format nutzt. Dies muss vereinheitlicht werden, damit die Daten gemeinsam verarbeitbar sind.

Datenwertkonflikt Maßeinheiten
Die Datensätze gehen neben dem Datum auch unterschiedlich mit den Maßeinheiten um, sowohl die Daten des Fitbits und des DWD liefern Maßeinheiten nur in Metadaten. Medisana hingegen handhabt dies anders und in den jeweiligen Datensätzen unterschiedlich. Bei den Gewichtsdaten werden die Maßeinheiten im Attributsnamen mit angegeben, bei den Blutdruckdaten steht bei Systole und Diastole die Maßeinheit im Wert, bei dem mit gemessenen Puls jedoch nicht, siehe Tabelle 2.4.

Fehlende Daten
Alle Körperdaten sind von fehlenden Daten betroffen. Dabei variieren die Gründe und Ausmaße. Bei den Fitbit Daten gibt es immer dann Lücken in den Daten, wenn der Sensor den Kontakt zur Haut verliert, dies betrifft hauptsächlich die Pulsdaten. Dieser Kontaktverlust kann verschiedene Gründe haben, Schweiß, Bewegung, schlechter Sitz oder Verschmutzung. Darüber hinaus werden keine Daten erfasst, wann immer der Sensor abgenommen wird. Dies muss er zum einen regelmäßig zum Laden und zum Duschen aber auch, wenn der Nutzer mit Wasser oder zu großer Hitze in Kontakt kommt, zum Beispiel beim Schwimmen, Tauchen, Schnorcheln, in der Sauna oder bei einem Sonnenbad in großer Hitze. Dazu kommt, dass es vorkommen kann, dass der Nutzer vergisst, den Sensor zu tragen. Im Messzeitraum gibt es auch eine große Lücke von etwas mehr als zwei Wochen, da vergessen wurde, den Sensor vor dem Schwimmen abzulegen und er dadurch beschädigt wurde. Dadurch konnten keine Daten erfasst werden, bis der Sensor ersetzt wurde. Bei fehlenden Daten nutzt Fitbit keinen Default-Fehlerwert, sondern lässt diese Datenpunkte aus. Dies betrifft Tagesdaten, genauso wie die Minutendaten die nicht erfasst werden konnten. Somit haben die wenigsten Tage wirklich 1440 Messungen für die Minutendaten.

Die Daten von Medisana sind ebenfalls von fehlenden Daten betroffen, die aber größtenteils darauf zurückgehen, dass der Nutzer nicht gemessen hat. Seltener sind technische Probleme, wie Synchronisation, Speicher oder schlechtes Akkumanagement. Es ist dem Blutdruckmessgerät leider erst anzusehen, dass der Akku schwach ist, wenn das Gerät nicht mehr misst. Dadurch fehlt in großen Zeitabständen hin und wieder ein Tageswert. Fehlende Daten werden auch bei Medisana nicht explizit gekennzeichnet, sondern zeichnen sich dadurch aus, dass es den Zeitstempel in den Daten nicht gibt. Besonders zu Beginn des Erfassungszeitraums gibt es leider lange Zeitabschnitte, in denen kaum Messwerte aufgenommen wurden. So gibt es für den September 2015 nur neun Gewichtsmessungen, für den Oktober sogar nur drei. Diese dünne Datenlage bessert sich erst Mitte 2016 und liegt an der stationären Natur der Waage. Im Gegensatz zum Blutdruckmessgerät lässt sie sich nicht ohne weiteres in einer Tasche transportieren.

Die untenstehende Tabelle zeigt das Ausmaß der Datenlücken. Die fehlenden Daten des Fitbits sind dabei relativ gering. Bei den Gewichts- und Blutdruck-Daten ist dies deutlich schwerwiegender. Das Problem, das zum Fehlen solch massiver Mengen an Datenpunkten geführt hat, wurde während des Erhebungszeitraums behoben, sodass die größten Lücken am Anfang des Erhebungszeitraumes sind, während im weiteren Verlauf die Daten sehr regelmäßig sind. Bei den Wetterdaten wird von keinen Fehlern gesprochen, obwohl bei zwei Datensätzen Fehler auftraten. die Daten fehlen jedoch nicht komplett, wie es bei Mediasana oder Fitbit der Fall wäre, sondern haben bei zwei Datensätzen in eher unwichtigen Attributen einen Platzhalter. Dies ist so verschwindend gering, dass im Text zumeist von keinen fehlenden Daten gesprochen wird.

Werte	**Optimale Anzahl Datenpunkte**	**Fehlende Anzahl Datenpunkte**	**Fehlende Daten in %**
Fitbit generell	1330	52	4%
Puls Minuten Daten	1872000	174854	9%
Blutdruck	1330	314	24%
Gewicht	1330	444	34%
Wetter	1330	2	0,2%

Tabelle 2.7: Fehlende Daten Übersicht.

Falsche Daten (Exaktheit)

Falsche Daten können aus unterschiedlichen Gründen in die Daten gelangen. Angefangen von ungenauen, fehlerhaften oder ungenügenden Sensoren über falsche Benutzung der Messgeräte bis hin zu Messfehlern oder Ungenauigkeiten durch die Datenübertragung.

Die Messgenauigkeit der Sensoren ist nicht mit medizinischen Geräten zu vergleichen und wird durch die Tätigkeiten und Verwendung beeinflusst. Das heißt, es ist zu jeder Zeit davon auszugehen, dass sich der reale Wert vom Messwert unterscheidet. Der Hersteller gibt nicht an, mit welcher Messungenauigkeit man rechnen muss, daher wurde eine Arbeit herangezogen, in der die Genauigkeit eines Fitbit Sensors geprüft wurde Benedetto u. a. (2018). Es handelt sich nicht um das gleiche Modell des Fitbits wie jenes, das in dieser Arbeit genutzt wurde, jedoch kann davon ausgegangen werden, dass die Genauigkeit sich in ähnlichen Größen bewegt. Dabei wurde eine Messungenauigkeit von ca. 6 bpm gegenüber eines Elektrokardiographen (EKG) festgestellt.

Medisana gibt in der Gebrauchsanleitung des Gerätes die Messabweichungen an. Für den Blutdruck gilt eine maximale Messabweichung des statistischen Drucks von ± 3 mmHg und für den Puls ± 5% des Wertes.

Neben der Messungenauigkeit kann es immer zu fehlerhaften oder ungenauen Daten kommen, wenn die Geräte falsch genutzt werden, nicht richtig angewendet oder die Anwendungshinweise nicht beachtet werden. Zum Beispiel verschlechtert reden, falsche Körperhaltung oder eine zu kurze Ruhephase vor dem Messen das Ergebnis der Blutdruck-Messung. Bei Fitbit können schnelle Handbewegungen durch Gewichtheben, Klatschen, Seilspringen etc. dazu führen, dass Schritte gezählt werden, die nicht gemacht wurden oder der Pulssensor den Kontakt verliert. Darüber hinaus werden bei manchen Aktivitäten Schritte fälschlicherweise gezählt, zum Beispiel Kochen, Autofahren oder Radfahren oder gar zu starkes Gestikulieren. Die Verfälschung dadurch hält sich jedoch in Grenzen, sollte aber bedacht werden.

Um grobe Messfehler und unwahrscheinliche Daten als solche zu erkennen, wurde ein Datenmodell erstellt, das im nächsten Abschnitt beschrieben wird. Mithilfe dieses Modells kann erkannt werden, ob ein Datum valide ist und den Ansprüchen des Modells entspricht oder nicht. Nach Anwendung dieses Modells auf die Daten wurde festgestellt, dass es keine invaliden Daten gibt. Alle Daten bewegen sich in einem plausiblen Rahmen.

2.4.4 Datenmodell

In diesem Abschnitt soll darauf eingegangen werden, wie das Datenmodell aussieht, das für die Daten, die dieser Arbeit zu Grunde liegen, genutzt wurde. Das Modell wird dabei nicht in seiner Gänze, sondern nur exemplarisch für einige Werte aufgeführt. Dabei ist zu beachten, dass die Grenzen und Werte aus unterschiedlichen Gründen gewählt wurden. Teilweise sind sie auf Werte der WHO zurückzuführen, teilweise auf das Vermögen der Sensoren und ebenso auf die Erfahrung mit dem Nutzer. Dieses Modell gilt also in gewisser Hinsicht nur für einen Nutzer und auch nur für eine begrenzte Zeit, da sich die Werte mit dem Alter und der Gesundheit verändern

können. Rückwirkend ist es aber sinnvoll, ein möglichst enges Modell zu entwickeln und fehlerhafte Daten auch dann zu erkennen, wenn sie für andere Nutzer vielleicht valide wären. Es gibt unterschiedliche Möglichkeiten, Datenmodelle zu erstellen, die die Validität prüfen. Hier wird mit harten Ober- und Untergrenzen gearbeitet, die, wenn möglich, mit einer validen Steigung unterstützt werden.

Puls

Erklärung:	Herzschlag vom Fitbit
Einheit:	bpm
Erhebungshäufigkeit:	Minütlich
Valide Untergrenze:	30
Valide Obergrenze:	230
Valide Steigung:	/
Missing Data:	Daten fehlen
Missing Data Bedeutung:	Das Fitbit wurde nicht getragen oder hat den Kontakt verloren

Die Angaben des Datenmodells beziehen sich hauptsächlich auf die Grenzwerte und mögliche Steigungen. Darüber hinaus beschreibt es, wie fehlende Daten (Missing Data) erkannt werden und welche Bedeutung sie haben können.

Die Werte beziehen sich hauptsächlich auf Werte, die für eine gesunde Person im Alter zwischen 20 und 30 möglich sind. Dadurch ergeben sich die oberen und die unteren Werte, die Steigung ist nicht angegeben, da sie nicht sinnvoll angegeben werden konnte. Das Herz kann innerhalb einer Minute bei hoher Belastung sehr viel schneller schlagen.

Ruhepuls

Erklärung:	Ruhepuls berechnet während des Nachtschlafs
Einheit:	bpm
Erhebungshäufigkeit:	Täglich
Valide Untergrenze:	30
Valide Obergrenze:	80
Valide Steigung:	<= 10
Missing Data:	Daten fehlen
Missing Data Bedeutung:	Das Fitbit wurde nicht getragen oder hat den Kontakt verloren

Gewicht

Erklärung:	Gewicht gemessen morgens auf der Medisana Waage
Einheit:	kg
Erhebungshäufigkeit:	Täglich
Valide Untergrenze:	40
Valide Obergrenze:	200
Valide Steigung:	< 6
Missing Data:	Daten fehlen
Missing Data Bedeutung:	Es wurde nicht gemessen

Blutdruck

Erklärung:	Systole und Diastole gemessen morgens mit einem Handgelenkmessgerät
Einheit:	mmHg
Erhebungshäufigkeit:	Täglich
Valide Untergrenze Systole:	40
Valide Untergrenze Diastole:	20
Valide Obergrenze Systole:	160
Valide Obergrenze Diastole:	110
Valide Steigung:	< 30
Missing Data:	Daten fehlen
Missing Data Bedeutung:	Es wurde nicht gemessen

Besonders der Blutdruck ist stark an den Nutzer angepasst, da sich zeigt, dass jener einen sehr niedrigen Blutdruck hat, sodass Werte, die laut WHO noch im Bereich der erreichbaren Werte liegen, hier schon ein Indiz dafür sind, dass falsch gemessen wurde.

Diese Anpassung an den Nutzer könnte noch weitaus mehr betrieben und die Grenzwerte immer dichter an die wahrscheinlichen Werte gezogen werden. Dies wurde im Rahmen dieser Arbeit jedoch nicht derart weit getrieben, da die Möglichkeit großer Ausreißer bestehen können sollte und im Rahmen dieser Arbeit kein Mehrwert daraus gezogen werden konnte.

2.4.5 Feature Selection

Feature Selection ist eine Methode des maschinellen Lernens, in der nur ein Teil der vorhandenen Features für einen Lernalgorithmus verwendet wird. Features sind dabei

neben den Werten selbst auch errechnete Werte aus den Grunddaten. Diese Art, den Datenbestand zu erweitern, soll auch hier genutzt werden. Dabei müssen die letztendlich verwendeten Features stark eingeschränkt werden, um den Umfang der Arbeit nicht zu sprengen. Die Möglichkeiten sind jedoch mannigfaltig und reichen von einfachen statistischen Features wie Mittelwerten bis hin zu komplexeren wie der Anzahl der Pulswerte eines Tages, die eine gewisse Grenze überschreiten. Ebenso zählt ein errechneter Wert, der bestimmt wie schnell der Puls nach einer Anstrengung wieder sinkt und somit die körperliche Fitness des Nutzers wiedergibt, zu den komplexen Features. In diesem Abschnitt soll auf die Features eingegangen werden, die im Umfang dieser Arbeit umsetzbar erscheinen.

Einfache statistische Werte
Es sollen statistische Werte berechnet werden, darunter Minimum, Maximum, Mittelwert bzw. Median und Standard-Abweichung. Dies soll für den gesamten Datensatz geschehen aber auch für die einzelnen Jahre. Dabei soll ein Gefühl für den Datensatz vermitteln werden um aufzuzeigen, in welchen Wertebereichen sich die einzelnen Attribute bewegen. Zudem sollen diese Werte in Analysen bzw. Visualisierungen fließen.

Werte aufteilen
Zur einfacheren Verarbeitung bietet es sich an manche Attribute, zum Beispiel das Datum, aufzuteilen, sodass neben dem einzelnen Zeitstempel auch der Wochentag, der Monat und das Jahr zur Verfügung stehen. Dies ermöglicht Analysen, die die Wochentage betreffen oder Monate bzw., Jahre gegenüberstellen.

Komplexe Werte
Neben der schon erwähnten Berechnung des Erholungspulses, kommen noch weitere Werte in Frage, die den Datensatz erweitern können. Zum Beispiel umfassen die Schlafdaten von Fitbit nicht nur den Nachtschlaf, sondern auch jeden anderen Schlaf, wie zum Beispiel Mittagsschlaf. Somit lässt sich daraus die Information ziehen, ob es einen Schlaf neben dem Nachtschlaf gab, um dies in den Datensatz aufzunehmen.

2.5 Fazit

In diesem Kapitel wurde das Umfeld von Quantified Self, intelligenten Spiegeln und dem Erkenntnisgewinn aus Daten(KDD) sowie die Datenqualität der vorliegenden Daten betrachtet. Dafür wurde ein Modell beschrieben, das zur Validitätsprüfung dient. Die Datenqualität wird als gut genug eingestuft, um sie mit in dieser Arbeit aufbringbarem Aufwand verwendbar zu machen.

Das in diesem Abschnitt beschriebene Datenmodell könnte später anhand der Analysen überarbeitet werden, um die Grenzen, anhand der Erkenntnisse über die Daten

und ihre Normbereiche in Bezug auf den Nutzer. enger zu ziehen. Dadurch könnten Ausreißer und Messfehler schneller und an den Nutzer angepasster gefunden werden.

Daten sollten zudem nicht isoliert betrachtet werden, es zeigte sich, dass der Kontext der Daten enorm wichtig ist. Jede Anomalie in den Daten sollte immer im Kontext mit weiteren Datenpunkten im Zusammenhang mit den hier bestehenden Möglichkeiten wie Wetter-, Orts-oder Kalenderdaten betrachtet werden. Dies muss bei der Auswertung mit einbezogen werden, denn der Umstand, in dem sich der Nutzer befindet, kann durchaus auch die Daten verändern.

Zusammenfassend werden aus den Analysen der einzelnen Bereiche alle Anforderungen an ein Quantified Self System mit Spiegeloberfläche zusammengetragen.

2.5.1 Anforderungen an das System

Neben den Anforderungen, die bereits im Abschnitt zum Spiegel 2.2.3 herausgearbeitet wurden, kommen weitere Anforderungen an das System hinzu, die sich aus der Analyse der unterschiedlichen Gebiete ergeben haben.

Das System muss so aufgebaut sein, dass ein vertrauenswürdiger Umgang mit den empfindlichen Daten gepflegt wird und für den Nutzer auch so einsehbar ist. Der Nutzer sollte dabei allzeit volle Kontrolle über seine Daten und das System haben. Er soll weder bevormundet noch übergangen werden. Das System als solches ist dazu da, dem Nutzer Erkenntnisse zu ermöglichen und muss somit auf ihn angepasst und individualisiert werden.

Um möglichst gute und weitgreifende Erkenntnisse zu ermöglichen, sowie individualisierbar zu sein, muss das System so gebaut werden, dass die Datenquellen erweiterbar und anpassbar sind. Darüber hinaus sollte das System nicht nur für einen einzigen Nutzer reagieren, sondern für alle Personen des Haushaltes. Es müssen also mehrere Nutzer unterstützt werden.

Um die Qualität der Nutzererfahrung zu erhöhen ist es sinnvoll, die Oberfläche individualisierbar zu gestalten, sodass der Nutzer selbst entscheiden kann, welche Daten er auf welche Art und Weise als erstes sieht. Darüber hinaus muss darauf geachtet werden, dass die Oberflächen eingängig und einfach zu bedienen sind. Die Daten sollten aus verschiedenen Quellen gezogen werden, um eine möglichst breite Datenbasis zu beziehen, dazu sollte die Anzahl der Quellen erweiterbar sein, um das System anpassbar an zukünftige Möglichkeiten und die Bedürfnisse des Nutzers zu gestalten. Dies unterstützt zudem das zufällige Entdecken von neuem Wissen, also das Serendipity-Prinzip. Darüber hinaus muss sorgsam ausgewählt werden, wie mit negativem Feedback umgegangen werden soll. Es soll dem Nutzer nicht vorenthalten, aber gleichsam auch nicht zu seinem Nachteil im falschen Moment ungefiltert angezeigt werden. Mit

negativen Daten und Feedback muss ein vorsichtiger Umgang gepflegt werden. Dabei sollten Systeme den Nutzer niemals bevormunden und ihm jederzeit die vollkommene Macht über das System und die Daten geben.

3 Entwurf und Implementierung

In diesem Kapitel wird darauf eingegangen, wie die Planung eines Systems zur Erweiterung der Selbstwahrnehmung aussehen könnte und wie es im Rahmen dieser Arbeit umgesetzt wurde. Dafür wurde das System, das bereits im vorangegangenen Bachelor Lüdemann (2016a) begonnen wurde, erweitert und modifiziert.

Darüber hinaus wurden Teile des Systems prototypisch durchgeführt, implementiert oder aufgebaut, um die generelle Umsetzbarkeit des Systems zu gewährleisten. Dabei war es nicht Ziel, ein funktionstüchtiges System zu implementieren, sondern die Teile des Systems zu untersuchen, bei denen Probleme auftreten könnten oder als besonders interessant für die Fragestellung eingestuft wurden. Im Folgenden wird die Architektur beschrieben und darauf eingegangen, welchen Stand der prototypische Aufbau des Spiegels hat. Darüber hinaus wird anhand des KDD Prozesses durch jedes Teil des Systems geführt, um zu beschreiben, welchen Stand es hat, wie es umgesetzt wurde und welche Schwierigkeiten sowie Erkenntnisse dabei aufkamen. Die Ergebnisse der Auswertungen und Erkenntnisse die Daten selbst betreffend werden zum Großteil im folgenden Kapitel 4 diskutiert. Die Komponenten werden jeweils in zwei Schritten beschrieben, zunächst wird auf die Planung der Komponente und ihrer Funktion eingegangen und weiterhin darauf, inwieweit die Komponente prototypisch umgesetzt wurde. Als letztes wird beschrieben, wie der technische Aufbau für das hier beschriebene System des Spiegels aussieht.

3.1 Architektur

In diesem Abschnitt wird auf die Anforderungen an die Architektur, grundlegende Architekturentscheidungen und wie sich dies am Ende in die Planung eingebracht hat, eingegangen. Darüber hinaus wird ein Augenmerk darauf gelegt, wie sich die in Lüdemann (2016a) beschriebene Architektur verändert hat.

Im Rahmen des Spiegelprojektes wurden einige Arbeiten verfasst, auf die hier Bezug genommen wird, da sie einige Bereiche bereits im Detail beschreiben. Diese Arbeit baut auf diesen Arbeiten auf.

3.1.1 Anforderungen an die Architektur

Aus dem Abschnitt 2.5.1 und der vorangegangenen Arbeit Lüdemann (2017a) ergeben sich an den Spiegel eine Reihe von Anforderungen mit Auswirkungen auf die Architektur. Dazu gehört, dass der Spiegel eine große Anzahl Sensoren und Datenquellen unterstützen soll, die sich auch in Zukunft noch erweitern lassen. Das bedeutet, die Architektur muss offen, erweiterbar und pluginfähig sein. Damit soll gewährleistet werden, dass auch in Zukunft weitere Module für neue Funktionalitäten, Anwendungen, Ausgaben oder Datenquellen angebunden werden können. Ebenso müssen bereits bestehende Module angepasst werden können, da sich der Bedarf oder die Umstände verändern können, zum Beispiel, weil sich die Quell-APIs oder die Datenstrukturen verändern. Ebenso können sich die Anforderungen an die Visualisierungen und Auswertungen ändern.

Um schnelle Einsichten in die Daten oder manuelle Eingaben zu ermöglichen ist es notwendig, das System nicht nur stationär auf dem Spiegel anzubieten. Es muss möglich sein, mit einer App oder Website zumindest Teile des Systems zu erreichen und zu bedienen. Dies erfordert eine strikte Trennung von Anzeige und Logik, sodass das Ausgabegerät wechseln kann.

Sofern mit dem Spiegelbild über eine Kamera gearbeitet wird, ist es wichtig, auf die Latenz zu achten, damit der Nutzer kein verzögertes oder stockendes Spiegelbild seines Selbst sieht. Zudem sollte aus Sicherheitsgründen der Livestream der Kamera nicht abgespeichert werden. Darüber hinaus muss genau modelliert werden, wann Daten von den APIs der Sensorhersteller abgefragt werden, um die Latenz der Datenanzeige möglichst gering zu halten. Um gleichsam nicht unnötig großen Datenverkehr zu erzeugen und Ressourcen zu verbrauchen muss darauf geachtet werden, in welchen Frequenzen Daten erhoben werden. Daten, die einmal am Tag erfasst werden, müssen weit seltener abgefragt werden als Daten, die minütlich erhoben werden. Es gilt, einen Weg zu finden, der mit möglichst wenig Abfragen eine möglichst schnelle Darstellung der benötigten Daten ermöglicht.

Das System sollte nicht für nur einen einzigen Nutzer ausgerichtet sein, sondern mit Daten und Individualisierungen verschiedener Nutzer umgehen können. Dafür muss die Möglichkeit bestehen, mehrere Nutzerkonten anzulegen, deren Daten separat gespeichert werden. Das System muss die unterschiedlichen Nutzer erkennen können.

Da bei Langzeitnutzung von mehreren Nutzern mit unbestimmt vielen Datenquellen durchaus große Datenmengen entstehen können, sollte das System verteilt geplant werden, um die Speicherung und Verarbeitung großer Datenmengen zu gewährleisten.

Die Anforderungen an Sicherheit und Verfügbarkeit des Systems sind besonders wichtig, da die verwendeten Daten sehr persönlich und heikel sind. Dadurch ergibt sich, dass die Daten dupliziert gehalten werden sollten, um Verluste zu vermeiden. Darüber

hinaus sollte eine sehr eingeschränkte Verbindung zum Internet gehalten und sehr strikte Maßregelungen des Datenflusses und der Rechte innerhalb des Systems geführt werden. Es sollte ebenso darauf geachtet werden, die Daten möglichst lokal zu speichern.

3.1.2 Grundlagen

Aus den Anforderungen ergibt sich, dass die Architektur verteilt geplant werden muss. Um die Architektur einordnen zu können, wurde im Rahmen der vorangegangenen Arbeit Lüdemann (2018) das Modell von Tanenbaum und van Steen (2008) herangezogen. Die untenstehende Abbildung 3.1 zeigt dies. In diesem Modell geht es um die vertikale Verteilung eines Systems und in welcher Schicht (Layer) der Schnitt zwischen dem Klienten(Client) und dem Server gemacht wird. Je nach Anwendung bietet sich eine andere Aufteilung an, bei der Architektur des Spiegelssystems ist dies der Thin Client (a). Dieser versammelt möglichst viel Logik auf dem Server und lässt den Klienten nur die Interaktion mit dem Nutzer und die Anzeige ausführen. Dies bringt eine Reihe von Vorteilen: die komplette Berechnung und Analyse der Daten, sowie das Speichern und Verarbeiten kann von leistungsfähigen Servern übernommen werden. Dadurch wird die Anzeige weniger ressourcenbelastend und kann auf weniger gut ausgestatteten Geräten wie Tablets oder Smartphones stattfinden. Im Thin Client werden sogar Teile der Anzeige vom Server übernommen, sodass einige Visualisierungen vom Server vorbereitet werden und nicht auf dem Endgerät. Darüber hinaus können so die Daten zentral gehalten und gesichert werden. An den Klienten wird nur das nötigste übermittelt und kein direkter Datenbankzugriff erlaubt. Des Weiteren sind Änderungen leichter, da das Herzstück des Systems auf dem Server liegt.

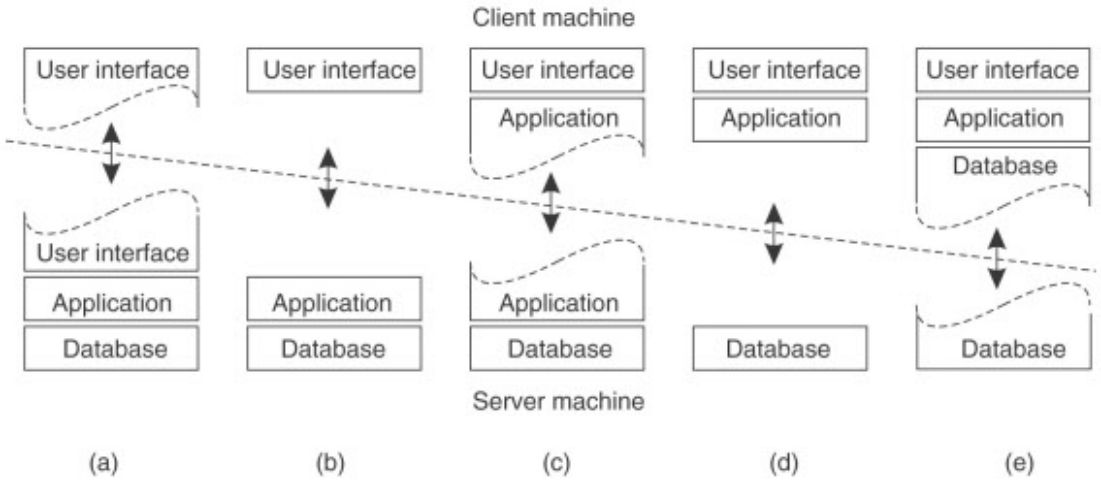

Abbildung 3.1: Drei-Ebenen-Architektur der vertikalen Verteilung entnommen aus Tanenbaum und van Steen (2008)

Diese Einteilung in die Schichten User Interface (Anzeige), Application (Logik) und Database (Persistenz) zieht sich durch die Architektur. Zwischen diesen Schichten muss kommuniziert werden, dies kann auf verschiedenste Arten geschehen. Um den Anforderungen gerecht zu werden, wird eine Kommunikation benötigt, die die Komponenten möglichst lose verbindet und somit auch eine Erweiterung der Komponenten und des Systems ermöglicht. Dafür bot sich die Middleware des Labors an, da sie eine publish-subscribe Anwendung ist, die ohne Queuing auskommt und dadurch eine Kommunikation in Echtzeit ermöglicht Eichler (2014). Die einzelnen Komponenten können als Agenten einer Gruppe beitreten, um dort Nachrichten zu lesen oder zu schreiben, dadurch ergibt sich eine lose und problemlos erweiterbare Kommunikation zwischen den Schichten des Systems. Die untenstehende Abbildung 3.2 verdeutlicht die Schichten des Systems und die Kommunikation zwischen ihnen.

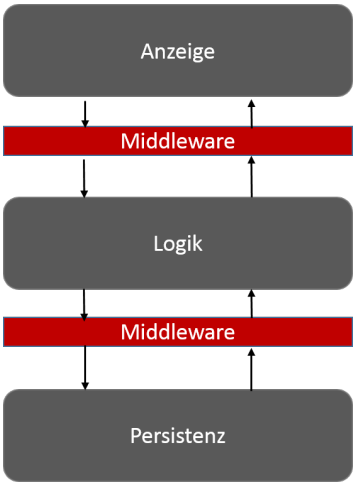

Abbildung 3.2: Kommunikation der Ebenen

3.1.3 Komponenten

In diesem Abschnitt wird darauf eingegangen, wie die Planung der Komponenten des Systems aussieht. Dafür wird in Abbildung 3.3 das Komponentendiagramm aufgezeigt, in dem die einzelnen Komponenten des Systems, deren Datenfluss und die externen Komponenten des Systems dargestellt sind. Bei den Komponenten wurde darauf geachtet, dass sie eine hohe inhärente Kohäsion aufweisen, sowie, dass die Kopplung zwischen den Komponenten möglichst gering bleibt. Es gibt zwei Schnittstellen nach außen, zum einen die Datenquellenkomponente, die mit den APIs der jeweiligen Quellen kommuniziert und zum anderen eine Schnittstelle zu Expertensystemen, um das für die Auswertung der Daten benötigte Expertenwissen heranzuziehen. Die Datenquellenkomponente ist hier verdeutlicht als Datenakkumulator dargestellt und reicht die Daten an die Datenbereinigungskomponente weiter, die dafür sorgt, dass die Daten bereinigt werden. Der Nutzer kommuniziert mit dem System über die Kommunikationskomponente, die verschiedene Nutzer-Schnittstellen anbieten soll. Sie soll derart erweiterbar sein, dass es je nach Anzeigemedium unterschiedliche Wege der Kommunikation geben kann. Die Anzeigekomponente verwaltet die unterschiedlichen Möglichkeiten der Anzeige und ermöglicht dem System, die Daten und Visualisierungen anzuzeigen. Die Visualisierungen selbst werden von der Visualisierungskomponente erstellt und verwaltet. Diese bezieht ihre Daten zum einen aus der Verarbeitungs- und Analysekomponente, die die größten Berechnungen und Auswertungen durchführt. Zum anderen aus der Datenbankkomponente, die die Kommunikation zwischen dem System und der Datenbank übernimmt. Die Transformationskomponente transformiert die Daten aus der Datenbank in eine für die Analyse- und Verarbeitungskomponente nützliche Form. Die Interpretationskomponente nutzt die aus dem System gewonnenen Berechnungen,

die Ergebnisse der Analysen und das externe Expertenwissen, um Erkenntnisse über die Daten zu entwickeln.

Die Komponenten der Architektur haben sich seit dem Stand der Bachelorarbeit Lüdemann (2016a) verändert. Dazu gehört, dass das Konzept der Device API Komponente deutlich erweitert wurde und so neben den Webschnittstellen der Wearables, Quellen modular genutzt werden sollen. Dazu gehören Wetter, Kalender sowie manuelle Daten. Die Komponente soll so aufgebaut sein, dass sie nachträglich jederzeit um Quellen erweitert werden kann, daher wurde sie in Datenquellenkomponente umbenannt. Darüber hinaus ist eine Anzeigekomponente dazugekommen, die sich darum kümmert, was und wie auf welchen Geräten angezeigt wird. Das untenstehende Bild verdeutlicht den aktuellen Planungsstand der Architektur. Das System ist im Großteil prototypisch umgesetzt, um die generelle Umsetzbarkeit zu untersuchen. Die Verbindungen in der Abbildung zwischen den Komponenten verdeutlichen den Datenfluss.

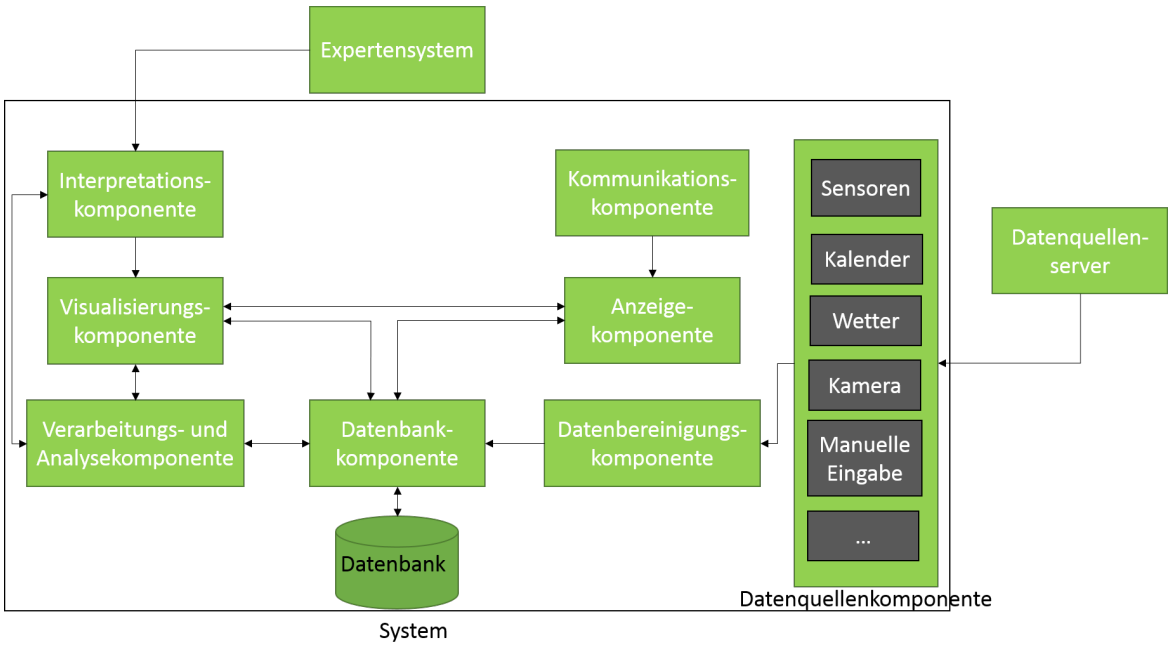

Abbildung 3.3: Die Komponentensicht des Spiegel Systems

Im weiteren Verlauf dieses Kapitels wird im Rahmen des KDD Prozesses genau auf die einzelnen Komponenten eingegangen und beschrieben, welchen Stand der Umsetzung sie haben.

