Vrije Universiteit Amsterdam

MSc Business Analytics

Research Paper

The Data Science of the Quantified Self

Author:
Niamh
Bermingham-McDonogh

Supervisor:
Prof. Gusz Eiben

December 14, 2015

Preface

This research paper has been written as part of the Business Analytics Master program at Vrije Universiteit Amsterdam. This programme aims to improve business processes through the application of mathematics, computer science and business management techniques. As part of this programme students are required to produce a research paper relating to a problem in these three areas. The research could be based on literature, but may also be extended with own research.

This paper seeks to outline the position of personal analytics and the Quantified Self within the field of Data Science, and to elaborate on the consequences this may have for business analytics. Research is conducted through an elaborate literature review to identify what the Quantified Self is, why a shift towards personal analytics is occurring, what technical and analytical barriers are presented, and what this means for the future of business analytics.

I wish to thank my supervisor, Prof. Gusz Eiben, for his help and guidance throughout this project. He always made time for me and provided helpful suggestions. I also wish to thank Dr. Mark Hoogendoorn for providing the data and further support.

Niamh Bermingham-McDonogh
December 2015

Contents

1 Introduction

The recent decade has seen a substantial growth in the need for business analytics. There has been a large growth in the amount of data being produced, with 90% of the world's data having been created within the last two years [26]. This new commodity has brought with it a wealth of opportunities and has caused the breakthrough of data science into many fields. The study of data through statistical analysis and mathematical modelling has been applied to a wide range of businesses with a view to gain insight into organisational structure and customer behaviour, and often patterns are discovered before the reasons become apparent. Business analytics offers businesses a greater understanding of how to meet customer needs and ensure growth; in essence, it allows businesses to optimise themselves in every aspect.

As with the personal computer and countless previous technologies, data technologies, such as analytics software and data collecting sensors, that were previously only available to large organisations, are now available for personal use. Furthermore, the desire to quantify attributes and to make data-based decisions is shifting from business to personal. People are taking personal growth and improvement into their own hands, and are turning to novel technologies to measure and quantify attributes in order to analyse them. This is a market that is currently being tapped into by the major technology firms and start-ups, as new wearable technologies and life-logging apps are released. This is evidence of an apparent shift from business analytics to personal analytics, leading to the formation of a group of likeminded individuals known as the Quantified Self (QS).

This paper seeks to outline the position of the Quantified Self within the field of data science, and to elaborate on the consequences this may have for business analytics. The Quantified Self is of particular interest from a business and informatics perspective, as many of the obstacles it faces mirror those faced by business analytics in its infancy. For this reason, this paper will focus on the business and informatics aspect of personal analytics, as opposed to the mathematic algorithms that sit behind. In particular, the following questions are addressed. What is the Quantified Self? Why is this shift towards personal analytics occurring? What motivates users to begin measuring and quantifying personal metrics? What technical barriers does the QS movement face? What does this mean for the future of analytics, in particular, business analytics?

This paper is structured as follows. Section 2 introduces the Quantified Self and some common definitions. A brief history of the Quantified Self is provided in Section 3, and Section 4 describes motivations for embarking on personal analytics. Section 5 provides example areas where self-tracking has proved popular and beneficial, and a selective list of QS tools. Section 6 highlights the data challenges that the Quantified Self faces, and explores some analytical methods. Some key players in the QS movement are highlighted in Section 7, and future of the Quantified Self is speculated in Section 8. Section 9 explores some data analytic techniques on an example QS dataset. Finally, in Section 10, the advantages and disadvantages of personal analytics are discussed, with some concluding remarks.

2 The Quantified Self

With powerful technology now ubiquitous within smart-phones and tablets, more people and things are connected to the Internet than ever. This has led to a rise in the number and the diversity of applications of connected devices, with an apparent shift towards those that offer the opportunity for self-improvement. This section explores the definition of the Quantified Self and the role it plays within the field of Data Science.

2.1 Definitions

Data Science is a term often found alongside "buzzwords" like *Data Analytics* and *Big Data*. These terms, and numerous others, are used interchangeably, creating confusion and resulting in different opinions on the definition of data science. In a broad sense, data science is the study of large, unstructured datasets to extract information and knowledge, on which to base decisions [23]. Unstructured data refers to data of all types: sound, textual, photos, and videos, to name a few. Originally, numerical data was analysed in the fields of Mathematics, Statistics, and Computer Science, and applied across various industries. Now, however, technology is rapidly evolving and becoming more readily available to those that previously had no access. Additionally, the Internet of Things is experiencing large growth with an increasing number of devices connected to the Internet. The amount of data being generated on a daily basis is growing aggressively, and awareness of the possible applications of data science is expanding.

Organisations across most sectors have realised the potential of using data science to drive the decision making process, and we are now witnessing a transition towards individuals taking responsibility for the data that they personally generate. Previously, data science was applied in large organisations seeking to minimise costs and maximise profits, for example, marketing departments would analyse their customer data to launch targeted campaigns, resulting in higher revenue. More recently, start-ups and SMEs are focussing their efforts on making data driven decisions. Now, however, tools are available for individuals to optimise several aspects of their own lives, resulting in a new and exciting area for the field of data science.

Big Data refers to large datasets with structures too complex that challenges arise when applying traditional data mining techniques [23]. It is widely believed that these challenges arise as a result of Big Data being high in volume, velocity and variety [17]. Around 2.5 exabytes of data are created each day (one exabyte is one billion gigabytes) and around 90% of the world's data was created in the last two years [26]. The velocity at which this data is being created is increasing as connected devices become the norm, and sharing personal information is socially accepted and trusted. Technology is being used in exciting and innovative ways, such that data is collected in a variety of unstructured forms. Collectively, this gives rise to a number of technical challenges when utilising traditional analytical methods [17]. Big Data also refers to the new supporting technologies that have been developed in order to support this transition, such as open source technologies like cloud computing and Hadoop[1], a framework for processing large amounts of data via distributed computing.

[1]Hadoop, `https://hadoop.apache.org/`, retrieved August 22, 2015.

Cloud Computing is the use of networks of remote servers hosted by the Internet to perform functions that would normally be performed locally [8]. This allows easier storage for large amounts of data, and enables the performance of large scale analytics that would otherwise require local computers with high processing power. The rise of cloud computing has helped raise awareness of the power of analytics by providing organisations the means with which to handle Big Data. As we progress towards creating large amounts of individual data and utilising personal analytics, the Cloud will play an important role in storing and sharing our information.

The Internet of Things is the concept of a network containing everyday objects that are connected, and are readable and controllable via the internet [28]. There is no doubt that we live in a very technological era, with 42.2% of the world's population, and 70.4% of Europe, using the Internet [1]. However, most of this usage is through familiar Internet-connected devices, such as computers, smartphones, and tablets. The Internet of Things encapsulates a wider spectrum of devices that, up until now, have not been connected, including buildings, home appliances, roads, and wearable technology. In 2008 the number of **things** connected to the Internet exceeded the number of **people** on earth [28], and by 2020 it is predicted there will be 50 billion Internet-connected objects [30].

Internet-connected devices hold microprocessor chips and sensors that record and transmit data of various forms, allowing for a wide range of applications that can be categorised into three classes: buildings, automotive and transport, and personal health and environment [28]. Within the buildings class we find homes and buildings that make use of temperature monitoring, connected security management, electricity monitoring and management, smart power meters, and remotely controlled devices such as lighting and appliances. In the automotive and transport class we find connected cars that use the Internet for information and entertainment applications, applications that direct users to open parking spots or available electric vehicle chargers, and traffic management systems. The final class, personal health and environment, has experienced a large growth in recent years and includes self-tracking through wearable technologies, clinical remote monitoring, Wi-Fi scales that record your data, and other biosensing applications [28].

Wearable Technology is technology or computers that are incorporated into wearable clothing or accessories [6]. In some cases, these technologies perform the same tasks that one would expect from a smart phone or tablet. A well-publicised example is Google Glass,[2] an optical head mounted display developed by Google in 2013, that allows users to perform tasks similar to those performed on a smart phone, hands-free via voice commands. However, with increasing popularity, wearable technologies perform specific tasks attributable to the personal health and environment class of the Internet of Things. These wearable technologies use sensors to record an abundance of information about an individual's sleep pattern, calorie intake, movement or heart rate, to name a few. Often, these technologies are paired with software that

[2] Google Glass, `https://developers.google.com/glass/`, retrieved June 18, 2015.

Figure 1: The official logo of the Quantified Self community

allows the user to upload the data and analyse their daily measurements.

2.2 The Quantified Self

The Quantified Self (QS) refers to a movement in which its participants track the biological, physical, behavioural, and/or environmental aspects of their everyday lives [29]. It incorporates technology into everyday tasks to acquire data and quantify each aspect. Quantified-Selfers (QSers) use various forms of *self-tracking* or *self-monitoring* on one or several aspects of their lives, either manually or through wearable technology, with the intention of discovering new information that motivates their future lifestyle decisions.

The term *Quantified Self* was first coined in 2007 by Gary Wolf and Kevin Kelly, alumni of the technology publication *Wired*. In 2008 Wolf and Kelly founded the company *Quantified Self*, with the aim "to help people get meaning out of their personal data" [2]. Since then, QS has developed into a community of users and makers of self-tracking tools, who collaborate and share ideas through local Meetup talks (organised via meetup.com[3]), blogs, and conferences. As by the organisation's mantra, seen in the official logo in Figure 1, QS is summarised as a group of people that seek to find "self knowledge through numbers" [2].

In 2012, Meetups were held in 50 cities around the world [4]. Currently, there are over 200 QS groups worldwide, with a total of 50,267 registered members, as seen in Figure 2, illustrating the rapid growth of self-tracking. The largest Meetups are in San Francisco, New York, and London, which is not surprising given that these cities are known to be technology hubs. Amsterdam hosts the 7th largest Meetup, with 1,154 members [3]. During these Meetups, QSers are invited to present their QS experiences by discussing what they did, how they did it, and what they learned [29]. It is clear from Figure 2 that QSers are concentrated in the developed countries, which is again not surprising given economic status and the availability of the relevant technologies in these areas.

QSers can be considered to fall on the spectrum of QS: from those that sporadically track and keep note of a few biometrics, to those that meticulously track all aspects of their lives and analyse their data regularly. Extreme QSers perform experiments and test hypothesis on themselves. However, this raises the question of whether one can truly perform a fair self experiment given the inability to be objective. If a QSer is knowingly tracking data about themselves then the outcomes could be a result of the placebo effect or the Hawthorne effect [29]. However, when an experiment is focussed on improving a personal aspect, does it really

[3] Quantified Self Meetups, `http://quantified-self.meetup.com/`, retrieved June 18, 2015.

Figure 2: Geographical information of QS Meetups[4]

matter how the outcome was reached? Either way, QS is a catalyst for a lifestyle change that has proved effective.

Rocket Fuel, a provider of artificial intelligence (AI) marketing and advertising solutions, investigated the use of QS tools in a survey entitled *Quantified Self (QS) Digital Tools survey for Consumer Packaged Goods (CPG)*, during Q4 2014 [5]. The survey focussed on highlighting key trends in the use of QS tools among U.S. consumers. It was found that 31% of U.S. consumers use a QS tool (a wearable device, app, or website) to track some form of biometric data. Furthermore, 25% of consumers who do not use a QS tool are interested in using one in the future. Information on the demographic profiles of QSers can be seen in Figure 3 [5]. There is a slightly higher proportion of female QSers (53%) than male (47%), and a similar balance among interested non-QSers (56% female and 44% male). The majority of QSers are caucasian, and the majority own a smartphone and/or a tablet, indicating an interest in technology. The majority of the QSers fall in the age range of 25-44 (58%), whereas the majority of the interested non-QSers fall in the higher age range of 45-65 (54%), perhaps indicating the willingness of younger generations to adopt new technology into their everyday lives more freely than older generations.

