Blockchain Technology for Data Management of Research Data:

A Systematic Review and Proposed Research Design

Author:	Ronan Duchemin
Student Number:	2578924
Contact information:	r.l.m.duchemin@student.vu.nl
Supervisor:	Prof. Dr. S. Friesike
Date of publication:	30 - 06 - 2018
Programme:	Bachelor of Business Administration
Specialization:	Digital Business Innovation

Abstract

An ascending amount of scientific research is not replicable. This study aims to identify how blockchain technology can impact data management of research data. This is done by examining how blockchain technology for data management of research data can result in a higher rate of replicable studies. Three causes of the replication crisis have been identified: a lack of data quality, indefinite methodology, and the publication bias. This study evaluates whether or not blockchain technology can help reduce this. To do this, six characteristics of the new emerging technology have been identified: decentralization, timing data, data handling, security and reliability, programmability, and consensus mechanisms. Furthermore, the potential challenges and benefits of implementing blockchain technology for data management in the scientific community have been systematically mapped. The challenges that have been identified are: speed of transaction handling, scalability, security, and limitations of private keys. A negative relationship between scalability and security has been identified during the mapping of the challenges of blockchain technology for data management in scientific research. The identified benefits are: decentralized management, cost-efficiency, immutability, and improved security and privacy. Blockchain for data management of scientific research can result in a higher rate of replicable studies because of three reasons. Firstly, the issue of indefinite methodology can be addressed through the programmability characteristic of the blockchain. Secondly, this same programmability characteristic can be used to address the publication bias. Finally, the data quality can be improved through the security, immutability, and timestamping characteristics, and by the appropriate use of private keys.

Keywords: Data management, Blockchain, Replication crisis, Data quality, Science, and Data collection

Contents

Chapter 1: Introduction and definitions

1.1: Introduction

With the growing interests for big data the volume of data for most scientific disciplines is increasing to petabytes, and for some disciplines the amount of data even exceeds an exabyte (Guo, 2015). There is a need for excellent data management that is becoming more and more important to the scientific community.

According to Pusztai, Hatzis and Andre (2013) finding positive results is a career must for most scholars. This can motivate scientists to perform some kind of scientific misconduct. Resulting in misallocated scientific resources by following leads suggested by fraudulent research (Stroebe, Postmes & Spears, 2012). A low quality of research data, indefinite or incomplete methodology, and the publication bias are all identified as causes of the replication crisis. A problem that is particularly damaging for the image of the medical (Ioannidis et al. 2005), and psychological (Open Science Collaboration, 2015) sciences.

Almost 10 years ago, Nakamoto (2008) introduced the concept of a decentralized distributed ledger, called blockchain, to administer the distribution of his newly developed cryptocurrency called Bitcoin. The past 10 years a lot of research has been published on this new technology. However, the available literature on blockchain for science is limited. Blockchain is a interdisciplinary concept with papers published in almost all of the scientific disciplines (Yli-Huumo, Choi, Park & Smolander, 2016).

The aim of this paper is to give an overview on how blockchain technology for data management can impact scientific research. This was decided upon after identifying the following problem:

'An ascending amount of scientific research is not replicable.'

This study could be of importance for researchers, publishers, and everybody that is contributing to the scientific community. Blockchain is identified as a possible solution to this problem. To further evaluate this solution the following research question will guide the direction of this study:

'How can blockchain for data management of research data result in a higher rate of replicable studies?'

In order to provide the best answer to this research question, some sub-questions have been formed to provide more insights into the research question. The sub-questions for this study are:

1) *'What causes the replication crisis?'*
2) *'What are the characteristics of blockchain technology?'*
3) *'What are potential challenges and benefits for the scientific community of using blockchain technology for data management?'*

The first sub-question is studied by means of a traditional review (Jesson, Matheson & Lacey 2013; Saunders, Lewis & Thornhill, 2016). The traditional review is chosen here to allow some flexibility in the selection of articles. The limitation in time and the broad scope of this subtopic are the main reasons for this decision. The literature on blockchain is studied by means of a systematic literature review. This allows for valid and reliable results of this study by using predetermined criteria for including and excluding articles (Jesson et al. 2013).

This study has informative value for researchers by providing new insights in a field where little research has been done yet. Firstly, the causes of the replication crisis will be identified. Furthermore, the characteristics of blockchain technology will be evaluated. Finally, a framework is introduced identifying the possible benefits and challenges that the scientific community will experience by adapting to blockchain technology.

This paper will start by providing some definitions to assist the reader through the vocabulary that is used in this paper. Afterwards, the methodology of both the traditional and the systematic literature review will be presented in chapter 2. Chapter 3 consists of a traditional review of the literature available on the replication crisis. Identifying various causes and effects of this problem and providing some information about the scale of it. Chapter 4 contains a systematic review of the literature on blockchain technology. Identifying six different characteristics of blockchain and providing some insights on the potential challenges and benefits that the scientific community could experience from using blockchain for data management. Chapter 5 introduces a proposal for future research, highlighting the need for additional research on the rationale of the included stakeholders to adopt to this new technology. Finally, based on the results from chapter 3 and chapter 4 a conclusion is presented in chapter 6. Followed by a discussion of the limitations and managerial implications.

1.2: Definitions

Although this study tries to explain the terminology used in the article, the complexity and novelty of this subject might call for some additional background knowledge. This paragraph serves as an overview of the definitions. This way, the reader can refer to this when necessary.

51% attack refers to a method that external attackers can use to change the consensus of the blockchain. If the attackers manage to provide more than half of the computer power the network is running on they will be able to make changes in the blockchain.

API is short for application programming interface. It consists of the rules and protocols within which programmers will have to develop their applications.

Blocks are new pieces of information that are added to the blockchain. For the Bitcoin network a new one MB block is generated every 10 minutes (Poon & Dryja, 2015). The blocks contain a header and a list of transactions. The header consists of some metadata about when the block was created and includes a hash from the previous block so any changes in earlier blocks will be noticed.

Data integrity is defined by Boritz (2015) as the maintenance and assurance of the accuracy and consistency of data over its entire life-cycle.

Data management is the administrative process of acquiring, storing, validating, processing, and protecting the data that is required to ensure the reliability, accessibility, and timeliness of the data for the data users (Galetto, 2018).

Hashes can be described as digital signatures of a piece of data (Hou, Wang & Liu, 2017). A hash is created through a cryptographic algorithm. For Bitcoin this is done with the SHA-256 algorithm. If anything about the data changes, even if only one letter, the hash value will change. This makes all changes to the data notable, and thus provides the immutable property of the blockchain.

Merkle tree refers to the data structure of the blockchain. After the transactions are confirmed and added to the blockchain, changes to the data can still be identified because it will change the 'hash' value. The Merkle trees hierarchy makes that a change in the bottom of the tree will still result in a change of the so called 'Merkle hash'. An example of a Merkle tree is displayed in figure 1 (appendix 2, Lee & Yang, 2018, p.8)

Mining is the process of validating transactions on a blockchain. After a transaction is successfully validated the miner gets a reward in the form of some tokens (e.g. Bitcoin). The word comes from the

comparison of Bitcoin with gold. Consensus mechanisms like Proof-of-Work and Proof-of-Stake are both a type of mining.

Nodes are the participants on a blockchain. They provide servers that secure a copy of the database (Hou et al. 2017; Wang et al. 2018). These are the 'miners' that confirm the transactions.

Private keys are used to decode an encrypted message intended for a specific recipient. Within the blockchain they can also be used to confirm transactions and access your wallet.

Proof-of-Work is one of the consensus mechanisms used to secure the blockchain. It is used for Bitcoin. It requires solving a complex algorithm that is hard to solve but easy to verify. So that once one of the miners succeeds in solving the puzzle the other miners can easily verify its correctness.

Proof-of-Stake is an alternative consensus mechanism that requires less energy. This is because not everyone is competing to solve the puzzle simultaneously, but one node is designated to do this by random chance, moderated by the wealth and the age of the nodes (Yeow, Gani, Ahmad, Rodrigues & Ko, 2017).

SHA-256 is a cryptographic hash algorithm designed by the NSA. The algorithm is used by bitcoin during the proof-of-work process, but also with the creation of a new wallet. This is to assure the security and privacy of the owner of the new wallet. Unlike encrypted information, a hashing function like SHA-256 cannot be decrypted. It is a one-way function. Given the data you can calculate the hash function. However, given the hash value it is still mathematically impossible to calculate the original data. This is due to the fact that the hash function is limited to 256 bits while that data file could be of any size (Yeow et al. 2017).

Smart contracts use computer language to replace legal language to record terms (Hou et al. 2017). They can be used to automate processes and confirm the execution of tasks.

Chapter 2: Methods

In this chapter the methodology used for this research will be described. First, the literature about the replication problem has been reviewed during the scope and map phase. The searches, and results from these searches have been carefully recorded. During this phase the knowledge of the subject was not yet sufficient to identify the appropriate inclusion and exclusion criteria. This is justified because this study is not trying to provide an evidence base for the existence of the replication crisis, but is moreover providing a solution to the problem. A method more similar to the traditional review has been applied to the literature on data manipulation and the replication crisis. The review of this literature has resulted in an understanding of these subjects and will be used to systematically evaluate the available literature on blockchain technology for data processing. This will be done to create an overview of the different characteristics that a blockchain possesses and to identify possible benefits and challenges from using the technology in the scientific community.

2.1: Research design

For this study a combination of a traditional review (Jesson et al. 2013; Saunders et al, 2016) and a systematic literature review methodology (Jesson et al. 2013; Webster, 2002; Okoli and Schabram, 2010; Rowley & Slack, 2004; Saunders et al. 2016) was used. The traditional review allows more room to the researcher for exploration and creativity because there is less of a predefined path that the researcher will need to follow. The systematic review originates from medical sciences where it was introduced as an attempt to improve the research process by synthesizing research in a systematic, transparent, and reproducible manner. After which it has found its way to other scientific disciplines (Tranfield, Denyer & Smart, 2003). The aim of using this methodology is to identify key scientific contributions and create an evidence base that exceeds those of a single paper (Dada, 2018). By using predetermined criteria for including and excluding studies, reproducibility of this study should be possible (Jesson et al. 2013). This study is based on secondary data.

2.2: Review protocol

The review protocol serves as a plan to execute in order to perform this literature review. By explicitly describing all the necessary steps the objectivity of this study will be ensured. This study follows eight steps as recommended by Jesson et al. (2013).

Note: Adapted from "Doing Your Literature Review: Traditional and systematic techniques.", by Jesson, J., Matheson, L., & Lacey, F. M., (2013). Royal New Zealand Foundation of the Blind, p. 104

2.2.1: Score and map

According to Jesson et al. (2013) the scoring phase can be compared to a traditional review, which has to be conducted before the systematic review. The traditional review is a less objective way of reviewing literature than the systematic review where the researcher has no defined path, which allows for creativity and exploration (Jesson et al. 2013). As stated by Tranfield et al. (2003) this is even more important for management research: *"Within management it will be necessary to conduct scoping studies to assess the relevance and size of the literature and to delimit the subject area or topic. Such studies need to consider cross-disciplinary perspectives and alternative ways in which a research topic has previously been tackled"* (p. 214).

This scoping process was the start of this study. Mainly academic literature was reviewed during this phase, but also some professional journals, internet sources, whitepapers, conferences, and (informational) books. The articles used in this phase have been subtracted through Libsearch and Google Scholar. Keywords that were used can be found in the table below (table 1). If a search provided a lot of results, only the first five pages of results would be reviewed during this phase of the study. This choice was made because of the limitation in time and was based on the idea that the search engines would provide the most relevant and most cited ideas in these first pages.

Table 1: Keywords Scoping search

Key words	Search terms	Filter	Database	# of publications found
Reproducibility AND problem	Title	Peer reviewed + English + Last 10 years	Libsearch	29
Falsify AND research	Title	Peer reviewed + English + Last 10 years	Libsearch	27
Falsified OR manipulated OR modified OR fabricated AND data AND science	Title	Peer reviewed + English + Last 10 years	Libsearch	55
Replication crisis	Any	Peer reviewed + English + Last 10 years	Libsearch	2.645
Replication crisis	Any	2009-2018	Google scholar	37.600
Scientific misconduct	Any	2009-2018	Google scholar	25.600
Blockchain AND Science	Any	2009-2018	Google scholar	8.230

During the generation of ideas, the working document of Bartling et al. (2017) about *'Blockchain for open science and knowledge creation'* has helped to form the basic ideas. This article provides a good overview of the developments and opportunities for blockchain within science. The other 17 articles were a result of the searches presented above. The articles that have been analyzed from the keyword searches above can be found in table 2 (see appendix 1).

Literature that was obligatory for courses at the VU will also be considered for this thesis. This literature was not excluded based on the year of publication. An overview of the literature used from previous courses can be found in table 3 (see appendix 1).

During the scope and map phase the searches aimed at explaining the replication crisis as a consequence of low data quality. The literature provided insights into what fields have been experiencing this replication crisis and identified some causes and effects from the replication crisis. This understanding of the replication crisis will help to focus the systematic search of the available literature on blockchain for data management.

2.2.2: Inclusion and exclusion criteria

The articles were scanned for in- and exclusion by one researcher as part of a bachelor thesis. The most important inclusion criterion was the relevance of the article for answering the research question or any of the sub-questions. The titles were first examined to evaluate this. For the articles of which the title did not provide enough information to exclude it, the abstract was evaluated. No limitations have been applied on the research design of the articles used for analysis. However, when using literature reviews as unit of analysis, backward citation searching was used to identify the original source of the message used (Webster & Watson, 2002). The article that presented the used information first was also evaluated on the inclusion and exclusion criteria presented in this sub-paragraph.

In order to build a strong evidence base for the research question some exclusion criteria had to be in place. The search boundaries were set to electronic databases. The articles used in this study have been subtracted through Web of Science. Articles reviewed in this study were gathered from the ten-year period 2009 – 2018. The data has been gathered from a point in time not long after Nakamoto (2008) first introduced the idea of blockchain as a peer-to-peer technology with his paper presenting the Bitcoin. Only articles written in the English language have been reviewed.

