UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics
and Computer Science (EEMCS)

NeoBank

Data Driven Banking:

Applying Big Data to accurately determine consumer creditworthiness

Author: Shen Yi Man

Final MSc Thesis
Business Information Technology
Track: IT Management & Innovation
September 2016

Supervisors University of Twente:

ir. K. Sikkel.

dr. H.C. van Beusichem

External Supervisor:

Anonymous

30-9-2016

Document Title:	Data Driven Banking: Applying Big Data to accurately determine consumer creditworthiness
Date:	30-9-2016
Author:	Y. Man (Shen Yi) y.man@alumnus.utwente.nl s1128337
Educational Institution:	University of Twente *The Netherlands*
Faculty:	Faculty of Electrical Engineering, Mathematics and Computer Science
Department:	Industrial Engineering and Business Information Systems
Educational Program:	MSc. Business Information Technology Specialization: IT Management & Innovation

Graduation Committee

Ir. Klaas Sikkel
Faculty of Electrical Engineering, Mathematics and Computer Science
Dpt. of Industrial Engineering and Business Information Systems
University of Twente, Zilverling 4102
k.sikkel@utwente.nl

UNIVERSITY OF TWENTE.

Dr.ing. Henry van Beusichem
Faculty of Behavioural, Management and Social Sciences
Dpt. of Finance & Accounting
University of Twente, Ravelijn 2315
h.c.vanbeusichem@utwente.nl

Anonymous
Project Manager
NeoBank The Netherlands
Consumptive Finance
anonymous@neobank.nl

NeoBank

N.E.O. Bank and its employees, divisions and products are fictional entities to replace a large bank in the Netherlands who shall remain anonymous in the continuance of this thesis.

Preface

While writing this foreword I am finally realizing that after six years of studying at the university, my academic career is coming to a close. It was an incredibly enjoyable period in which I grew a lot and learned that there is much more to life than I could have imagined beforehand. With my internship at NeoBank coming to an end, this concludes another chapter of my life. One that I will always remember with a smile on my face.

This master thesis is the end result of the past six months I spent at NeoBank for my graduation project. Obstacles appeared in the course of this graduation project which I found quite challenging at times, but I'm glad to have done it altogether. It offered me the chance to learn a lot on the subjects of Big Data analytics, credit scoring and the banking sector. My master in Business Information Technology at the University of Twente will officially be completed at approval of this thesis. It also marks the beginning of my professional career which will start after a short vacation in Asia.

I would like to express my gratitude to all the people that supported me during this journey and that helped me finalize this last project. I'd like to thank my friends and family for all their relentless nagging, support and understanding. It provided me with motivation and inspiration at difficult times. In particular, I'd like to thank all my supervisors for their input, feedback and wisdom. Thank you Klaas, for assuming a leading role and helping me through each step of the way to this final product. My thanks to you Henk, for helping me in the initial stages of the thesis. Thank you Henry, for jumping in on such a short notice and helping me put the last hand to the project. I'd like to thank Anonymous for the support and guidance during this pleasant period at Neo. Whenever I was stuck in a frame of thought, I could discuss things with you over the phone or in person. A special thanks to all the great colleagues at NeoBank that provided me with the help and information needed to complete my thesis. I do hope that this report will be of good use and interesting to read.

Enjoy!

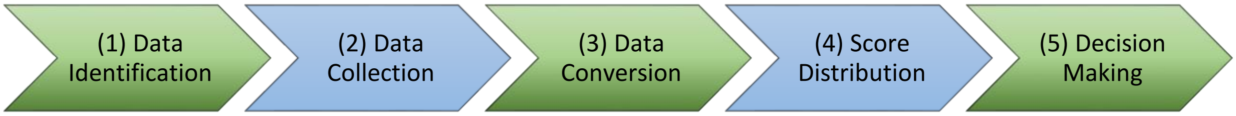
Shen Yi Man

Nijmegen, September 2016

Executive Summary

Financial institutions judge consumer creditworthiness on frequent basis. Errors and inaccuracies in this process cause an increased value of outstanding loans which will not be recuperated by banks due to default. Failing to comply with payment obligation can mark consumers for years, lowering their consumer creditworthiness and making it even more difficult for them to obtain a future loan. These are concerns of both authorities and banks when designing a financial product and its application process.

To solve these problems accompanied by structural consumer debt, we turn to Big Data analytics. Conventional credit scoring methods at traditional banks are becoming less relevant in today's age of massive data generation. Millennials are well-connected and more digitized than ever before. This leads to new possibilities when looking at the contents of new data and the applications that are possible with thorough analysis of great representative quantities. In specific, Machine Learning can be used to greatly improve three of the current five steps in which a credit scoring process is structured.



Within Data Identification, new relevant data variables can be detected and used as proxies to measure the two components of creditworthiness: the ability to repay and the willingness to repay. These data variables can be used to enhance a credit scoring or risk model which is used in Data Conversion to compute a consumer credit score. The model is used to improve the credit scoring process on a list of conditions (Subsection 3.1.1). It is also possible in this step to let the model build and enhance itself through Machine Learning algorithms. The last step Decision Making can be improved by creating a proprietary automatic decision making algorithm through Machine Learning which will streamline the underwriting process. After an initial model is created by using the training data, it has to be validated and tested to measure its performance. The validation step is used to enhance and calibrate the model before it is practically tested or completely discarded. The accuracy is determined by taking historical data sets in which the result is known and comparing these true numbers with results generated by the new model.

	Predicted Result		
Actual Result		No Default	Default
	No Default	True Positive (n)	False Negative (n)
	Default	False Positive (n)	True Negative (n)

During the data identification process, new credit models are built for testing which make use of newly discovered data variables. These new variables need to be validated through the *Six-Point FICO Test* before integration and testing within the model. These points are the following.

1. **Regulatory compliance** – All data sources and data variables must comply with the legislation.
2. **Depth of information** – This factor covers the detail and context of data variables. The richer the data, the more accurate the score will be when computed. High quality data must be acquirable.
3. **Scope and consistency of coverage** – To be relevant, the data source must cover a large percentage of the population. Format consistency is for operating, analyzing and storing the data.
4. **Accuracy** – The incoming data must be validated and tested on basis of historical data.
5. **Predictiveness** – To add value to credit risk models, data variables must be proven predictive towards consumer repayment behavior. This can be tested in practice through Machine Learning.
6. **Additive Value (Orthogonality)** – Data must be uniquely additive and not "double counted".

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

The Big Data maturity had been qualitatively measured in order to gauge the possibilities in implementing Machine Learning at NeoBank. The assessment shows that the technical requirements are currently easily fulfilled to start implementation. However, if the demand continues to grow of Big Data oriented projects, the department will soon be short on hands to be able to capitalize on each opportunity. This research forms the basis of a recommendations plan (Section 6.2) to improve the Big Data maturity of the whole of NeoBank based on the TDWI Big Data maturity model assessment.

EVALUATED PARTIES	DIMENSIONS					TOTAL
	Organization	Infrastructure	Data Management	Analytics	Governance	Big Data Maturity
NeoBank Company-wide	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
DDA Department	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

In the High Level Solutions which are provided additionally, strategic designs are explained which make use of the technology to gain a competitive advantage towards the market. These solutions intersect with different interests of various stakeholders and qualitative criteria. Extended drafts have been made to describe the scenario in which these HLSs could be implemented. The solutions had been ordered in level of disruptiveness. The first solutions in Personalization, Automatic Client Appraisal and Budget Counseling are more feasible to occur in the future according to market research than the last HLS: the IOIS platform. This is largely due to the high dependency on joint funding and the collaborative research effort of inter-organizational information systems. This risky endeavor would require extensive funding and commitment of all parties involved. Most banks would rather depend on their own internal system.

Table of Contents

List of Figures

List of Abbreviations

AFM	**A**utoriteit **F**inanciële **M**arkten (**A**uthority for the **F**inancial **M**arkets)
AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etworks
AWS	**A**mazon **W**eb **S**ervices
BTS	**B**inding **T**echnical **S**tandards
CBR	**C**ase **B**ased **R**easoning
CDR	**C**all **D**etail **R**ecords
CI	**C**ustomer **I**ntelligence
CoE	**C**enter **o**f **E**xcellence
CRR/CRD IV	**C**apital **R**equirements **R**egulation and **D**irective IV
DDA	**D**ata **D**riven **A**nalytics (NeoBank)
DNB	**D**e **N**ederlandsche **B**ank (**D**utch **N**ational **B**ank)
DSRM	**D**esign **S**cience **R**esearch **M**ethodology
EAD	**E**xposure **A**t **D**efault
EBA	**E**uropean **B**anking **A**uthority
FCRA	**F**air **C**redit **R**eporting **A**ct
GRC	**G**overnance, **R**isk management and **C**ompliance
HDFS	**H**adoop **D**istributed **F**ile **S**ystem
HLS	**H**igh **L**evel **S**olution
ID3	**I**terative **D**ichotomiser **3**
IDB	**I**nter-American **D**evelopment **B**ank
IOIS	**I**nter-**O**rganizational **I**nformation **S**ystem
LGD	**L**oss **G**iven **D**efault
LML	**L**ifelong **M**achine **L**earning
LTI	**L**oan-**T**o-**I**ncome
LTV	**L**oan-**T**o-**V**alue
MDA	**M**ulti-**D**iscriminant **A**nalysis
ML	**M**achine **L**earning
NEOFC	**NEO** **F**ast **C**redit
NVB	**N**ederlandse **V**ereniging van **B**anken
P2P	**P**eer-**to**-**P**eer
PD	**P**robability of **D**efault
PFC	**P**aleo **F**ast **C**redit
PMO	**P**rogram **M**anagement **O**ffice
ROC	**R**eceiver **O**perating **C**haracteristic
ROI	**R**eturn **O**n **I**nvestment
SEPA	**S**ingle **E**uropean **P**ayment **A**rea
STP	**S**traight **T**hrough **P**rocessing
SVM	**S**upport **V**ector **M**achines
TDWI	**T**he **D**ata **W**arehouse **I**nstitute
TILA	**T**ruth **I**n **L**ending **A**ct
VfN	**V**ereniging van **f**inancieringsondernemingen in **N**ederland
Wbp	**W**et **b**escherming **p**ersoonsgegevens
Wck	**W**et op het **c**onsumenten**k**rediet
Wft	**W**et op het **f**inancieel **t**oezicht
WSBI	**W**orld **S**avings and Retail **B**anking **I**nstitute

1 Introduction

Latest research has shown that consumer debt is structurally growing worldwide and has a certain correlation with the economic prosperity (Brown, Stein, & Zafar, 2015). Keynesian theory suggest that lending money is beneficial towards the economy as it leads to more expenditure. The increase in expenditure leads to increased production and growing industries. Growing industries lead to increased employment rates and provide a stimulus to the economy. Financial institutions play an essential role in this process as they collect idle savings and redistribute these funds in an uncertain environment. In spite of precautions, some consumers still loan to the extent they structurally cannot pay back the money they are indebted. When an economic crisis occurs, this massive scale occurrence is called an "economic credit bubble". The value of assets deviate from the intrinsic value as the obtained credit of consumers deviates from the actual creditworthiness. Consumers spend money they do not actually own and in the prospect of paying back, fall behind in economic wealth and stay indebted. Detailed and accurate risk assessment is of key importance to prevent this from occurring.

In the Netherlands, financial authorities such as the "Autoriteit Financiële Markten" (AFM), The Dutch Bank (DNB), and the government come into play when the risk of such an unfortunate event grows. These institutions create laws and guidelines that limit the playing field of banks in order to protect consumers. They supervise over all national financial institutions to keep relevant parties in check. The main goal of a financial firm will always be to generate value and earn profit in order to guarantee its existence. However, ethical conduct and a positive impact on the society are also primary goals for a bank. This sometimes results in a conflict of interest between various key stakeholders.

1.1 Problem Statement

In this specific case of consumer credit, NeoBank in the Netherlands released a financial product called "NEO Fast Credit" (NEOFC). This income-based credit was designed as a short-term high interest loan which facilitates small abrupt payments in a convenient manner. The utility of this product lies in the fact that relatively small credit can be borrowed without a hassle for a short period of time in a consumer friendly way. The credit has to be completely paid off every three months, after which a new loan cycle can be started. The application process has been streamlined by implementing an automated superficial income-test without credit scoring model.

Recently, the AFM collided with the "Nederlandse Vereniging van Banken" (NVB) in a dispute to extend the requirements before granting this type of short-term credit to an individual. The AFM argues that this financial product requires a wider client evaluation based on calculations in order to reduce the risk of structural consumer debt. The NVB disagrees, as this type of credit legally does not have to comply with the "Wet consumentenkrediet" (Wck). This law obligates extensive terms of client evaluation for financial products with a lending period of more than three months. Furthermore, extensive evaluation implicates increased transaction and overhead costs to facilitate and process such an income and expenses test. Moreover, it reduces the utility and consumer friendliness of this financial product as an extensive screening in its current form delays the application.

There is a concerning and growing issue of consumer debt which can only be solved by gathering true, accurate and timely data of consumers. Improving risk assessment entails using more information to construct a complete and relevant consumer profile. In this research project, the possibilities are explored in solving the problem of structural consumer debt through the use of Big Data.

The main question is if it is possible to use Big Data to accurately determine the creditworthiness of consumers. In other words, we research the potential of using Big Data to predict bad loans and structural debt beforehand. The last step is to proactively react to such cases with front-end applications and initiatives to reduce consumer debt in an ethical way.

1.2 Research Objectives

Banks are interested in convenient and quick application processes for loans to attract many borrowers and gain performance in a competitive financial environment (Chen & Lin, 2015). In order for them to reduce consumer default rate and increase stability, it is important to improve the accuracy of the loan approval process. This thesis aims to contribute by analyzing the current capacity of Big Data for the purpose of accurately determining creditworthiness. Individuals and enterprises alike generate data which can be used to form a credit score. Aside from momentarily computing a credit score to grant or reject a loan, there is potential in using this data for other front end applications such as financially monitoring and counseling individuals. Concretely speaking, at the end of this research we hope to offer NeoBank multiple High Level Solutions across a spectrum of different levels of Big Data use. These choices allow NeoBank to solve the problem at hand and grant credit in an ethically responsible way. This morality to banking, or rather moral authority, ensures credit is only granted to the creditworthy and not to individuals who might hurt their own economic position with it (Polillo, 2011). Exclusion and boundaries must be set; this thesis researches if Big Data is suitable for these operations.

1.3 Research Questions

In this section a few formal research questions are formulated to address the problem context mentioned in the introduction. By answering these knowledge questions, steps are made to come closer to concrete designs with a specific purpose in reducing consumer debt and bad loans.

1. **How can Big Data be used to determine the creditworthiness of consumers?**
   a. What type of data is relevant in determining consumer creditworthiness?
   b. How can the accuracy of consumer creditworthiness be determined?
   c. What are the requirements in deploying Big Data to determine client creditworthiness?
   d. How do other lenders determine consumer creditworthiness?
2. What is the Big Data maturity of NeoBank?
   a. Which data are available internally at NeoBank?
3. How does NeoBank currently determine consumer creditworthiness?
   a. What is the structure of traditional credit scoring?
   b. How can the scoring process distinguish between ability and willingness to repay?

The first question formulated is the main research question. This research is dedicated to discovering how Big Data can be used to better establish the creditworthiness of consumers. The other questions aid by elaborating on secondary conditions and current capacities to better indicate limits in the design specifications. When these questions are answered above, the article turns to the concrete design question to solve the problem at hand.

- **How can NeoBank make better use of Big Data in determining consumer creditworthiness?**

Answering this question results in the final design choices which NeoBank can opt to invest in.

1.4 Research Approach

Initially, an academic literature review is conducted to define the terms creditworthiness and Big Data and to establish their components. Consequently, the conditions of an effective credit scoring process are determined. Lastly, the advances in Big Data technology and known applications of Big Data to determine creditworthiness are analyzed. This may be forms of credit scoring, microfinancing or granting loans. Case studies are examined in order to discover methods and challenges behind the computing of credit scores. An overview is formed of the potential of Big Data to determine creditworthiness. Afterwards, technical and organizational requirements are set up for the actual implementation of such applications.

Simultaneously, an internal research is conducted where semi structured interviews are held with experts at NeoBank to map the current Big Data capabilities at the firm. The current credit scoring process is analyzed to discover if there are compatible parts in this process. The gap between ideal maturity and current capabilities is mapped after this part of the analysis. Aside from this, an external qualitative research is conducted at different companies from various markets to discover contemporary advancements on the field of credit scoring based on real-time or historical data. Most of these companies are from the IT or Fintech sector, as innovation is plentiful there. Some traditional organizations are also taken into consideration for comparison. On basis of the market, the potential, the requirements and the current state of Big Data at NeoBank itself, a handful of High Level Solutions is given to satisfy the interests of different stakeholders on various levels. Three major parties of interest are defined; the society, NeoBank and authorities. Design preferences are shown based on each of these stakeholders and a final recommendation is given based on the current situation at NeoBank and the markets.

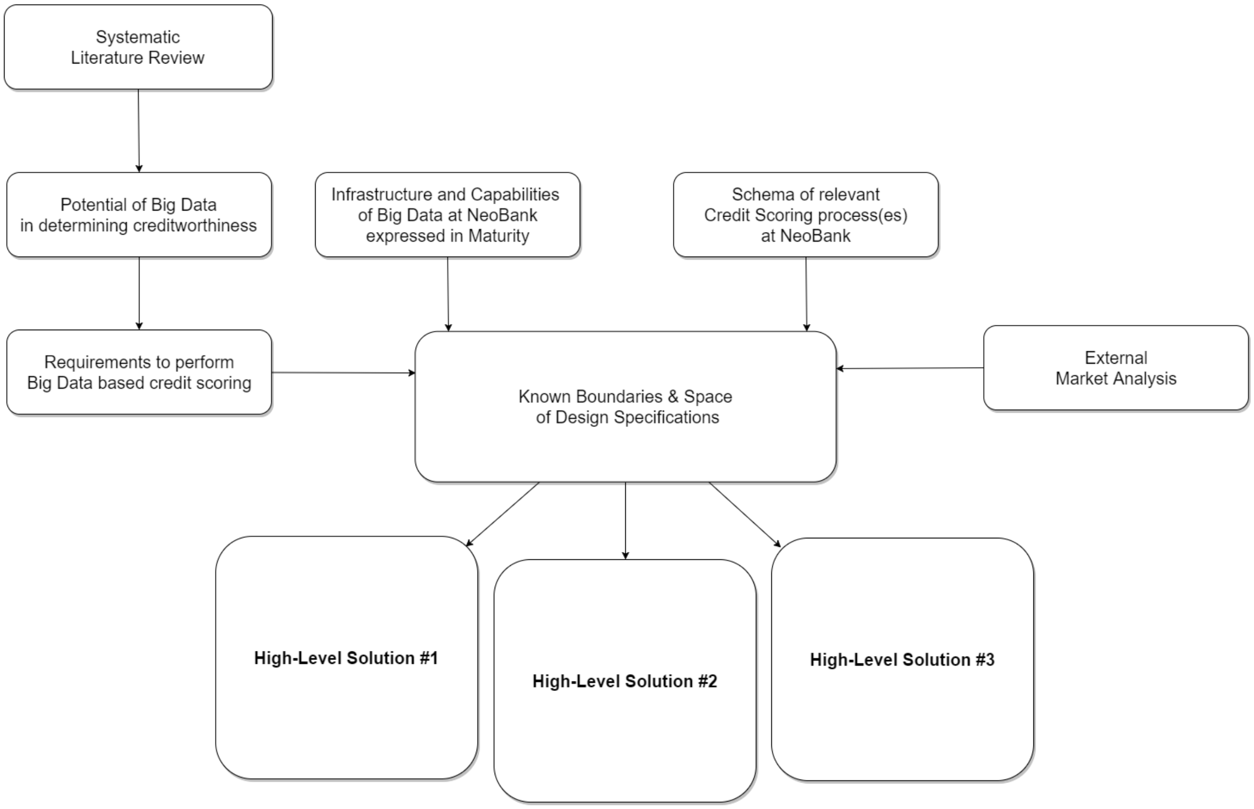


FIGURE 1. INTERMEDIATE RESEARCH RESULTS

1.5 Reading Guide

This first chapter aimed to introduce the problem context and offer some background information on the research project. The following parts of the thesis will discuss the following. The second chapter elaborates on the methodologies used to gather and analyze qualitative data and how solutions were generated. Continuing on this, the third chapter treats the academic literature review and non-academic literature gathered to help answer the research questions. The fourth chapter shows the results through the external and internal research – which consists mostly of models and insights from semi structured interviews. The fifth chapter proposes multiple drafts of High Level Solutions based on the potential, requirements and capabilities. Following this, the final chapter discusses the overall contribution of this thesis and addresses the validation and limitations within the research. Furthermore, it concludes our thesis by offering specific recommendations to NeoBank and suggesting areas with future research possibilities.

Parts of this version of the thesis have been restricted due to confidentiality. For the unrestricted version please contact Shen Yi Man for further discussion on authorization.

2 Methodology

In this chapter the methodologies are explained that are used in this thesis. The various methods are discussed which serve to gather information and to come to deeper qualitative insights. Due to the nature of our topic, a strictly scientific literature review would exclude useful information. Various search engines like Scopus, ScienceDirect, Google Scholar and Web of Knowledge have been deployed to find relevant articles. Various news articles on Fintech companies and whitepapers have been considered in our analysis as well. Furthermore, a number of charts is produced to describe the process in our study.

2.1 Systematic Literature Review

Initially Scopus was used to find relevant scientific articles on determining creditworthiness through Big Data. The search methodology introduced by Wolfswinkel et al. (2011) is used to obtain an initial superset of articles through the use of the following search string:

TITLE-ABS-KEY("**Big Data**" AND ("**Credit worthiness**" OR "**Credit***") AND NOT “**Medic***”) AND (LIMIT-TO(PUBYEAR,2016) OR LIMIT-TO(PUBYEAR,2015) OR LIMIT-TO(PUBYEAR,2014) OR LIMIT-TO(PUBYEAR,2013) OR LIMIT-TO(PUBYEAR,2012)).

This search query generated a result of a mere 87 complying documents in Scopus (June 2016). The group was then filtered through various criteria as relevance, validity and timeliness. Because of the rapid progression in the technology of Big Data, only recent articles from after 2012 were taken into consideration. Through the use of SFX (Full text linking) within Scopus, many of the articles were obtained with the University License. Others were found by using other search engines like ScienceDirect, Google Scholar and Web of Knowledge. The very few unobtainable papers have eventually been excluded of this review. The papers were then further analyzed on relevance of their content to answer our research questions. To conclude, backward and forward citation added a small number of scientific articles. The various steps of the search strategy carried out are illustrated once more in the following image.

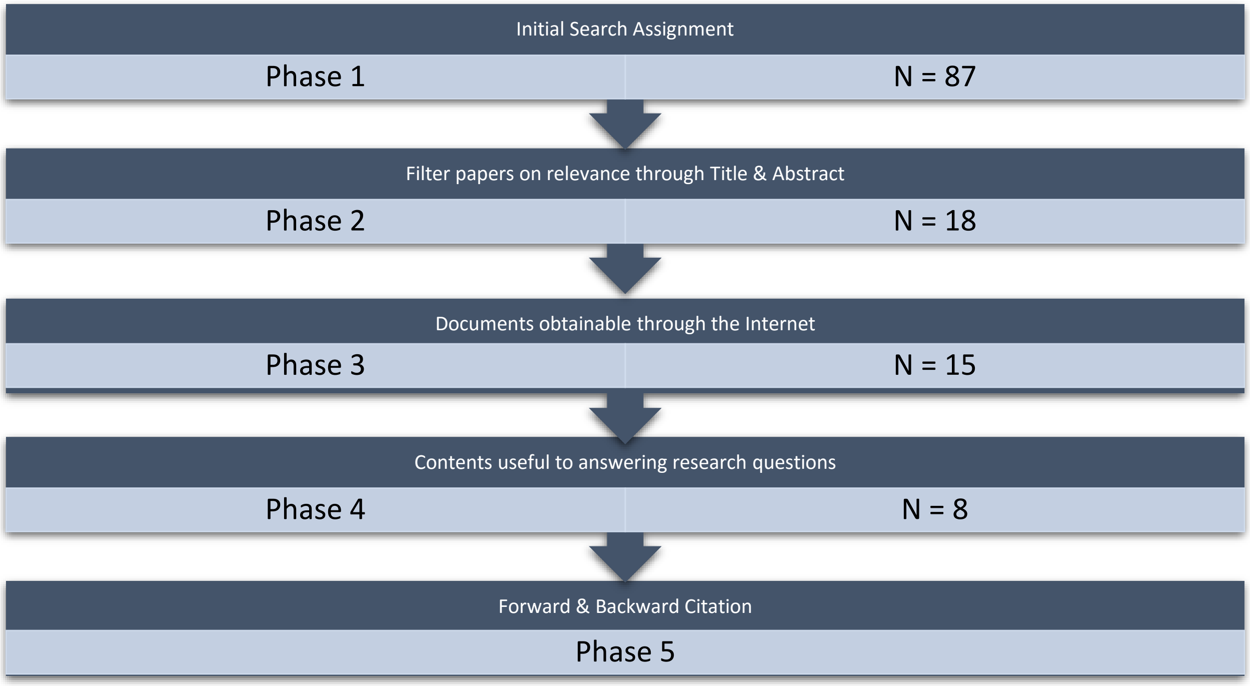

Initial Search Assignment
Phase 1
N = 87
Filter papers on relevance through Title & Abstract
Phase 2
N = 18
Documents obtainable through the Internet
Phase 3
N = 15
Contents useful to answering research questions
Phase 4
N = 8
Forward & Backward Citation
Phase 5

FIGURE 2. FILTERING PROCESS OF LITERATURE REVIEW

2.2 Non-Scientific Literature Review

Because the topic of Big Data and its application to determine creditworthiness is relatively new, other generic search engines like Google have also been deployed. These engines are used with the specific purpose of finding relevant whitepapers of large firms, articles of Fintech start-ups and governmental reports. The validity of these papers is not always ensured, especially in comparison with their academic counterpart. However, these articles are published more frequently and in higher numbers on topics such as Big Data and the practical application of other new technologies. Searches were conducted in the period March to July 2016. Various combinations of the following key words were used: **Big Data**, **Creditworthiness**, **Financial Sector**, **Retail Banking**, **Financial Services**, **Credit Scoring**.

The whitepapers are obtained from multinational enterprises or consultancy bureaus that publish frequently on IT. These organizations publish their reports in order to share knowledge and stimulate further research on certain topics. Other organizations include governmental institutions and financial authorities that have a more conservative and critic view on the implementation of new technology. The treated whitepapers show valuable insights on how financial institutions can deal with the implementation of Big Data in pursuit of various data driven trends. They also elaborate on the constraints and limitations of using these innovative techniques for certain goals.

Many papers and reports on Fintech organizations explain the mechanism behind innovative applications of successful start-ups all over the world. Certain cases also treat the practical results of applying the technology on test groups. Searches are conducted on interesting Fintech companies mentioned in articles, like ZestFinance, Cignifi, Earnest, Credit Karma, Upstart, SoFi and Common Bond among others. These companies are continuously researched to discover more about the technologies used in the financial applications and their business model.

2.3 External Information Acquisition

Attempts to contact external parties were made through the common channels of the organization such as phone contact, company mail and web contact forms. If this proves inefficient, personal contact is made with employees through LinkedIn connections. In this approach, initial contact is made by using a 300-word limited message to briefly explain the background and inquire for a dialogue on the research subject. A second message is sent on the platform to elaborate on the research and to request cooperation from the external party through answering a list of qualitative questions. The dialogue in which the questions are answered can be held through an audio call, video call, chat or mail exchange. The detail and extent to which is answered can vary due to non-disclosure agreements signed. Furthermore, data is anonymized upon request and a version of the end-report will be published on the UT Essays site for them to look into.