Arguably the most important result of the study is the impact that QS has made on its users. Nine in ten QSers say that self-tracking has influenced their health and fitness choices, resulting in increased awareness of habits and progress towards health goals, and an improvement in lifestyle choices [5]. Rocket Fuel further explores how QSers are more likely to respond to direct marketing approaches as they are more open to sharing data with digital advertisers.

[4] *ibid.*

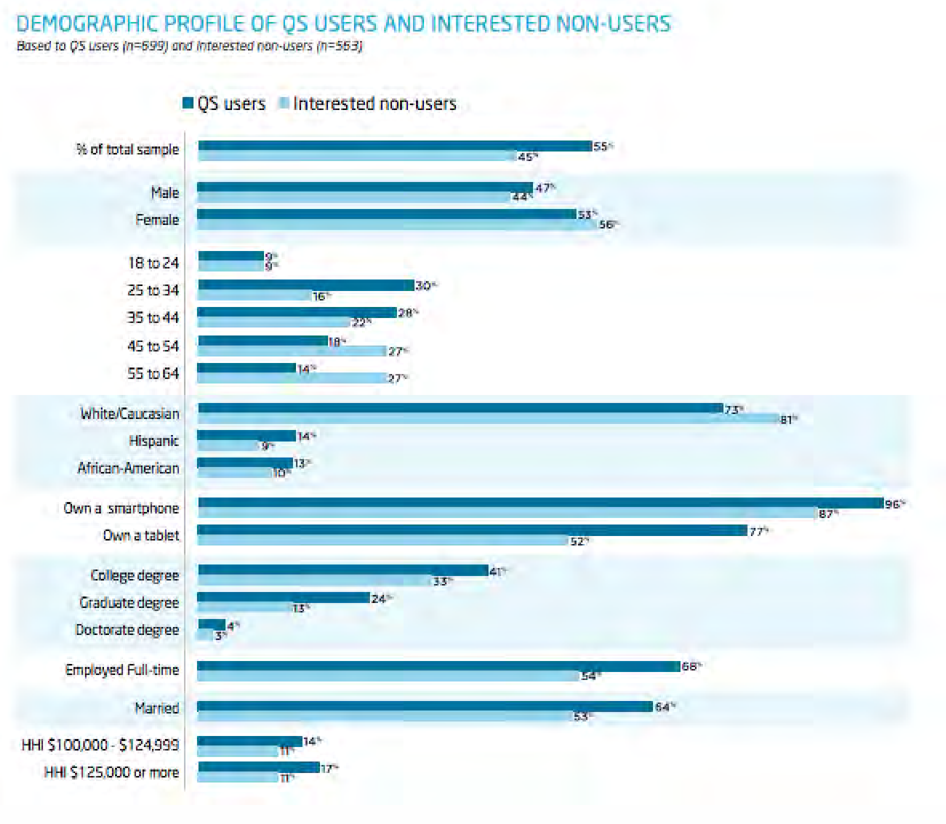

Figure 3: Demographic profile of QSers and interested non-QSers [5]

3 A Brief History of the Quantified Self

The foundations of QS are not as modern as one may think. Some believe that the origins of QS stem from Stoicism, the practice of moral perfection dating back to as early as 3rd century BC [27]. Fundamentally, self-quantifying relies on tracking aspects of an individual's life over a period of time. For years people have tracked their weight using scales, or kept a record of their finances with receipts and check books, or tracked their calorie intake. For centuries people have even tracked thier moods, feelings and experiences using diaries and journals, examples of life-logging. The transition from these examples to QS is a natural progression that coincides with the development of technology [34].

According to Wolf, there are four particular factors within technology that are attributable to the development of self-tracking [34]. First, electronic sensors that enable data collection almost instantaneously are getting smaller and more powerful. Second, mobile phone technology has progressed to the point that powerful devices can be carried around on person all day. Third, social media has become the norm and people are choosing to share information about themselves with their peers and with the public. Fourth, the development of cloud computing has enabled more data to be easily stored, managed, and processed [34]. These attributes explain how the development of the Quantified Self has become possible, but they do not explain why an interest in self-quantifying has emerged. The development of these factors has sparked an interest in the possible new applications of mobile and wearable technology, which has acted as a catalyst for an interest in self-tracking and the QS movement.

3.1 Timeline of Events

1600s
Sanctorious of Padua (1561-1636), an Italian physiologist, physician, and professor, invented medical devices and is said to have introduced the quantitive approach to medicine. Sanctorious tracked his weight and bodily functions every day for 30 years. He constructed a chair that responded to the amount of food he ate, where if he ate too much, the chair would drop out of reach of the food [27]. This is an early example of devices being used to influence the user's behaviour.

1700s:
Benjamin Franklin (1706 - 1790), one of the Founding Fathers of the United States of America, created the "13 Virtues" and meticulously tracked his personal accordance to them on a daily basis, with the goal of morally optimising his choices. He found that by tracking and grading his actions he was able to improve over time [27].

1800s:
1885 saw the invention of the first penny-scales in Germany, and they were available to the general public in supermarkets, and other public spaces [12]. Shortly after, in the early 20th century, weighing scales were a common item found in households, and it was the first time that people could accurately quantify their weight, which had previously been subjective.

1900s: In the 1960s, the first consumer pedometer, named manpo-kei, was created and marketed in Japan [31]. This is the first example of wearable device, used to measure an aspect of fitness and to motivate its users. However, early mechanical models of the pedometer resulted in inaccuracies, and new electronic pedometers were developed in the late 1980s [31]. Professional athletes also began to monitor biometrics through the use of wearable devices, in order to improve their bodies and optimise their performance.

2000s:
The late 1900s saw rapid development in mobile phone technology and by 2002 smartphones were increasing in popularity, especially among professionals. In 2007, Apple released the iPhone to the market, and from then on smartphones were common amongst the population of developed countries. In 2004, Facebook was founded, sparking the digital revolution of social networking. Wearable devices that were once only available to professional athletes became available and affordable to a much broader market, which, accompanied by the development of data technologies previously discussed, resulted in the birth of The Quantified Self.

4 Motivation

Although there has been great interest in what QS is and how it has grown to become a worldwide movement, there has been little focus the motivating factors of its followers. The general consensus amongst various technology publications, is that people choose to track their biometrics as a form of self-optimisation, to better themselves both physically and mentally, to reach certain lifestyle goals, or for medical purposes [4, 34]. Gimpel et al., however, have developed a "five-factor framework of self-tracking motivations" based on the study of 150 QSers [14]. The five motivational aspects are: *self-entertainment*, *self-association*, *self-design*, *self-discipline*, and *self-healing*, and are comprised of a total of 19 motivational components. These factors and components are summarised in this section.

4.1 Self-entertainment

The first motivational factor addresses the entertainment value of self-tracking. The components associated with this factor are centred around the enjoyability of self-tracking as a whole. This includes those who enjoy the novelty of the innovative technical devices that are associated with QS, and those who enjoy the analytical side of experimenting with their personal data [14]. Self-entertainment also encapsulates those who are motivated by the curiosity of new technology and the possibility of discovering new information.

4.2 Self-association

Motivation by self-association includes those motivated by the prospect of comparing themselves against those within the QS community. This is an interesting motivational factor as it is more focussed on the state of a QSer in relation to the QS community, as opposed to one's self. This follows from the idea that without community there is no individuality as there is none against which to draw a comparison. Also included are those who wish to share their data and their success with others for information purposes or to inspire [14].

4.3 Self-design

Self-optimisation is the driving force behind the motivational factor self-design. This includes those who are motivated by the prospect of having more control over certain lifestyle factors, such as exercise and diet, and who want more from the body and their mind. These people want to achieve optimal mental, cognitive, and/or physical state [14].

4.4 Self-discipline

Motivation by self-discipline includes those who are motivated by working towards a goal, and implementing a level of discipline in their life. This includes the idea of gamification, where typical elements found within games, such as point scoring, levels, competition, are incorporated into self-tracking, to motivate the users into continued participation and improvement [14].

4.5 Self-healing

The final motivating factor is self-healing. This factor is associated with those who are motivated to start self-tracking for health purposes, either they were encouraged to do so by their healthcare provider, or chose to do so independently in order to monitor illnesses and symptoms. Among those motivated by self-healing, there appeared to be a lack of trust in the healthcare system and standard medicine, and an interest in personalised medicine, which will be discussed further in Section 5 [14].

5 Tools

The QS website includes a comprehensive guide of 505 self-tracking tools that are currently on the market [2]. These tools are tagged with keywords that identify functionalities and technical aspects, shown in the frequency word cloud in Figure 4. The functionalities can be summarised by the following categories: fitness and health, medical, life-logging, money, and productivity. This section provides some examples of QS tools to illustrate how they are used within the QS community.

Figure 4: Tags of QS Tools

QS tools come in the form of wearable devices, smartphone and tablet apps, or websites. They collect data either passively, without input from the user, such as distance travelled by GPS, or body temperature by wearable sensor, or actively, where input is required from the user, such as mood ratings or weight. According to the Rocket Fuel survey, 16% of U.S. consumers own a wearable device, of which 51% use in conjunction with an app to track the metrics recorded [5]. A quarter of wearable device owners use both an app and a website, and a similar proportion (24%) use only a website in conjunction with their wearable device. On the other hand, 29% of U.S. consumers use an app or a website that is not associated with a wearable device, and 14% stated they were likely to purchase a wearable tracking device within the six months after the survey was conducted [5]. This illustrates the increasing popularity of wearable devices, as they become more available and more affordable.

5.1 Fitness and Health

It is clear from Figure 4 that fitness and health are the two most frequent functionality tags. As previously stated, one of the main motivations for self-quantifying is the idea of *self-*

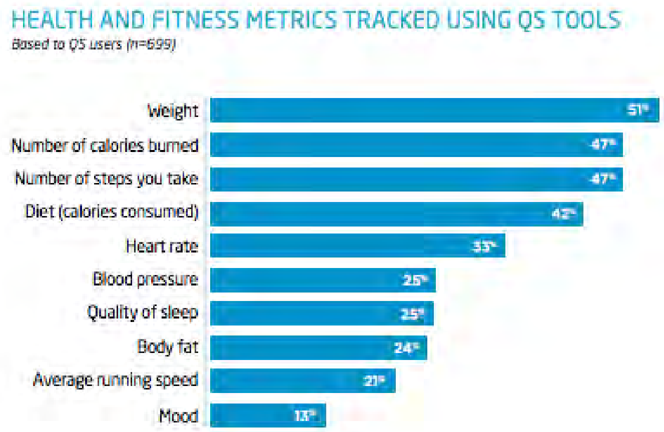

Figure 5: Metrics tracked using QS tools [5]

optimisation, where users seek to improve themselves physically and mentally. Tools found in the fitness and health category provide the user with a platform on which to track various biometrics, allowing them keep track of their progress, and aiding them in reaching their goal of becoming healthier and fitter. Figure 5 shows the health and fitness metrics tracked using QS tools according to the Rocket Fuel survey [5]. Most of these metrics are associated with fitness and weight loss, however there are some associated with overall well-being, such as quality of sleep and mood. These are interesting biometrics to measure, as previously there was no method with which to quantify them.

Fitbit

Fitbit[5] is a wearable device that wirelessly tracks the everyday activities of the wearer. The different Fitbit products available can be seen in Figure 6. The basic functionalities of the Fitbit include tracking steps, calories, distance, and active minutes. More advanced models also track sleep patterns, number of floors climbed, and heart rate. The devices are synced wirelessly to an app and Fitbit dashboard where the user can view their stats, log additional measurements, connect with other Fitbit users and receive notifications to maintain motivation. According to the Rocket Fuel survey, Fitbit is the most popular fitness and health wearable device amongst QSers, and the most well known device amongst non-QSers [5].