3.1.4 Toolset

In diesem Abschnitt soll zusammenfassend darauf eingegangen werden, welches Toolset im Rahmen des Systems genutzt wurde und wofür.

Im Rahmen dieser Arbeit wurden für unterschiedliche Aufgaben diverse Programme genutzt und getestet. Für die Vorverarbeitung und Bereinigung der Daten wurde zum Großteil selbstgeschriebener Code verwendet. Darüber hinaus aber auch Microsoft Excel, da die Bearbeitung einiger Datentypen mit diesem Programm besser vorgenommen werden kann. Ebenso wurde ein Textbearbeitungsprogramm verwendet, um die Daten während des Prozesses zu überprüfen und an den Zeilenabständen und Umbrüchen zu arbeiten, dafür wurde Notepad ++ genutzt.

Zur Persistierung wurde in erster Linie das Filesystem des genutzten Computers verwendet und darüber hinaus eine InfluxDB, die in Docker aufgesetzt wurde. Diese Datenbank eignet sich sehr gut, um Zeitreihendaten zu persistieren und an Programme weiterzuleiten, die auf die Verarbeitung von Zeitreihendaten spezialisiert sind. Darunter Grafana, eine Oberfläche um Daten aus der Influx DB anzuzeigen und in diversen Visualisierungen zu sichten. Für weitere Visualisierungen wurde ebenso Excel verwendet.

Außerdem wurde KNIME genutzt, ein Programm für interaktive Datenanalyse, ähnlich dem Programm Rapidminer. Damit wurden die Daten sowohl verarbeitet und bereinigt wie auch analysiert und in Visualisierungen angezeigt. Dabei ist die Verwendung von KNIME angenehm einfach, da diverse Module aneinandergereiht werden können, die die Daten verarbeiten. Durch die Möglichkeit, eigene Module mit Code einzufügen, können auch Funktionalitäten eng an die eigenen Daten und Fragestellungen angepasst wie auch vermeintlich fehlende Module ersetzt werden.

3.2 Datenselektion und -integration

Im ersten Schritt des KDD Prozesses, der technisch umgesetzt wird, werden die Daten ausgewählt und erhoben. Die Auswahl der Daten wurde im Abschnitt 2.4 ausführlich beschrieben, sodass hier nur kurz darauf eingegangen wird. Die Datenerfassung ist im Großteil eine Weiterführung der Erfassung aus Lüdemann (2016a). Das heißt, dass ein Fitbit für Schlaf, Puls und Aktivitätsdaten, ein Blutdruckmessgerät und eine Analysewaage verwendet wurde. Darüber hinaus wurden Kontextdaten erfasst, die zur manuellen Kontextualisierung herangezogen werden, ohne in die maschinelle Auswertung einzufließen. Dies betrifft vor allem Kalenderdaten, also das Wissen darum, was an einem Tag geschehen ist. Dabei gibt der Kalender nur einen Hinweis, da er nicht mit wissenschaftlicher Präzision geführt wurde. Im Verlauf der Auswertung wurde an

die Daten der Ort, an dem sich der Nutzer primär aufhielt, annotiert. Darüber hinaus wurden die Wetterdaten herangezogen und flossen in die Auswertung ein.

Es muss neben der Auswahl der Daten und ihrer Quellen entschieden werden, wie häufig Daten erhoben werden. So wurden zu Beginn der Erhebung Blutdruckdaten zweimal am Tag gemessen, morgens und abends. Nach einer kurzen Testphase wurde die Erhebungsfrequenz auf einmal am Tag reduziert. Dadurch ergibt sich, dass die Blutdruck- und Gewichtsdaten einmal am Tag und die Fitbitdaten laufend erhoben werden. Die Wetterdaten sind ebenfalls tägliche Werte.

Der Aufbau und die Syntax der Rohdaten ist im Abschnitt 2.4.1 und 2.4.2 beschrieben. Weitere Daten sind konzeptionell angedacht, wurden jedoch nicht erhoben. Trotz der bestehenden Implementierung der API Schnittstelle wurden die Daten nicht über die Zeit regelmäßig, sondern im Rahmen dieser Arbeit manuell als Ganzes heruntergeladen. Dabei wurden zum Teil die Implementierungen genutzt aber nicht ausschließlich. Der 8. Februar 2019 wurde als Datenstopp gewählt, sodass Daten vom 20. Juli 2015 bis zum 8. Februar 2019 vorliegen. Das ergibt 1300 Tage Daten, die heruntergeladen und zusammengeführt wurden.

Datenquellen Komponente

Hier wird beschrieben, wie die Komponente geplant ist, die die Datenquellen akkumuliert und verarbeitet, sowie welche Teile davon implementiert wurden.

Entwurf
Die Datenquellenkomponente soll die einzelnen Quellen anbinden und die Daten aus ihnen ziehen sowie an die Datenbereinigungskomponente weitergeben. Dabei soll sie modular aufgebaut sein, um später einfacher Quellen hinzuzufügen, zu bearbeiten oder entfernen zu können. Die Abfrage der Daten geschieht in regelmäßigen Abständen, die möglichst an die Art und Erhebungshäufigkeit der Daten angepasst sind. Dabei muss eine Balance zwischen den Anforderungen, Daten möglichst schnell zu beziehen und möglichst geringen Datenverkehr zu erzeugen, gewahrt werden. Außerdem ist zu beachten, dass die meisten APIs Zugriffsbeschränkungen haben, die einschränken, wie viele Abfragen pro Stunde möglich sind. Ist die Abfragehäufigkeit höher als diese Beschränkung, erhält man für diese Stunde keine Daten mehr. Es muss also vermieden werden, zu viele Anfragen zu stellen, nur um die Daten möglichst zeitnah zu beziehen.

Die Datenquellenkomponente steht somit im Kontakt mit dem Internet und den APIs der Hersteller. Dafür muss sie mit der Technologie der Webschnittstellen arbeiten können. Die meisten APIs sind durch OAuth2 abgesichert. Dafür muss sich die Komponente beim Server authentifizieren, um dort die Daten des Benutzers abrufen zu

dürfen. Das heißt, jeder Benutzer benötigt einen Account bei der Datenquelle, der berechtigt ist, Daten über die API abzurufen. Die jeweiligen Token, die aus den Credentials erzeugt werden, müssen dann im System hinterlegt werden, damit dieses sich authentifizieren kann. Mit den Token kann die Komponente das OAuth2-Protokoll des Servers durchlaufen, sich authentifizieren und die Daten abrufen. Die abgerufenen Daten werden dann als Dateien im Filesystem gespeichert, wo sie von der Datenbereinigungskomponente aufgerufen werden können.

Implementierung
Die Implementierung der Datenquellenkomponente ist im Großteil aus Lüdemann (2016a) übernommen und hat sich seither nicht verändert.

Sowohl die Fitbit- wie auch die Medisana API arbeiten mit dem OAuth2-Protokoll. Für die Wetterdaten wurde im Rahmen dieser Arbeit keine Implementierung genutzt, sondern die Daten manuell vom bereitgestellten FTP-Server gezogen.

3.3 Datenvorverarbeitung und -bereinigung

Die nächste Phase des KDD Prozesses nimmt die erhobenen Daten und bereinigt sie von fehlerhaften, überflüssigen oder falsch formatierten Daten. Im Abschnitt 2.3.3 und 2.3.4 wurde ausführlich darauf eingegangen, welche generellen Probleme Daten mitbringen können, die direkt aus den Hersteller APIs kommen. Im Abschnitt 2.4.3 wurde expliziter beschrieben, welche Probleme bei den hier verwendeten Daten vorliegen. Im Rahmen der Datenvorverarbeitung und -bereinigung müssen diese Probleme behoben werden. Dies geschah mit einer Reihe unterschiedlicher Werkzeuge und in verschiedenen Schritten. Da der KDD Prozess ein iterativer ist, wurde mehrfach zu dieser Phase zurückgekehrt um die Bereinigung, mit Wissen, das sich aus der Data Mining Phase ergeben hat, zu justieren.

Datenbereinigungskomponente

Die Datenbereinigungskomponente sorgt dafür, dass die Daten aus den APIs so bereinigt werden, dass sie in das System integriert werden können. Dabei geht es in erster Linie darum, die Daten zu bereinigen, Dopplungen und überflüssige Datenpunkte zu entfernen und fehlende Daten zu erzeugen bzw. zu interpolieren und als solche kenntlich zu machen. Dadurch wird gewährleistet, dass es für jedes Datum des Messzeitraums Daten gibt. Dabei wird noch nicht darauf eingegangen, wie die unterschiedlichen Analyseverfahren die Daten benötigen. Die Transformation der gesäuberten Rohdaten geschieht in der Transformationskomponente 3.4.

Entwurf
Die Daten, die aus den Schnittstellen abgerufen werden, müssen bereinigt werden, um Inkonsistenzen, fehlende Daten, schlechte Formatierung und Datenfehler zu beheben sowie die Daten in ein einheitliches Format zu überführen. Darüber hinaus müssen doppelte Daten aussortiert werden. Dies soll in der Datenbereinigungskomponente geschehen. Die Daten werden aus der API übergeben, verarbeitet, bereinigt und dann an die Persistenz weitergeleitet. Da die Bereinigung von Quelle zu Quelle variiert, muss diese Komponente für jede weitere Quelle angepasst werden und somit möglichst erweiterbar konstruiert sein.

Implementierung
Die Datenbereinigungskomponente ist in großen Teilen in Java implementiert, die einzelnen Teile müssen jedoch manuell angestoßen werden, da es noch keine Automatisierung gibt. Einige Teile der Bereinigung wurden komplett manuell durchgeführt. Die Datenbereinigungskomponente, die bereits in Lüdemann (2016a) erstellt wurde, wurde weiter verwendet und erweitert.

In dem schon bestehenden Teil werden die JSON Daten aus Fitbit und Medisana verarbeitet und von unnötigen Metadaten und Parametern befreit. Nachdem die Fitbitdaten konvertiert und bereinigt wurden, werden fehlende Daten ersetzt. Diese sind daran zu erkennen, dass der Zeitstempel nicht existiert, also kein Eintrag vorgenommen wurde. Entweder fehlt dadurch der ganze Tag mitsamt der Datei oder im Datenstrom fehlen die entsprechenden Einträge. Bei den Minutendaten werden die Datensätze mit den fehlenden Minuten aufgefüllt. Dafür wird der Attributwert auf 0 gesetzt, um in weiteren Schritten fehlende Daten zu erkennen und verarbeiten zu können. Bei den Tagesdaten werden fehlende Tage erzeugt und die Werte ebenfalls ausgenullt. Bei Medisana wird das gleiche mit fehlenden Daten getan.

Bei den Medisanadaten muss das Datumsformat angepasst werden, da es zusammen mit der Uhrzeit gespeichert wird. Das heißt, zuerst muss die Uhrzeit extrahiert und als einzelner Parameter gespeichert und dann das Datum auf das festgelegte gemeinsame Datumsformat konvertiert werden. Darüber hinaus muss aus den Medisanadaten die Maßeinheit herausgezogen werden, da sie bei den Blutdruckdaten mit im Attributswert steht. Im Gegensatz zu den Fitbitdaten weisen die Gewichts- und Blutdruckdaten von Medisana Dopplungen und vor allem überflüssige Datenpunkte auf. Nach eingehender Sichtung der Daten konnte bisher kein Muster erkannt werden, nach dem die Daten automatisch gelöscht werden können, sodass bisher die Dopplungen und überflüssigen Daten manuell entfernt werden müssen, um sicherzugehen, die richtigen Daten zu behalten.

Die Wetterdaten werden ebenfalls von unnötigen Parametern bereinigt sowie die Kürzel der Attributsnamen in sprechende Namen umbenannt. Die Daten werden, nachdem sie bereinigt und konvertiert wurden, durch Semikolon voneinander getrennt in nach Datensatz separierten Dateien abgelegt.

Datenbankkomponente

Nachdem die Daten bereinigt wurden, werden sie an die Persistenz weitergegeben und gespeichert. Darüber hinaus bietet die Datenbank die Möglichkeit, die Daten von verschiedenen Punkten des Systems abzurufen und anzureichern. Das heißt, neben den bereinigten Rohdaten liegen auch erzeugte Mehrwerte und berechnete Werte in der Datenbank.

Entwurf
Die Datenbankkomponente kümmert sich um die Kommunikation zwischen der Logik und der Persistenz. Dabei ist es sinnvoll, mehrere Datenbanken zu nutzen. Zum einen sollte eine dokumentenorientierte Datenbank genutzt werden, um die bearbeiteten Rohdaten unabhängig persistieren zu können, um bei Speicher- oder Berechnungsfehlern die Daten in ihrer unveränderten Form wieder herstellen zu können. Zum anderen sollte eine Datenbank genutzt werden, die die Zugriffe des Systems schnell bearbeiten und im Zweifel mit Änderungen im Datenschema umgehen kann, falls sich die Datenquellen ändern. Darüber hinaus können Datenbanken genutzt werden, die für spezielle Analysen von Vorteil sind oder die Ergebnisse der Analysen speichern. Die Datenbankkomponente sollte so gestaltet sein, dass bei Bedarf neue Datenbanken hinzukommen können und sie sollte Datenbanken unterstützen, die eine verteilte Speicherung der Daten ermöglichen.

Implementierung
Im Rahmen dieser Arbeit wurden die Daten in ihrer bearbeiteten Form auf dem Filesystem des Computers gespeichert, auf dem die Datenbearbeitung und Analyse stattfindet. Die Datenbankkomponente wurde im groben entworfen und geplant, jedoch nur zu einem kleinen Teil umgesetzt.

Dieser Teil ist eine analysespezifische Datenbank, die besonders beim Auswerten von Zeitreihendaten verwendet wird. Die Daten mussten für die Influx DB in ein spezielles Format transformiert werden, das im Abschnitt 3.4 näher beschrieben wird. Die Datenbank selbst wurde auf dem System installiert, auf dem auch die anderen Auswertungen und Datenbearbeitungen durchgeführt wurden, um die Nützlichkeit und Umsetzbarkeit der Datenbank und der dazugehörenden Datenverarbeitung und Visualisierung zu testen. Das Aufsetzten der Datenbank wurde mithilfe von Docker durchgeführt, um schnelles Aufsetzen und Testen zu ermöglichen. Die transformierten Daten wurden in die Datenbank mithilfe der Influx DB internen Sprache eingelesen. Dabei war darauf zu achten, nicht zu viele Daten auf einmal einzulesen, da die Schnittstelle einen recht eng gesetzten Timeout vorweist, der den Import abbricht, sollte eine Datei länger als fünf Sekunden benötigen um importiert zu werden. Die Datenbank wird also im ersten Schritt nur manuell genutzt, um ihren Nutzen zu evaluieren.

3.4 Datentransformation

In der Phase der Datentransformation werden die Daten von ihrem bereinigten Zustand in einen für die jeweiligen Analysen benötigten transformiert. Dabei hängt die Transformation sowohl von den eigentlichen Daten, wie von den anzuwendenden Analysen, Verfahren und Algorithmen ab, da diese sich stark darin unterscheiden können, welche Formate der Daten sie benötigen. Ebenfalls ist der Umfang, den die Daten haben dürfen oder müssen, variabel. Neben einer Umformatierung kann es somit nötig sein, Daten zu akkumulieren oder Ausschnitte zu wählen. Darüber hinaus werden die Daten im Rahmen des Kontextes bereinigt. Anders als in der Bereinigungskomponente, in der die Daten aufgrund ihrer technischen Fehler bereinigt werden, bedeutet dies, in der Transformationskomponente fehlende Daten zu interpolieren, zu diskretisieren oder die Frequenz der Daten anzupassen.

Transformationskomponente

In diesem Abschnitt wird darauf eingegangen, wie die Transformationskomponente geplant wurde, sowie wie und in welchem Umfang sie bereits umgesetzt wurde.

Entwurf
Die Transformationskomponente bezieht ihre Daten aus der Datenbankkomponente und formatiert sie je nach Anfrage passend für die Analyse. Sowohl die Daten als auch die verwendeten Analysen und Werkzeuge können sich verändern, dies zieht eine notwendige Veränderung der Transformationskomponente nach sich. Daher sollte sie möglichst erweiter- und veränderbar geplant sein. Darüber hinaus muss die Transformationskomponente mit verschiedensten Daten umgehen können, sie sowohl umformatieren wie auch die Dimensionen der Daten einschränken können.

Die Daten müssen gegebenenfalls in verschiedene Formate umformatiert werden, dies bedeutet einfache Änderungen wie die Reihenfolge und Benennung der Attribute wie auch Änderungen der Trennzeichen (Semikolon, Komma, Tabstopp, etc.) oder Umformatierung des Datums bis hin zu komplexeren Umformatierungen, wie das Ändern der Datentypen wie zum Beispiel, von numerischen Werten zu ordinalen. Neben der Umformatierung ist es oft notwendig, fehlende Daten zu interpolieren oder den Datenraum zu extrapolieren. Ebenso kann es vorkommen, dass bei zu vielen vorliegenden Datenpunkten die Werte aggregiert werden müssen.

Implementierung
Die Transformationskomponente wurde in Teilen umgesetzt, ist als komplette Komponente jedoch nicht implementiert worden. Unterschiedliche Aufgaben der Komponente wurden durch verschiedene Werkzeuge umgesetzt.

Im Abschnitt über die Datenbanken wurde die Zeitreihendatenbank InfluxDB erwähnt, für die die Daten in ein eigenes Format gebracht werden mussten. Die Influx DB benötigt neben einem speziellen Datenschema auch angepasste Daten, beziehungsweise Formate. Der untenstehende Ausschnitt 3.1 zeigt das Datenformat, das benötigt wird, um Daten in die InfluxDB zu importieren. Jeder Datensatz bekommt ein Label, in diesem Beispiel *bloodpresure*. Dieses Label fungiert als Tabellenname, darüber hinaus werden die Daten einem User zugeordnet. Das System muss mit mehreren Usern umgehen können, da es zurzeit nur Daten für einen Nutzer gibt, wird immer *user=01* gesetzt. Danach folgen die eigentlichen Daten, die Attribute Systole, Diastole und Puls mit ihren jeweiligen Ausprägungen und als letztes der Zeitstempel, der für die Influx DB im Sekunden-Format vorliegen muss. Diese Formatierung wurde in Java implementiert.

```
2 bloodpresure,user=01 systole=114,diastole=73,puls=71 1433252340
3 bloodpresure,user=01 systole=113,diastole=70,puls=112 1433262780
4 bloodpresure,user=01 systole=112,diastole=70,puls=98 1433276820
```

Listing 3.1: Beispiel der Blutdruckdaten Formatierung Influx DB

Neben der Influx DB wurde KNIME für weitere Analysen verwendet. Um die Daten in KNIME einlesen zu können, mussten sie ebenfalls in ein spezielles Format gebracht werden. Dieses benötigt ebenfalls ein spezielles Datumsformat und zwar yyyy-MM-dd sowie eine Datenseparierung durch Semikolons in Textfiles. Es müssen diverse Arten Daten in KNIME importiert werden, im Rahmen dieser Arbeit wurde das Einlesen aus Textfiles gewählt. Das dafür nötige Format wird in der untenstehenden Abbildung 3.2 aufgezeigt. Hier stehen die Daten mit dem Attributsnamen als Header in einer durch Semikolon separierten Formatierung. Das Datum leitet den jeweiligen Datensatz ein, welcher die einzelnen Attribute beinhaltet. Diese Umformatierung der Daten wurde ebenfalls in Java implementiert.

```
5 Datum;Uhrzeit;Systole;Diastole;Puls
6 2015-07-20;10:42;101;61;62
7 2015-07-21;08:40;96;59;72
8 2015-07-22;08:14;106;62;71
```

Listing 3.2: Beispiel der Blutdruckdaten Formatierung KNIME

Daten mit Stellen hinter dem Komma, wie zum Beispiel das Gewicht, müssen so gespeichert werden, dass das Komma durch einen Punkt ersetzt wird, da KNIME mit dem englischen Format arbeitet. Um erweiterte Analysen zu ermöglichen, wurde das Datum sowohl in einem einheitlichen Format gespeichert wie auch nach den einzelnen Komponenten aufgeteilt, damit Wochentage, Monate und Jahre gegenüber gestellt werden können. Dieses Aufteilen des Datums geschieht allerdings erst in KNIME und nicht schon in der Vorformatierung. Um mit den Daten arbeiten zu können, wurden

die einzelnen Datenarten in getrennten Dokumenten in KNIME eingelesen und über das Datum zusammengefügt. Dieser Schritt zeigt das untenstehende Bild 3.4 als einen Ausschnitt aus KNIME.

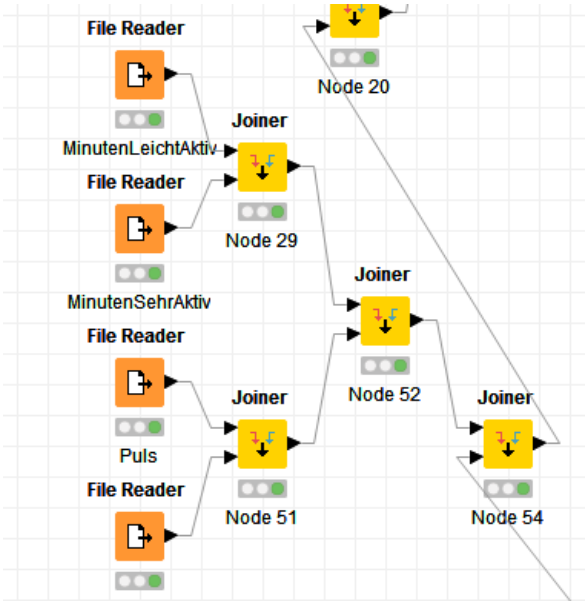

Abbildung 3.4: Ausschnitt aus dem Zusammenführen der Daten in KNIME

Die File Reader Nodes lesen den jeweiligen Datensatz ein, in diesem Beispiel die *Minuten Leicht Aktiv*, die *Minuten Sehr Aktiv*, den *Puls* und die Wetterdaten. Sie alle haben als erstes Attribut, wie in 3.2 gezeigt, das Datum. Über dieses Datum führen die Join Nodes die Daten zusammen. Am Ende entsteht eine große Tabelle, in der alle Daten dem jeweiligen Datum zugeordnet werden.

Die Transformation des Datums zu seinen einzelnen Komponenten zeigt die untenstehende Abbildung 3.5, die ebenfalls einen Ausschnitt aus KNIME zeigt. Dort ist zu sehen, wie der finale Join, der eine Tabelle erzeugt in der alle Daten vorhanden sind, in die Transformation führt. Zuerst werden die Spalten für die jeweiligen Werte erzeugt, Wochentag, Monat und Jahr. Danach wird das Datum, das im Datensatz als String gespeichert ist, zu einem internen Datumsformat konvertiert, um im nächsten Schritt aus dem Datum jeweils den Wochentag oder Monat zu ziehen und zu speichern. Das Extrahieren des Jahres ist auf diesem Bild nicht zu sehen, da es aufgrund der Übersichtlichkeit verschoben wurde.

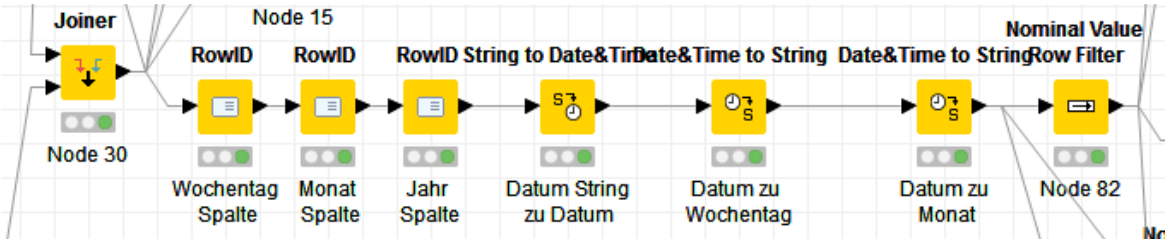

Abbildung 3.5: Ausschnitt aus der Aufsplittung des Datums in KNIME

In der Bereinigung der Daten wurden fehlende Daten mit Nullen aufgefüllt. Für manche Darstellungen und Auswertungen ist dies unpraktisch, also wurden die ausgenullten Daten mithilfe der Missing Value Node in KNIME interpoliert. Die untenstehende Abbildung 3.6 zeigt diesen Teil des Workflows in KNIME. Die betroffenen Daten sind Gewichts- und Blutdruckdaten. Auch bei den Fitbitdaten gibt es fehlende Daten, diese treten jedoch in viel kleinerer Menge auf, sodass sie weniger Probleme in der Verarbeitung darstellen. Jedoch können diese in einem weiteren Schritt interpoliert werden, um einen vollständigeren Datensatz zu erzeugen.

KNIME bietet diverse Möglichkeiten, mit fehlenden Daten umzugehen. Die Missing Value Node ermöglicht ein einfaches Ersetzen des Wertes durch verschiedene Werte. Zum Beispiel einen festen vom Nutzer festgelegten Wert, einem Minimum oder Maximum des eigenen Wertebereiches oder dem vorangegangenen, nachfolgenden oder dem am häufigsten vorkommenden Wert. Darüber hinaus kann die Node einen Wert berechnen, entweder durch lineare beziehungsweise durchschnittliche Interpolation(linear Interpolation, average Interpolation) oder einen gleitenden Durchschnitt (moving Average). Welches dieser Mittel gewählt werden sollte hängt stark davon ab, wie die Werte weiter verarbeitet werden und die innere Struktur aussieht. Im Rahmen dieser Arbeit wurden die Möglichkeiten getestet, bis sich eine funktionale Konstellation ergeben hat. Dabei war darauf zu achten, ob die Methode sinnvolle Werte ergab, die ohne große Sprünge zu den vorhergehenden und nachfolgenden Werten stehen sowie sich auch ändern. Oft gab es das Problem, dass Abschnitte mit Konstanten gefüllt wurden, die teilweise im Kontrast zu den existierenden Werten darum herum standen, obwohl Verfahren für variable Werte genutzt wurden.

Bei den Blutdruckdaten wurde für die Systole und Diastole die lineare Interpolation gewählt, für den Puls ein Moving Average mit einem Lookahead von 40 nach vorne und 40 zurück. Diese Zahl musste so hoch gewählt werden, da es abschnittsweise fast ganze Monate lang keine Daten gibt. Die Uhrzeit, die zu jedem Messwert gehört, wurde mit einem fixen Wert gefüllt, der so gewählt ist, dass er leicht erkennbar ist und sonst im Datensatz nicht vorkommt, da zu dieser Uhrzeit kein Blutdruck gemessen wird. Er steht konstant auf 2:22 Uhr, dadurch sind die ersetzten Datensätze erkennbar.

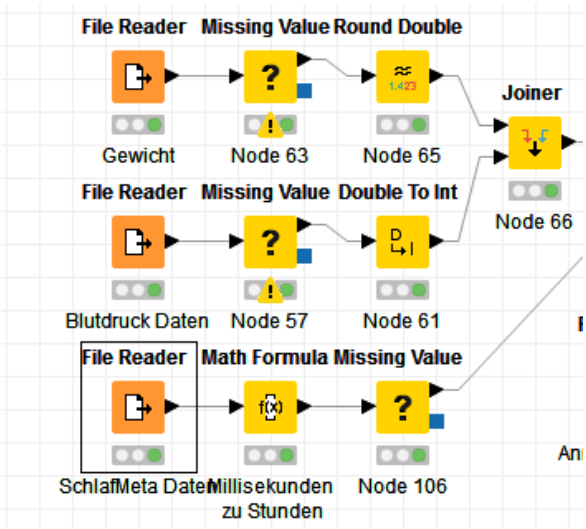

Abbildung 3.6: Ausschnitt aus der Behandlung fehlender Daten in KNIME

Bei den Gewichtsdaten wurde aus denselben Gründen für die Uhrzeit der gleiche Wert genommen wie für den Blutdruck. Das Körpergewicht wurde linear interpoliert und alle weiteren Werte außer der BMI, also Knochenmasse, Körperfett, Körperwasser und Muskelmasse, wurden mittels des Moving Average mit einem Lookahead von 40 ersetzt. Bei dem BMI wurde der Next Value genommen. Dies ist im Prinzip ungenau, da der BMI ein aus den Körperdaten errechneter Wert ist und er dadurch inkonsistent wird. Da er aber nicht weiter verwendet wird, fällt dies nicht ins Gewicht. Sollte er später verwendet werden, sollte er für die fehlenden Daten nach den Regeln des BMI berechnet werden.

Um die Daten zu visualisieren und Erkenntnisse dadurch zu verdeutlichen und zu generieren, mussten die Daten geglättet werden, da die Rohdaten zu starken Schwankungen unterliegen, die bei der Visualisierung Probleme verursachen. Für die Glättung wurde ein einfaches Verfahren verwendet, es wird ein gewichteter gleitender Durchschnitt (WMA) der Zeitreihendaten gebildet. Dabei hat das Fenster der Betrachtung eine Größe von $t = 2$, es werden also zwei Werte vor und zwei Werte nach dem zu berechnenden Wert mit einbezogen. Die Wahl der Fenstergröße bestimmt, wie sehr der Wert an die Kurve angepasst wird. Bei der Wahl eines zu kleinen Fensters ist die Glättung minimal, bei der Wahl eines zu großen Fensters werden die Werte zu sehr an den Gesamt-Mittelwert angeglichen. Die Gewichtung kann dabei frei gewählt werden, um den zeitlichen Verlauf darzustellen. Dabei muss jedoch darauf geachtet werden, dass durch die Summe der Gewichtungen geteilt wird. Der WMA mit einem t von 2 bestimmt sich durch $WMA_x = \frac{w_1 \cdot x_{-t} + w_2 \cdot x_{-t-1} + x + w_3 \cdot x_{+t-1} + w_4 \cdot x_{+t}}{w_1 + w_2 + w_3 + w_4}$. Diese Formel wurde auf den jeweiligen Datenbestand angewandt, der darüber geglättet wird.

3.5 Data Mining

In der Data Mining Phase des KDD Prozesses werden die Daten nach der Bereinigung, Persistierung und Transformation analysiert. Dies kann auf ganz unterschiedliche Weise geschehen. Es können einfache bis komplexe mathematische Verfahren genutzt werden, diverse Analysen und Algorithmen oder maschinelles Lernen. Ebenso können manuelle Schritte dazu verwendet werden, die Daten besser zu verstehen und Ergebnisse zu ermöglichen. Die hier verwendeten Verfahren sind stark gemischt, da mithilfe von KNIME einige Algorithmen, ebenso wie einfache mathematische Berechnungen, verwendet wurden. Darüber hinaus wurden mithilfe von Visualisierungen einige Aspekte der Daten näher untersucht. Dies geschah im Rahmen einer explorativen Datenanalyse, da es keine konkrete Fragestellung gab, sondern es darum ging, möglichst viele Informationen aus den Daten zu bekommen, die für den Nutzer hilfreich oder nützlich sein können. Eine konkrete Fragestellung wäre an dieser Stelle zum Beispiel: Verändert sich die Schlafdauer nach erhöhtem Koffeinkonsum?

Aufgrund der Verflechtung von maschineller und manueller Auswertung wird neben der Verarbeitungs- und Analysekomponente hier auch die Visualisierungskomponente behandelt.

Verarbeitungs- und Analysekomponente

In diesem Abschnitt wird auf die Verarbeitungs- und Analysekomponente eingegangen. Dabei wird betrachtet, wie sie geplant und wie sie am Ende in welchen Teilen und wie umgesetzt wurde.

Entwurf
Die Verarbeitungs- und Analysekomponente wendet mathematische Methoden und Algorithmen zusammen mit diversen Auswertungsverfahren an, um aus den Daten einen Mehrwert für den Nutzer oder Zwischenergebnisse zu generieren. Dies können einfache mathematische Anwendungen sein wie Mittelwerte, Mediane, Steigungen oder Trends. Es können aber auch komplexere Werte berechnet werden. Darüber hinaus können Korrelationen, Veränderungen über Zeit und Auslöser für Datenveränderungen bestimmt werden. Außerdem können komplexere Analysen verwendet werden, wie Cluster Analysen oder die Klassifikation der Daten um Ähnlichkeiten festzustellen, sowie eine Assoziationsanalyse. Welche Verfahren angewendet werden können und sollten, ergibt sich dabei aus der Natur der Daten und den Fragestellungen, sowie dem Aufwand, der betrieben werden kann. Daraus ergibt sich, dass die Komponente so geplant ist, dass sie im Nachhinein erweiterbar ist, um sich wechselnden Daten oder Fragestellungen anzupassen. Sie nimmt die Daten aus der Datenbankkomponente sowie der Transformationskomponente und wendet diverse Analysen darauf an. Die Ergebnisse

werden an die Visualisierungskomponente weitergegeben, damit diese sie anschaulich visualisiert und dem Nutzer anzeigt, um ihm weitere Erkenntnisse zu ermöglichen.

Implementierung
Die Umsetzung der Komponente ist in Teilen prototypisch geschehen, um eine Einsicht zu ermöglichen, welche Wege gangbar und sinnvoll sind. Dafür wurden diverse Werkzeuge verwendet und ganz unterschiedliche Herangehensweisen gewählt, die im Folgenden beschrieben werden.

Statistische Auswertung
In erster Instanz wurden statistische Werte über die Daten erzeugt, um ein Gefühl für die Daten zu entwickeln und Grundlagen zu haben. Dafür wurde die Statistics Node von KNIME genutzt. Sie liefert eine Reihe statistischer Werte, die weiter genutzt werden können, darunter Minimum, Maximum, Mittelwert, Median, Standardabweichung, Varianz, Schiefe und ein Histogramm der Werteverteilung. Diese Node könnte auch die fehlenden Daten zählen, jedoch werden fehlende Daten in diesem Datensatz mit 0 angegeben und nicht dadurch, dass der Wert tatsächlich fehlt, da die meisten Nodes nicht damit umgehen können, dass Werte fehlen. Somit kann die Statistic Node in einem so späten Zeitpunkt des KNIME Ablaufs keine Aussagen über fehlende Daten treffen. Sie wurde teilweise an einzelnen Quellen im früheren Ablauf genutzt, um derartige Auswertungen zu erhalten.

Durch die Angabe des Wertebereichs (Minimum, Maximum) kann ein Gefühl dafür vermittelt werden, in welchem Bereich sich die jeweiligen Werte bewegen. Durch das Histogramm ist zu erkennen, wo sich gehäuft Werte finden und welche Ausprägungen eher Einzelfälle sind. Die untenstehende Abbildung zeigt ein Beispiel für ein solches Histogramm.