Furthermore, there were three phases during the screening of the articles where articles have been excluded. First, the title of the article was evaluated. If the article could not be excluded based on the title, the abstract was read. If there was still no reason to exclude the article, the article was included in the second phase. During the second phase the articles full text was read to make sure the article contained relevant information to help answer any of the research questions. Articles that were not included in the universities subscription and required a purchase have also been excluded during this phase. Finally, the article influence (AI) score of the journals has been evaluated in the third phase. This is only done for the academic literature to further assure the quality of the scientific articles. No hard

cap will be set for the AI score. However, journals scoring under 70 all have been individually evaluated on their quality and contributions in sub-paragraph 2.2.4.

2.2.3: Search and screen

Jesson et al. (2013) emphasize that you have to document all the decisions you make. This way the process is transparent and can therefore be replicated by other researchers. This sub-paragraph will describe these decisions.

Some of the articles analyzed during the scope and map phase also used some relevant theories and ideas about blockchain. For those articles, the relevant citations were reviewed by going backward (Webster, 2002). In this phase this has been done for three articles. Namely, *'How blockchain-timestamped protocols could improve the trustworthiness of medical science.'*, *'Scientific Misconduct and the Myth of Self-Correction in Science.'*, and '*Reproducibility in Science Improving the Standard for Basic and Preclinical Research.',* which resulted in a total of 12 additional articles. An overview of the additional articles can be found in table 4, table 5, and table 6 (see appendix 1).

An overview of the conducted searches providing information about their dates, amount of results, and the number of used articles is presented in the table below (table 7). The in- and exclusion criteria as mentioned above were used to select relevant articles.

Table 7: documentation of searches

Search #	Search keywords	Date of search	# of results	After first phase	After second phase	# of used articles
1	Blockchain AND data	04/06/2018	88	12	8	8
2	Blockchain AND (Attributes OR characteristics OR features)	13/06/2018	30	7	5	

The first search on the Web of Science provided 88 results. 12 of these results seemed relevant after evaluating the titles and abstracts. Eight articles have been selected after reviewing the full articles and van be found in table 8 (see appendix 1).

Two of the articles reviewed during this search contained relevant theories from other authors that have been included in this study through backward citation reviewing. The articles used for citation reviewing are '*Lightweight and Manageable Digital Evidence Preservation System on Bitcoin.'* and *'Blockchain technology for improving clinical research quality.'*. The citation review resulted in four additional articles that were reviewed. An overview of the additional articles can be found in table 9 and table 10 (see appendix 1).

The second search resulted in a total of 30 articles. Seven of these articles have been evaluated after the first phase in which the titles and abstracts were studied. One of these articles was not available within the universities subscription and one of the articles was excluded after reading the full text. This resulted in a total of five new articles for the literature review. An overview of the selected articles can be found in table 11 (see appendix 1).

2.2.4: Quality appraisal

All the literature that has been obtained was reviewed on their quality before the literature review is conducted. Journals have been evaluated on their Article Influencer (AI) score. According to Eigenfactor the AI score provides the average influence score of a journal over a five-year period. The articles are scored as a ratio with one being the highest and zero being the lowest possible score. Leading articles typically score a 0,7 or higher. An overview of the used academic journals and their Article Influencer score can be found in table 12 (see appendix 1). Journals marked red have been excluded from this review. This process is described below.

All the articles that did not score above 70 on the Article Influencer percentile have been evaluated individually on their reliability, validity, and qualitative contributions. Because a lot of the articles were published very recently a non-existing AI score could be caused by the algorithm that calculates this score over a five-year period. Next to the journal, the individual article has been re-evaluated to assure the quality.

Exclusion

The article from *'F1000Research'* was taken offline with the following statement: *"The authors have taken this decision after considering the methodological concerns raised by a peer reviewer during the post-publication open peer review process."* when revisiting their website on June 15th 2018. The articles from *'Industrial and Organizational Psychology'*, *'Science and Engineering Ethics'*, and *'The Forestry Chronicle'* have all been excluded from the literature review based on their low AI score.

Inclusion

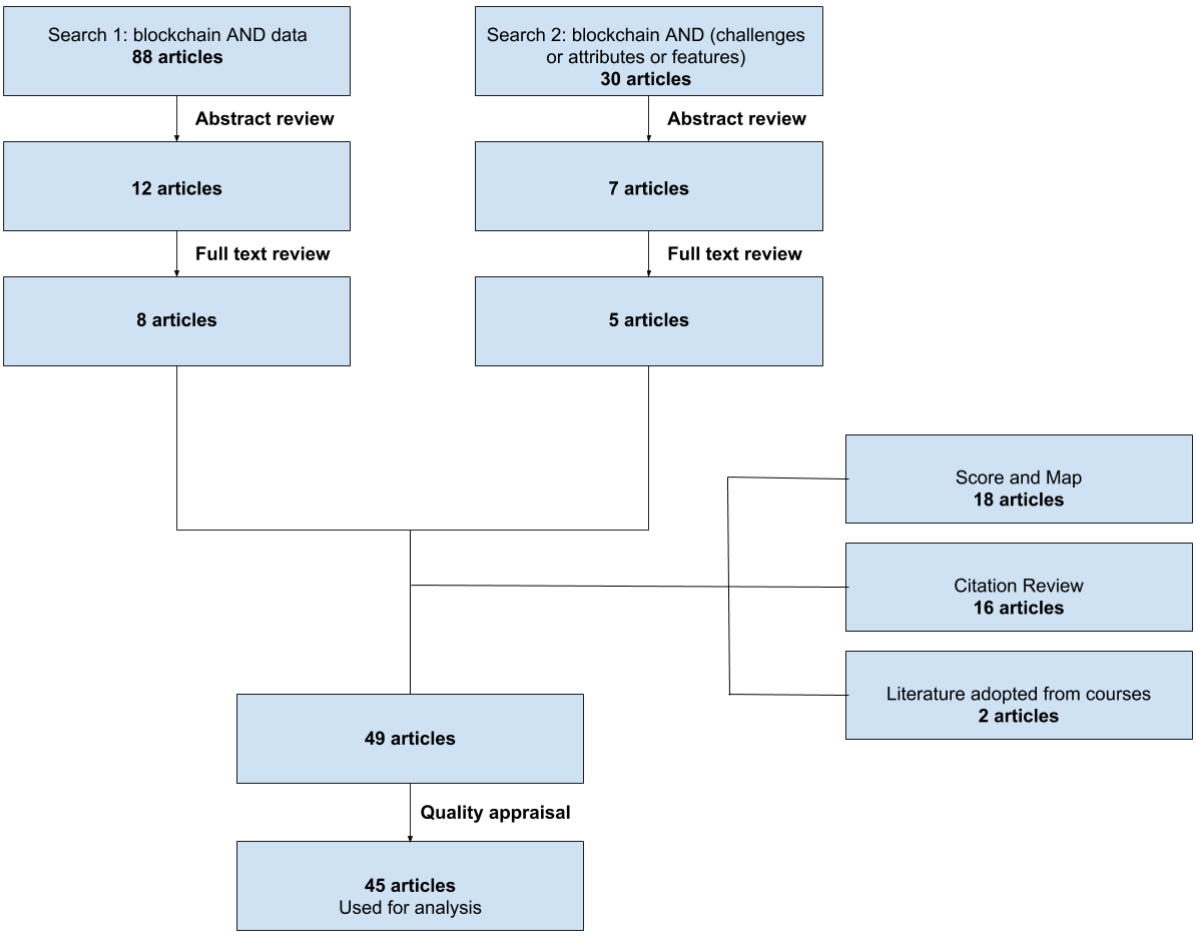
The article from *'IEEE Access'* has been maintained. This journal has only been around since 2013, the minimum requirement for having an AI score. However, it already scores a 67, while leading journals typically score a 70 or higher. The article from *'Environments'* has been maintained for analysis. It was an opinion article that provided a lot of insights on possible challenges that could occur implementing blockchain. Articles from the *'International Journal of Distributed Sensor Networks',* the *'International Journal of Energy Research',* and the *'Journal of Computer Science and Technology'* have all been included because they provide an application of Blockchain technology in a

somewhat similar way, and this provided valuable information for this review. The article from Engelhardt (2017) that was published in *'Technology and Innovation Management Review'* has been included although no AI score could be found of the journal. The article from *'Telecommunications Policy'* has been included due to its valuable insights to security and privacy benefits of using blockchain technology.

2.2.5: Data extraction

The process described above has resulted in a total of 45 articles that have been used for analysis. A systematic overview of the selection process is presented in figure 2 below.

Figure 2: Systematic overview of the selection process



Chapter 3: Results of the Traditional Review on the Replication Crisis

In this chapter, the replication crisis will be studied by conducting a traditional review (Jesson et al. 2013). This chapter includes information about the size of the problem, causes of the replication crisis, and the effects of non-replicable research will be discussed.

3.1: Size of the problem

Replication of research protects against false positives. If a finding cannot be replicated it is most likely that it was a false positive. Using a level of significance of 0,05 you would expect the number of false positives to be around five percent. In practice, this percentage is a lot higher.

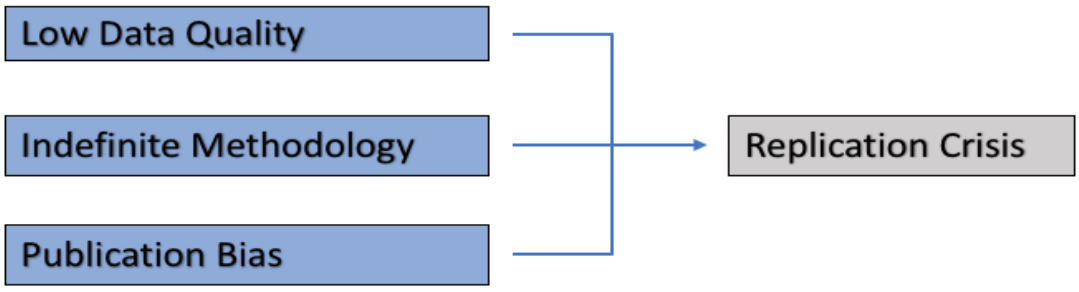
The Open Science Collaboration (2015) tested 100 publications of three leading journals in psychology. Only 36 percent of the findings that they tested could be replicated. The reviewed journals followed by their percentage of replicable articles are: Journal of Personality and Social Psychology (23%), Journal of Experimental Psychology (48%), Psychological Science: social articles (29%), and Psychological Science: cognitive articles (53%).

Medical scientific publications have been coping with a similar problem. Lack of reproducibility has been studied extensively in this field. Ioannidis et al. (2005) estimated a rate of about 80% non-reproducible studies within medical sciences. According to Benchoufi and Ravaud (2017) this high rate of non-reproducible studies may be related to several types of errors, misconduct or fraud.

3.2: Causes of the replication problem

Why are so many findings unable to be replicated? In the literature, three main causes for the replication problem have been found (see figure 3). Most importantly, a lack of data quality can lead to irreplicable results. Secondly, indefinite methodologies make it hard for reviewers to reproduce research correctly, thus leading to a replication problem. Finally, journals seem to prefer confirmed hypothesis, leading to a certain need for the researchers to confirm their hypothesis. These causes of the replication crisis will be further discussed in this paragraph.

Figure 3: Causes of the replication crisis



3.2.1: Data quality

High quality data is data that is fit for use by data consumers (Strong, Lee & Wang, 1997). Following their definition of high quality data, they evaluate data quality (DQ) on four categories. These categories are: Intrinsic DQ, Accessibility DQ, Contextual DQ, and Representational DQ. An overview of the categories and their dimensions can be found in the table below (table 12, strong et al. 1997, p. 104).

Table 12: DQ categories and dimensions

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-added, Timeliness, completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

Note: Adapted from "Data quality in context.", by Strong, D. M., Lee, Y. W., & Wang, R. Y., (1997). Communications of the ACM, 40(5), p. 104

Low quality data is one of the primary causes of the replication crisis. A low score on any of the four categories could result in irreplicable results. Some examples will be discussed to illustrate this.

Intrinsic DQ refers to the accuracy and objectivity of data (strong et al. 1997). A couple of situations came forward in the literature that could result in low intrinsic DQ. The most remarkable is scientific misconduct. Another reason why intrinsic DQ could drop is that respondents filling in surveys often follow a specific strategy which can result in disparate answers (Blasius & Thiessen, 2015).

The National Science Foundation (2001) defined scientific misconduct as fabrication, falsification, or plagiarism in proposing, performing, or reviewing research or in reporting research results. This misconduct must be done intentionally, knowingly, or in disregard of the accepted practices. Fanelli (2009) conducted a meta-analysis on survey data to investigate how big the scope of this problem is. His study showed that about two percent of scientists admitted to have fabricated, falsified or modified data or results at least once. A much bigger portion of researches admitted to other questionable research practices like *'dropping data points based on a gut feeling'*, and *'changing the design, methodology or results of a study in response to pressures from a funding source'*. Fanelli (2009) also argues that we can assume these numbers to be even higher, because most scientists would not admit to something like this even if anonymity is guaranteed. Stroebe et al. (2012) created a list with 41 scientists who conducted serious scientific misconduct. One of them is Diederik Stapel, an internationally renowned Dutch social psychologist who received multiple prestigious awards for his research. He admitted to have faked the results of a number of studies after some of his students reported this to the university (Bhattacharjee, 2013). Also, Mark Hauser (a popular Harvard professor), and Karen Ruggiero (University of Texas) have been caught falsifying their research data. This shows that scientific misconduct is happening in all levels of the scientific community.

Accessibility DQ achieves a high score when the data is easy to access for the researcher, but the data is kept secure. Improving data security usually comes with a decrease in accessibility (strong et al. 1997).

Contextual DQ refers to the completeness of a data set (Strong et al. 1997). Plant and Parker (2013) argue that for biological research, the data submitted with the results often is the climatic stage of multiple experiments. They add to this that within the current publication model, no space exists to include these auxiliary data sets with the final publication. With the shift towards an online publication model however, there really is no excuse to exclude these supplementary materials. The authors then describe high quality data as carefully, rigorously achieved and precisely controlled data sets. Also Nekrutenko and Taylor (2012) conclude that a lot of published paper have incomplete data.