Apple Watch

In April 2015, Apple released the Apple Watch[7], a wearable computer with similar capabilities to the iPhone, with the addition of sensors to track the number of steps taken, heart rate, and calories burned. Through its built in Activity app incentive is provided visually: the Apple Watch will measure the amount of time the user has spent standing, exercising and/or moving, and gradually update a ring accordingly. The user can see how close he is to achieving his goal throughout the day, until the goal has been reached and the ring shows a closed circle.

[5] Fitbit Products, `https://www.fitbit.com/uk/compare`, retrieved June 20, 2015.
[7] Apple Watch, `http://www.apple.com/watch/`, retrieve September 10, 2015.

Figure 6: Fitbit Products[6]

Furthermore, when the user reaches personal goals he is rewarded with personalised badges that can be shared with friends, an example of gamification and the use of social pressure to act as an incentive. Apple is a big player in the technology industry, and its recent entry into the QS market is bound to propel the QS movement into the mainstream.

MyFitnessPal

MyFitenessPal[8] (MFP) is a website and an app that is not associated with a wearable device. MFP is a platform where users can create profiles, set goals, and log food that they have eaten and any exercise they have completed. It holds a large database of external metrics, such as food and nutritional information, and sends notifications to users to remind them to log any entires, making the process of self-tracking simple and straight forward. MFP also incorporates a social media aspect, where users can create a profile and share their progress via social networking platforms, or follow other MFP users.

MoodPanda

MoodPanda[9] is a QS mood tracking app and mental health support community, where users can rate their happiness on a scale of 0-10, with the option of including short descriptions explaining their mood. This simple model allows a user to actively track their mood and identify key determinants. It is the first of its kind that quantifies and tracks the mental health of QSers.

5.2 Medical

According to a study on the use of self-tracking among adults with chronic illnesses, carried out in 2014 by the Pew Research Center, 45% of U.S. adults live with one or more chronic illnesses, including high blood pressure, diabetes, heart disease, or cancer [13]. It was found that 80% of adults living with two or more chronic illnesses regularly tracked biometrics including, weight, exercise, blood pressure, blood sugars etc., compared with 70% of those living with one chronic illness and 61% of those with no chronic illness, although the majority of which did not use the aid of technology. Self-tracking can help patients monitor and manage their conditions, and help to identify any external factors that could trigger flare-ups. Further, it is speculated

[8]MyFitnessPal, `https://www.myfitnesspal.com/`, retrieved June 22, 2015.
[9]MoodPanda, `http://moodpanda.com/about.aspx`, retrieved June 22, 2015.

that self-tracking among patients has the power to change the landscape of medical practice and the doctor-patient relationship as we currently know [22].

The use of personal data within medicine can be grouped into three areas. First, through self-tracking tools patients are taking their health into their own hands, to discover elements about their illness and its indicators that would otherwise remain unknown. As with wearable technologies, medical technologies are becoming more advanced, smaller, and more usable, such that patients are now able to measure their biometrics at home, as opposed to in a clinic. These metrics can be tracked against lifestyle metrics, such as food, exercise, and sleep, to highlight correlations and improve lifestyle decisions. This form of *personalised medicine* does not offer the chance for patients to cure themselves, rather it lets them manage their condition and regain control of their lives [22]. Additionally, these tools aid those who suffer from less serious health concerns, such as headaches and fatigue, that affect quality of life [25].

Second, self-tracking offers physicians, other medical professionals, and care-givers, the opportunity to continuously monitor patients, and alter treatments and dosages accordingly [22]. This form of tracking has also proven effective for mental health conditions [4]. Patients can be notified by a mood tracking tool, such as MoodPanda, to update their moods a number of times a day. A notification can then be sent to a care giver or close relative, under the permission of the user, who can offer real time support when needed [4]. This data centric patient-doctor relationship would rely on sufficient data storage and aggregation, from which several issues arise, as discussed in Section 6.

Third, the data that is produced through personalised medicine can be aggregated to form a larger data store that can be used for research purposes [22]. Previously, medical research has had to rely heavily on patient surveys and trust that they are answered correctly and honestly. In the future, the patient data could be made available to researchers to generate information and verify studies.

CureTogether

CureTogether[10] is a website that provides a platform on which patients can share quantitive information about themselves and their treatments, and compare and discuss with others. They can track their illness and compare with similar illnesses to identify factors that might have an effect, expanding the knowledge base from doctors to the general public, an example of crowdsourcing. CureTogether also acts as a patient-contributed database for research projects.

Vitality

Vitality[11] produce consumer products, namely the GlowCap, a prescription bottle cap, and the GlowPack, a pouch, that audibly and visually reminds patients to take their medication. The products are wirelessly connected to the internet, and glow and produce sound when the user is scheduled to receive a dose of medication. They also come with plug in devices so there is no forgetting should the bottle by stored in a cupboard. If the patient does not take the medication within a certain amount of time, a text message is then sent to the user, or an automatic call is made to the user's home phone as a reminder. Further, there is a button

[10] CureTogether, `http://curetogether.com`, retrieved June 23, 2015.
[11] Vitality, `http://www.vitality.net`, retrieved June 23, 2015.

at the bottom of the GlowCap that, when pressed, notifies the local pharmacy via a mobile broadband connection to confirm your refill.

SenseWear

SenseWear[12] is an armband created for patients such that their doctor can receive real-time information about their biometrics. The SenseWear passively measures energy expenditure, sleep duration and efficiency, motion, and skin temperature to name a few. The data produced by the armband can be analysed by the doctor and shared with the patient. This encourages patients to comply with the lifestyle recommendations of the doctor, but is also a reliable method for research groups to collect data without much active input from the patients.

Ginger.io

Ginger.io[13] is a mobile phone app that collects data about patients, both passively and actively. It can monitor how often the user texts, talks on the phone, when patients lock and unlock their phones, and distance travelled, to name a few, and is used to predict if patient prone to mental illness might become ill. It has been used in the healthcare industry to monitor new mothers to predict and identify cases of postpartum depression, which is currently an under-diagnosed condition [33]. The tool notifies the care providers when it feels the patient's health may be at risk, who can then provide the support needed.

5.3 Life-logging

Life-logging refers to keeping track of many aspects of ones life for personal interest, to gain insights, or for memory purposes. QSers can track their lives through any inputs, such as text entries, photographs, or numeric data. In many ways, life-logging has been present for many years, through journal and diary writing, photograph albums, and video recordings. Now most life-logging is done digitally, and also includes a social aspect, as we share more information on social networks.

Facebook

One of the most popular forms of digital life-logging is the social networking site Facebook[14]. Users can connect with each other and share information about their thoughts, location, photographs, videos and events. Facebook also has the option to connect with other websites and applications, such that the user can build an extensive digital profile and network.

Momento

Momento[15] is a digital journal app that aggregates entries and information into one platform, to record activities and memories. The user can input text entries about their day, tag people they have met and places they have visited, use GPS to track their movements, and attach photos. Momento also provides to option export the data for analysis or to store elsewhere.

[12] SenseWear, `http://sensewear.bodymedia.com`, retrieved June 23, 2015.
[13] Ginger.io, `https://ginger.io`, retrieved June 23, 2015.
[14] Facebook, `https://www.facebook.com/`, retrieved June 23, 2015.
[15] Momento, `http://www.momentoapp.com/`, retrieved June 23, 2015.

5.4 Money

By its very nature, personal expenditure is quantifiable and is one of the earliest self-tracking techniques. Previously, people relied on balancing checkbooks and bank records, whereas now accounts are automatically updated with a list of transactions. The QS movement incorporates monetary metrics, offering QSers the opportunity to track this information and analyse it to highlight patterns in their spending habits. Further, there are opportunities for QSers to set goals in spending and saving, and record their purchases, allowing them to see if they are on track. The success of these monetary tracking products relates to human behaviours of realising information when facts are written down and difficult to ignore [4].

Mint

Mint[16] is a website that automatically merges all of your financial transactions and accounts in one interface. The user can explore their spending habits through various data visualisations, and set spending and saving goals.

5.5 Productivity

There are a number of tools that are available for tracking productivity, both actively and passively. These tools measure how long you spend doing various tasks: working, checking emails, browsing the internet, Facebook etc. This allows the user insights into where and when they are most productive, and they can then deduce reasonings as to why that might be the case. Again, seeing the facts in front of you are difficult to ignore, and as a results QSers can increase their productivity by aiming to decrease the amount of time wasted on unimportant tasks.

RescueTime

RescueTime[17] is an application that tracks computer activities and creates a weekly report summarising the time spent of different applications. It offers the opportunity to create goals for time spent on productive activities and distracting activities.

[16] Mint, `https://www.mint.com/`, retrieved June 23, 2015.
[17] ResceTime, `https://www.rescuetime.com/dashboard`, retrieved June 23, 2015.

6 Data

The tools outlined in Section 5 illustrate the ease at which QSers can collect vast amounts of personal data. However, barriers still arise throughout the self-tracking and analysing process. As with any process that involves human input, there is room for human error, for example forgetting to turn on a wearable device, forgetting to start recording from an application, or manually inputting data incorrectly, which will lead to incomplete datasets [18]. Furthermore, QSers that are measuring a variety of metrics using many tools, may encounter obstacles when their data is stored in different types of databases, and in different formats. Finally, the level of insights that a QSer can attain depends on their knowledge of data analytics and algorithms [29]. Some QSers may need an analytics interface that calculates the key patterns and results automatically, whereas others may prefer to handle the raw data and perform personalised analysis [24].

The ideal QS experience would include total automatic and immediate data collection across all possible self-tracking platforms, with streamlined data integration and aggregation, and an appropriate analysis platform. This leads to apparent similarities between QS, Big Data and Business Analysis, in the hurdles with which they are faced. It is difficult to tell exactly how much data is generated through self-tracking. Cisco forecasts that by 2018 mobile phone traffic will exceed 190 exabytes (approximately 11 times more than in 2013), and attributes this surge to the increase in number of things connected to the internet [7]. Furthermore, they predict that by 2019, two-thirds of all data across the internet will originate from devices that are not PCs, in particular smartphones will experience a growth of 63% and machine-to-machine devices (including wearables) will experience a growth of 71%. Although QS data on an individual level is not comparable to Big Data in size and the velocity at which data is being produced, considering all QS data as a whole, it is clear that it can be considered to be Big Data. In this section the typical data management problems that QSers face are reviewed: data storage, integration, aggregation, and analysis.

6.1 Data Storage

As the popularity of self-tracking apps continues to increase, resulting in a surge in the volume of personal data, QS tools and their providers must offer transparency on where consumer data is stored and who has access to it. Often the data that is created by the self-tracking tools is first temporarily stored on the device, then transferred to a cloud-based storage system. Once the data arrives at the cloud-based storage, it is preprocessed and stored in a central database. The problem is that once the data is stored in the Cloud, QSers currently have very little control or visibility of how it is managed, and what is being done to defend against cyber-attacks [10]. The security of ones data is further explored in Section 5. Furthermore, QSers are faced with the difficulty of storing their own data for personal analysis purposes, as their data may be too large to manage locally. Swan states that a short term solution for this problem would be cloud based services for individuals' self-tracking data storage, which allows easy data integration and aggregation [29]. Furthermore, QS tools store their data differently, some store data on devices or in databases, whereas others use cloud based systems [24]. This

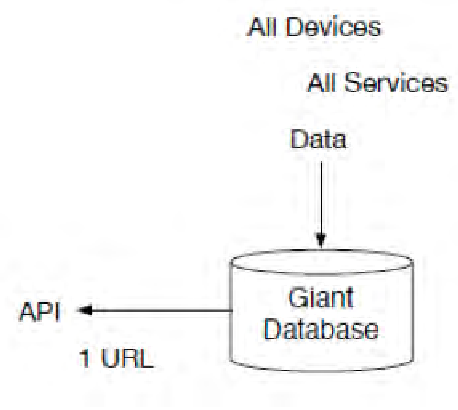

Figure 7: Ideal data storage and integration [20]

adds to the complexity of data integration and the overall process of self-quantifying.