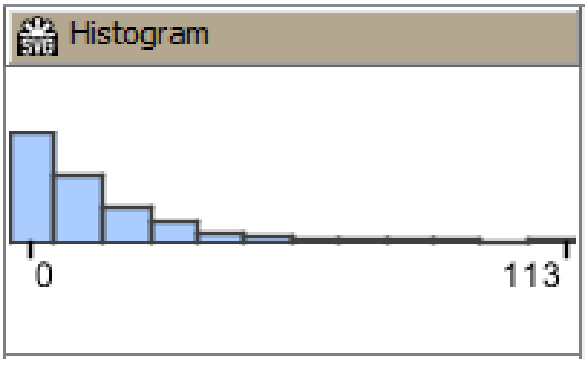

Abbildung 3.7: Ausschnitt dem Statistic Node in KNIME für Minuten Aktiv

Anhand dieses Histogramms ist zu erkennen, dass die Ausprägung der Daten zwischen 0 und 113 liegt. Dabei gibt es eine enorme Häufung bei 0 und nahen Werten. Dies

sind in diesem Fall keine fehlenden Daten, sondern einfach Tage, an denen kein Sport getrieben wurde. Zu sehen ist, dass eine geringe Anzahl an aktiver Minuten häufig vorkommt, aber es schnell weniger wird, je mehr Minuten es werden. Der aktivste erfasste Tag verzeichnet somit 113 Minuten Aktivität.

Sowohl einige der statistischen Werte wie auch diese Histogramme könnten dem interessierten Nutzer angezeigt werden, um bei ihm ein gesteigertes Verständnis der Daten zu fördern.

Cluster

Die Clusteranalyse wurde ebenfalls mit KNIME durchgeführt um zu testen, ob es möglich ist, Tage zu bestimmen, die einander ähneln. Also ob sich Cluster bilden, die an einem oder mehreren Attributen hängen und dadurch Gruppen von Tagen ergeben, die einander ähnlicher sind als andere. Die untenstehende Abbildung 3.8 zeigt den Aufbau der Nodes, die in KNIME verwendet wurden, um dies zu testen. Dabei wird eine Reihe von Nodes benötigt, die zum einen die Daten so vorbereiten, dass der genutzte k-Means Algorithmus mit ihnen arbeiten kann, zum anderen einige, um das Ergebnis klarer aufzeigen zu können. Die Daten werden durch die Normalizer Node normalisiert, um von den Algorithmen verarbeitet werden zu können, die normalisierte Daten benötigen. Diese werden dann dem Cluster Assigner und dem k-Means Algorithmus weitergegeben. Der k-Means Algorithmus gibt seine Ergebnisse ebenfalls an den Cluster Assigner weiter, der dann den Datensatz mit den jeweiligen Clustern erweitert. Dieses Ergebnis wird zum einen in einer Datei im Filesystem abgespeichert und zum anderen an Nodes weitergegeben, die es visualisieren. Dazu gehört die Color Manager Node, in der jedem Cluster eine eigene Farbe zugewiesen wird und dem Shape Manager, in dem einem Attribut, zum Beispiel dem Wochentag, dem Monat oder dem Jahr, eine eigene Form gegeben werden kann. Das sorgt dafür, dass am Ende in der Scatterplott Node mehr Dimensionen in der Visualisierung gesichtet werden können.

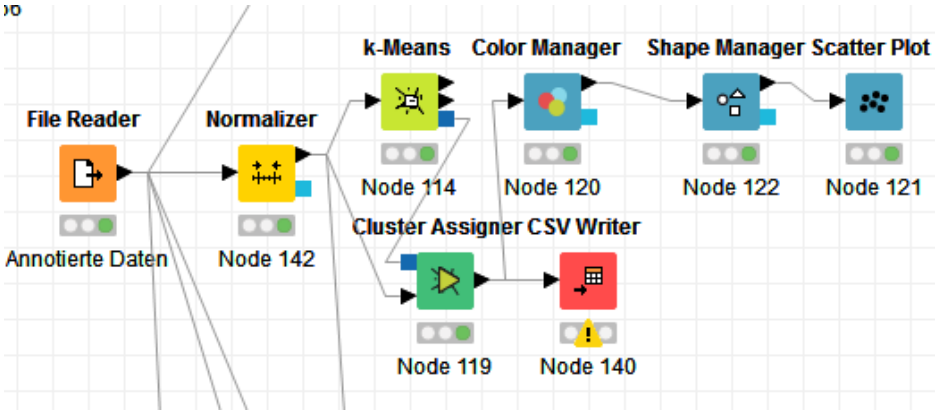

Abbildung 3.8: Cluster Algorithmus k-Means in KNIME

Der Aufwand, der in diese Analyse gesteckt wurde, musste leider in einem engen Rahmen gehalten werden, sodass nur eine begrenzte Anzahl an Testdurchläufen mit den Attributen durchgeführt werden konnte. Aufgrund der Beschaffenheit von KNIME war es hingegen recht einfach, weitere Cluster Algorithmen mit sehr ähnlichen Abläufen wie den oben gezeigten durchzuführen, sodass ebenfalls der FuzzyC-Means, k-Medoids und der DBSCAN Algorithmus genutzt wurden. Für einige dieser Algorithmen musste durch eine Numeric Distances Node eine Distanzfunktion für die Daten verwendet werden. Die Ergebnisse, die diese Analysen ergaben, werden im Abschnitt 4.2.2 besprochen. Die Algorithmen wurden mit diversen Modifikationen durchgeführt, um unterschiedliche Ergebnisse zu erhalten. Beim k-Means Algorithmus werden neben den zu bewertenden Attributen auch die Anzahl der angestrebten Cluster und die Anzahl der Iterationen des Algorithmus angegeben. Bei den Tests wurde immer eine möglichst hohe Zahl an Iterationen verwendet, also 99, die Zahl der Cluster wurde verändert und schwankte zwischen 2 und 7. Bei den meisten Versuchen wurden alle numerischen Attribute verwendet. Bei einigen wenigen wurden die Schritte oder andere einzelne Werte ausgelassen, aber zumeist wurden möglichst alle Attribute verwendet. Beim Fuzzy c-Means wurden ebenfalls die Anzahl der Cluster bei verschiedenen Durchläufen angepasst. Beim k-Medoids Algorithmus wurde der Partition Count (k) angepasst, von 2 bis 6. Beim DBSCAN wurde das Epsilon der Distanz angepasst, um zusammengehörende Punkte zu finden.

Assoziation
Die Assoziationsanalyse wird genutzt, um Zusammenhänge in Datensätzen zu untersuchen. Dabei arbeitet diese Analyse auf ordinären Daten, um häufiges gemeinsames Vorkommen zu untersuchen. In den hier vorliegenden Daten sind die meisten Werte jedoch nicht ordinär, sodass sie erst aufwändig umformatiert werden müssten. Nach einer ersten Einschätzung wurde dieser Aufwand jedoch als zu hoch für den eventuellen Nutzen dieser Analyse eingeschätzt. Sie könnte jedoch in folgenden Arbeiten hinzugenommen werden.

Klassifikation
Bei der Klassifikation geht es darum, die Daten in Klassen einzuteilen. Dies geschieht, indem Regeln beziehungsweise Kriterien für die Klassen entworfen werden und dann jeder Datensatz auf diese Regeln geprüft wird. Daraus können Entscheidungen getroffen werden. War es ein sportlicher Tag? War es ein Tag, an dem es dem Nutzer gut ging? War es ein Schummeltag, also ein Tag, an dem der Nutzer aus irgendwelchen Gründen seine eigenen Ziele ignoriert hat? Die Möglichkeiten der Klassifikation sind derart groß, dass sie ohne eine konkrete Fragestellung schwierig in Grenzen zu halten sind. Sollte also eine konkrete Fragestellung auftreten, wie zum Beispiel: Was begünstigt Schummeltage?, dann kann es durchaus sinnvoll sein, die Daten nach solchen Tagen zu klassifizieren und ggf. Auslöser zu suchen. Für die hier durchgeführte explorative Datenanalyse wurde der Aufwand als zu hoch eingeschätzt, der Nutzen als zu niedrig.

Zeitreihendatenbank
Die Zeitreihendatenbank Influx DB wurde, nachdem die Daten darin eingefügt waren, mit der Grafana Oberfläche verbunden, die verschiedene Möglichkeiten bietet, die Daten anzuzeigen und zu untersuchen. Mit dem hier gewählten Setup war jedoch kein großer Analysemehrwert zu erarbeiten, sodass die Grafana als Anzeigemodul durchaus interessant ist, sich für die Analyse aber auf andere Programme beschränkt wurde. Die Möglichkeiten der Anzeige in Grafana sind gut, aber da sich die Datenbasis schwer erweitern lässt und die Anzeigewerkzeuge durchaus eingeschränkt sind, wurde für die weitere explorative Datenanalyse Excel verwendet. Mit weiteren Programmen, die zusammen mit der Influx DB und Grafana verwendet werden können, könnte dieses Potential genutzt werden, jedoch wurde dies im Rahmen dieser Arbeit nicht weiter betrachtet.

Korrelationen
Die Betrachtung der Korrelation der einzelnen Werte wurde in KNIME umgesetzt. Dort gibt es eine Linear Korrelation Node, welche die lineare Korrelation der Attribute auswertet und anzeigt. Dabei wird für jedes Attributspaar der Korrelationskoeffizient gebildet, der zwischen -1 und 1 liegt. Dies ergibt bei vielen Attributen eine Matrix, die die Node anzeigt und farbig kennzeichnet. Diese Matrix ist in Abbildung 4.4 gezeigt. Durch die Betrachtung des Korrelationskoeffizienten können Auswirkungen eines Attributs auf ein anderes betrachtet werden. So könnten zum Beispiel Verbindungen zwischen dem Wetter und der Schrittzahl erkannt werden, wenn sie existieren, zum Beispiel in Form von: Steigt die Temperatur, steigt auch die Schrittzahl. Es ist jedoch nicht einfach, die Korrelation richtig zu interpretieren, da auch zufällige Korrelate auftreten können oder die Auswirkungen nicht direkt auftreten, sondern zeitversetzt. Es handelt sich dennoch um ein hilfreiches Mittel, einen Einblick in die Struktur der Daten zu erhalten.

Visualisierungskomponente

Die Visualisierungs Komponente gehört zum Data Mining Abschnitt, da durch die explorative Herangehensweise einige der Analysen auf dem Visualisieren der Daten fußen. Darüber hinaus wird beschrieben, welche Visualisierungen verwendet und für den Benutzer als nützlich eingestuft wurden. Des Weiteren ist die Visualisierung ein wichtiger Schritt, um dem Nutzer eine selbstständige Datenanalyse zu ermöglichen.

Entwurf
Die Visualisierungskomponente stellt die Daten aus der Transformations-, Datenbank- oder Analysekomponente auf Nachfrage des Nutzers in verschiedenen Arten bereit. Dazu gehören diverse Graphen, Diagramme oder auch Tabellen, seien es Datenausschnitte, komplette Übersichten oder berechnete Werte. Der Sinn dieser Komponente ist es, dem Nutzer eine möglichst umfassende und verständliche Ansicht seiner Daten

zu ermöglichen. Diese sollen angepasst und personalisiert werden können. Je nachdem, was der Nutzer gerne sehen möchte, sollen die Standardvisualisierungen, die der Nutzer als erstes zu sehen bekommt, angepasst werden können. Darüber hinaus ist es wichtig, dem Nutzer das Gefühl zu geben, volle Kontrolle und Einsicht in seine eigenen Daten zu haben, daher sollen durchaus auch große Tabellen mit Rohdaten anzeigbar sein. Um den möglichen Veränderungen der Ansprüche an die Visualisierungen entgegenkommen zu können, muss die Komponente gut erweiterbar gestaltet werden.

Implementierung
Für die Visualisierungskomponente wurden mithilfe verschiedenster Werkzeuge diverse Visualisierungen evaluiert, um ihre Nützlichkeit und Umsetzbarkeit zu testen. Dabei wurde neben der Grafana Oberfläche für die Zeitreihendatenbank InfluxDB, auch KNIME, sowie Microsoft Excel, genutzt. Durch die explorative Herangehensweise bei der Datenanalyse wurden die Visualisierungen nicht nur dazu genutzt, um die Daten verständlich und brauchbar anzuzeigen, sondern auch, um neue Erkenntnisse zu generieren. Dafür mussten die Daten teilweise transformiert oder erweitert werden. Dies wurde bereits in der Transformationskomponente beschrieben. Im Folgenden wird darauf eingegangen, welche Visualisierungen evaluiert wurden.

Parallele Koordinaten
KNIME bietet ein Node, das diese Art der Visualisierung ermöglicht. Dabei handelt es sich um eine interaktive Version für hochdimensionale Daten. Dabei werden die einzelnen Attribute als vertikale Achsen aufgetragen. Die Datensätze verbinden diese Achsen und formen dadurch Strukturen, die es ermöglichen, die Natur der Daten besser zu verstehen. Die untenstehende Abbildung 3.9 zeigt ein Beispiel dieser Visualisierung. Die Visualisierung selbst kann als Mittel genutzt werden, dem Nutzer ein eigenes Datenverständnis zu ermöglichen, da durch interaktives Erkunden die Struktur der Daten aufgezeigt und Zusammenhänge abgebildet werden können. Im untenstehenden Beispiel sind Körpergewicht, Knochenmasse, Muskelmasse und Diastole abgebildet. Sie zeigen, neben dem Wertebereich in dem sich die Daten bewegen, wie ein Datensatz von Achse zu Achse aussieht. Die weiß-beige gestrichelte Linie zeigt einen Datensatz, der bei einem sehr niedrigen Gewicht eine ebenso geringe Knochenmasse, dafür aber sehr hohe Muskelmasse, aufweist. Seine Diastole liegt dabei im Mittelfeld. Darüber hinaus ist zu sehen, wie sich die Daten auf den Wertebereich aufteilen. Während die Diastole sehr stark verteilt ist, zentriert sich die Knochenmasse auf zwei Werte. Anhand dieses Beispiels ist zu sehen, dass diese Art der Visualisierung, vor allem wenn sie Interaktiv ist, dem Nutzer ein Werkzeug liefert, ein Gefühl für die Natur der Daten zu entwickeln.

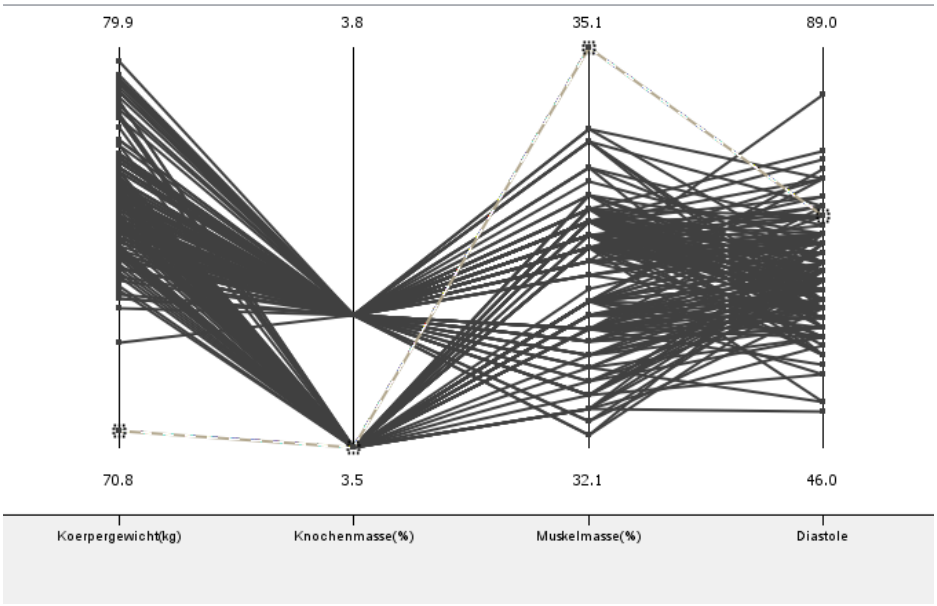

Abbildung 3.9: Beispielhafte Ansicht der parallelen Koordinaten in KNIME

Scatter Plot

Ein Scatter Plot bzw. Streudiagramm erzeugt Punktwolken mit zwei Dimensionen. Diese Visualisierung wurde in KNIME mit der Scatter Plot Node getestet. Durch sie kann die Abhängigkeitsstruktur der Werte zweier Attribute zueinander untersucht werden. Die Verwendung von Farben und Formen erlaubt das Betrachten von bis zu vier Dimensionen. Anhand der Verteilung der Datenpunkte können Rückschlüsse auf die Daten getroffen werden. Diese Rückschlüsse können die Abhängigkeiten, wie auch die Verteilung der Daten, betreffen. Ausreißer bei der Verteilung der Daten im Werteraum können Informationen liefern. Das untenstehende Bild 3.10 zeigt eine solche Abbildung als Beispiel. Dabei ist die Abhängigkeit der *Minuten Aktiv* zum *Ruhepuls* zu sehen, sowie die Regressionsgerade. Die untenstehende Abbildung zeigt neben der Verteilung der Werte im Wertebereich auch die Ausreißer, sowie die Tendenz dazu, dass höhere *Minuten Aktiv* zu leicht gehobenem *Ruhepuls* führen. Dies kann aber auch bloß so wirken, weil es recht wenig Datensätze mit hohen *Minuten Aktiv* Werten gibt und diese sich alle eher im mittleren Ruhepuls Wertebereich aufhalten. Darüber hinaus ist hier mit Ungenauigkeiten zu rechnen, da Ruhepulswerte jeweils von der Nacht auf einen Tag gemessen werden, an dem die aktiven Minuten erfasst werden. Das heißt, hier ist nicht zu sehen, wie der Ruhepuls auf den Sport reagiert, dies kann zudem einige Tage dauern.

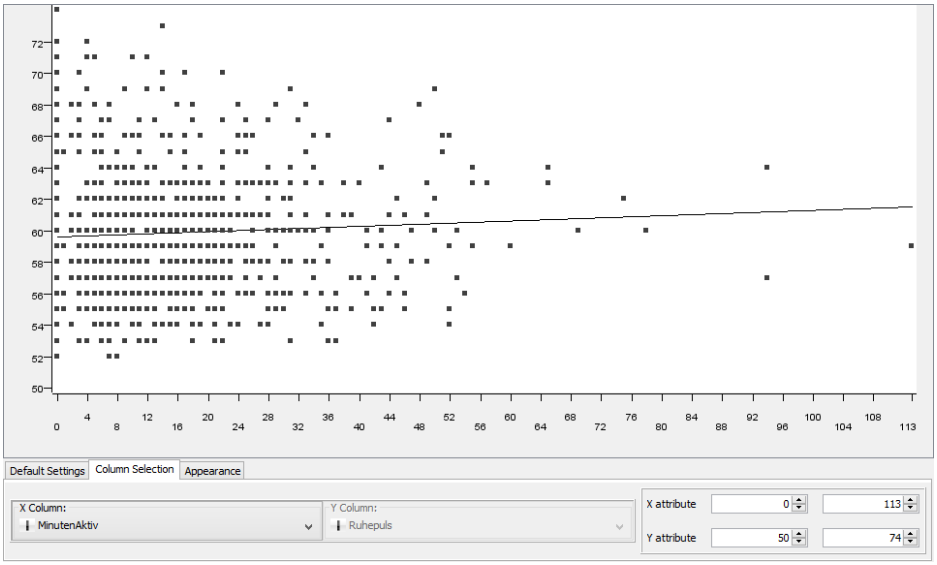

Abbildung 3.10: Beispielhafte Ansicht eines Scatter Plots in KNIME

Neben der Verwendung zweier stetig verteilter Werte kann ein Streudiagramm auch mit einem ordinalen und einem numerischen Wert verwendet werden. Dies sorgt dafür, dass Wochentage, Jahre oder Monate gegenübergestellt werden können. In der untenstehenden Abbildung 3.11 ist zu sehen, wie die Schritte auf die Wochentage aufgeteilt sind. Durch den großen Wertebereich kommt es wenig zu starken Häufungen, es zeichnet sich aber ab, dass die meisten Datenpunkte zwischen 2000 und 10000 liegen. Die genaue Verteilung variiert dabei an den Wochentagen.

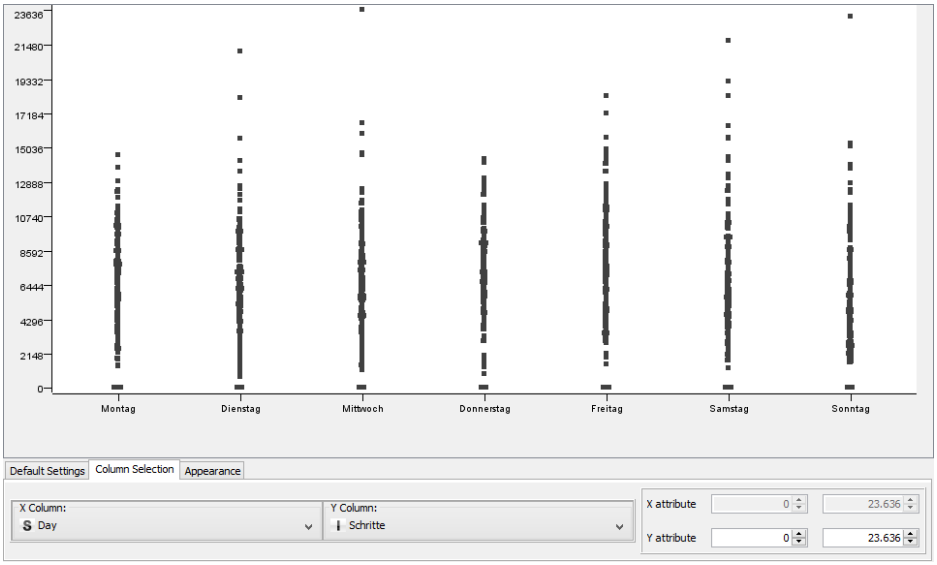

Abbildung 3.11: Beispielhafte Ansicht eines Scatter Plots mit ordinalen Werten in KNIME

Scatter Matrix

Die Scatter Matrix ist eine Matrix aus Streudiagrammen, die nicht nur das gewählte Streudiagram zeigt, sondern alle Diagramme mit den Attributen in allen möglichen Konstellationen. Das heißt, in einer Streumatrix ist jedes Matrixelement *Eij* ein Streudiagramm der Spalten i und j, wobei die Werte der i-ten Spalte auf der x-Achse und die Werte der j-ten Spalte auf der y-Achse angezeigt werden, während die Koordinaten abwechselnd auf allen Seiten der Darstellung angezeigt werden. [17]

Durch diese Konstellation ergibt sich ein Überblick über die Zusammenhänge zwischen verschiedenen Daten, da in einer Matrix mehr als zwei Attribute aufgezeichnet werden können. Die untenstehende Abbildung 3.12 zeigt eine solche Matrix. Sie ist entnommen aus KNIME und stellt drei Attribute gegenüber: Kalorien, Aktivitätenkalorien und Schritte. Dabei ist zu sehen, dass zwischen allen drei Attributen eine starke lineare Abhängigkeit besteht. Verglichen mit der breit gestreuten Punktwolke in Abbildung 3.10, ist hier eine viel deutlichere Abhängigkeit zu erkennen.

[17] Beschreibung entnommen aus der Node Beschreibung von KNIME

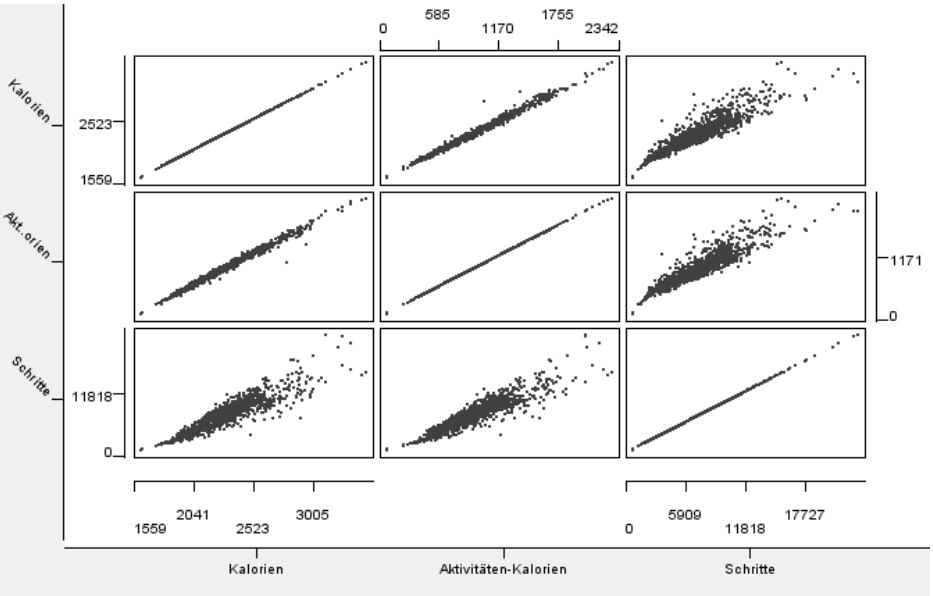

Abbildung 3.12: Beispielhafte Ansicht einer Scatter Matrix in KNIME

Durch die Möglichkeit, mehr Attribute zu vergleichen, ist diese Matrix, vor allem wenn sie interaktiv genutzt wird, ein gutes Mittel um Verständnis für die Struktur und inhärenten Abhängigkeiten der Daten zu entwickeln.

Histogramme

Das Histogramm kann in KNIME interaktiv genutzt werden, um Daten als Balkendiagramm anzuzeigen, das ist gut dazu geeignet, die Häufigkeitsverteilung der Attribute darzustellen. Das heißt, es kann dazu dienen, die Daten besser zu verstehen und sowohl Ober- wie auch Untergrenzen abzustecken sowie ein Gefühl dafür zu entwickeln, in welchen Wertebereichen sich Daten bewegen. Ein Beispiel dafür wurde bereits in Abbildung 3.7 aufgezeigt. Durch die Möglichkeit, die Art der Aggregation zu wählen, können darüber hinaus auch Werte verglichen werden. Bei der Standardeinstellung dem Row Count (Zeilenanzahl) ist dies bei dem hier vorliegenden Datensatz nicht sinnvoll. Wird Attribut x und y mit dem Aggregationsmittel Row Count betrachtet, heißt das, dass aufgezeigt wird, wie sich das Attribut y auf den Datensätzen in denen das Attribut x vorkommt verteilt. Da es aber aufgrund der Datenstruktur auf dem gesamten Datensatz immer alle Attribute geben muss, ist diese Anzeige äquivalent zur Betrachtung der Werteverteilung von y auf dem gesamten Datensatz. Wird jedoch als Aggregator der Durchschnitt genutzt, kann betrachtet werden, wie die Attribute zueinander stehen, zum Beispiel Diastole und Systole, das untenstehende Histogramm 3.13 zeigt auf, dass die Diastole im Durchschnitt bei ca. 52 liegt, wenn die Systole im Bereich 85 bis 91 liegt. Bei der Verteilung von Diastole und Systole ist dies erwartungsgemäß, da bei einem gesunden Blutdruck die Diastole immer unter der Systole liegt.

Daraus ergibt sich die aufsteigende Natur der gezeigten Balkenansicht. Bei anderen Werten kann dabei der Erkenntnisgewinn größer sein. Es gilt also, auf der Suche nach interessanten Ergebnissen diverse Attribute gegenüberzustellen.

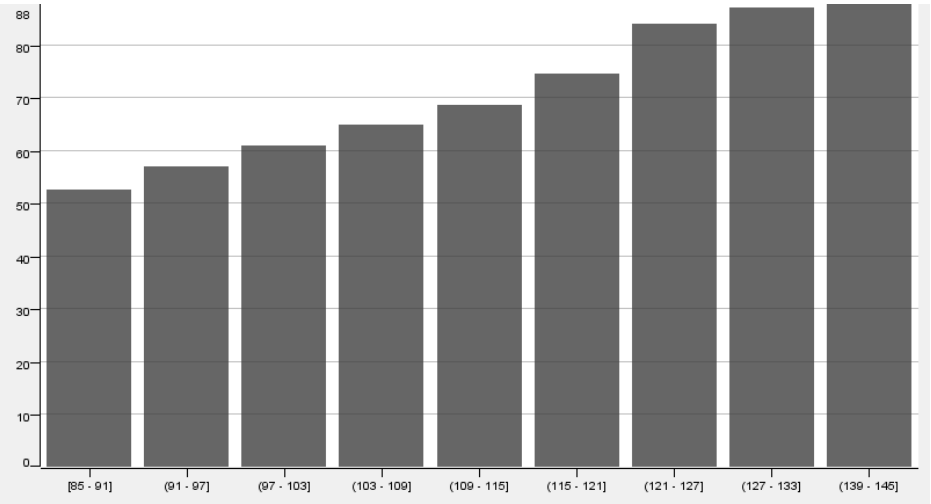

Abbildung 3.13: Beispielhafte Ansicht eines Histogramms mit Systole und Diastole, über den Durchschnitt aggregiert.

Im Prinzip ist ein Histogramm für die Verwendung mit metrischen Daten gedacht. Jedoch kann bei der Verwendung zweier Attribute ein ordinales genutzt werden. In diesem Fall wäre dies der Wochentag, Monat oder das Jahr, in Betrachtung mit einem beliebigen anderen Attribut. So kann aufgezeigt werden, wie viele Stunden im Durchschnitt an den Wochentagen, Monaten oder den Jahren geschlafen wurde, wie viele Schritte pro Jahr gegangen wurden, was der aktivste aller Monate ist und so weiter. Die unten stehende Abbildung 3.14 zeigt die durchschnittliche Schrittverteilung der Jahre.

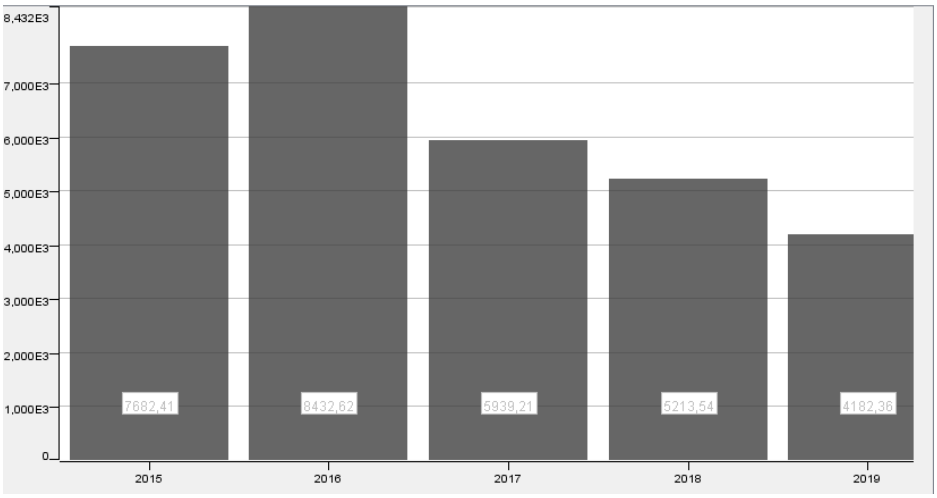

Abbildung 3.14: Beispielhafte Ansicht eines Histogramms mit dem Durchschnitt der Schritte pro Jahr

Liniendiagramme

Neben den spezielleren bereits vorgestellten Diagrammen und Visualisierungen wurden ebenso Liniendiagramme verwendet, um unterschiedlichste Sachverhalte zu veranschaulichen oder neue Informationen zu gewinnen. Diese Art der Diagramme wurde mit Excel erstellt, da dort die Freiheit in der Gestaltung, sowohl der Diagramme, wie auch der Bearbeitung der Daten, größer ist als bei KNIME oder Grafana. Die Liniendiagramme wurden genutzt, um zum Beispiel den Verlauf eines einzelnen Attributs über Zeit genau darzustellen oder um den Verlauf eines Jahres zu untersuchen. Ebenso wurden die Jahre verschiedener Attribute gegenübergestellt, um saisonale Veränderungen besser zu erkennen. Die Abbildung 6.3 zeigt eine solche Gegenüberstellung. Es war möglich und notwendig, Datenreihen um einige Tage zu verschieben, um unterschiedlich schnell reagierende Werte besser untersuchen zu können. So wurde zum Beispiel der Blutdruck mit dem Gewicht versetzt verglichen, weil das Gewicht langsamer auf Veränderungen reagiert als der Blutdruck.

Neben diesen Visualisierungen wurde das Liniendiagramm in Excel dazu verwendet, einen Normalwert aufzuzeigen, also den Bereich zu bestimmen, in dem ein Wert als normal gewertet werden kann, da die meisten Werte innerhalb dieses Bereiches liegen. Dadurch wird schneller klar, welche Werte Ausreißer sind und es wird ein Gefühl dafür vermittelt, ob ein personalisierter Normalwert anders definiert ist, als ein allgemeingültiger Normalwert. Dafür wurden neben der Wertekurve eine Konstante für den Mittelwert sowie den Mittelwert plus und minus Standardabweichung angezeigt. Dadurch ergibt sich ein Bereich zwischen den zwei äußeren Konstanten, der als Normalbereich angesehen werden kann. Normal ist hier nicht als gesund oder optimal

definiert und erhebt keinerlei medizinischen Anspruch sondern ist anhand der Datenbetrachtung und der Verteilung der Werte definiert. Abbildung 6.4 im Anhang zeigt ein solches Diagramm mit Normalskala. Dabei ist zu beachten, dass der Mittelwert sowie die Standardabweichung auf den gesamten Erhebungszeitraum bezogen ist.

3.6 Dateninterpretation und Evaluation

In der Interpretations- und Evaluationsphase des KDD Prozesses werden die Ergebnisse herangezogen, die in den vorherigen Phasen generiert wurden, um daraus Wissen und Erkenntnisse zu entwickelt. Dies ist die entscheidende Phase, in der Ergebnisse diskutiert werden, um aus ihnen einen Mehrwert zu entwickeln. Die Ergebnisse müssen hinsichtlich der Nützlichkeit sowie Glaubwürdigkeit und dem Handlungsbedarf betrachtet und bewertet werden. Waren die Ergebnisse zufriedenstellend? Wurde die Fragestellung beantwortet? Wenn nein, zu welcher Phase des Prozesses muss zurückgegangen werden, um den Ansatz zu korrigieren? Waren die Miningverfahren geeignet? Lagen die richtigen Daten in einer passenden Granularität vor etc.? Das heißt, in dieser Phase werden sowohl die Ergebnisse auf ihrer technischen Ebene wie auch hinsichtlich dem Nutzen für den Benutzer betrachtet.