The following external market groups have been approached (Full list in Appendix A):

- Fintech organizations (e.g. Cignifi, InVenture)
- Mail Order Credit Companies (e.g. Wehkamp, Lacent)
- Internet giants (e.g. Amazon, ANT Financial)
- Public organs (e.g. Bureau Krediet Registratie)
- Credit rating agencies (e.g. TransUnion, Equifax)
- Consumer data broker (e.g. Acxiom, Datalogix)
- Academic Legal Experts

2.4 Semi-Structured Interviews

A qualitative research was conducted to explore the boundaries and possibilities of NeoBank in determining creditworthiness through the use of Big Data technology. The reason this inquiry approach was taken on internal and external basis is twofold. The goal of the internal interviews is to generate an understanding of the internal situation at NeoBank. This considers the structure of the established order in credit scoring and the organization's current capabilities expressed in Big Data maturity. Externally, the in-depth interviews serve to obtain new insights of the credit scoring market and to establish the conditions and legal boundaries of data-driven applications.

Semi-structured interviews are conducted on basis of the methodology given by Cohen & Crabtree (2006). This method was chosen because qualitative information is required while dealing with a limited number of respondents. Formal interviews are organized in person where possible and a list of mostly open-ended questions (Appendix B) is used in a given order. This is done to set a trajectory of topics for further exploration during the discussion between interviewer and respondent. An open interview is conducted where the interviewee can stray from the initial topic to more interest bearing areas. Detail and depth is essential in gathering qualitative information, therefore follow-up questions are asked frequently. Upon agreement beforehand, the audio is recorded which serves as a transcript to facilitate analysis.

Selection of Respondents

Internal interviews are held with Data Driven Analytics (DDA) members, credit model builders and financial product developers. The respondents are spread throughout departments in the organization. The DDA department is chosen as this is the designated department of Big Data use at NeoBank. The financial product developers are chosen whom are involved with our case of "NEO Fast Credit", as they hold information on the credit scoring process. The credit model builders are chosen as they are knowledgeable about the origin of the credit models. The results serve to generate an overview on the current processes, infrastructure and the organization's compatibility with Big Data in this context. Various types of respondents are chosen for interviews depending on what design aspect is researched.

As mentioned in section 2.3, external companies were approached through a variety of channels to research the current Fintech market. The interviews are held with various companies from different sectors to come to insights on financial products, credit scoring applications and future developments. Further interviews are held externally with legal experts to estimate the legal boundaries in which NeoBank can operate considering data in the Netherlands and the EU.

Operationalization

The list of open-ended questions (Appendix B) was generated on basis of some of the research questions composed in section 1.3. The interviews have a list of specific goal to work towards, the questions and topics served as a roadmap to guide the conversation. However, the respondent had the freedom to change the topic or put emphasis on different parts of the dialogue. This is also done to establish relations in data and discover underlying information during the interview. Time is limited in an interview as most of the respondents reply during worktime. Due to this uneven prioritization, lacking information in certain topics would have to be compensated by (1) Follow-up correspondence or (2) Other interviews.

The interviews as displayed in Appendix B are categorized in three main topics. The main research questions and the interview goals related to the interview questions are illustrated below.

1. Internal Big Data Maturity

Related Research Questions:

- What is the Big Data maturity of NeoBank?
  - Which data are available internally at NeoBank?

Interview Goals:

- The interview questions mainly serve to qualitatively judge criteria of the TDWI Big Data maturity model (Subsection 3.1.5, Appendix C).
- Determine the current internal analytical and computing capacities in terms of Big Data.
- Map the current IT Infrastructure and the support for Big Data analytics.
- Inquire about data management and data governance on local and company-wide level
- Discuss the strategic, tactical and operational internal vision on Big Data.
- Discover bottlenecks due to organizational structure or processes in corresponding dimensions.

---

2. Internal Credit Scoring Processes

Related Research Questions:

- How does NeoBank currently determine consumer creditworthiness?
  - What is the structure of traditional credit scoring?
  - How can the scoring process distinguish between ability and willingness to repay?

Interview Goals:

- Map the steps of NeoBank's relevant credit scoring processes from start till end.
- Acquire background information on how used parameters, coefficients, algorithms and formulas in credit scoring models are established.
- Determine the compatibility of Big Data Analytics with current credit scoring processes.
- Discuss the expanding potential of Big Data usage in financial product design.

---

3. External Market Review

Related Research Questions:

- How do other lenders determine consumer creditworthiness?
  - What type of data is used by third party underwriters?
  - How is the performance established of credit scoring?

Interview Goals:

- Obtain practical information on data scoring processes at third parties.
  - Determine which data variables are representative in computing credit scores.
- Converse about the accuracy in which creditworthiness can and should be determined.
- Converse about the potential and challenges of Big Data on determining creditworthiness
- Discuss the requirements for using Big Data to assess credit scores.
- Discuss the European market and limitations within legislation.

Analysis Methodology

The data generated through conducting semi-structured interviews can be of complex nature. In order to thoroughly analyze the interviews, steps are used from the methodology provided by Burnard (1991). A more practical version is deducted from his fourteen stages of analysis which is more compatible with our current research set-up. The seven steps of the systematic approach are explained as follows.

1. Potential categories are established before and during the interview to separate the collected data structurally under themes. This can also be done with color highlighting.
2. Of each interview, a(n) (audio) transcript is made to analyze the content in detail. Additional notes are made to add context and immerse oneself in the perspective of the interviewee.
3. Open Coding is applied where transcripts are thoughtfully processed and headings are created to include all the content except for filler material the author calls "*dross*". At the end of this step, all the data has been initially categorized.
4. Optional: For the purpose of answering the research questions, all categories created in previous steps are revised and similar categories are assimilated into one. Some relevant categories can be grouped into one broad category. Irrelevant or non-applicable categories can be discarded.
5. Iteration step. The transcripts are thoughtfully read once again to see if the whole content is covered with these categories. Adjustments are made when necessary.
6. A second compact version of the transcriptions is made consisting of all the "coded parts". To ensure that the essence is maintained. The context (question) of the coded text is taken into consideration and additional notes are made to relate data where necessary.
7. Application of all relevant collected data takes place. This can be done by making a matrix illustration, a relevant chart, by writing up information coherently in a chapter or by using the insights to draft a design.

The data between interviews is reflected upon each other when new themes are discovered during the interview, or when coming to insights are after multiple interviews have been held. Inconsistencies are cross-checked with the facts or double questioned afterwards in a follow up through mail or in person.

2.5 Drafting Solutions

In the course of this graduation project, the initial steps in the cycle of the Design Science Research Methodology by Pfeffers et al. (2007) are conducted. This methodology is highly compatible with our research because it offers a scientific approach to a design problem. The goal is to design multiple "artifacts" or High Level Solutions (HLSs) which can solve the underlying problem of growing consumer debt while simultaneously dealing with the current problem context of multiple stakeholders. Considering the data aspect of this problem, there is potentially a substantial improvement in accuracy in which creditworthiness can be determined. The objective is to design an artifact which can reduce the consumer debt by improving the credit scoring and underwriting process while complying with legislation. The further specifics of these artifact designs are given in a chapter five. Several stakeholders have different interests that are satisfied by each of these designs.

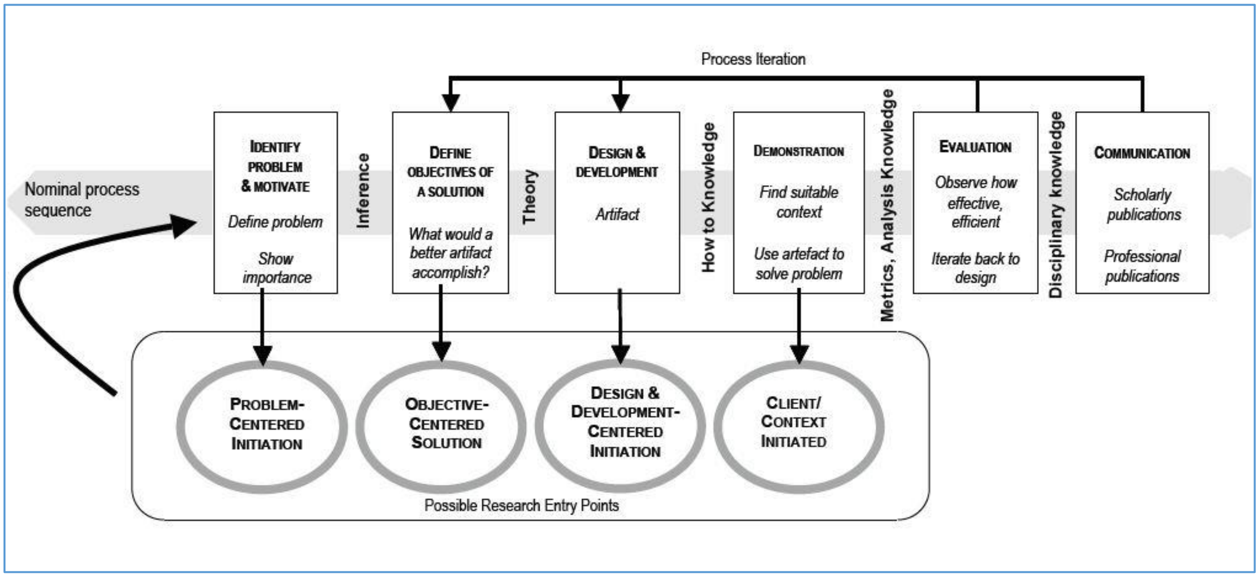

Process Iteration
Nominal process sequence
IDENTIFY PROBLEM & MOTIVATE
Define problem
Show importance
Inference
DEFINE OBJECTIVES OF A SOLUTION
What would a better artifact accomplish?
Theory
DESIGN & DEVELOPMENT
Artifact
How to Knowledge
DEMONSTRATION
Find suitable context
Use artefact to solve problem
Metrics, Analysis Knowledge
EVALUATION
Observe how effective, efficient
Iterate back to design
Disciplinary knowledge
COMMUNICATION
Scholarly publications
Professional publications
PROBLEM-CENTERED INITIATION
OBJECTIVE-CENTERED SOLUTION
DESIGN & DEVELOPMENT-CENTERED INITIATION
CLIENT/ CONTEXT INITIATED
Possible Research Entry Points

FIGURE 3. THE DSRM PROCESS MODEL (PFEFFERS ET AL., 2007)

1. **Identify Problem & Motivate**: The research and practical motivations that stem out of the problem are identified and analyzed through internal research. Multiple stakeholders and their interests are identified. Formal research questions are formulated oriented on solutions.
2. **Define Objectives of a Solution**: The goal of this research and the design phase is elaborated on. Information is collected internally and from the literature and external market to assess the limitations and boundaries of the design phase.
3. **Design & Development Stage**: Three HLS drafts are presented, based on a core solution (Big Data).
4. **Validation**: Validation takes place by expert review before possible future implementation.
5. **Demonstration**: The practical implementation of the solution, possibly in future research.
6. **Evaluation**: After and during implementation, a performance measurement takes place.
7. **Communication**: Publications of results, unlikely to occur in the confidential case of banking.

In this research we aspire to pass through a shortened design iteration within DSRM. The scope will be primarily laid upon a stakeholder analysis and an artifact design (HLS) due to prioritization and limitations. After the designs of the HLSs are finalized, they are validated through an expert review panel. The implementation step of feasible solutions could be conducted in a future project.

A High Level Solution template will be used in the artifact design and development stage. This template is inspired by the HLS template NeoBank uses internally as a standard for business analysts to create business solutions. Designing solutions based on this improves the practical usability of the proposed solutions in terms of familiarity and comprehensibility. This version of the HLS template is enriched with additional variables to define the context in detailed fashion and provide further arguments to the creation of the scenarios. The results of the research questions will be used to fill out the template and provide new entry points for further experimentation and research for NeoBank. The drafted solutions are validated with internal legal counsel to determine the feasibility with regard to data and privacy legislation. Actual implementation might occur in a different project which would require approval of management first.

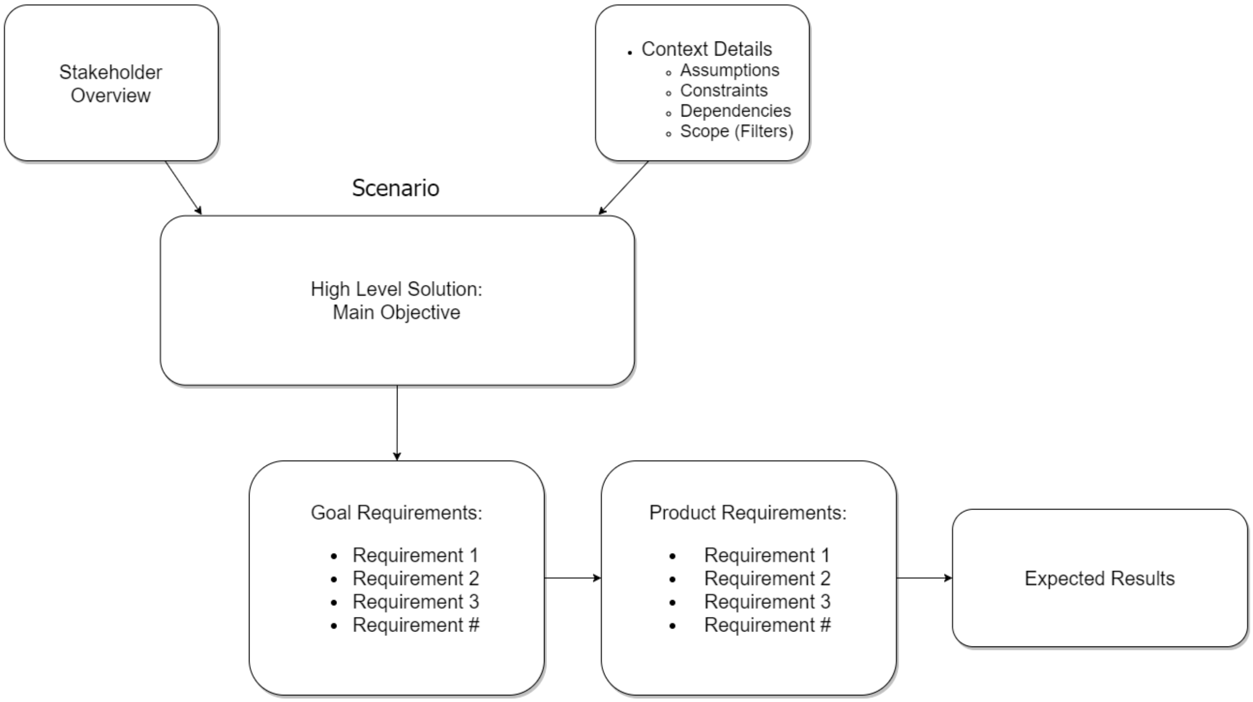


FIGURE 4. HIGH LEVEL SOLUTION TEMPLATE

- **Main Objective HLS -** Describes the objective and the resulting products of the project on main features. This is functionality described from the customer point of view. Also describes the added value the front-end application generates for the customer.
  - **Goal Requirements** - List of requirements that have to be reached before the main objective of the HLS can be completed.
    - **Product Requirements** – List of product requirements for the artifact/HLS in order to fulfill the main objective within the boundaries of the context.
      - **Expected Results** - Prognosis of the aimed effects of implementation.
- **Context Detail -** The context detail provides further information and arguments which specify the creation of the scenarios in which the HLS may or may not be feasible.
  - **Assumptions –** Lists the assumptions made in the scenario of this HLS.
  - **Constraints** - Describes restrictions in the design and known limitations which may not be dealt with by the firm’s current capabilities, architecture and infrastructure.
  - **Dependencies –** Lists the dependent conditions in which the HLS is designed.
  - **Scope** - Describes what is in scope of the project on main features.

Key Stakeholder	Interests	Opposing Arguments	Impact of HLS	Influence over HLS

FIGURE 5. STAKEHOLDER OVERVIEW TEMPLATE

The stakeholder overview is used to illustrate the primary stakeholders within the HLS project. It can also be used to formulate strategies to deal with blocking stakeholders or the negative impact of an HLS.

Based on this template, what immediately catches the eye is that the design template itself does not stake one initial problem. This is done as the common underlying problem is the growth of structural consumer debt. The High Level Solutions all treat this problem in a different way through a front-end application. Practically speaking for NeoBank, there are multiple stakeholders with various interests considering the use of these various data-driven applications which determine client creditworthiness. A global overview is given with regard to each of these solutions through the HLS template.

The HLS is made by the business analyst in consultation of many team members. The acceptant of the HLS is the architect. Its purpose is to determine the solution direction on main features with all the parties involved so that Solution Architect and Product Owner can approve it. It is also a first step towards feasibility determination. In order to capture the full compatibility with the HLS model of NeoBank, all Business Modules involved need to be listed as well. They are categorized as new, existing and to be modified or existing and no change needed. Relevant channel switches need to be stated as well for administrative (and security) purposes. For instance, from face to face to internet secure. Visual aid should be provided such as illustrations of the relevant states of the orderline. Finally, the chosen implementation process is stated which can vary between Pilot, Trail and/or Large Scale. A concrete explanation is given on choices made by the project. Implementation of an HLS can have great impact on an organization and should therefore be consistent with the integrated momentum of change throughout the firm. Misaligned HLSs with the strategic vision of the organization or departments can cause disputes over resources, troubled employees and diminishes the potential benefits of the project. It is recommended that for complex projects, multiple alternatives are proposed with different degrees of change so that a suitable solution can be chosen for each scenario.

3 Literature Review

In this chapter an overview will be given of the knowledge available in the literature on determining creditworthiness. This chapter is divided in two sections; academic literature and non-scientific literature. The first part consists of scientific articles exclusively found through academic search engines. These have been selected on criteria specified in the search strategy previously explained in section 2.1. The second part consists of non-scientific literature found in accredited whitepapers published by multinationals and articles on Fintech companies. The separation is made due to the difference in validity of the literature.

3.1 Academic Literature

This section discusses the validated scientific addition to our research. As a starting point, it is useful to establish the definition of the term creditworthiness. Creditworthiness is defined in the literature as "*the intrinsic quality of people and businesses reflected in their* ***ability*** *and* ***willingness*** *to fulfil their business obligations"* (Safi & Lin, 2014). There is a notable distinction between the ability and the willingness, as the focus of financial institutions is currently primarily on the financial capacity of clients to repay loans and not necessarily their habits.

It is noteworthy that the definition of creditworthiness is less elaborate when considering common use in the society[1]. The word is often used in this significance in common situations, for instance when applying for a loan or signing a mortgage. A more financially accurate definition is available to financially oriented organizations[2]. This definition is a more extended version which includes the potential consequence of default. It includes the societal meaning and links it with the systems and processes of the financial world. This version concurs with the definition found in the literature by Mavlutova et al. (2014), which mention in their paper that the view of economists with respect to creditworthiness is classified in two elements. The emphasis lies on the moral aspects of the borrowers due to the influence their moral compass has on their willingness to repay. Furthermore, the papers assume that the basis of creditworthiness is "*the ability to generate profits for servicing obligations*". This would be accompanied by a continuous absence of default and also an effective use of borrowed resources. For the remainder of this thesis we will use the broad definition and take all aspects of the definitions into account. This offers a wider financial context and includes the psychological aspect.

The term Big Data was first defined as *"the collection of large modern data sets that are difficult to process using on-hand data management tools or traditional data processing applications due to the sheer size and complexity."* The term is currently also a popular way to refer to the technology used to deal with the massive digital information available in both many forms integrated from multiple, diverse, dynamic sources of information (Srinivasa & Metha, 2014). At the META Group which has now been renamed Gartner, the authors Beyer & Laney (2012) characterized Big Data with three dimensions. According to them, Big Data was defined as *"high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."* The fourth and fifth V of Value and Veracity have been added in adaptations of these dimensions that describe Big Data and its components. Big Data Analytics is defined as involving analysis of huge data in order to unmask valuable patterns and information (Hafiz, et al., 2015). With the two critical concepts of this thesis now defined to a minimum level of detail, it is now possible to start answering the research questions of this thesis based on the available literature.

[1] Merriam-Webster (2016) defines creditworthiness as *"the extent of being financially sound enough to justify the extension of credit*."

[2] Investopedia (2016) defines creditworthiness as "*a valuation performed by lenders that determines the possibility a borrower may default on his debt obligations. It considers factors, such as repayment history and credit score. Lending institutions also consider the amount of available assets and the amount of liabilities to determine the probability of a customer's default.*

3.1.1 Structure of Credit Scoring

Essentially, determining creditworthiness is the process of *predicting* the risk of default. In order to perform in a risk environment, banks must be able to predict the likelihood that debt will be repaid in spite of information-asymmetry. The more accurately it is done, the better one has determined creditworthiness loyal to its definition. This means that many factors are taken into account that are relevant to the ability and the willingness to repay. Polillo (2011) mentions that creditors must not only know what the transaction is which is asked to be financed and how it is likely to turn out. They also need to know the customer, his business and even his private habits in order to sketch a clear picture of each instance. In the United States the financial system works with a FICO score for credit and mortgages. In the past, this scoring process used to be a subjective analysis by local credit officers based on documents and observations during personal appointments. Nowadays it consists of largely automated credit assessments by banks and credit scoring agencies.

Credit scoring is roughly structured by five general steps which can be all executed by one entity or by various entities. When the various steps in credit scoring are operated by different entities, there is usually a value exchange for the tasks performed. The credit scoring structure is as follows.

1. **Relevant Data Identification** – Potent data proxies are identified to determine creditworthiness.
2. **Data Collection** – Data is gathered on individuals, procured from different data brokers or exchanged through potential data partnerships.
3. **Data Conversion** – The vast amount of data on individuals and the market is converted to scores in digits, ratios or decisions through the use of algorithms and credit models.
4. **Score Distribution** – The computed scores are distributed when necessary to the decision makers or requestors. It is of importance that only the rightful proprietors gain access.
5. **Decision Making** – On basis of insights gathered by the credit score and possibly secondary information, decisions are made with regard to these respective individuals or households.

The distinguishing factor of credit scorers mutually, is the ability to convert the data into a meaningful score on which a well-founded decision can be made. Currently, the credit scoring process is predominantly built upon widely available predictive modeling techniques which are used in statistics and computer science. The credit model which is kept behind dictates the moving parts within the credit scoring process. Einav et al. (2015) mention that large data sets are usually taken, detailed to the individual which contains a key outcome-to-predict with a rich list of many potential regressors. Creators of algorithms then deploy state-of-the-art predictive models to select independent variables and through this process build the most successful predictive model. However, within these models the behavioral response is not registered. This means that the current risk scores do not capture heterogeneity across each individuals' unique behavioral response to the financial product. Indirectly this means that while a person might be fully capable to earn back and pay the money, that person's eventual default risk is also dependable on its actual behavior and willingness. This can also be influenced by personal situations where money is urgently needed and spent. These are economic choices which are heterogeneous across individual consumers. The authors argue that risk scores are not merely statistical objects but are influenced heavily by economic behavior and circumstances as well.

This issue is also regarded by Zarsky (2016) who addresses two crucial assumptions when allowing algorithms to sort, govern or decide on issues related to human behavior. Indirectly, the assumption is made that human behavior is consistent and with sufficient data, human behavior becomes predictable.

The extent to which human nature is predictable is not the topic of this thesis. However, we do treat the question of how accurate creditworthiness is determinable. When designing a financial product, one must recognize the limited errors such predictions might entail in its totality.

Citron & Pasquale (2014) believe that credit scores have a very large impact on the lives of consumers as they dictate whether a person gets credit or not. The article warns of scores becoming self-fulfilling prophecies, causing the financial distress rather than merely indicating it. The authors call for increased transparency in the workings of credit score calculation and introduction of continuous oversight to protect human values against arbitrary and inaccurate decision making. Information provided by clients can be inaccurate and biased while the predictive algorithms and source code of scoring processes are systematic. They make mention of the fact that the credit scoring systems in the United States in current form are black box assessments and shrouded in obscurity. The credit scoring agencies defend themselves by saying that their algorithms are trade secrets and publishing them would mean giving up their competitive advantage. There is however a possibility to publish a rough outline of what is being measured in order to let authorities review it upon fairness. Scored consumers would also be able to make objections and defend themselves, however fraud and system gaming would also become an increasing concern.

On basis of the literature, we have determined a list of conditions that credit scoring processes must comply with, in order to ethically determine consumer creditworthiness. These conditions are as follows;

- **Transparency** – In order to safeguard against abuse and to ensure fairness and validity, a certain level of transparency must be the basis of a credit scoring process. This can be built in by publishing which data is taken into account so that individuals can influence this. Another option is through interactive modeling. This shows and potentially teaches individuals what alterations in behavior cause an increased risk in credit default and thus change in score.
- **Participation** – The consumer must be willing to transfer data and influence the scoring system. The only way to truly reduce consumer debt is to educate, counsel and inform consumers financially. This can be done by active participation and keeping the credit application process truthful and transparent. Participation is positively reinforced in lacking areas.
- **Fairness** – This condition was earlier mentioned in transparency and ensures a stable process without frustration from the clients' side. This means that scores must be susceptible to improvements in behavior and loyal to the definition of creditworthiness.
- **Accountability** – By considering consumers accountable for their own score and the scoring party accountable for the argumentation of the verdict, the scoring system will have effect. This will ensure that the consumer understands how the credit score has come about and how it can be affected by oneself or the market (Citron & Pasquale, 2014).
- **Accuracy** – The accuracy of a credit score indicates the actual depiction true to a client's creditworthiness. By underfinancing or over financing consumers, additional problems come into play. The condition of accuracy will be further elaborated on in a latter subsection of this chapter.
- **Fraud Resistance –**Resistance to fraud can be achieved through looking at an enormous number of data points to create fraud proof customer profiles. However, transparency can enable criminals to manipulate profiles to their will. Blind automatization is a lurking danger.
- **Legislative Compliance** – Each area of the data scoring process and each country has their own laws and regulation with regard to credit scoring and data collection. Banks and financial institutions need to comply and function effectively in these legal frameworks (Zarsky, 2016).

3.1.2 Relevant Data

In order to measure creditworthiness, relevant data is required to determine the probability of default. The consumer creditworthiness in basis consists of the ability and the willingness to repay, relevant data variables are divided between these components. The ability to repay is traditionally measured by financial data that shows the consumer's existing capital and financial health. The willingness to repay is a complex behavioral factor and thus difficult to express. Traditionally this is done by verifying past on-time payments and loan applications, from duration to capacity. Baer et al. (2013) mention that Big Data proxy measures which are used in determining creditworthiness, can be split in three categories. The risk is calculated by building a credit risk model through the use of various data variables in these categories.

- **Identity** – Information verifying the personal identity that is used to prevent fraud.
- **Ability to Repay** – Current net income, current debt load, fixed expenditures, ratios such as loan-to-income (LTI), loan-to-value (LTV), regularity of cash flow and collateral net worth. Non-financial data include past and present online shopping time, habits and reputation.
- **Willingness to Repay** – Payment histories (e.g. utility bills), Credit history and experience at different banks with previous loans. Non-financial data include habits and reputation once more, type of items purchased, books read and customer reviews written.

We add the factor of **Circumstantial Risk**, which is the situational financial environment in which the consumer is navigating. This usually gives rise to the sudden motives of credit lending, or explains the sudden drops in creditworthiness. Circumstantial risk should be accounted for by assessing the client's current personal situation and environment for more accurate predictions. The "Theory of Planned Behavior" states that individuals are rational but their intention is always influenced by subjective norms and situational attitude. This indicates that consumers will not always fulfill their debt obligations when the perceived costs of repayment exceed the benefits. Client default can occur in spite of the consumer's complying ability to repay. This increases the importance of data points describing the other components.

Additionally, the Big Data industry distinguishes between structured, semi-structured and unstructured data. Structured data is data that fits in relational tables and is highly organized. Semi-structured data does not necessarily follow relational patterns and formal structures but contains elements of arrangement. Examples are XML and JSON files. Unstructured data is the most unique type of data and thus the most defiant to analysis. It has no pre-defined data model and can range from text documents to video files (Srinivasa & Metha, 2014). Another similar distinction the authors Wang et al. (2013) make in their paper is between hard credit information and soft credit information. Hard credit information is financial data and mostly structured. Soft credit information is non-financial or non-traditional data and can be represented in unstructured form as well. The paper argues that soft credit information is needed to comprehend the creditworthiness in today's vast and dynamic eCommerce market. This is in line with the earlier theory mentioned that the creditworthiness is representable by the components ability and willingness to repay. Most of the data used to determine the ability to repay is structured and the willingness to repay is mostly expressed by analyzing unstructured data. The latter analysis increases in importance when considering a client with a high risk profile.