6.2 Data Integration

In addition to heterogeneous data storage, the use of different QS tools will result in heterogenous data. Data from different sources exist in different forms, for example, biometric, sound, images, textual, and GPS data, that may not lend themselves to straightforward integration. Furthermore, different self-tracking tools may record data of the same type differently. Ideally, the QSer will be able to take all data from each of its sources and consolidate it into one main database, on which to perform their analysis, as seen in Figure 7 [20]. Unfortunately, obstacles may be encountered due to difficulty in accessing some local and distributed data sources. Some QS tool providers allow for the user to export their data into Excel or CSV files [20]. However, some only allow users to access the data through an API, making it much more difficult to integrate with data from other sources. Some providers even only allow access through a third party [24].

Figures 8 and 9 show the mappings from devices to applications to services of FitBit and and RescueTime respectively [20]. As we can see, the mapping process is unique to each tool. In this case both tools allow for CSV export, with FitBit offering it as a premium service. FitBit limits the access of its API to registered third part applications and partner programs, whereas RescueTime is much more flexible allowing its API to accessed by any user-authorised application. If an experienced QSer wanted to integrate his FitBit data with their RescueTime data he could export the data to a CSV (at the expense of paying for the FitBit premium package) and perform his analysis. However, a QSer with limited data analytic skills would need to go through one of FitBit's approved third party applications and sync his RescueTime data. Finding an application appropriate for both sets of data is a task in itself, and would be of greater magnitude the more QS data tools involved. This highlights the data integration problems QSers are currently faced with.

It is important that QSers can integrate the different streams of data in order to reach the full potential of QS. The primary goal within data integration is to find an effective way to

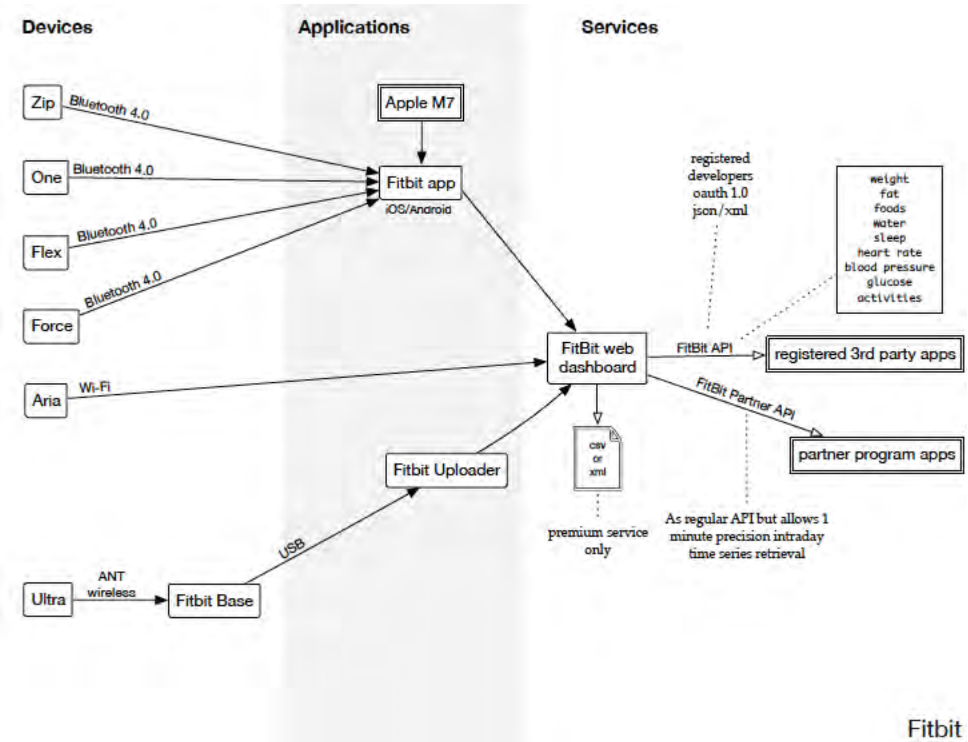

Figure 8: FitBit API [20]

merge and integrate all streams of personal data. For the healthcare industry this translates to merging traditional medical data, genomic data, and QS data, which will provide a comprehensive view of patients' health [29]. Full data integration could see individuals' gaining access to their own extensive personal dataset, including, for example, social media data, financial data, and biometric data, providing the opportunity for full self analysis and insights that previously have not been possible to attain.

6.3 Data Aggregation

Although the Quantified Self focuses on data representation of an individual, there is still a need for effective data aggregation of multiple QSers. Returning to the motivations for self-tracking discussed in Section 4, QSers may wish to compare themselves with others. This would require a repository of other QSer's data of the metrics with which they wish to compare. Some tools provide dashboards for their users to compare and rank themselves against the other users of that tool, however, comparisons across tools becomes much more difficult due to ownership and privacy issues of such data. It is the providers of QS tools that hold a powerful position by sitting on large aggregated databases, that can be viewed as knowledge and potentially sold, as will be discussed in Section 8.

The aggregation of QS data allows society as whole an overall view of smaller groups of people to gain a better understanding. The most prominent, and arguably the most important, example is within the field of medicine. The aggregation of QS data has potential to change and improve the concept of preventative medicine. Aggregating multiple streams of personal QS

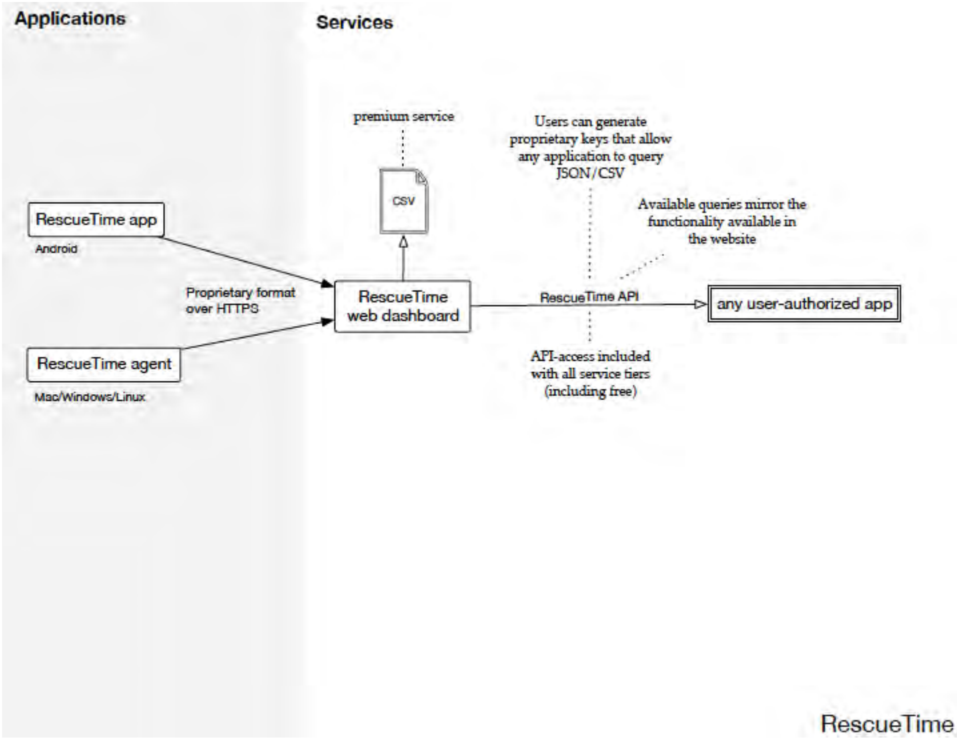

Figure 9: RescueTime API [20]

data, similar to CureTogether, could result in QS applications that offer automatic suggestions based on the information and insights gained from the wider dataset [29]. For example, if there existed a large integrated and aggregated dataset, containing lifestyle data and medical data, of a group of patients suffering with the same chronic condition, real-time suggestions could be sent to users depending as a result of the input data and the insights discovered. In some sense, this contradicts the idea of personalised medicine, as it offers suggestions based on others' experiences. However, by incorporating QS data with traditional medical data, suggestions can be made more specifically for clusters of patients within the overall dataset.

6.4 Data Analysis

The next problem that arises is how QSers can make sense of all the data that they have collected. As with most data analysis projects, a key issue is extracting the relevant information from the noise [29]. Many QS tools are linked to a personalised dashboard that provide key statistics and visualisations of the data they create, enabling users to reflect and make relevant lifestyle choices. However, as discussed in preceding sections, there is a gap in the market for an affordable tool and service that can store and integrate all personal data streams, in addition to environmental data streams, to provide information and insights from all perspectives. It is the responsibility of the QSer to apply analytical techniques to capitalise on their data. Some basic analytical techniques are describe in this section.

6.4.1 Exploratory Analysis

Prior to any formal statistical analysis and testing, it is beneficial for the user to explore the dataset and highlight the key characteristics. The goal of such exploratory data analysis (EDA) is to highlight any patterns and relationships, by exploring all perspectives of the data, to generate hypothesis and develop models [11].

EDA primarily involves calculating the summary statistics and generating various graphical representations of the data. Summary statistics include the *mean*, *median*, *mode*, *minimum*, *maximum*, *range*, *variance*, and *skewness*, amongst others. Further, this information can be illustrated graphically, examples of which are outlined below.

Box plots provide analysts with a visual spread of the data. One can compare the median, quartiles and range, and highlight any outliers. Further, box plots allow the user to group the data according to categorical attributes, in order to compare the spread of the data for each category [11]. For example, a QSer may be interested in whether drinking alcohol had an affect on their total length of deep sleep. The data can be grouped into two categories, days on which they consumed alcohol, and days on which they didn't, allowing the user to quickly identify if there is grounds for further statistical tests.

Scatter plots display individual data points according to two variables and are useful in identifying possible relationships [11]. Returning to the previous example, a QSer may be interested in how different volumes of alcohol affect their total length of deep sleep. In this case, the volume of alcohol consumed can be plot against the length of deep sleep. A scatter plot will highlight linear or non-linear relationships, randomness, clusters, and whether the effect plateaus after a certain volume.

Line plots connect the data points for one variable over a period of time, to produce a continuous line. Line plots can be used to highlight time periods where unexpected measurements are recorded for the variable of interest, which could be further investigated. A QSer could, for example, plot their length of deep sleep over time to identify any trends or seasonality. If they spot an unexpected result that significantly deviates from the rest of the data, they could focus their investigation on identifying the contributing factors.

Bar charts illustrate the numerical values of a categorical variable, where the lengths of the bars are proportional to the values that they represent, thus making it easy to compare categories. Returning to the example, a bar chart could be used to illustrate the number of different types of drinks consumed (e.g. wine, spirits, non-alcoholic etc.), increasing awareness to drinking habits.

6.4.2 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique that reduces the number of dimensions of a dataset by transforming correlated variables to uncorrelated variables, known as the principal components. The principal components are ordered, such that the first few are accountable for the most of the variation found in the entire dataset. PCA can also be used for identifying associations between variables within the dataset, that would otherwise be difficult to visualise [16].