Interpretationskomponente

Entwurf
Die Interpretationskomponente soll die Daten und Ergebnisse interpretieren und mit Wissen aus dem Expertensystem anreichern. Daraus sollen Erkenntnisse gezogen sowie dem Benutzer die Möglichkeit gegeben werden, die eigenen Daten mithilfe des Expertenwissens zu bewerten. Dabei soll darauf geachtet werden, keine vorschnelle Bewertung durchzuführen, die der Nutzer vielleicht gar nicht möchte. Besonders das Bewerten der Daten kann ein sehr sensibles Feld sein, das für den Nutzer kontraproduktiv umgesetzt werden kann, wenn er zu stark mit ungewollten negativen Reizen konfrontiert wird. Das heißt, das Augenmerk soll darauf liegen, dem Nutzer ein ausgewogenes und möglichst umfangreiches Bild der Daten zu ermöglichen. Allerdings sollte das System in der Lage sein, Besonderheiten der Nutzer zu erkennen und sie wiederzugeben. Wo finden sich Auffälligkeiten in den Daten, welche Bedeutung könnten sie haben? Welche Veränderungen sind in jüngster Zeit aufgetreten? Gibt es Werte, die der Nutzer besonders im Auge haben möchte? Kann das Expertensystem Ratschläge, Informationen oder Erkenntnisse über die Daten liefern? Des Weiteren sollte das System lernen, was für den Nutzer normal ist, also welche Werte in welchen Bereichen zu erwarten sind und dann darauf reagieren können, wenn diese Werte ihren Normalbereich verlassen. Es ist wichtig, individualisierten und vertrauenswürdigen Umgang

mit den Daten zu pflegen und die Erkenntnisse und das Wohl des Nutzers in den Vordergrund zu stellen.

Implementierung
Die Komponente wurde nicht umgesetzt und auch eine Anbindung an ein Expertensystem gibt es nicht. Jegliche Interpretation der Daten wurde händisch mit Expertenwissen aus diversen Quellen ermittelt.

3.7 Anzeige und Kommunikation mit dem Nutzer

Neben den Komponenten, die sich stark am KDD Prozess orientieren, benötigt das System Komponenten, die es zu einem vollständigen nutzerzentrierten System machen. Diese Komponenten finden sich allerdings nicht im KDD Prozess wieder, da sie nicht direkt an der Datenverarbeitung beteiligt sind. Dies ist zum einen die Kommunikationskomponente, die sich um jegliche Kommunikation mit dem Nutzer kümmert sowie die Anzeigekomponente, die die Anzeige aller Daten, Visualisierungen, Erkenntnisse und Ergebnisse orchestriert.

Anzeigekomponente

Entwurf
Die Anzeigekomponente soll für den Nutzer jeweils individualisiert die Daten und Visualisierungen anzeigen. Da es eine Mischung aus Quellen gibt sowie die Art und Weise der Anzeige wichtig ist, ist dafür eine eigene Komponente geplant, die die Arten der Anzeigen ermöglicht und die Sichten (Views) integriert. Ebenso soll sie die Vorlieben der Ansicht des Nutzers speichern und ihm eine individualisierte Sicht auf die eigenen Daten ermöglichen. Sie nimmt neben den Rohdaten auch verarbeitete an, um sie genauso wie Visualisierungen anzeigen zu können. Neben der Möglichkeit, die Daten zu sichten und zu verstehen, muss die Anzeigekomponente Views haben, in denen der Nutzer das System bedienen und Daten einpflegen kann, damit qualitative Daten abgefragt werden können, die später den Datenbestand erweitern.

Neben den verschiedenen Views für unterschiedliche Aufgaben im System muss diese Komponente verschiedene Ausgabegeräte unterstützen, damit der Nutzer neben dem Spiegel selbst auch ein Tablet oder Smartphone sowie eine Website nutzen kann.

Implementierung
Die Anzeigekomponente wurde in Teilen prototypisch umgesetzt, um die Anbindung an das System sowie die Anzeige auf dem Spiegel zu testen und in ihrer Umsetzbarkeit zu überprüfen.

Dafür wurde ein rudimentäres Weboverlay in HTML implementiert, das über dem Spiegelbild angezeigt werden kann und dabei ausgesuchte Daten aus dem Backend des Systems präsentiert. Diese Umsetzung erfolgte in C#, das nötige Videobild wurde aus der Kinect bezogen. Dabei wurde der Fokus nicht auf schöne Visualisierung oder Übersichtlichkeit gelegt. Es ging in erster Linie darum, prototypisch Daten über die Middleware vom Server auf dem Spiegel anzuzeigen. Somit sind es eingeschränkte Rohdaten, die zu sehen sind, jedoch gibt es eine Anbindung des Javasciptes des Overlays zur Middleware des Labors. Wie der Spiegel im Laboraufbau mit den Overlays aussah, zeigt das Bild 3.15

Die erste Umsetzung erfolgte im Rahmen des Grundprojektes Lüdemann (2017a) und hatte die Form einer statischen Anzeige, die mit dem Kamerabild verbunden wurde, um auf dem Spiegel angezeigt zu werden. In der Arbeit des Hauptprojektes Lüdemann (2018) wurde beschrieben, wie die Anbindung an die Middleware und die damit verbundene Datenübertragung realisiert wurde.

Kommunikationskomponente

Die Kommunikations- oder auch Interaktionskomponente regelt die Möglichkeiten des Benutzers, mit dem System zu interagieren.

Entwurf
Die Komponente sollte modular und erweiterbar aufgebaut werden, um weitere Möglichkeiten der Bedienung einbauen zu können. Dabei soll die Kommunikation zwischen dem Nutzer und den unterschiedlichen Sichten auf unterschiedlichen Geräten auch differenziert bearbeitet werden können. Während der Spiegel selbst durch Touch, Sprache, Gesten oder einem Second Screen benutzt werden könnte, ist eine App immer per Touch und eine Website zumeist traditionell benutzbar. Das heißt, die Komponente sollte unabhängig von der Art der jeweiligen Bedienung implementiert werden, sodass die Arten der Kommunikation als eigene Module implementiert werden oder dem Endgerät obliegen. Der Blickwinkel sollte hierbei jedoch eher auf multiplen Modalitäten liegen als auf speziellen Technologien oder Hardware, wie auch im Smartkom Forschungsprojekt Wahlster (2003), wo dadurch eine viel höhere Flexibilität erreicht wird. Die Verwendung eines tragbaren Gerätes, wie einem Smartphone oder Tablet, bietet sich dabei an, da sie gut verfügbar sind und den Nutzer nicht zu sehr bei der Verwendung des Spiegels einschränken. Dafür kann der Spiegel aus der Entfernung zuverlässig und sicher über eine App bedient werden, die dazu weitere Informationen anzeigen könnte sowie eine Datenaufnahme vom Nutzer ermöglicht.

Implementierung
Die Kommunikation mit dem Nutzer wurde im Rahmen dieser Arbeit nicht als Modul

entwickelt, sondern behelfsmäßig über den Server beziehungsweise Desktop-PC der den Spiegel betreibt, realisiert.

3.8 Der Spiegel

In diesem Abschnitt wird darauf eingegangen, warum sich ein Spiegel als Oberfläche eignet und wie der technische Laboraufbau umgesetzt wurde.

Ein Spiegel als beispielhafte Anzeige

Für das System wären neben dem Spiegel auch andere Anzeigearten denkbar. Das Herzstück des Systems umfasst das Sammeln und Auswerten der Daten. Die Ergebnisse daraus könnten dem Benutzer auch mit einer App oder auf einer Website angezeigt und zugänglich gemacht werden. Jedoch ist ein Spiegel seit jeher ein Ort, sich selbst zu begegnen, das Sichtbare aufzuzeigen. Im Abschnitt 2.2 wurde bereits darauf eingegangen, warum ein Spiegel sich besonders eignet, um Körperdaten anzuzeigen. Zum einen wegen seiner Symbolik für die Menschen, zum anderen weil die Anzeige der Daten neben dem Körper beziehungsweise dem Spiegelbild des Körpers dem Nutzer hilft, Verbindungen zu ziehen und weil es die eigentliche Funktion eines Spiegels aufgreift, vertieft und erweitert. Ein Spiegel der nächsten Generation im Rahmen der Smart Objects, der nicht nur der Begegnung der eigenen äußeren Erscheinung dient, sondern einen tieferen Einblick und ein tieferes Verständnis liefert. Ein Blick in die Black Box Körper, eine Möglichkeit das Unsichtbare sichtbar zu machen und ein Blick unter die Haut. Dabei muss im Prinzip kein neues Verhalten erlernt werden, um das System zu nutzen, denn die meisten Menschen blicken ohnehin täglich in den Spiegel. Der tägliche Blick in den Spiegel kann somit nicht nur dem Herrichten dienen, sondern auch, um zu überprüfen, ob Abweichungen vom Standard zu sehen sind, optisch wie auch im digitalen Abbild seiner Selbst.

Der Laboraufbau

Die Umsetzung als Laboraufbau diente dazu, sicherzustellen, dass der Aufbau konzeptionell durchführbar ist. Dabei wurde der Fokus nicht darauf gelegt, ästhetischen oder praktischen Ansprüchen zu genügen, von einer Endfassung ist der Aufbau noch entfernt. Aktuell entspricht sie dem Stand, der in Lüdemann (2017a) beschrieben wurde.

Für den Spiegel wird eine Spiegelfläche benötigt, diese kann, wie in Abschnitt 2.2 beschrieben, ein Bildschirm oder ein Spiegelglas mit Bildschirm dahinter sein. Darüber

hinaus werden Kameras benötigt, um die Funktionalitäten mit Bildverarbeitung zu gewährleisten oder um das Spiegelbild abzunehmen, falls ein Bildschirm ohne Spiegel verwendet wird. Die Daten müssen von einem leistungsstarken Rechner verarbeitet werden. Je nach Art des Spiegels benötigt der Nutzer eine Möglichkeit, mit dem System zu interagieren, sei es ein Touch-Bildschirm, ein Zweitgerät wie ein Smartphone oder ein Tablet. Ebenso wäre eine Bedienung über Sprachbefehle oder Gesten denkbar. Jedoch muss die Technik vorhanden sein, die diese Kommunikation annimmt.

Bei der Laborumsetzung wurde ein Bildschirm für den Versuchsaufbau genutzt, in weiteren Stadien der Entwicklung kann dann auf einen halb durchlässigen Spiegel mit Bildschirm umgestiegen werden. Der Bildschirm mit einer Diagonale von 79 Zoll wird vertikal genutzt, um einem Ganzkörperspiegel zu ähneln. Das Bild erfasst eine Kinect, die neben dem Bildschirm platziert wurde. Aufgrund der Größe des Bildschirmes ist eine Platzierung über oder unter dem Bildschirm nicht möglich, da sich die Perspektive des Spiegelbildes dadurch zu sehr verschiebt. Auch neben dem Spiegel gibt es eine leichte Verschiebung, diese ist aber gering genug, um im Laboraufbau nicht zu stören. Die Verarbeitung der Daten übernimmt ein Desktop-PC, der in der Nähe des Spiegels steht. Die untenstehende Abbildung 3.15 zeigt den Versuchsaufbau im Labor.

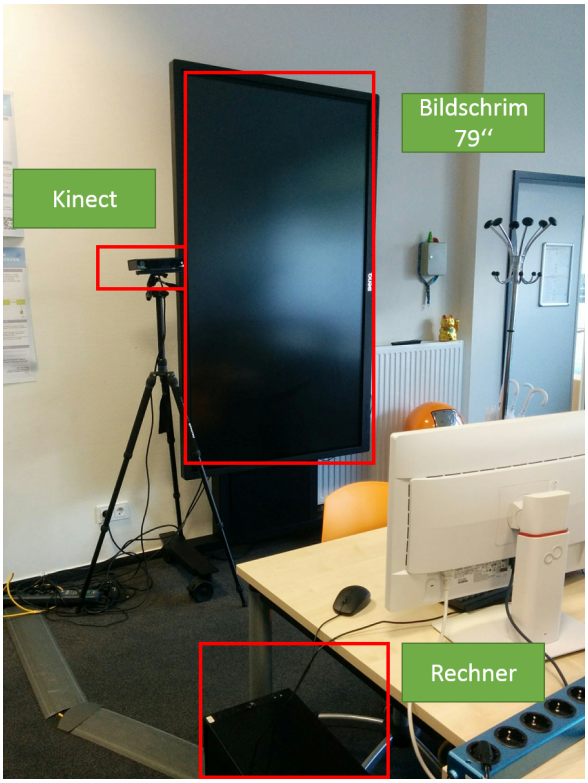

Abbildung 3.15: Der Versuchsaufbau des Spiegels im Labor.

Die Bedienung des Spiegels läuft in dieser Version des Aufbaus über den Rechner selbst. Dies sollte durch eine andere Bedienungsart ersetzt werden, da sie hier nur für Testzwecke genutzt wurde und für jede Bedienung den Nutzer zwingt, sich vom Spiegel abzuwenden.

3.9 Fazit

In diesem Kapitel wurde gezeigt, dass ein Quantified Self System, das die Daten aus Sensoren extrahiert, bereinigt und für Auswertungen transformiert umsetzbar ist. Dazu können Analysen angewendet werden, um die Daten auszuwerten und neue Erkenntnisse zu erzeugen. Die Umsetzung als Ganzkörperspiegel ist ebenso möglich und gezeigt worden. Dabei haben sich in der Verarbeitung der Daten einige Probleme ergeben, die zum Großteil bereits in der Analyse der Daten abzusehen waren. Im Umfang und der Funktionalität des Systems konnte alles wie geplant getestet und umgesetzt werden, sodass es nun einen laboratorischen Stand gibt, der jeden Aspekt testet aber kein vollständiges System erzeugt.

In zukünftigen Arbeiten sollte zudem ein halbdurchlässiger Spiegel mit Bildschirm und integrierter Kamera genutzt werden, da dadurch die Nutzererfahrung verbessert wird.

Mit mehr Zeit und Ressourcen wäre das System, das hier laboratorisch implementiert und evaluiert wurde, über den hier aufgezeigten Stand hinaus implementier- und umsetzbar. Die meisten Aspekte konnten im Rahmen dieser Arbeit zumeist nur angerissen werden. Mit mehr Zeit könnte dabei jeweils tiefer gegangen werden, um noch mehr Aspekte auszuarbeiten und das System zu vervollständigen und zu verbessern.

Die Datenanalyse könnte von mehr Daten, präziseren Auswertungen und Beobachtungen sowie mehr Durchläufen mit veränderten Umständen in den Verfahren profitieren. Dabei ist unter anderem die zeitliche Dimension der Daten gemeint, weitere Daten über einen großen Zeitraum könnten wiederkehrende Muster verdeutlichen aber auch mehr Quellen. Je mehr Daten korreliert und im Zusammenhang gesetzt werden können, umso mehr Aspekte können betrachtet werden.

4 Evaluation

In diesem Kapitel werden die Ergebnisse der Analysen und der Arbeit diskutiert und evaluiert. Da es in der Arbeit ergebnisoffene Fragestellungen und Analysen gab, werden hier einzelne Aspekte exemplarisch ausgewertet und diskutiert, um ein Bild davon zu geben, wie eine Auswertung der Daten aussieht ohne den Rahmen dieser Arbeit zu sprengen. Dieses Kapitel ist ähnlich wie das Entwurfs- und Implementationskapitel aufgebaut. Als erstes wird die Datenerhebung und Bereinigung evaluiert. Es soll betrachtet werden, welche Probleme aufgetreten sind, was gut funktioniert hat, was hätte besser gemacht werden können und welche Sensoren oder Daten noch von Vorteil gewesen wären. Danach wird die Verarbeitung der Daten unter sehr ähnlichen Gesichtspunkten besprochen, um dann die Ergebnisse der Analysen zu diskutieren. Des Weiteren wird auf die Datenqualität eingegangen und abschließend die Arbeit sowie der technische Aufbau evaluiert.

4.1 Datenerhebung und Bereinigung

In diesem Abschnitt wird die Datenerhebung und Bereinigung evaluiert. Dabei wird sowohl auf technische, wie auch menschliche Aspekte eingegangen sowie Aspekte aus der Literaturrecherche den Erfahrungen der Datenerhebung gegenübergestellt. Dieses Wissen fließt in die Evaluation der Sensoren ein. Danach wird die Datenextraktion evaluiert sowie Probleme und Besonderheiten der Extraktion aus den Hersteller APIs diskutiert. Als letztes wird das Augenmerk auf die Bereinigung und Verarbeitung der Daten gelegt, inwieweit dies automatisiert wurde und welche Probleme dabei auftraten.

Datenerhebung

Die Daten wurden über einen recht langen Zeitraum vom Juli 2015 bis zum Februar 2019 erhoben. Dabei wurde auf manuelle Daten großteils verzichtet, da sich sowohl in den Testphasen der vorangegangen Bachelorarbeit Lüdemann (2016a), sowie in der Literaturrecherche 2.2 ergeben hat, dass dies oft ein Problem in der Erhebung darstellt. Manuelle Daten sind oft zu zeitaufwändig, um vom Nutzer angenehm erfasst werden zu

können. Es gibt Mittel und Wege, die Erhebung so einfach wie möglich zu machen, zum Beispiel indem einfache Skalaabfragen über eine App gemacht werden, die den Nutzer erinnert und ihm mit wenigen Klicks eine Erhebung ermöglicht. Dies funktioniert, es muss jedoch die Technik dafür vorhanden sein. Soll der Nutzer aufwändige Daten wie die Ernährung in Apps erfassen, kommt es schnell zu Datenungenauigkeiten und Ausfällen, bis der Nutzer das Erfassen ganz abbricht. Daher wurde von vornherein auf diese Art der Erhebung verzichtet und nur ein Kalender möglichst genau geführt. Da dieser aber zum einen sehr persönliche Informationen enthält und zum anderen nicht zufriedenstellend genau geführt wurde, fließt er nur zum Überprüfen und in Kontext setzen der automatisch erfassten Werte ein.

In Abschnitt 2.1 und 2.2 wurde darauf eingegangen, welche Kriterien des Erfassens und der Anzeige in der vorliegenden Literatur und in den vorangegangenen Arbeiten beschrieben werden. Diese Kriterien werden hier noch einmal aufgenommen. Es ist wichtig, dass Daten so automatisch wie möglich erhoben werden. Jede vom Nutzer notwendige Interaktion sollte sich problemlos in die tägliche Routine eingliedern, ohne großen Mehraufwand zu erfordern. Besonders bei Langzeiterfassung können schon kleine Hürden zu Datenausfällen führen, angefangen bei der Notwendigkeit, ein Gerät zum Duschen oder Aufladen abzulegen. Dabei kann ein Nutzer schnell dazu neigen, das Wiederanlegen zu vergessen. Bis hin zur Mobilität der Geräte, ein viel reisender Nutzer wird seine Körperwaage nicht mitnehmen, seien es Berufsreisende oder einfach Nutzer die aktiv viel reisen. Bei ihnen werden sich immer Datenausfälle dadurch ergeben, dass sie nicht immer zu Hause sind. Es muss aber gar nicht beruflich gereist werden, um auf Probleme zu stoßen. In den Zeiten, in denen immer mehr Menschen alleine Wohnen, kommt es immer häufiger vor, dass in Partnerschaften an zwei unterschiedlichen Orten geschlafen wird. Dies führt, wenn man Geräte nicht doppelt anschafft, unweigerlich zu Datenausfällen.

Darüber hinaus müssen Punkte wie die Tragbarkeit betrachtet werden. Geräte müssen sich möglichst komfortabel und an den Kleidungsstil anpassbar gestalten, sonst werden sie nur getragen, wenn das Outfit es zulässt. Die Geräte müssen sich insgesamt möglichst gut an das Leben des Nutzers anpassen sowie hautverträglich und angenehm im Tragegefühl sein und auch nachts nicht stören, sonst ergeben sich schnell Gründe, den Sensor nicht anzulegen. Es ist für die Qualität der Daten unabdingbar, dass sie möglichst regelmäßig erhoben werden, je größer die Lücken zwischen den Datensätzen sind, umso weniger Aussagen können getroffen werden. Daher ist es notwendig, die Sensoren so zu gestalten, dass sie sich problemlos und angenehm in möglichst viele Lebenslagen integrieren lassen, ohne den Nutzer damit zu konfrontieren, viel darüber nachdenken zu müssen, wie und wann Daten erfasst werden. Selbst wenn diese Faktoren beachtet werden, kann es passieren, dass ein Nutzer nach ein paar Wochen das Interesse verliert. Der Novelty Effekt bindet die Nutzer einige Zeit an einen interessanten neuen Sensor, sobald er aber zur Gewohnheit wird, sinkt die Motivation, sich damit zu beschäftigen Koch u. a. (2018). Das heißt, dass das Erheben bis zu dem

Punkt, an dem der Novelty Effekt nachlässt, zur Routine geworden sein muss, um langfristig durchgeführt zu werden.

Abgesehen davon, dass das manuelle Erfassen von Daten auf lange Sicht praktisch nicht durchzuführen ist, sofern der Nutzer nicht eine sehr hohe intrinsische Motivation dazu hat, wären diese Daten von großem Vorteil, angefangen bei subjektiven Wahrnehmungen um die automatisch erhobenen Werte zu evaluieren, sei es eine subjektive Bewertung der Schlafqualität, der Müdigkeit und des Stressfaktors über den Tag oder eine Einordnung der Stimmung. Ebenso wären Ernährungsinformationen sehr wertvoll, neben der Nahrung und der Energiemenge wäre auch die Art der Nahrungsmittel interessant, sowie Alkohol- und Koffeinkonsum. Diese Informationen könnten in den Kontext zu Schlafqualität, Gewicht, Blutdruck und Stimmung gesetzt werden, um tiefere Einsichten zu ermöglichen. Ebenfalls wäre es interessant, Informationen über Krankheiten und Medikationen zu erhalten um ggf. Veränderungen in den Daten, die durch Krankheit hervorgerufen werden, benennen zu können.

Neben der Tatsache, dass die Sensoren mehr in den Routinen des Alltags verschwinden müssen, kommt hinzu, dass sie zuverlässiger werden müssen. Das Laden der Geräte, Übertragen der Daten und die Benutzungserfahrung müssen einfach, komfortabel und vor allem zuverlässig sein. Übertragungsfehler, Bedienungsprobleme und komplizierte Prozesse werden schnell zu einem Grund, Daten nicht zu erfassen.

Alles in allem war das Erfassen mit den drei hier genutzten Geräten, dem Fittnessarmand, der Analysewaage und dem Handgelenkblutdruckmessgerät bereits recht komfortabel, in einem längeren Zeitraum gut durchführbar und wurde weitergeführt. Jedoch ist zu bedenken, dass durchaus eine starke intrinsische Motivation zu Grunde lag. Daher war nicht zu erwarten, dass der Nutzer das Interesse am Erheben verlieren könnte und gewillt war, auch Probleme und Schwierigkeiten zu überwinden.

Sensoren

In diesem Abschnitt soll explizit auf die genutzten Sensoren eingegangen werden. Dabei werden die Vor- und Nachteile diskutiert, die sich ergaben und die Sensoren auf die Kriterien hin betrachtet, die im Datenerhebungsabschnitt beschrieben wurden.

Fitbit Fitnessarmband
Das Firnessarmband, das getragen wurde, war im ersten Abschnitt des Zeitraumes ein Fitbit Surge, dieses musste jedoch nach ca. sieben Monaten ausgetauscht werden, da es durch Wasser beschädigt wurde. Das zweite Gerät ist ein Fitbit Blaze. Zu beiden Geräten kann man sagen, dass sie angenehm zu tragen sind. Im Vergleich zum Surge hat das Blaze das Problem, dass der Bildschirm sehr hell ist, was besonders nachts dazu führen kann, dass es unangenehm blendet, wenn man das Handgelenk bewegt.

Die beiden Geräte sind durch verschiedene Armbänder, die zusätzlich erworben werden müssen, sehr angenehm an verschiedene Lebenslagen, den persönlichen Stil und Tragekomfort anpassbar. Dies jedoch nur äußerlich, da die Oberfläche der App und des Gerätes selbst nur geringfügig modifizierbar und praktisch nicht individualisierbar sind. Es können verschiedene Layouts gewählt werden, doch sind diese nicht anpassbar, sodass man mit den von der Firma vorgegebenen Oberflächen leben muss. Ebenso ist nicht konfigurierbar, wie lange das Display angeschaltet bleibt, nachdem man es durch die Handgelenkbewegung oder per Knopfdruck aktiviert hat. Dies ist besonders beim Sport störend, da sich das Display immer wieder ausschaltet, sodass man die Werte wie Zeit und Puls nicht mehr sehen kann. Besonders bei Sportarten, bei denen man die Hände nicht frei hat wie Gewichtheben oder Eigengewichtsübungen, ist das sehr störend.

Neben den Problemen in der Nutzung verliert das Fitbit recht schnell den Kontakt, sodass der Puls nicht gemessen wird. Dies kann durch falsche Trageposition passieren aber auch durch bestimmte Bewegungen oder wenn der Puls außerhalb einer Sportmessung zu schnell steigt. Teilweise wird der Kontakt auch ohne erkennbaren Grund unterbrochen. Dies führt immer wieder zu Messausfällen sowie dazu, dass man gerade dann wenn es wirklich interessant ist, zum Beispiel nach einem Sprint zum Bus oder schnellem Treppensteigen, keine Möglichkeit hat, den Puls zu verfolgen, wenn man vorher nicht explizit eine Sportmessung gestartet hat.

Ein weiterer negativer Punkt ist, dass das Gerät alle paar Tage aufgeladen werden muss und dafür ein spezielles Ladekabel benötigt, welches von Produktreihe zu Produktreihe variiert und nur von Fitbit genutzt wird. Vergisst man also sein Ladegerät auf Reisen, ist es praktisch unmöglich, dies irgendwo zu leihen oder zu ersetzen, ohne ein neues zu kaufen. Ebenso können die Ladegeräte nicht weiter benutzt werden, wenn die Produktlinie gewechselt wird, wie hier vom Surge auf das Blaze.

Die App des Gerätes, in dem die Daten einsichtig sind, nutzt relativ starke Datenaggregation und Gamification. Der Nutzer wird nicht mit Rohdaten belastet und bekommt die Daten bereits kontextualisiert und bewertet angezeigt. Dies kann je nach Einstellung des Nutzers positiv oder negativ sein. Zum einen können große Tabellen und unbewertete Rohdaten den Nutzer verwirren und ihm das Einschätzen seiner Werte erschweren, zum anderen können sie ihm auch das Gefühl der Bevormundung und des Datenvorenthaltens geben. Dabei ist das Augenmerk der App nicht auf das neutrale Erfassen von Daten gerichtet, sondern eher auf die Optimierung des Nutzers. Der Nutzer wird stark durch visuelle Reize und Abzeichen belohnt, wenn er das vom Hersteller vorgegebene Ziel erreicht. Praktisch jeder Wert kann mit einem virtuellen Ziel belegt werden, dessen Erreichen dann durch leuchtend grüne Farben und eine kleine Animation belohnt wird. Diese Ziele sind vom Nutzer anpassbar, lassen sich jedoch nicht ausstellen. Neben der Belohnung gibt es auch negative visuelle Reize durch rote oder gelbe Farben oder unvollständig ausgefüllte Ringe, wenn die Werte nicht erreicht

werden. Dabei ist durchaus fraglich, welcher Sinn hinter einem Ziel für Treppensteigen steht. Warum sollte ein Nutzer mindestens 10 Treppen am Tag steigen, wenn seine Umgebung zum Beispiel gar keine Treppen beinhaltet? Hinzu kommt, dass zumindest der Surge Sensor die Treppen nicht richtig gemessen hat, sodass das Ziel eigentlich immer erfüllt wurde, weil unverhältnismäßig viele Treppen gemessen wurden. Bis zu 602 waren pro Tag gemessen, obwohl nur 10-20 gegangen wurden.

In der App ist anpassbar, welche Werte auf den ersten Blick einsichtig sind. Somit kann der Nutzer für ihn weniger relevante Daten weiter in den Hintergrund der App schieben. Die untenstehende Abbildung 4.1 zeigt die App, wie sie während des Erfassungszeitraums ausgesehen hat. Die unvollständigen Kreise weisen darauf hin, dass Tagesziele noch nicht erreicht wurden. Wenn sie hingegen erreicht werden, ist der Balken wie beim Schlafwert grün, das Ziel ist hier mindestens sieben Stunden zu schlafen, da es bereits vom Nutzer von acht auf sieben Stunden angepasst wurde.

Abbildung 4.1: Übersicht der Fitbit-App, entnommen der FitbitApp auf dem Smartphone

Die Datendarstellung in der App ist sehr übersichtlich und ansprechend gestaltet, jedoch ist eine tiefere Einsicht in die Daten oft nicht möglich. Es wird viel mit aggregierten und berechneten Daten gearbeitet, deren Zusammensetzung nicht immer ersichtlich ist. Der neueste Wert ist der Schlafindex. Er kam im Spätsommer 2019 in

die deutsche App und gibt eine Zahl zwischen 0 und 100 als Bewertung für den Schlaf an. Dabei ist ein Wert unter 60 als wenig bezeichnet, zwischen 60 und 79 als akzeptabel, zwischen 80 und 90 gut und darüber exzellent. Wie sich dieser Wert jedoch zusammensetzt und was er bedeutet, wird dem Nutzer nicht mitgeteilt. Er erhält bloß eine hoch aggregierte Information über einen so komplexen Vorgang wie den Schlaf. Durch ein wenig Recherche im Internet lässt sich ermitteln, wie sich der Wert im Groben zusammen setzt, aber wirklich genaue Informationen erhält der Nutzer über seine persönlichen Daten nicht. Die untenstehende Abbildung 4.2 zeigt, wie die App nach dem Einführen des Schlafindexes aussieht und wie prominent dieser in der App platziert wird, obwohl dem Nutzer keine Einsicht in die Bedeutung dieses Wertes gegeben wird.

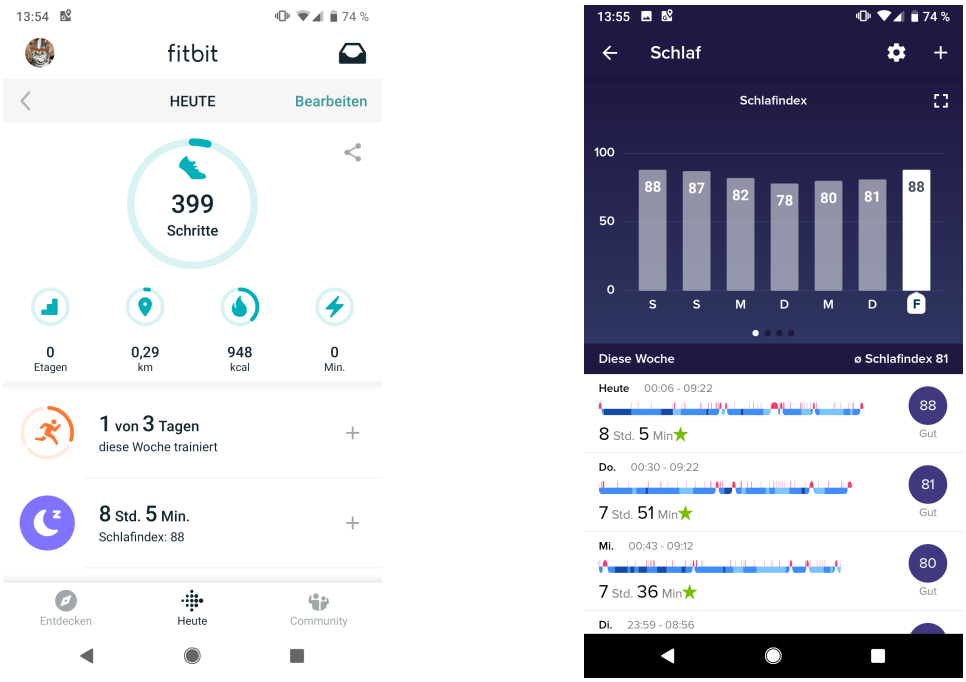

Abbildung 4.2: Neue Ansicht und Schlafindex der Fitbit App, entnommen der FitbitApp auf dem Smartphone

Dieser Trend der Datenaggregation und Bevormundung der Nutzer scheint sich leider langsam durchzusetzen. Darüber hinaus gibt es in den Vereinigten Staaten bereits ein sogenanntes Fitbit Premium Abonnement FitbitPremium, welches dem Nutzer für eine monatliche Zusatzzahlung mehr Informationen und Daten einsehen lässt. Dies wird hier als sehr kritisch und bedenklich bewertet, da es immerhin um eigene persönliche Daten geht, die nicht durch eine Bezahlschranke vor den Nutzern verborgen sein dürften. Sollte dieser Dienst auch in Deutschland eingeführt werden, muss

dringend noch einmal evaluiert werden, ob eine weitere Benutzung eines Fitbits ratsam ist.

Analysewaage
Die über den gesamten Messzeitraum verwendete Waage ist die Analysewaage Medisana BS 440 Connect. Die Waage ist angenehm einfach und von mehreren Nutzern über verschiedene Profile nutzbar. Nach einem einfachen Einstellen der Profile erkennt sie, wenn sich die Profile genug unterscheiden, wer auf der Waage steht, ohne, dass es jedes Mal explizit angeben werden muss. Die Datenverbindung zur App funktioniert über Bluetooth schnell und zuverlässig. Die Laufzeit des Gerätes ist zudem angenehm lang. Leider entwickelte das Gerät während des Erfassungszeitraums kleinere Mängel, so muss das Gewicht zum Beispiel immer zweimal gemessen werden, weil die erste Messung immer 0,8 - 1,5 kg höher ist als der reale Wert. Dies wurde mithilfe eines zweiten Gerätes überprüft. Darüber hinaus leidet die Messgenauigkeit, wenn die Leistung der Batterie abnimmt.