Representational DQ is high when the data is easy to understand and interpret for the data consumer (Strong et al. 1997). Marino (2014) called attention to the poor understanding of statistical tools a lot of researches seem to have according to his study. They also identified that emerging technologies in genetics, like 'omics', lead to data sets which the majority of researchers are unable to interpret, or in some cases the statistical methods for inference has not even been developed yet. Macleod (2011) has also appealed for increased statistical rigor and Pusztai et al. (2013) emphasize the complexity of biomedical research and therefore, the lack of statistical skills that many laboratory and clinically trained scientist seem to have.

3.2.2: Indefinite methodology

According to Henderson, Kimmelman, Fergusson, Grimshaw and Hackamet (2013) there is a need to formalize the process of guideline development within clinical science and experiments. They state that very few of the guidelines in the sample they reviewed used an explicit methodology, also the use of evidence to support recommendations was sporadic. The absence of an explicit mythology makes it hard for reviewers and other researchers to replicate a study. Begley and Ioannidis (2014) call for a rethink in research methods and increased standardization of research processes as a solution to the replication problem.

3.2.3: Publication bias

The publication bias refers to the fact that the decision to publish a paper often depends on getting positive research finding. According to Pusztai et al. (2013) the salaries of most scientists at research-focused universities are partly or even completely self-funded through research grants. Because of this, obtaining grant funding has become a career must for every scientist.

According to Simmons, Nelson and Simonsohn (2011), in many cases a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. Since it is uncommon for prestigious journals to publish null findings or exact replications, researchers have little incentive to even attempt them. *"Our goal as scientists is not to publish as many articles as we can, but to discover and disseminate truth. Many of us— and this includes the three authors of this article—often lose sight of this goal, yielding to the pressure to do whatever is justifiable to compile a set of studies that we can publish. This is not driven by a willingness to deceive but by the self-serving interpretation of ambiguity, which enables us to convince ourselves that whichever decisions produced the most publishable outcome must have also been the most appropriate"* (Simmons et al. 2011, p. 1365)

3.3: The effects of incorrectly published results

Publishing false results leads to unnecessary costs (Smith & Houghton, 2013; Simmons et al. 2011). The most significant cost being the potential for misallocation of research resources. According to Smith and Houghton (2013) this may also lead to the allocation of patients to clinical trials with little to no chance of success within the medical field. Simmons et al. (2011) add that false publications can motivate investments in fruitless research programs and could lead to ineffective policy changes. A (scientific) field that is known for publishing false positives loses its credibility from doing this.

Chapter 4: Results of the Systematic Review on Blockchain for Data Management in Scientific Research

In the first paragraph of this chapter, literature on blockchain technology is reviewed to provide an overview of the characteristics of blockchain and discuss the various design decisions when starting a blockchain network. In the second paragraph, this study will look into the possibilities of using blockchain for data management in scientific research. In this last paragraph the design decisions will be discussed for such a network and the benefits and challenges of switching to such a network will be reviewed.

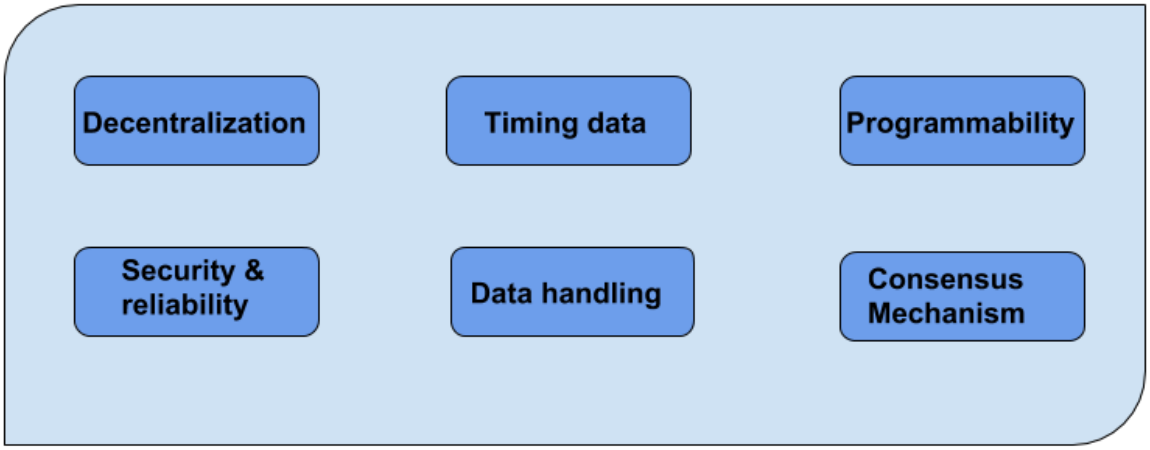
4.1: Blockchain

Blockchain, also called a *'distributed ledger'*, can be described as a large, secure, decentralized and public datastore of ordered records or events, called blocks. According to Hou et al. (2017) the blockchain database consists of two different sorts of records: blocks and transactions. Blocks consist of batches of valid transactions that are hashed and encoded, for Bitcoin by the SHA-256 cryptographic algorithm, into a Merkle tree. All of these blocks include a timestamp which is linked to a previous block (Benchoufi & Ravaud, 2017). The concept is still very new and was first used as a public ledger for the Bitcoin by Nakamoto (2008) when he published his paper introducing the world's first cryptocurrency. In the past 10 years a lot of new ideas and concepts have been introduced, but the growth in capabilities and cost-effectiveness are not expected to slow down anytime soon (Lin et al. 2017). Engelhardt (2017) argues that blockchain technology is best applicable to projects where there is more than one stakeholder, more trust than currently exists is required, a intermediary that could be omitted to increase trust or efficiency, a need exists for reliable tracking of information, and there is a need for reliable data over time. In this paragraph the characteristics of a blockchain will first be elaborated on. Afterwards, the design decisions for setting up a blockchain network will be discussed.

4.1.1: Characteristics of blockchain

This study has examined the literature on blockchain to identify common characteristics of blockchain technology. This has been done systematically by mapping the identified characteristics in table 13 (see appendix 1). The identified characteristics are: decentralization, timing data, programmability, security and reliability, data handling, and consensus mechanisms (see figure 4).

Figure 4: Characteristics of blockchain



Decentralization refers to the consensus that blockchains are distributed ledgers, multiple parties are holding the data. There is no single source that claims authority over the true data (Engelhardt, 2017). The blockchain uses pure mathematical methods to establish trust relations among distributed nodes, instead of trusting central institutions. Resulting in a decentralized and trustworthy system (Zheng et al. 2018). Transactions can be verified in a secure and traceable way without the need of a third party (Benchoufi & Ravaud, 2017; Lin et al. 2017). According to Kshetri (2017) the decentralized nature of blockchain is likely to result in a low susceptibility to manipulation and forgery by malicious participants.

Timing data refers to the traceability of data and transactions within a blockchain network. "*Blockchain stores data with a timestamped block structure. Thus, it adds time dimension to data, and has extremely strong verifiability and traceability."* (Zheng et al. 2018, p. 558) Because of this timestamp, users will have the ability to verify that uploaded data is theirs. This way of proving that something on the blockchain was yours is described by Wang et al. (2018) as proof of existence. Benchoufi & Ravaud (2017) refer to the relevancy of this application for health care data.

Data handling for blockchain is done in such a way that the data provider can keep some autonomy over its data. Kshetri (2017) talked about the possibility for individuals to control their own personal data. Also, the decentralized aspect of blockchain in combination with some of its security protocols makes it nearly impossible to change data without leaving a trace. Thus, data on the blockchain is immutable (Lee & Yang, 2018; Benchoufi & Ravaud, 2017). The decentralized feature of blockchain requires all the data to be stored at all the administrators (miners) computer, and thus the information on the blockchain is publicly available. However, privacy can still be secured by using encryption to secure the data to only be accessible with a private key (Kogure, Kamakura, Shima & Kubo, 2017; Engelhardt, 2017; Wang et al. 2018; Lee & Yang, 2018).

Security and reliability of a blockchain are related very closely. Security is ensured by an asymmetric cryptographic principle to encrypt data. *"The consensus algorithm forms a computing power to resist external attack, and prevent the blockchain data from being tampered. Thus, blockchain has higher security."* (Zheng et al. 2018, p. 559) According to Hou et al. (2017) the activities of any of the nodes (copies of the database) in the system are constantly being monitored by the rest of the network. Security of permissionless blockchains like bitcoin has is guaranteed through consensus mechanisms like Proof-of-Work. According to Lee and Yang (2018) this process, also referred to as 'mining' has to be kept difficult to assure security. Engelhardt (2017) argues that the security from external threats is being enforced by the fact that an external attacker would need to attack a majority of the nodes to harm the system.

Programmability of the blockchain makes it possible to build applications, cryptocurrencies and smart-contracts. Smart contracts use computer language to replace legal language to record terms (Hou et al. 2017). They can validate the completion of certain criteria and can be executed automatically (Benchoufi & Ravaud, 2017; Hou et al. 2017). Smart contracts can result in major savings on contract drafting, opportunistic behavior, and regulatory costs (Hou et al. 2017; Sklaroff, 2017). According to Yli-Huumo et al. (2016) the API that bitcoin uses to develop services is difficult to use.

Consensus mechanisms are the pre-defined process of verification for transactions. In the literature two sorts of consensus mechanisms where identified: Proof-of-Work (PoW) and Proof-of-Stake (PoS). Bitcoin uses PoW to secure its network. According to Nakamoto (2008) it involves scanning for a value that when 'hashed' begins with a number of zero bits. The 'hashing' is done with an encryption algorithm. For bitcoin the used algorithm is SHA-256. This is a military cryptographic algorithm. It creates an almost unique 256-bit (32-byte) signature for a piece of text, unconditionally of how much text the input is. The signature created is also referred to as a hash. The average work required for PoW grows exponential with the number of zero bits required. The article of Lee and Yang (2018) also discusses the exponential growth of work required by increasing the amount of zero bits. PoW consists of an intensive computerial hashing task that is being regulated by the *'blockchain difficulty'* that controls the average time spent by miners to create a new block. Once the miner manages to create a new block, it gets broadcasted to all the other miners, who will accept this block as the latest and start mining for new ones (Gaetani et al. 2017). The first one to solve the puzzle gets rewarded in Bitcoins (Kogure et al. 2017; Zheng et al. 2018). PoS is a cheaper variation of the PoW mechanism. It required less energy and computer power (Yeow et al. 2017; Hou et al. 2017).

4.1.2: Blockchain design decisions

In this sub-paragraph the decisions regarding the design of a blockchain network will be discussed. First, the differences between permissioned and permissionless blockchains will be discussed. Secondly, the distinction between a public and a private blockchain will be made. Finally, the concept of smart contracts will be briefly discussed.

Permissioned and permissionless blockchains

When setting up a blockchain network you will need to decide on whether you want a permissioned or a permissionless blockchain in an early stage. Therefore, it is important to understand what their possible benefits and drawbacks are.

Permissionless blockchains can be secured by everyone. They follow the earlier described proof of work approach in combination with some incentives for people to provide the work (typically in the form of tokens) to prevent attacks on the network (Bartlin et al. 2017; Gaetani et al. 2017; Yeow et al. 2017). Bitcoin is an example of a permissionless blockchain.

A permissioned blockchain implies that only specified parties can provide calculation power to maintain the blockchain (Bartlin et al. 2017; Engelhardt, 2017; Gaetani et al. 2017; Yeow et al. 2017). A permissioned blockchain results in less decentralization. However, this does not mean that the parties providing *'cryptographic power'* have any control over the content that is on the blockchain (Bartlin et al. 2017). Ripple is an example of a permissioned blockchain. According to Yeow et al. (2017) the use of trusted parties for the verification process makes the threat of a Sybil attack disappear and eliminates the mining requirements that constitute the economic incentive. This allows for the adoption of other consensus protocols. Because of the reasons described above a permissioned blockchain typically operates faster and more efficient than a permissionless one.

Public and private blockchains

Next to the decision on who secures the blockchain, decisions have to be made on who can use the blockchain. This is what distinguishes private blockchains from public blockchains.

Public blockchains have no restrictions on who is allowed to use the blockchain. It can be used by everybody. Most existing blockchains are public blockchains. Even though everybody can use the blockchain, data can still be kept private through encryption (Bartlin et al. 2017).

A private blockchain can only be used by certain parties or people. You will typically see these for company related blockchains. For example, the Port of Rotterdam is experimenting with blockchain to

keep track of information about the distribution of inventory. They use a private blockchain that only allows the Port of Rotterdam, the suppliers, the buyers and the logistic partners to see this information.

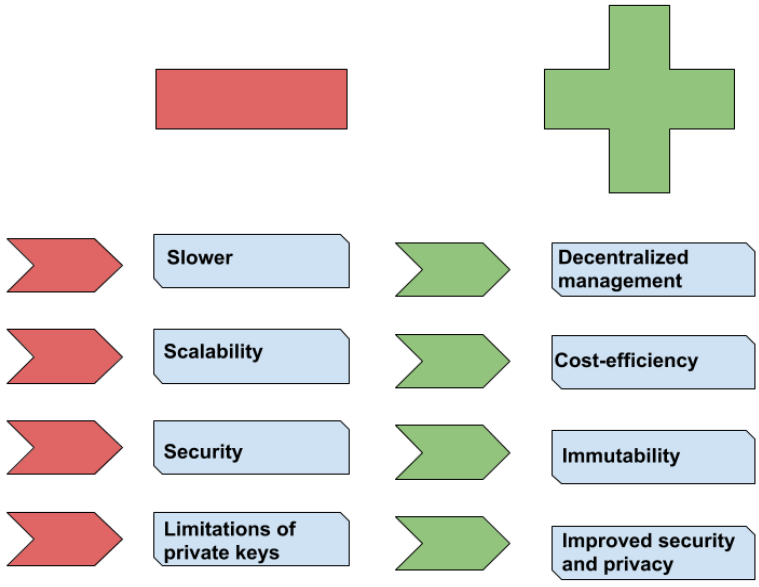
Smart contracts

A smart contract allows for automatic contract execution given predetermined conditions (Kogure et al. 2017). These conditions can be made obligatory to assure the completion of certain processes: *"Smart Contracts enable the validation of a step with the only condition that every preceding step has been fully validated."* (Benchoufi & Ravaud, 2017, p. 4)

4.2: Blockchain for data management in scientific research

This paragraph will provide an overview on how blockchain technology could assist the data collection process of the research cycle. First, the design decisions will be discussed. After which the possible challenges and benefits of using blockchain technology for data management in scientific research will be discussed (see figure 5).