Let a data set consist of n observations and p correlated variables, X_1, X_2,..., X_p. The aim of PCA is to find a new set of uncorrelated variables Y_1, Y_2,..., Y_q, where, in the case of dimension reduction, $q < p$, and Y_1, Y_2,..., Y_q account for most of the total variance. Each principal component Y_i is a linear combination of X_1, X_2,..., X_p. The sample covariance matrix $\mathbf{Q}$ is defined with entries as follows:

$$q_{ij} = \frac{1}{n-1} \sum_{k+1}^{n} (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}), \tag{1}$$

for $i = 1, 2, ..., p$, $j = 1, 2, ..., p$, and where $\bar{x_j} = \frac{1}{n} \sum_{k=1}^{n} x_{jk}$ is the sample mean for variable j. Let Y be a linear combination of X_1, X_2,..., X_p, such that $Y = \mathbf{a}^T\mathbf{X}$. The aim is to find the Y that is accountable for the maximum variance. To calculate $\mathrm{Var}(Y)$, $\mathrm{Cov}(\mathbf{X})$ is needed, which is estimated by $\mathbf{Q}$:

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbf{a}^T\mathbf{X}) \approx \mathbf{a}^T\mathbf{Q}\mathbf{a}. \tag{2}$$

To find the linear combination that accounts for the maximum variance, the following constrained optimisation problem is considered: maximise $\mathbf{a}^T\mathbf{Q}\mathbf{a}$ where $\mathbf{a}^T\mathbf{a} = 1$. Solving this optimisation problem, we find that $\mathbf{a}$ is the eigenvector of $\mathbf{Q}$ with largest corresponding eigenvalue. This eigenvalue represents the variance of the linear combination [16].

To summarise, the first principal component is the linear combination of the original variables X_1, X_2,..., X_p, where the weights are determined by the eigenvector of $\mathbf{Q}$ corresponding to the largest eigenvalue. The second principal component is the linear combination with weights determined by the eigenvector corresponding to the second largest eigenvalue, and so on [16].

PCA is a useful statistical technique for QS, as it reduces the dimensionality of the dataset significantly, without sacrificing a lot of information. If a QSer measures numerous aspects of their life, the dimensionality of their personal dataset can grow to be very large quite quickly. Presuming the QSer has limited data analysis tools available, PCA can reduce the dimensionality, making the dataset more manageable for analysis.

6.4.3 Correlation Analysis

Correlation analysis refers to the study of the relationships between variables of observed data. With respect to QS, correlations can be very useful to identify lifestyle factors that are affecting the variable in which you are interested in monitoring or optimising, or if a change in one variable coincides with a change in another variable [9]. Correlation can either be *linear*, when the ratio of change remains constant between the two variables, or *non-linear*, when the ratio of change is not constant. Correlation is also either *positive*, when an increase (decrease) in one variable coincides with an increase (decrease) in the other variable, or *negative*, when an increase (decrease) in one variable coincides with a decrease (increase) in the other [9]. The degree of correlation is calculated by the correlation coefficient, which can measure linear association or non-linear association, depending on the correlation measure used. The two most common measures are Pearson's correlation coefficient and Spearman's rank correlation coefficient.

Pearson's Correlation Coefficient

Pearson's correlation coefficient between variables X and Y is defined as follows:

$$r = \frac{\text{Covariance between X and Y}}{\text{(Standard deviation of X)(Standard deviation of Y)}} \tag{3}$$

$$\Rightarrow r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}}, \tag{4}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of variables X and Y respectively [9]. This results in a value between -1 and 1, enumerating degrees of correlation. When $r = 1$ there is perfect positive correlation: a change in one variable always coincides with a change in the other variable in the same direction. When $r = -1$ there is perfect negative correlation: a change in one variable always coincides with a change in the other variable in the opposite direction. When $r = 0$ there is no correlation. The closer $|r|$ is to 1, the stronger the correlation between the two variables. Pearson's correlation coefficient, however, does not offer any information on whether the correlation between two variables is due to a change in one variable directly causing a change in the other, or whether both variables are mutually affected by an external factor or variable. Furthermore, Pearson's correlation coefficient only determines linear correlation between two variables, and does not determine non-linear correlation.

Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient measures the correlation between two variables based on their rank, i.e. the order of values when sorted. Spearman's rank correlation coefficient can be used to measure non-linear correlation, and the correlation between variables that are not quantitive but can be ranked in some form. For example, a QSer may record their moods (happy, ecstatic, excited, irratic, annoyed, grumpy etc) and rank them in order of happiness [9]. Spearman's rank correlation coefficient between variables X and Y is defined as follows:

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)}, \tag{5}$$

where N is the number of pairs, $D = R_x - R_y$, and R_x and R_y denote the ranks in x and y (1 for highest rank, 2 for second highest etc) respectively [9]. This results in a value between -1 and 1, as before, and can be interpreted similarly to Pearson's correlation coefficient.

6.4.4 Regression Analysis

Regression analysis refers to a statistical technique that models the relationships of variables mathematically. The aim is to model a dependent variable, depending on several independent variables, and in doing so, describe the relationship between them [9]. Therefore, regression can be used to predict the value of one variable based on its relationship and correlation of dependant variables.

Similar to correlation analysis, the relationship between a dependent variable and the independent variables can be *linear* or *non-linear*. Linear regression models data where the dependent variable changes at a constant rate for each unit change of independent variables, and can therefore be represented graphically by a straight line. Non-linear regression is more general and is built to model the data as closely as possible. This can be represented graphically by a curve other than a straight line [9]. Further, if one independent variable is used to model the dependent variable, it is said to be *simple regression*, whereas if two or more independent variables are used to model the dependent variable, it is said to be *multiple regression* [9].

One of the more popular methods for building a regression model is the *Method of Least Squares*. Let Y be the variable we wish to model, the dependent variable, and let $X_1, X_2, ..., X_k$ be the k independent variables. To model this as a linear regression, Y can be calculated as follows:

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + .. + a_k x_k, \tag{6}$$

where $\hat{y}$ is the predicted dependent variable value, based on $x_1, x_2, ..., x_k$, the independent variable values, and $a_0, a_1, ..., a_k$ are the weights. The weights are determined by minimising the sum of square errors between the actual values y and the predicted values $\hat{y}$ (n observations):

$$\text{minimise} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

Once the regression model is built, predications can be made for future values. This could be of great use to QSers. To reach a certain value of one of their biometrics, the QSer can continue the regression line to deduce the values of the dependent variables needed in order to reach their desired outcome.

6.5 Summary

There are still numerous issues regarding QS data in general that need to be resolved for the analytical side of QS to become as streamlined as possible. To resolve the issue of heterogeneous data types and storage, QS tool providers need to work together to ensure that the integration of all QS data is possible. However, with the increase in popularity of the QS movement and the rapid of advancement in data management technology, a solution should be on the horizon.

In terms of analytics, data visualisation is the first key step to gaining insights from QS data. Effective visualisation show immediate relationships, but can also raise questions and highlight areas for further investigation. Furthermore, analysis of QS data can occur on several layers. At the bottom layer QS devices and services can analyse the raw data for each individual based on the metrics that they measure. One level higher is the integration all data sources and external factors, for example, geographical location and socio-economical factors, for a more complete analyses on an individual level. Finally, another level higher is the aggregation of all data of that type, and the opportunity to calculate summary statistics to analyse groups of QSers.

7 Key Players in the Quantified Self Movement

The QS community is a network of users, developers, healthcare professionals, and investors, interested in the possibilities of self-tracking. At the core of this network is the Quantified Self company that hosts the Meetups and provides the platform for communication between users developers. This section explores the key players within the QS movement and how they relate to each other.

7.1 Users

The demographic of QSers has changed as the QS movement has developed. Self-tracking was previously a task undertaken by those who sought health and medical benefits to improve chronic illnesses, and by professional athletes seeking to better their performance and physical state. In its infancy, the QS movement expanded to include those with a keen interest in technology, and those with an interest in the social aspects of self-tracking [34]. However, as self-tracking tools become more affordable and available, and awareness of QS is expanding, interest in self-tracking is reaching a wider audience. As the notion of self-tracking becomes the norm, data analytics that was previously found in the context of business is shifting towards the users, to form a new area of personal analytics.

7.2 Developers

7.2.1 Tech Companies

The majority of the tools stem from technology companies that have the expertise in hardware, software and data aggregation. Apple recently released the Apple Watch[18] which integrates the features of a smartphone, and a heart rate sensor in a wearable device. Jawbone[19], a technology company that produces audio devices and wearable products, is a key played in fitness wearables. Similarly, Fitbit[20] are a technology company solely concentrated consumer wearable devices. Finally, Samsung are an example of an electronics company that has expanded their products into QS and wearable device, with the Samsung GearFit,[21] a wearable fitness band.

7.2.2 Fitness Companies

Fitness companies are also developing their own apps and tools to be used within the QS community. Nike have produced the Nike+[22] fitness app to encourage users to reach their fitness goals using social and gamification techniques, and the Nike+ Fuelband, a wearable tracking device (although it is no longer on the market).

[18] Apple Watch, `https://www.apple.com/watch/`, retrieved June 26, 2015.
[19] Jawbone, `https://jawbone.com/`, retrieved June 26, 2015.
[20] Fitbit, `https://www.fitbit.com/uk`, June 26, 2015.
[21] Samsung GearFit, `http://www.samsung.com/global/microsite/gear/gearfit_features.html`, last retrieved June 26, 2015.
[22] Nike+, `http://www.nikeplus.com.br/`, retrieved June 26, 2015.

7.2.3 Start-ups

There are many start-ups and independent app developers producing tools for QS and to enhance the functionalities of wearble devices. Many of the apps and tools discussed in Section 5 were produced by start-ups, for example MoodPanda and Momento. Investors are interested in self-tracking and QS projects, and are investing more in start-ups with innovative QS ideas. This is encouraging for those interested in innovation within healthcare, as it is a sector in which funding is often difficult to obtain [15].

7.3 Research Organisations

The Quantified Self as an organisation acts as a research group: they facilitate the exposure of new projects, and offer a platform for the development of new ideas [2]. The Quantified Self Institute[23] is a multidisciplinary research group with a focus of self-tracking for personalized medicine, based in Groningen, the Netherlands. They work closely with research, health and industry organisations to find the meaning and usefulness of self-tracking to the healthcare industry. DIYgenomics[24] are a research organisation with an interest in the application of QS and self-tracking in the healthcare industry, and personalised medicine. They are currently running a number of crowdsources research studies, illustrating the transition from the quantified self to the quantified them, discussed in Section 8.

[23] Quantified Self Institue, http://www.qsinstitute.org/, retrieved June 26, 2015.
[24] DIYgenomics, http://diygenomics.org/, retrieved June 26, 2015.

8 The Quantified Future

Since the start of the QS movement in 2007, self-tracking has developed from an activity in which a small proportion of the the population were taking part, to a worldwide movement in the development of a new self-optimising state. These QS years have seen a rapid development of technology and a growth in wearable technology introduced to the consumer market. The fast pace at which this new and exciting field is growing leads to the question: what is on the horizon for QS, both in the near and distant future? In this section some interesting predictions on what the future may hold for QS are discussed.

8.1 Future QS products

Swan believes that QS will shift towards becoming the Qualitative Self, where QSers will focus on tracking subjective metrics such as, mood, happiness, emotions, and their productivity [29]. This include the tools discussed in Section 3, where QSers can actively monitor these metrics using descriptive words and phrases, or by translating to a quantitative scale. However, there have also been significant developments into sensors that can passively extract these biometrics. An example is the mapping of EEG (electroencephalography) brain signals onto key emotions, which can then be used to determine happiness and moods. Therefore, there is the possibility of qualitative data being produced passively by real time with the improvement of technology [29].