Neben dem Gerät gibt es von Medisana eine App, die jedoch deutlich weniger designt ist als die von Fitbit. Das Augenmerk liegt bei ihr nicht zwingend auf perfekter Nutzbarkeit, sondern auf Datentransparenz. Die Daten, die das Gerät erhebt, sind in der App alle einsehbar, sowohl übersichtlich als Tagesausschnitt wie auch in tabellarischer Form. Es wird auf hoch aggregierte Werte verzichtet und dafür dem Nutzer das Gefühl vermittelt, die Kontrolle über seine Daten zu haben. In der App können auch alle Daten bearbeitet und gelöscht werden. Sie ist dabei leider ein wenig unübersichtlicher und schwieriger zu benutzen als die von Fitbit.

Es wäre von Vorteil, wenn die jeweiligen Geräte sich Informationen aus der App holen könnten, im Augenblick sieht es so aus, als gäbe es nur in eine Richtung Informationsfluss, vom Gerät in die App. Dies kann dann zu Problemen führen, wenn das Datum, das auf der Waage eingegeben wurde, nicht mit dem aktuellen Datum übereinstimmt. Auch die Daten über die Profile können nur auf dem Gerät verändert werden, dabei ist die Waage deutlich schwieriger zu bedienen als es die App wäre. Es wäre also deutlich komfortabler und weniger fehleranfällig, wenn die Geräte sich die Datums- und Profilinformationen über die App auf dem Smartphone holen würden. Alles in allem ist das Gerät jedoch zu empfehlen und Medisana als Firma scheint sehr vertrauenswürdig und benutzerorientiert mit den Daten umzugehen.

Blutdruckmessgerät
Das Blutdruckmessgerät ist ebenfalls ein Gerät von Medisana, es handelt sich dabei um das Handgelenkblutdruckmessgerät BW 300 Connect. Es ist ebenfalls angenehm einfach zu benutzen. Dazu kommt, dass die Akkulaufzeit sehr lang ist und das Gerät über einen standardisierten Mini-USB-Stecker geladen werden kann. Allerdings warnt das Gerät den Nutzer nicht davor, dass der Akku bald leer ist, der Nutzer kann dies erst merken, wenn das Gerät nicht mehr messen kann. Dadurch geht eigentlich immer eine Messung verloren. Darüber hinaus ist die Datenübertragung problematischer als

bei der Waage, das Gerät ist sehr anfällig für falsche Datums- und Uhrzeiteinstellungen, dadurch schlägt die Verbindung fehl ohne das dem Nutzer bewusst ist, wieso. Die Übertragung dauert zudem viel länger als bei der Waage und bricht ohne erkennbaren Grund gelegentlich ab. Das heißt, dass es für dieses Gerät wirklich von Vorteil wäre, wenn es sich das Datum über die Datenverbindung holt sowie wenn sich die Datenübertragung signifikant verbessern würde.

Die Oberflächen der Geräte und der App sind optisch nicht individualisierbar und beinhalten so gut wie keine Gamifikation. Die Werte werden farblich bewertet angezeigt, sodass wünschenswerte Werte grün und schlechte Werte magenta angezeigt werden, diese Bewertung wird laut der Hersteller den Angaben der WHO entnommen. Die untenstehende Abbildung 4.3 zeigt die Medisana-App und ihr Erscheinungsbild. Zum einen gibt es eine recht übersichtliche Einstiegsseite, auf der die letzten Messungen der jeweiligen Geräte gezeigt werden, zum anderen können weitere Informationen mittels eines Diagrammes oder einer ausführlichen tabellarischen Ansicht betrachtet werden.

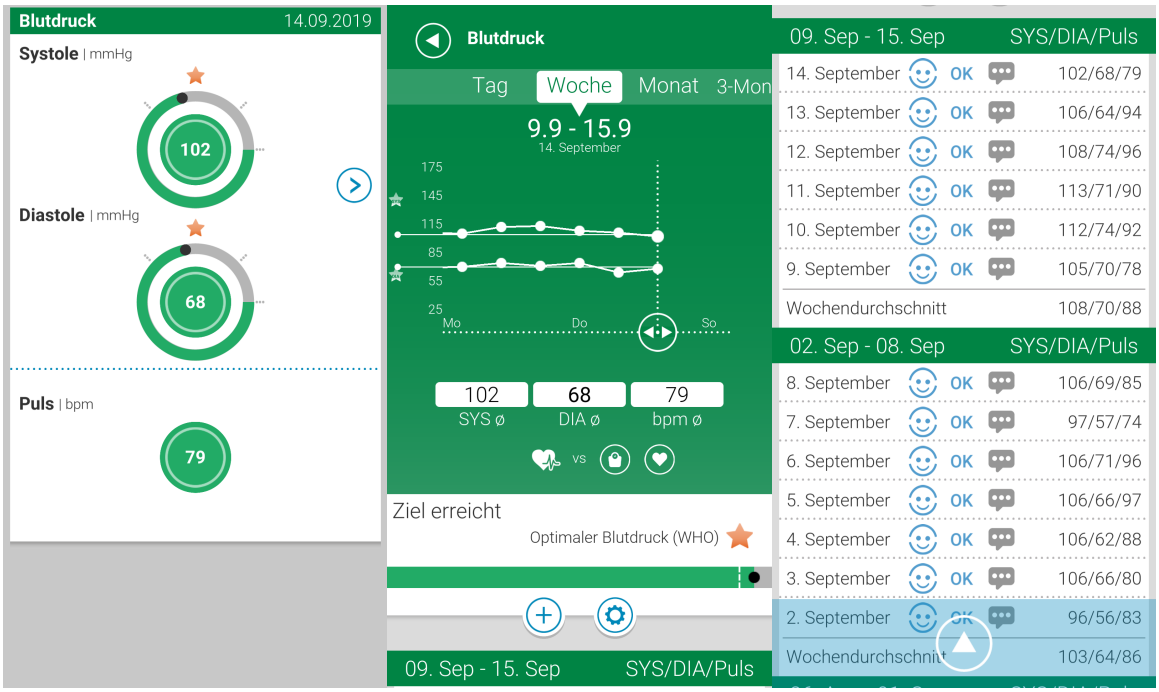

Abbildung 4.3: Medsana-App Blutdrucksübersicht, Graph und tabellarische Ansicht, entnommen aus der MedisanaApp

Weitere Sensoren

Neben den verwendeten Sensoren waren auch eine Reihe anderer Sensoren im

Gespräch, für die Arbeit genutzt zu werden, da sie interessante Werte beisteuern könnten. Dabei könnten diese Daten zur weiteren Kontextualisierbarkeit oder Vollständigkeit der vorhandenen Daten beitragen. Dazu gehören zum Beispiel Luftqualitätssensoren für das Schlaf- und Arbeitszimmer. Im Schlafzimmer könnten sie dazu genutzt werden, den Zusammenhang der Luftqualität auf die Schlafqualität zu beobachten. Dazu wären auch Temperaturinformationen im Schlafzimmer interessant, sowie ein Lautstärkesensor, um zum Beispiel den Einfluss von Straßenlärm, Nachbarn, Schnarchen oder Lärm anderer Art auf den Schlaf zu untersuchen. So könnte zum Beispiel besser untersucht werden, warum der Nutzer aufwacht oder die Schlafphase wechselt. Weitere Aufschlüsse über die Veränderung von Schlafphasen oder den Grund für Wachphasen könnten die Schlafdaten des Partners liefern. Weckt man sich gegenseitig auf, schläft man unruhig, weil der Partner unruhig ist, und weitere Faktoren. Im Arbeitszimmer könnten sehr ähnliche Sensoren interessant sein, um die Konzentrationsfähigkeit, den Grad der Ermüdung und ähnliches zu untersuchen.

Ebenfalls wäre es interessant, zu den Daten regelmäßige Körpertemperaturwerte zu haben, auch hier wäre die Entwicklung einer Normalskala durchaus interessant, um zum Beispiel bei Krankheit besser einschätzen zu können, ob die Körpertemperatur vom zu erwartenden Wert abweicht. Auch die Leistungsfähigkeit und das Wohlbefinden könnten im Zusammenhang mit der Körpertemperatur spannend sein.

Für die Anzeige auf dem Spiegel wäre eine Infrarotkamera interessant, da darüber diverse Informationen generiert und abgenommen werden können. Neben den Emotionen und der Stressentdeckung könnten Trainingserfolge im Spiegel betrachtet werden. Wenn sich also direkt nach einem Training vor den Spiegel gestellt wird, könnte die Infrarotkamera stark durchblutete Muskeln aufzeigen, um dem Nutzer so ein Gefühl davon zu vermitteln, welche Muskeln er gerade trainiert hat. Auch die Körpertemperatur bzw. die Temperatur der Haut könnte dadurch regelmäßig gemessen und zur Analyse herangezogen werden.

Datenextraktion

Der nächste Schritt nach der Erhebung der Daten ist die Extraktion aus den APIs der Hersteller. Wie dabei im Detail vorgegangen wurde, wurde in Abschnitt 3.2 beschrieben. Hier wird darauf eingegangen, was gut und was schlecht war, sowie, was in zukünftigen Arbeiten besser gemacht werden könnte.

Ein Problem an der Fitbi-API ist ihr begrenzter Zugriff, einem Nutzer ist nur dann über die API Zugriff erlaubt, wenn er über E-Mail nachfragt und zumindest vorgibt, es für einen wissenschaftlichen oder ähnlichen Grund zu nutzen. Als die API-Anbindung implementiert wurde, war dies der einzige Weg, an einen Großteil der Daten zu kommen, da der Download auf der Homepage noch zu vernachlässigen war. Mittlerweile

ist dieser besser und umfangreicher geworden und zumindest für historische Daten eine valide Option, da über die API durch die technischen Einschränkungen und von der API gesetzten Anfragebeschränkungen das Downloaden der historischen Daten sehr lange dauert und mit einem hohen Aufwand verbunden ist. Bei der geplanten täglichen Datenabfrage ist eine Abfrage über die API zu empfehlen und von Vorteil. Dabei ist darauf zu achten, so viel wie möglich zu automatisieren. Dies wurde noch nicht umgesetzt, ist aber nach den hier vorgenommenen Einschätzungen machbar.

Die API von Medisana ist für alle Nutzer nutzbar, sofern sie sich dafür eingetragen haben, hat aber ähnliche Anfrageprobleme für große Mengen historischer Daten. Die Homepage unterstützt allerdings schon seit Beginn des Erfassungszeitraums einen guten manuellen Download aller Daten. Dieser wurde hier letztendlich für die historischen Daten genutzt. Für die aktuellen Daten während der Laufzeit des Systems ist eine automatisierte Abfrage über die API zu empfehlen und als umsetzbar eingestuft. Dabei muss darauf geachtet werden, dass die API mit anderen Zeitformaten arbeitet als der manuelle Download und dies weitere Einschränkungen der Datenqualität und einen Mehraufwand beim Bereinigen erzeugt. Die Abfrage über die API wurde implementiert und ist in der vorangegangenen Bachelorarbeit Lüdemann (2016a) genauer beschrieben.

Bereinigung und Verarbeitung

Nachdem die Daten extrahiert wurden, werden sie bereinigt und verarbeitet. Dies geschah zum Großteil nach Bedarf und hinsichtlich der im weiteren Verlauf genutzten Analysen. Dieser Schritt ist enorm aufwändig und von großer Wichtigkeit. Zum einen ist eine gute Datenqualität unabdingbar, um gute Ergebnisse erzielen zu können, und zum anderen benötigen die meisten Analysen die Daten in einem bestimmten Format, um arbeiten zu können. Das heißt, dass die Bereinigung und Verarbeitung sehr sorgfältig und überlegt durchgeführt werden muss. Ein einheitliches Datenformat muss geplant und erzeugt werden, es muss definiert werden, wie mit fehlenden und fehlerhaften Daten umgegangen werden soll. Dafür bietet es sich an, ein Modell dafür zu erstellen, wann Daten als fehlerhaft eingestuft werden. Es muss entschieden werden, ob das Fehlen von Daten eine Information beinhaltet, die verwendet werden soll oder ob die Daten einfach ersetzt werden können und wenn ja, wie.

In dieser Arbeit wurde ein Modell für plausible Daten erzeugt, um mit diesem falsche Daten erkennen zu können, es ist in Abschnitt 2.4.4 beschrieben. Dadurch wurden alle gemessenen Daten als plausibel erachtet, es empfiehlt sich immer, ein Modell zu erstellen. In dieser Arbeit hat das Fehlen der Daten meist einen Grund, seien es Fehler in den Messungen, Urlaub, Probleme mit den Sensoren oder ähnliches. Somit wurde diese Information als wertvoll erachtet und die Datenlücken, die geschlossen wurden, stets so gekennzeichnet, dass sie erkennbar blieben. Auch dies ist zu empfehlen, da

besonders bei Körperdaten das Fehlen immer Rückschlüsse auf das Verhalten des Nutzers, besondere Ereignisse oder Probleme mit den Sensoren geben könnte.

Da die Daten aus unterschiedlichen Quellen unterschiedliche Datumsformate vorweisen, war ein wichtiger Arbeitsschritt, diese Formate zu vereinheitlichen. Dies wurde im Rahmen dieser Arbeit nicht vollständig automatisiert, für weitere Arbeiten ist es aber dringend zu empfehlen, da es gleichzeitig fundamental wie auch aufwändig ist.

Im Rahmen dieser Arbeit wurde auch Aufwand betrieben, der am Ende nicht zu Erkenntnissen über die Daten selbst beigetragen hat, da zur explorativen Analyse auch das Ausprobieren gehört. Dazu gehörte der große Aufwand, die Daten für die InfluxDB zu transformieren. Sofern in weiteren Arbeiten kein deutlicher Nutzen dieser Datenbank und/oder der dazugehörigen Tools nachzuweisen ist, ist davon abzuraten, sie weiter zu verwenden.

Aufgrund der besonderen Natur der Daten gab es einige Punkte, an denen Entscheidungen getroffen werden mussten. Dazu gehört, dass Messungen, die kurz nach Mitternacht vor dem Schlafen erfasst wurden, zum Tag davor zählen und nicht für den Tag gewertet werden, an dem sie erfasst wurden. Des Weiteren haben die Jahre unterschiedlich viele Tage, weil das Jahr 2016 ein Schaltjahr war, das somit einen 29. Februar, also ein Tag mehr, hatte. Dies ist ein Problem für die graphische Gegenüberstellung. Um die Daten gegenüberstellen zu können, hätte somit ein künstlicher Tag in jedes andere Jahr eingefügt werden müssen, der interpoliert oder ausgenullt wird. Hier wurde sich dazu entschieden, den 29.Februar in der Datengegenüberstellung wie in Abbildung 4.6 geschehen, herauszunehmen.

4.2 Datenauswertung und -interpretation

In diesem Abschnitt werden einige Ergebnisse der einzelnen Analyseverfahren aufgezeigt und diskutiert. Darüber hinaus wird das Analyseverfahren an sich diskutiert und evaluiert.

4.2.1 Korrelationsanalyse

Bei der Korrelationsmatrix 4.4 ist zu sehen, in welchem Maße die Werte numerisch korrelieren. Dabei gilt ein Wert von 1 als starkes positives und -1 als starkes negatives Korrelat. Die Werte sind in der untenstehenden Abbildung 4.4 als 1 blau, -1 rot und die Werte dazwischen als farbliche Abstufungen zu sehen. Bei der Analyse gibt es erwartete und unerwartete Korrelate sowie Scheinkorrelate, also Korrelate, die zwar berechnet einen hohen Wert haben, in der Realität aber nicht zusammenhängen. In der unterstehenden Korrelationsmatrix sind in der Diagonalen die Werte alle dunkelblau,

da dies die Korrelation des Wertes x mit dem Wert x ist. An dieser Achse ist die Matrix gespiegelt, daher ist es auch nur notwendig, die Wertbezeichnungen rechts lesen zu können, da die oben-stehenden ihnen entsprechen.

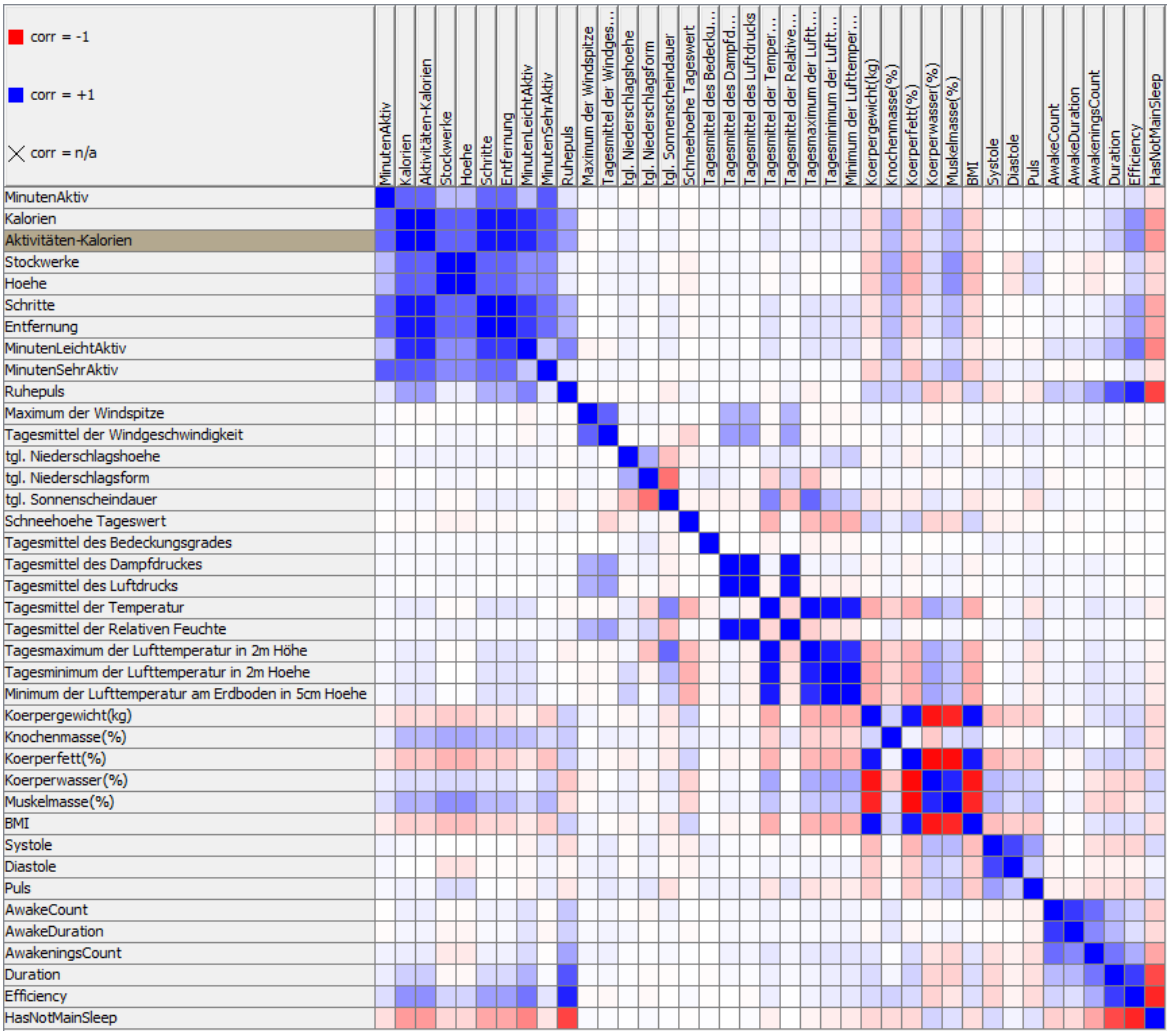

Abbildung 4.4: Korrelationsmatrix aller Werte

Bei der Interpretation des Korrelationskoeffizienten wurde der Maßstab von Chan (2003) verwendet, da dieser sich auf die Auswertung medizinischer und körperlicher Daten bezieht. In der folgenden Tabelle wird aufgezeigt, wie dort die Werte interpretiert werden.

Korrelationskoeffizient ±	Zusammenhang
0	Keiner
0,1 - 0,2	Schwach
0,3 - 0,5	Leicht
0,6 - 0,7	Stark
0,8 - 1	Sehr stark

Tabelle 4.1: Die Interpretation des Korrelationskoeffizienten nach Chan (2003)

Erwartete sinnvolle Korrelate

Erwartete sinnvolle Korrelate sind nützlich um zu sehen, dass die Werte von einer brauchbaren Qualität sind. Sie bestätigen vorhandenes Wissen. Wenn sie nicht auftreten ist zu überprüfen, ob die Werte oder Analysen fehlerhaft sind. Sie liefern allerdings kein neues interessantes Wissen, sondern sind ein guter Indikator für die Validität der Ergebnisse. Einige Beispiele hierfür sind Korrelate innerhalb der Werte aus dem Fitbit, zum Beispiel Entfernung und Schritte. Sie korrelieren mit einem Wert von 0.999, dies ist erwartet, da das Fitbit die zurückgelegte Entfernung unter anderem durch die Schritte festlegt. Hier sollen einige erwartete Korrelate exemplarisch gezeigt werden.

Weitere erwartete Korrelationen sind zum Beispiel *Aktivitäten-Kalorien* und *Kalorien*, selbstverständlich steigt der Kalorienwert(am Tag verbrauchte Kalorien), wenn mehr Kalorien in Aktivitäten verbraucht werden. Sehr ähnlich verhält es sich mit *Minuten Leicht Aktiv* und *Aktivitäten-Kalorien* bzw. Kalorien da sich die *Aktivitäten-Kalorien* aus den Aktivitäten ergeben, also aus den *Minuten Leicht Aktiv*, *Minuten Aktiv* und *Minuten Sehr Aktiv*, es ist also zu erwarten, dass diese Werte positiv zusammen hängen. Eine weitere erwartete Korrelation findet sich zwischen der Höhe und den Stockwerken, diese hängen stark zusammen, da Stockwerke ein aus der Höhenmessung berechneter Wert ist.

In der Tabelle werden einige erwartete Korrelate und ihre Werte aufgeführt.

Wert 1	**Wert 2**	**Korrelations-koeffizient**
Schritte	Entfernung	0.999
Kalorien	Aktivitäten-Kalorien	0.994
Kalorien	Minuten Leicht Aktiv	0.833
Aktivitäten-Kalorien	Minuten Leicht Aktiv	0.865
Höhe	Stockwerke	1
Aktivitäten-Kalorien	Efficiency(Schlaf)	0.437
Körpergewicht	Körperfett	0.928
Körpergewicht	Muskelmasse	-0.861

Tabelle 4.2: Korrelationskoeffizient bei erwarteten Werten

Erwartete aber nicht nachgewiesene Korrelate

Es gab Korrelate, die aufgrund der Dateneigenschaften erwartet wurden, aber in der Korrelationsmatrix nicht zu sehen sind oder in einem viel geringeren Ausmaß als gedacht auftreten. Dies kann unterschiedliche Gründe haben, es könnte sein, dass es kein Korrelat gibt oder die Werte nicht in geeigneter Qualität, Granularität oder Masse vorliegen. Ebenso kann es sein, dass sie mit einem Zeitdelta korrelieren. Unterschiedliche Werte haben divergierende Reaktionszeiten. So kann zum Beispiel Puls und Blutdruck relativ direkt auf veränderte Umstände reagieren, während das Gewicht und der Ruhepuls nur langsam reagieren und sich verändern.

Die untenstehende Tabelle 4.3 zeigt die erwarteten Korrelate mit ihren überraschend niedrigen Ergebnissen.

Eine zu erwartende Korrelation, die auftreten konnte, war, dass die Anzahl der Schritte am Tag steigt, wenn mehr Sonnenstunden am Tag auftreten oder die Temperatur steigt. Also, dass sich mehr bewegt wird, wenn das Wetter gut ist, dies ist den Daten jedoch so nicht zu entnehmen, die Korrelation liegt unter bzw. knapp über 0.1 und ist damit laut der Bewertung aus Tabelle 4.1 als schwach zu werten. Ein weiteres Korrelat, das nicht in dem Maße auftrat, wie es erwartet wurde, ist die sportliche Aktivität im Zusammenhang mit dem Körpergewicht oder der Muskelmasse, auf das Gewicht hat die Aktivität nur eine verschwindend geringe Auswirkung, auf die Muskelmasse immerhin ein wenig.

Wert 1	**Wert 2**	**Korrelations-koeffizient**
Ruhepuls	Aktivitäten-Kalorien	0.38
Kalorien	Körpergewicht	-0.152
Aktivitäten-Kalorien	Körpergewicht	-0.138
Minuten Aktiv	Körpergewicht	-0.073
Aktivitäten-Kalorien	Muskelmasse	0.288
tgl. Sonnenscheindauer	Aktivitäten-Kalorien	0.037
tgl. Sonnenscheindauer	Schritte	0.059
Tagesmaximum der Luft-temperatur in 2m Höhe	Schritte	0.112

Tabelle 4.3: Korrelationskoeffizient bei erwartet aber nicht nachgewiesen

Nicht erwartete Korrelate

Die Werte, bei denen der Korrelationskoeffizient hoch ist, bei denen es aber nicht erwartet wurde, sind die interessantesten Werte, da hier neue Erkenntnisse möglich sind. Dabei muss entschieden werden, ob es einen kausalen Zusammenhang geben kann oder ob es wahrscheinlicher ist, dass die Werte durch Zufall einen hohen Koeffizienten haben. In der untenstehenden Tabelle 4.4 ist aufgezeigt, welche Werte unerwartet korrelieren.

Das Ergebnis der *Aktivitäten-Kalorien* mit dem Ruhepuls ist doppelt überraschend, da es zum einen nur eine leichte Korrelation gibt, wobei man davon ausgehen könnte, dass ein sportlicherer Lebensstil den Ruhepuls verändert und zum anderen korreliert er anders herum als gedacht. Obwohl man davon ausgeht, dass mehr Sport den Ruhepuls senkt, da sich der Kreislauf und das Herz stärken, ist es hier anders herum. Mehr Sport führt zu einem leichten Anstieg des Ruhepulses. Dies ist zum einen ein Hinweis darauf, dass überprüft werden könnte, ob die Korrelation mit einem zeitlichen Versatz stärker wird, da der Ruhepuls sich nicht stark ändert. Zum anderen ist es ein Hinweis darauf, dass die Grundannahme vielleicht zu überdenken ist. Zusammen mit dem Wissen, dass der Ruhepuls recht niedrig ist, [18] kann die Vermutung aufgestellt werden, dass es daraufhinweist, dass mehr Sport beim Nutzer dazu führt, dass das Herz schneller schlägt und dichter an den normal Bereich heranSkommt. Erst wenn der Körper sehr sportlich ist, wird der Ruhepuls vermutlich wieder sinken. Dies ist nicht als allgemeine Entdeckung zu sehen, sondern auf den Körper des Nutzers bezogen.

[18] Optimaler Ruhepuls einer Person mit dem Alter, Geschlecht und Fitnessstand des Nutzers liegt bei 73-76 bpm PulsNormWerte. Der hier gemessene Durchschnitt liegt bei 57 bpm und damit außerhalb des normalen Bereichs auf der Skala.

Der Wert *HasNoMainSleep* gibt an, ob es einen Schlaf gibt, der nicht der Nachtschlaf ist, also zum Beispiel Mittagsschlaf. Im Zusammenhang mit diesem Wert gibt es eine hohe Korrelation zum Ruhepuls, der im Nachtschlaf gemessen wird. Das heißt, wenn Mittagsschlaf gehalten wird, steigt der Ruhepuls und er sinkt, wenn keiner gehalten wird. Die Datenbasis ist an dieser Stelle allerdings recht dünn, da in dem Datensatz von 1300 Nächten nur 176 mit vorangegangenem Mittagsschlaf sind. Des Weiteren kann sich durchaus eine Unschärfe und Übertreibung in den Korrelaten abzeichnen, da der *HasNoMainSleep* Wert ein boolescher ist und alle anderen Werte Integer mit deutlich größeren Wertebereichen sind.

Ein weiterer interessanter Punkt ist das starke Korrelat zwischen der Schlafeffizienz und dem Ruhepuls, dieses liegt mit 0.827 bereits im Bereich der sehr starken Korrelationen. Es ist zudem verwunderlich, dass es eine positive Korrelation ist, also dass ein längerer Schlaf in einem höheren Ruhepuls resultiert.

Wert 1	**Wert 2**	**Korrelations-koeffizient**
Aktivitäten-Kalorien	Ruhepuls	0.379
Stockwerke	Muskelmasse	0.434
Ruhepuls	HasNoMainsSleep	-0.738
Ruhepuls	Efficiency(Schlaf)	0.827

Tabelle 4.4: Korrelationskoeffizient bei nicht erwarteten Werten

Evaluation der Analysemethode

Die Korrelationsanalyse wurde in KNIME realisiert und war nach dem Bereinigen und Transformieren der Daten sehr einfach umzusetzen. Die Auswertung hingegen ist aufwändiger, da neben der akkuraten Sichtung der Ergebnisse auch eine Interpretation nötig ist, die ein gewisses Domänenwissen voraussetzt. Mit diesem Wissen muss bestimmt werden, ab wann ein Korrelat interessant ist und was es aussagen könnte. Durch die Betrachtung der erwarteten Korrelate ist diese Analyse ein gutes Mittel, um die Glaubwürdigkeit der Daten zu prüfen. Sind zu viele der erwarteten Korrelationen nicht vorhanden, ist dies ein guter Hinweis darauf, dass die Daten fehlerhaft, zu ungenau oder falsch sein könnten. Darüber hinaus können unerwartete Korrelate auftreten, die nach ausgiebiger Prüfung Erkenntnisse bringen. Dadurch ist diese Analyse zu empfehlen. Es könnten einzelne Phasen des Erhebungszeitraums näher betrachtet werden, um zu untersuchen, wie sich die Korrelationskoeffizienten verändern, wenn Ausschnitte von einem Jahr oder Monat gewählt werden, da sich aus den Ergebnissen zeigte, dass die Jahre sich durchaus unterscheiden. Korrelationen könnten sich aufgrund von einem veränderten Körper- und Gesundheitszustand also auch ändern.

4.2.2 Clusteranalyse

Die Daten wurden, wie in 3.5 beschrieben, mit verschiedenen Clusteralgorithmen in diversen Durchläufen analysiert, um herauszufinden, ob sie Cluster bilden. Dabei wurden diverse Durchläufe unternommen, um zum einen möglichst klaren Cluster zu definieren und zum anderen, um herauszufinden, nach welchen Attributen das Verfahren die Cluster bildet. Dabei ist bei dem Großteil aller Durchläufe kein klares Ergebnis entstanden, die meisten Clusteralgorithmen können auf diesen Daten keine Cluster finden. Für sie wäre es besser, wenn die Daten noch einmal explizit für die Verfahren transformiert und ausgewählt werden. Das heißt, es müssten mehr Durchläufe mit Teilmengen der Attribute durchgeführt werden. Der k-Medoids Clusteralgorithmus findet dahingegen Cluster. Bei einem k von 6, also sechs Medoiden, die die Zentren der Cluster bilden, zeichnen sich sechs Cluster ab. Die untenstehende Abbildung 4.5 zeigt die Cluster, die zu sehen sind, wenn Schritte und Kalorien aufgetragen werden. Dabei sind die verschiedenen Farben die jeweiligen Cluster und die Symbole die Wochentage. Dies ist zwar ein besseres Ergebnis als bei den anderen Algorithmen, jedoch definieren sich die Cluster primär durch die Anzahl der Schritte, die gegangen wurden, während andere Attribute eher vernachlässigt werden. Somit ist auch dieses Ergebnis eher zu vernachlässigen. Die Daten sind so wie sie vorliegen nicht sinnvoll zu clustern, es müssten deutlich erweiterte und tiefere Tests durchgeführt werden, um eventuell nützliche Ergebnisse zu erhalten.

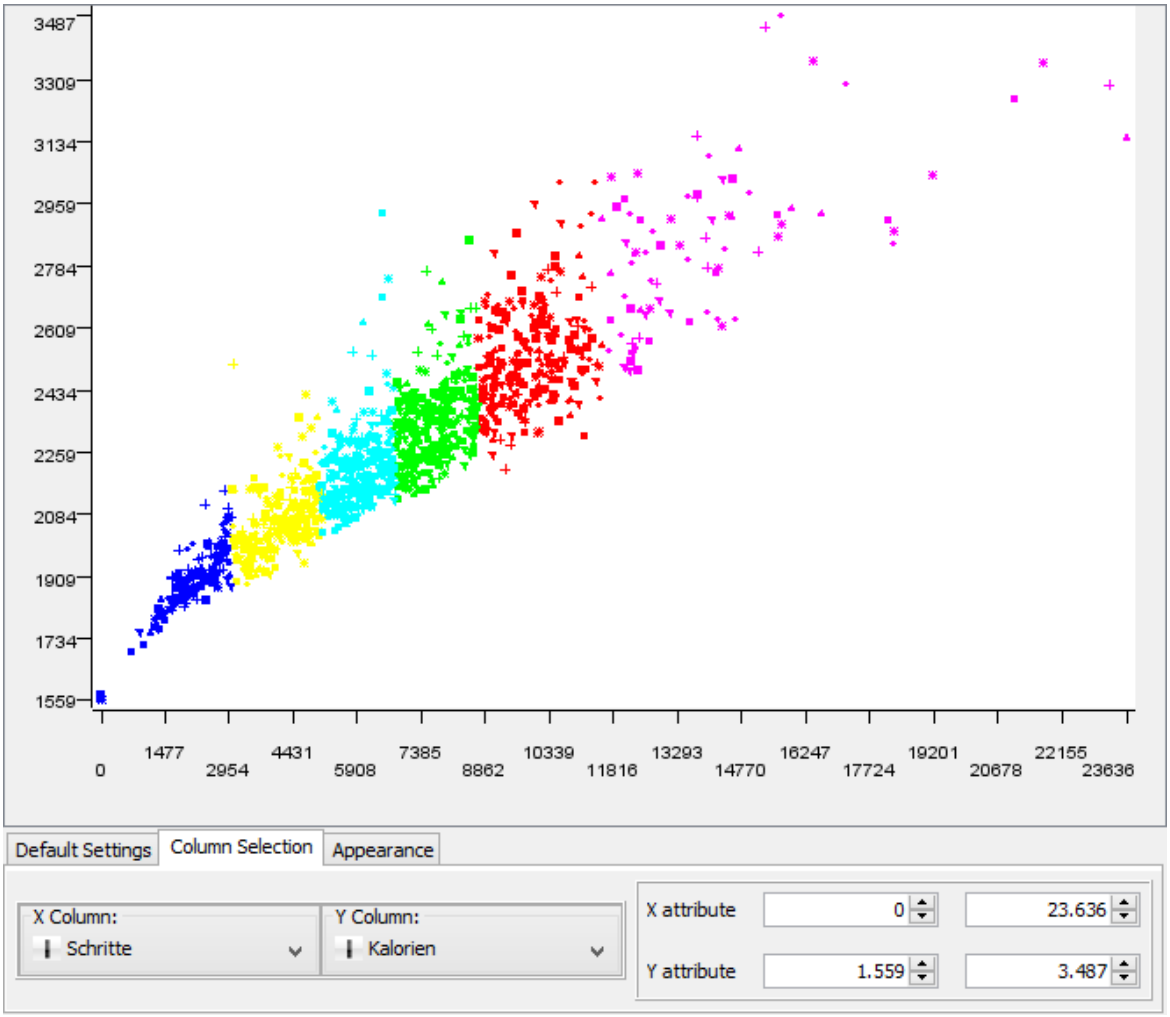

Abbildung 4.5: K-Medoids Clusterübersicht

Evaluation der Analysemethode

Die Clusteranalyse hat in den hier durchgeführten Versuchsdurchläufen keine neuen Erkenntnisse ermöglicht. Allerdings ist die Umsetzung durch KNIME relativ einfach. Der Versuchsaufbau ist, sobald die Daten bereinigt und transformiert in KNIME vorliegen, angenehm schnell realisiert. Der eigentliche Aufwand entsteht bei den Versuchsdurchläufen, deren Parameter angepasst werden müssen, um dann die Ergebnisse durchzugehen und zu interpretieren. Im Rahmen dieser Arbeit war die Clusteranalyse kein Erfolg, dies könnte durch einen Mehraufwand und einen erhöhten Fokus auf die Analyse ggf. verändert werden. Durch gut gewählte Untermengen und Datenforma-

te könnten die Daten so vorbereitet werden, dass sie der Clusteranalyse eine bessere Grundlage bieten.