Figure 5: Challenges and benefits for the scientific community of using blockchain for data management in scientific research



4.2.1: Design decisions

Bartlin et al. (2017) argue that it is best to choose permissioned over permissionless. For an application in science it could still provide more decentralization then existing information systems. Multiple universities and other trusted parties could provide the *'cryptographic power'* to secure the blockchain database. According to Bartlin et al. (2017) *"They cannot censor or approve beyond the defined protocol in the blockchain system (Not like trusted third parties in the current sense such as service providers (e.g. journal publishers, universities, centralized data repositories, etc.))."* (p. 10)

The blockchain should be public to allow all researchers and other users to use the application. According to Bartlin et al. (2017) this is applicable for both public and non-public data, they add that it is important to understand that *"public/private says nothing about who will be able to read the content. For example, a public blockchain can still be used to secure non-public research data."* (p. 10)

There could be some predetermined smart contracts in place to increase the trust in research even more. Benchoufi and Ravaud (2017) provide two examples on how smart contracts could increase trust in medical research. First, they propose to set up a smart contract that verifies that the designed methodology has been followed. The authors would then present the publication itself and the set of blocks that constitute the smart contract to the publisher to provide proof that the study was conducted according to the presented methodology. Second, they call for a smart contract that allows for patient inclusion in a data set, immediately after they have consented for the use of their data.

4.2.2: Challenges to the scientific community of using blockchain technology for data management

Blockchain technology is still in in its infancy. In the past ten years some challenges have been identified. The focus for the identification of challenges was the impact of blockchain on the whole scientific community. The challenges identified during this literature review are: the speed of the blockchain networks transaction handling, scalability, security, and individual responsibilities of users (see figure 5). A systematic overview of the identified challenges can be found in table 14 (see appendix 1).

Speed of transaction handling

Transactions are grouped together in a 'block' awaiting confirmation from the consensus mechanism. According to Yeow et al. (2017) block creation rate is 10 minutes per block. Compared to centralized transaction handlers, blockchain is not very effective in handling transaction (Kongure et al. 2017). The popular payment company Visa peaked at 47.000 transactions per second during the holidays of 2013 while in 2015 bitcoin could barely manage to get up to seven transactions per second (Poon & Dryja, 2015). This is caused by three aspects of blockchain. Namely: cryptographic verification (Kongure et al. 2017; Linn et al, 2017), consensus mechanisms, and redundancy (Linn et al, 2017). Increasing the speed of transactions could be done, but typically comes with a trade-off with data integrity, because the demands for proof of work would have to be reduced (Gaetani et al. 2017).

Scalability

The Bitcoin protocol is forced to have a very slow block creation rate of 10 minutes per block. The size of a block is also limited because the delay among nodes with a larger block size will cause more

undesirable forks to occur. (Yeow et al. 2017) According to the authors an increase in either the block size or the creation rate could result in limited guarantees of transaction irreversibility and it would reduce the required computational power to launch attacks on the system to below 50 percent. Furthermore, Lee and Yang (2018) identified long delays in transaction confirmation and called attention to the low scalability.

Security

Some security concerns exist with this decentralized peer-to-peer technology. Due to the democratic characteristics of the blockchain decisions are made by a majority of computer power. Gaetani et al. (2017) talk about the possibility of a '51% attack'. If an attacker manages to get more than half of the computer power they could change the existing protocols of the blockchain. Another security issue is that all information (public and non-public) has to be saved on all the 'miners' computers, and thus, is publicly available (Kogure et al. 2017).

Limitations of private keys

Non-public data is protected by encryption. To access this data the user will need to provide a private key to solve the encryption. Two risks emerge with this security measure. First, if the key gets stolen, the user cannot decrypt the data (Engelhardt, 2017; Kogure et al, 2017). Second, if the key is stolen the data will be at risk of decryption permanently because of the immutable characteristics of the blockchain (Kogure et al. 2017).

4.2.3: Benefits to the scientific community of using blockchain technology for data management

Although some challenges are very serious. A lot of benefits come forward in the literature. Again, the focus of this study was to identify the benefits for the full scientific community. The four benefits that have been identified are: decentralized management, cost-efficiency, immutability, and improved security and privacy (see figure 5). A systematic overview of the identified benefits of blockchain technology for data management can be found in table 15 (see appendix 1).

Decentralized Management

Blockchain takes away the need for a trusted third party because there is no single source that claims authority of the data, instead the data is being maintained by the community of users, and trust is encoded in the protocol (Lin et al. 2017; Engelhardt, 2017; Benchoufi and Ravaud, 2017; Hou et al. 2017; Kuo, Kim & Ohno-Machado, 2017). According to Zheng et al. (2017) the mathematical methods that blockchain use establish trust relations between the different nodes, forming a decentralized and trustworthy network.

Cost-efficiency

According to Lin et al. (2017) *"Blockchain competitiveness and cost-effectiveness are likely to increase following three laws: (1) Moore's law, i.e., time required for data processing halves every 18 months; (2) Kryder's Law, i.e., data storage halves every year; and (3) Nielsen's Law, bandwidth doubles every two years."* (p. 2) Engelhardt (2017) argues that the full potential value of blockchain has not been reached: *"Blockchains, also called distributed ledgers, enable a combination of cost reduction and increased accessibility to information by connecting stakeholders directly without requirements for third-party brokers, potentially giving better results at lower costs."* (p. 22)

Immutability

According to the literature (Lin et al. 2017; Lee & Yang, 2018; Engelhardt, 2017; Benchoufi and Ravaud, 2017; Zheng et al. 2017; Gaetani, 2017; Gipp Meuschke & Gernandt, 2015) information stored on the blockchain can be described as immutable. This means that it is not possible to change information without leaving a trace. Furthermore, records can be added, but can never be removed (Engelhardt, 2017). The timestamp that a user leaves on the blockchain when using it could also be used by researchers, artists or anyone else to prove that an idea was originally theirs (Gipp et al. 2015).

Improved Security and Privacy

Blockchain assures the security of data and data quality in a couple of ways. By using encryption only, the holders of the correct cryptographic keys can access the information (Engelhardt, 2017). Lee and Yang (2018, p. 11) add to this that *"When the data is stored on the system, the system uses blockchain technology to protect the privacy and correctness of the data, so as to improve users' trust in the system. Through blockchain technology, any modification records of data are able to be tracked."*

Chapter 5: Proposed Research Design

This study identified different causes of the replication crisis. This was done to illustrate the need for higher data quality in scientific research. Afterwards, the characteristics of blockchain technology have been identified and were systematically mapped. Finally, the potential benefits and challenges of using blockchain technology for data management in scientific research were also identified and mapped. This chapter will provide directions for further research on this topic. Two proposals are made. The first proposal is to study the potential challenges and benefits of adopting to blockchain technology for the individual researcher. The second proposal aims to identify all stakeholders that are involved in the scientific process and to identify their rationale for adopting to this new technology.

5.1: The Individual Researcher

This study provides an overview of the potential challenges and benefits for the scientific community as a collective identity from using blockchain for data management. However, the majority of scholars is responsible for their own research protocol and design. Moreover, they typically look for a publisher and funding by themselves. The individual researcher often is the decision maker in this situation. Knowing this, more research on potential challenges and benefits for the individual researcher of using blockchain technology for their data handling could help to provide insights on how to increase the adoption speed to this new technology.

After identifying the benefits of using blockchain some questions arose about why the scientific community did not yet adopt the technology to its process. Some research was done to map the existing initiatives for Blockchain for Science. DEIP (2017) is a decentralized blockchain based platform that focuses on reinventing the way research funding is divided. Another initiative that aims more at preserving data quality is one from Elosua, Brede, Ritola & Botev (2018) that tries to incorporate a blockchain based database of academic literature in combination with artificial intelligence based tools to assist users of the database in finding relevant data. They accuse existing publishers of limiting our knowledge creation by maintaining high prices to access academic literature. Another initiative that aims to increase data quality is DAT. Ogden, McKelvey and Madsen (2018) introduce a new data handling protocol that shows a lot of similarities to those of blockchain based applications.

As shown above, the possibility for the individual researcher to use a blockchain based platform to conduct their research on exists. Although the platforms already have a user base and the low adoption rate could be amortized to the infantry of the companies providing the service. The identification of the challenges and benefits of using blockchain technology for data management that the individual researcher could experience can provide valuable insights to increase the adoption rate.

5.1.1: Research Design

This sub-paragraph presents the proposed research questions and the corresponding research design. The problem statement of the study will be:

'The individual researcher does not seem to adopt to blockchain technology.'

The aim of this research will be to identify challenges and benefits the individual researcher experiences when adapting to blockchain technology for data management. To do so it will answer the following research questions:

1) '*What are the benefits for the individual researcher of using blockchain technology for data management?'*
2) '*What are the challenges for the individual researcher of using blockchain technology for data management?'*

The proposed research design is a sequential exploratory research design (Saunders et al. 2016). To identify the possible challenges and benefits it is recommended to use a qualitative study. Conducting interviews with scholars and preferably letting them use a potential blockchain based service to make sure they know what they are talking about. Furthermore, focus groups could be used as an additional form of data acquisition. After coding the interviews to identify challenges and benefits of using blockchain technology for data management for the individual researcher. The identified challenges and benefits should be tested through a survey. This survey should be filled in by a larger group of scholars to increase the reliability of the findings in the first phase (Saunders et al. 2016).

The approach described here is called the partially integrated mixed methods research approach (Nastasi & Hitchcock, 2015; Saunders et al. 2016). This refers to the fact that the study consists of multiple (two in this case) phases. Both phases using their own type of data. In this case the first phase uses qualitative data to generate a theory and the second phase quantitative data to confirm the reliability of these findings.

The research philosophy could be described as pragmatist. Pragmatists understand that there are a lot of different ways to interpret the world and to perform research. Thus, no single point of view could ever provide the entire picture (Saunders et al. 2016). This does not imply that pragmatists will always use a multi-method research design. They will however use a method that enables credible, well-founded, relevant, and reliable data to be gathered (Kelemen & Rumens, 2008).

This study serves two main purposes. Firstly, there is a exploratory purpose when identifying the possible challenges and benefits. *"An exploratory study is a valuable means to ask open questions to discover what is happening and gain insights about a topic of interest."* (Saunders et al. 2016, p. 174) The interviews conducted in the exploratory phase are typically somewhat unstructured. Therefore, the quality of the interviewees and other participants has a major impact on the quality of your data (Saunders et al. 2016).

Secondly, the purpose of the second phase of this study is to find out how well these identified challenges and benefits represent the average individual scholar. Research with such a purpose is described as evaluative studies (Saunders et al. 2016).

5.1.2: Quality of the Research Design

Reliability and validity are key ingredients for a high-quality research paper. This sub-paragraph will discuss the reliability and validity issues of the proposed research design.

Reliability refers to consistency and replication (Saunders et al. 2016). If the same methodology could be applied and the results are the same, you could say it is a reliable research finding. To increase the internal reliability of this paper is already being enforced by the two-phase process of the study. The findings in the first phase will be tested during the second phase. The internal reliability could be increased even more by using more than one researcher to conduct the interviews and analyze the data (Saunders et al. 2016). External reliability can be tested after the research is completed. Assuming that the methodology of the research provides sufficient information to other scholars to replicate the research.

"Validity refers to the appropriateness of the measures used, accuracy of the analysis of the results and generalisability of the findings" (Saunders et al. 2016, p. 202). It can be grouped in three types of validity. The first type is measurement validity and refer to whether or not the measures that are being used for analysis actually measure what they are intended to measure (Saunders et al. 2016). The second type is internal validity. *"Internal validity is established when the research accurately demonstrates a casual relationship between two variables."* (Saunders et al. 2016, p. 203) For this research this would typically be achieved when suggestions to attract more scholars to use blockchain technology are being tested. In the questionnaire some statistical tests can be included to exclude invalid surveys. An example of an appropriate test is the *'Cronbach's Alpha'*. External validity refers to the extent that a research finding can be generalized to other settings or groups. This would also be evaluated after the research is completed.

5.2: Identifying all Involved Stakeholders involved in the publication process and finding their Rationale for Implementing Blockchain

Some new initiatives have already been implemented to improve the research quality. Some examples are 'similarity check' (Watson, 2017) and 'iThenticate' ("Prevent Plagiarism in Published Work.", 2018) to check for plagiarism, and 'Crossmark' (Meddings, 2017) to monitor whether content has been updated or corrected and to give access to other valuable metadata. The Crossmark application provides metadata that would also be available when using blockchain for data management. The differentiator being that, with Crossmark there is still a central third party that secures the data. Giving the system a single point of failure resulting in a lower level of security and reliability (e.g. Lin et al. 2017; Engelhardt, 2017; Hou et al. 2017).

This study found that the research and publication process often involve a good number of stakeholders. This makes it complicated to implement changes. It might be even harder to set a standard format of for example the methodology. Due to the various interests of the stakeholders they might have a different idea about what a good methodology is. It also seems like the publishers are the most powerful stakeholders right now. They have been making a very profitable business out of knowledge creation for a while now. Although the movement towards online publications makes me question whether these publishers are really irreplaceable.

It seems like the pie has been cut in unfair pieces and the parties with the bigger pieces refuse to consider another pie. They seem satisfied with the distribution as it is right now. This research aims to identify these parties by examining the relation between profit margin and resistance to change.

In order to successfully implement a new technology in the publication industry it is important to identify the barriers the stakeholders might face implementing this new technology. Addressing these barriers individually will allow for more consensus to adapting to the new technology. During this study some stakeholder have already been identified. However, it is important that none are missed. Thus, it is recommended to search the literature for some additional insights on this.

5.2.1: Research Design

This sub-paragraph presents the proposed problem statement and research questions followed by the proposed research design for this study. The problem statement is:

'Implementing change in the publication industry is hard due to the number of stakeholders involved'

The focus of this study will be to identify the interests of the various stakeholders involved in the publication process for implementing blockchain for data management. To do this the following research question is proposed:

'What moves the various stakeholders to accept blockchain for data handling in the publishing process?'