Data can be obtained directly from clients, by exploring internally or procuring externally. Examples of external data partnerships are telecommunications providers, utility companies, wholesale suppliers, large retailers and governmental organs. These organizations would provide mobile-phone usage patterns, utility payment history, retail purchase history, tax payment consistency, demographic information and historical data on governmental support. The acquired data needs to be checked if it is legally permitted to take into use in order to estimate default risk. Third party data are usually preferred over the potential borrower due to the low trustworthiness of verifying oneself when applying for consumer credit. One of the key issues in this practical application is the reliability and validity of the factors used to determine creditworthiness. This can only be tested by using the relevant data variables itself in a proof of concept or case and analyzing the data afterwards. This has been done plentiful in the past by various Fintech companies that are discussed in a later section of this chapter.

Wang et al. (2013) mention that community reputation and built historical transaction history can help evaluate trustworthiness. Historical fulfillment of contracts and payment regularity also vouches for willingness to repay and trustworthiness as mentioned earlier. A similar result was published by Lin et al. (2009) in which was proven that the relational aspects of peer to peer lending are significant predictors of lending outcomes. While a system purely based on social network would lack sophisticated risk assessment, the soft information provided in social capital is substantial enough to be included. Large, transparent and verifiable relational networks are associated with high creditworthiness of individuals. An important condition is that the consumer is willing to disclose as much information as possible for the institution to determine his creditworthiness accurately. This means that in social capital, the cooperation is required of related individuals to verify creditworthiness. Transparent information reduces the probability of default by enabling lenders to monitor the customer's behavior, activity and social capital. More potential data variables that can be used to determine creditworthiness will be mentioned later in section 3.2, which discusses contemporary Fintech firms. More relevant data variables are being discovered as Machine Learning techniques are deployed like Active Learning. In this technique, the objective is to minimize three different costs: false positives (extending a loan to someone who defaults), false negatives (failing to give a loan to someone who would not have defaulted) and data labelling costs.

3.1.3 Applying Big Data Technology

Traditionally speaking, consumer credit scores are realized by using statistically based risk models and long established financial decision systems. Information would be requested by the financial institution from consumers and where possible also from third party data brokers. This data provided in text documents, images or through interviews would then be interpreted and entered by administrators in their own database. A process which is by current standards inefficient, costly and prone to human error. Moreover, there is no guarantee that the data provided is sufficient and accurate. There is sensitivity to fraud when false data is provided to influence the model. Once the data is processed, a snapshot credit score can be computed, on which a credit would be granted or rejected. In this evaluation, the bank would be oblivious to the customer's activity. There are no records of consumer spending habits while banks have internal client data that can function as digital traces. For instance, bank card transactions can be interpreted to deduce behavioral patterns (Sobolevsky, et al., 2014).

The banking industry nowadays has already moved partially from physical to digital environment by offering their financial products and services online through various applications. However, most of the large banks still evaluate consumer creditworthiness by using traditional financial data and non-dynamic data provided by clients in the procedure. This traditional approval-based credit risk management process is becoming outdated and less relevant, but banks have trouble adopting new technology. The eCommerce market and Fintech companies bring new real-time challenges which require up to date dynamic information. Beforehand this large scale and real time analysis was not possible due to technical limitations and lack of data. In this day and age, more data than ever is collected on individuals everywhere. Big Data technology has progressed remarkably to enable the affordable storage and computing power to conduct advanced analysis (Wang, Li, & Lin, 2013).

Traditional consumer scoring processes are becoming less compatible with the fast online environment. Manually it can take weeks and in worse cases even months for an inquiry to be completed. By moving to the online environment, banks are expected to adjust by reducing opacity and introducing interactive and personalized processes, products and services. Moving from an approval based consumer loan system to an accurate credit scoring system can show strong positive effects on the profitability. A better accuracy can be acquired on high risk clients and lower risk clients are allowed to obtain a larger loan capacity. Improved credit scoring also allows for a better fit of financial products and services which has benefits for the client and borrower such as lower buffers needed and lower rates (Einav, Jenkins, & Levin, 2013).

Machine Learning

One of the most important and widely used Big Data analysis tools is Machine Learning (ML). Machine Learning is the field of study that uses autonomic algorithms or sets of rules to learn and predict from large amounts of data. Computers are programmed to optimize a performance criterion using example data or past experience (Alpaydin, 2014). The behavioral aspect of creditworthiness is dependent on an enormous amount of data points per individual. The technology offers ways to cope with these complex behaviors which describe the willingness to repay, and paths ways to predict the future ability to repay. Application of analytical methods such as traditional statistics or machine learning to large databases is called data mining. Data mining programs have the potential to optimize predictive algorithms which predict types of behavior suggesting loan repayment. These determine which adjustments were most effective in the past. Iterations are continuously run to further optimize in tasks like minimizing default rate (Citron & Pasquale, 2014). In the paper by Hafiz et al. (2015) Lifelong Machine Learning (LML) is similarly mentioned as a means for improving machine learning algorithms with newly added data. This also deals with the major challenge of constantly evolving financial ratios and market sentiment. Additional challenges which are introduced by using this method are the reassurance of continuous validity and the scalability. Especially for individual consumer profiles, the creditworthiness can change and differ a lot after large sudden purchases, changes in purchase patterns or financial and personal crises. There is a distinction between three types of Machine Learning which is shown below (Kotsiantis, 2007).

- **Supervised ML** – Consists mainly of predictive models that work with labeled historical training data. This type of ML is driven by an optimization task given up front by the builder. In other words, instances are given with known labels and response variable. It primarily makes use of regression and classification techniques.
- **Unsupervised ML** – Consists of descriptive models without a target and hierarchy in features that work with unlabeled data. The focus of this type of ML is rather exploratory, practitioners hope to discover new relevant classes of items. It primarily makes use of clustering (e.g. K-Means), density estimation and dimensionality reduction (e.g. feature extraction) techniques.
- **Reinforcement Learning** – This type of ML focuses on taking specific decisions on basis of requirements that measure and optimize performance. The machine trains itself continuously by interacting with the environment in trial and error fashion and evolves on autonomic basis.

In the wake of upcoming technology as Machine Learning and Artificial Intelligence (AI), military scholars introduced a “technological due process” model to deal with automatized decision making. This process reintroduces human oversight and expert review, which is essential when technology is implemented to make impactful decisions. This “Human-on-the-loop” approach ensures that the decision algorithms won’t carry the unique responsibility of these important outcomes (Citron & Pasquale, 2014).

Machine Learning and statistical modeling are similar in their objective to extract information from data, but ML copes with the vast scale of data and has minimal human involvement in its model building process. In fact, ML makes use of statistics in building mathematical models as it makes inference on sample data. Both are used in the Bankruptcy Prediction Model developed by Hafiz et al. (2015). This model was used to determine the creditworthiness of companies, expressed in a bankruptcy rate. This can be compared with modeling the creditworthiness of consumers expressed in the default rate. The difference would lie in the used data points and the volatility of an entity as an organization and that of a human being.

A few of the widely used generic Machine Learning techniques used in predictive models are briefly explained in Appendix J. The best choice of algorithm always depends on the task at hand. This is why the data scientist is so important, in order to assess the situation. Tasks include making sure the data is correctly structured, identifying biases and choosing how to use which tooling. There are certain characteristics that suit certain algorithms better nearly all of the time.

In ML when problems are scaled to Big Data, usually a tailored combination of algorithms is needed to find the most optimal solution. Hafiz et al. (2015) state that a suitable platform to execute Machine Learning for classification problems such as credit scoring would be Apache Mahout with MapReduce as programming model. Algorithms are run on this system to classify the data collected in the Hadoop Distributed File System (HDFS). These algorithms would then be tasked to compute and predict the default rate of consumers. Predictive algorithms can automate the learning process by determining which adjustments in parameters and data variables worked best in the past. After a certain threshold the process would be able to recommend further iterations for testing itself. This self-optimization process will eventually narrow down on the most accurate types of behavior that suggest eventual loan repayment or default. In order for this to work, substantial improvements must be recognized by the machine through rules and definitions. An "understanding" of its own problem solving process is needed so that predictive algorithms may evolve to develop an AI that guides their evolution.

ML has been applied to the credit scoring process for quite some time in *classification* problems. Banks have always needed to predict the risk associated with loans to predict the margin of profit and to not grant a loan above a customer's financial capacity. Certain criteria would also indicate red flags in identity fraud detection and anomaly detection. Traditional data used includes income, savings, collaterals, profession, age, past financial history and records of past loans. The goal is to use designated training data to infer relation between a customer's attributes and the risk. The machine learning algorithm fits a model to the past data to calculate the risk for a new application. It is important to not overfit to the training data or underfit to generalizations, this is known as inductive bias. Discriminant functions separate clients in different classes on basis of outcome prediction. This is an example of Supervised ML, when applied to structured and additionally unstructured data on enormous scale and fast pace, it can produce useful insights. New interesting innovations come from applications in Unsupervised ML and Reinforcement Learning. New customer segments and non-traditional data variables can be found from the masses of consumer data generated in this age. Real-time automatic decision making is becoming more important to deal with the fast pace and scale of business (Alpaydin, 2014).

In general, it can be said that descriptive analytics is used to discover new insights and data variables that are related to creditworthiness. Predictive analytics can be used to predict the probability of default on basis of the available data. The prospect exists that the internet finance of the future consists mainly out of firms that leverage Big Data analysis tools (Yu, 2015). The ability of data collection and analysis is imperative to ensure the sharpest of margins through accurate prediction. Value based innovations using Big Data applications will become a distinguishing factor within the market. This will be achieved through the constant drive provided by up and coming Fintech starters that challenge the status quo. The author mentions two main purposes of Big Data in finance. Firstly, to effectively alleviate the problem of information asymmetry between supply and demand on a more accurate basis. Secondly, Big Data also reduces transaction costs and financial risks. Real-time monitoring can take place to reduce the risk of bad debts by updating the results of risk assessments continuously. Which is an aspect that comes back when assessing the creditworthiness of individuals throughout their loan period.

Predictive Behavior Modeling

The progress in mobile technology has led to rapid developments in contactless payment. These millions of transactions register place, time and amount of money paid. Singh et al. (2015) found that financial outcomes for consumers are complexly related to their human consumption habits across space and time. The authors use continuously generated streams of geo-mobile data to determine diversity, loyalty and regularity in order to construct a financial and behavioral profile. The predictions are inspired on biologically inherent behavioral traits of animals, namely exploration, engagement and elasticity. The financial outcomes of individuals are categorized in three classes: (1) overspending, (2) late payment and (3) in financial trouble. Machine learning can be used to distribute consumers over these financial states. This would then be treated as a predictive classification problem in which the probability is computed of consumers in landing in one of these classes over the course of a period.

Through the use of continuous financial prediction, banks can handle accordingly to each unique situation by providing counsel, mitigating overspending and preventing delinquency. With additional insights, risks can be avoided and the number of “bad loans” granted can be reduced. Singh et al. (2015) analyzed the detailed data of consumers’ shopping behavior in order to predict financial outcomes of these individuals. The study was conducted by analyzing an anonymized credit-card transaction database from a large financial institution. The generated models were 30% to 49% better than traditional demographic models for predicting financial outcomes. Habit markers and historic shopping data are also harder to manipulate than a social-economic profile. Financially healthy consumers tend to shop at diverse locations but at consistent times in the week. The three “favorite” shopping locations account for approximately 90% of all shopping expenditure. This indicates sustained familiarity, healthy routines and employment. The behavioral mobility features (diversity, loyalty and regularity) were found to be significantly related to the financial outcomes. Consumers with high regularity were more likely to pay their bills on time. Consumers with a high level of diversity or loyalty were less likely to overspend, but more likely to miss payments or get into financial trouble.

Ideally speaking, creditworthiness of individuals would be determined by the bank based on the data of their mobility traces and behavioral patterns rather than merely on financial statistics or collaterals. Rich profiles can be continuously created by the transaction data of customers so that financial troubles can be mitigated by giving timely financial advice. However, the tracking and monitoring of all these data can be a serious infringement on the privacy of individuals. Financial institutions need to make sure they comply with the laws and regulation regarding consumer privacy. Where needed, individuals must sign the right terms of agreement for use of certain data intensive financial products like budget counseling. The bank needs to determine which level of detail it would track and gather data, in order to achieve full functionality without upsetting the potential users. Future research is needed to map the legal framework in which Big Data can operate and how this inhibits the development of Big Data applications in finance.

3.1.4 Accuracy of Creditworthiness

The accuracy of creditworthiness can be improved by using Big Data to discover new variables and calculating new parameters that can better express the creditworthiness of a client. These new data variables are then used in a model, which is tested on predictive accuracy. Circumstantial risk such as sudden unemployment and divorce are near impossible to predict in an accurate manner, not even by using Big Data. The only way to compensate for this is to monitor and update creditworthiness timely. Simultaneously, credit model builders do not have the illusion to be able to model this and thus these special cases should be excluded from data sets to maintain the representability.

In order to measure the actual creditworthiness of a Machine Learning based prediction model, the outcomes of the model are compared with the reality by running a test with historical data. It is important that this data set is representative and obtained from the same source as the data set on which the model is built, the training data set. The same data set can obviously not be used because the predictive model was based on it. Common practice is that the initial sample data is split into three partitions in a certain proportion. The training data is the largest set and is used to build and enhance models. The validation data set is used to optimize parameter settings and select the best model built. The test data is strictly used to estimate the accuracy. Example methods of estimation are Holdout, Bootstrap and Cross Validation (Appendix K). The model's objective is to minimize the amount of false predictions. False positives cause banks to grant loans which a client cannot burden, and exposes the bank to unsalvageable outstanding loans. False negatives cause banks to lose out on potential clients, and forces the client to look elsewhere for a (riskier) loan. A Receiver Operating Characteristic curve can be plotted with these outcomes to illustrate the trade off and compare the performance of models. The accuracy is expressed by dividing the amount of true results by the total amount of results. Notice that the below results are only applicable to the final decision of granting or rejecting a loan, and not to determine more complex attributes of a loan such as limit height and risk percentage (Alpaydin, 2014).

	Predicted Result		
Actual Result		No Default	Default
	No Default	True Positive (n)	False Negative (n)
	Default	False Positive (n)	True Negative (n)

FIGURE 6. COMPUTING THE ACCURACY OF A CREDIT SCORING MODEL (CONFUSION MATRIX)

Zarsky (2016) mentions that data can be wrong or that the processing of data can contain errors. The scoring process on individual level can still generate an accurate credit score for applicants. On aggregated level, the errors will have small effect on the total numbers when dealing with a robust model. A potential possibility in improving the accuracy of credit scores is to heighten the transparency. The author argues that transparency allows individuals to understand how their credit score is built and provides them with additional insight in their data. This will allow them to be able to correct data when needed. However, offering clients full disclosure is a complex issue, for banks and clients alike. A client's creditworthiness is a dynamic attribute which is subject to change over time. For a result to be both timely and accurate, it would require frequent and complete updates on many parameters or real time data capture and analysis. In the past this would not have been possible due to the overhead and lack of data. Big Data analysis can be deployed to reduce transaction costs and financial risks by monitoring clients through complete customer profiles. This reduces the risk of bad structural debts and defaults occurring and creditworthy clients will be able to obtain a loan easier. The human error factor can be reduced by replacing hierarchal decision-making by decision-rules with objective data as input.

3.1.5 Big Data Maturity

This subsection aims to introduce a model which is used to determine the maturity of Big Data. The Data Warehouse Institute (TDWI) Big Data maturity model is used to indicate the average progress and potential of Big Data. The maturity of Big Data is defined by Halper & Krishnan (2014) as "the evolution of an organization to integrate, manage and leverage all relevant internal and external data sources. It means creating an innovative ecosystem, delivering insightful business value and enabling impactful transformation." The maturity level factors in the combination of people, processes and technology measured through qualitative criteria. The establishment of Big Data maturity is useful for organizations when implementing new Big Data projects. It can be used to structure, measure and monitor the progress towards their strategic and high level goals. The five stages of the maturity model are illustrated in the figure below and are further described in this subsection.

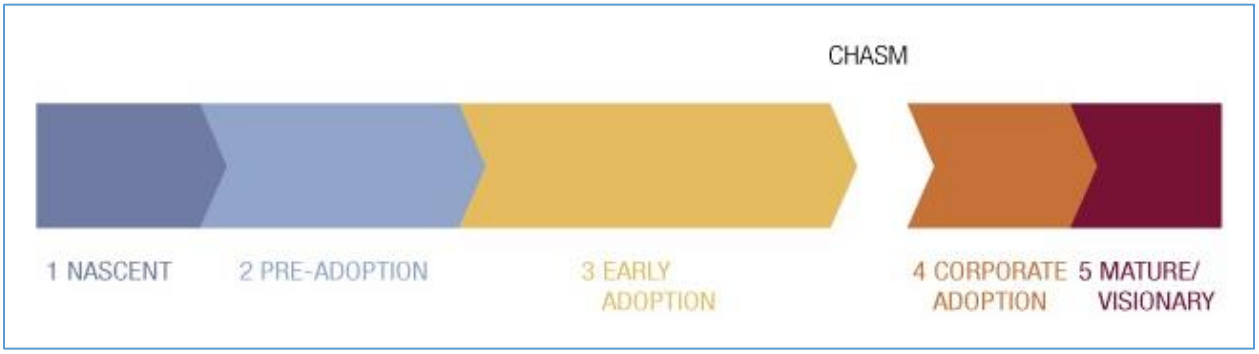


FIGURE 7. STAGES OF BIG DATA MATURITY (HALPER & KRISHNAN, 2014)

1. **Nascent** – This phase is characterized by the low awareness on Big Data and its functionality. There is no real data management strategy or support from the management. Data might be sparsely gathered but without an overall plan to generate value. Analytics might be used by certain subdivisions of the firm, while only using structured data.
2. **Pre-Adoption** – Research is done within the company on Big Data. Staff members might be attending conferences and awareness in general is increased. The organization will start thinking about adopting new technology such as Hadoop to support Big Data. There are likely a few visionaries, champions and promoters of Big Data which operate on departmental level. The board and CIO might agree on running proofs of concept to experiment with Big Data. The organization will use data as part of its decision making process but the company in its entirety is not data driven yet. At this stage, the organization will have typically initiated exploring advanced analytics because it is accentuated at certain parts of the company who envision the value of it. Special commissions, committees or divisions of data analytics might be founded.
3. **Early Adoption** – A common characteristic is that an organization has run one or two proof of concepts that have been completely adopted and are now production ready. The firm and employees are excited about the prospect of Big Data and the technology starts being included in the strategy. Big Data initiatives are moving at a fast pace and issues are known in data quality and security, but not necessarily addressed. Common implemented Big Data infrastructure include NoSQL databases and Hadoop clusters. This is the longest phase of the maturity model and many companies won't cross the "chasm" due to a lack of data driven policy. Using large internal sources, different types of data will be used in various departments in the organization for isolated purposes.

The Chasm – This gap between phase three and four represents a collection of challenges that companies have to face when transitioning. Challenges are met in: funding and business involvement, data management and governance, architecture, human resources and company political issues. The best way to make the transition happen, is to stimulate both top-down and bottom-up input and involvement.

4. **Corporate Adoption** – This period is characterized by company-wide influence of Big Data, the so called data driven environment. The data has more end-users and generates more value. Employees from every hierarchy and process in the chain are much stronger involved with the value that is extracted from the collected data. In this stage of maturity, the company has been convinced that Big Data analytics can be a competitive advantage. Structural funding has been established and a steady return on investment has been achieved. The infrastructure is fully implemented with a typical first tier production class cluster installed and maintained in the data center. A range of technology can be used within this phase like Hadoop, NoSQL databases, cloud-based techniques and data warehouses. Operational maturity is achieved and the firm can run multiple workloads on a cluster. At this point collaboration occurs between departments internally and between companies externally to share knowledge and data. Collection of new data and data analysis both occur rapidly. The organization is data driven at this stage and improvement focused through the use of Big Data. In terms of governance, companies will have structure and a special program with committees that research if data collection is done within the boundaries. They also ensure that the governance policy is aligned with Big Data operations.
5. **Mature/Visionary** – The very few enormous internet firms that have made it to visionary maturity for Big Data are experts in executing data oriented projects. Management is convinced and views Big Data analytics as critical asset on decision making. Collaboration is expanded and the company is always looking for new ways and business opportunities to use Big Data. Infrastructure is further improved in terms of security, disaster recovery, backup and performance management. The data storage, analytics and infrastructure are all state of the art and strategically aligned.

This maturity model is a framework that can be used as a benchmark for comparison when looking at an organization’s progression in Big Data. As organizations move through the steps of maturity, they gain additional value from the use of Big Data. Companies learn to better manage and make use of available data through this growth. The chasm between early adoption and corporate adoption indicates the difficulty of many firms to make the transition to this phase. On the other end of the spectrum, there are only a few companies belonging to the visionary stage of maturity. These are the Big Data internet giants that generate and analyze masses of data to create a competitive advantage. Concrete examples are Facebook, Yahoo, Google and Amazon. It is very important to note that while a company might have a decent maturity in one area relevant to Big Data, it might still be lagging behind in another. For example, a mature Big Data infrastructure and a nascent data management strategy. This is expressed by the Big Data criteria shown in the different dimensions in the figure below. The impact of having different levels of maturity in these dimensions differs for each industry branch and type of company. However, the extent of this difference is not the subject of this thesis.

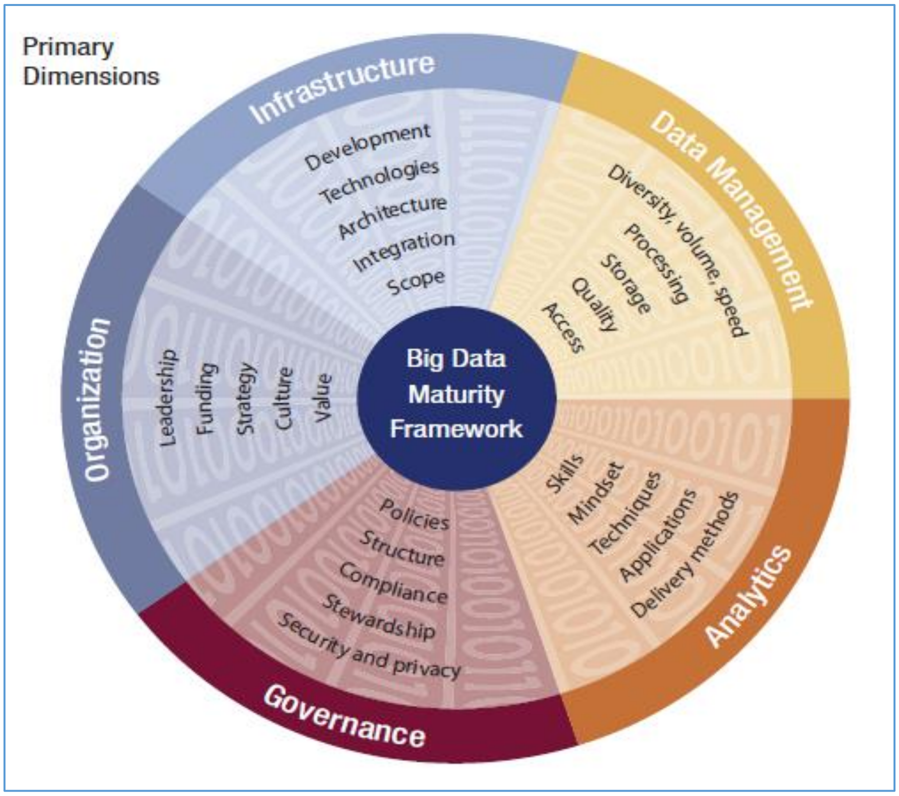


FIGURE 8. BIG DATA MATURITY ASSESSMENT CRITERIA (HALPER & KRISHNAN, 2014)

Each of the criteria within the five dimensions is ranked with a score from one to ten through qualitative or quantitative research. The scores of each criterion within the dimension are then accumulated to form a score for the dimension in total. A final verdict is given on the firm on basis of the average.

STAGE	SCORE PER DIMENSION
NASCENT	<15
PRE-ADOPTION	16–25
EARLY ADOPTION	26–35
CORPORATE ADOPTION	36–45
MATURE	46–49
VISIONARY	50

FIGURE 9. MATURITY SCORING TABLE (HALPER & KRISHNAN, 2014)

This model is used to distinguish between companies where Big Data tools have been widely adopted and traditional firms which still have steps to take. In large enterprises, there might be a lot of local departmental solutions and proof of concepts without improving the maturity substantially. As there are advancements in Big Data technology and the threshold lowers, more enterprises will see adoption in the future. The expectancy is that it will be a slow process as there are too many challenges and opponents restraining wide adoption of Big Data. The risks are quite high for implementation as it is costly to build from scratch and maintain afterwards. To see a decent return on investment, front-end Big Data applications will have to be developed and used successfully. Companies that are not specialized in IT don't have the staff to test and develop on efficient scale. They lack in developers, data scientists and distributed systems architects to really be able to produce valuable solutions in conjunction with legacy. In the analysis chapter, research is conducted to see how NeoBank fares internally in this maturity model.

3.2 Non-Academic Literature

The non-scientific literature review is primarily consulted to gain insight on how the Fintech technology has developed. Other (Fintech) organizations are consulted to determine how credit is scored in varying industries. The focus is of course on the determination of consumer creditworthiness. The information in this section is used to discover how these methods and technologies are compatible with predicting consumer creditworthiness true to its definition. The compliance to the conditions determined in the first section of this chapter is taken into consideration. Practical cases, business reports, news articles and whitepapers were taken as source.

3.2.1 Consultancy and IT Company Whitepapers

As a part of the external market review, this subsection will treat some of the whitepapers and nonscientific reports collected from various consultancy bureaus, large IT enterprises and financial service providers. This subsection is organized by the major themes represented in the papers published by these companies. These articles illustrate the situation of the current financial market and how technology is gaining in importance. Valuable insights are shared on how financial institutions can capitalize on Big Data technology to pursue a variety of business goals. The articles note how the lending and underwriting process will change over the coming years, driven by innovations in data technology.

Financial Inclusion in Emerging Countries

A large portion of consumers in emerging countries have difficulties in obtaining a loan. In first world countries there are also underbanked consumers, but to a much different extent. In this case, China is taken as an example as only 20% of the adult population has a credit score. In the banks' traditional decision systems, there is a lack of historical credit data on the majority of consumers. Upon requesting files from the clients themselves, it becomes apparent that there is no trustworthy credit history of these individuals at all. This emerging middle class is in dire need of a credit scoring system that will allow fair approval of loans based on their actual creditworthiness. To achieve this, Kshetri (2016) researched the possibility of using Big Data in the determination of creditworthiness of consumers and SMEs. Proxy measures were used to determine a credit applicant's ability and willingness to repay. Data can be gathered mostly internally by banks, but also externally through social media and mobile phone usage patterns. The author mentions the benefits of using rich unstructured data next to structured data.

In his article, Kshetri (2016) highlights practical cases in which Big Data is being used to determine creditworthiness. The Chinese internet giant Alibaba already uses its own credit scoring system based on retail data and eCommerce transactions. The Alibaba Group had 300 million registered users and 37 million small businesses in 2015. Alibaba started assessing the creditworthiness of SMEs in 2007. Sesame (Zhima) Credit was founded in January 2015, a credit scoring system that offers businesses and consumers to construct and access their credit profiles. The system draws internal data from the Open Data processing center and the rest of the Alibaba ecosystem. External data is procured from partners and both online and offline data is used. Examples are court reports, debt deficiency, late returns of rented cars and transactions registered on Alipay. Tencent's WeBank is another example of a Chinese internet giant that branched out to internet finance. They work with an underwriting system that requires clients to upload a frontal picture of their face. The system matches the image with data provided by the Ministry of Public Security to verify the consumer's identity. The system then gives the individual a credit score rating based on a large number of sources. Examples are online shopping history, activities on social networks, website games and other online activity.

Besides domestic enterprises, JD-ZestFinance Gaia is a joint venture by Chinese online retailer JD.com and ZestFinance, an American lender startup which will be elaborated on in subsection 3.2.2. In short, the company combines Machine Learning with Big Data scale analysis to assess a client's creditworthiness. ZestFinance claim to use tens of thousands of data points to assess potential borrowers. Algorithms are used to compute the risk by using past and present shopping data of the second largest internet retailer in China. JD.com already offers online credit for purchases on their websites based on these computations.