4.2.3 Explorative Datenanalyse

Die Daten wurden explorativ analysiert und untersucht, um Auffälligkeiten, Besonderheiten und Zusammenhänge durch Beobachtung und Visualisierung zu finden und aufzuzeigen. Dabei wurden zum Teil auf Verdacht hin verschiedene Datenstränge auf verschiedene Arten visualisiert, um Eigenschaften sichtbar zu machen und zeitliche Zusammenhänge zu erkennen. Einige der dabei entstandenen Visualisierungen, die als interessant eingestufte Erkenntnisse liefern, werden hier gezeigt und beschrieben.

Bei der Gegenüberstellung des Gewichts des kompletten Messzeitraums ist zu sehen, dass das Gewicht einen deutlichen Einknick in allen Jahren, in denen es um den August herum erfasst wurde, erfährt, siehe Abbildung 6.5. Das Jahr 2016 ist leider in den Gewichtswerten sehr unvollständig und zu großen Teilen interpoliert, so finden sich im August bloß zwei Messungen. In den Jahren, in denen die Messungen deutlich öfter bis täglich erfolgen, ist aber gut zu sehen, dass der Wert um den 28.08-30.08 deutlich bis extrem abnimmt. Um dafür eine Erklärung zu finden, wurde der Kalender zur Rate gezogen, in dem neben den Terminen auch Notizen vorhanden sind. Dabei ergibt sich, dass die stärkste Veränderung im Jahr 2015 durch einen starken Stressfaktor zu erklären ist. In den anderen beiden Jahren ergibt sich im Kalender kein eindeutiger Hinweis darauf, woher diese starke Veränderung kommt. Gleichsam ist zu sehen, dass im Allgemeinen das Gewicht im Winter höher ist als im Sommer. Die untenstehende Abbildung 4.6 zeigt den Verlauf der Messung über das ganze Jahr zur besseren Übersicht. In der Abbildung 4.7 wurde der auffällige Abschnitt rund um den August vergrößert dargestellt.

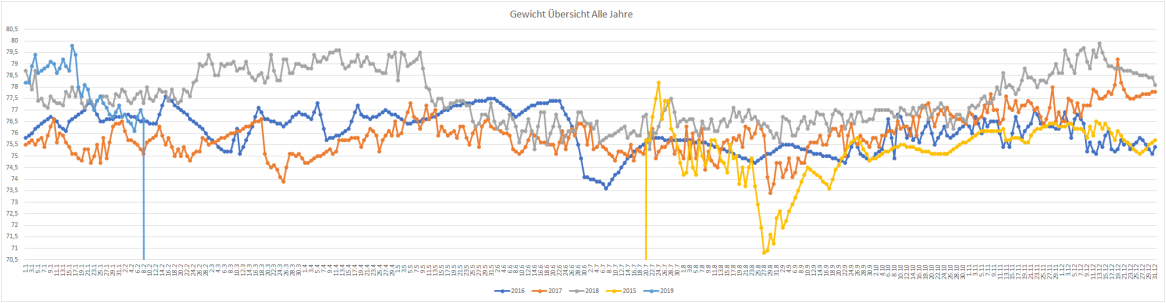

Abbildung 4.6: Gegenüberstellung aller Jahre, Gewicht in kg

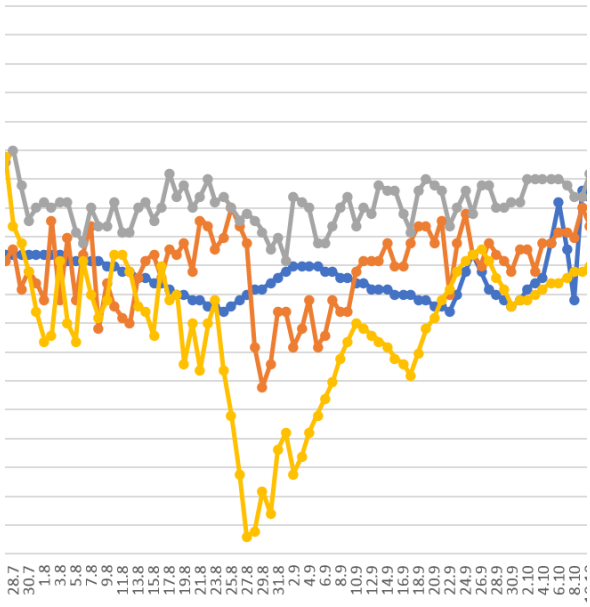

Abbildung 4.7: Gegenüberstellung des Gewichts aller Jahre Ausschnitt August

Eine andere Art der Visualisierung wurde angewendet, um aufzuzeigen, wie die Bewegung an unterschiedlichen Wochentagen variiert. Die sich im Anhang befindende Abbildung 6.1 zeigt einen Scatter Plot mit den *Minuten Sehr Aktiv* Wert pro Wochentag für alle Jahre. Dabei ist zu sehen, dass nicht nur die Spitzen schwanken, sondern auch die Basis, also die Werte mit 0 aktiven Minuten. Am Wochenende ist diese Basis viel breiter als zum Beispiel am Mittwoch. Das heißt, es gibt viel mehr Tage am Wochenende, an denen keinerlei sportliche Aktivität stattfand. Dabei ist Mittwoch scheinbar der Tag mit der meisten regelmäßigen sportlichen Betätigung. Betrachtet man die Ausreißer, gibt es Tage mit einer sehr hohen Aktivität, drei Tage zeigen über 90 Minuten hohe Aktivität, also einen Puls über 165. Dies entspricht eineinhalb Stunden Höchstleistung und erscheint sehr viel. Diese Art der Visualisierung eignet sich, um Ausreißer schnell entdecken und überprüfen zu können. Zusammen mit anderen Datenpunkten kann mithilfe des Kalenders bestimmt werden, ob es sich um einen fehlerhaften Wert handelt oder ob er valide scheint.

Neben Auffälligkeiten in den Daten sollte aufgezeigt werden, was eigentlich 'normal' ist. Die WHO gibt zwar vor, welche Werte für Menschen in welchem Alter als 'normal' oder 'gesund' gelten, jedoch ist aufgefallen, dass die erhobenen Werte von den Angaben der WHO teilweise sehr deutlich abweichen. Daher wurde für jeweils ein Jahr der Median bzw. Mittelwert eines Messwertes genommen und die jeweilige Standardabweichung dazu gerechnet, abgezogen und dies jeweils als Linie dargestellt. In der untenstehenden Abbildung zeigt sich somit, was im Jahr 2017 als 'normal' für die Systole gemessen wurde. Normal wird hier definiert als der, in der Gesamtheit der Daten

häufigste Werte, also unabhängig von einer Expertenmeinung bezogen auf die Datenlage. Dabei stellt sich heraus, dass der Normalbereich der Systole zwischen 104 mmHg und 112 mmHg liegt. Laut WHO liegt ein normaler systolischer Blutdruck bei 120-129 mmHg BlutdruckNachWHO. Dabei wird ebenfalls angegeben, dass ein optimaler Blutdruck unter 120 mmHg liegen sollte. Somit ist der gemessene Blutdruck durchaus als optimal zu bezeichnen. Dabei ist es aber auch wichtig, sich die Ausreißer bzw. die Höchstwerte anzuschauen. So erkennt man in Abbildung 4.8 am 23.09.2017 einen Wert, der stark erhöht ist und weit aus dem Normalbereich steigt. Dieser Wert liegt bei 126 mmHg und damit noch im Bereich des, laut WHO normalen, Blutdrucks. Diese Werte werden allerdings als extrem unangenehm empfunden und sind körperlich als zu hoher Blutdruck zu fühlen, sie teilen sogar Symptome einer Hypertonie, also eines zu hohen Blutdrucks. Diese sind unter anderem Schwindel, Ohrensausen(Tinnitus) und Übelkeit BlutdruckSymptome. Darüber hinaus wird ein extremes Unwohlsein empfunden. Daraus schließt sich, dass dieser Bereich als 'nicht gut' einzuschätzen ist. Zur besseren Übersicht ist die Abbildung im Anhang 6.4 vergrößert dargestellt.

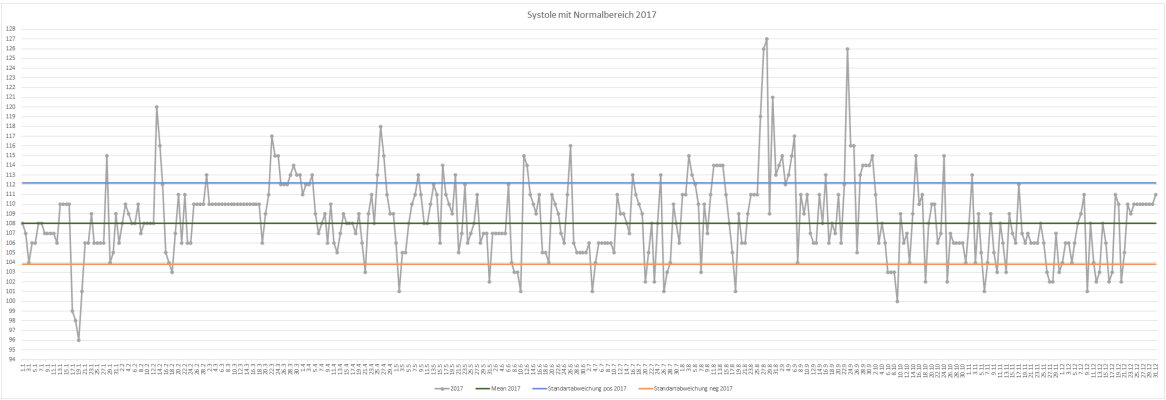

Abbildung 4.8: Systole 2017 mit Normalskala

Die Normalskala wurde für jedes Jahr einzeln aufgezeichnet, da sich schnell abzeichnete, dass sich die Mittelwerte über die Jahre veränderten. Beim Ruhepuls zum Beispiel ist der Mittelwert über den gesamten Zeitraum 57,4, die einzelnen Jahre, beginnend bei 2015, listen sich jedoch wie folgt auf: 63,3; 58,5; 57,8; 60,1; 57,5 [19]. Bei einem so kleinen Wertebereich sind auch kleine Veränderungen schon ausschlaggebend. Daher wurde eine jährliche Normalskala gebildet, um die Skala und Ausreißer genauer bestimmen zu können. Darüber hinaus könnten auch monatliche Skalen gebildet werden, da dies noch genauer wäre.

[19] Hier fällt vielleicht auf, dass die Werte der einzelnen Jahre höher sind als der Durchschnitt aller Jahre. Dies ist ein Beispiel für das Simpson-Pardoxon und kein Berechnungsfehler.

Bei der Betrachtung der Graphen, die dabei entstanden, die Normalskala der Werte aufzuzeigen, fiel auf, dass sich Gewicht und Blutdruck in gewissen Punkten auffällig ähnelten. Die Korrelationsmatrix zeigte jedoch einen eher niedrigen Korrelationskoeffizienten von 0.25. Da Gewicht und Blutdruck unterschiedlich schnell schwanken können, wurde eine Visualisierung erstellt, in der das Gewicht um drei Tage nach vorne gezogen wurde. Diese zeigt die untenstehende Abbildung 4.9. Zu sehen sind die besonders auffälligen Teile vom 8.03-10.04 und 20.08-10.09, bei denen sich die beiden Werte recht stark gegenteilig gegenüberstehen. Es zeigen sich auch immer wieder kleinere Bereiche, in denen dieses Verhalten zu sehen ist. Es ist dabei aber keine generelle Tendenz. In großen Zeitabschnitten bewegen sich die Werte relativ unabhängig. Auch hier wurde der Kalender zu Rate gezogen, um eine Idee davon zu entwickeln, woher diese Auffälligkeiten in den Daten kommen. Zum einen kann man gut erkennen, dass die Daten zwischen dem 02.02 bis zum 16.03 interpoliert sind, da dort aufgrund eines Urlaubs keine Daten erhoben wurden. Danach ergibt sich im Kalender kein eindeutiger Grund, der diese Auffälligkeit erkläre könnte. Es ist zu vermuten, dass es mit den starken Rückenbeschwerden und darauf folgende Belastung durch Stress zurückzuführen ist. Bei der zweiten Anomalie ist dahingegen ein Stressfaktor klar erkennbar und der Grund für die starke Gewichtsabnahme und die Steigung des Blutdrucks. Zur besseren Leserlichkeit ist die Abbildung im Anhang 6.6 vergrößert angefügt.

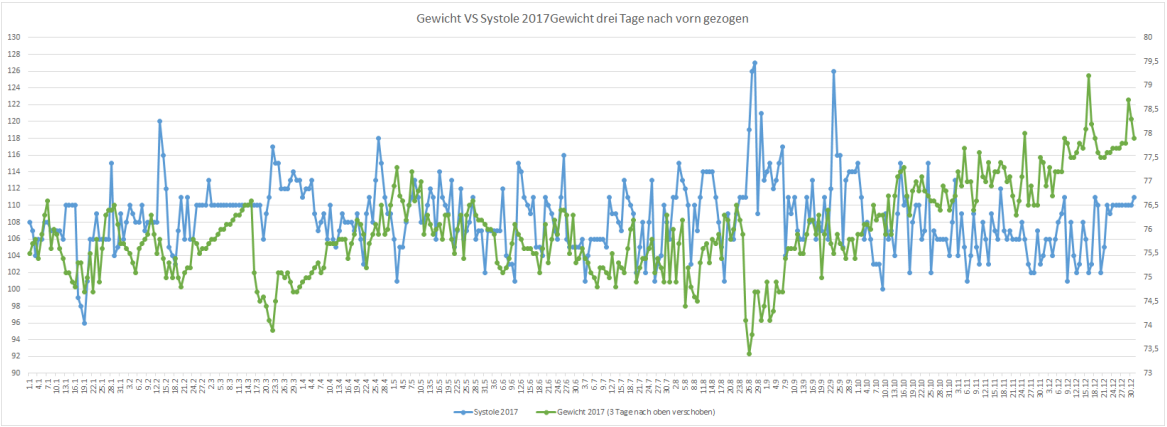

Abbildung 4.9: 2017 Gewicht vs Systole, drei Tage gegeneinander verschoben

Neben der leicht versetzten Reaktion des Gewichts auf Umstände, die den Blutdruck verändern, gibt es auch andere Werte, die eventuell voneinander abhängen aber unterschiedlich schnell reagieren. So wurde aufgrund der Vermutung, dass Sport den Ruhepuls verändert, auch die Aktivität in Form des Wertes *Minuten Leicht Aktiv* mit dem Ruhepuls betrachtet. Da die *Minuten Leicht Aktiv* sehr schwanken und einen recht großen Wertebereich im Vergleich zum Ruhepuls haben, mussten sie geglättet

werden, um ein übersichtliches Diagramm zu ergeben. Bei genauer Betrachtung ist zu sehen, dass bei steigender Bewegung auch der Ruhepuls steigt. Dabei reagiert der Ruhepuls oft ein wenig versetzt. Steigt der Aktivitätslevel für einige Zeit, so braucht der Ruhepuls ein paar Tage, um auch zu steigen. Gleichsam sinkt der Puls wieder, wenn der Aktivitätslevel absinkt. Dies erscheint auf den ersten Blick seltsam, da sie dem allgemeinen Konsens, wer Sport treibt hat einen niedrigeren Ruhepuls, widerspricht PulsNormWerte. In Abschnitt 4.2.1 wurde ein Erklärungsversuch unternommen, warum die Werte sich hier so verhalten. Das hier besprochene Diagramm 4.10 unterstützt dabei die Annahme. Auch diese Abbildung ist vergrößert im Anhang 6.7 zu finden.

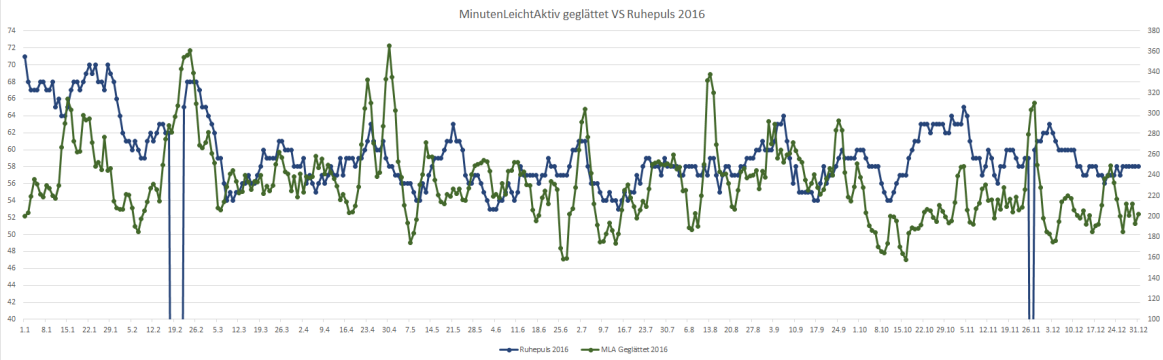

Abbildung 4.10: Minuten Leicht Aktiv versus Ruhepuls 2016

In einem Scatterplot wurde die Verteilung des Ruhepulses aller Jahre gegenübergestellt. Dabei ist gut zu erkennen, wie er sich im Laufe der Jahre verändert. Besonders abweichend erscheint hier das Jahr 2015, aus dem leider erst ab dem siebten Monat Daten vorliegen. Die genauen Gründe für das Verschieben über die Jahre sind aus den Daten nicht entnehmbar aber hängen vermutlich mit dem Bewegungs- und Stresslevel zusammen. Auffällig ist weiterhin, dass es in jedem Jahr immer einen nennenswerten Anteil an Werten gibt, die außerhalb eines gemeinhin als 'normal' bezeichneten Ruhepulses liegen, der zwischen 60 und 80 Schlägen pro Minute liegen würde. Der gemessene Ruhepuls liegt stellenweise bei 52 Schlägen pro Minute und dies, obwohl keineswegs von einer Sportlichkeit auszugehen ist, bei der solche Werte erwartet werden würden. Tatsächlich ist es eher so, dass in den Jahren, in denen mehr Sport getrieben wird, viel häufiger hohe Werte gemessen werden. Dies unterstützt die Annahme, die in Abschnitt 4.2.1 getroffen wurde.

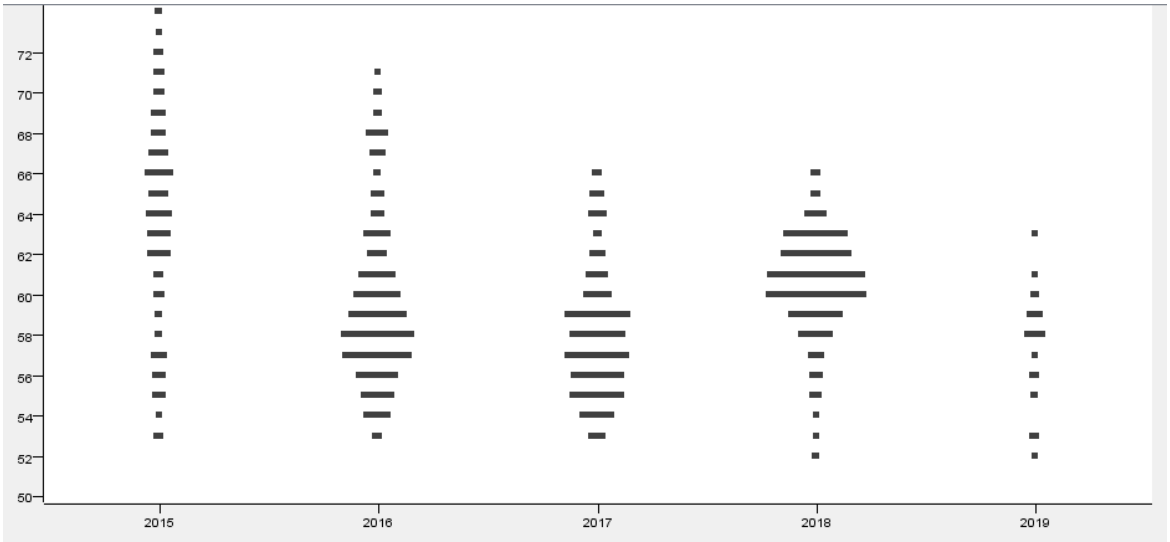

Abbildung 4.11: Ruhepuls Vergleich alle Jahre

Die Veränderung der Sportlichkeit verdeutlicht die sich im Anhang befindende Abbildung 6.2, in der zum Vergleich aufgetragen ist, wie viele *Minuten Sehr Aktiv* pro Jahr gemessen wurden. *Minuten Sehr Aktiv* beschreiben die Minuten, in denen der Nutzer einen sehr hohen Puls bei sportlichen Aktivitäten hat. Zu beachten ist dabei, dass das Jahr 2015 nur mit etwas mehr als fünf Monaten Messung vorhanden ist und somit nur ein Eindruck davon gewonnen werden kann, wie es im kompletten Jahr ausgesehen hat. Selbst mit dem Ausschnitt den die Daten liefern ist zu sehen, dass im Jahr 2015 weit häufiger und viel länger ein so hoher Puls erreicht wurde, sodass er als sehr aktiv gilt. Über die Jahre ist zu sehen, wie die sportliche Betätigung immer weiter abnimmt. Auch in dieser Abbildung sind die Ausreißer nach oben sehr gut zu sehen, sodass die Abbildung genutzt werden kann, diese zu untersuchen.

In einem Balkendiagramm aus KNIME wurde aufgezeigt, wie sich die Monate in Bezug auf die durchschnittlich gegangenen Schritte unterscheiden. Dabei fällt schnell auf, dass im Februar am wenigsten gegangen wird und im Mai am meisten. Insgesamt ist der Durchschnitt unter der Empfehlung von 10000 Schritten am Tag. Dies ist nicht verwunderlich, da durch die Arbeit und das Studium sehr viel gesessen wurde und kein Augenmerk darauf lag, diese Grenze zu erreichen. Es wurde erwartet, dass sich deutlich abzeichnet, dass im Winter weniger gegangen wird als im Sommer, dies bildet sich aber nur geringfügig ab. So wurden im November fast so viele Schritte gegangen wie im Juli. Da die Achsenbeschriftung hier nicht zu lesen ist, wurde die Abbildung vergrößert dem Anhang beigefügt 6.8 und hier zur Übersicht gezeigt, da die Bewegung der Balken dennoch verdeutlicht, wie die Schrittzahlen schwanken.

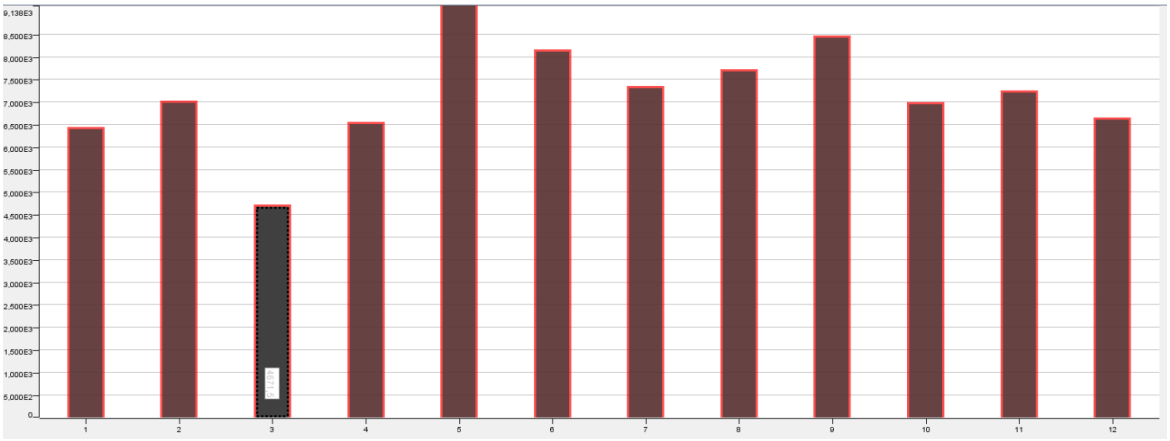

Abbildung 4.12: Schritte, Durchschnitt pro Monat

Um noch einmal zu verdeutlichen, wie sensibel Körperdaten sind, wird hier ein Beispiel gegeben, wie aus den Gewichtsmessungen eine ganz andere sensible Information gewonnen werden kann. Besonders in den ersten Jahren der Datenerfassung gibt es in den Gewichtsdaten immer wieder fehlende Daten. Dies ist dem Umstand geschuldet, dass die Waage ein relativ stationärer Sensor ist. Während das Handgelenkblutdruckmessgerät immer mitgenommen werden konnte, stand die Waage im eigenen Badezimmer. So kann im Datenbestand sehr gut nachvollzogen werden, wann außerhäusig und wann zuhause geschlafen wurde. Das heißt, dass allein aus der Tatsache, ob ein Wert erhoben wurde oder nicht mit ein wenig Metawissen ein grobes Bewegungsprofil erstellt werden kann. So können aus scheinbar harmlosen Daten wie der Erfassungshäufigkeit, der Uhrzeit und ähnlichen Metadaten über die Werte schon sensible Informationen generiert werden. Es müssen nicht einmal die Messwerte selbst übermittelt werden, damit vermeintliche Dritte daraus Informationen generieren können, die ihnen eigentlich nicht übermittelt werden sollten.

Darüber hinaus kann dieses Mehrwissen auch für diese Arbeit genutzt werden, da diese Information herangezogen werden könnten, um Schlafdaten danach zu klassifizieren, wie sich die Schlafsituation ändert. Ab einem gewissen Punkt in den Daten wurde die Waage bewegt, ab diesem Tag werden die Messungen deutlich häufiger und regelmäßiger. Bemerkenswert ist dabei, dass auf den Tag genau bestimmt werden kann, wann der Standort der Waage verändert wurde.

Evaluation der Analysemethode

Die explorative Datenanalyse hat im Rahmen dieser Arbeit einige Erkenntnisse erbracht. Dabei ist sie aber sehr aufwändig und benötigt immer neue Umsetzungen. Je

nach Fragestellung können schnelle Anpassungen zum Erfolg führen aber auch sehr aufwändige Anpassungen der Datenmenge, Formate oder des Ausschnitts notwendig sein. Darüber hinaus kann die Berechnung neuer Werte schwierig sein. Im Rahmen dieser Arbeit wurde nicht jede Visualisierung systematisch durchgeführt, sondern für Werte, die als potentiell relevanter als andere eingestuft wurden. Diese Visualisierungen und Auswertungen könnten durchaus auf den gesamten Datenbestand ausgeweitet werden, um weitere Erkenntnisse zu erlangen oder bestehende mit tieferer Grundlage zu versehen. Einige der Visualisierungen aus KNIME sind zwar einfach zu realisieren, da aber die Visualisierung selbst oft nicht so übersichtlich und praktikabel ist, könnten diese Teile in einem anderen Programm nachgebildet werden, um bessere Visualisierungen zu erhalten. Dies betrifft zum Beispiel die Visualisierungen, die Monate, Jahre oder Tage beinhalten, da diese oft nicht in der richtigen Reihenfolge sind, was für eine intuitive Exploration sehr nachteilig ist. Dies scheint auch nicht trivial lösbar zu sein. Darüber hinaus wurden die Möglichkeiten der Visualisierung in KNIME nicht ausgeschöpft. Der Großteil der Ergebnisse stammt aus Gegenüberstellungen von Daten in Excel. Dazu gehörten das Gegenüberstellen von den jeweiligen Jahren eines Wertes sowie die Aufzeichnung eines Jahres mit verschiedenen Werten. Dabei ist das Erstellen der Visualisierungen immer mit einem gewissen Aufwand verbunden, zudem muss jede Gegenüberstellung sorgsam studiert und betrachtet werden, was den eigentlichen Aufwand ausmacht.

4.2.4 Weitere Analyse

Neben den hier durchgeführten Analysen soll kurz darauf eingegangen werden, welche Berechnungen oder Analysen noch durchgeführt werden könnten, da sie von Interesse sein könnten. Dabei werden hier nur einige Beispiele aufgeführt.

Zum einen könnte der Erholungspuls berechnet werden, da dieser ein Hinweis auf den Grad der körperlichen Fitness ist. Er definiert sich darüber, wie schnell der Puls nach einer körperlichen Anstrengung wieder sinkt. Je schneller der Puls wieder auf den normalen Puls sinkt, umso trainierter ist das Herz. Darüber hinaus könnten die Daten stärker klassifiziert werden, angefangen mit einfachen Regeln der Klassifikation könnten sich daraus komplexe Regeln entwickeln, die dazu Erkenntnisse erbringen. So könnten gute, sportliche, schlechte, arbeitsame oder freie Tage klassifiziert werden. Aber auch einzelne Aspekte wie zum Beispiel: Was ist gutes Wetter? Wann ist ein Ausreißer in den Daten? Reicht dafür ein einzelner, auffälliger Wert oder ist es erst ein Ausreißer, wenn bestimmte andere Werte ebenfalls auffällig sind? können klassifiziert werden.

Zur besseren Ansicht der Veränderungen über Zeit könnte die erste Ableitung der Daten gebildet und angezeigt werden. Darüber hinaus könnten mehr Metainformationen herausgearbeitet werden, um die Natur der Daten besser abbilden zu können,

dazu gehören Extrembereiche, die den Normalbereich umgeben, Fehlerhäufigkeiten, Informationen darüber, wie viel Prozent der Werte im Normalbereich liegen und ähnliches.

4.3 Datenqualität

Die Datenqualität variiert zwischen den einzelnen Sensoren und war, wie zu erwarten, nie von einer Güte, in der sie sofort hätte verarbeitet werden können. Ebenso war zu erwarten, dass die Sensoren keineswegs medizinisch genaue Daten liefern. Die Sensoren sind zum einen vom Werk aus zu einem gewissen Grad ungenau, wie schon in Abschnitt 2.4.3 erläutert. Zum anderen muss mit technischen Problemen und Problemen in der Nutzung gerechnet werden. Es kann immer wieder zu falschen Daten kommen, wenn das Fitbit nicht schnell genug den Puls anpasst, was zum Beispiel beim Treppensteigen oft der Fall ist. Der Puls steigt dabei meist zu schnell, als das die Technik diese Steigung nachvollziehen kann. Die Analysewaage hat einen Fehler entwickelt, der den Nutzer dazu zwingt, jede Messung zweimal durchzuführen, um ein richtiges Ergebnis zu erhalten, macht er dies nicht, ist der Wert zu hoch und somit fehlerhaft. Beim Blutdruckmessgerät ist die Pulsmessung zumeist deutlich höher[20] als jene, die der Fitbit durchführt. Diese falschen Daten wurden während des Erfassens wenn möglich umgangen, die Pulsdaten des Blutdruckmessgerätes werden vernachlässigt, beim Wiegen wird immer darauf geachtet, den richtigen Wert zu messen, und beim Fitbit muss damit gearbeitet werden, dass die Spitzen im Alltag leider fehlen.

Neben den falschen Daten, die nur schwerlich verbessert werden können, sich aber im Rahmen halten, gibt es einige fehlende Daten, die zum einen technisch bedingt sind, seien es Probleme bei der Übertragung, Messfehler weil der Sensor den Kontakt verliert oder weil der Akku des Gerätes leer ist. Dazu kommen vom Nutzer verschuldete fehlende Daten, weil die Messung oder das Anlegen des Bandes vergessen wurde, weil eine Aktivitätsmessung nicht begonnen wird oder weil das Ladegerät des Sensors nicht mit in den Urlaub genommen wurde. Dies kann durch Interpolation der Daten ausgeglichen werden, um die Datenqualität zu erhöhen.

Überflüssige Messungen mussten bei den Medisanageräten sorgfältig und aufwändig aussortiert werden. Besonders zu Beginn der Datenerhebung sind viele überflüssige Messungen nutzerbedingt, jedoch kommt es besonders beim Blutdruckmessgerät auch später noch ohne Zutun des Nutzers zu Duplikaten, die aufwändig bereinigt werden müssen. Das Armband von Fitbit erzeugt hingegen keinerlei Duplikate.

Ein großes Problem in der Datenqualität war der inkonsistente Umgang mit Datumsformaten und Maßeinheiten. Um die Daten auf einen gemeinsamen Stand zu bringen,

[20] Bis zu 40 bpm

mussten die Datumsformate fast aller Daten vereinheitlicht werden und die Maßeinheiten aus einem Teil der Medisanadaten extrahiert werden, da sie Teil des Attributs sind. Ebenso ist die Auflösung der Daten unterschiedlich, es gibt Minuten, Stunden und Tagesdaten. Um Auswertungen durchzuführen, müssen die Daten dieselbe Auflösung haben, sodass zumeist mit Tagesdaten gearbeitet wurde.