To assist the researchers in answering this research question the following sub-questions will be addressed:

1) *'Who are the stakeholders involved in the research and publishing process?'*
2) *'Is there a positive relationship between profit margin and resistance to change?'*
    a) *'What are the profit margins of the various stakeholders'*
    b) *'To what extent are the stakeholders resisting to change?'*
3) *'What are the benefits for the various stakeholders of implementing blockchain for data management in the research and publishing process?'*
4) *'What are the drawbacks for the various stakeholders of implementing blockchain for data management in the research and publishing process?'*

A small additional literature search will be conducted to identify all stakeholders participating in the research and/ or publishing processes. After doing this, the proposed design is a concurrent research design. Being quantitative and qualitative at the same time. Although the focus in on the qualitative aspects of the study. The quantitative aspect is to calculate the second sub-question *'Is there a positive relationship between profit margin and resistance to change?'* This is to test the hypothesis that the stakeholders with the highest profit margins will show the most resistance to change. The remainder of this study will follow an interpretive research philosophy. This is due to the fact that the researcher needs to make sense of the socially constructed and subjective meanings expressed about the topic that is being studied (Saunders et al. 2016).

A multi-method qualitative method is recommended for this study. This involves using more than one method for data collection (Saunders et al. 2016). Data collection could be done by conducting interviews, participating in meetings, shadowing employees, and other methods that could provide valuable information about the different stakeholders' practices, processes, and interests.

This study serves an exploratory purpose. Explaining what to address to make blockchain an attractive alternative for all stakeholders. Because the interviews do not include a lot of structure it is important to ensure that the participants are familiar with the subject to attain a high quality of data (Saunders et al.

2016). An advantage of this explanatory research is that it is very flexible and adaptable to changes. The insights that are gained during the gathering of data should be taken seriously and the researchers should be open to change their research based on these new insights (Saunders et al. 2016).

5.2.2: Quality of the Research Design

The reliability and validity issues of this proposed research design will be discussed in this sub-paragraph.

For the reliability of this research it is important that the results are replicable. External reliability refers to the ability to be replicated by other researchers. To attain this, it is important that the methodology and the interview guide is complete and the process the researchers went through while conducting the study is transparent (Saunders et al. 2016). To ensure the external reliability even more it is important to interview multiple people with similar jobs at the same stakeholder. If a certain remark is being made by multiple people within a firm, chances are that the results will be replicable in future research. To increase the internal reliability, it is recommended to conduct interviews with a minimum of two researchers. It is also important to have a general agreement on how to code the collected data. To ensure that you have a similar idea about this it could be useful to code the first interview together. After this, the researchers could individually code another article and compare their results.

The measurement validity is an important factor in this study. A lot of assumptions and interpretations will have to be made. Therefore, it is important to evaluate the usefulness for answering any of the research question for every piece of data that is being acquired. If a certain quote comes forward in multiple interviews this will increase the measurement validity. Similar to the proposed measure for external reliability, the internal validity can be improved by interviewing multiple people with similar jobs at the same stakeholder. This will create more valid evidence for the information identified during the interviews. External validity is partly accounted for in the study design by identifying the relationships, challenges, and benefits independently for all stakeholders. If certain arguments are generalizable for all stakeholders this shows some external validity. External validity could also be found in the ability to replicate the findings of this study to another industry than the scientific industry (Saunders et al. 2016).

Chapter 6: Conclusion and discussion

6.1: Conclusion

The aim of this paper was to give an overview on how blockchain technology for data management can impact scientific research. To do so, knowledge about blockchain and the scientific field was required. A literature research was conducted to obtain this knowledge.

The literature review on the scientific field helped to address the first sub-question; identifying the causes of the replication crisis. The main cause of the replication crisis is a lack in data quality (DQ). Four categories of DQ are identified. A low quality on any of the categories can result in non-replicable research findings. The first category is intrinsic DQ. A low score here can be the result of scientific misconduct, data fabrication or survey strategies (Blasius & Thiessen, 2015). The second category is accessibility DQ, referring to the trade-off between the accessibility and security of the data. The third category identified is about the completeness of a data set and is called contextual DQ. A lot of research publishes only the datasets that lead to success. Leaving out the datasets they used earlier that did not deliver the required results (Nekrutenko & Taylor, 2012; Plant & Parker, 2013). The last identified category is representational DQ. This category is about the ability for the data consumer to interpret and understand the data. The literature refers to this problem with complicated statistical analysis where the researchers often do not possess the required statistical skills (Marino, 2014; Macleod, 2011; Pusztai et al. 2013).

Two other causes of the replication crisis that have been identified are: indefinite methodology, and the publication bias. To replicate a research, you have to execute all the steps called upon in the methods. If the methods are incomplete or inaccurate the research cannot be replicated correctly (Begley & Ioannidis, 2014; Henderson et al. 2013). The publication bias refers to the fact that the decision to publish a paper often depends on getting positive results. The fact that researchers' salaries often depend on publishing papers could motivate scholars to conduct in scientific misconduct to confirm their hypothesis (Pusztai et al. 2013).

Publishing false results leads to unnecessary costs. Some examples are misallocation of research resources and ineffective policy changes (Simmons et al. 2011). False results in the medical field can result in harmful medication on the market (Smith & Houghton, 2013).

The systematic review on the available literature on blockchain was executed to address the second and third sub-questions: identifying the characteristics of blockchain technology; and mapping the possible challenges and benefits for the scientific community of using blockchain for data management. Also

contributing to the research question; how blockchain for data management of research data can result in a higher rate of replicable studies. The identified characteristics are: decentralization, timing data, programmability, security and reliability, data handling, and consensus mechanisms. The challenges for the scientific community of implementing blockchain for data management are: speed of transaction handling, scalability, security, and limitations of private keys. The potential benefits are: decentralized management, cost-efficiency, immutability, and improved security and privacy.

Blockchain can be characterized as decentralized. Meaning that there is no central third-party that secures the data, but instead a distributed ledger is being secured by multiple participants (nodes) of the network (e.g. Engelhardt, 2017; Lin et al. 2017; Benchoufi & Ravaud, 2017). Another characteristic is the ability to timestamp data, adding a time dimension to the data. Contributing to data verifiability and traceability (Zheng et al. 2018). Additional to the ability to time data, the way blockchain handles data is different from existing data management systems. Allowing individuals to have autonomy over their personal data (Kshetri, 2017), making it impossible for users to change data without leaving a trace (Lee & Yang, 2018; Benchoufi & Ravaud, 2017), and still guaranteeing privacy through encryption and public keys although the data processing process is publicly available for transparency (Kogure et al. 2017; Engelhardt, 2017; Wang et al. 2018; Lee & Yang, 2018). Blockchain can be described as secure and reliable due to its cryptographic security and consensus mechanism (Hou et al. 2017; Zheng et al 2018). Because a copy of the blockchain is being kept on all the nodes, an external attacker would need to change the majority of the nodes to make a change. This is also referred to as a 51% attack (Engelhardt, 2017). The programmability of the blockchain makes it possible to build applications, cryptocurrencies and smart-contracts on it. Smart contracts use computer language to replace legal language to record terms (Hou et al. 2017). For science applications, smart contracts can serve to standardize methodology, and to assure that the methodology is being followed (Benchoufi & Ravaud, 2017). Consensus mechanisms are the pre-defined process of verification for transactions. Two sorts of consensus mechanisms are identified: Proof of Work (PoW), and Proof of Stake (PoS) (Yeow et al. 2017; Hou et al. 2017).

Compared to the existing trusted third parties, blockchain is not very efficient with handling transactions (Kongure et al. 2017). Transactions need to be grouped together in a 'block' and wait for consensus by the network. Thus, leading to long transaction times. Increasing transaction speed can be done, but it will reduce the data integrity because it requires lower demands for PoW. This is caused by the security protocols that the blockchain needs. The Bitcoin network for example was not able to process more than seven transactions per second as of 2015 (Poon & Dryja, 2015). With the increase of Central Processing Unit (CPU) and Graphical Processing Unit (GPU) power, and with the raise of quantum computing, some questions arise about the security of the encryption of the data. Although, as of today, no one has succeeded to crack the encryption. Finally, some limitations of the private keys used to access encrypted

data have been identified. One being that losing a private key will result in permanently losing access to the data being secured by the key (Engelhardt, 2017). The other limitation is that when the key gets stolen. The data will forever be at risk of decryption. This is because of the immutable attribute of the blockchain (Kogure et al. 2017). Although some articles called attention to the security problems, and the limitations of private keys. These challenges seem somewhat farfetched, and more research is required to evaluate the impact of these challenges.

A negative relationship between scalability and security has been identified during the mapping of challenges for blockchain for data management in scientific research. The literature (Lin et al. 2017; Yeow et al. 2017) implies that an increase in scalability (increasing the size or the throughput time of a block) will result in a situation where a smaller amount of computing power is needed to perform attacks on the network, thus leading to a lower level of security. However, choosing for a permissioned blockchain network allows you to have control over who can secure the blockchain. Within the scientific community there are already some trusted parties like universities, publishers, and research groups that could provide the computer power. Reducing the external security risks. Thus, allowing for more scalable design decisions in the consensus protocols.

On the blockchain, data is being maintained by the community of users. Trust is encoded in the protocol. This takes away the need for a trusted third party and is also referred to as decentralized management. Lin et al. (2017) expect the competitiveness and cost-effectiveness of blockchain to increase following Moore's Law, Kryder's law, and Nielsen's Law. Although concrete evidence of this is still lacking there seems to be an agreement that the full potential of blockchain has not yet been achieved, and the networks will operate more efficient as time passes by. The most important benefit for data management for the scientific community is the immutability of data. Records can be added, but can never be removed (Engelhardt, 2017). Timestamps allows creators of intellectual property to prove an idea was theirs. Timestamps also ensure that data uploaded cannot be changed by researchers without leaving a trace. Making sure that a researcher cannot make a few minor changes in his data set to achieve his 0,05 level of significance for example. Another feature of blockchain that secures the privacy of data is the use of cryptography to ensure that private data is only available to people that own a private key. Allowing only those with a private key to encode the encrypted data.

This study will now provide an answer to the research question; how blockchain for data management of scientific research can result in a higher rate of replicable studies. Firstly, the issue of indefinite methodology can be addressed through the programmability characteristic of the blockchain. For example, smart contracts can be programmed to only accept the publication of a paper if specific demands of the methodology have been followed. This way a certain quality of methodology can be maintained. Providing researchers with sufficient information to review or reproduce the article.

Furthermore, this same programmability characteristic can be used to address the publication bias. By programming certain demands to the articles methodology, and possibly some other demands, into a smart contract. Publishers can automate the publish decision and minimize the publication bias. Although the technology provides this possibility, no evidence has been found that this will happen. More research about this specific application is required to say something about the feasibility.

Finally, the DQ can be improved by using the new technology. Intrinsic DQ will be higher due to the immutability of the data. Researcher cannot change the data without leaving a trace after the data got timestamped. The accessibility DQ can be improved through the encryption and private key security features of the blockchain. Allowing researchers to share data in a secure and transparent way, only with the people they intend to share it with. This allows for data sharing in earlier stages. Thus, leading to more reliable results. Contextual DQ will also be higher if all the data sets are being timestamped before performing analysis on them. Creating transparency about the earlier used data sets and their reasons of failing contribute to the knowledge within the scientific field and help researchers to stop following false leads. The final category of DQ is about the ability of researches to interpret the data and cannot really be improved through a new data management system. However, showing an improvement to three out of the four categories of the most important cause of the replication crisis is definitely a worthy result.

6.2: Discussion

Blockchain for data management in science brings a lot of possibilities. The amount of stakeholders involved in the research and publication processes makes it complicated to implement changes or to keep a certain standard. A lot of research groups, journals, publishers, and universities have their own protocols and standards to assure the quality of their work. Therefore, it is important to identify the rationale for these different stakeholders to adopt to this new technology. That is also why the proposed research design aims to identify these rationales.

Some limitations have been identified when coding the articles to identify characteristics, benefits and challenges of blockchain. The potential increase in cost-effectiveness of blockchain technology is only backed up by a limited amount of literature and is mainly based on assumptions. Thus, further research is required to test the increase in cost effectiveness. Moreover, the limitations of private keys present the private key as a weak characteristic of the blockchain. However, the private keys create a lot of new opportunities to for example give people autonomy over their personal data and allow researchers to share their data with selected groups of people. The amount of literature presenting the private key as a challenge to blockchain is limited. For the characteristics of blockchain the choice has been made to

keep timing data separate from data handling. Although timing data is one of the aspects of the data handling characteristics. The timestamping characteristic is so important for science that the decision has been made to keep it separate. This allowed to code more specific information to the timing characteristic.

6.3: Managerial Implications

The results show that the medical and psychological field are in need of better data management. This indicates that companies building blockchain applications for data management in scientific research should focus on these fields first. The results also show possibilities to improve data autonomy. This allows individuals to control the use of their personal data at any moment. This can be of interest to the medical field that requires access to a lot of sensitive personal data. Furthermore, the results show that using smart contracts can serve as a cost-efficient way to confirm the completion of pre-specified tasks. The results also suggest that this can be used to confirm the use of correct and complete methodology. Furthermore, the results suggest the possibility to automate exclusion decisions of literature based on pre-programmed criteria. Finally, the results show that the data quality will increase from using blockchain for data collection. The results confirm that scholars cannot make changes in their data without leaving a trace. They also confirm the ability for researchers to share data in a secure and transparent way, only with those whom it is intended for. Allowing for more data sharing in scientific research.