All of these developments have led to plans made by China's State Council to issue a state governed Social Credit System by 2020. They envision that every adult in China would have a credit code linked to their identity card. This would be derived by many elements such as financial standing, criminal record and social media behavior. The goal of this system would be to provide the Chinese population with justifiable and trustworthy credit scores. This would reduce the odds that a bad loan would be granted and increase the chances for entrepreneurial minds in the population. In total, China has given ten private companies permission to launch internet-based credit rating systems to innovate in these areas. The complete value in loans granted through these companies is not significant in relation to the total value of loans made by banks in China. It adds value in a different way as it makes small credit available to a large portion of the population, whereas it was not possible for them to lend credit beforehand. By enabling them to lend money, these Fintech firms are allowing them to build their own credit history. This in turn can be used to obtain a bigger loan at the traditional bank in the future. However, these developments require special circumstances such as the scale of online retailers in China like Alibaba to generate, store and analyze their own Big Data within a fully mature architecture. Traditional banks will find it difficult to reach this level of representable versatility in data. Other reasons would lie in law and regulation regarding data privacy and the compliance to laws. Providing financial products and services as an IT company has different conditions and legislation in effect throughout different parts of the world.

Selecting Non-Traditional Data Variables

Research by FICO (2015) on alternative data to expand credit access has concluded that it is possible to reliably and accurately score creditworthiness with the use of non-traditional data. These methods contribute to additional financial inclusion of the population, as so called "thin-file" clients can obtain initial credit and further build up their credit history. In the United States, 22% of the consumer population doesn't have sufficient registered or accredited details to generate a credit score, the data is too sparse, old or non-existent. This group of 25 million consumers contains many creditworthy individuals. Inaccurate credit scoring creates higher default rates, but also lower lending volume than necessary for optimal economic performance. Debtors receive lower credit space than needed or are overloaded with more than they can handle. A lack of data flowing in files or an inactive credit history should not automatically translate in a negative score, but the disadvantages are apparent. The biggest problem is that this low-income consumer group needs a loan or credit to start building this history that is lacking.

Research also pointed out that unscorable credit applicants were three times as likely to default as scorable applicants. However, there is a large difference in credit risk between these unscorable applicants. Some are new to credit, some are retired and inactive and some have lost access to credit due to previous misconduct. It is also notable that the type of credit sought after, differs per consumer group. When complementing the lacking bureau data with telecom data a substantial improvement was seen in the Gini index, indicating better predictive performance. However, the authors warn that not all non-traditional data necessarily improves the accuracy of credit scoring. The data variables are held to the *FICO Six-Point Test* which include the following criteria:

1. **Regulatory compliance** – All data sources and data variables must comply with the legal frameworks that apply to them. The scorer must have an infrastructure that supports validation on a large scale. Used data variables must be grounded and justifiable towards all parties.
2. **Depth of information** – This factor covers the detail and context of data variables. The richer the data, the more accurate the score. When data is lacking in quality, the credit score will also suffer.
3. **Scope and consistency of coverage** – The alternative data source must cover a large percentage of the population in order to have high utility. Consistency in format is required for practically operating, analyzing and storing the data.
4. **Accuracy** – As mentioned earlier, the incoming data must be validated, tested and verified.
5. **Predictiveness** – In order to add value to a credit risk model, data must be proven predictive towards future consumer repayment behavior. This can be tested by using Big Data.
6. **Additive Value (Orthogonality)** – Data must be uniquely additive and not "double counted".

The case study results show that more than 50% of the previously unscorable applicants could be accurately assessed after implementation of complying non-traditional data variables. FICO revealed that many of these consumers would otherwise be stuck in the vicious circle of credit disability. A large part enters the credit mainstream and raises their credit scores to even higher levels through the years.

Citihub (Tivey, 2015) recommends establishing a *Center of Excellence* (CoE) for data analysis within each organization. This would consist of a division of teams that help identify Big Data projects and specify business goals for the technology. Responsibilities would also include detecting existing data quality issues and lack in capability needed to achieve these goals. They suggest using the agile development approach for Big Data research, in order to steepen the learning curve. New ideas and environments should be tested out quickly in iterations after which new data variables can be used to improve credit modelling processes in practice. The goal is to create an accurate customer profile in which creditworthiness is indicated together with preferences in order to personalize financial products and services. By doing so, customers will be bound by loyalty and the churn rate will decrease. This customer centric view should however not abandon traditional metrics to gauge ability to repay, but enhance existing processes.

In light of this detail, they illustrate in their case study how client account and transaction data can be mined to detect a changing environment. This factor had been named *circumstantial risk* in our thesis. The ideal picture is to even detect monumental life events such as marriage, divorce, promotion, retirement and starting a family. The data is then used to predict the likelihood in which a client would be susceptible to certain financial products in order to enhance the marketing channels. Client profiles are constructed and additionally enriched by using data from third parties such as social media, telecom or (online) retail. NG Data (2014) mentions the synergy of Big Data solutions with the increasing use of mobile wallets. This captures the required data in detail together with surrounding context as location and time. InsideBIGDATA (Guttierez, 2014) reaffirms the increasing trend in which predictive credit risk models are used, fed by large historical data sets. The models are used to predict the future payment behavior of clients and the client's propensity towards delinquency or payment. Credit risk management can also occur at the potential opening of new accounts. Big Data can be used to analyze the creditworthiness and compute the risk of default of new applicants to determine high risk profiles. Loan and credit decisions can be made in seconds by using automated processes based on Machine Learning algorithms. In the new age, scoring decisions are made based on data from various non-traditional sources like social media, internet retail, governmental databases, mobile and location data. The capacity of more than 10.000 data points in real-time is stated.

Growth of Big Data use in Finance: Personalization

Data generation is growing at enormous speed and by harvesting and acting upon it, organizations can bring upon hugely impactful changes to society. Traditionally speaking, the banking sector has always been data intensive. The advancement in technology, Internet of Things (IoT) and mobile devices offers even more potential data to act upon. Big Data can help exercise excellent Governance, Risk management and Compliance (GRC) as it can generate complete overviews of the entire firm. Making decisions on basis of client creditworthiness is indirectly a GRC activity due to the legislation, a varying risk appetite and the financial buffers that are maintained for it. Threat and fraud can also be identified by using anomaly detection on large data sets (Hewlett-Packard, 2013).

Customers provide valuable data to banks on regular basis. This can be in the form of structured data in transactions and logs, but also unstructured (audio) data from customer services. Historical data can give strong indications in individual preferences, buying patterns and pricing strategies. By using all of the internal customer data at the disposal of a bank, a complete profile can be sketched of the client-bank relationship. The logical next step would be *personalization*. The design of engaging financial products and services, marketed directly to these clients and their needs. By estimating the risk more accurately by using Big Data, rate optimization can be applied to financial products. Additionally, costs can be reduced by optimizing operational efficiency and reducing overhead of internal processes. Priority points are centralized access to essential data and the removal of data silos in order to get information where it is needed to act upon data-driven ideas (Hewlett-Packard, 2013).

Benefits when succeeding are new profit through increased credit and debit card usage and stronger customer engagement, loyalty and retention. Strong relationships with third party organizations and collaborations that enable better qualitative cross and up selling. Insights acquired from analyzing internal and external data include detailed shopping behavior, influence of social media on shopping, tendency to mobile channels, price sensitivity, hobbies, interests and many more preferences. Many insights can already be obtained by analyzing internal transaction (context) data. This data can then be deployed to customize client approach and offer them personalized deals to enhance the customer experience at the right time in the right place and in the right form (NG Data, 2014).

One of the reasons that the personalization trend has gained so much traction is the so called Millennial consumer population. Innotribe (McAuley & Weiner, 2015) claims that this generation of consumers is less loyal and more skeptical towards financial institutions. Their survey conducted with over 10.000 respondents concluded that half of all Millennials didn't think their bank offered anything different than their competitors. Millennials trust in technology and are interested in digital products that are relevant to their daily life. This generation shift and change in preferences of large consumer populations is also noted by the WEF (2015). Results of this rapid adoption of technology are the previously mentioned data driven trends and innovations in mobile banking, virtual banks and banking as platforms. The WEF believes that customer expectations will continue to rise which will change the role of the bank. They will need to provide a whole digital customer experience and a customized suite of financial products and services. Another option for traditional banks is outsourcing the digital customer experience by partnering with trustworthy non-traditional providers while focusing on manufacturing financial products. This could however result in a power struggle between the tech firm and the bank. This also reduces the control that the bank has over the customer experience and possibly reduces its direct access to customer data.

Disruption in the Financial Services Market

An article published by the World Economic Forum (WEF) explores the transformative potential of Fintech starters and technological innovation on contemporary business models in financial services. They believe that the disruption in this sector will be expressed by a continuous pressure towards innovation. This will reshape customer behaviors, business models and the long-term structure of the financial services industry. Various stakeholders need to collaborate and come to comprehend how new innovations will alter the risk profile of the industry. The WEF project identified eleven key clusters in which innovation takes place. Two of these clusters will be discussed as they are relevant to our topic of lending. The researchers believe that consumer lending will be transformed due to new ways of credit evaluation and origination. It will become more difficult to determine creditworthiness when many alternative platforms exist that each provide unique credit to track. Another disruptive point in the market is that new entrants are more specialized and consumer focused, offering stand-alone products. New innovations are also cutting out the traditional banks' role as intermediary and offering lower prices and higher returns to customers with their products (World Economic Forum, 2015).

The WEF believes that after the financial crisis in 2008, banks have lowered their risk appetites significantly. This gave alternative peer-to-peer lending platforms that can cater to each risk appetite, a chance to experience large growth. Other Fintech companies use alternative scoring methods and various automated processes to offer loans to a broader customer base. In comparison with new innovative lending processes the traditional lending models have limited access, slow manual processing speed that puts a strain on customers, models that margin for error, limited control and low returns on savings. The report predicts that the future will bring more accurate and streamlined underwriting which will increase the accessibility of credit to a broad audience. It is also said that the accuracy of risk profiling will reduce the costs for borrowers. *"Emerging alternative lending models create both competitive threats and evolutionary opportunities for financial institutions, making it important for incumbent institutions and alternative platforms to develop more integrated partnerships and learn from and share each other's capabilities."* This is in line with the most positive of the three future scenarios mentioned by the report.

1. **Traditional intermediaries are replaced by alternative platforms.**
   The traditional lending function of banks is replaced due to the cheap and effective alternative methods of other platforms. This could happen if traditional banks are too entrenched by their legacy systems and conservative attitude to keep up with changing consumer demands. And alternative platforms continue unhindered by legislation or client skepticism in their growth.
2. **Traditional intermediaries are complemented by alternative platforms.**
   Both parties coexist side by side, as banks will primarily cater to loyal low risk appetite clients based on trust. Alternative platforms will cater to high risk appetite clients based on accessibility. The industry will become more diversified while the disruption of traditional models is limited.
3. **Traditional intermediaries transform their own processes.**
   This option would become realistic when sufficient pressure is coming from alternative platforms to substitute the credit lending process of banks completely. It would however require banks to be capable of investing in new business processes and IT infrastructure. A different option would require the bank to acquire the alternative platform in itself.

The scenarios are speculative and have certain assumptions and requirements in market conditions to come true. It is however sure that alternative platforms will continue to pressure traditional banks in their lending process as they cater to a substantial group of underserved consumers.

Challenges and Obstacles

Cognizant, Marketforce & Pegasystems (2016) believe that the current way of retail banking is being revolutionized. Established business models are being replaced by smarter solutions that leverage technology. As the versatility in different options grows, customers are becoming less loyal and more value oriented. Especially the previously mentioned generation of Millennials is increasingly digitized and interested in additional added value like self-service, personalization and chatbots. In order for traditional banks to compete, they need to invest and capitalize on these data driven trends. According to the results of the questionnaires held with financial executives, *“93% agree that finding innovative ways to provide value adding services to customers based on data-driven insights will be crucial to long-term success”.* Moreover, *“75% of the respondents are expecting to offer full personalization and 83% are planning to predict individual requirements within five years.”*

Access to sufficiently rich customer data is one of the prerequisites and a fair share of organizations is experiencing difficulties acquiring this asset. The use of IoT is currently gaining market terrain and mobile payment is also becoming more popular. More people are getting “connected” every day and this indirectly means that more data is being generated by and of consumers. This can form the basis for applications in finance like dynamic pricing, need prediction and personalization. Due to the risk of data breaches and high level of data regulation, consumers need to give consent to allow access to their data. This can only happen through building trust and showing the consumers what added value and insights an application can offer. Transparency in policy and processes is key to earn and keep the trust of clients. Damage to reputation and bad branding can be catastrophic.

Another obstacle is the cautious attitude of most traditional banks. While all signs are indicate that the financial market will be significantly disrupted as illustrated by the WEF, banks still cling to their conservative policies. The Cognizant report states; *“only 30% of the respondents said their organization would tolerate a one out of two failure rate on innovation pilot projects and more than 60% believe their governing board should set the failure rate for innovation pilots much lower, below 30%.”* This runs opposite to the overwhelming opinion of nearly all organizations: that innovation is necessary in retail finance and that it can only be achieved by thinking beyond traditional industry standards.

In the InsideBIGDATA article, Guttierez (2014) also recognizes the cautious mindset in the financial sector which is wary of change through the adoption of new technologies. Besides the conservative culture at banks, the whitepaper also addresses the banks’ issue of low Customer Intelligence (CI) level. They believe that this is one of the main reasons that in spite of an abundance of data, little is actually capitalized upon. Banks need to look at their own customer retention rate and prevent being outperformed by competitors. Customer centric companies like banks will eventually need to adjust with the times and offer effective personalized products and services. In order to stay competitive, additional focus must be placed on indicators in Customer Relations such as churn rate and percentage cross and up-selling. More and more financial institutions are also aspiring to implement a real-time view of data in order to gain in effectivity. As the data is sometimes unavailable or the analysis and storage have requirements that are out of reach, banks can collaborate with third parties to use their scoring services.

In the article of NG Data (NG Data, 2014) the author argues that traditional banks are still doing too little with their data while Fintech organizations pose a realistic competitive threat. The growth in eCommerce and rise of compliance regulation form challenges and opportunities to invest in technology. Banks need to capitalize on this, especially because they have distinct inherent advantages. These advantages include the abundance of data and local retail presence to build relationships with merchants and consumers. Banks also have flexibility in spending capital and an established earned trust from consumers due to strong branding. However, this reputation needs to be upheld when participating in data intensive activities. *"Over 48% of the consumers trust banks the most as their mobile banking provider, when compared to other providers such as payment systems, mobile operators, retailers and technology companies."*.

Research and investments are needed in Big Data technology. Capable technical staff has to be contracted as most traditional banks are lacking in this field of expertise. Collaborative relationships with retailers are also an option when enhancing marketing programs through behavioral data. A huge barrier is also the large amount of data silos in the bank's IT infrastructure. With each system and data store holding different information with unique access rights. Incompatibility of data format and legacy systems is a big problem and prevents coherent reform which leads to the fragmented IT landscape that most banks have. Traditional banks are usually large inflexible corporations which require time to adopt new technologies. To capitalize on this technology, different departments like marketing will need to work with data analysts, product designers and potentially even other organizations in harmony. Collaboration is needed if models are to be created, tested, validated, verified and deployed. Data scientists work together with the risk department, underwriters and the legal department. Financial companies need to make sure that they won't offer financial products or services that have a disparate impact or break other laws. This will be discussed more in subsection 3.2.3 later on (NG Data, 2014).

3.2.2 Fintech Organizations

This subsection will exclusively treat Fintech firms and their lending activities. These whitepapers treat certain cases that form the existential ground that these companies built their business model on. Some papers feature the technology, business plan and mission of these organizations. They have been collected and bundled to obtain an overview of the various Fintech applications in the area of lending and determining creditworthiness. These various organizations are considered next to NeoBank to assess the compatibility of each application in the current company environment. All Fintech companies mentioned in this subsection use their own method to score creditworthiness, the companies are sorted by initial strategic goal and vision. It is notable that most successful organizations are expanding their activities.

Financial Inclusion of Thin File Consumers

Companies like Cignifi, InVenture and Lenddo are mainly focused on using alternative scoring methods based on non-traditional data to achieve financial inclusion in emerging markets. The organizations use different types of alternative data which is consistently generated by the consumer population ranging from mobile data to social media data. Their primary target group is the financially underserved population of mostly emerging markets due to a lack of historical credit data.

The organization Cignifi is a Fintech B2B company that analyzes mobile data patterns to compute accurate credit scores. After reaching out to Cignifi, two whitepapers were obtained from the Fintech company. Both papers describe the challenge and opportunity which ultimately form the company's mission.

The emerging markets in the world have a growing burgeoning middle class which lack credit history to determine their creditworthiness through traditional means. This inhibits them from contributing more to the economy as they lack financial access to make a bigger impact. Banks also have a large potential customer base, but the costs of traditional customer acquisition are quite high and inefficient. The solution to this was found by Cignifi in credit scoring based on an alternative source of non-traditional data: Mobile phone usage and patterns are analyzed as they are highly predictive of behavior. Credit risk is assessed through behavior models based on mobile phone data in the form of Call Detail Records (CDR). "*The Cignifi Risk Score measures the risk of default (60 days or more past due) at or before six months with a range of 300 to 800. A score of 800 indicates no measurable risk of default. Risk approximately doubles for each 60-point decrease in the score. The granular, numeric scale and format of the Cignifi Risk Score enables easy integration into existing underwriting platforms and tools.*" (Cignifi, 2012).

The Cignifi platform, in which the risk assessment is offered, allows financial institutions to qualify prospective clients and match them with a suitable financial product. The company also offers a "Response Score" which enables creditors to measure the probability that a consumer will both procure and use a particular financial product or service. Customer acquisition can then use this data to focus on attractive creditworthy customers with high response rates. Because Cignifi uses dynamic data, changes in customer behavior, creditworthiness and consumptive appetite are also updated every two to four weeks. This offers the opportunity of preemptively budget counseling or cross selling new products.

The scores cannot be reverse engineered to original data and the privacy of mobile-users is respected as no content-information is used. The CDR data is provided by telecom companies whilst protected by encryption and anonymization until consent is given. Old traditional credit scores require payment history over twelve months or longer to forecast credit behavior, Cignifi can compute a credit score with as little as the most recent month of call history. The platform had been tested in a commercial pilot in 2011 and proven successful by using their scores as significant indicators. Cignifi sought out the emerging Brazilian market in collaboration with the telecom provider Oi and the Inter-American Development Bank (IDB). Practically speaking, Cignifi's platform offered opportunities to financial institutions in enhancing existing loan portfolios by reducing overall default rate and activating more customers. Customer acquisition costs could also be greatly reduced by focusing on highly potent customers and filtering low-response individuals beforehand. The platform can be of use to mobile network operators by increasing the average revenue per client and loyalty due to the byproducts linked to them.

A similar pilot project was run in Ghana while collaborating with the World Savings and Retail Banking Institute (WSBI). The data suggested opportunities to promote financial inclusion of unbanked individuals in Ghana. Locally the two entities partnered up with Airtel Ghana and HFC Bank in order to obtain mobile data and improve the bank's underwriting process. Cignifi analyzed both transaction and savings data of HFC's accountholders and continued to place them in customer segments based on activity and balance. Then the CDR data of Airtel was taken for behavioral analysis to uncover insights based on a three-month period. Highly correlated variables were discovered in this endeavor and the article concluded that mobile data patterns are a highly suitable way to determine creditworthiness of the unbanked. Another application is to target high potential groups with tailored financial service or product offers. The client's profile is further constructed when more interaction takes place with the financial institutions. Usage of financial products will keep growing, creating a richer database along the way (Cignifi, 2014).

Refinancing Student Loans

SoFi, Earnest and CommonBond are examples of Fintech organizations that operate in a niche-market. The companies specialize in taking over student loans by refinancing governmental loans or providing personal loans to students. In a way these companies can also be seen as having a P2P lending system as investors and alumni provide the firm with capital. These companies pride themselves in using other types of cumulative data to judge creditworthiness like savings patterns, investments, education and career trajectory to produce a more representative potential rating. Most of the information and documents needed is uploaded by the consumer online for verification or access is granted to the source. Each client's data is run through a series of predictive analytics and algorithms to narrow in on people who show great financial responsibility and potential.

On their homepage, Earnest states that the many data points captured and analyzed allows them to reduce risk of fraud and default. This ultimately reduces the costs which allows them to offer low rates. This data exchange requires a strict security and privacy policy in order for clients to trust them. CommonBond takes the earning of clients' trust a step further by distinguishing themselves with a more thorough personalized customer relationship management. They try to establish a level of community through personalized application processes and even dinners organized between borrowers and the firm. The company also runs special social projects and donates to fund education in third world countries. This personal level builds loyalty early in the careers of students and young professionals.

These U.S. based Fintech organizations entered this niche as the outstanding student loan debts have quadrupled in the past decade and will continue to grow together with tuition fees. It is the highest consumer liability after mortgages in the U.S. (SoFi, 2015). Statistics show that the higher one's level of education, the lower the unemployment rate and the higher the median salary. Benefits that draw consumers to refinancing are saving money, lower monthly burdens, shorter loan terms, variable instead of fixed loan rate or vice versa and consolidation benefits like overview and consistency.

Peer-to-Peer Lending

P2P Lending platforms have been established as early as 2006 and have gained momentum in the past years. Traditional banks like Commerzbank are also launching their variation of these platforms (Main Funders) to take part in this innovative lending and investing model. The WEF (3.2.1) explained that these models have gained in popularity due to the various risk appetites that they satisfy through their platform. Examples of successful consumer P2P lending platforms are Funding Circle, Lending Club, Prosper and Upstart. A handful of P2P platforms is also active in the Netherlands.

The interesting part is how these platforms verify and test the creditworthiness of creditors and debtors. As it turns out, this does not differ much from the way traditional banks score credit. There are subtle differences, as Lending Club takes the FICO credit score of an individual and computes a rate based on the height of the loan and the motive. Their stated mission is to transform the banking system to make credit more affordable and investing more rewarding. The Lending Club partners with small and large national banks which utilize the technology and lending programs. The partners state that sound risk management is conducted, aligned with the bank's overall business strategy. Prosper's credit scoring system is similar and mainly based on external credit scoring. However, they do have their own proprietary rating system, a so-called Prosper Rating for easy internal evaluation after a client has built sufficient history.

The Funding Circle's credit scoring process starts by requesting a borrower to upload around five documents online. These are in the categories tax returns, bank statements, credit reports and additional data like bankruptcy history. These are the initial documents that indicate an individual's financial capacity in creditworthiness. To be eligible for a loan, they require a minimum 620 FICO score and at least 2 years of operating history. They claim to be more thorough in their credit scoring analysis.

Upstart uses FICO score and credit history in their analysis, but complements this with important detailed information in other areas. At first they require clients to have a minimum of 640 FICO score, but they will consider clients without sufficient credit history. They request documents to verify an individual's educational background, area of study and job history. Similar to the other firms, they also request the motive of the credit application. Similar to some of the student financers, Upstart also takes the individual's prospect into account and not merely the current financial state of a consumer. In short, these additional factors allow many consumers to obtain a credit, but traditional data still matters a lot.

Consumer Credit Scoring Services

ZestFinance, Big Data Scoring, Think Finance Inc. and Credit Karma are different specialists in credit scoring through alternative and non-traditional data. These three firms offer their services to consumers (B2C) or other financial institutions (B2B). Credit Karma is focused on informing and educating potential borrowers. After giving Credit Karma access to your data, the application generates two free credit scores (Equifax and Transunion). The distinguishing factor of their scoring model is that they are completely transparent in their various scoring parameters. A score is generated for payment history, age of credit history, credit card utilization and number of credit inquiries by third parties. The weight of each of these parameters and how you rank next to your peers in specific areas is also shown. The application shows you what is tracked and how you can improve which empowers you in influencing and building up your credit score. It has macro and micro overviews which can help you manage your budget and credit. Credit Karma also gives an overview of suitable financial products like credit cards, auto loans and insurances.

Big Data Scoring is a cloud based credit scoring service provider that collaborates with financial institutions and other lenders to improve the efficiency of the lending process. The accuracy is improved to lower the default rate and the chance that creditworthy applications are rejected. This is done by analyzing large data sets from many different sources such as social media, government databases and statistics offices. Details as device usage, web browsing behavior, habitation and location are analyzed through proprietary algorithms and included in the computation of creditworthiness. The results are provided to a lender and credit decisions are made on basis of it. A large case study was conducted with a Central European bank, active in over seven countries. The Big Data Scoring solution was implemented and the bank's loan underwriting process improved in accuracy. Initially there was an improvement of 14.7% for new clients, the models continued to improve and refine and were on course to reach a 26% improvement.

ZestFinance started out as a credit scoring endeavor to provide smarter loans to payday type loan borrowers. They use their own proprietary algorithms which are developed and refined through the use of Big Data and Machine Learning. They expanded to target near-prime borrowers as well. In order to build their credit histories they send data to major credit bureaus upon payment. This way they target the underserved with fair and transparent credit, whom usually make use of their services to consolidate multiple debts or medical expenses into a loan. ZestFinance doesn't look primarily at traditional credit scores but compares data from different sources to spot inconsistencies. Various data points are collected and analyzed in relation with one another. They argue that traditional credit-score systems use less than 50 data points to determine creditworthiness while there's much more public data available on any given person. This puts a lot of stress and dependence on the sparse amount of data that is available on clients, misjudging certain clients that are actually creditworthy. ZestFinance licenses their platform and credit scoring services to other lenders and financial institutions next to offering their own financial products. Their most recent large venture is with JD.com, one of China's largest online retailers.

Think Finance Inc. is a B2B online credit scoring software and service provider through their lending platform Cortex. The company is active in the consumer lending industry as their tools help manage and optimize loan portfolios. The company boasts fraud mitigation and compliance management through established credit policies. The platform includes web integration, decision engines and a loan management system built to facilitate online lending. The mother company also offers various financial products through ThinkCash, quid.co.uk and GreatPlainsLending.

Payday Loans

Some Fintech companies are specialized in providing short term payday loans at high rates while using Big Data to score creditworthiness. Examples of these firms are LendUp and Wonga. Some other before mentioned Fintech firms like ZestFinance and Think Finance Inc. have also combined their credit scoring activities with payday loans in a previous stage.

The startup LendUp is an organization focused on redefining the way underbanked consumers in the U.S. access financial services through credit and financial education. The company developed a proprietary platform to lower costs and to provide transparent credit-building opportunities for consumers with limited options in the traditional banking system. The first product which was introduced was called LendUp Ladder, a socially responsible loan. It differentiated from traditional loan products as it was designed to allow consumers to access short-term credit and build positive credit histories over time without ending up in structural debt. The following four steps are taken in order to achieve this.

(1) Customers that need more time to repay receive a reinforcing payment plan to aid them instead of penalties. (2) Loyal and dependent customers are given access to lower rates over time. (3) Positive payment behavior is reported to mainstream credit bureaus so that positive credit history is built up. Lastly, (4) LendUp believes financial education is imperative for the wellbeing of individuals. The company therefore offers free financial education courses through their channels. LendUp rewards clients that educate themselves by following these courses. Incentives are offered in the form of "creditworthiness points" which gives access to better rates for future loans at LendUp.

In the case study published by Databricks on LendUp, we learn that LendUp relies on improving credit models rapidly to algorithmically approve loans for their clients. Machine Learning based models are applied to a broad spectrum of data types. In this early stage they experienced a bottleneck which was caused by a single-machine based work approach. Feature extraction was performed on a single machine in Amazon Web Services (AWS). Feature extraction relied heavily on processing XML documents and these semi-structured documents required a long processing time. Furthermore, efficiency was lowered due to poor integration of data storage, feature extraction, modeling and analytical tools. They solved these issues by implementing instant Apache Spark clusters in an integrated workplace which improved productivity. Incremental improvements help make the models of LendUp more accurate and therefore allow more credit healthy people to be financially included with tailored products while default rates are lowered. Improving credit models requires extracting feature information from large data sets, training and evaluating Machine Learning models and analyzing the results by interrogating and visualizing data. This higher productivity rate means that more iterations can be run in shorter timespan which results in better credit models being created at a significantly faster pace (Databricks, 2016).