Swan also argues that as humans are naturally programmed to think qualitatively as opposed to quantitatively, there is a gap in the current market for tools that incorporate both dimensions. Tools could have statistical elements and data driven recommendations, but produce qualitative output. Swan provides an example of users logging emotional reactions to aspects of a neighbourhood (e.g. potholes) that also incorporate location. These emotional metrics could be aggregated to produce a new variable that, in turn, could be used in studies to provide another viewpoint of that neighbourhood [29]. This illustrates the possibility of the natural progression from QS to the Quantified Us or the Quantified Them.

A more radical proposition, which was first proposed by Kevin Kelly and further discussed by Swan, is the development of a person's individual state to include external factors in which technology and the surrounding environment are incorporated, via "exosenses" [29]. An example is wearable temperature sensors that are linked to a thermostat that will automatically adjust the temperature of the environment depending on if the user is too hot or too cold. Swan describes this expansion of the self to become the exoself, a fourth-person perspective [29]. This would see the Quantified Self and the Internet of Things merge into one connected state.

8.2 Quantified Healthcare

As previously discussed, it is expected that there will be a shift towards personalised medicine within the healthcare industry. However, problems exist amongst the integration of data driven solutions that would need to be addressed in order for its success. Neff argues that there are four elements that need change in order for QS activities and data solutions to be connected and used within healthcare [21]. First, the ambiguity of data privacy needs to be addressed. Without clear regulations regarding the ownership and availability of personal health data, entrepreneurs and innovators may be discouraged to develop QS healthcare products by data security concerns. Second, tools need to be designed with both the current users, and the current healthcare professionals in mind, as opposed to tools that are built with solely the patient in mind. Third, new tools are needed to increase patient-doctor communication, to make it feasible and easier for patients to share data and information with their doctor on a regular basis. Finally, Neff calls for a policy within healthcare that is open to the technological adoption and advancement, and for greater focus on preventative healthcare [21]. A major change as such is likely to be met with opposition, especially those who believe and trust in traditional medical practices.

8.3 Quantified Insurance

As the QS movement expands, more people will be collecting and tracking their biometrics that, up until now, have remained private from insurance companies. Health insurers are now leveraging the fact that there is the opportunity to access hard facts concerning an individuals health and lifestyle choices. This can be viewed as either an advantage or a disadvantage. It can be considered an advantage for the QSer if they can prove to their insurer that they are leading a positive lifestyle that warrants them a superior insurance deal. This was the case for Andreas Schreiber who, after suffering a stroke, uses QS tools to better his health. This enables him to prove to his health insurers that he is doing everything in his power to recover [24]. However, not everyone would view this development in a positive light and, again, the issue of data ownership comes into play. Who owns the data and what right do they have to sell it on? Furthermore, what right do the insurance companies have to use this data for health risk profiling insights?

8.4 Legal and Privacy

As referred to throughout this paper, there are complex issues regarding the ownership and privacy of one's data. By using QS tools to track all aspects of one's life, QSers are placing their trust in these service providers and are somewhat blind to the exact ownership of their highly sensitive data. Symantec research showed that 20% of QS apps used unsecure methods to transmit user credentials. Although these apps and services may require username and password to access information stored on the cloud, these details are begin transmitted in clear text, without encryption. Therefore, this channel is at risk from cyber-attacks. Not only does this put the QSers data from this particular account at risk, but the user may also be putting his other accounts, for example email and banking, at risk, due to commonly reused login credentials and detailed online portfolios [10].

The Symantec research also found that 52% of apps did not have privacy policies available to its users, either at the time of signing up to the service, or throughout use. Privacy policies are necessary to show clearly who owns the data that is produced. The lack of such a policy is alarming, not only in determining present ownership, but also in determining who owns the data should company be sold. These apps and services currently own a large amount of valuable data that, in the future, could be sold to other organisations, for example, insurance companies, regardless of users preference [10].

8.5 Business and Personal Analytics

The data that is generated by QS tools is also very valuable to business. Currently organisations are utilising their own customer and market data to gain insights into their business and to make data-driven decisions. With the availability of QS data that could potentially be bought, organisations, researchers, and marketeers could gain even more insights into their customers and further communities. This would see the focus of business analytics shift from internal to external data. Furthermore, the increase in popularity of QS will significantly contribute to the world's volume of data, increasing the need for analysts.

We can see similarities between QS and the infancy of business analytics, with regard to their objectives and data obstacles. Currently, QS tools are providing the service of personal analyst to QSers. However, as discussed, there is a gap in the market for a tool to integrate all forms of a QSers data to provide an overall analysis. Perhaps the future may experience a shift in business analytics tools to include personal analytics, whereby tools similar to SAS[25], SPSS[26], RJMetrics[27], and Tableau[28], offer services that cater for the Quantifies Self.

8.6 Quantified Workplace

A further possibility for the future is the expansion of the Quantified Self to the Quantified Workplace. Kris Duggan, cofounder and CEO of BetterWorks, a company that applies the theories of QS the work place, believes it is the future of human resource management [19]. Speaking on Forbes: The Future of Work podcast, Duggan explains that millennials are the driving force behind the shift towards the quantified workplace. Today's workers much more geared towards measurements and quantifying aspects of their lives, due to the current nature of social media and technology, for example, Facebook likes and Twitter followers. The workplace needs to apply the gamification and tracking elements of QS in order to keep employees motivated and productive. Duggan explains that making workers' individual goals public, and assigning a score to reflect the their progress, maximises the chance of reaching their goal [19].

Unsurprisingly, there is a resistance towards employers keeping too close an eye on their employees. It is important that organisations maintain an open environment concerning everyone's data. To gain trust and to further motivate, top-level management should also make their data available, creating a flat structure within the organisation, in terms of openness of information. It is also important for organisations to be open about their motivations for quantifying. Organisations can gain insights and knowledge from the data that can then be shared with employees, such as how an individual worker compares to their colleagues or their

[25]SAS, https://www.sas.com/nl_nl/home.html/, retrieved September 29, 2015.
[26]SPSS, http://www-01.ibm.com/software/analytics/spss/, retrieved September 29, 2015
[27]RJMetrics, https://rjmetrics.com/, retrieved September 29, 2015
[28]Tableau, http://www.tableau.com/, retrieved September 29, 2015

historical self [19]. The expansion of QS to the quantified workplace is another example of the shift towards Quantified Us.

8.7 Quantified Charity

Unicef has launched the 'Wearables for Good Challenge', where they have invited professionals and enthusiasts in the fields of technology, design, and science to address social problems using wearable technology and data [32]. Organisations are looking for new ways to apply the theory and methodologies of QS to new ventures. Unicef believes that wearables can be used not only to track and monitor ones health, but also to acquire data on, for example, water purity and economical factors, in less developed countries. In the healthcare industry, there are opportunities to apply self-tracking to wearable technology to improve communication links between patient and doctor, and to closely monitor patients in low resource environments [32]. There are less technology and connected devices in less developed countries, and therefore less data available. Data can offer a new perspective on social issues, and can therefore highlight where changes need to be made and how best to make them.

9 QS Data Analytics Example

To further explore the analytical side of the QS process, this section uses an example dataset[29] to apply the methods outlined in Section 6. The dataset contains quantified self data from 27 anonymous individuals, and 19 variables. Immediately, the problem of data integration becomes clear. The data from each of the variables are collected at different time intervals, with some metrics being recorded every 2 hours on the hour, and others being recorded to the second, for example, logging the time a phone call is accepted. This makes it very difficult to compare variables along a time scale without pre-processing the data. In this section the data is analysed firstly from an individual perspective, by investigating only one QSer's data.[30]

9.1 Quantified Self Perspective

A subset of the data attributed to one QSer (AS14.31) is taken. A difficulty when analysing QS data is that the frequency of the data points vary significantly. Table 1 lists the variables in this dataset, alongside the corresponding number of entries. The variables monitored contain a mix of passively observed, such as metrics attributed to the use of the QSer's mobile phone, and those tracked with user input, such as mood, arousal and valence. From this data available, we can investigate the relationship between phone usage and general mood and emotions.

Table 1: Table of QS data and the frequency of data entries

Variable	**Freq**
activity	793
appCat.builtin	4038
appCat.communication	1891
appCat.entertainment	472
appCat.finance	50
appCat.game	0
appCat.office	105
appCat.other	296
appCat.social	192
appCat.travel	30
appCat.unknown	1
appCat.utilities	171
appCat.weather	0
call	173
circumplex.arousal	198
circumplex.valence	198
mood	198
screen	3005
sms	78

9.1.1 Preprocessing the data

The most important process within data analytics is preprocessing the data so that is in a format ready for analytical use. Some preprocessing steps are outlines as follows.

[29] Dataset provided by Dr. Mark Hoogendoorn, VU, Amsterdam
[30] All analysis is carried using Excel and R

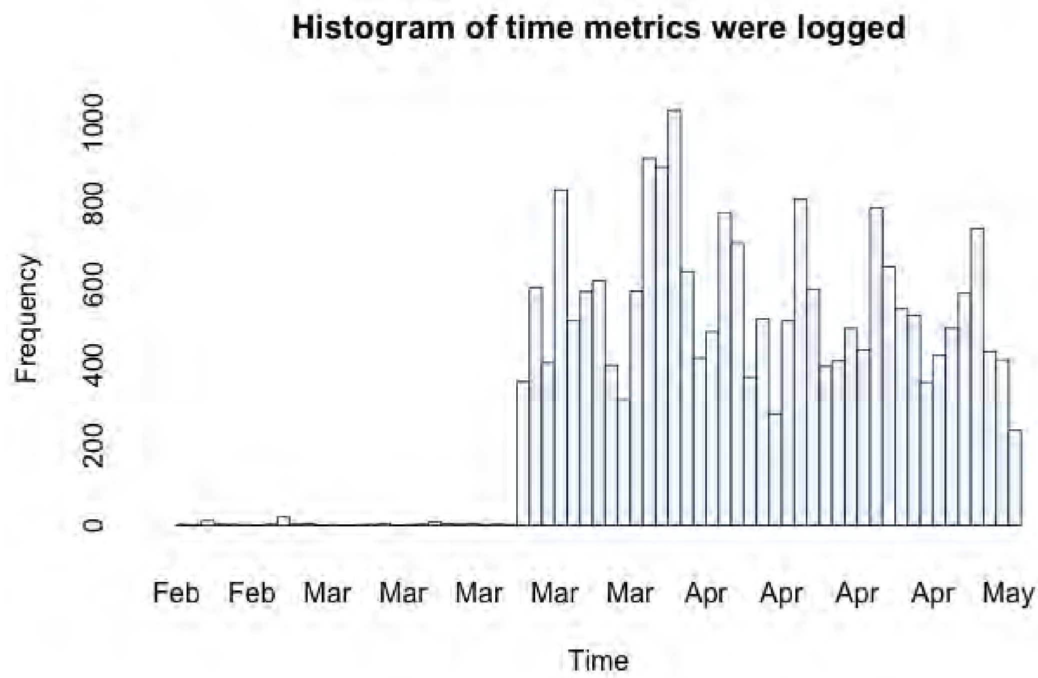

Figure 10: Frequency of data collection over time

Timeframes All metrics were recorded on different time frames. The metrics requiring user input (mood, arousal and valence) were recorded sporadically on the hour, where as those recorded passively were recorded in real time. Figure 10 shows frequency of the timestamps, and as we can see the majority of the data was recorded between the middle of March and the end of May, with very few data entries before then. In order to make the data more readable and ready for analysis, the timestamps were rounded up to the nearest hour, using the logic that one's interaction with their phone only affects their mood in the future, and not in the past. To achieve this the data was reordered to show for each unique timestamp, the average value within that hour and the count of entries for that hour. This reduced the data from 11,055 unique timestamps to 983. However, re-ordering the data this way also increased the number of attributes from 5 to 28. Alternatively, one could extended the data by taking the last value for mood, valence, and arousal and imputed these values for the next timestamps of the dataset, until a new value was recorded. For the purpose of exploratory data analysis, both the raw dataset and the adjusted dataset were used.