References

Bartlin, Sönke, & et contributors to living document. (2017). *Blockchain for Open Science and Knowledge Creation*. 10.5281/zenodo.401369

Bartlin, Sönke, & et contributors to living document (10). (2017). *Blockchain for Open Science and Knowledge Creation*. 10.5281/zenodo.401369

Beck, R., Czepluch, J. S., Lollike, N., & Malone, S. (2016). *BLOCKCHAIN – THE GATEWAY TO TRUST- FREE CRYPTOGRAPHIC TRANSACTIONS.* AIS Electronic Library. Retrieved May 22, 2018, from http://aisel.aisnet.org/ecis2016_rp/153

Begley, C. G., & Ioannidis, J. P. (2014). *Reproducibility in Science: Improving the Standard for Basic and Preclinical Research*. Circulation Research, 116(1), 116-126. doi:10.1161/circresaha.114.303819

Benchoufi, M., & Ravaud, P. (2017). *Blockchain technology for improving clinical research quality*. Trials, 18(1). doi:10.1186/s13063-017-2035-z

Benchoufi, M., & Ravaud, P. (2017). *Blockchain technology for improving clinical research quality* (4). Trials, 18(1). doi:10.1186/s13063-017-2035-z

Bhattacharjee, Y. (2013, April 26). *The Mind of a Con Man.* The New York Times. Retrieved June 1, 2018, from http://web.missouri.edu/~segerti/capstone/StapelLying.pdf

Blasius, J., & Thiessen, V. (2015). *Should we trust survey data? Assessing response simplification and data fabrication*. Social Science Research, 52, 479-493. doi:10.1016/j.ssresearch.2015.03.006

Boritz, J. E. (2005). *IS practitioners views on core concepts of information integrity*. International Journal of Accounting Information Systems, 6(4), 260-279. doi:10.1016/j.accinf.2005.07.001

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). *Power failure: Why small sample size undermines the reliability of neuroscience*. Nature Reviews Neuroscience, 14(5), 365-376. doi:10.1038/nrn3475

Decentralized research platform. (2017). Retrieved June 13, 2018, from https://deip.world/

Earp, B. D., & Trafimow, D. (2015). *Replication, falsification, and the crisis of confidence in social psychology*. Frontiers in Psychology, 6. doi:10.3389/fpsyg.2015.00621

Eigenfactor: *Revealing the Structure of Science*. (n.d.). Retrieved June 11, 2018, from http://www.eigenfactor.org/

Elosua, J., Brede, A. S., Ritola, M., & Botev, V. (2018). *Home – Iris.ai - Your Science Assistant*. Retrieved June 13, 2018, from https://iris.ai/

Engelhardt, M. A. (2017). *Hitching Healthcare to the Chain: An Introduction to Blockchain Technology in the Healthcare Sector*. Technology Innovation Management Review, 7(10), 22-34. doi:10.22215/timreview/1111

Engelhardt, M. A. (2017). *Hitching Healthcare to the Chain: An Introduction to Blockchain Technology in the Healthcare Sector* (22). Technology Innovation Management Review, 7(10), 22-34. doi:10.22215/timreview/1111

Fanelli D (2009) *How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data*. PLoS ONE 4(5): e5738. https://doi.org/10.1371/journal.pone.0005738

Gaetani, E., Aniello, L., Baldoni, R., Lombardi, F., Margheri, A., & Sassone, V. (2017). *Blockchain-based Database to Ensure Data Integrity in Cloud Computing Environments*. ITA-SEC, 1816. Retrieved from https://CEUR-WS.org.

Gipp, B., Meuschke, N., & Gernandt, A. (2015). *Decentralized Trusted Timestamping using the Crypto Currency Bitcoin*. IConference. Retrieved from https://arxiv.org/abs/1502.04015.

Guo, H. (2015). *Big data for scientific research and discovery*. International Journal of Digital Earth,8(1), 1-2. doi:10.1080/17538947.2015.1015942

Henderson, V. C., Kimmelman, J., Fergusson, D., Grimshaw, J. M., & Hackam, D. G. (2013). *Threats to Validity in the Design and Conduct of Preclinical Efficacy Studies: A Systematic Review of Guidelines for In Vivo Animal Experiments.* PLoS Medicine, 10(7). doi:10.1371/journal.pmed.1001489

Hou, J., Wang, H., & Liu, P. (2018). *Applying the blockchain technology to promote the development of distributed photovoltaic in China*. International Journal of Energy Research, 42(6), 2050-2069. doi:10.1002/er.3984

Ioannidis, J. P. (2007). *Why Most Published Research Findings Are False: Authors Reply to Goodman and Greenland*. PLoS Medicine, 4(6). doi:10.1371/journal.pmed.0040215

Irving, G., & Holden, J. (2016). *How blockchain-timestamped protocols could improve the trustworthiness of medical science*. F1000Research, 5, 222. doi:10.12688/f1000research.8114.2

Jesson, J., Matheson, L., & Lacey, F. M. (2013). *Doing your literature review: Traditional and systematic techniques*. Auckland, N.Z.: Royal New Zealand Foundation of the Blind.

Kelemen, M., & Rumens, N. (2008). *An Introduction to Critical Management Research*. London: SAGE Publications. http://dx.doi.org/10.4135/9780857024336

Kepes, S., & Mcdaniel, M. A. (2013). *How Trustworthy Is the Scientific Literature in Industrial and Organizational Psychology?* Industrial and Organizational Psychology, 6(03), 252-268. doi:10.1111/iops.12045

Kingori, P., & Gerrets, R. (2016). *Morals, morale and motivations in data fabrication: Medical research fieldworkers views and practices in two Sub-Saharan African contexts*. Social Science & Medicine, 166, 150-159. doi:10.1016/j.socscimed.2016.08.019

Kogure, J., Kamakura, K., Shima, T., & Kubo, T. (2017). *Blockchain Technology for Next Generation ICT*. FUJITSU Sci. Tech., 53(5), 56-61. Retrieved from http://www.fujitsu.com/global/documents/about/resources/publications/fstj/archives/vol53-5/paper09.pdf

Kshetri, N. (2017). *Blockchains roles in strengthening cybersecurity and protecting privacy*. Telecommunications Policy, 41(10), 1027-1038. doi:10.1016/j.telpol.2017.09.003

Kuo, T., Kim, H., & Ohno-Machado, L. (2017). *Blockchain distributed ledger technologies for biomedical and health care applications*. Journal of the American Medical Informatics Association,24(6), 1211-1220. doi:10.1093/jamia/ocx068

Lee, S. H., & Yang, C. S. (2018). *Fingernail analysis management system using microscopy sensor and blockchain technology.* International Journal of Distributed Sensor Networks, 14(3), 155014771876704. doi:10.1177/1550147718767044

Lee, S. H., & Yang, C. S. (2018). *Fingernail analysis management system using microscopy sensor and blockchain technology* (11). International Journal of Distributed Sensor Networks, 14(3), 155014771876704. doi:10.1177/1550147718767044

Lin, Y., Petway, J., Anthony, J., Mukhtar, H., Liao, S., Chou, C., & Ho, Y. (2017). *Blockchain: The Evolutionary Next Step for ICT E-Agriculture*. Environments, 4(3), 50. doi:10.3390/environments4030050

Lin, Y., Petway, J., Anthony, J., Mukhtar, H., Liao, S., Chou, C., & Ho, Y. (2017). *Blockchain: The Evolutionary Next Step for ICT E-Agriculture* (2). Environments, 4(3), 50. doi:10.3390/environments4030050

Macleod, M. (2011). *Why animal research needs to improve.* Nature, 477(7366), 511-511. doi:10.1038/477511a

Marino, M. J. (2014). *The use and misuse of statistical methodologies in pharmacology research*. Biochemical Pharmacology, 87(1), 78-92. doi:10.1016/j.bcp.2013.05.017

Meddings, K. (2017, January 30). *Crossmark - Crossref*. Retrieved June 20, 2018, from https://www.crossref.org/services/crossmark/

Nakamoto, S. (2008). *"Bitcoin P2P e-cash paper"*. Bitcoin.org. Retrieved May 22, 2018, from https://bitcoin.org/bitcoin.pdf.

Nastasi, B. K., & Hitchcock, J. H. (2015). *Mixed methods research and culture-specific interventions: Program design and evaluation*. Thousand Oaks: SAGE Publications.

National Science Foundation. (2001). *New research misconduct policies*. Retrieved from http://www.nsf.gov/oig/session.pdf

Nekrutenko, A., & Taylor, J. (2012). *Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility.* Nature Reviews Genetics, 13(9), 667-672. doi:10.1038/nrg3305

Ogden, M., McKelvey, K., & Madsen, M. B. (2018). *Dat - Distributed Dataset Synchronization And Versioning*. Retrieved June 14, 2018, from https://www.datprotocol.com/

Okoli, C., Schabram, K. (2010). *"A Guide to Conducting a Systematic Literature Review of Information Systems Research"*. Sprouts: Working Papers on Information Systems, 10(26). http://sprouts.aisnet.org/10-26

Open Science Collaboration. (2015). *Estimating the reproducibility of psychological science*. Science, 349(6251). doi:10.1126/science.aac4716

Plant, A. L., & Parker, G. C. (2013). *Translating Stem Cell Research from the Bench to the Clinic: A Need for Better Quality Data.* Stem Cells and Development, 22(18), 2457-2458. doi:10.1089/scd.2013.0188

Poon, J., & Dryja, T. (2015). *The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments*. Retrieved from https://www.weusecoins.com/assets/pdf/library/Lightning%20Network%20Whitepaper.pdf.

"Prevent Plagiarism in Published Work." (2018). Retrieved June 20, 2018, from http://www.ithenticate.com/

Pusztai, L., Hatzis, C., & Andre, F. (2013). *Reproducibility of research and preclinical validation: Problems and solutions*. Nature Reviews Clinical Oncology, 10(12), 720-724. doi:10.1038/nrclinonc.2013.171

Resnik, D. B. (2013). *Data Fabrication and Falsification and Empiricist Philosophy of Science*. Science and Engineering Ethics, 20(2), 423-431. doi:10.1007/s11948-013-9466-z

Rowley, J., Slack, F. (2004) *"Conducting a literature review"*. Management Research News, Vol. 27 Issue: 6, pp.31-39, https://doi.org/10.1108/01409170410784185

Saunders, M. N., Lewis, P., & Thornhill, A. (2016). *Research methods for business students*. Harlow, Essex, England: Pearson Education Limited.

Saunders, M. N., Lewis, P., & Thornhill, A. (2016). *Research methods for business students* (174). Harlow, Essex, England: Pearson Education Limited.

Saunders, M. N., Lewis, P., & Thornhill, A. (2016). *Research methods for business students* (202). Harlow, Essex, England: Pearson Education Limited.

Saunders, M. N., Lewis, P., & Thornhill, A. (2016). *Research methods for business students* (203). Harlow, Essex, England: Pearson Education Limited.

Shrout, P. E., & Rodgers, J. L. (2018). *Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis*. Annual Review of Psychology, 69(1), 487-510. doi:10.1146/annurev-psych-122216-011845

Simmons, J., Nelson, L., & Simonsohn, U. (2011). *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*. PsycEXTRA Dataset. doi:10.1037/e519702015-014

Simmons, J., Nelson, L., & Simonsohn, U. (2011). *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant* (1365). PsycEXTRA Dataset. doi:10.1037/e519702015-014

Sklaroff, Jeremy M. (2018). *Smart Contracts and the Cost of Inflexibility.* Prize Winning Papers. 9. https://scholarship.law.upenn.edu/prize_papers/9

Slade, E., Drysdale, H., & Goldacre, B. (2015). *Discrepancies Between Prespecified and Reported Outcomes*. Annals of Internal Medicine, 164(5), 374. doi:10.7326/l15-0615

Smith, M. A., & Houghton, P. (2013). *A Proposal Regarding Reporting of In Vitro Testing Results*. Clinical Cancer Research, 19(11), 2828-2833. doi:10.1158/1078-0432.ccr-13-0043

Snow, P., Deery, B., Lu, J., Johnston, D., & Kirby, P. (2014). *Business Processes Secured by Immutable Audit Trails on the Blockchain*. Factom. Retrieved from www.factom.org.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. Perspectives on Psychological Science,7(6), 670-688. doi:10.1177/1745691612460687

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). *Data quality in context*. Communications of the ACM, 40(5), 103-110. doi:10.1145/253769.253804

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). *Data quality in context* (104). Communications of the ACM, 40(5), 103-110. doi:10.1145/253769.253804

Sugden, L. A., Tackett, M. R., Savva, Y. A., Thompson, W. A., & Lawrence, C. E. (2013). *Assessing the validity and reproducibility of genome-scale predictions*. Bioinformatics, 29(22), 2844-2851. doi:10.1093/bioinformatics/btt508

Tranfield, D. , Denyer, D. & Smart, P. (2003), *Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review*. British Journal of Management, 14: 207-222. doi:10.1111/1467-8551.00375

Tranfield, D. , Denyer, D. and Smart, P. (2003), *Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review* (214). British Journal of Management, 14: 207-222. doi:10.1111/1467-8551.00375

Vaux, D. L. (2013). *Know when your numbers are significant*. The Forestry Chronicle, 89(01), 7-9. doi:10.5558/tfc2013-003

Watson, M. (2017, November 22). *Similarity Check - Crossref*. Retrieved June 20, 2018, from https://www.crossref.org/services/similarity-check/

Wang, M., Wu, Q., Qin, B., Wang, Q., Liu, J., & Guan, Z. (2018). *Lightweight and Manageable Digital Evidence Preservation System on Bitcoin*. Journal of Computer Science and Technology, 33(3), 568-586. doi:10.1007/s11390-018-1841-4

Webster, J., & Watson, R. (2002). *Analyzing the Past to Prepare for the Future: Writing a Literature Review*. MIS Quarterly, 26(2), Xiii-Xxiii. Retrieved from http://www.jstor.org/stable/4132319

Winquist, R. J., Mullane, K., & Williams, M. (2014). *The fall and rise of pharmacology – (Re-)defining the discipline?* Biochemical Pharmacology, 87(1), 4-24. doi:10.1016/j.bcp.2013.09.011

Yeow, K., Gani, A., Ahmad, R. W., Rodrigues, J. J., & Ko, K. (2018). *Decentralized Consensus for Edge-Centric Internet of Things: A Review, Taxonomy, and Research Issues*. IEEE Access, 6, 1513-1524. doi:10.1109/access.2017.2779263

Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). *Where Is Current Research on Blockchain Technology?—*A Systematic Review. Plos One, 11(10). doi:10.1371/journal.pone.0163477

Zheng, B., Zhu, L., Shen, M., Gao, F., Zhang, C., Li, Y., & Yang, J. (2018). *Scalable and Privacy-Preserving Data Sharing Based on Blockchain*. Journal of Computer Science and Technology, 33(3), 557-567. doi:10.1007/s11390-018-1840-5

Zheng, B., Zhu, L., Shen, M., Gao, F., Zhang, C., Li, Y., & Yang, J. (2018). *Scalable and Privacy-Preserving Data Sharing Based on Blockchain* (558). Journal of Computer Science and Technology, 33(3), 557-567. doi:10.1007/s11390-018-1840-5