Wonga is a UK based Fintech firm which offers short term payday loans since 2007. They use fully automated risk processing technology based on alternative scoring methods to approve loans for clients in six countries on three continents. They offer reinvented short term and secure loans by giving applicants the options with regard to exact amount of credit they want to borrow and transparently showing a clear price based on the loan length. They require clients to provide them digitally with ID number, employment details, monthly income and expenses as well as bank account information. Afterwards, a decision is automatically generated as whether the loan will be approved or not. After this, a last verification step is conducted of the borrower's income to safeguard against fraud. Three of the most recent pay slips or bank statements are checked which should clearly show personal details and income of the last three months. The cash deposit afterwards is within minutes, making it easy to loan cash in a rapid fashion. However, these short-term loans do have high rates and therefore the process and the terms are communicated transparently and extensively beforehand.

Additional costs will be mounted when exceeding the loan period. Wonga will try to extend an olive branch and collaborate with the borrower to come to an agreement. Service fees will continue to mount for up until 90 days, after which the account may be passed on to an external collections partner. Moreover, credit bureaus will record the malpractice for several years and Wonga will also blacklist the client and ban them from lending from them again. In 2013, Wonga additionally expanded their activities by offering later bill payments through acquiring Billpay, a German online payment group. They are mostly still focused on providing emergency cash when needed for a maximum loan period of 30 days.

3.2.3 Financial Authorities & Government

This subsection treats the general opinion of financial authorities and the government with regard to the increasing use of Big Data in financial products and services. Technology is playing a more prominent role in the design, development and utility of these products. This causes a dependency on technologies, especially those which have a substantial influence on the lives of consumers. While the industry is convinced that Big Data can bring value to consumers by for example accurately assessing creditworthiness, there are two sides to every story. The opposition claims that data should not be collected and used with disrespect to their relevance and accuracy as the chances of misinformation and inconsistencies are too high. The current laws do not account for all the technological advancements made in the area of impactful data collection and analysis.

The White House

A report published by the White House (2016) on Big Data raises concerns that the use of Big Data might include the malicious potential of encoding discrimination in automated decisions. When this bias is inadvertently introduced to the data, it can have negative consequences for individuals and population segments. In specific to lending and scoring credit, this is dangerous when the technology is being used to deny whole low-income communities credit as they might end up in a downward spiral. On a more positive note, Big Data introduces the opportunity of fair automatized credit decision making and increasing financial inclusion of low-income communities. Public data is gathered by many major analytics companies that use data for process improvement and customization. However, other uses might be unwarranted, intentionally invasive or discriminatory through human bias. According to the authors, a principle of "*equal opportunity by design*" needs to be developed besides the ethical norm within computer science. It is a mistake to assume that Big Data is always unbiased due to the scale and algorithmic systems. Bias can come from (merely) using certain parameters within an algorithm and by over- or underweighing certain data used as inputs. Human bias and error can be passed through to an algorithm by wrong qualitative choices made in selecting data variables. This can also happen by using incomplete, incorrect or outdated data or by only gathering the data from strongly favored or biased groups. Perpetuation of bias can occur when bias is introduced within an iteration and repeats and reinforces itself.

Another problem with Big Data algorithms is the opaque nature of the mechanisms behind the decisions. The so called black box which inhibits the affected individuals in their ability to detect and seek correction of errors and bias when they occur. Flaws include poorly designed matching systems, overly narrowing personalization and recommendation services, correlation with reckless assumption of causation and lacking data sets that disproportionately map certain populations. The priority of the White House is to ensure that data is being used legally and ethically. The end goal should be to advance democratic principles such as equality and opportunity. With regard to using Big Data to determine creditworthiness, the White House realizes there is an important opportunity. Namely, the possibility to offer affordable credit to underserved Americans without a large credit history file. This gives the economy a stimulus while credit offers growth opportunities to the population, which would otherwise be impossible. Multiple sources of non-traditional data can construct a credit profile. For example, by including utility and telecom payments in the computation of creditworthiness, 70% of the unscorable population would be eligible and 64% would see improvement of scores. A challenge with using more data in general and unconventional data is that inaccuracies will become more likely with the diminishing ability to dispute these errors. The government believes there is a need for more regulation, testing and validation before deploying these algorithms to avoid discriminatory effects and inaccuracies.

The U.S. Department of the Treasury

In the whitepaper published by the U.S. Treasury Department (2016), the government reports on the evolving online lending marketplaces. The benefits and risks are highlighted of the increasing use of technology in the financial markets. The authors repeat previously mentioned dangers of using of Big Data and innovative modeling techniques to enhance their underwriting process. The applications hold a lot of promise but also carry a lot of risk. Data-driven algorithms may improve efficiency and effectiveness of credit assessments. However, they also risk leaving a disparate impact in society and might violate fair lending legislation due to discriminative algorithms. The opacity is one of the major problems as impacted individuals cannot object to these automatic decisions due to the low transparency. Traditional financial institutions need to be wary not to be over reliant on this technology.

Other risk points include that credit applicants lacking a large digital footprint be put at disadvantage and vulnerable customer segments may be susceptible to predatory lending practices. The use of social media data in the underwriting process may be damaging to the professional reputation of banks. Critics also note that Fintech firms are treated under different legal frameworks than the traditional financial institutions. While the default rate might seem under control now, the new models and underwriting tools remain untested through deteriorating economic conditions or even an economic recession.

The Federal Trade Commission

The analysis conducted by the Federal Trade Commission in the U.S. resulted in a report about the use of Big Data consumer data and its impact on low-income and underserved populations. Before writing this report, the FTC organized a public workshop which brought together stakeholders to discuss the potential of Big Data in alternative credit scoring using non-traditional data. This report was partially written on basis of the insights gathered during those workshops. The goal of this report was to give an overview of the benefits and risks of using Big Data analytics for consumer credit scoring.

Customized financial products and services are effectively matched with consumers and lower income communities can be better financially included. This would normally not be possible with scoring methods based on traditional banking data. The authors point out however that inaccurate and wrong use of Big Data can have detrimental effects like preemptive exclusion under discriminative factors. The article mentions that companies have to adhere to the FCRA so that accuracy is maximized in reports and transparency is improved. When deploying non-traditional data this can be complicated. Consumers should at all times be provided access so they can see their own information and correct it.

Additionally, the Federal Trade Commission Act prohibits unfair or deceptive practices. The act obligates companies to refrain from violating material promises to consumers in disclaimers. For instance, the terms of agreement stating that data will not be shared with third parties. There are certainly concerns of biased data being used and overstated within the analyses. To maximize the benefits and limit the harms of data the FTC recommends enterprises to critically revise the representativeness of the data sets and models being used. Inaccuracy, overreliance and bias are problems that need to be treated by devising company strategies to overcome them. It is only by acting on Big Data in this way that value can be generated ethically and correctly according to the law and standards (The Federal Trade Commission, 2016).

The National Consumer Law Center

The NCLC is arguably the biggest critic of using Big Data in determining consumer creditworthiness based on a report they published in 2014. In this study, they investigated various large “data broker” companies and seven loan products that use Big Data technology in their underwriting process. The largest critique points and objections are given on transparency and accuracy. The authors believe that these core values of consumer protection laws are nearly impossible to achieve when using Big Data. In a field test with fifteen volunteers, reports were obtained from some of the data brokers eBureau, ID Analytics, Intelius, Axciom and more. There was an astonishing amount of inaccuracies discovered ranging from wrong email to income and educational discrepancies or even mixed identities. These inaccuracies potentially play a great role in determining creditworthiness which is detrimental for both creditors and debtors. When evaluating the Big Data loan products, they discovered that many of the products offered identify themselves as a payday loan alternative. Analysis concluded that they are in fact better alternatives than payday loans, however the annual interest rates still varied from 134% till 749%. Most of these companies offering these products do not offer 90+ days for the repayment period or 30 days per $100 borrowed. These are characteristics of normal payday loans. Instead customers usually enter a period where they mount higher rates until the debt is paid or an arrangement is struck.

The issues previously mentioned by the other governmental organs are repeated in this paper once more. The opacity within the Big Data credit models doesn’t allow consumers to dispute their credit score. This is directly against the obligations of the FCRA which states that consumers have the right to know and influence the information contained in a credit file at the consumer reporting agency. This makes these products high risk. The lack of transparency within the Machine Learning models also allows these companies to potentially score with high discriminatory impact or even racial profiling. In conclusion the NCLC states that these new innovative loan products are not living up to their expectations and they are operating in a legal grey area. Some firms argue that a lot of the data variables tracked are merely used to detect fraud and not to determine creditworthiness, thus they should not be subjected to the same legislation. The NCLC therefore urges the government to sharpen their policies through a few suggestions including a mandatory registry for consumer reporting agencies so that consumers can know who has their data. They would also like to see more legislation that obliges Fintech organizations to guarantee certain thresholds of accuracy. Transparency must also be promoted, in order to allow clients to dispute wrongful data used within their scoring of creditworthiness (National Consumer Law Center, 2014).

3.3 Literary Insights on Research Questions

This section will focus on answering three of the initially established sub-research questions through insights gathered within the literature study. The answers to these questions will help establish the potential of Big Data in determining creditworthiness. This summary also serves to illustrate the prior requirements needed before an organization can deploy Big Data to generate credit scores. The following sub research questions were answered.

a) What type of data is relevant in establishing consumer creditworthiness?
b) How can the accuracy of consumer creditworthiness be determined?
c) What are the requirements in deploying Big Data to determine creditworthiness?

The literature has proven that non-traditional data variables are relevant in establishing the willingness of clients to repay. Multiple cases have shown that call detail records, utility bill payment, purchase patterns and so forth are all representable proxies for human behavior. Practical studies have shown that non-traditional data can be used stand-alone to make decisions on consumptive credit distribution when there is no credit history available. However, in large financial products that carry heavy burden, the greatest accuracy is required. In this case, an enhancement should be used of non-traditional alternative data in combination with traditional structured data. By deploying Machine Learning on Big Data, more relevant variables are discovered with the passing of time to enhance credit models. By using more data, a vaster overlap can be created of data variables that express the creditworthiness loyal to its definition. This can greatly reduce the risk in which financial products and services are distributed. In short, the potential of Big Data in establishing consumer creditworthiness is enormous and the various front end applications that can be made based on this are plentiful. Some examples will be discussed in chapter five, where the High Level Solutions will be explained.

Based on the literature provided in chapter three, we have sketched four rough levels of accuracy in which creditworthiness can be established. We can measure the accuracy of the underwriting process by looking at the relation between true results and total results within a confusion matrix. This can be tested by applying different processes and models on historical data and comparing the results of the dependent variables with the independent (target) variable, which is default rate in this case.

1. **Superficial** – Usually enough momentary data to grant a credit in a certain specific moment. Consists merely of hard, structured, non-dynamic financial information. Obtained through procurement, historical transaction data collection or form request.
2. **Adequate** – Consists of various datatypes including financial, demographic, personal and behavioral data to create a risk-graded snapshot of the creditworthiness of an individual.
3. **Accurate** – Transparent process of validated soft and hard data collection in order to predict risk of default. Enough data to proactively react in different modules to high-risk individuals.
4. **Extended** - Real time data on both willingness and capacity for payment extracted from soft and hard data. Mostly unstructured data is obtained to determine the willingness accurately obtained by interacting with the client possibly in collaboration with third parties.

On individual level, it is difficult to measure the accuracy of creditworthiness if not on a historical basis. On the aggregate level, models can be tested on sample populations in order to field test the accuracy which would be expressed in default rate and approved customers. Before field testing, historical testing is obliged to avoid premature risk.

The requirements to perform Big Data based credit scoring can be divided in three categories. Namely, Capacity, Compliance and Social-Economic (Societal) Acceptance. Different front-end applications that are based on Big Data credit scoring have different technical requirements and different stakeholder opinions. The basis in improving the credit scoring process lies in exploratory research through Machine Learning. Additional new data variables can be identified to improve the accuracy of credit models by analyzing existing data sets. New relevant data variables are subdivided in the components of creditworthiness: Ability to Repay, Willingness to Repay, Circumstantial Risk and Identity (to reduce fraud). Due to the scale of the data sets, Big Data tools such as Hadoop or NoSQL need to be used to lay the foundation of Big Data at enterprises where this is not the case yet. The correct infrastructure, tools and architecture need to be implemented to support a Big Data analytics program containing the right engineers and analysts with adequate skillsets. These are the minimal requirements in terms of capacity. In the design of a good credit scoring process, additional criteria were found in the literature. A list of these criteria was formulated in subsection 3.1.1. A list of Machine Learning techniques has been provided in Appendix J and the most widely used validation techniques are described in Appendix K.

On the matter of compliance, the legislation varies throughout the different countries and branches in which enterprises operate. The relatively mature U.S. legislation on the data-driven financial market had been analyzed, as this is where most of the relevant Fintech firms operate. There is an overall skeptical opinion on Big Data based credit scoring, as it reduces the transparency of the processes. It is in contradiction with certain laws that dictate the rights of consumers to dispute their credit score. This is to avoid the possibility of the credit score having a dispersive effect on the society. This is also known as the "disparate impact theory". Discriminative parameters on basis of sensitive personal information such as religion and race are forbidden. But the governmental authorities are in fear of whole low-income communities being denied credit due to unethical parameters or bias. This can lead to downward spirals and a reinforcing negative effect. Equal opportunity by design is a main point on the agenda of many legislators that allow Big Data based decision making such as credit scoring. Other authorities such as the NCLC have large doubts with regard to the accuracy of these methods. Also, there are many parts of the population lacking a digital footprint which puts them at disadvantage regarding non-traditional data.

The regulation on consumer data collection and use is different in many branches, countries and under certain conditions as collaboration or procurement. The same can be said of social-economic acceptance, as the tradeoff of added value in exchange for certain data is a deliberation that can vary for each and every individual. The "general opinion" on what is desirable differs throughout the many cultures and nations in the world. This is partially the reason why the European Union experiences difficulties in establishing union-wide laws on data privacy and use.

Now that we have established the potential of Big Data in credit scoring and the requirements prior to using this technology, we will start analyzing the concrete situation at NeoBank. In the next chapter we will discuss the practical qualitative research conducted to illustrate the Big Data maturity of NeoBank and relevant data-driven front-end applications using Big Data in the Fintech market.

4 Analysis & Results

This chapter will regard the analyses of the qualitative research conducted through semi-structured interviews on external and internal basis. Next to this, we consider the implications of the available academic and non-academic literature and prepare to answer research questions on basis of these results. These findings will be used as a basis to draft the high level solutions that are explained in chapter five.

4.1 Big Data Maturity of NeoBank

----------------------------------- **Access Restricted Due to Confidentiality** ------------------------------------

4.1.1 NeoBank-wide Assessment

------------------------------------ **Access Restricted Due to Confidentiality** ------------------------------------

4.1.2 Data Driven Analytics Assessment

------------------------------------ **Access Restricted Due to Confidentiality** ------------------------------------

4.1.3 Big Data Maturity in the Financial Sector

The maturity of the financial sector regarding Big Data has been established through a literature review. We note that there is a valid difference between most traditional banks and the avant-garde Fintech firms that use the latest technology. Due to this strong difference we felt a need to express this in the external assessment and the eventual verdict.

When the TDWI Big Data maturity model is applied to the financial sector as a whole, we can conclude the following for each of the dimensions based on the papers collected in an earlier conducted systematic literature review (Man, 2016). The Fintech firms are assessed in each second paragraph. Their evaluation is based on the articles gathered and analyzed in subsection 3.2.2. Additionally, some insights were gathered through a limited number of interviews conducted externally.

Organization: Most firms in the financial sector are aware of the availability of Big Data analytics. In general, the financial organizations include this technology in their strategic vision. However, concrete plans and funding are hard to achieve because of critics and slow penetration of new technologies. There are usually a number of individuals who are strongly in favor of Big Data implementation. These knowledgeable champions or promoters can provide organizations with leadership on achieving their own Big Data program. Usually the enthusiasm comes from a small group of local visionaries that implement some proof of concepts. A change in business culture is however not that simple, value and data driven culture has to be integrated slowly.

In this perspective, the Fintech organizations have a strong advantage as their organization is completely built around the use of such technology for their products and services. Rather, it can also be seen as their competitive advantage, what distinguishes them from the traditional firms. The culture is data-driven from the start and due to the relative small size coordination in value extraction from these projects is more efficient.

Infrastructure: The IT backbone is usually not entirely advanced at financial firms, there is a certain lack of expertise and data scientists are scarce. There are sometimes problems with the coherency of the IT architecture because of the prevailing legacy systems. This is problematic for the support of a Big Data initiative due to the incompatibility. Usually there is a specialized department that oversees Big Data and its infrastructure. Some form of specialized Big Data technology like Hadoop or NoSQL is usually already adopted to reinforce Big Data analytics.

With regard to Fintech organizations, the architecture is completely supportive towards the initiative of Big Data as most of the companies were founded with the premises of pursuing these ambitions. There is a strong coordinated effort and the company has decided beforehand what Big Data tools, software and hardware needs to be in place. The only varying part is how these firms deal with the expansion of their infrastructure under the rapid growth of their firms.

Data management: The management of data differs a lot between companies in the financial sector. However, the data is initially structured and usually obtained from within the company. More advanced applications like the market trade exchange require data collection and analytics in higher velocity. Data analytics is involved in a great number of projects, but certainly not all of them. Financial firms already consider data quality, integration and storage issues when managing their Big Data. Traditionally speaking financial institutions already have a lot of data and client data, but the degree in which this is utilized varies very much throughout the industry.

Fintech firms use structured, unstructured and semi-structured data for their appliances. For instance, some firms need vast amounts of data to efficiently apply Machine Learning algorithms to improve their risk models. Data is constantly gathered or procured by the firms for further processing and value generation. Some firms do not have the capital to store all the data but use cloud services like AWS to solve this.

Analytics: Firms in the financial sector mostly have a Big Data department or a few committees that are specialized in advanced data analytics. These departments have been set up after initial backing by management due to convincing pilot tests. Theory successfully tested in a proof of concept is funded but in most companies it has no concrete implications for the other departments besides the board. Analytics works in silos and the knowledge does not always end up where it can be fully exploited.

There are undoubtedly differences in the top of the food chain of Fintech organizations and starters when considering data analytics. Some are further progressed in their development of state of the art algorithms to accurately assess risk for insurances or predict default rates. Due to the amount of qualitative data scientists being scarce, the risk of having an asset poached is also high in the top of these industries.

Governance: Due to laws and legislation of late, the financial firm is forced to think of a good coherent data governance strategy to legally comply and support Big Data implementation. However, structure is very limited and policies are minimalistic. Companies are hopeful that innovation will spawn from the governmental demand to have data issues in order.

The papers treated in subsection 3.2.3 sketch an image of most Fintech companies operating in a grey area due to the limited legislation developed for these new technological applications. There are a lot of disputes and many governmental reports state that there are worries about the impact of Big Data decision making when it can be potentially erroneous, exploitative or abusive.

Overall Big Data Maturity: The financial sector operates slowly in their adoption of Big Data. Most of the large organizations are reserved in their use of these new technologies due to their conservative business culture and legacy systems. A lot of traditional financial institutes do not feel the immediate threat of Fintech organizations yet. Most companies do have Big Data included in their strategic plan, and intend to invest in this technology the coming years. Most companies are however far from data driven development and structurally delivering value out of collected and procured data.

Fintech organizations are at the top end of the maturity spectrum as they are the main users of Big Data and innovative technologies. These companies were founded and built around this technology. Due to the relative small size and flexible business structure, adoption and scaling of new hardware and software is easier. There is no old legacy architecture to which these firms are bound and t mindset within these companies is very progressive. Fintech organizations focus on a specific type of product or platform in which they satisfy the needs of their customers. They base their strategy, mission and vision on this unique added value that is offered to the market. However, not all companies know how to deal with the laws and legislation. This is because of the fact that the regulators and authorities are also in the process of regulating these new applications of vast amounts of data and differing interpretations of laws.

4.1.4 Final Assessment

The dispersion of Big Data maturity throughout the whole of a large corporation as NeoBank has forced us to make two internal assessments. One of the assessments is of the Data Driven Analytics department, which has the highest concentration of Big Data utility and awareness internally. It serves as an internal benchmark for NeoBank to strive for. This way we can also compare the internal maturity with the industry standard and compare it with the competing Fintech organizations.

EVALUATED PARTIES	**DIMENSIONS**					**TOTAL**
	Organization	**Infrastructure**	**Data Management**	**Analytics**	**Governance**	**Big Data Maturity**
NeoBank Company-wide	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
DDA Department	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Financial Sector	Early Adoption	Pre-Adoption	Early Adoption	Nascent	Pre-Adoption	Pre-Adoption (High)
Fintech Firms	Mature	Corporate Adoption(High)	Mature (High)	Mature	Chasm	Mature

Figure 10. Big Data Maturity Final Assessment

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

4.2 Traditional Credit Scoring at NeoBank

This section treats the credit scoring processes and models at NeoBank relevant to our case. Most credit scoring processes are set in motion when a client applies for a financial product or service. Depending on the request, different tests can be conducted by the bank in order to satisfy information requirements before extending product or service. NeoBank also initiates certain screening tests before accepting new clients to see if they adhere to certain healthy client conditions. These are the front-facing credit scoring processes. Behind these processes there are credit scoring models that are built in compliance with regulation in order to create buffers to deal with the default risk associated with the front-end activities. It is notable that most of the traditional credit scoring structure is similar to that of most large traditional banking institutes.

In order to map the structure of traditional credit scoring at NeoBank, interviews were held with five individuals related to the development of financial products or the models behind them. Two (former) product owners, an optimization manager, operational risk manager and credit modeler were interviewed. The results and the schematics made based on these interviews can be found below.

4.2.1 Credit Scoring Processes

Each financial product is linked with its own credit scoring processes to assess clients before distribution. These processes start when the applicant applies for a type of credit through the procedural channels. The processes of two tests that are relevant to our NeoBank case are further explained in this subsection.

Income and Expenses test

----------------------------------- **Access Restricted Due to Confidentiality** -----------------------------------

Automated Income test

----------------------------------- **Access Restricted Due to Confidentiality** -----------------------------------

4.2.2 Models within Credit Scoring

----------------------------------- **Access Restricted Due to Confidentiality** -----------------------------------

4.3 External Market Review

This section aims to offer an overview of the external market which was sketched roughly by conducting external qualitative interviews in combination with extensive market research. With the results of our research we hope to answer questions on how Fintech organizations and other traditional financial institutions determine consumer creditworthiness. Further elaboration is given on what requirements there are to implement Big Data into the scoring process and what potential data variables can be used. An additional subsection is dedicated to treat the legal impediments that are encountered.

4.3.1 Overview of Data-Driven Lenders

There are five different main causes in which current Fintech lenders use Big Data driven credit scoring processes. There is also a certain amount of overlap in some organizations that pursue multiple goals. These are the following.

- Improving financial inclusion in emerging countries
- (Re)financing student loans
- Providing payday loans
- Providing peer-to-peer lending
- Providing consumer credit scoring service

Full financial inclusion is one of the main drivers of the projects that use alternative sources of data instead of the traditional credit history and financial capacity tests. Providing formal access to financial services to the whole world population is a challenge which is not directly relevant to NeoBank in the Netherlands. The near whole of the Dutch population has access to financial services based on their credit records. Exceptions to this rule are expatriates and very young adults that make use of financial products that require extensive screening. Applications that requires similar data and fuel the investments in Big Data are (1) determining client receptivity towards marketing offers and (2) the capacity utilization rate of financial products. This results in offering personalized financial products, targeted marketing and aligned communications. It is expected that in order to achieve this, the risk department needs to enhance their collaboration with the marketing department. Front and back office need to work hand in hand to provide the backbone for client visible applications. The organization also needs to employ data specialists and invest in capable hardware for each department that can potentially generate significant value.

The refinancing of student loans in the Netherlands is out of the question as there is no market for this. The government offers student loans at superior low rates and there is no incentive to refinance as students are given a maximum of 35 years (previously 15) to pay back the loans at a constant low rate. However, financing students themselves is an interesting market in order to bind loyal customers to the firm. Various Fintech firms are now offering financial products while taking the potential creditworthiness of clients into consideration through Big Data analytics. Students and young adults are part of the Millennial consumer population. This indicates that in general they make use of new technology and generate a lot of alternative data while they lack formal credit history in comparison with full grown adults. However, the students are unlikely to apply for loans outside the governmental lending system. If they do, the chances are high that their limit has been reached and that it considers a high-risk individual.

The annual percentage rate of payday loans is much higher than other types of consumer credits, second only to pawn loans. Consumers usually contract a payday loan to pay for unexpected emergency expenses in a convenient manner. The loan is then paid off in the short term. The Fintech firms that grant these type of loans use automatized decision making systems to approve loans or reject loan applicants. These systems are needed because the underwriting process cannot cost the firm too much money. The decision systems do not necessarily take alternative data variables into consideration but some firms do make use of Big Data to optimize their decision algorithms. Most firms require you to fill out multiple online forms on employment, monthly expenses and personal details. Each country has a different application process due to the difference in legislation and standard data which is available to the organization through partnerships or public governmental data. In general, payday loans ("flitskredieten") are much less regulated and require less screening from companies. On these types of loans, there is different legislation than say a personal loan or other credit lines due to the nature and characteristics of the credit.

Peer-to-peer (P2P) lending is a constant feedback loop of multiple parties that make use of the exchange platform. Consumers, business owners and investors meet their offer and demand on this platform. Nearly all P2P lenders in the U.S. require borrowers to use a FICO credit score from a credit scoring agency to express their creditworthiness. Most P2P lending platforms in the U.S. work together with one of the large credit scoring agency. In other countries, P2P lenders focus mainly on bringing investors and small business owners together. They require the borrowers to upload a questionnaire, digital forms, personal credit history, ratings or credit scores to prove their creditworthiness. Additionally, these firms verify client identity by requesting digital documents such as ID card, salary slips, employer's declaration, bank statements and so forth. Based on the eventual calculated credit score, the borrower is placed in a risk category. The investor decides if the risk taken by investing is acceptable, bigger risks yield higher interests. Regulation and legislation on peer-to-peer lending are currently nearly nonexistent in the Netherlands, but heavily monitored by the AFM. There are a few P2P lenders active in the Dutch market, some of them have a small segment in which they carry these activities out next to their main operations in crowdfunding. Examples are Sameningeld B.V. (Mortgages), Geldvoorelkaar.nl, Lendico and Lendex.

Big Data based credit scoring service providers collaborate with financial institutions in their endeavor to reduce risk of default and score consumers accurately. The bank would contract these firms to improve the current credit scoring processes by enhancing them with alternative sources of data. These service providers offer cloud-based solutions to improve the underwriting process. "Common" alternative sources used by different branches are mobile data, social media data and utility payment data. However, public governmental data, data procured from statistics offices and third party data brokers are also used. Traditional credit scoring uses dozens of data variables and focuses on mapping a client's ability to repay while Big Data credit scoring uses thousands of data variables and maps the behavior of clients too.

One of the main questions the industry asks themselves is how to package and sell accurate creditworthiness to consumers. There is a need to convince the clients to allow access to their data for mutual benefit. In order to do so, consumers must see the added value in ceding this information. Bankers must oblige themselves to practice ethically in their process of determining accurate creditworthiness. If not done correctly and effectively, the control over money lending will be contested and eventually usurped by organizations drawing from other sources of authority. The government and financial authorities need to regulate "bad loans" and make sure clients are able to escape negative spirals. The objective of Fintech firms must be to guarantee customers better service through personalized loans with proper counseling. All the while business in the financial sector must be conducted with responsibility.

4.3.2 External Interview Insights

Attempts were made to conduct interviews with external parties to gather qualitative data on the market. This had led to limited success as it was heavily notable that most Fintech organizations operate with secrecy considering their used data points and algorithms. Employees that work with sensitive data are highly aware of the non-disclosure agreement they signed with their contractor. It is a reminder that they work with important and sensitive client data in a controversial area. But also an area in which firms operate in a highly competitive environment and distinguish themselves with their intellectual property in the form of used models and algorithms. This subsection treats the insights gathered from the external interviews in relation with the previous literature study conducted. These insights are used to produce scenarios in which Big Data scoring would be desirable with diverse front-end applications.

Out of the roughly 200 individuals approached from 45 financial firms and six universities (Appendix A), six individuals responded positively. Among the respondents were a data engineer, data scientist, analyst, software developer, account manager and business manager. Two were willing to answer some questions superficially in the chat client under the condition the data would be anonymized. Three were willing to do an interview through a video call. One respondent answered the questions and mailed them through a document. Most of the respondents were only willing to discuss the external market instead of their employer or limited themselves to public information considering their employer. An overview is made in this subsection of the most notable comments and insights achieved through these dialogues.