Dahingegen wurden die Ausreißer untersucht und als plausibel bewertet. Die Wetterdaten hatten eine große Integrität, bei ihnen mussten nur überschüssige Daten entfernt, das Datum angepasst und die Attribute umbenannt werden.

4.4 Systemevaluation

Zum Abschluss der Evaluation soll das System als Ganzes und hinsichtlich der Fragestellung evaluiert werden. Konnte ein Mehrwert aus den Daten generiert werden, der dem Nutzer Erkenntnisse über sich selbst brachte, die er vorher nicht hatte? Ist das Zentralisieren und Auswerten der Daten sinnvoll und auch für alle Nutzer zu empfehlen? Ist ein System umsetzbar,, das die Daten zentralisiert und dem Nutzer gesammelt anzeigt?

Zum einen ist das System wie es im Abschnitt 3.1 geplant und in Abbildung 3.3 aufgezeigt wurde, als Ganzes generell als durchführbar zu erachten. Die einzelnen Aspekte und Teile des Systems wurden im Verlauf dieser Arbeit untersucht und in Teilen umgesetzt, daraus ergab sich die generelle Umsetzbarkeit dieses Systems. Inwieweit das Analysieren und Visualisieren automatisierbar ist, ist eine Frage für weitere Arbeiten, da hier das Augenmerk mehr auf den Ergebnissen als auf der Automatisierung lag. Ebenso kann keine Aussage über die Anbindung an Expertensysteme getroffen werden, da sie nur visionär betrachtet wurden. Dahingegen hat sich die Extraktion, Bereinigung und Vorverarbeitung der Daten durchaus als umsetzbar und automatisierbar gezeigt. Die Struktur des Systems ist erweiterbar angelegt, um auch zukünftige Sensoren mit neuen Modulen implementieren zu können. Darüber hinaus ist die Anbindung des Systems an die Middleware des Labors getestet worden und die Anbindung eines Spiegels mit Datenoverlay an die Middleware und das hier beschriebene System.

Die Analyse der Daten lieferte einige sehr interessante Ergebnisse, die dem Nutzer so vorher nicht bewusst waren und lieferten ihm eine neue und tiefere Sicht und Einsicht in sich selbst. Inwieweit dies generalisierbar und auf weitere Nutzer übertragbar wäre ist nicht geklärt. Es kann vermutet werden, dass bei genügend Daten und Analyseaufwand stets ein Erkenntnisgewinn zu erwarten ist. Allein aus dem Grund, dass es eine ganz neue Sicht, frei von menschlichem Vergessen, Verklärung und subjektiver Wahrnehmung auf den Körper und das Selbst liefert. Dabei liefert die Analyse eine Möglichkeit, langfristige Zusammenhänge zwischen wenig greifbaren Merkmalen zu

finden. Es stellte sich heraus, dass die Messungen und das Gefühl nicht selten divergieren, es ist schwer, den Blutdruck einzuschätzen, und selbst beim Gewicht trügt das Gefühl hin und wieder. Momente, in denen Herzklopfen oder starke Trägheit empfunden wurden, zeigte der Pulssensor etwas anderes. Die Wahrnehmung des eigenen Körpers scheint nicht perfekt zu sein und kann durch diese Messungen erweitert werden. Allein daraus können Erkenntnisse entstehen, die zusammen mit den Auswertungen dem Nutzer helfen können. Im Rahmen dieser Arbeit wurde kein Optimierungszwang durch das Erheben und Analysieren empfunden, was bei der Literaturrecherche als mögliche Gefahr eingestuft wurde.

Dabei ist aber zu bemerken, dass der hier durchgeführte Aufwand keineswegs für alle Nutzer von Wearables zu empfehlen ist, da er ohne Programmier- und Datenverarbeitungsfähigkeiten schwerlich bis gar nicht durchzuführen ist. Von dieser Hürde abgesehen ist es jedoch durchaus zu empfehlen, die Daten fern der Herstellerdatenbanken zu zentralisieren, da die Analysemethoden der Sensorenhersteller zumeist wenig einsichtig und nicht bearbeitbar oder individualisierbar sind. Um erweiterte Einsichten und Möglichkeiten auf den eigenen Daten zu haben, muss man sie eigentlich immer zentralisieren und ggf. durch Werkzeuge von Drittanbietern verarbeiten oder wie hier gezeigt selbst verarbeiten und analysieren.

In Bezug auf die Datenlage war sie durch den Zeitraum und die Anzahl der Sensoren recht gut, es wäre aber durchaus interessant, noch mehr Daten über einen deutlich längeren Zeitraum zu haben sowie Daten von mehr Sensoren, um mehr Kontext herstellen zu können.

Ein Spiegel als Anzeige

Als Anzeigefläche und Interaktionsschnittpunkt zwischen dem System und dem Nutzer wurde ein Spiegel gewählt. Die technische Umsetzbarkeit dieser Anzeige wurde im Rahmen dieser Arbeit gezeigt und als empfehlenswert eingestuft. Dazu ist festzuhalten, dass ein Spiegel nicht die einzige Nutzerschnittstelle eines solchen Systems sein sollte, da es Aspekte des Ablaufes gibt, die auf einem Secondscreen, in einer App oder als Webapplication besser aufgehoben sind. Darüber hinaus eignet sich ein Spiegel jedoch hervorragend als Ort der Begegnung mit sich selbst, seinen Daten und um als Nutzer ein besseres Bild von sich zu bekommen. Zum einen ist dies vorteilhaft, da der Spiegel schon immer ein Gegenstand der Selbsterkenntnis war, zum anderen eignet sich der große Bildschirm beziehungsweise die Anzeigefläche der Spiegelfläche hervorragend, um Daten anzuzeigen, die im direkten Kontext zum Körper stehen. Im Laboraufbau wurde ein großer Bildschirm mit einer Kamera verwendet, für weitere Arbeiten ist allerdings ein halb durchlässiger Spiegel mit Bildschirmen und integrierter Kamera zu empfehlen, da dadurch der Nutzer weit weniger durch Bildverzögerungen oder

den ungewohnten Bildwinkel verwirrt wird. Eine Kamera wird für Bildverarbeitung, Nutzererkennung und weitere Analysen empfohlen.

4.5 Fazit

In diesem Kapitel wurden einige der Ergebnisse diskutiert und aufgezeigt sowie der Systemaufbau evaluiert. Einige der vielversprechenden Analysen haben sich als nicht funktional beziehungsweise erfolgreich erwiesen, wie zum Beispiel die Clusteranalyse. Um diese erfolgreicher zu machen, müsste der Aufwand der Tests und der Datentransformation erhöht werden, allerdings ist selbst dann ungewiss, ob die Natur der Daten dabei Ergebnisse zulässt. Dafür zeigt die explorative Datenanalyse ein paar sehr interessante und unerwartete Ergebnisse. Mit gesteigertem Aufwand und mehr Zeit könnten dort durchaus weitere Ergebnisse erzielt werden. Darüber hinaus könnten sowohl längerfristige Daten interessant sein sowie mehr Quellen. Obwohl manuelle Daten einen großen Mehraufwand bedeuten und geeignete Oberflächen erfordern, könnten sie dazu dienen, die Daten selbst sowie die Ergebnisse besser einzuordnen und in Kontext setzen zu können. Anhand der vorliegenden Ergebnisse lässt sich darin durchaus Potential erwarten.

Im hier gezeigten Aufbau ist davon auszugehen, dass auch andere Nutzer prinzipiell quantifiziert werden können, wenn die Daten vorliegen. Inwieweit dies auch bei anderen Nutzern zu neuen Erkenntnissen führen kann, ist jedoch nicht abzusehen. Außerdem ist nicht klar, ob der Systemaufbau auf andere Lebensszenarien übertragbar ist, da trotz Vereinfachung der Erhebung viele Probleme entstehen können, wenn sich die Lebensumstände des Nutzers verändern.

5 Zusammenfassung und Ausblick

Diese Arbeit hat sich mit einem Quantified Self System beschäftigt, von der Datenerhebung über die Verarbeitung bis hin zur Analyse, Interpretation und Visualisierung der Daten. Dazu wurde als Interaktionspunkt zwischen Nutzer und System ein intelligenter Spiegel betrachtet. In diesem Kapitel soll abschließend die Arbeit noch einmal zusammengefasst werden, um daraus ein Fazit zu ziehen und einen Ausblick zu geben.

5.1 Zusammenfassung

In Kapitel 1 wurde die Arbeit in der Informatik und in den Bereich der Data Science eingeordnet. Die Motivation, die in der Idee und der Arbeit steckt, wurde ebenso beschrieben wie die grundlegende Fragestellung. Lässt sich ein Quantified Self System bauen, das mithilfe von Mitteln der Datenverarbeitung Teilaspekte des Lebens quantifizieren kann und darüber hinaus dem Nutzer Selbsterkenntnisse liefert?

In Kapitel 2 der Analyse wurde das wissenschaftliche Umfeld analysiert. Dabei wurden die Bereiche des Quantified Self, der intelligenten Spiegel und der Datenverarbeitung im Rahmen des KDD betrachtet. Darüber hinaus wurden die vorliegenden Daten sowie ihre Qualität analysiert und bewertet. Dazu wurde unter anderem ein Validitätsmodell aufgebaut und vorgestellt. Aus den Analysen der einzelnen Felder und der Daten ergaben sich die Anforderungen an das geplante System.

In Kapitel 3 wurde die Planung des Systems mit Architektur und Einarbeitung der Anforderungen, die sich in Kapitel 2 ergeben haben, beschrieben. Darüber hinaus wurde die Umsetzung des Systems in seinen einzelnen Komponenten beschrieben. Darauf folgt die Beschreibung der Durchführung der Analysen sowie dem technischen Aufbau des Spiegels als Anzeige- und Kommunikationselement des Systems.

Das Kapitel 4 dient der Evaluation und Diskussion der Ergebnisse, die sich sowohl im Rahmen der Umsetzung des Systems ergaben wie auch aus den Analysen der Daten. Diese Ergebnisse wurden darüber hinaus interpretiert und kontextualisiert, um die Frage zu beantworten, die in der Einleitung gestellt wurde. Ein System zur Quantifizierung einiger Aspekte des Lebens ist mit Wearables umsetzbar. Darüber hinaus war

es möglich, dem Nutzer Erkenntnisse und Einsichten zu ermöglichen, die er vorher nicht hatte.

5.2 Fazit

Angefangen bei der Erhebung der Daten hat sich ergeben, dass das Erheben für den Nutzer möglichst komfortabel und in sein Leben integrierbar sein muss. Dabei ist manuelles Erfassen so gut es geht zu vereinfachen oder ganz darauf zu verzichten, da daraus eine ganze Reihe von Problemen entsteht. Des Weiteren müssen Sensoren zuverlässig, einfach und komfortabel sein. Es ist wichtig, dass sie zuverlässige und möglichst akkurate Daten liefern, da falsche Messungen den Nutzer irreführen und beunruhigen können.

Nach der Datenerhebung müssen die Daten extrahiert werden, in diesem Fall aus den APIs der Wearableshersteller. Dies ist teilweise sehr schwierig und ohne erweiterte technische Fähigkeiten nicht möglich. Nicht alle Hersteller empfinden die Daten, die der Nutzer erhebt, als Eigentum des Nutzers, sondern als Firmeneigentum und handeln unter anderem damit. Somit ist es nicht immer im Interesse der Firmen, dem Nutzer auf einfache Art alle Daten so zur Verfügung zu stellen, dass jeder einfach an seine Daten kommt. Die neue Datenschutzrichtlinie hat dies jedoch für den Nutzer deutlich vereinfacht.

Die Daten müssen, nachdem sie extrahiert sind, bereinigt, verarbeitet und transformiert werden. Diese Arbeitsschritte sind von größter Wichtigkeit und mit größter Sorgfalt auszuführen, da die zu erwartenden Ergebnisse sehr stark von der Qualität der Daten und den Entscheidungen in der Bereinigung abhängen. Es zeigt sich, dass der größte Aufwand im Bereich der Data Science die Verarbeitung und Bereinigung der Daten ist. Das Aufsetzen der Analyse und die Durchführung der Testdurchläufe sind ebenfalls aufwändig, bauen jedoch auf allen vorherigen Phasen auf. Neben der Verbesserung der Qualität ist auch die Kontextualisierung der Daten und der Ergebnisse ein großer und wichtiger Aspekt, um die Ergebnisse interpretieren zu können. Die Interpretation ist dabei weder trivial noch einfach und erfordert ein großes Maß an Domänenwissen und Datenverständnis. Die Qualität der Ergebnisse ist auch an die Wahl der Analysemethoden und Algorithmen gebunden.

Nach der Analyse müssen die Ergebnisse aufbereitet und visualisiert werden, um dem Nutzer angezeigt werden zu können. Dabei ist auch hier ein großes Augenmerk auf Verständlichkeit, Übersicht und Funktion zu richten. Der Nutzer kann nur schwer Informationen oder gar Erkenntnisse gewinnen, wenn die Visualisierungen nicht gut sind oder seine Informationssuche nicht unterstützen. Dies ist unter anderem einer der Punkte, in denen aktuelle Apps Defizite aufweisen.

Ebenso kann als Fazit gezogen werden, dass mit den Mitteln, die hier verwendet wurden, ein System gebaut werden kann, das Aspekte des Lebens quantifizieren kann. Das Auswerten dieser Daten führte zu Erkenntnissen für den Nutzer, die vorher nicht vorhanden waren. Darüber hinaus besteht jedoch keine zwingende Generalisierbarkeit. Das System als solches ist auf andere Nutzer durchaus anwendbar, inwieweit daraus Erkenntnisse erhoben werden können, liegt allerdings immer an der Art der Daten. Es ist davon auszugehen, dass es etwas zu entdecken gibt, ob es aber in den Ausmaßen auftritt wie sie hier zu finden waren, ist nicht abzusehen. Jedoch können andere Lebensumstände dazu führen, dass das Erheben der Daten nicht ausreichend gut funktioniert. Das heißt, auch auf das Leben des Nutzers ist das System nicht zwingend generalisierbar und bedarf höchstwahrscheinlich Anpassungen.

Ein am Anfang der Arbeit angesprochener Punkt war die gefühlte Unzuverlässigkeit von Gefühlen. Es ist nicht immer einfach, zwischen dem gefühlten Zustand und der Realität zu vermitteln. Die Technik kann hier als Vermittler auftreten und ist es im Rahmen dieser Arbeit auch in Teilen bereits. Besonders der Schlaf ist schwer selbst zu empfinden, da man sich nicht immer daran erinnert, aufgewacht zu sein und zumeist auch nicht, wann man eingeschlafen ist. Die Messungen über den Schlaf sind somit ein guter Hinweis. Dagegen ist es aber auch immer mit Vorsicht zu genießen, da die Sensoren noch nicht zuverlässig genug sind. Es muss also vom Nutzer immer kritisch bewertet werden, wie glaubhaft ein Sensor ist und ob er dem Sensor zutraut, glaubwürdiger als sein Gefühl zu sein. Das heißt, um die Lücke zwischen Gefühl und Realität zu überbrücken, braucht es in erster Linie bessere, zuverlässigere und genauere Sensoren. Darüber hinaus ist allerdings nicht abzusehen, welche Daten noch erhoben werden müssen, um dem Nutzer ein vollständigeres digitales Spiegelbild liefern zu können. Neben dem eigenen Misstrauen gegenüber dem Gefühl gibt es auch Dinge, die schlichtweg nicht erfühlt werden können. Ist ein Blutdruck zu hoch, merkt der Betroffene das, stimmen jedoch manche Blutwerte nicht, so erkennt der Mensch die Symptome nicht, ordnet sie anderen Dingen zu oder hat schlicht keine erkennbaren Symptome, bis der Wert so kritisch ist, dass es schlimme Folgen hat. Bei Diabetes ist dies zum Beispiel so, ohne die Messungen ist es für die Betroffenen praktisch unmöglich zu sagen, wie ihre Werte sind. Durch die Messungen jedoch kann im besten Fall eine neue Sensibilität erzeugt werden, die die Betroffenen unterstützt, ein besseres Körpergefühl zu entwickeln oder zumindest den Körper und die Umstände nicht mehr als Gegner zu sehen. Neben dem Beispiel der chronischen Krankheiten ist es jedoch auch so, dass die Körperwahrnehmung der Menschen stark schwankt, neben Menschen mit einem sehr guten Körpergefühl gibt es welche, die Signale des Körpers schwer deuten oder kaum wahrnehmen können. Für diese ist die Technik und auch die Datenerhebung ein großartiges Mittel, um das fehlende, trügende oder falsche Gefühl zu ersetzen. Im besten Fall kann dabei durch das Datenerheben eine Körperwahrnehmung trainiert werden. Im schlimmsten Fall könnte aber eine Abhängigkeit zum Werkzeug erzeugt werden. Inwieweit das eine realistische Angst ist und ein Problem darstellt, ist aber noch nicht

abzusehen. Sicher ist, dass chronisch Kranke schon jetzt in einer Abhängigkeit den Messungen gegenüber stehen und dies auch schon seit langem sind.

Eine offene Debatte bleibt, ob Quantified Self einen Nutzer automatisch zum Optimieren seiner Daten zwingt. In der hier vorliegenden Arbeit war dies über die Anfangsphase hinaus, in der die Sensoren neu waren, nicht der Fall. Nachdem der Novelty Effekt nachgelassen hatte, hat der Nutzer sein Verhalten nicht mehr den Sensoren angepasst und sich auch nicht zum Optimieren gezwungen gefühlt. Die Sensoren waren ein Mittel zur Beobachtung. Jedoch sind die meisten Apps so aufgebaut, dass sie zum Optimieren animieren, es ist also nicht unwahrscheinlich, dass ein Nutzer, der aus Neugierde anfängt Daten zu erheben, von der Natur der Apps dazu animiert wird, sich zu optimieren oder jemand, der einen Aspekt optimieren wollte, dies auf weitere Aspekte und Werte ausdehnt.

5.3 Ausblick

Im Laufe dieser Arbeit haben sich viele weitere Fragen aufgetan sowie Stoff für viele weitere Arbeiten, sowohl technisch wie auch konzeptionell. Das System kann erweitert, verbessert und weiter ausgearbeitet werden. Angefangen mit der Verwendung anderer Hardware für den Spiegel über die Erweiterung der Sensoren mit anderen Kameras, wie einer Infrarotkamera, bis hin zur Erweiterung der Software. Dies könnten Funktionalitäten für Gesichts- und Spracherkennung sein wie auch weitere Analysen, ein verbesserter Datenfluss und eine geeignete Anzeige der Daten und Ergebnisse für den Nutzer. Weitere Quellen und ein System, um vom Nutzer manuelle Daten komfortabel abzufragen, wären ein guter nächster Schritt, um die vorliegenden Daten besser zu kontextualisieren. Dabei ist nicht abzusehen, ob mehr Daten wirklich mehr Erkenntnisse bringen können. Bei einigen Daten ist zu vermuten, dass sie durchaus noch spannende Aspekte darstellen, die durchaus eine Bereicherung sein können, wie Daten über Alkoholkonsum, Krankheiten, Emotionen und ähnliche in der Arbeit beschriebene Aspekte. Aber es ist nicht allgemeingültig, dass mehr Daten mehr Erkenntnisse bringen. Dies kann dazu von Nutzer zu Nutzer und Lebenssituation zu Lebenssituation schwanken. Daten, die bei dem einen Nutzer große Erkenntnisse liefern, müssen dies bei einem anderen Nutzer nicht. Gleichermaßen ist nicht abzusehen, wie genau und vertrauenswürdig die Sensoren in Zukunft werden und welche Auswirkung dies auf Arbeiten, wie die hier gezeigte, hat. Es ist zu vermuten, dass es positive Effekte auf die Datenqualität und somit Auswertungsmöglichkeiten hat. Aber das wird erst die Zukunft zeigen. Ein möglicher weiterer Schritt könnte es sein, einen Experten im Rahmen von Sportwissenschaften oder Medizin zu konsultieren, um die Validität der Erkenntnisse zu überprüfen und neben der Technik auch die qualitative Auswertung der Daten zu verbessern.

Es bleibt somit abzuwarten, was die Zukunft im Rahmen der Sensorik bringt, aber auch im Diskurs über Quantified Self. Wie werden sich das Vermessen des Selbst und die Sicht der Gesellschaft darauf verändern? Wird die Technik das Körpergefühl ersetzen und die Menschen vom Erfassen der Daten abhängig machen oder zu einem neuen präziseren Körpergefühl beitragen? Werden politische Entwicklungen die Menschen vor der neuen Technik und ihre Folgen zurückschrecken lassen oder sie gar zur Pflicht machen? In den Arbeiten von Deborah Lupton wird immer wieder die Vision von Human Enhancement angesprochen, für die schon jetzt Grundsteine gelegt wurden. Kann es sein, dass der Wunsch der Menschen nach besseren Sinnen, verbesserter Gesundheit und dem Ausweichen unangenehmer Situationen soweit führt, dass der Körper selbst durch Technik nicht nur verbessert, sondern auch erweitert wird? Bessere Sinne finden sich in vielen Geschichten über optimierte Menschen wieder, größere Stärke oder ein Adlerblick locken so manchen. Gleichzeitig lockt auch die Hoffnung, unangenehme Prozeduren auslassen zu können, wenn die Wahl besteht zwischen einem Chip oder ähnlichem, der einem einmal implementiert wird und dadurch Blutuntersuchungen obsolet macht und den unangenehmen Blutuntersuchungen selbst, die regelmäßig durchgeführt werden müssen, ist es nicht unwahrscheinlich, dass eine Vielzahl sich für den Chip entscheidet. Auch jetzt verbessert sich der Mensch durch Technik, Hörgeräte, Herzschrittmacher, neue Gelenke oder gar intelligente Prothesen. All das fällt schon in den Bereich Human Enhancement, es bleibt abzuwarten, wie sich dieses Feld weiter entwickelt und welche Neuerungen es auch für Quantified Self bringt. Somit ist Quantified Self in einem Themenbereich angesiedelt, der sich im Augenblick und auch in Zukunft rasant entwickelt und sowohl technisch wie auch gesellschaftlich Neuerungen bringt. Somit verändert die Datafizierung des Menschen eine Menge, eine interessante Frage wäre, inwieweit diese Datafizierung nicht nur die Sicht auf den Körper und das Selbst verändert, sondern auch auf die Umwelt, mit der wir in Wechselwirkung stehen. Kann die Verdatung unseres Lebensraumes in Form von Geräuschpegeln, Luftqualität und ähnlichen Dingen in Bezug auf Quantified Self dazu führen, dass wir uns in unserer Umwelt anders wahrnehmen? Dass das Verhalten sich gegebenenfalls anpasst und ein stärkeres politisches Engagement zum Beispiel für verbesserte Umweltbedingungen zu Tage treten kann? Nehmen wir die Umwelt wegen ihrer Daten vielleicht irgendwann als Teil von uns selbst wahr? Dies sind Fragen, deren Antworten bisher nicht abzusehen sind und die wohl erst in Zukunft beantwortet werden können.

Literaturverzeichnis

[AppleMirror] *The Apple Mirror*. http://www.rafaeldymek.com/portfolio/apple-mirror/. – Besucht am: 24.02.2017

[BlutdruckNachWHO] *Blutdruck Tabellen nach WHO und der deutschen Hochdruckliga*. https://www.blutdruckdaten.de/lexikon/blutdruck-normalwerte.html. – Besucht am: 29.10.2019

[DBSCAN] *DBSCAN Algorithmus in KNIME*. https://nodepit.com/node/org.knime.base.node.mine.dbscan.DBSCANNodeFactory. – Besucht am: 29.10.2019

[Docker] *Docker*. https://www.docker.com/. – Besucht am: 14.08.2019

[FitbitApp] *Fitbit App*. https://www.fitbit.com/de/app. – Besucht am: 15.08.2019

[FitbitPremium] *Fitbit Premium*. https://investor.fitbit.com/press/press-releases/press-release-details/2019/Fitbit-Launches-Fitbit-Premium-New-Health-and-Fitness-Subscription-Service/default.aspx. – Besucht am: 13.09.2019

[FuzzyC-Means] *Fuzzy c-Means Algorithmus in KNIME*. https://nodepit.com/node/org.knime.base.node.mine.cluster.fuzzycmeans.FuzzyClusterNodeFactory2. – Besucht am: 29.10.2019

[k-Medoids] *Fuzzy c-Means Algorithmus in KNIME*. https://nodepit.com/node/org.knime.base.node.mine.cluster.fuzzycmeans.FuzzyClusterNodeFactory2. – Besucht am: 29.10.2019

[Grafana] *Grafana*. https://grafana.com/. – Besucht am: 14.08.2019

[InfluxDB] *Influx DB*. https://www.influxdata.com/. – Besucht am: 14.08.2019

[k-Means] *k-Means Algorithmus in KNIME*. https://nodepit.com/node/org.knime.base.node.mine.cluster.kmeans.ClusterNodeFactory2. – Besucht am: 29.10.2019

[dwd] *Klimadaten Deutschland*. https://www.dwd.de/DE/leistungen/klimadatendeutschland/klimadatendeutschland.html. – Besucht am: 31.05.2019

[KNIME] *KNIME*. https://www.knime.com/. – Besucht am: 14.08.2019

[MagicMirror] *The Magic Mirror*. http://www.magicmirror.me/. – Besucht am: 24.02.2017

[MedisanaApp] *Medisana App*. https://www.medisana.de/VitaDock-App-2-0.html. – Besucht am: 16.09.2019

[Excel] *Microsoft Excel*. https://products.office.com/de-de/excel. – Besucht am: 14.08.2019

[MySmartMirror] *My Smart Mirror*. https://www.mysmartmirror.co.uk/. – Besucht am: 04.10.2019

[Naked] *The Naked Mirror*. https://naked.fit/. – Besucht am: 24.02.2017

[Notepad] *Notepad ++*. https://notepad-plus-plus.org/. – Besucht am: 14.08.2019

[PulsNormWerte] *Puls Normwerte nach Geschlecht, Alter und Fitness Level*. https://www.cardiosecur.com/de/ihr-herz/fachartikel-rund-um-das-herz/das-gesunde-herz. – Besucht am: 29.10.2019

[Rapidminer] *RapidMiner*. https://rapidminer.com/. – Besucht am: 14.08.2019

[Statistics] *Statistische Auswertungen in KNIME*. https://nodepit.com/node/org.knime.base.node.stats.viz.extended.ExtendedStatisticsNodeFactory. – Besucht am: 12.11.2019

[BlutdruckSymptome] *Symptome von zu hohem Blutdruck*. https://www.blutdruckdaten.de/lexikon/blutdruck-symptome.html. – Besucht am: 29.10.2019

[Athira u. a. 2016] Athira, S. ; Francis, F. ; Raphel, R. ; Sachin, N. S. ; Porinchu, S. ; Francis, S.: Smart mirror: A novel framework for interactive display. In: *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, March 2016, S. 1–6

[Beierle und Isberner 2006] Beierle, Christoph ; Isberner, Gabriele K.: *Methoden wissensbasierter Systeme : Grundlagen, Algorithmen, Anwendungen*. 3., überarb. u. erw. Aufl. Braunschweig [u.a.] : Viewg, 2006

[Benedetto u. a. 2018] Benedetto, Simone ; Caldato, Christian ; Bazzan, Elia ; Greenwood, Darren C. ; Pensabene, Virginia ; Actis, Paolo: Assessment of the Fitbit Charge 2 for monitoring heart rate. In: *PLOS ONE* 13 (2018), 02, Nr. 2, S. 1–10. – URL https://doi.org/10.1371/journal.pone.0192691

[Bentley u. a. 2013] Bentley, Frank ; Tollmar, Konrad ; Stephenson, Peter ; Levy, Laura ; Jones, Brian ; Robertson, Scott ; Price, Ed ; Catrambone, Richard ; Wilson, Jeff: Health Mashups: Presenting Statistical Patterns Between Wellbeing Data and Context in Natural Language to Promote Behavior Change. In: *ACM Trans. Comput.-Hum. Interact.* 20 (2013), November, Nr. 5, S. 30:1–30:27. – URL http://doi.acm.org/10.1145/2503823. – ISSN 1073-0516

[Besserer u. a. 2016] Besserer, Daniel ; Bäurle, Johannes ; Nikic, Alexander ; Honold, Frank ; Schüssel, Felix ; Weber, Michael: Fitmirror: A Smart Mirror for Positive Affect in Everyday User Morning Routines. In: *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. New York, NY, USA : ACM, 2016 (MA3HMI '16), S. 48–55. – URL http://doi.acm.org/10.1145/3011263.3011265. – ISBN 978-1-4503-4562-0

[Chan 2003] Chan, Y. H.: Biostatistics 104: Correlational Analysis. In: *Singapore medical journal* 44 (2003), December, Nr. 12, S. 614–619. – URL http://www.sma.org.sg/smj/4412/4412bs1.pdf

[Choe u. a. 2014] Choe, Eun K. ; Lee, Nicole B. ; Lee, Bongshin ; Pratt, Wanda ; Kientz, Julie A.: Understanding Quantified-selfers' Practices in Collecting and Exploring Personal Data. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2014 (CHI '14), S. 1143–1152. – URL http://doi.acm.org/10.1145/2556288.2557372. – ISBN 978-1-4503-2473-1

[Cleve und Lämmel 2014] Cleve, J. ; Lämmel, U.: *Data Mining*. De Gruyter, 2014 (De Gruyter Studium). – URL https://books.google.de/books?id=g0TpBQAAQBAJ. – ISBN 9783486720341

[Cvetkoska u. a. 2017] Cvetkoska, B. ; Marina, N. ; Bogatinoska, D. C. ; Mitreski, Z.: Smart mirror E-health assistant - Posture analyze algorithm proposed model for upright posture. In: *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, July 2017, S. 507–512

[Kersten-van Dijk u. a. 2017] Dijk, Elisabeth T. Kersten-van ; Westerink, Joyce H. D. M. ; Beute, Femke ; IJsselsteijn, Wijnand A.: Personal Informatics, Self-Insight, and Behavior Change: A Critical Review of Current Literature. In: *Hum.-Comput. Interact.* 32 (2017), November, Nr. 5-6, S. 268–296. – URL https://doi.org/10.1080/07370024.2016.1276456. – ISSN 0737-0024

[Eichler 2014] Eichler, Tobias: Agentenbasierte Middleware zur Entwicklerunterstützung in einem Smart-Home-Labor. (2014). – URL http://users.informatik.haw-hamburg.de/~ubicomp/arbeiten/master/eichler.pdf

[Emmert 2013] Emmert, Bianca: Sind wir bereit für den Big-Data-Lifestile? (2013). – URL http://igrowdigital.com/wp-content/uploads/2013/12/BA_QS_Big-Data_17-09-2013.pdf

[Fayyad u. a. 1996] Fayyad, Usama M. ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: Advances in Knowledge Discovery and Data Mining. In: Fayyad, Usama M. (Hrsg.) ; Piatetsky-Shapiro, Gregory (Hrsg.) ; Smyth, Padhraic (Hrsg.) ; Uthurusamy, Ramasamy (Hrsg.): *From Data Mining to Knowledge Discovery: An Overview*. Menlo Park, CA, USA : American Association for Artificial Intelligence, 1996, S. 1–34. – URL http://dl.acm.org/citation.cfm?id=257938.257942. – ISBN 0-262-56097-6

[Felizi und Varon 2018] Felizi, Natasha ; Varon, Joana: *MENSTRUAPPS – How to turn your period into money (for others)*. CHUPADADOS. 2018. – URL https://chupadados.codingrights.org/en/menstruapps-como-transformar-sua-menstruacao-em-dinheiro-para-os-outros/. – Zugriffsdatum: 23.04.2018

[Fipps 2018] Fipps, Paul: *NOTICE OF DATA BREACH*. MyFitnessPal Website. 2018. – URL https://content.myfitnesspal.com/security-information/notice.html. – Zugriffsdatum: 23.04.2018

[Focus] Focus: *Ïntelligente Tinte": Tattoos zeigen, wann Diabetiker zum Insulin greifen müssen*. https://www.focus.de/gesundheit/news/gesundheitsinformationen-auf-der-haut-tattoos-mit-intelligenter-tinte-zeigen-ges id_7662148.html. – Besucht am: 16.10.2019

[Fujinami 2010] Fujinami, Kaori: A Case Study on Information Presentation to Increase Awareness of Walking Exercise in Everyday Life. In: *International Journal of Smart Home* 4 (2010), S. 11–26. – URL https://www.earticle.net/Article/A148072

[Fujinami u. a. 2005] Fujinami, Kaori ; Kawsar, Fahim ; Nakajima, Tatsuo: AwareMirror: A Personalized Display Using a Mirror. In: Gellersen, Hans W. (Hrsg.) ; Want, Roy (Hrsg.) ; Schmidt, Albrecht (Hrsg.): *Pervasive Computing*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2005, S. 315–332. – ISBN 978-3-540-32034-0

[Gold u. a. 2016] Gold, D. ; Sollinger, D. ; Indratmo: SmartReflect: A modular smart mirror application platform. In: *2016 IEEE 7th Annual Information Tech-*

nology, Electronics and Mobile Communication Conference (IEMCON), Oct 2016, S. 1–7

[Hanfeld 2015] Hanfeld, Michael: *Punkte für gefälliges Verhalten.* Frankfurter Allgemeine vom 10.10.2015. 2015. – URL http://www.faz.net/medien/punktrichter-citizen-score-ueberwachung-in-china-13848403.html. – Zugriffsdatum: 09.02.2016

[Harrison u. a. 2014] Harrison, Daniel ; Marshall, Paul ; Berthouze, Nadia ; Bird, Jon: Tracking Physical Activity: Problems Related to Running Longitudinal Studies with Commercial Devices. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication.* New York, NY, USA : ACM, 2014 (UbiComp '14 Adjunct), S. 699–702. – URL http://doi.acm.org/10.1145/2638728.2641320. – ISBN 978-1-4503-3047-3

[Hossain u. a. 2007] Hossain, M. A. ; Atrey, P. K. ; Saddik, A. E.: Smart mirror for ambient home environment. In: *2007 3rd IET International Conference on Intelligent Environments*, Sept 2007, S. 589–596. – ISSN 0537-9989