Zheng, B., Zhu, L., Shen, M., Gao, F., Zhang, C., Li, Y., & Yang, J. (2018). *Scalable and Privacy-Preserving Data Sharing Based on Blockchain* (559). Journal of Computer Science and Technology, 33(3), 557-567. doi:10.1007/s11390-018-1840-5

Appendices

1: Tables

Table 2: Articles from keyword searches

Authors	Article name	Published
Bartling, Sönke & et contributors to living document	Blockchain for Open Science and Knowledge Creation.	2017
Begley & Ioannidis	Reproducibility in Science: Improving the Standard for Basic and Preclinical Research.	2014
Beck, Czepluch, Lollike & Malone	Blockchain – The Gateway to Trust - Free Cryptographic Transactions.	2016
Blasius & Thiessen	Should we trust survey data? Assessing response simplification and data fabrication.	2015
Earp & Trafimow	Replication, falsification, and the crisis of confidence in social psychology.	2015
Fanelli	How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data	2009
Irving & Holden	How blockchain-timestamped protocols could improve the trustworthiness of medical science.	2016
Kingor & Gerrets	Morals, morale and motivations in data fabrication: Medical research fieldworkers views and practices in two Sub-Saharan African contexts. Social Science & Medicine.	2016
Macleod	Why animal research needs to improve.	2011
Open Science Collaboration.	Estimating the reproducibility of psychological science.	2015
Pusztai, Hatzis & Andre	Reproducibility of research and preclinical validation: Problems and solutions.	2013
Resnik	Data Fabrication and Falsification and Empiricist Philosophy of Science.	2013
Shrout, & Rodgers	Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis.	2018
Simmons, Nelson & Simonsohn	False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.	2011
Stroebe, Postmes, & Spears	Scientific Misconduct and the Myth of Self-Correction in Science.	2012
Vaux	Know when your numbers are significant.	2013
Yli-Huumo, Ko, Choi, Park & Smolander	Where Is Current Research on Blockchain Technology?—A Systematic Review.	2016
Nakamoto	"Bitcoin P2P e-cash paper".	2008

Table 3: literature adopted from courses

Authors	Article/ book name	Published
Strong, Lee & Wang	Data quality in context.	1997
Saunders, Lewis & Thornhill	Research Methods for Business Students.	2016

Table 4: Citation review: *'How blockchain-timestamped protocols could improve the trustworthiness of medical science'*

Authors	Article name	Published
Slade, Drysdale & Goldacre	Discrepancies Between Prespecified and Reported Outcomes.	2015

Table 5: Citation review: *'Scientific Misconduct and the Myth of Self-Correction in Science.'*

Authors	Article name	Published
National Science Foundation	New research misconduct policies.	2001
Bhattacharjee	The Mind of a Con Man	2013

Table 6: Citation review: '*Reproducibility in Science Improving the Standard for Basic and Preclinical Research.'*

Authors	Article name	Published
Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson & Munafò	Power failure: Why small sample size undermines the reliability of neuroscience.	2013
Henderson, Kimmelman, Fergusson, Grimshaw & Hackam	Threats to Validity in the Design and Conduct of Preclinical Efficacy Studies: A Systematic Review of Guidelines for In Vivo Animal Experiments.	2013
Winquist, Mullane & Williams	The fall and rise of pharmacology – (Re-)defining the discipline?	2014
Marino	The use and misuse of statistical methodologies in pharmacology research.	2014
Sugden, Tackett, Savva, Thompson & Lawrence	Assessing the validity and reproducibility of genome-scale predictions.	2013
Plant & Parker	Translating Stem Cell Research from the Bench to the Clinic: A Need for Better Quality Data.	2013
Smith & Houghton	A Proposal Regarding Reporting of In Vitro Testing Results. Clinical Cancer Research	2013
Nekrutenko & Taylor	Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility.	2012
Kepes & Mcdaniel	How Trustworthy Is the Scientific Literature in Industrial and Organizational Psychology?	2013

Table 8: Articles acquired through search 1: (Blockchain AND data)

Authors	Article name	Published
Benchoufi & Ravaud	Blockchain technology for improving clinical research quality.	2017
Engelhardt	Hitching Healthcare to the Chain: An Introduction to Blockchain Technology in the Healthcare Sector.	2017
Gaetani, Aniello, Baldoni, Lombardi, Margheri & Sassone	Blockchain-based Database to Ensure Data Integrity in Cloud Computing Environments.	2017
Kogure, Kamakura, Shima & Kubo	Blockchain Technology for Next Generation.	2017
Lee & Yang	Fingernail analysis management system using microscopy sensor and blockchain technology.	2018
Lin, Petway, Anthony, Mukhtar, Liao, Chou & Ho	Blockchain: The Evolutionary Next Step for ICT E-Agriculture.	2017
Wang, Wu, Qin, Wang, Liu & Guan	Lightweight and Manageable Digital Evidence Preservation System on Bitcoin.	2018
Zheng, Zhu, Shen, Gao, Zhang, Li & Yang	Scalable and Privacy-Preserving Data Sharing Based on Blockchain.	2018

Table 9: Citation review: '*Lightweight and Manageable Digital Evidence Preservation System on Bitcoin.'*

Authors	Article name	Published
Gipp, Meuschke & Gernandt	Decentralized Trusted Timestamping using the Crypto Currency Bitcoin.	2015
Poon & Dryja	The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments.	2015
Snow, Deery, Lu, Johnston & Kirby	Business Processes Secured by Immutable Audit Trails on the Blockchain.	2014

Table 10: Citation review: '*Blockchain technology for improving clinical research quality.'*

Authors	Article name	Published
Ioannidis	Why Most Published Research Findings Are False: Authors Reply to Goodman and Greenland	2007

Table 11: Articles acquired through search 2: (Blockchain AND (Attributes OR characteristics OR features))

Authors	Article name	Published
Hou, Wang & Liu	Applying the blockchain technology to promote the development of distributed photovoltaic in China.	2018
Kshetri	Blockchains roles in strengthening cybersecurity and protecting privacy.	2017
Kuo, Kim & Ohno-Machado	Blockchain distributed ledger technologies for biomedical and health care applications.	2017
Sklaroff	Smart Contracts and the Cost of Inflexibility.	2018
Yeow, Gani, Ahmad, Rodrigues & Ko	Decentralized Consensus for Edge-Centric Internet of Things: A Review, Taxonomy, and Research Issues.	2018

Table 12: Article influencer scores of journals used in this study

Journal name	Article Influencer (AI) percentile
Annals of Internal Medicine	99
Annual Review of Psychology	100
Biochemical Pharmacology	87
Bioinformatics	97
British Journal of Management	77
Circulation Research	98
Clinical Cancer Research	97
Communications of the ACM	95
Environments	No score available
F1000Research	No score available
Frontiers in Psychology	80
IEEE Access	67
Industrial and Organizational Psychology	33
International Journal of Distributed Sensor Networks	19
International Journal of Energy Research	48
Journal of Computer Science and Technology	21
Journal of the American Medical Informatics Association	86
Nature	100
Nature Reviews Clinical Oncology	99
Nature Reviews Genetics	100
Nature Reviews Neuroscience	100
Perspectives on Psychological Science	99
PLoS Medicine	99
PLoS ONE	83
Science	100
Science and Engineering Ethics	40
Social Science & Medicine	87
Social Science Research	83
Stem Cells and Development	80
Technology Innovation Management Review	No score available
Telecommunications Policy	50
The Forestry Chronicle	28
Trials	76

Table 13: Characteristics of blockchain

Characteristics of blockchain technology						
Article	Identified characteristics					
	Decentralization	**Timing data**	**Data handling**	**Security & reliability**	**Programmability**	**Consensus mechanism**
Hou et al. (2017, p. 2063, p. 2059, p. 2061)	"Blockchain relies on the various nodes for achieving the maintenance of the system and ensuring the authenticity of information transmission. It stores data on the basis of a distributed structure, so the entire network has no centralized hardware or management agency."		"The 2 sides of the data exchange in the blockchain system can be anonymous. The nodes in the system can exchange data without knowing each other's identity and personal information, so the privacy of each participating node is protected."	"The activities of any node in the system are monitored by the whole network. And the database uses distributed storage, so that each participating nodes can get a complete copy of the database."	"It is a contract that uses a computer language to replace a legal language to record terms. Smart contracts can be executed automatically by a computing system. Potential benefits of smart contracts include reduced contract signing, enforcement, and regulatory costs."	"The idea of blockchain consensus mechanism is as follows: When the block data of a node change and the change is transmitted to all participating nodes, all the nodes should judge whether this change is effective through calculating and processing based on certain rules and mechanisms. The common consensus mechanisms are Proof of Work, Proof of Stake, and Delegate Proof of Stake."
Sklaroff (2017, p. 263)					"'Smartcontracts' are decentralized agreements built in computer code and stored on a blockchain. Proponents imagine a future where commerce takes place exclusively using smartcontracts, avoiding the high costs of contract drafting, judicial intervention, opportunistic behavior, and the inherent ambiguities of written language."	
Yeow et al. (2017, p. 1521, p. 1522)						"Proof of Work (PoW): … To be exact, for any assembled block that qualifies as a subsequently mined block on the network, any node

						must search for the correct nonce (random number) in the block header…. *Proof-of-stake* (PoS) is a cheaper substitute for PoW and requires much less CPU computation in mining."
Kshetri (2017, p. 1027, p. 1036)	"Using practical applications and real-world examples, the paper argues that blockchain's decentralized feature is likely to result in a low susceptibility to manipulation and forgery by malicious participants."		"Among the most promising is that individuals are able to control their own personal data. For instance, after certifiers such as a government agency provide the subject with a digitally signed copy of a document (e.g., driving license), and put it on blockchain, they no longer have access to the data."	"Especially blockchain's decentralized, and consensus driven structures are likely to provide more secure approach when the network size increases exponentially."		
Kuo et al. (2017, p. 1212)						"To solve the double-spending problem, each computation node in the blockchain network not only needs to store every transaction to enable the distributed verification of the transactions, but also to follow a distributed timestamp mechanism to determine which transactions should be accepted and which should be rejected"
Lee & Yang (2018, p. 11, p. 8)	"TICT agricultural systems with blockchain infrastructures are therefore immutable		"When the data is stored on the system, the system uses blockchain technology to protect the privacy and correctness of the data, so	"Bitcoin designs mining to be very difficult, and there is a very important reason in fact: to avoid the arbitrary generation of blocks."	"The first block of any blockchain system (Block 0) is called Genesis Block. It uses Merkle Tree42 when building the	"As for Bitcoin's blockchain system, the block generation process is known as mining. Therefore, if we want to

	and decentralized record management systems."		as to improve users' trust in the system. Through blockchain technology, any modification records of data are able to be tracked."		blockchain. Merkle Tree, also known as Hash Tree, is the binary tree that stores hash values. The value of the Merkle Tree's leaf node is the cell data of the dataset or the hash of the cell data. The value of the non-leaf node is the hash value of all its child nodes. Any block scattered in the system has a number. This number is the order in which the blocks are generated. The way in which the blocks are generated can be designed by the system developers of the blockchain."	achieve a blockchain, we should establish the Genesis Block first. Genesis Block will have its own hash value, and this value is generated by the hash algorithm, so it is also called hash ID. Using the algorithm, it will produce Block 1 through calculation."
Lin et al. (2017, p.9 p. 8)	"ICT's contribution to digital democratization has progressed from trusted closed and centralized networks, to open access centralized cloud computing, and now to blockchain distributed networks that do not require public trust in a centralized authority."	"As the name implies, a blockchain is organized in a linear sequence of smaller encrypted datasets called 'blocks', which contain timestamped batches of transactions."	"When ICT e-agricultural systems with blockchain infrastructures are immutable and decentralized record management systems, baseline agricultural environmental data is safeguarded for farmers, NGOs, stakeholders, consumers, and decision makers who participate in transparent data management."		"Furthermore, Ethereum blockchain technology can auto-execute programmed transactions (i.e., 'smart contracts') that are secure and censorship resistant."	
Wang et al. (2018, p. 568)		"The transparency used for audit, which relates to the proof of	"The anonymity is naturally formed from the cryptographic design, since the cipher evidence under encrypted cryptosystem			

		existence, comes from instant timestamps and irreversible hash functions in mature blockchain network."	and hash-based functions leakages nothing to the public."			
Engelhardt (2017, p. 23, p. 24)	"There is no single source that claims authority over the true data, which is in- stead declared by consensus amongst the multiple parties holding the data."		"Information in each block can be encrypted such that only the holders of the correct cryptographic keys can access the information in it. Blockchains are thus *private."*	"This arrangement protects the data from tampering not just by individual keepers of the blockchain, but also external attempts at damage. In one example, the decentralization of blockchain solutions would offer intrinsic protection against assaults such as the recent WannaCry ransomware attacks because the blockchain would only be affected if simultaneously attacked at many sites."	"Additional rules, often referred to as smart contracts, can be built into these decentralized, immutable, private, and trusted ledgers to regulate how the data can be used. Smart contracts are not a core feature of every blockchain, but are often central to their use in the complex world of healthcare."	
Benchoufi & Ravaud (2017, p. 1, p.2, p.4)	"Blockchain is known to be the technology powering Bitcoin, as an open, distributed public ledger recording all the Bitcoin transactions in a secure and verifiable way, without the need for a third party to process payments."	"This information is publicly transparent; any user owns a copy of the proof of the time-stamped data."	"First, data integrity is ensured by the cryptographic validation of each transaction. This is key to ensuring the sincerity of data — limiting data falsification, data 'beautification' and in some sense data invention."		"Practically, Smart Contracts enable the validation of a step with the only condition that every preceding step has been fully validated."	