Predictive Behavioral Data Variables

Millennials, expatriates and exceptional thin-file clients have always been the most difficult groups to score credit in first world countries due to their lack of credit history. Too little traditional data would mean a heavier emphasis on the available data which can be incomplete or erroneous in the representation. Financial institutions have too little go to on and in this process decline financially healthy individuals. Another possibility is that they offer an unfit loan with either too little or too much credit. This translates in missed opportunities for banks and exposure to bad debt for consumers. Scoring methods based on psychological evaluation exist, but are unrealistic options on large scale due to the costly and inefficient nature compared with scoring methods using Big Data.

A concrete example of supplementing data would be mobile data procured through a data partnership with a telecommunications operator. Relevant data would be the amount of calls made, the time of call, to whom the calls are, the duration of calls, how often someone texts, who they are texting, texting at what time of the day, your location (cell tower). The prepaid top up rate, amount and consistency is also a strong indicator for prepaid clients. Postpaid clients could be gauged by using regularity of payment and phone bill size. The predictability is quite high and consistent as one can tell that phone usage behavior does not change that much over time. Despite this, historical data of many sequential months are used to map this behavior initially and it is constantly updated with new data. As an extra validation step on accuracy and timeliness, U.S. based Fintech firms enter collaboration with credit scoring agencies. To counteract manipulation by clients, extra steps are taken by testing through historical trends and peer validation. Receiver operating characteristics (ROCs maximize true positives and minimize false positives) and Gini coefficients are computed to test for accuracy and models are recalibrated regularly. Case studies in emerging markets using mobile data to score credit have shown default rates between 3% and 20% depending on loan size, business climate and risk taken. Higher risk will allow creditors to charge more interest and while low risk credit customers are a more stable source of income.

Many Fintech firms have taken into consideration that the newest and largest generation of millennials have come into adulthood with an immediate online presence, in contrary to previous generations. Practical applications already show that online browsing data, emails and social media data can enhance the accuracy in which creditworthiness can be expressed. Transaction data can also be stripped to indicate value, location, time and sometimes even the goods purchased. Focusing merely on certain data can overstate the representativeness of the data. For example, social media connections have low thresholds and are quite gameable whereas having someone's number and calling regularly is a more testimonial connection. Service providers are starting to use many of these alternative data sources to reinforce traditional credit scoring process. Some Fintech firms are convinced that it should eventually become industry standard practice to use non-traditional data additively to better detect fraud and improve accuracy. Used in conjunction, richer customer profiles can be created which aid in the understanding of their financial needs. Clients are given a bigger opportunity to lend custom financial products with a great reduction of default risk. By analyzing household accounts, banks can help in the identification of spending patterns. This allows customers to acquire insights on their financial habits and provides the bank with an overview of the customer's true chance in paying back loans. Furthermore, it could also help in identifying opportunities in future cross-selling and up-selling. However, large corporations are more conservative in their adaption of new solutions. They require more proof and reassurance that the solution is safe and that it has added value. Implementation is a time consuming process for large corporations and the stakes are much higher due to the large scale.

In one case, the new experimental data models have proven to reduce credit losses for banks substantially in lower income segments. Numbers vary from 20 to 50 percent with an increase of twice the application approval rate. The default rate is lowered and market penetration is increased for financial products offered through these data models. However positive this may seem for the financial institutions and the (low-income) consumers obtaining credit, there is a certain amount of controversy. The risk models and monitoring have to ensure that the credit granted to an individual can be paid back in a responsible way. Authorities are wary of over lending and aggressive loan pushing practices by banks.

Market Trends

Traditional banks and lenders are ethically obliged to provide a responsible and sound loan application process. However, they are unable to do this because most traditional financial institutions still rely on legacy systems of credit scoring. This is mostly due to the rigid organizational structure, the large size and the conservative culture of these financial firms. Despite the fact that financial institutions have more data at disposal than ever before, their Big Data maturity is lacking. Some large banks in the U.S. are already transferring servicing rights to other organizations specialized in treating high risk and default loans. Beforehand, detailed analyses are required of various data points that treat the history of the loan, the borrower and the loan itself.

One of the trends that data-driven organizations are following is customization and personalization of products and services. Authorities have high interests in governing the lending process to protect the interests of consumers. In their case, consumers should be allowed to contract an affordable loan when in need. If the rates and their creditworthiness do not allow this, a customized loan product offer should be available after a detailed assessment of the current and potential credit score. Financial education, transparency and client interaction might be a centralized topic in the coming years as the main concern of authorities lies at the interests of consumers. Clients are also becoming increasingly demanding and expect tools and services together with their financial products.

Most of the credit scoring in consumptive finance is becoming automatized with decision rules. Adding more non-traditional data is a good fit and can be done so by editing the credit model. The speed at which credit scores can be provided still heavily relies on the company, the operations, the used technology and the credit product itself. Analysts are also focused on building data models that predict the likelihood of loans becoming delinquent in order to proactively intervene in according way. A lot of funds are also invested in fraud detection. Scoring service providers are also thinking about working together with telecom providers, eCommerce firms and internet retailers next to banks and insurance companies to leverage non-traditional data to score client creditworthiness. In fact, this is possible because there is a need in these industries to predict the behavioral aspect of clients using data.

Fintech firms don't see banks as the main competitor but rather other Fintech peers, it is a rather specialized niche market. The competition between peers is enormous as firms are very secretive in respect to their Big Data operations. This causes a lot of redundant research and reinventing the wheel in terms of working credit scoring parameters, models and proprietary algorithms. Traditional banks score credit in certain authorized models and processes. In their opinion, banks won't change that legacy easily. Moreover, outsourcing this to other third party companies is out of the question as they need direct control of the risk they are willing to assume. A collaboration with a credit scoring service provider is not unthinkable, as it is the value proposition of some organizations. Data partnerships might especially become increasingly important when using alternative data to compute the creditworthiness. Companies in the financial sector might work together with retail in this case. Various markets are turning to personalization of the customer journey to improve service and quality of products. In this case the customer is identified, classified and approached in a certain way with custom products and services. This is currently already happening in simple forms within retail, but the financial market is also beginning.

European countries have quite evolved markets and there is a lot of conventional data at hand already of a large part of the consumer population. Big Data can be used to determine credit worthiness by analyzing and helping understand client behavior. This is similar to how credit scoring agencies profile an individual but also how investigators profile fraudsters or criminals. Using Machine Learning, data between credible individuals and credit risk individuals are mapped. Using this, a certain pattern can occur which machines can use as a basis for determining which one is which. This correlation is used to determine new data variables that might influence the outcome of a loan. When asked about an integral credit score, or a generic model that fits for all, respondents answered skeptically. Some Fintech organizations have experimented with the idea and discovered that both customer segments and product types are too different and varying to be able to make use of a one size fits all model. In this sense it makes sense that these products have their own creditworthiness requirements. One opinion is that financial products have a distinct effect on each individual, burden-wise and relief-wise, and this effect is not expressible in a simple integer. Accuracy is hard to determine the case deals with humans. Any person's behavior has a high tendency to change on critical situations. For example, an honest and credible individual suddenly was forced to take a loan by force for some emergency reasons. That individual might also take out a loan because he wants to obtain an MBA or own a house, which are different circumstances. There are situation and discrepancies to the rule that have to be coped with within the credit scoring process.

4.3.3 Legal Impediments

Financial organizations that use technology to determine creditworthiness are challenged by the regulation imposed on them. This subsection starts by highlighting the greatest legal obstacles in Big Data implementation that organizations face within the U.S. boundaries. These laws pose strict rules on how data can be used under which conditions and sets boundaries to what data may be collected and stored. In subsection 3.2.3 it was already concluded that the many laws in effect by U.S. jurisdiction inhibit the activities of data-driven Fintech organizations. Due to the relatively innovative nature of the market, new legislation is enacted and adjustments are made on past laws. This section aims to project this U.S. sentiment towards pending European legislation and the Dutch legal system, where data laws are still following up on the status quo of what is socially acceptable in the Dutch market and society.

Legislation in the U.S.

They United States have their own unique legal system and thorough legislation that protects consumer privacy and security. Some of the major governmental authorities and their standpoints were discussed previously in subsection 3.2.3. In short, there are five main objections to using Big Data driven credit scoring. These are the following.

- **Privacy and Security Concerns** – Companies that collect and analyze data on individuals might do this on unwarranted basis. These actions might be intentionally invasive as data might be used with malicious intent such as inciting consumers to take unfair loans. Vulnerable customer segments identified may be susceptible to predatory lending practices. Companies must refrain from violating material promises to consumers in disclaimers.
- **Dispersive Effects** – The authorities are afraid that much of the non-traditional data used could cause a bigger divide between the rich and poor. The fear is that low-income communities will fall further behind due to the inability to get a loan, which is caused by the new data. Moreover, credit applicants lacking a large digital footprint could be put at relative disadvantage
- **Lack of Accuracy** – It is noted in many reports that there are sporadic inconsistencies and misinformation in the collected and used data. Big Data usually solves this sensitivity issue by aggregating large amounts of data. But the potential consequences for individuals can be enormous, especially when making decisions on the border of a threshold.
- **Lack of Transparency** - There is a lack of transparency in the market due to proprietary models and algorithms. Authorities argue that the ability to glance over records and correct errors is imperative to a fair process, especially when inaccuracies in data collection are frequent.
- **Wrongful Discrimination** – Human bias is involved in the development phase of credit models and algorithms. It is wrong to think that decisions made by our systems are unbiased as there might not be “equal opportunity by design”. This means that certain wrongful parameters might be selected for use and others might be over- or under-weighed. Perpetuation of bias can occur.

There are aspects of legislation that influence the organization’s processes heavily like the obligation to anonymize personal data. An example is the Truth in Lending Act (TILA) and ongoing discussions with regulatory bodies like the CFPB. On the other hand, to weed out fraudsters whom work around and game the system requires thorough verification of identity by using various data variables. Especially when using alternative data in a transparent way, those who familiarize them with how the system works, might try to abuse it. In this case, historical data of longer periods are used to protect the system from ill intended.

The Equal Credit Opportunity Act makes it illegal for creditors to discriminate against applicants on basis of race, religion, national origin, sex, marital status, age, or receiving public assistance. The legal cases that have been brought under attention due to this act often convict human bias while making credit decisions. For this specific reason, one would think that algorithmic decision making would be an objective alternative to avoid this act altogether. However, in the new era, algorithmic decision making is based on Machine Learning and the way modelers program the algorithms. Some priorities, weighing, constraints or conditions might cause dispersity on basis of wrongful discrimination. Organizations might involuntarily use proxies for race, sex, indebtedness etc. within Big Data sets and then draw correlations and make conclusions that have discriminatory effects.

Legislation in the Netherlands (EU)

The Big Data legislation within the Netherlands is in the early stages of development and a constantly evolving topic. It is especially difficult to make a judgement in the area of client data usage for credit scoring in Fintech as there is lack of precedence in comparison with the U.S. market. It is important to see that much of these laws are European based and due to this standard, it is difficult to create and pass an overarching law which fits all the societies within the European Union. The current activity is based mostly on European guidelines instead of laws which are enforced. This is embodied in the Data Protection Directive (95/46/EC) on which most EU member states' privacy laws are currently based (European Parliament, 1995). Recently in 2016, the European Parliament has published the official texts of the new general data protection Regulation which overrides the guidelines of the 95/46/EC directive, these are replaced by the norms indicated in the Directive 2016/680 (European Parliament, 2016). This new European legislation must be transposed into each member states' national law by May 2018. Failure to comply can result in fines up to four percent of worldwide revenues. The new law mainly introduces the penalty system, obligatory reporting of data leaks, obligatory employment of data protection officers, and regular data protection impact assessments to ensure compliance.

Due to the difficulty of standardization and uncertainty in legal boundaries, large traditional banks maintain a conservative attitude in their use of these new technologies. In terms of innovation, the type of data used within the models is bound by known legal and regulatory frameworks. An example is location data, which cannot be used without requesting explicit permission of the client. The Dutch privacy law ("Wet Bescherming Persoonsgegevens Wbp") is based on the old Data Protection Directive (95/46/EC). It is the most confronting law which data oriented Fintech organizations encounter within the Netherlands. The Wbp dictates that that organizations must handle personal data of consumers carefully. This includes the collection, storage, processing, transfer and further use of this sensitive data. The law specifies the rights of the consumer to have insight in their data and the ability to correct it, similar to the FCRA. Before handling consumer data, permission must be granted through an agreement or it must be compatible with the original goal of a previous agreement. This also applies to reuse of previously collected data. There are however slight exceptions to the rule such as national security and delinquency investigation. The "Autoriteit Consument & Markt" (ACM) and "Autoriteit Persoonsgegevens" (AP) work together to regulate the Dutch market in the area of data privacy and security. They produce national guidelines on privacy and data usage to protect consumers.

Considering the fear of dispersity effects similar to the occurrence in the U.S., some respondents answered that it can also be envisioned the other way around. The more data you have on your target groups, the more accurate you can determine the creditworthiness of various people. Of course this only considers relevant data that is tested and validated thoroughly. Additionally, it opens up the market and increases financial inclusion to the benefit of many individuals who would otherwise not have access to credit due to shortage of historical data. Young underbanked population, individuals paid in cash, expatriates and all others previously discussed. This in turn allows them to build up more historical data to take a more active and valuable role in the society and economy. There are of course two sides of this story, some individuals on basis of the data simply do not deserve a credit due to previous malpractice and increased risk of default. This would create issues for the lender as well because "bankruptcy" causes shortage of cash and late paybacks cause the buffers of banks to be drained. The consumer would gain from being simply rejected instead of ending up in debt prison for substantial time and worsening credit score. The more enterprises know, the better they could help with sound judgement and argumentation. In cases of high default risk, a solution would be automatically personalizing the loan to satisfy the immediate needs of the client. At the same time this will prohibit them from burdening themselves with too much debt weight which would jeopardize the safety of their financial situation and creditworthiness.

There are however workarounds for many of the rules imposed on organizations. Data anonymization until access to decryption is granted by clients is one of them. Processing of prior anonymized data can be done to analyze the patterns of aggregated data sets and avoid harming privacy of individuals. Not transferring ownership of certain data also simplifies legislation. Arrangements and data partnerships can be made to make sure sensitive data doesn't leave its original environment. All conversion, analysis and further interaction takes place through cloud services. By operating on the data like this, no privacy and data security laws are violated. Each country's legislation differs, some countries including the Netherlands don't even allow interaction through the cloud. Because it considers personal data on consumers, the sensitivity of this kind of data does not allow it to leave the environment. They require a server on the premises or some kind of alternative assurance to the authorities that the data is protected.

In an interview with an internal legal expert on data driven credit scoring, it was concluded that the issues on privacy, security and dispersive effects would be limited within the Netherlands. Bigger issues are the remaining three factors; the lack of accuracy, the lack of transparency and wrongful discrimination. These directly infringe some of the consumer rights dictated by the Wbp, such as the right to have transparent insight within processes and the ability to correct erroneous data. Without transparency, it is also not possible to retrace the original purpose of the collected data. These are direct challenges within the process of using Big Data for credit scoring purposes.

On the other hand, the financial industry is under the assumption that in two years the European Union will implement PSD2. The data would then be open to the public. This overall availability will facilitate the collection and storage of data for further analysis purposes. This will allow organizations to use this data to further develop and perfect algorithms to further understand, accurately describe and ultimately predict human behavior. Of course this is an area of uncertainties and much can happen in the period before the directive is implemented. Most companies try not to let this political game influence their day to day operations. Strategically however, most companies have this scenario in the back of their mind. The same can be said for the enforcement of the general data protection regulation in the EU. It is uncertain if the regulation will lead to consistency in the whole EU. However, with the implementation of the regulation and directive, steps are taken to standardize data protection and punish violating parties

5 High Level Solutions

The purpose of this chapter is to describe the High Level Solutions for three Big Data based front-end applications envisioned which leverage the use of Big Data in establishing creditworthiness. In this chapter some scenarios will be presented in which NeoBank can implement the HLSs which satisfy the wishes of specific stakeholders. The results of the previous chapters will be used as input to establish context details and requirements on qualitative basis. This section is dedicated to drafting strategic solutions that carry many interests and fulfill a variety of needs in the market and of stakeholders. The strategic designs take the limitations of NeoBank into account by elaborating on constraints and assumptions. The feasibility is evaluated in an expert review. In the last chapter of this thesis, a recommendation will be given to indicate which course of action is the most suitable one for NeoBank.

The practical design question of this thesis is:

How can NeoBank make better use of Big Data in determining consumer creditworthiness?

In many scenarios, there are contradictory pressures that influence the boundaries of innovative credit scoring applications using Big Data. Market workings require the bank to offer competitive products and thus attractive loans with competitive specifications. The other pressure comes from the state which restrains the bank in its activities and heightens the requirements and obligations before granting credits. An example of a High Level Solution which gives both these stakeholders what they want is to lower the threshold of granting credits but heighten the degree of monitoring and budget counseling as to reduce the chances of structural debt. Internally, conservative stakeholders might be convinced by running a proof of concept on a small loan product. The problem which should be illustrated is the performance gap between two different models. It must be proven it could be done much more efficient when more data variables are used. This urgency can be illustrated by showing how the market is developing according to the scenarios of the WEF, and estimations by other Fintech and IT firms (Subsection 3.2.1, Disruption). They state that traditional intermediaries risk losing a substantial market share or even full replacement by Fintech firms on the long term if continued in the same trend. Regain of control of the market can be established by realizing transformation of own products and processes.

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

5.1 Machine Learning Basis

It can be stated upfront that Big Data can be used to improve three of the steps in which the credit scoring process is structured, these are tagged green in the figure below. Most prominently, the relevant data identification process can be enriched by discovering new relevant data variables through patterns in extreme large data sets by using Machine Learning. The newly discovered data variables and parameters can then be used to improve the models behind the data conversion. This is illustrated in the diagram of Appendix I. Data on individuals can then be used on large scale to map an accurate and rich client profile on financial behavior and ability to repay.

------------------------------------ **Paragraph Deleted Due to Confidentiality** ------------------------------------

The collection, storage and procurement of the required data is less of a contemporary issue due to the advancement in (Big Data) technology. The vast amount of data generated by connected clients and the availability of data partnerships facilitate this as well. The decision making process during customer appraisal can also be automated with the help of automatized decision making algorithms. These algorithms can also be enhanced by training them through supervised ML. This contributes to cost reduction and fast effective customer service.

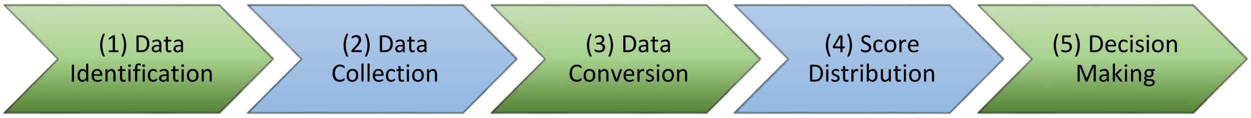


FIGURE 11. POTENTIAL IMPROVEMENT IN CREDIT SCORING STEPS

There are three approaches NeoBank can take in their implementation of Machine Learning. As shown in the external market review, there are possibilities in partnerships with “case proven” Fintech firms specialized in developing risk algorithms and data scoring processes using non-traditional data variables. The partnership can be on basis of (1) outsourcing function or (2) collaboration in the development and enrichment process. The latter option might be preferred as banks would not want to relinquish control of their risk processes. They might not even let external parties influence the amount of risk taken. The third option is that (3) NeoBank focuses on their own infrastructure and staff, by investing in their CoE DDA. Knowledge can be brought in by data science consulting firms and research is done internally.

------------------------------------ **Paragraph Deleted Due to Confidentiality** ------------------------------------

There are two concrete options when considering Machine Learning implementation at NeoBank. (1) Use ML to identify and provide the most relevant, efficient and predictive data variables for the credit model builders to enhance their statistical models. (2) Use ML to build, optimize and test the predictive risk and credit models directly. This second option also includes the training of automatic decision making algorithms, which should classify consumers accurately in risk categories which can determine aspects of the product such as maximum limit.

The literature research offered us a list of criteria which could rate the quality of a credit scoring process. When Machine Learning is applied to these criteria, there are positive and negative influences towards the quality of the scoring process as there are trade-offs depending on how ML is applied in the front-end.

- The **transparency** decreases substantially due to the ML concept which is called black-box learning. It is quite difficult to explain an algorithm and the decisions that are made through a model which is optimized by a programmed machine.

- In terms of prediction and modeling, ML has the potential to deliver a superior **accuracy** in comparison with traditional statistical models. However, this depends on the quality of the model and algorithm which can be heightened by using larger and varying training data sets.
- Due to impactful decisions being made with a model that learns from patterns within data, the justification is increasingly important. **Accountability** fails to be established at times with current models. With Machine Learning models it is expected to be worse due to causal relations fading.
- Machine Learning requires substantial data from large numbers of customers to succeed. In most cases the data cannot lawfully be gathered unless legal consent is given. Therefore, the **participation** increases in importance when using ML.
- The **fairness** of credit scoring processes must be guaranteed at all times. Objective decision making algorithms and credit scoring models are systematic and thus fair in their approach in judging individuals. However, the concept of “equal opportunity by design” must be complied to. This ensures that there are no dispersive or unlawful discriminatory variables used in the model.
- In order to implement Machine Learning, NeoBank must make sure that the process does not interfere with **legislative compliance** considering privacy and data laws. This includes making sure commercial use is initially ruled out and the points mentioned in subsection 4.3.3 are lived up to.
- By deploying a large number of data variables which are taken in consideration in the ML generated models, customer profiles will be more difficult to manipulate. Efforts can also be made to deploy ML for increased **fraud resistance**. This particular process is called anomaly detection.

Additionally, FICO also published a paper on the *Six-Point FICO Test* which can be used to validate the quality of newly discovered data variables through ML for credit models (Subsection 3.2.1).

This chapter will proceed by offering three High Level Solutions in addition to basis of using machine learning to determine creditworthiness more accurately. These HLSs leverage aspects of the advantages that Big Data driven models grant the organization. The solutions are sorted in level of disruption.

The HLS models are individually validated by expert review in a number of areas (Section 6.1). These are compliance to laws and overall feasibility. The societal desirability is treated in the stakeholder analyses. An initial conclusion had been made in these areas based on the results of the qualitative research and literature reviews. This review step is taken to increase the validity and the feasibility of the design solutions provided, before considering actual implementation.

5.2 Automatic Personalization and Client Appraisal

Scenario: In recent years, research has pointed out that in terms of financial inclusion, 99% of the adult (15+) population in the Netherlands owns a bank account. Of those people, 98% owns a debit card, of which 94% uses it to make payments. Of the people that own a bank account, 60% uses it to receive their structural loan and 63% uses their account to pay utility bills. Internet payment or web purchases were made by 68% of the adult population (The World Bank, 2014). These numbers show a steady increase compared with the previous percentages published in 2011. The statistics clearly indicate that the amount of detailed data generated by account transactions made by the Dutch population is only increasing due to the digitization. This means that a more qualitative client profile can be shaped by using this vast internal data at disposal to traditional banks such as NeoBank.

The World Bank database also shows us that 27% of the adult population borrowed money in the year 2014. However, a mere 12,6% borrowed from a financial institution, which is less than the half. The other 13,2% borrowed from either friends or family on informal basis. Less than a percent of the population made use of private informal lenders. It is notable that higher income individuals borrow more frequently. The fact that less than the half of the borrowers approaches a financial institution might be related to the bureaucracy and complex, needlessly long application processes. It may also be attributed to the fact that consumers doubt they will be granted a loan given their credit history. In any case there is room for improvement as NeoBank's role as a lender can be expanded.

One of the common goals of the financial authorities and NeoBank is financial inclusion of the Dutch population. In order to achieve improvement in this aspect within a saturated and well-connected market, a financial institution needs to provide loans when it is a necessity. The focus must be shifted to thin-file adults, expatriates and borderline creditworthy clients. This means that clients with less than excellent ratings also need custom financial products that can satisfy their needs. These products however need to be paid off in time. In order to achieve excellent market penetration, an accurate establishment of creditworthiness is imperative in order to score clients on a narrower basis. A possible solution is the use of data to enable personalization of products. The aim however is to also distribute products that make profit in order to guarantee the sustainability of the bank. This is why this HLS does not merely treat the personalization of financial products and services, but also the automatization of client appraisal to increase volumes. An indirect consequence of larger volumes of data streams is that more data is collected of interested market segments and the financial product itself.

----------------------------------- Text Deleted Due to Confidentiality -----------------------------------

Tracking, analyzing and computing information on clients on basis of Big Data can be a controversial topic. Especially when it considers applications for commercial use. It is however frowned upon by the authorities and data governors. Similar to the sentiment in the U.S., they are afraid that financial institutes might start exploiting susceptible low-income clients with bad loans. Controversy also arises when obtaining credit is perceived as "too easy", or the easy way out of trouble. There are also societal disputes on how long data can be stored, and how long a person can be accounted for their past in creditworthiness. Exclusion on basis of this data has a negative undertone, and is difficult to explain due to the complex model and algorithms. Personalization is also difficult due to different rates on account of historical data from the far past. Therefore, a simple, fast and transparent classification process is desired in risk profiles to reduce the impact of such harsh data and decisions. In terms of the qualitative criteria, this HLS would score well on participation, accuracy and fairness.

- **Main Objective HLS –** Increased market penetration through a streamlined and fully STP automated application processes of customized financial products and services. The process makes use of machine learning to train proprietary automatic decision making algorithms.
- **Expected Results** – More spontaneous borrowers, increases financial inclusion, better fit of customer and consumer product and thus increased utilization of limits, more consumptive expenditure in society, lower impact of default (EAD and LGD). Accuracy improves to tier three: "Accurate" (Section 3.3).

Goal Requirements	Product Requirements
• Increased customer satisfaction	• Machine Learning application
• Increased utilization rate of products	• Automatic decision-making algorithms
• Increase in number of applicants	• Individual based computation of risk
• Increase in low income clients	• Interdepartmental collaboration
• Increase in loyal young potentials	• External unstructured data (partnerships)
• Decrease in default rate	• Approval of internal management
• Decrease in churn rate	• Approval of financial authorities

Context Detail

- **Assumptions –** Technical requirements and legacy systems do not form an unsurmountable hurdle. The consumer market is interested in personalization. Senior clients are bondable.
- **Constraints** – Commercial image. Impossible to discern default due to personalized offer.
- **Dependencies –** Requires approval of the board, legislative authorities and the desire of consumers to disclose their personal data from various sources to allow personalization.
- **Scope** – Initially only implemented for consumptive finance products that have less extensive tests and requirements.

Key Stakeholder	Interests	Opposing Arguments	Impact of HLS	Influence over HLS
Financial Authorities	Consumer well-being. Balanced lenders market.	The initiative increases unnecessary lending. Incites high risk borrowers.	Low	High
Society	Consumer friendly and effective process focused on providing quality customer experience.	Infringement of privacy.	Low	Medium
Financial Institutions	Reduction of overhead costs. Increase of application influx due to lower threshold. More clients due to customized offers.	The current way works as well. Funding will not allow it. Too many systems linked with application process. Lowers "personal bond" with clients, scares seniors.	High	High
Low-class income individuals	Easier and faster application process. Lower income threshold due to custom offer.	Sensitive to impulsive lending which can be harmful.	High	Medium
Thin-file consumers	Lower threshold to apply for a credit.	Possible disagreement on models and algorithms on which creditworthiness is based.	Medium	Medium

Figure 12. Stakeholder Overview HLS 1

5.3 Budget Counseling

Scenario: One of the main stakeholder groups, the financial authorities, primarily want financial institutions to guarantee ethical procedures when offering financial products and services. One of the themes of the AFM is to protect clients from themselves (Appendix H). Predictive models can be very accurate in determining the payment potential of clients but this does not guarantee a sound and timely willingness to repay. Mistakes are made when granting credit to people that cannot effectively and responsibly manage their finances. In this case, the focus should be moved to improving the financial literacy of clients and budget counseling. High risk profiles should not be neglected and avoided, but confronted with heavier monitoring and frequent feedback. This scenario describes a prioritization of current resources in funding a loan product in which borrower education is the central topic. Other products won't have similar effects as the market is saturated with unsecured high-risk, high-interest payday and pawn loans. Prudent and educated consumers will know to avoid shark lenders, but there are still groups in the population who are receptive towards this, depending on the situation.