Data cleaning The process of data cleaning involves removing outliers from the dataset and handling missing values. To remove outliers, the spread of each variable was investigated individually. Outliers were removed when they deviated from the rest of the data substantially. However most variables remained untouched, as it is difficult to get a multidimensional view on the data to surely identify outliers.

Figure 11 shows the percentage of missing data for each variable. There is a high percentage

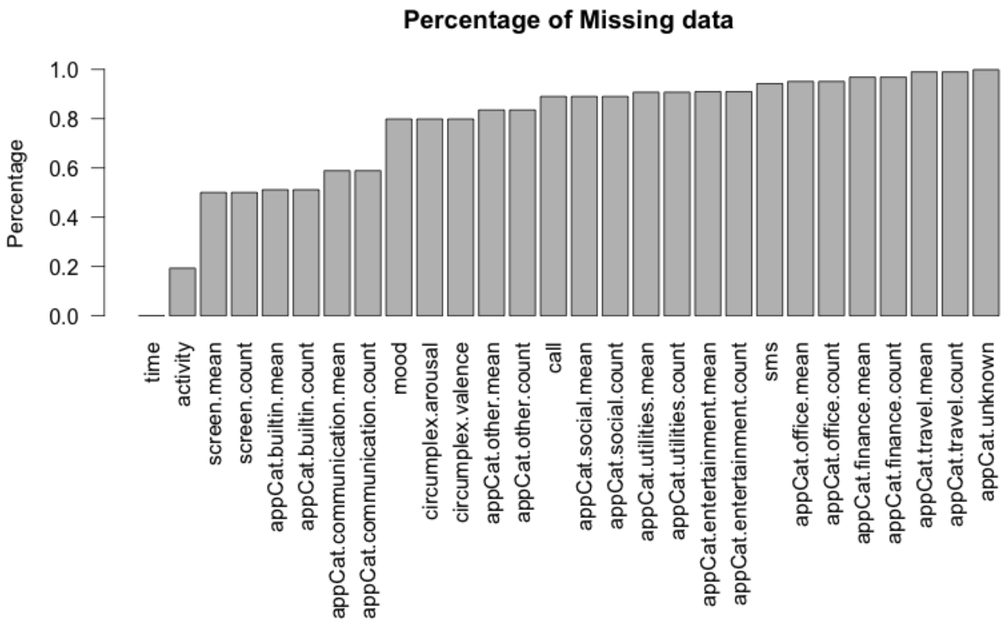

Figure 11: Proportion of missing data in each attribute

of missing data for the variables that measure app usage, the reason for which is unclear. The user could have forgotten to switch a particular app on, or perhaps they stopped recording a metric for a certain amount of time. One could assume that these variables only take measurements when the app is in use, and do not record usage as zero for when the app is not in use. However, imputing a value of zero for all missing data points has the potential to skew the data. Another option is to impute the variables with the median (or other statistic), but this imply use of the app at each timeframe. A more precise option would be to predict the values using the values that are available. However, this dataset may struggle predicting precise values for the missing values given the percentage of missing values in each variable.

Data cleaning procedures can be quite laborious, and QSers will be faced with these challenges when integrating and analysing their own data from various data sources. Therefore, it is necessary for a service and solution to be brought to the market to ensure ease of use for various QS tools.

Feature engineering It may be of interest to include extra features that have been inferred from the existing dataset. In the reshaping of the dataset new features were already included: the mean values and counts for the app related data. Intuitively, these values could be of interest as a QSer's mood may decrease dramatically from receiving numerous phone calls in a short period of time. Alternatively, a feature could be introduced to record the time until the last call was received. Further features introduced were the day of the week and the time of day.

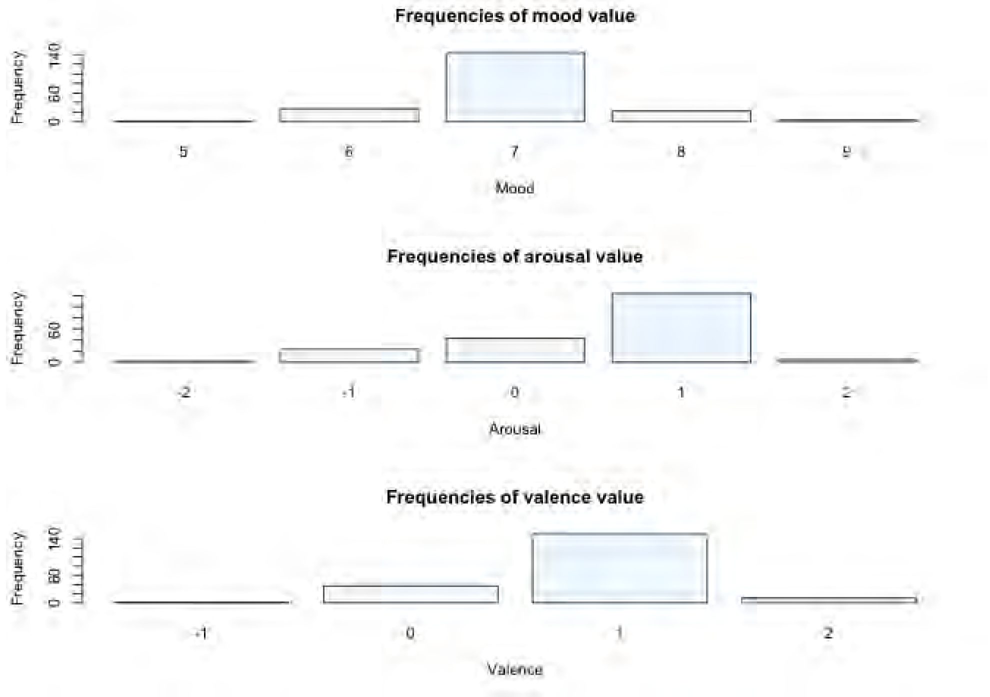

Figure 12: Frequencies of mood, arousal, and valence variables

9.1.2 Exploratory analysis

Of particular interest in this case, is the users mood, arousal and valence. By measuring these values against the use of the applications, one can identify what contributes to a happier mood or a lower mood. First the raw data was analysed. Figure, 12 shows the frequencies of each emotion, the majority of which are positive. Figure 13 shows each emotion over time, from the beginning of April until the middle of May. For all three, the QSer experiences quite volatile emotions until around April 20, where the graphs begin to level out. Data is still being collected at frequent intervals during these times, and the QSer's emotions seem to level out.

For the purpose of this report, I investigated the variable `mood`. Figures 14 and 15 show the plots of mood against a selection of variables. In Figure 14 a weak positive correlation can be seen between `mood` and `activity`, and between `mood` and `screen.count`. Likewise, in Figure 15 a weak positive correlation can be seen between `mood` and `communiction.count`

I ran a random forest with 100 trees, and two variables randomly sampled at each split, with all variables as inputs, to gain a better understanding of the variables that are associated with `mood`. For this, I replaced all missing values with the value 0, under the assumption of no usage for that time. Figure 16 shows the importance plot, with importance based on permutation on the left, and based on Gini on the right (the least important variables were removed from the plot for clarity). As we can see, variables that are considered important in predicting a QSer's mood (ignoring date) are `activity`, `screen.mean`, `screen.count`, and `builtin.count`. As we can see, variables that are considered important in predicting a QSer's mood (ignoring date) are `activity`, `screen.mean`, `screen.count`, and `builtin.count`.

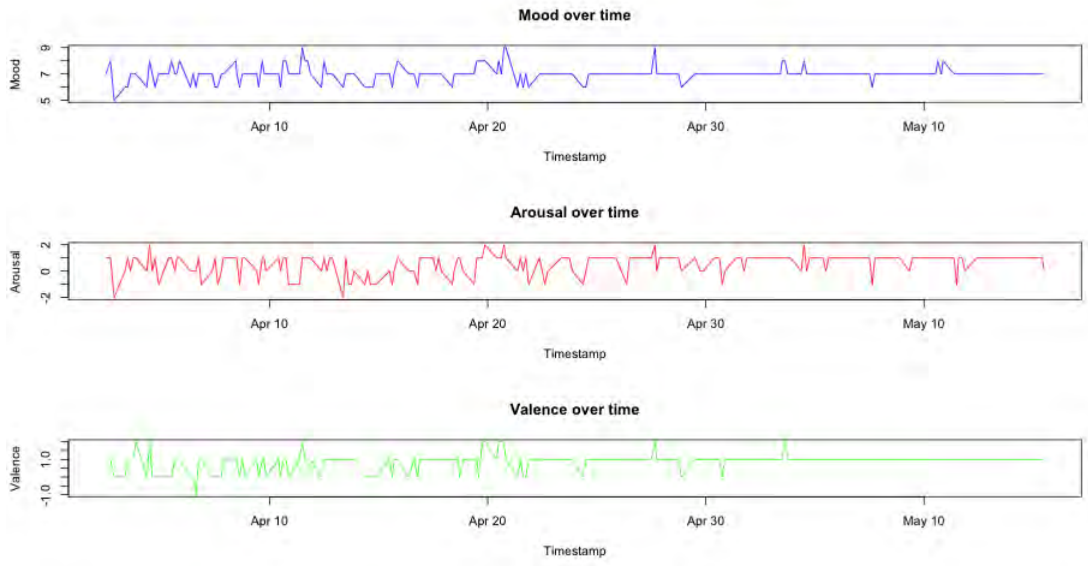

Figure 13: Mood, arousal, and valence over time

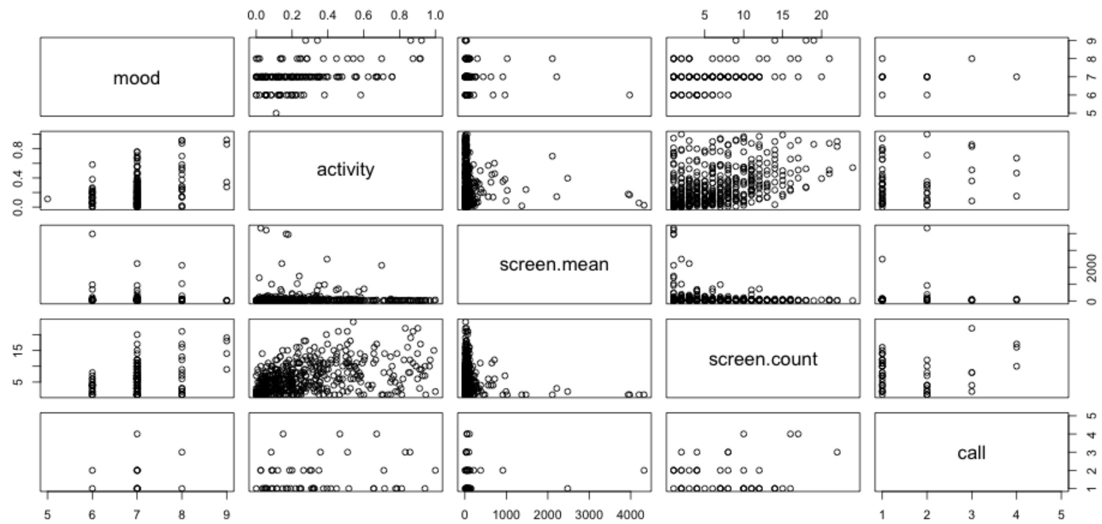

Figure 14: Plot of mood value against `activity`, `screen.mean`, `screen.count` and `call`