Zheng et al. (2018, p. 557, p. 558, p. 559)	"The validation, bookkeeping, storage, maintenance, and transmission of blockchain data are based on a distributed system structure. The blockchain uses pure mathematical methods instead of central institutions to establish trust relations among distributed nodes, thus forming a decentralized and trustworthy distributed system."	"Blockchain stores data with a time-stamped block structure. Thus, it adds time dimension to data, and has extremely strong verifiability and traceability."		"Blockchain adopts asymmetric cryptography principle to encrypt data. The consensus algorithm forms a computing power to resist external attack, and prevent the blockchain data from being tampered. Thus, blockchain has higher security."	"Blockchain can provide flexible script code system, and support users to create advanced smart contracts, currencies or other decentralized applications. For example, Ethereum that provides Turing-complete script language for the user to build any smart contract or transaction type can be precisely defined."	"The blockchain system adopts specific economic incentive mechanism to ensure that all nodes in the distributed system can participate in the verification process of data blocks (such as bitcoin mining process). The new block is added to blockchain through the consensus algorithm."
Kogure et al. (2017, p. 55, p. 58)			"To realize privacy protection in a blockchain, encryption can be used to maintain confidentiality when information is made accessible to certain concerned individuals."	"To avoid the duplicate payment problem, a transaction is broadcast through a peer to peer (P2P) Bitcoin network, and its validity is checked by all the partici-pants in the network."		"New blocks are created and managed by 'miners.' A miner who constructs a block with a hash value below a certain threshold receives bitcoin as an incentive. The miners 'race' to construct blocks."

Table 14: Challenges of blockchain

Challenges of blockchain				
Article	Challenge			
	Slower	**Scalability**	**Security**	**Limitations of private keys**
Lin et al. (2017, p. 8)	"Currently, blockchain networks will always be slower than centralized databases because of three i.e., 'security through transparency' additional processes per transaction it undertakes: cryptographic verification, consensus mechanisms and redundancy"	"Scalability remains the all-inclusive and fundamental problem within a Nakamoto-consensus blockchain network. While it is generally understood that as a network gains users, the network itself gains value, the issue of scaling is unresolved. Blockchain protocols face serious scalability obstacles since transaction processing rates are limited by block size and block interval such that larger block sizes improve throughput while slowing propagation within the network and, minimizing block intervals reduces latency while degrading system stability with increased branching."	"At the core of the problem of reducing blocktime is the issue of security since the faster the blocktime, the more centralization of processing is needed."	
Lee and Yang (2018, p. 3)	"…long delays in transaction confirmation…"	"…broadcast transactions of low scalability…"		
Engelhardt (2017, p. 31)				"If a key is lost, then the data it accesses becomes irretrievable. Losing access to a lifetime of health information through the loss of one of these keys is unacceptable, and solutions will have to be implemented to reconnect users with their data. "
Kogure et al. (2017, p. 58)	"A blockchain transaction takes more time to process than a conventional transaction as validation by all network participants is required. Therefore, for those applications involving a great number of transactions, improving		"Blockchain technology is characterized by ensuring security through verification of the transaction information by all network participants, thus preventing illegal activity. This enables an individual outside the blockchain to access a transaction executed within the blockchain. In some cases for some applications, transaction information is not accessible to an individual unless the individual is directly	"If an individual loses the key, that individual cannot decrypt the data. Furthermore, if the key is stolen, the information is exposed to the never ending risk of decryption because the data in a blockchain cannot be deleted or modified."

	processing speed is a challenge."		involved in the transaction. A significant challenge relates to the method used for guaranteeing the confidentiality of a transaction by using a blockchain."	
Gaetani et al. (2017, p. 4)	"However, PoW-based blockchains have a main drawback: performance. This lack of performance is mainly due to the broadcasting latency of blocks on the network and the time-intensive task of PoW."		"…hence it is practically non-repudiable and persistent (unless an attacker has the majority of miners' hash power that are able to create a fork of the chain)."	
Poon and Dryja (2015, p. 1)		"The payment network Visa achieved 47,000 peak transactions per second (tps) on its network during the 2013 holidays, and currently averages hundreds of millions per day. Currently, Bitcoin supports less than 7 trans- actions per second with a 1 megabyte block limit."		
Yeow et al. (2017, p. 1517, p. 1516)		"Adopting an increase in either block size or block creation rate to boost the throughput will render Nakamoto's primary assurance obsolete, thereby un-honest nodes would only need less than 50 percent of the computational power to launch attacks on the system."	"Economies of scale in PoW have therefore caused mining power to fall to fewer individu- als than originally intended. ... Nonetheless, this centralization risk is mitigated by the argument that such attack will not be in line with miners' long-term economic interests."	
Kuo et al. (2017, p. 1217)		"The transaction time of blockchain can be long, depending on the proto- col, and such a speed constraint may limit the scalability of blockchain-based applications."	"The last challenge is the threat of a 51% attack. A blockchain network may suffer from the "51% attack," which happens when there are fewer honest nodes than malicious ones in the whole network, and thus the whole network is taken over by the malicious attackers"	"Also, even if a user is "anonymized" by using hash values as addresses, the user may still be reidentified through inspection and analysis of the publicly available transaction information on the blockchain network, and therefore the block- chain network only provides "pseudonymity."

Table 15: Benefits of blockchain

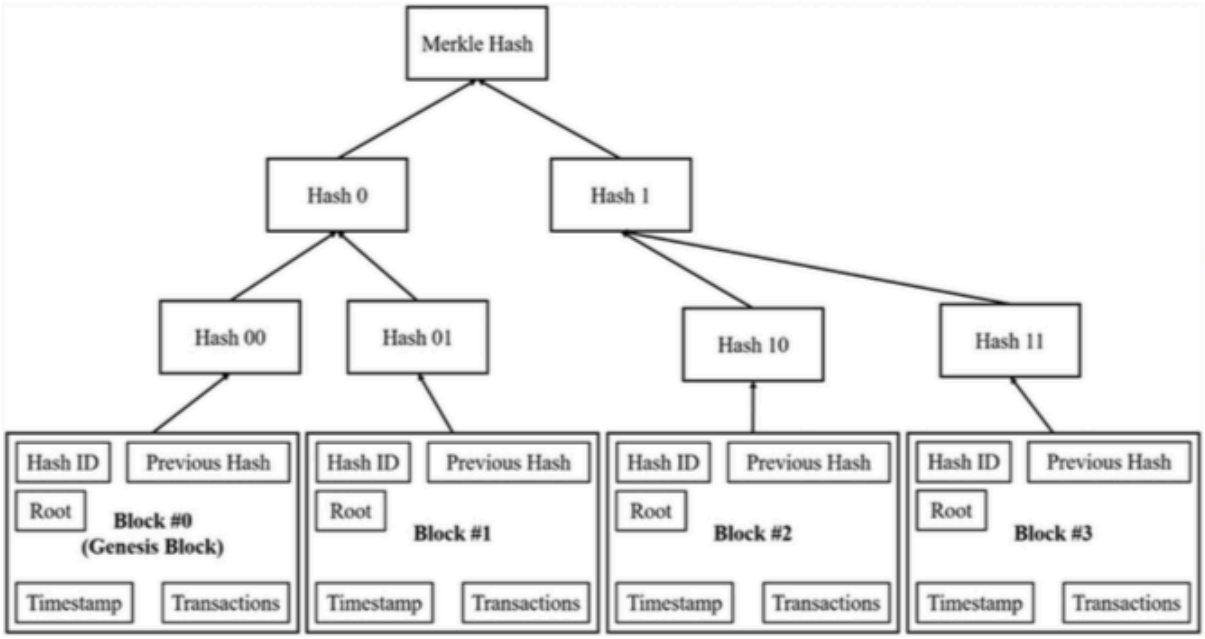
Benefits of blockchain				
Article	Benefits			
	Decentralized management	**Cost-efficiency**	**Immutability**	**Improved security & privacy**
Lin et al. (2017, p. 9)	"…and now to blockchain distributed networks that do not require public trust in a centralized authority."	"Blockchain competitiveness and cost-effectiveness are likely to increase following three laws: (1) Moore's law, i.e., time required for data processing halves every 18 months; (2) Kryder's Law, i.e., data storage halves every year; and (3) Nielsen's Law, bandwidth doubles every two years."	"ICT e-agricultural systems with blockchain infrastructures are immutable…"	
(Lee & Yang, 2018, p. 1; p. 11)			"…in which any change or manipulation can be recorded and tracked, and the data security is improved. "	*"When the data is stored on the system, the system uses blockchain technology to protect the privacy and correctness of the data, so as to improve users' trust in the system. Through blockchain technology, any modification records of data are able to be tracked."*
Engelhardt (2017, p. 23; p. 22; p. 24; p. 24)	"There is no single source that claims authority over the true data, which is instead declared by consensus amongst the multiple parties holding the data. Because of this, blockchains are referred to as *decentralized."*	"Blockchains, also called distributed ledgers, enable a combination of cost reduction and increased accessibility to information by connecting stakeholders directly without requirements for third-party brokers, potentially giving better results at lower costs."	"Records can only be added, never removed, and only by consensus of the maintainers of the distributed copies. Blockchains are thus *immutable."*	"Information in each block can be encrypted such that only the holders of the correct cryptographic keys can access the information in it. Blockchains are thus *private."*
Benchoufi and Ravaud (2017, p. 2; p. 1)	"The datastore is owned by no one, is controlled by users and is not ruled by any trusted third party or central regulatory instance. In fact, trust is encoded in the protocol and maintained by the community of users."		"Regarding inviolability and historicity of data, it follows that Blockchain ensures that events are tracked in their correct chronological order, which largely prevents a posteriori reconstruction analysis."	"Blockchain technology can be considered a basis for improved clinical research methodology and a step toward better transparency to improve trust within research communities and between research and patient communities."

Zheng et al. (2017, p. 558; p. 557)	"The blockchain uses pure mathematical methods instead of central institutions to establish trust relations among distributed nodes, thus forming a decentralized and trustworthy distributed system."		"We use blockchain to prevent the shared data from being tampered, and use the Paillier cryptosystem to realize the confidentiality of the shared data."	"Blockchain adopts asymmetric cryptography principle to encrypt data. The consensus algorithm forms a computing power to resist external attack, and prevent the blockchain data from being tampered. Thus, blockchain has higher security."
Gaetani et al. (2017, p. 4)			"PoW-based blockchains enjoy many fascinating properties related to data integrity, which follow from the mining process and from the full replication of the blockchain on a large number of nodes."	
Gipp et al. (2015, p. 4; p. 1)	"The service presented is non-commercial and uses decentralized trusted timestamping enabled by the distributed and cryptographically validated block chain of the digital currency Bitcoin."		"Users can then retrieve and verify the timestamps that have been committed to the block chain. The non-commercial service enables anyone, e.g., researchers, authors, journalists, students, or artists, to prove that they were in possession of certain information at a given point in time."	
Hou et al. (2017, p. 2063, p. 2059	"Blockchain technology is called trustless in the sense of not needing to trust the counterparty but instead trusting the blockchain software system."	"Moreover, with the characteristics of point-to-point transactions and programmability, blockchain can greatly reduce the transaction costs and improve the transaction efficiency, avoid cumbersome process of centralized liquidation and delivery, achieve convenient and efficient transactions of distributed PV power, and solve the problem that the recovery cycle of the cost for distributed PV power is long."	"The second key benefit is the immutable audit trail. DDBMSs support create, read, update, and delete functions like all database systems, while blockchain only supports create and read functions."	"The rules of the entire blockchain system are open and transparent, and all the data content is also open to the public. In addition, since all nodes in the system can act as "supervisors," there is no need to worry about the fraud issue."
Kshetri (2017, p. 1036)	"Some of the key security challenges associated with the cloud can be addressed by using the decentralized, autonomous,			"Among the most promising is that individuals are able to control their own personal data. For instance, after certifiers such as a government agency provide the subject with a digitally signed copy of a document (e.g., driving license), and put it on

	and trustless capabilities of blockchain."			blockchain, they no longer have access to the data."
Kuo et al. (2017, p. 1211, p. 1212, p. 1214)	"It should be noted that the central intermediary is not desired because it creates a single-point-of-failure…" & "The first key benefit of blockchain is decentralized management. DDBMSs are logically centralized-managed (ie, users logically feel they are operating a centralized database, but the underlying machines can be physically distributed), while blockchain is a peer- to-peer, decentralized database management system."		"An additional benefit of the proof-of-work consensus protocol used in blockchain is the ability to resolve disagreement of the chains, and thus let blockchains be immutable audit trails."	"The final key benefit of blockchain is related to the improved security and privacy using cryptographic algorithms."

2. Figures

Figure 1: Merkle tree



Note: Reproduced from "Fingernail analysis management system using microscopy sensor and blockchain technology.", by Lee, S. H., & Yang, C. S., (2018 International Journal of Distributed Sensor Networks, 14(3), p. 8

3. Epilogue Reflection

First of all, I want to thank the reader for taking the time to read this paper to its full extend. This epilogue will reflect back on the process I went through while conducting this thesis project. I also want to thank Prof. Dr. Sascha Friesike for providing guidance during this process. During the feedback sessions we had planned you have provided me with valuable insights in the field which have helped me a lot.

I have always been an enthusiast of new innovative technologies. This made me try to build my own websites when I was only nine years old and assemble personal computers from scratch with my dad not much later. The concept of blockchain intrigued me from the start. Although, it took some time for me to understand the full potential of the technology.

Writing a literature review was not my first choice. I preferred finding a topic and/ or a dataset within *'blockchain for science'* that allowed me to perform a quantitative empirical study. This has been the goal the first month of the research project. Finding this was harder than I expected and after a while I decided to write a literature review, which is something I was not looking forward to due to my slow reading speed and dyslexia. The concept of a systematic literature review was still unfamiliar to me and it required some studying to get familiar with this. Getting familiar with it could be described as a process. During this process there were some moments where I felt I already had a full understanding of the subject. Because of this, the literature search on the replication crisis, although well documented, was not completely systematic. I did not have explicit inclusion and exclusion criteria yet and for some more generic searches I did not review all results but only the first five pages of the search engine results. Hence, the decision to examine the replication crisis by means of traditional review. I believe writing the systematic review has provided me with a lot of new insights about knowledge creation and research philosophy. This might also be due to the literature search on scientific misconduct and the replication crisis, but it helped me to understand the importance of a complete and clear methodology.

Although I am very pleased with the result of this study. I believe the order in which the literature review has been conducted could have been better. In this study I narrowed my application of the technology down to an application aiming to solve the replication problem. I think that studying the technology itself first would have been a better choice. This would allow me to choose an application of the technology based on a full understanding of the characteristics of the technology.