An aspect of this high level solution is the use of active continuous prediction models. The creditworthiness is tracked on frequent basis instead of determining a momentary state during the application. Contact can be established with the client and through this interaction, improved financial products can be offered and the liquidity of the client can be safeguarded and grown. Many respondents acknowledged that the trend exists of a desire to counsel, monitor and guide consumers better in their financial management and decision making. Creditworthiness is an integral concept of this process, as understanding of it allows financial advisors to cater to the specific needs of consumers. When the financial burden starts causing clients to default in small credits, it is expectable that high value products such as mortgages will follow in suit. Proactive measures with the client can lower the risk of downward spirals leading to such unfortunate events unfolding.

In most contemporary cases, traditional banks make a withdrawing motion when they deem a client unfit to bear the weight of a financial product. Authorities oblige financial institutions to be strict in their procedural application process. However, it might be too late to bind potentially wealthy clients when help cannot be offered to them when it is needed the most. By merely offering unsecured loans to borderline clients, the chances are high that the consumer will get into much more trouble than before without the proper guidance and financial education. Predictive technology has advanced very much, but the consequences of wrong prediction have to be constrained. Moreover, it is difficult to accurately indicate credit potential and link that with circumstantial financial situations of clients. Particular rare and grave instances such as divorce, death and unemployment are not representative for the rest of the population. These instances should be filtered and considered as outliers when constructing the credit and risk models.

This HLS argues that most financial problems causing default in consumptive finance can be prevented or solved by financial education and monitoring clients. The interactive element of this solution grants clients transparency in their credit score and an increased ability to influence it. This process involves them more in their development in financial standing. NeoBank can choose to implement this solution by offering more essential dashboard information, proactively providing financial advice and by providing incentives for clients to work on their creditworthiness. In terms of the qualitative criteria for credit scoring processes, this HLS is the most complete as it treats all the qualitative aspect except for legislative compliance. This criterion will be assessed in the validation step of this HLS.

- **Main Objective HLS –** A monitoring system is implemented in conjunction with balanced risk credit products in order to supervise the repayment process. Warnings are issued and adjustments are offered to the type of loan product. Interventions are made in timely fashion once possible default has been detected. The system aims to improve financial literacy by offering tools to budget and dashboard overviews to analyze their own spending patterns.
- **Expected Results** – Financial inclusion of more low-income consumers. Decrease in default rate through Dutch society. Less people that need to contract loans due to proper budgeting. Better prioritization of expenditure by consumers. Increase in lifelong loyal consumers bound to the bank. Accuracy improves to tier four: "Extended" (Section 3.3).

Goal Requirements	Product Requirements
• Increased financial literacy of clients • Increase of inquiries in • Decrease of churn rate • Increase of loyal customers • Increase of customer satisfaction • Decrease of prolonged default cases • Decrease of fully rejected clients.	• Machine Learning driven personalization • Web personalization algorithms • Recommender systems (ML) • Customer feedback on application • Data on customer preferences • Integration of BKR Data • Data proxies for income and behavior

Context Detail

- **Assumptions –** Under agreement, clients will honestly disclose all personal data of multiple bank accounts and financial standings within household or assets in possession. Senior clients will be able to operate this technology in the simple form of a mobile app or through channels such as personal financial advisors.
- **Constraints** – Client data run on limited data sets due to partitioning of organizational structure. Limitation in collaboration due to various unique financial products. A differing vision of how client creditworthiness should be represented. Client target groups differ by definition and there is a lot of legacy to work around as well as company processes are designed around it. Lack of information of overall financial health of household. Multiple accounts at different banks. Cash money and assets are not always declared and collateral is only part of one's funds. Limits imposed by anonymization and aggregated views of bank wide consumers.
- **Dependencies –** Management buy-in. Value of ML must be proven to decision makers. Conservative attitude must be conquered to renew and innovate through visualization and hard numbers. Desire in society must be demonstrable before it can be implemented. Consumers must disclose their data and allow tracking and monitoring within a disclaimer signed beforehand.
- **Scope** – Bank limited to supervising, educating and advising role. Final freedom of choice is with consumer. Obligations are eventually contractually imposed by bank.

Key Stakeholder	Interests	Opposing Arguments	Impact of HLS	Influence over HLS
Government	Protecting consumers. Reducing structural debt.	(Unnecessary) interference with Dutch households.	Low	High
Financial Authorities	Privacy right. Market sustainability. Transparency. Fairness. Progression.	Ease of manipulation.	Medium	High
High Risk Individuals	Financial inclusion. Customized offers. Lowered risk.	Handling rejection by black box approach.	High	Medium
Bank Conservatives	Preserving the status quo. Maintaining current feasible and concrete solutions. Avoiding unnecessary risk.	Difficulty in explaining algorithms. Difficulty in formulating causal relationships within data. Understanding your model vs. accuracy of the model.	High	High
Society	Decrease in default within population. Increased financial education. Heightened awareness.	Deliberation if it is culturally desirable in NL. Using data for commercial use. Privacy infringement. May be perceived as restrictions on financial freedom.	Medium	High
Millennial Consumer Population	Digitization. Comfort of products and services. Good customer experience	Overly invasive use of data. Privacy infringement.	Medium	High

FIGURE 13. STAKEHOLDER OVERVIEW HLS 2

Integral Interactive Scoring Overview

In continuation on the budget counseling topic, an integral interactive scoring overview is suggested. The bank lacks an integral overview of creditworthiness as not all systems internally are connected to see if a household has multiple financial products. The bank would make use of a creditworthiness portfolio of client – Every time a client applies for a financial product, the creditworthiness is evaluated in some aspects. These aspects can be tracked and taken into account while monitoring or applying for larger and future products and services.

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

Both the department of consumptive finance and mortgages conduct research on clients to discover parameters that indicate customer default. They do this partially together due to the correlation between default of loans and inability to pay mortgage. The data is used to distinguish between high risk profiles and low risk "safe" clients. In spite of collaboration on research efforts, both departments still build their own credit model and use their own tests for their financial product to determine creditworthiness.

Mockup: A visual prototype has been sketched which can be seen in the illustration below. As stated in the main objective and the scenario of this HLS, the interactive aspect of building credit history adds to the financial education of clients. It is important for clients to comprehend the basic components of their credit score, and how it can be influenced. The bank needs to communicate transparently and stay connected with their client. Frequent updates on creditworthiness is one way to do this while increasing the awareness of clients. A dashboard overview also helps clients to financially budget in fixed periods.

The client is offered valuable insight in their own spending patterns and actions. A program or mobile application like this can improve the comprehensibility of intangible assets such as creditworthiness. Moreover, an incentive is offered to clients to improve their credit score in better personalized offers. The consumer needs to know that the bank can help as a partner in finance instead of being traditionally marketed as a means to achieve goals.

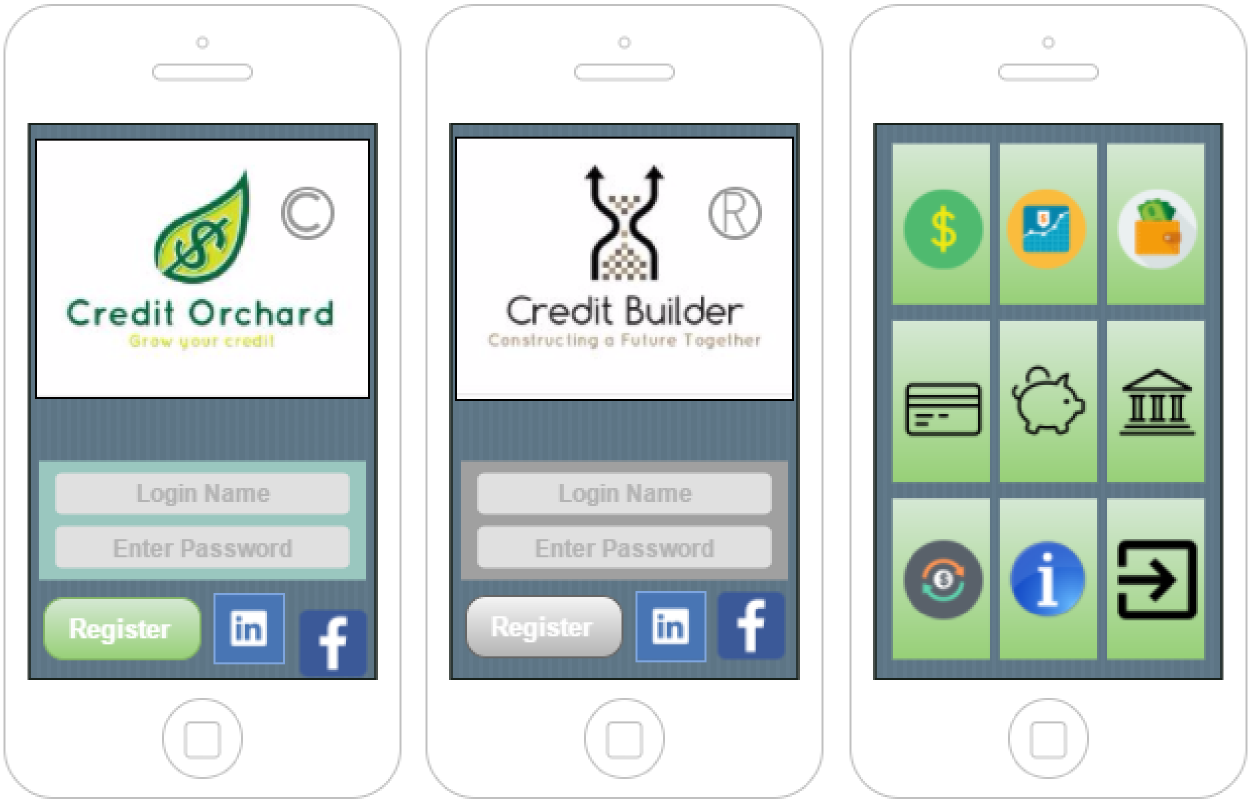
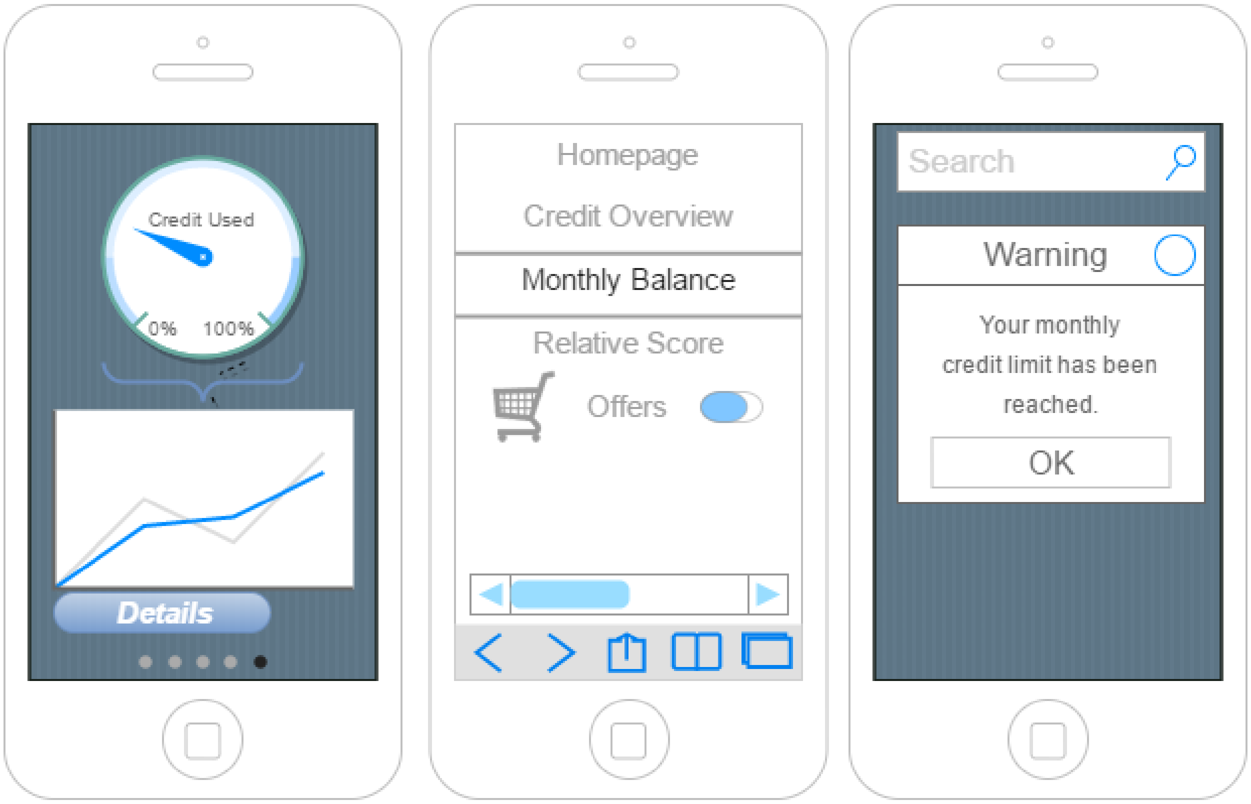


FIGURE 14. MOCKUP OF DATA DRIVEN APP – LOGIN & NAVIGATION PAGE



Your monthly credit limit has been reached.

FIGURE 15. MOCKUP OF DATA DRIVEN APP – VIEWS

The above images show a mockup of an interactive, educative and transparent application for financial products. Overviews can be generated of the client's outstanding credit, balances and existing accounts. If these do not exist yet, they can apply for them through an automated procedure in which their computed score is updated. Rejected clients can be offered a personalized credit product with a lower risk of default, to prevent churn to underwriters of riskier loans. A unified integral credit score is used in the application, which tracks consistent payment behavior. Warnings can be issued when limits are close to being reached or (personal) obligations are not met in time. A relative score can be generated of creditworthiness in comparison with all other users (anonymous). Additional data can be mounted through registering through Facebook or LinkedIn. Bank account data can be linked to generate overviews of monthly expenditure and to indicate patterns and trends. Additional detailed information can be offered on the client's credit score and how it has been affected in the past period. Recommendations can be given on how to improve the credit score. Of course the application itself will not be made in this specific form, but the mockup serves to indicate the idea which could reach a large part of the digitizing consumer population and their wishes.

5.4 Collaboration and Standardization (IOIS)

Scenario: Ideally speaking the bank would be able to perform an analysis with all the available client data in order to generate an enriched client profile which allows to make credit decisions easily. This requires the constant procurement or generation of relevant and accurate data. The only collaboration that is currently in effect is the national institute "Bureau Krediet Registratie in Tiel". A collective effort amounting in a centralized databank would require the BKR to further this exchange of banking information between financial institutions. All the banks are interlinked with the national central organ to register all loans, payment behavior and misconduct of past years of clients. There are also obligations to verify if potential clients are not on blacklists which prohibit banks to finance them irrespectively towards their creditworthiness. BKR is usually at the start of any application process for a financial product or service. Another opportunity for increased collaboration would be PSD2, which contributes by opening access to all transaction data.

When approving or rejecting a credit, at the moment of application the bank usually does not have a complete data picture of the client. Therefore, requests are sent to the client to provide the bank with a complete data picture, unless the individual is already a client at merely that one bank. In which case the bank can use historical data. Other data on clients will have to be procured and managed. This High Level Solution suggests that a collaborative bank data system should be created to load In the transaction data of the main bank account when the customer has an account at a different bank. This Inter-Organizational Information System (IOIS) improves transparency enormously, but it has many constraints and dependencies. The system would affect the entire financial sector within the Netherlands.

One of the reasons that the financial market has a critical attitude towards mutual collaboration is the many difficulties in determining who takes the lead, who finances the general system, how and what data is channeled to each other and many other issues that come into play. The question is if the financial market is ready for such an abundant and extensive collaborative effort using data. Most large banks are satisfied with their current position and would not want to give up their competitive advantage or take unnecessary risks. Culturally and legally speaking there are also objections when turning to data ownership and privacy issues. However, financial authorities also have an interest in this data centralization as this facilitates many of their auditing processes.

In the ideal scenario in which the data is fully disclosed and shared between all banks. The databank allows for collaboration in the development of standardized variables and accurate credit risk models as well. The incentive to develop such a standard model would however be contradictive with collaboration as proprietary models are currently the competitive advantage and distinguishing factor of banks. In terms of standardization there is also a lot of room for improvement in the Netherlands in comparison to the U.S. model. The possibility should be researched how to implement a standardized scoring system (3S). This scenario also suggests the Netherlands adapts to the credit scoring standards that are used in the U.S. Instead of scoring credit every time when a consumer applies for a credit, banks can mutually invest in a standardized credit score to sell to clients. In this case, all financial institutions would agree to certain basic compatibility models of creditworthiness, which allow them to take over each other's credit reports. The goal is to strive in improving the accuracy of consumer creditworthiness and eradicate fraudulent practices. In terms of the qualitative criterions mentioned in 3.1.1, this HLS focuses on improving transparency, accuracy, legislative compliance by facilitating audit and fraud resistance.

- **Main Objective HLS –** Construct a centralized horizontal inter-organizational information system which allows banks to mutually share data and knowledge on clients, portfolios and credit-risk models. This electronic data interchange allows banks to transfer and obtain data of clients freely. Standardized compatible credit score which allows banks to take over each other's credit reports for increased efficient utility and decreased labor and redundant testing. Allows for better forecasting of creditworthiness development in clients.
- **Expected Results** – Common collective efforts to stimulate economy and focus on financially serving clients to their best needs. Reduction in overhead and bank-wide storage and hardware costs, economies of scale. Increased resilience and flexibility in missing data, once adopted it creates a decreased risk level. Increased insight and oversight by financial authorities to detect possible pending crises in society. Exchange of technology and a collective effort towards advancement in multiple efforts such as security, modeling, consumer fraud detection etc. Accuracy improves to tier four: "Extended" (Section 3.3).

Goal Requirements	Product Requirements
• Nationwide implementation of overarching database system • Increased collaboration and innovation • Increased level of available client data • Clean high quality data recorded in structured and enhanced fashion • Future company processes built on IOIS • Break through initial investment barrier by dividing risk among participants • Break through company data silos • Increased comprehensibility of data scoring due to standardization	• Low adoption threshold • Compatibility with existing data systems • Compatibility of software • Compatibility of policies and procedures • Compatibility with existing credit models • Transaction cost economy for maintenance of IOIS • Initial distribution of costs for all participants vs Buy-in for later partakers • Minimal IT threshold of system members • Minimal network security levels • Receptive to data conversion

Context Detail

- **Assumptions –** Financial authorities and clients will allow the interchange of client data due to the beneficial nature of applications. Banks will not use the data to exploit clients. Consumers buy in to the idea that centralized collective data is safer and will lead to client benefits.
- **Constraints** – Funding, and distribution of costs to implement and maintain. Requires collaboration from all major parties or its effect will diminish substantially. Difficult to verify if collaborating parties have shared all data at disposal. High levels of complexity lead to low comprehensibility and causes parties to lose interest. Adoption of an IOIS for large rigid enterprises is huge commitment, there might be integration issues.
- **Dependencies –** Enough incentive for banks to buy in the concept, highly dependent on resources and funding. Benefits must outweigh the many fortunes invested in current standards and systems. Engineers and data scientists would have to rely on each other for collaboration. Banks have to act in good faith and avoid becoming suspicious of competitors withholding data. Benefits have to be measurable or at least visible to convince banks. Banks have to agree on long-term data partnerships with each other which includes substantial investments.
- **Scope** – Limited to consumer client data, exempting business clients which have more objection to all banks having insight in their financials. Also, the standardization only applies to the Netherlands and not the European Union as there are too many different systems, financial products and models which are used. Due to the scale of the project and the possible consequences, a trial should be run first before full commitment.

Key Stakeholder	Interests	Opposing Arguments	Impact of HLS	Influence over HLS
Government	Protecting the society and its consumers.	Security of the centralized data system. Lay-off in overhead due to technology. Feasibility of nation-wide standard.	Medium	Medium
Financial Authorities	Increased oversight through technology. Ease of auditing due to standardization. Market sustainability , transparency fairness, innovation.	No free lunch; there is not one model that works best for all possible situation. Unlevelled playing field for new financial entrants. Danger of concentration of power. Danger in exploitation of data.	High	High
Consumers	Centralized Information. Ease of transfer.	Privacy infringement.	Medium	Low
Dutch Association of Banks	Increased access to data enabling innovation. Decreased costs in overhead, IT infrastructure and storage. Functional benefit. Increased effectivity, efficiency and flexibility of data.	Banks reluctant to give up their current advantage in data assets. Costly nation-wide collaborative project which is risky and could fail. Collaboration on less involving terms desired. True benefits yet to be proven through ROI. Banks reluctant to give up prior investments in risk models and database systems.	High	High

FIGURE 16. STAKEHOLDER OVERVIEW HLS 3

Envisioned IOIS Architecture

A common draft of the envisioned IOIS Architecture is sketched in the illustration on the next page. The centralized database is created to pool information resources. Knowledge sharing on national level primarily serves to facilitate the second step in the credit scoring process: Data Collection. Auditors and the government can access the data as well for certain tasks through specific authorization and legal basis. The data is accessed through a secure overarching database management system. The trusted environment and the data can only be accessed by authorized parties such as member banks. As stated earlier the IOIS consists mainly of a database system that facilitates electronic data exchange of client data for banks. This data is used to collaborate on accurate credit risk models and algorithms that better describe human behavior and to identify relevant data variables. Furthermore, the data can be used for joint operations on detecting and reducing client fraud, and to innovate on certain data intensive client applications. By sharing data on frequent basis the risk is avoided of using data which has lost relevance through the aging process.

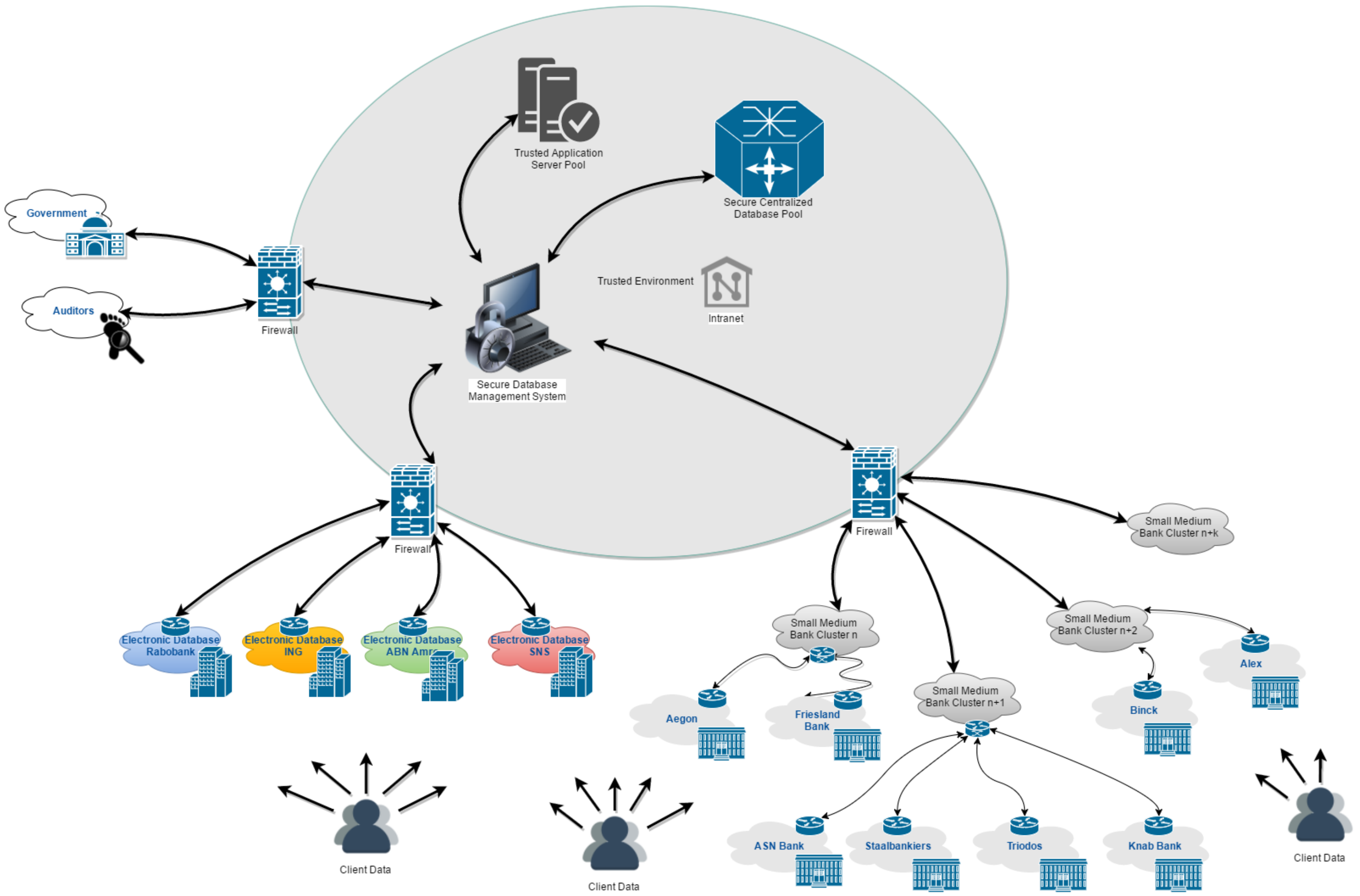


FIGURE 17. ROUGH ARCHITECTURE SKETCH IOIS

6 Conclusion & Discussion

In recent years, the use of Big Data technology has increased remarkably in many industries due to the increase in accessibility. Storage and computing power to conduct advanced analytics have become affordable and many optimization projects have proven themselves successful. This design research project can be roughly divided into four main pieces of work. Initially, the literature has been explored to find theoretical background to support the hypothesis that Big Data can be used to better determine consumer creditworthiness. Both practical and scientific theory suggest that the use of Big Data analytics has an inherent high potential in improving contemporary credit scoring processes. This is partially due to the predictive and modelling nature of computing a credit score. The underwriter is essentially forecasting the likeliness that a loan will be repaid. Within this prediction of creditworthiness, certain components must be taken into account. Some of which are structured like spendable income and some which are more complex and dependent such as behavior and preference. Various case studies have shown that alternative non-traditional data can be highly predictive of the behavioral aspect of creditworthiness, such as mobile data and social media data. This is a vast improvement of the traditional credit score cards which widely use statistical techniques to compute credit scores that are mainly based on the payment capacity. These practices can still be found at most traditional banks.

The external market research had been conducted to create an overview of the data-driven innovations within the financial lending sector. It was notable that many Fintech firms are using data-driven technology to gain a competitive advantage, and act in secrecy to protect their intellectual property. Especially Fintech organizations in the U.S. focus on incorporating Big Data in their credit scoring process. The credit scoring agency FICO are focused more on the use of data variables that express the willingness to repay. FICO published the Six-Point Test on research basis in order to judge new data variables on their quality to be used in predictive credit and risk models. Fintech organizations have business models and goals such as financial inclusion of emerging markets, refinancing student loans or credit scoring services. In general, what makes them successful is an alternative proprietary way to score creditworthiness.

Internally, the Big Data maturity of NeoBank had been determined to estimate the potential in adopting new applications of the technology in processes. There were requirements to adopt Big Data analytics and eventually Machine Learning to perform optimization of credit scoring. These were divided in three categories: Capacity, Compliance and Social-Economic Acceptance.

------------------------------------ Text Deleted Due to Confidentiality ------------------------------------

An efficient way to predict this sentiment for Europe, was to look at the developments of the matured U.S. market. This indicates the primary fears of the society and government which include opacity, consumer rights to dispute, disparate impact theory, equal opportunity by design, unlawful discriminative parameters and even basic doubts on the accuracy. Finally, the Social-Economic-Acceptance differs per individual and can be wavered by allowing clients to disclose their data by choice in exchange for functional benefits.

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

Through the use of observations and semi structured interviews, the internal credit scoring process had also been mapped. Similar to many traditional banks, the credit scoring process at NeoBank resembles five structured steps in determining consumer creditworthiness. These are (1) Data Identification, (2) Data

Collection, (3) Data Conversion, (4) Score Distribution and (5) Decision Making. It had been made explicit that at NeoBank, Machine Learning could be adopted to improve the processes in steps one, three and five.