[Institute 2018] Institute, AV Test The Independet IT-Security: *Fitness Trackers – 13 Wearables in a Security Test.* https://www.av-test.org/de/news/fitness-tracker-13-wearables-im-sicherheitstest/. 05 2018. – Besucht am: 16.10.2019

[quantified self institute 2016] institute quantified self: *What is quantified self.* Quantified Self Institute. 2016. – URL http://qsinstitute.com/about/what-is-quantified-self/. – Zugriffsdatum: 16.07.2018

[Johri u. a. 2018] Johri, A. ; Jafri, S. ; Wahi, R. N. ; Pandey, D.: Smart Mirror: A time-saving and Affordable Assistant. In: *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec 2018, S. 1–4

[Kamal u. a. 2010] Kamal, Noreen ; Fels, Sidney ; Ho, Kendall: Online Social Networks for Personal Informatics to Promote Positive Health Behavior. In: *Proceedings of Second ACM SIGMM Workshop on Social Media.* New York, NY, USA : ACM, 2010 (WSM '10), S. 47–52. – URL http://doi.acm.org/10.1145/1878151.1878167. – ISBN 978-1-4503-0173-2

[Kamenz 2014] Kamenz: Quantified Self - Anspruch und Realitaet. In: *Master Informatik an der HAW Grundseminar* (2014), S. 6. – URL https://users.informatik.haw-hamburg.de/~ubicomp/projekte/master14-15-gsm/berichte.html

[Koch u. a. 2018] Koch, Michael ; Luck, Kai von ; Schwarzer, Jan ; Draheim, Susanne: The Novelty Effect in Large Display Deployments – Experiences and

Lessons-Learned for Evaluating Prototypes, European Society for Socially Embedded Technologies (EUSSET), 2018, S. 65–93. – ISSN 2510-2591

[Lüdemann 2016a] Lüdemann, Maria: Data Minung auf Consumer Sensor Daten für Quantified Self. (2016). – URL http://users.informatik.haw-hamburg.de/~ubicomp/arbeiten/bachelor/luedemann.pdf

[Lüdemann 2016b] Lüdemann, Maria: Quantifed Self nicht nur zum Selbstzweck. In: *Grundseminar Arbeit published an der HAW Hamburg* (2016). – URL http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2015-gsem/luedemann/bericht.pdf

[Lüdemann 2017a] Lüdemann, Maria: Der Intelligente Spiegel - Ein Companion zur Unterstützung der Selbstwahrnehmung Prototyp 1.0. In: *Grundprojekt Arbeit published an der HAW Hamburg* (2017). – URL http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2016-hsem/luedemann/bericht.pdf

[Lüdemann 2017b] Lüdemann, Maria: Der intelligente Spiegel- Ein Companion zur Unterstützung der Selbstwahrnehmung. In: *Hauptseminar Arbeit published an der HAW Hamburg* (2017). – URL http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2016-hsem/luedemann/bericht.pdf

[Lüdemann 2018] Lüdemann, Maria: Der Intelligente Spiegel - Ein Companion zur Unterstützung der Selbstwahrnehmung Prototyp 1.0. In: *Hauptprojekt Arbeit published an der HAW Hamburg* (2018). – URL https://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2018-proj/luedemann.pdf

[Li u. a. 2010] Li, Ian ; Dey, Anind ; Forlizzi, Jodi: A Stage-based Model of Personal Informatics Systems. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2010 (CHI '10), S. 557–566. – URL http://doi.acm.org/10.1145/1753326.1753409. – ISBN 978-1-60558-929-9

[Li u. a. 2011] Li, Ian ; Dey, Anind K. ; Forlizzi, Jodi: Understanding My Data, Myself: Supporting Self-reflection with Ubicomp Technologies. In: *Proceedings of the 13th International Conference on Ubiquitous Computing*. New York, NY, USA : ACM, 2011 (UbiComp '11), S. 405–414. – URL http://doi.acm.org/10.1145/2030112.2030166. – ISBN 978-1-4503-0630-0

[Lupton 2013] Lupton, Deborah: Quantifying the body: monitoring and measuring health in the age of mHealth technologies. In: *Critical Public Health* 23 (2013), Nr. 4, S. 393–403. – URL http://dx.doi.org/10.1080/09581596.2013.794931

[Lupton 2014] Lupton, Deborah: Self-tracking Cultures: Towards a Sociology of Personal Informatics. In: *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design*. New York, NY, USA : ACM, 2014 (OzCHI '14), S. 77–86. – URL http://doi.acm.org/10.1145/2686612.2686623. – ISBN 978-1-4503-0653-9

[Lupton 2016] Lupton, Deborah: *The Quantified Self*. 1st. Polity Press, 2016. – ISBN 1509500596, 9781509500598

[Lupton 2018] Lupton, Deborah: Self-Tracking. (2018), 02, S. 9. – URL https://www.researchgate.net/publication/323402296_Self-Tracking

[MacLeod u. a. 2013] MacLeod, Haley ; Tang, Anthony ; Carpendale, Sheelagh: Personal Informatics in Chronic Illness Management. In: *Proceedings of Graphics Interface 2013*. Toronto, Ont., Canada, Canada : Canadian Information Processing Society, 2013 (GI '13), S. 149–156. – URL http://dl.acm.org/citation.cfm?id=2532129.2532155. – ISBN 978-1-4822-1680-6

[Meetup 2018] Meetup: *Quantified self Meetups World Wide*. Meetup. 2018. – URL https://www.meetup.com/topics/quantified-self/. – Zugriffsdatum: 16.07.2018

[Mentis u. a. 2017] Mentis, Helena M. ; Komlodi, Anita ; Schrader, Katrina ; Phipps, Michael ; Gruber-Baldini, Ann ; Yarbrough, Karen ; Shulman, Lisa: Crafting a View of Self-Tracking Data in the Clinical Visit. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2017 (CHI '17), S. 5800–5812. – URL http://doi.acm.org/10.1145/3025453.3025589. – ISBN 978-1-4503-4655-9

[Neuringer 1981] Neuringer: Self-experimentation: a call for change. In: *Behaviorism - Springer* (1981), S. 16. – URL http://www.reed.edu/psychology/docs/SelfExperimentation.pdf

[Njaka u. a. 2018] Njaka, A. C. ; Li, N. ; Li, L.: Voice Controlled Smart Mirror with Multifactor Authentication. In: *2018 IEEE International Smart Cities Conference (ISC2)*, Sep. 2018, S. 1–8

[Plass-Flessenkämpfer 2015] Plass-Flessenkämpfer, Benedikt: *„Citizen Score": China bewertet seine Bürger und ihre Lebensweise*. Wired vom 07.10.2015. 2015. – URL https://www.wired.de/collection/latest/china-fuhrt-citizen-scores-ein-um-seine-burger-nach-ihrer-lebensweise-zu-bewerte – Zugriffsdatum: 09.02.2016

[Raisinghani u. a. 2006] Raisinghani, Mahesh ; Benoit, Ally ; Ding, Jianchun ; Gomez, Maria ; Gupta, Kanak ; Gusila, Victor ; Power, Daniel ; Schmedding,

Oliver: Ambient Intelligence: Changing Forms of Human-Computer Interaction and their Social Implications. In: *Journal of Digital Information* 5 (2006), Nr. 4. – URL https://journals.tdl.org/jodi/index.php/jodi/article/view/149. – ISSN 1368-7506

[Response 2014] Response, Symantec S.: *How safe is your quantified self? Tracking, monitoring, and wearable tech.* Symantec Official Blog. 2014. – URL http://www.symantec.com/connect/blogs/how-safe-your-quantified-self-tracking-monitoring-and-wearable-tech. – Zugriffsdatum: 23.04.2018

[Rogerson u. a. 2016] Rogerson, David ; Soltani, Hora ; Copeland, Robert: The weight-loss experience: a qualitative exploration. In: *BMC Public Health* 16 (2016), May, Nr. 1, S. 371. – URL https://doi.org/10.1186/s12889-016-3045-6. – ISSN 1471-2458

[Rooksby u. a. 2014] Rooksby, John ; Rost, Mattias ; Morrison, Alistair ; Chalmers, Matthew C.: Personal Tracking As Lived Informatics. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2014 (CHI '14), S. 1163–1172. – URL http://doi.acm.org/10.1145/2556288.2557039. – ISBN 978-1-4503-2473-1

[Self 2018] Self, Quantified: *Guide to Self-Tracking Tools*. QS Homepage. 2018. – URL http://quantifiedself.com/guide/. – Zugriffsdatum: 13.08.2018

[Spiegel 2014] Spiegel, Online: *Datenklau: Hacker stehlen Daten von 4,5 Millionen US-Patienten.* Spiegel Online. 2014. – URL http://www.spiegel.de/netzwelt/netzpolitik/us-krankenhaeuser-hacker-stehlen-daten-von-4-5-millionen-patienten-a-986804.html. – Zugriffsdatum: 23.04.2018

[Statista] Statista: *Prognose Smartphone Nutzer Weltweit 2016 bis 2021.* https://de.statista.com/statistik/daten/studie/309656/umfrage/prognose-zur-anzahl-der-smartphone-nutzer-weltweit/. – Besucht am: 14.10.2019

[Swan 2013] Swan: The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. In: *Big Data Volume: 1 Issue 2: June 18, 2013*. USA : Mary Ann Liebert, Inc., publishers, 2013 (Big Data 06.13), S. 85–99. – URL http://dx.doi.org/10.1089/big.2012.0002

[Tanenbaum und van Steen 2008] Tanenbaum, A.S. ; Steen, M. van: *Verteilte Systeme: Prinzipien und Paradigmen*. Pearson Studium, 2008 (It Informatik). – URL https://books.google.de/books?id=V6I6PQAACAAJ. – ISBN 9783827372932

[Tukey 1962] Tukey, John W.: The Future of Data Analysis. In: *Ann. Math. Statist.* 33 (1962), 03, Nr. 1, S. 1–67. – URL https://doi.org/10.1214/aoms/1177704711

[Tukey 1977] Tukey, John W.: *Exploratory data analysis*. Addison-Wesley, 1977 (Addison-Wesley series in behavioral science : quantitative methods). – URL http://www.worldcat.org/oclc/03058187. – ISBN 0201076160

[Wahlster 2003] Wahlster, et a.: *SmartKom - Dialogische Mensch-Maschinen-Interaktion durch koordinierte Analyse und Generierung multipler Modalitäten*. DFKI Website. 2003. – URL http://www.smartkom.org/start_de.html. – Zugriffsdatum: 14.08.2017

[West u. a. 2016] West, Peter ; Giordano, Richard ; Van Kleek, Max ; Shadbolt, Nigel: The Quantified Patient in the Doctor's Office: Challenges & Opportunities. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2016 (CHI '16), S. 3066–3078. – URL http://doi.acm.org/10.1145/2858036.2858445. – ISBN 978-1-4503-3362-7

[Whitson 2013] Whitson, Jennifer R.: Gaming the Quantified Self, URL https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/gaming, 2013, S. 163–176

[Whooley u. a. 2014] Whooley, Mark ; Ploderer, Bernd ; Gray, Kathleen: On the Integration of Self-tracking Data Amongst Quantified Self Members. In: *Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014 - Sand, Sea and Sky - Holiday HCI*. UK : BCS, 2014 (BCS-HCI '14), S. 151–160. – URL http://dx.doi.org/10.14236/ewic/hci2014.16

[Wiedemann 2016] Wiedemann, Lisa: Datensätze der Selbstbeobachtung - Daten verkörpern und Leib vergessen!? In: *Lifelogging Digitale Selbstverbesserung und Lebensprotokollierung zwischen disruptiver Technologie und kulturellem Wandel*, Springer VS, 2016, S. 65–93. – ISBN 978-3-658-10415-3

[Wolf 2010] Wolf, Gary: *The Data Driven Life*. The New York Times Magazine 28.04.2010. 2010. – URL https://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html. – Zugriffsdatum: 20.07.2018

[Yu u. a. 2012] Yu, Yuan-Chih ; D. You, Shingchern ; Tsai, Dwen-Ren: Magic Mirror Table for Social-Emotion Alleviation in the Smart Home. In: *IEEE Transactions on Consumer Electronics - IEEE TRANS CONSUM ELECTRON* 58 (2012), 02, S. 126–131

[Yusri u. a. 2017] Yusri, M. M. ; Kasim, S. ; Hassan, R. ; Abdullah, Z. ; Ruslai, H. ; Jahidin, K. ; Arshad, M. S.: Smart mirror for smart life. In: *2017 6th ICT International Student Project Conference (ICT-ISPC)*, May 2017, S. 1–5

6 Anhang

6.1 Beispielhafte Visualiationen

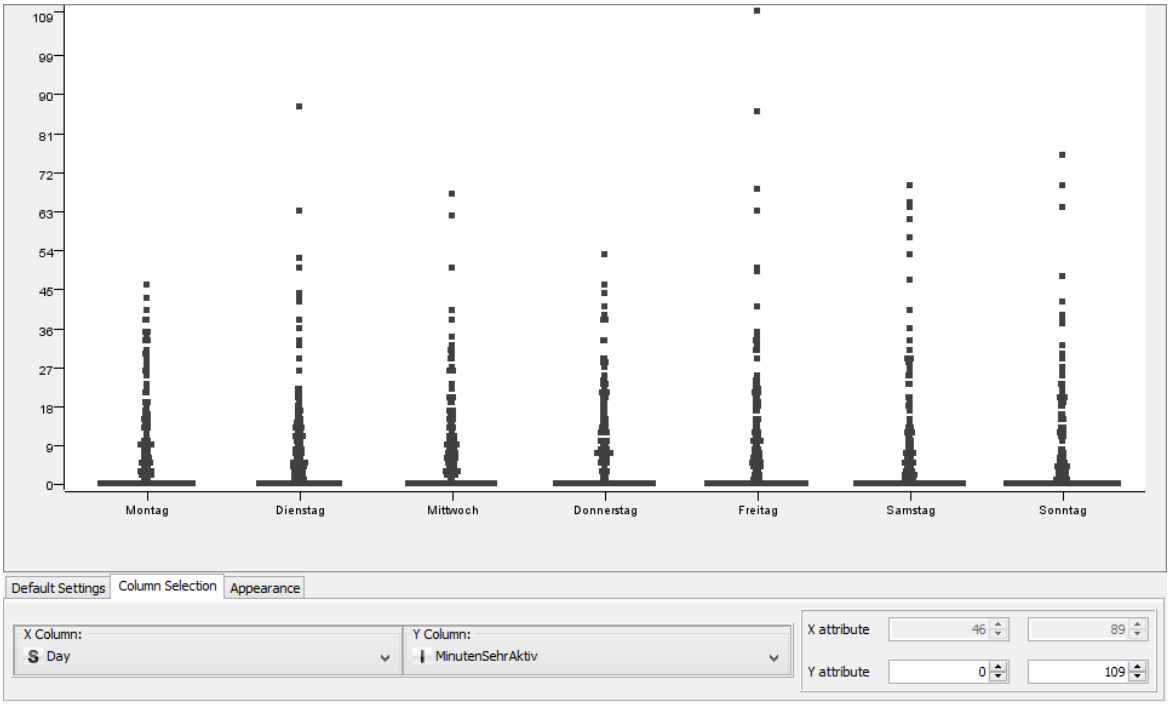

Abbildung 6.1: Minuten Aktiv pro Wochentag Alle Jahre

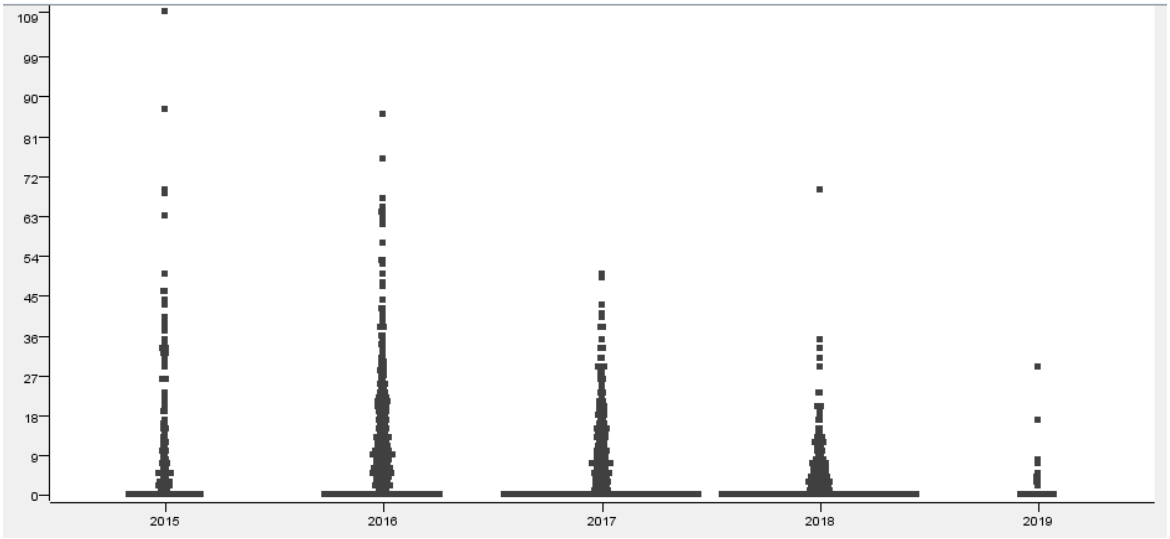

Abbildung 6.2: Minuten Sehr Aktiv Vergleich aller Jahre

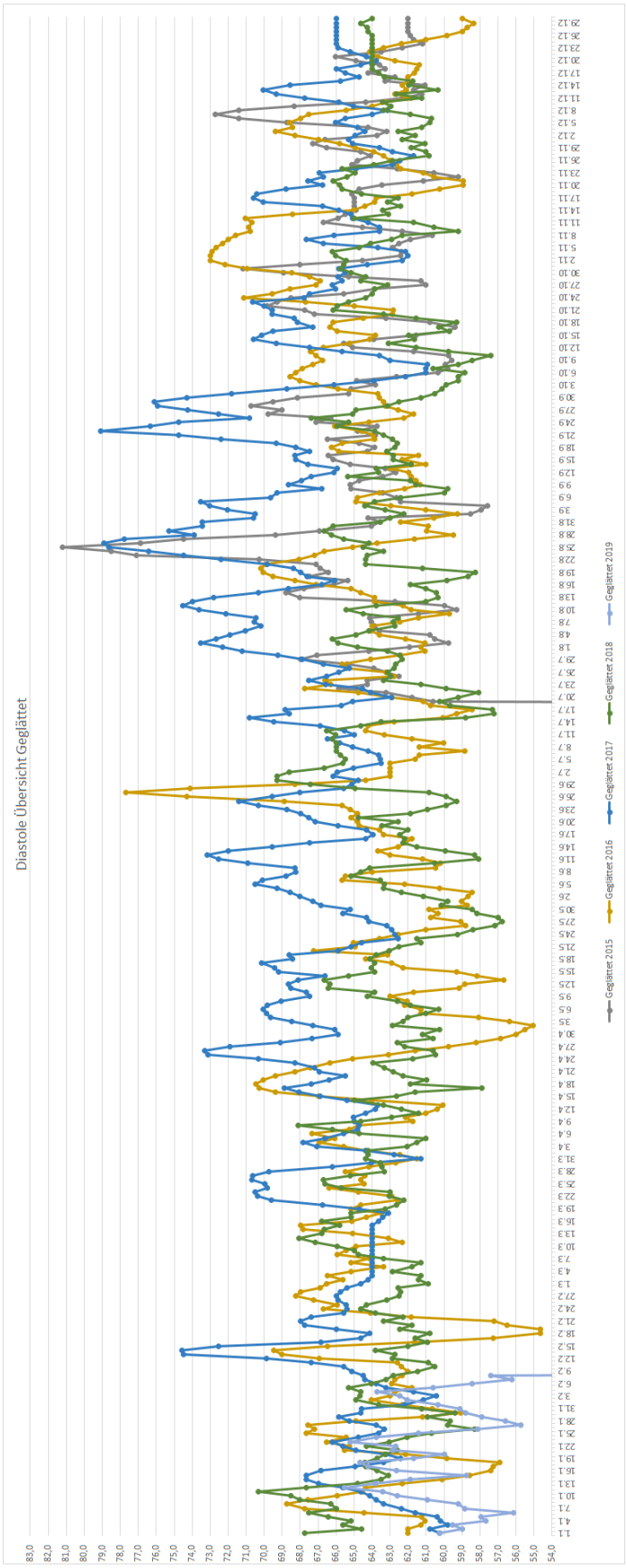

Abbildung 6.3: Gegenüberstellung der Diastole, geglättet über alle Jahre

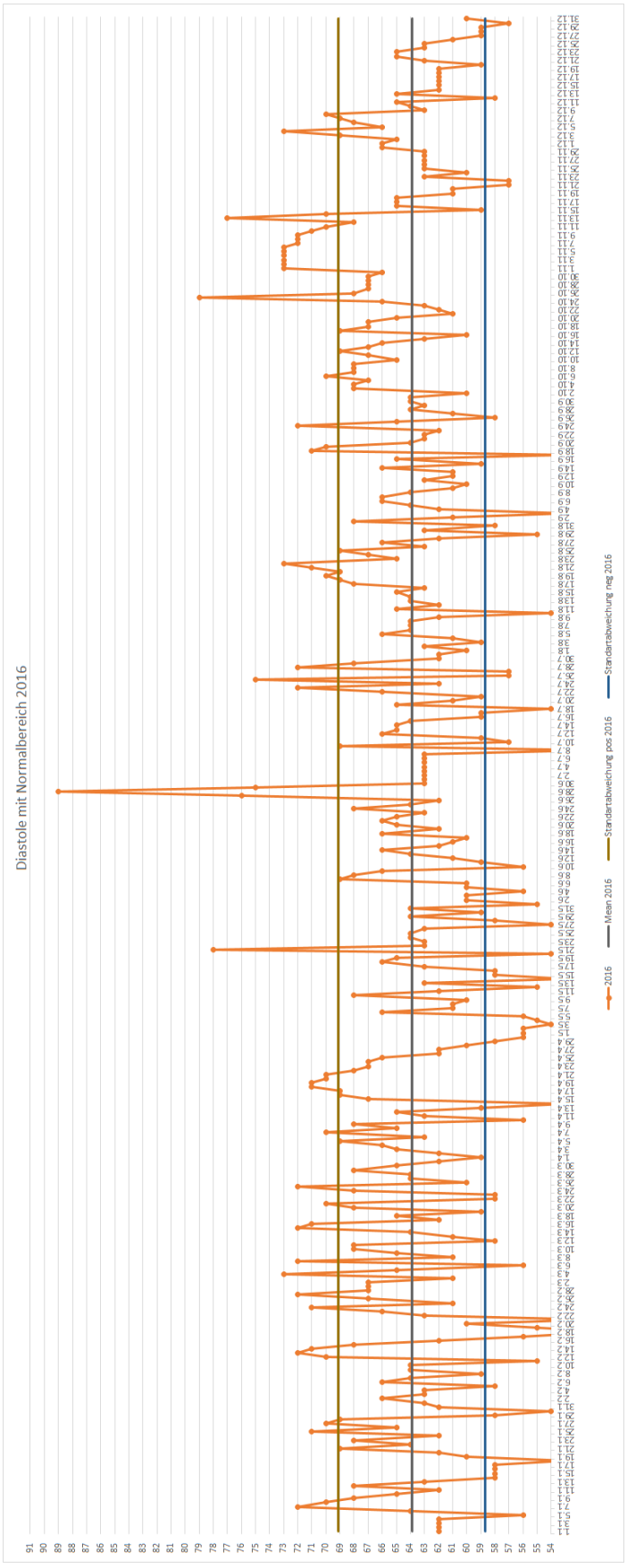

Abbildung 6.4: Diastole 2016 mit Normalskala

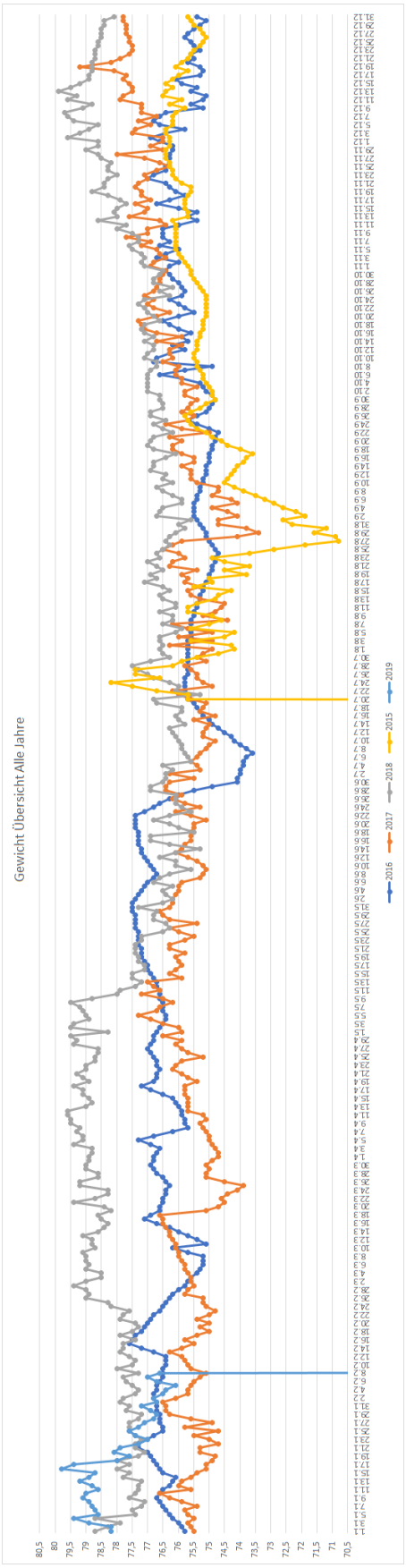

Abbildung 6.5: Gegenüberstellung des Gewichts aller Jahre

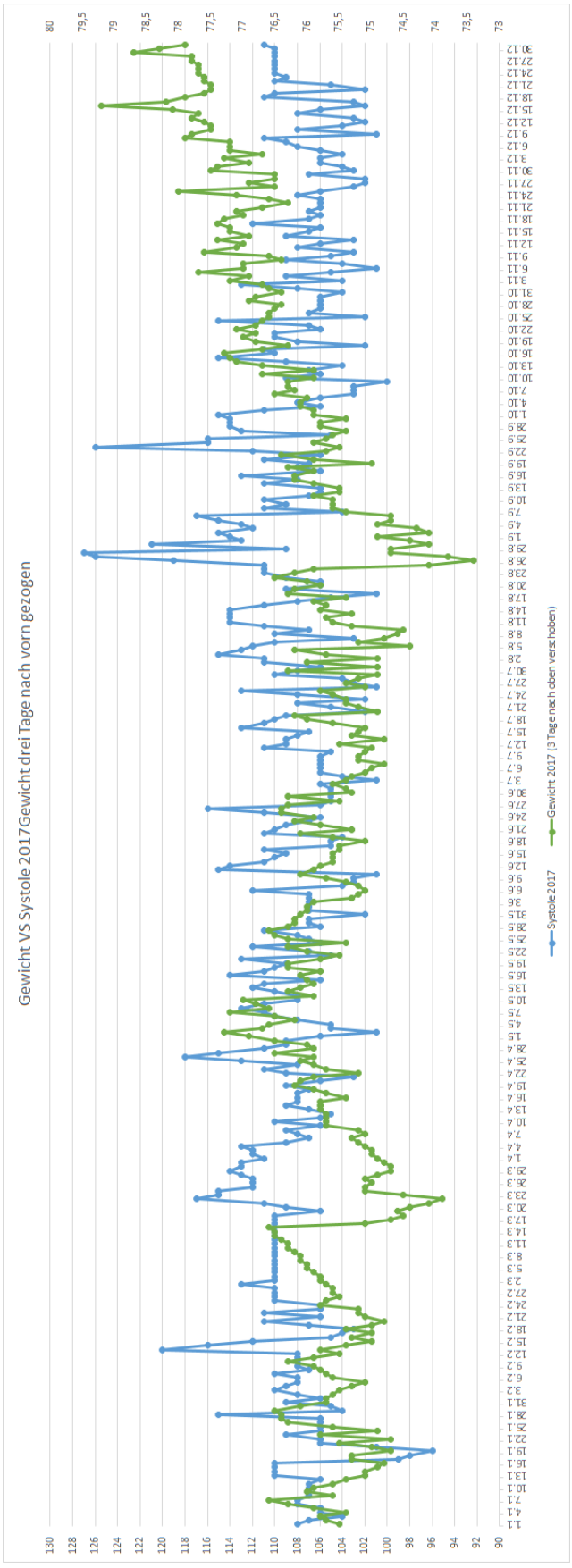

Abbildung 6.6: Systole mit um drei Tage verschobenem Gewicht 2017

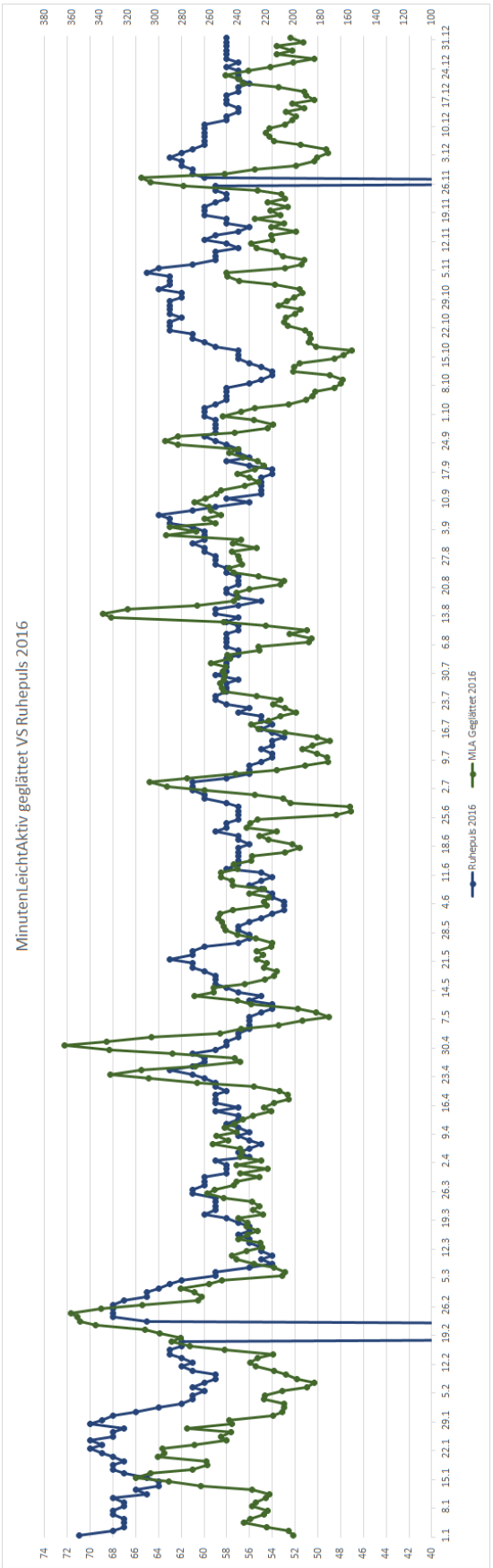

Abbildung 6.7: Ruhepuls und Minuten Leicht Aktiv 2016

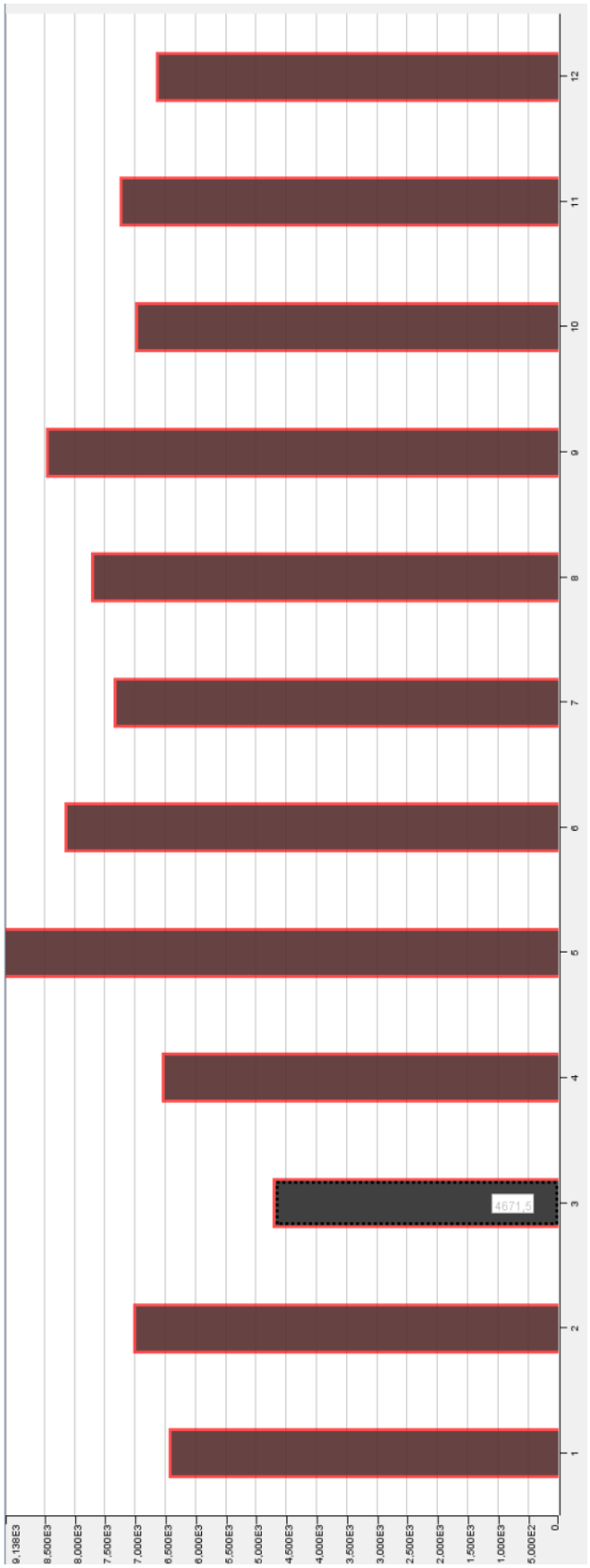

Abbildung 6.8: Schritte, Durchschnitt pro Monat

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit im Sinne der Prüfungsordnung nach §24(5) ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 19. Mai 2020	
Ort, Datum	Unterschrift