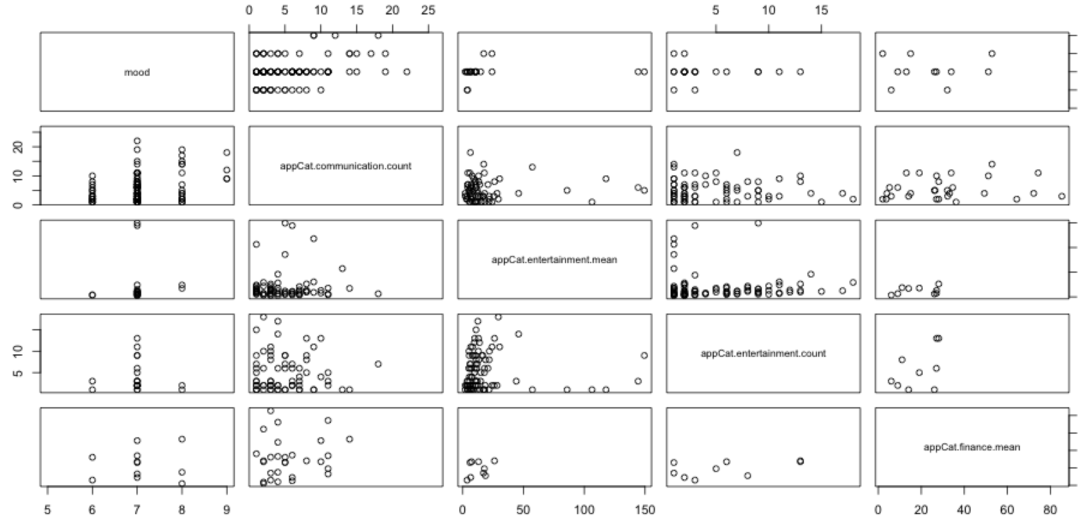

Figure 15: Plot of mood value against `communication.count`, `communication.mean`, `entertainment.count` and `finance.mean`

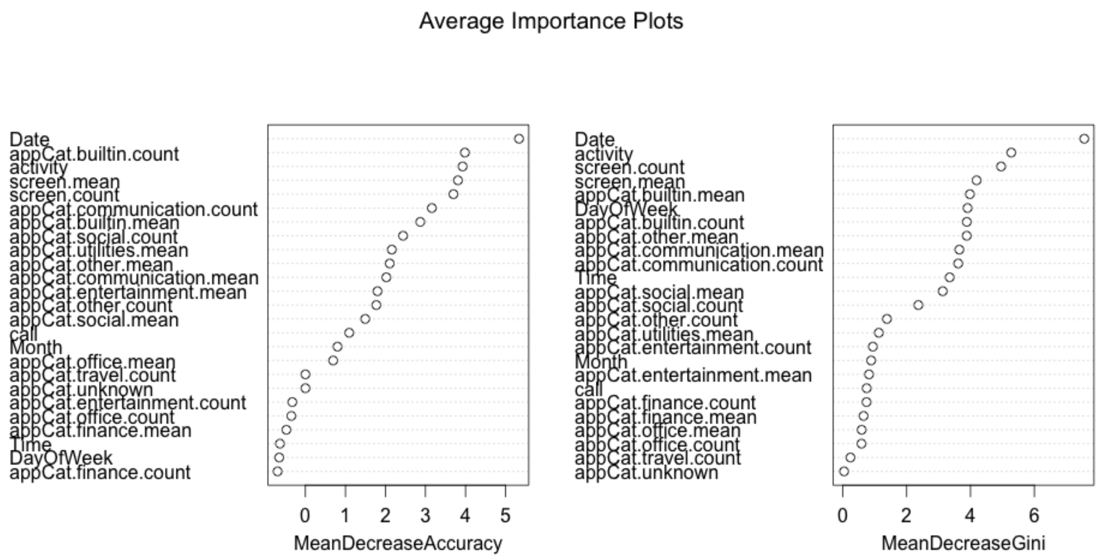

Figure 16: Importance plot in predicting mood

9.2 The Quantified Us/Them Perspective

As previously mentioned, the dataset originally contained information from 27 individuals and 19 variables. This amassed to a raw dataset of 376,912 entries. This data could provide a great insight into the collective behaviours of QSers, and allow for the comparison between individuals. However, when preprocessing data of this size I quickly encountered some problems, in particular memory problems when reshaping the dataset. This highlights data storage and memory obstacles that QSers face, as data of this size could easily be reached on an individual basis within a year, perhaps even 6 months.

This shows the wealth of information to which the builders of QS tools have access. They are in a position to model and predict the mood and emotional state of users by integrating and aggregating all data available. This could then be incorporated into marketing campaigns, for example, to manipulate users purchasing habits. However, it could also be used to aid individuals suffering from depression, by predicting when their mood will decrease and allow carers to help mitigate against this.

10 Conclusion

The aim of this research paper was to explore the methodologies and theories behind the Quantified Self, and to investigate the type of analytics and data management that it requires.

The Quantified Self is a movement that connects humans to the Internet of Things. There are many tools available that measure and quantify various aspects of oneself, allowing users to analyse their own biometric data. This highlights the similarities between business analytics and the Quantified Self, as both are centred around exploratory analyses and making decisions based on the data. As shown throughout this paper, the Quantified Self is now becoming mainstream and will play a major role in people's lifestyles and how they store their information. An increase in the availability of wearable technologies will see the QS movement contribute largely to Big Data science, and with this comes the obstacles of data management: storage, integration, aggregation and analytics. Business analytics will also shift to include data that is created through personal analytics, and businesses will be able to see even more insights into their customers. Data mining itself is a relatively new field of research, but there is little known about the mining of biometric data. The healthcare industry is changing to utilise QS data, and with it will come new theories and methodologies about the analysis of biometric data.

There are many advantages to the Quantified Self: self knowledge, self improvement, possible cheaper insurance to name a few. However, there are also many disadvantages: people may become too focussed on their data to the point of obsession, and there is ambiguity amongst the ownership of QS data and what this could lead to. Arguably the biggest advantage is the positive effect it can have on the health industry, both in preventative medicine and in real-time information updates for the patient-doctor relationship. Although the Quantified Self is focussed at an individual level, there is already evidence of it expanding to a broader level to include the Quantified Us and the Quantified Them. Eventually the term 'quantified' will no longer be needed to distinguish between the norm, as data and analytics will become the driving force in everyday life.

When working through the QS data in Section 9, I encountered many of the difficulties within data integration, aggregation and analysis discussed in this paper. Considering this is a small dataset, consisting of data collected over a period of four months, it is clear to see that QS data mirrors Big Data in the problems that it faces, but on a smaller scale. There is a gap in the market for the relevant tools to streamline the QS process, so that users can begin to analyse data across sources and fully utilise their information.

Most research into the Quantified Self has focussed on the social impact continuous self-tracking at an individual level, and the hurdles that it face. There is scope for future research on how this will impact society and the way business conduct their business analytics, as touched on in this paper. There is also scope for researchers to develop new data mining techniques specific for the integration of biometric and QS data, as this appears to be the biggest challenge when analysing QS data.

11 References

[1] Internet world statistics. http://www.internetworldstats.com/stats.htm. [Online; accessed June 15].

[2] The quantified self. http://quantifiedself.com/. [Online; accessed June 21].

[3] Quantified self meetups. http://quantified-self.meetup.com/. [Online; accessed June 18].

[4] The quantified self: Counting every moment. http://www.economist.com/node/21548493, March 2012. [Online; accessed June 18].

[5] Rocket fuel: Quantified self digital tools. http://quantifiedself.com/docs/RocketFuel_Quantified_Self_Research.pdf, 2014. [Online; accessed June 21].

[6] Wearable devices: Introduction to wearable technology. http://www.wearabledevices.com/what-is-a-wearable-device/, March 2014. [Online; accessed June 15].

[7] The zettabyte era - trends and analysis. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html, May 2015. [Online; accessed September 20].

[8] M. Armbrust et al. A view of cloud computing. *Coomun. ACM*, 53(4):50–58, 2010.

[9] A. Awasthis and S. Chaudhary. *Engineering Mathematics*. Darbose, Inc., India, 2009.

[10] M. Barcena, C. Wueest, and H. Lau. How safe is your quantified self? http://www.symantec.com/connect/blogs/how-safe-your-quantified-self-tracking-monitoring-and-wearable-tech?linkId=9093734, August 2014. [Online; accessed June 29].

[11] J. Behrens. Principles and procedures of exploratory data analysis. *Psychological Methods*, 2(2):131–160, 1997.

[12] K. Crawford, J. Lingel, and T. Karppi. Out metrics, ourselves: A hundred years of self-tracking from the weight scale to the wrist wearable device. *European Journal of Cultural Studies*, 18, 2015.

[13] S. Fow and M. Duggan. The diagnosis difference. Technical report, Pew Research Center, November 2013.

[14] H. Gimpel, M. Nißen, and R. Görlitz. Quantifying the quantified self: A study on the motivations of patients to track their own health. In *Thirty Fourth International Conference on Information Systems*, Milan, 2013.

[15] T. Hay. The rise of the 'quantified self' in health care. http://blogs.wsj.com/venturecapital/2013/08/13/the-rise-of-the-quantified-self-in-health-care/, Aug 2013. [Online; accessed June 26].

[16] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1 edition, 1986.

[17] D. Laney. 3d data management: Controlling data volume, velocity and variety. Technical report, META Group, 2001.

[18] I. Li, A. Dey, and J. Forlizzi. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010.

[19] J. Morgan. From the 'qualified self' to the 'qualified workplace'. http://www.forbes.com/sites/jacobmorgan/2015/01/22/from-the-quantified-self-to-the-quantified-workplace/, January 22 2015. [Online; accessed June 25].

[20] D. Nafus and R. Barooah. Qseu14 breakout: Mapping data access. http://quantifiedself.com/2014/06/qseu14-breakout-mapping-data-access/, June 2014. [Online; accessed June 20].

[21] G. Neff. Why big data won't cure us. *Big Data*, 1(3):117–123, 2013.

[22] J. Norris. Self-tracking may become a key element of personalized medicine. http://www.ucsf.edu/news/2012/10/12913/self-tracking-may-become-key-element-personalized-medicine, October 2012. [Online; accessed June 23].

[23] F. Provost and T. Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1:51–59, 2013.

[24] A. Schreiber. Quantified self: Analyzing the big data of our daily life. http://www.pyvideo.org/video/3077/quantified-self-analyzing-the-big-data-of-our-da, July 2014. [Online; accessed June 20].

[25] E. Singer. The measured life. *MIT Technology Review*, June 2011.

[26] SINTEF. Big data, for better or worse: 90% of world's data generated over last two years. www.sciencedaily.com/releases/2013/05/130522085217.htm, 2013. [Online; accessed August 22].

[27] M. Stempeck. Questioning the quantified self as it marches towards mainstream. http://www.mattstempeck.com/2014/02/28/questioning-quantified-self-as-it-marches-towards-the-mainstream/, February 2014. [Online; accessed June 23].

[28] M. Swan. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks*, 1(3):217, 2012.

[29] M. Swan. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 2, 2013.

[30] K. Tillman. How many internet connections are in the world? right. now. http://blogs.cisco.com/news/cisco-connections-counter, June 2013. [Online; accessed June 15].

[31] C. Tudor-Locke. Taking steps toward increase physcial activity: Using pedometers to measure and motivate. *President's Council on Physical Fitness and Sports Research Digest*, June 2002.

[32] Unicef. The wearables for good challenge. http://wearablesforgood.com/, 2015. [Online; accessed June 26].

[33] J. Walker. Can a smartphone tell if you're depressed? http://www.wsj.com/articles/can-a-smartphone-tell-if-youre-depressed-1420499238, January 2015. [Online; accessed June 23].

[34] G. Wolf. The data-driven life. http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html, April 2010. [Online; accessed June 15].