------------------------------------ Text Deleted Due to Confidentiality ------------------------------------

The current income test has experienced a few problems which are believed to be solved with the application of Machine Learning; The AFM deems the current test to be insufficiently accurate due to the pure measurement of income, and there is fear of narrowing legislation by other financial authorities. There is however reason for fear as machine learning models and algorithms are difficult to justify and find causal relations (black box). In order to minimally affect the consumer friendly application process, ML can be used with (public) non-traditional data to detect predictive data variables which can help model client creditworthiness with the available data. Lack of available or disclosed data regarding the actual payment capacity could be replaced or enhanced with predictive proxies such as job function, employer, duration of employment etc. This model can then be used to make an accurate estimation of the credit limit. This limit can be revisited every month and estimated more accurately to increase the fit on used capital which decreases the height of the buffers needed. ML can also be used to detect fraud with anomaly detection of data sets.

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

On ethical basis, consumers must be able to protect themselves from bias and inaccuracies from the data acquisition phase. This means that transparency is one of the biggest requirements, together with participation. Accurate credit scores are obviously not synonymous to fair credit, as correlations can lead to a credit extension policy that unintentionally harm large groups of applicants causing them to fall further behind (disparate impact theory). Exploitation of this technology can occur by targeting susceptible clients to predatory lending schemes. This leads us to the point where we have to choose in each situation if the trade-off between model comprehensibility and prediction accuracy is a correct one.

Focusing solely on non-traditional data variables has been successful in pilots in emerging countries. However, this was done due to the complete lack of a credit history of many clients there. As most clients in the western world have records, the method should be used to enhance the existing credit models to better predict the human behavioral aspect of creditworthiness. This aspect is becoming increasingly important as many financial organizations are adopting dynamic data in the developing market. Market research by the WEF predicts that traditional intermediaries would lose large portions of their market share to Fintech organizations continuing on the same trend.

------------------------------------ Text Deleted Due to Confidentiality ------------------------------------

Some supporters caution against regulating online lending too strictly and too early. They fear this would cut off innovation that could provide valuable financial products to individuals and small business owners. On the other hand, there is a major concern that when left unregulated, online lending at the rate that it is growing might lead us to another financial crisis. A deliberation exists on what levels regulation should be implemented by who. Authorities also have yet to decide how much transparency should be required in a market where the distinguishing factor is part of these sensitive algorithms and models.

On the matter of ethics of using Big Data within the lending process, we need to distinguish between the goals of the creditor and the goals of the debtor and society. The creditor’s main interests are accurate data, profit and compliance. The goal within the society is to create a financial system that does not worsen current inequalities or lead consumers into debt spirals. Most of the financial institutions have profit and survival at its core as an organization; the latter goal pertains to the society, but also the government and financial authorities.

6.1 Validation

This section will treat the validation of the core solution explained by the ML Improvement Process for Credit Scoring (Appendix I) and the three other High Level Solutions of chapter five. The validation has been parts on legislative and overall feasibility. The validation was conducted by consulting internal experts in each field to review each case. Other validation methods were unfit or not possible due to the circumstances and characteristics of this project.

Machine Learning Basis

On legislative front, implementing Machine Learning in credit scoring faces many challenges due to the failure to immediately comply to some of the previously mentioned components of the Dutch Privacy Law (Wbp, Subsection 4.3.3). Machine Learning increases the opacity of credit scoring processes by introducing the black-box learning concept. NeoBank must provide clear explanations to clients which is complicated when complex algorithms are involved. The internal legal expert mentioned that article 14 of the GDPR states that NeoBank is obliged to inform data subjects on "the existence of automated decision making". This includes profiling (Article 22, Paragrah 1 and 4) and at least in those cases, meaningful information about the logic involved as well as the significance and the envisioned consequences of such processing for data subjects. Only then the data subjects would be able to make an informed decision on whether to give consent or not for their data to be used for certain benefits. From a privacy perspective, consent given by data subjects is only valid when the data subject is fully informed of the implications at the time of agreement. The data subject must have the possibility to withdraw its consent and consent should be freely given at all times. Article 42 of the Wbp indicates that it is prohibited to automatically process data if this can have legal consequences for individuals without human intermediation. This article is based on article 22 of the GDPR. It is therefore of importance that within automatized processes, there is a Human-on-the-loop to verify and intermediate.

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

Automatic Personalization and Client Appraisal

The automation of the client appraisal and personalization process of financial products is an ongoing process with different levels of autonomy. Client appraisal is a sensitive area as financial market authorities are strict in their legislation on screening and loan incitement. Moreover, consumer data authorities won't allow the commercial use of collected transaction data currently held by the bank. This leads to the point where consumers must agree beforehand that their extensive data would be used for credit scoring and financial product specification. There are limited possibilities in personalization beforehand due to the juridical requirements as the privacy is a fundamental right in the Dutch constitution. However, this consent can be implemented in the client appraisal process by informing the client well during the application process of the concerned financial product. In this case, previously collected data for different purposes such as transaction data can be used with the rightful consent. It is however difficult to say if the previously collected and available data would be enough for useful thorough personalization. More data could be requested by the bank to enhance the process, but this would impede the speed and the automatized aspect of the process.

Budget Counseling

------------------------------------ Paragraph Deleted Due to Confidentiality ------------------------------------

Especially the real-time analyses of thousands of clients would require better technical specifications and larger scale infrastructure. The previously mentioned legal constraints for the Machine Learning basis also apply here such as the Wdp and GDPR.

Collaboration and Standardization (IOIS)

One of the main legal objections of this HLS is the antitrust act. It is prohibited for competitors to share sensitive commercial information in which clients could be exploited. This is to avoid the forming of cartels through the available information and collaboration. Competitors must each independently form their own commercial strategy and must not do so in deliberation with each other. This would mean that the IOIS must prevent direct and indirect commercial sensitive data from being exchanged. Examples are which financial products are purchased most under which specifications. This can be solved by anonymizing the source of the data so that competitors remain anonymous.

The overall feasibility of this High Level Solution is the lowest due to the fact that multiple critical stakeholders are involved with their own separate interests. Especially the amount of banks agreeing to such a centralized system would need to be enormous, and therefore the need for such a system would have to be greater. If any one of the big national banks would disagree to support or fund such a project, it would already lack the scale to be efficient. Moreover, the differences in prior investments in data analytics and Big Data are too large. It is likely that large banks would disagree and small banks would not be able to help fund such as system or reorganize their internal processes to be compatible. The interesting part of this HLS is the difference which is accentuated between the U.S. market and the Dutch and European markets which handle data in a very different way. This is due to the different legal frameworks, societal norms and values and age-old standardization which has become an industry itself in the form of many data brokers and credit scoring agencies. This is the legacy on which traditional banks and new underwriters build their credit scoring processes around. It is difficult to envision such a landscape in the Netherlands, yet the market wants to bring and adapt these technologies as well. The compatibility of these technologies has yet to be proven and its effects will be seen in the coming years.

6.2 Recommendations

------------------------------------ Access Restricted Due to Confidentiality ------------------------------------

Big Data Maturity

------------------------------------ Access Restricted Due to Confidentiality ------------------------------------

6.3 Limitations of the Study

As mentioned in the methodology chapter, this research project has not conducted all of the steps in the DSRM methodology. This is done due to decisions in project scope, limitations in time and resources but also the difficulty in concretely implementing a risky technological solution in a conservative context such as traditional credit risk models at large financial institutions. As a result, the implementation and evaluation step might be conducted in separate projects instead with in the future.

The conducting of semi-structured interviews is a qualitative research method that has its limitations. Respondents can be biased and involuntarily subjective in their answers due to their position and previous experience. Due to this, a variety of experts had been chosen to interview to cover all of the positions within the DDA department. Due to the limited size of the department, a qualitative research method was deemed more appropriate and insightful than a quantitative one. Face to face interviews were held internally to optimize the quality of communication. Externally, video and audio interviews were conducted due to the remote location of respondents and the difference in time-zones. E-mail and chat correspondence was also used to gather data of external respondents that declined to cooperate through an interview. In the ideal setting, multiple personal interviews would be preferred to ensure that the respondents have told everything that was intended with the aid of visual cues and context.

Validation of the High Level Solutions has occurred in limited sense by expert panel review. Internal and external experts on areas as credit scoring processes, risk models, Big Data and European financial legislation were approached. These experts run simulations in their thought process to estimate the feasibility and effects of these solutions in an interactive dialogue. This method of validation and verification is heavily dependent on the level of expertise of the panelists and the accuracy of their review. Additional validation was done on the results of the maturity model as numerical grade representation of a qualitative research method was difficult to establish. However, the wide margins between the maturity tiers solve the impact of small error naturally within the model itself.

Content wise, one can argue that despite an accurate credit scoring process, there will always be clients that will get into financial troubles. That can partially be the fault of inaccurate credit scoring, the approval of unbearable credit weight but also due to unexpected personal circumstances. A credit is traditionally granted on basis of the past or through predictive models, this cannot guarantee the future. Financial institutions do not have the illusion to bring eventual default rate to an absolute zero. However, the future of Fintech aspires to be able to predict, track and improve their client's financial position by monitoring many factors and providing correct stimulus where needed. Life changing events such as divorce and lay-off are near impossible to predict. Each person also reacts differently to such an event which makes it difficult to predict the financial impact. Refractory events are excluded from the data set which is used for Machine Learning as they are considered outliers.

The criteria used in this research were of qualitative nature and obtained from models and articles in the scientific literature. It is difficult to state that these are mutually exclusive towards other undiscovered or unmentioned criteria which could also affect the quality of new data variables or credit scoring processes. Despite being thorough, it is difficult to say that these criteria have covered all of the aspects.

6.4 Future Work

It is highly recommended that in future work a specialized team on law and legislation will be composed to research the legal framework in which Big Data organizations in Europe can operate. As mentioned, the problem with the limited amount of research in this area is that not many applications have been colliding with European legislation yet. Pending legislation in 2018 will pose new challenges to the teams responsible for creating data-driven applications in Fintech. Other grey areas in which these companies operate have not yet been treated by the law. Moreover, the continuation of the design science iteration of this graduation project could be found in a future project assignment which would include small scale implementation and evaluation.

Other points of interest are determining which concrete Machine Learning tools and algorithms are most suitable for the task of identifying new data variables for creditworthiness. And which methods can be best used to develop a model and iteratively improve it. The technology keeps developing and new expertise engages problems every day to obtain additional data and knowledge. The general improvement model given in Appendix I can then be further specified for each unique project and organization.

Additional future research is welcome on reducing fraud in banking as borrowing individuals are motivated to obtain money and they might try this through irregular and irresponsible channels and situations. A person might say or claim different things in moments of pressure and need. Criminals and structural fraudulent clients might also try to game financial products in order to unlawfully obtain credit.

This study has considered the possibilities of using Big Data within the current capacity of technology to determine client creditworthiness. Certain legal frameworks and expert opinions have been asked to validate these solutions. However, the third pillar on which the success of a solution rests is the acceptance of use by the society. An extended survey or quantitative interview can be conducted to evaluate the acceptance and desirability of such trade-offs between value and disclosure within Dutch society. Privacy and security are two sensitive topics which consumers approach in a skeptical manner.

Another research topic which can be interesting is the duration in which credit history is representable. The law states that all credit history should be annulled after five years, but when considering the behavioral aspects of individuals, this still might be representable information. This dilemma considers the duration in which history should be taken into account for the goal of accuracy and what is societally desirable in terms of privacy and “starting over” in life.

Bibliography

Alnafoosi, A. B., & Steinbach, T. (2013). An integrated framework for evaluating big-data storage solutions-IDA case study. *Science and Information Conference (SAI)* (pp. 947-956). London: IEEE.

Alpaydin, E. (2014). *Introduction to Machine Learning.* Cambridge: The MIT Press.

Autoriteit Financiële Markten. (2016, August 25). *Toezichtprioriteiten.* Retrieved from AFM.nl: https://www.afm.nl/nl-nl/over-afm/prioriteiten

Baer, T., Goland, T., & Schiff, R. (2013). *New credit-risk models for the unbanked.* New York City: McKinsey.

Beyer, M. A., & Laney, D. (2012). The Importance of 'Big Data': a definition. *Stamford, CT: Gartner*, 2014-2018.

Brown, M., Stein, S., & Zafar, B. (2015). The Impact of Housing Markets on Consumer Debt:Credit Report Evidence from 1999 to 2012. *Journal of Money, Credit and Banking*, 175-213.

Burnard, P. (1991). A method of analysing interview transcripts in qualitative research. *Nurse Education Today*, 461-466.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. *Journal of data warehousing*.

Chen, Q.-L., & Lin, J.-B. (2015). Integration of Business Intelligence and CRM in Banks: An Empirical Study of SOM Applied in Personal Customer Loans in Taiwan. *International Conference on Fuzzy Theory and Its Applications (iFUZZY)* (pp. 68-73). Istanbul: IEEE.

Cignifi. (2012). *Building the Bridge to New Customers in Brazil.* Cambridge: Cignifi.

Cignifi. (2014). *Mobile Phone Data as the Key to Promoting Financial Inclusion.* Cambridge: Cignifi & The World Savings and Retail Banking Institute (WSBI).

Citron, D. K., & Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89-121.

Cognizant, Marketforce, Pegasystems. (2016). *The Future of Retail Financial Services.* London: Marketforce Business Media Ltd.

Cohen, D., & Crabtree, B. (2006). Qualitative research guidelines project. *Princeton Review*.

Databricks. (2016). *Customer Case Study: LendUp.* San Francisco: Databricks.com.

De Nederlandse Overheid. (2016, 8 11). *Wet op het financieel toezicht.* Retrieved from wetten.overheid.nl: http://wetten.overheid.nl/BWBR0020368/2016-08-11

Docupace. (2016, August 24). *STP Network*. Retrieved from https://docupace.com/wp-content/uploads/2015/05/STP_Graphic_052615_STP-Graphic-Desktop1.png

Dunn, L. F., & Mirzaie, I. A. (2016). Consumer Debt Stress, Changes in Household Debt, and the Great Recession. *Economic Inquiry*, 201-214.

Einav, L., Finkelstein, A., Kluender, R., & Schrimpf, P. (2015). Beyond Statistics: The Economic Content of Risk Scores. *National Bureau of Economic Research - Working Paper Series*, 1-34.

Einav, L., Jenkins, M., & Levin, J. (2013). The impact of credit scoring on consumer lending. *RAND Journal of Economics*, 249-274.

European Banking Authority. (2016, August 24). *Implementing Basel III Europe.* Retrieved from EBA Europe - Regulation and Policy: http://www.eba.europa.eu/regulation-and-policy/implementing-basel-iii-europe

European Commission. (2016, August 24). *EC Europa - CRD IV FAQ Memo.* Retrieved from EC Europa: http://europa.eu/rapid/press-release_MEMO-13-272_en.htm

European Commission. (2016, August 24). *EC Europa - Legislation in Force.* Retrieved from EC Europa: http://ec.europa.eu/finance/bank/regcapital/legislation-in-force/index_en.htm#maincontentSec2

European Parliament. (1995). Directive 95/46/EC. *Official Journal of the European Communities*, 281/31-281-50.

European Parliament. (2016). Regulation (EU) 2016/679 & Directive (EU) 2016/680. *Official Journal of the European Union*, 119/1-119/131.

FICO Insights. (2015). *Can Alternative Data Expand Credit Access?* San Jose: Fair Isaac Corporation.

Financial Market Lawyers. (2016, August 25). *Basel III and Capital Requirements Directive IV.* Retrieved from FMLAAA: http://www.fmlaaa.com/images/cms/Diagrammen/35b.png

Guttierez, D. (2014). *Big Data for Finance.* Portland: Dell, Intel.

Hafiz, A., Lukumon, O., Muhammad, B., Olugbenga, A., Hakeem, O., & Saheed, A. (2015). Bankruptcy Prediction of Construction Businesses: Towards a Big Data Analytics Approach. *IEEE First International Conference on Big Data Computing Service and Applications* (pp. 347-352). Bristol: IEEE.

Halper, F., & Krishnan, K. (2014). *TDWI Big Data Maturity Model Guide Interpreting Your Assessment Score.* TDWI Benchmark Guide.

Hewlett-Packard. (2013). *Capitalize on Big Data in financial services.* Palo Alto: Hewlett-Packard.

Investopedia. (2016, 6 16). *Investopedia - Creditworthiness*. Retrieved from Investopedia.com: http://www.investopedia.com/terms/c/credit-worthiness.asp

Investopedia. (2016, August 24). *Straight Through Processing*. Retrieved from Investopedia.com: http://www.investopedia.com/terms/s/straightthroughprocessing.asp

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 1137-1145.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 249-268.

Kshetri, N. (2016). Big data's role in expanding access to financial services in China. *International Journal of Information Management*, 297-308.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2009). Judging Borrowers by the Company They Keep: Social Networks and adverse selection in online peer-to-peer lending. *SSRN eLibrary*.

Man, Y. (2016). *A Literature Review of Big Data in the Financial Sector.* Enschede: University of Twente.

Mavlutova, I., Zalitis, U., Mavlutova, A., & Mavlutov, B. (2014). Evaluation of Enterprise Solvency in Lending Practice of Commercial Banks: Evidence from Latvia. *Business and Uncertainty: Challenges for Emerging Markets*, 90-108.

McAuley, D., & Weiner, S. (2015). *The Millennial Generation and the Future of Finance:.* Innotribe.

Merriam-Webster. (2016, 6 16). *Merriam-Webster Dictionary: Creditworthiness*. Retrieved from Merriam-Webster.com: http://www.merriam-webster.com/dictionary/creditworthy

National Consumer Law Center. (2014). *Big Data: A Big Disappointment for Scoring Consumer Credit Risk.* Boston: NCLC.

NG Data. (2014). *Leveraging Big Data and Customer Relationships: How Banks Can Benefit from the Mobile Wallet Opportunity.* Gent: NGData.com.

Pfeffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 45-77.

Polillo, S. (2011). Money, Moral Authority, and the Politics of Creditworthiness. *American Sociological Review*, 437-464.

Pressman, S., & Scott, R. (2009). Consumer Debt and the Measurement of Poverty and Inequality in the US. *Review of Social Economy*, 127-148.

Rijksoverheid. (2016, August 25). *Wet op het financieel toezicht (Wft).* Retrieved from Rijksoverheid.nl: https://www.rijksoverheid.nl/onderwerpen/financiele-sector/inhoud/toezicht-op-de-financiele-sector/wet-op-het-financieel-toezicht-wft

Safi, R., & Lin, Z. (2014). Using Non-Financial Data to Assess the Creditworthiness of Businesses in Online Trade. *PACIS*, 206.

Sharma, S., & Manga, V. (2015). Technology and Trends to Handle Big Data: Survey. *Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on* (pp. 266-271). Rohtak: IEEE.

Singh, V. K., Bozkaya, B., & Pentland, A. (2015). Money Walks: Implicit Mobility Behavior and Well-Being. *PLoS ONE*, 1-17.

Sobolevsky, S., Sitko, I., Tachet des Combes, R., Hawelka, B., Murillo Arias, J., & Ratti, C. (2014). Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in Spain. *Big Data Congress* (pp. 136-143). Anchorage: IEEE.

SoFi. (2015). *The Financial Planner's Guide to Student Loan Refinancing.* San Francisco: Social Finance, Inc.

Srinivasa, S., & Metha, S. (2014). *Big Data Analytics.* New Delhi: Springer.

The Federal Trade Commission. (2016). *Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues.* Washington D.C.: FTC.

The U.S. Department of the Treasury. (2016). *Opportunities and Challenges in Online Marketplace Lending.* Washington D.C.: US Government.

The White House: Executive Office of the President. (2016). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.* Washington D.C.: US Government.

The World Bank. (2014). *Financial Inclusion Data - The Netherlands*. Retrieved from The World Bank Data Topics: http://datatopics.worldbank.org/financialinclusion/country/netherlands

Tivey, I. (2015). *Social Media Monitoring: Using Big Data Techniques to Enhance Customer Analytics.* Zürich: Citihub Consulting.

Wang, Y., Li, S., & Lin, Z. (2013). Revealing Key Non-financial Factors for Online Credit-Scoring in e-Financing. *IEEE*, 547-552.

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2011). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, 1-11.

World Economic Forum. (2015). *The Future of Financial Services. How disruptive innovations are reshaping the way financial services are structured, provisioned and consumed.* Davos: WEF.

Yin, S., & Kaynak, O. (2015). Big Data for Modern Industry: Challenges and Trends [Point of View]. *Proceedings of the IEEE* (pp. 143-146). IEEE.

Yu, Q. (2015). The Research on the Development of China's Internet Finance. *International Conference on Education, Management and Computing Technology* (pp. 1426-1429). Tianjin: Atlantis Press.

Zarsky, T. (2016). An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *SAGE*, 118-132.

Appendices

Appendix A – Approached Organizations

This appendix shows a list of approached organizations for the external market review. The organizations are sorted by type of company.

Retailers:

- Wehkamp
- Lacent
- Bol.com

Public Organs:

- Bureau Krediet Registratie

Internet Giants:

- ANT Financial Service Group
- Alibaba
- Amazon
- Yahoo

Traditional Credit Scoring Agencies:

- Transunion
- Equifax
- Experian

Data Brokers:

- Axciom
- Corelogic
- Datalogix
- EBureau
- ID Analytics
- Intelius

Universities:

- Cambridge University
- Queen Mary University of London
- University of Edinburgh
- Brunel University London
- Radbout Universiteit
- Universiteit Twente

Fintech Organizations:

- Cignifi
- InVenture
- Lenddo
- ZestFinance
- Big Data Scoring
- Credit Karma
- OnDeck
- Kabbage
- Better Finance Inc.
- Fundera
- Biz2Credit
- Kiva
- Funding Circle
- Lending Club
- Prosper
- Sameningeld B.V.
- Lendico
- Lendex
- Geldvoorelkaar.nl
- Kreditech
- Spotcap NL
- Upstart
- LendUp
- Think Finance Inc.
- Wonga
- Social Finance (SoFi)
- Earnest
- CommonBond

Appendix B – Interview Questions

The method concerned with conducting semi-structured interviews allows the respondent to create their own trajectory to more interest bearing topics. This causes the interviews to not always be conducted in the set order with the initially proposed questions. Most respondents provide additional information after follow-up questions that was are not included in the preset interview questions. They provide additional insight when commented and questioned further upon.

Goals – Big Data Maturity

- Determine the current internal analytical and computing capacities in terms of Big Data.
- Map the current IT Infrastructure and the support for Big Data analytics.
- Inquire about data management and data governance on local and company-wide level
- Discuss the strategic, tactical and operational internal vision on Big Data.
- Discover bottlenecks due to organizational structure or processes in corresponding dimensions.

Interview Questions – Big Data Maturity

- As an introduction, could you tell us a bit more about yourself and what you do exactly at the Data Driven Analytics Department at NeoBank?
- What are the operational, tactical and strategic tasks of Data Driven Analytics?
- What have the proofs of concepts concretely generated from a few years ago?
- Could you tell me a bit more about the current Big Data capacities of NeoBank?
    - What data points are being collected of the current clients?
    - What are the current and future plans of DDA and NeoBank?
    - How is NeoBank approaching real-time analytics of Big Data?
    - What is concretely the value generated by using Big Data?
- How does the organizational structure of NeoBank influence the use of Big Data?
- How would you approximate the Big Data maturity of NeoBank?
    - Does the DDA department operate alone or is it well involved and connected with the whole of the organization?
    - In your opinion, is NeoBank data-driven? Why (not)?
        - How does NeoBank deal with the conservative culture and mindset?
    - Who is currently using the gathered and generated information?
    - How would you rate the analytical capabilities of NeoBank/DDA?
    - Could you draw the Big Data infrastructure of NeoBank?
        - Do the Legacy Systems inhibit the implementation of Big Data?
    - Is there a lack of funding, business involvement or staff to further exploit Big Data? Is there sufficient top-down and bottom-up involvement?
    - Does NeoBank calculate the return on investments on Big Data projects?
- Which measures has NeoBank taken in relation to the upcoming European legislation on data privacy and ownership? (Mifid II, PSD2, FATCA, CRS, XS24)
    - How is compliance to security & privacy regulated internally?
- Is there knowledge being shared between organizations and banks in relation to Big Data?
- How do you believe Big Data could be used to improve the credit scoring process?

Goals – Internal Credit Scoring Processes

- Map the steps of NeoBank credit scoring processes from start till end (Specifically PFC & NEOFC).
- Acquire background information on how used parameters, coefficients, algorithms and formulas in credit scoring models are established.
- Determine the compatibility of Big Data Analytics with current credit scoring processes.
- Discuss the expanding potential of Big Data usage in financial product design.

Interview Questions – Internal Credit Scoring Processes

- Could you tell us a bit more about yourself and your role in the design of NEO Fast Credit?
    - Why in specific was "Paleo Fast Credit" (PFC) formed to NEOFC?
- What do you know about the various processes to determine creditworthiness at NeoBank?
    - How accurate is credit currently determined at NeoBank?
    - How is the accuracy measured of the credit scoring process?
    - What number of consumers is structurally indebted or encounters payment problems when being granted a credit? (Longer than three months)
        - What processes currently run to enable consumers to get rid of their debt?
        - Are payment problems easy to predict with the current credit scoring system?
    - On what time basis or regularity is the creditworthiness determined?
    - What are currently the biggest challenges in determining creditworthiness?
    - What are the costs when determining credit worthiness?
- What are the elemental components of a credit scoring process? What are those of a model?
    - Could you dissect the "Inkomens & Lastentoets" in components as an example?
    - What is the required time, counting from moment of applying till granted credit?
- How is a credit scoring algorithm tested before it is used in the market?
- Creditworthiness is defined as "*the intrinsic quality of people and businesses reflected in their ability and willingness to fulfil their business obligations*", are both parts measured and how does the measuring happen at NeoBank?
- Credit Scoring Agencies in the U.S. use integral scores to express the creditworthiness of clients. These scores are used in the application process of financial products. The algorithms behind these products are held secret. Has there been thought of an integral credit score at NeoBank?
    - Would this be possible or desirable in the future in the Netherlands?
- What are the challenges in specifying and developing a financial product?
    - How is the deliberation made between sufficient screening and consumer friendly product in product design?
- Are there instances of fraud and human error when granting a credit to consumers or determining credit worthiness? What causes these errors?
- What are the biggest thresholds and problems that consumers experience when applying for a credit, and NEOFC in specific?
- Reducing the default rate starts at rejecting or approving a financial product. In your opinion, what parameters and data variables should be looked at besides the current ones?
- How could the current credit scoring process be further improved by the use of Big Data?

Goals – External Research

- Obtain practical information on data scoring processes at third parties.
  - Determine which data variables are used to compute credit scores.
- Converse about the accuracy in which creditworthiness can and should be determined.
- Converse about the potential and challenges of Big Data on determining creditworthiness
- Discuss the requirements for using Big Data to assess credit scores.
- Discuss the European market and limitations within legislation.

Interview Questions – External Research

- Could you in short tell me a bit more about yourself and the work you do at *?
- Could you tell me a bit more about what * is and does?
  - How does * determine creditworthiness?
  - What data variables does * look at?
  - Where does * offers its products and to whom?
- Looking at one of the definitions of creditworthiness; *the intrinsic quality of people and businesses reflected in their* ***ability*** *and* ***willingness*** *to fulfil their business obligations.* Does * take both of these factors into account when scoring credit? Why/how?
- The accuracy and timeliness of creditworthiness is a major issue in the utility of credit scores, does * take these parameters into account?
- Other Fintech companies like * have been appearing that also score credit on basis of Big Data, what distinguishes * in this market?
- I've noted that * collaborates with his partner *, which is also a U.S. based FICO scoring agency. What type of collaboration is this?
  - Do you see traditional credit scorers as competition?
  - Do you see potential at banks performing the same operations? Why yes/not?
- In Europe, many of the financial products that are offered have their own credit scoring process linked to the application. This causes a lot of inefficiency and delay in applications for products. Does * think of data partnerships with banks to optimize such processes?
  - In these kinds of markets where credit scoring agencies are not common like in the U.S., how do you increase the utility of your product?
  - How do you think the lending industry and credit scoring processes will evolve with the advancement of technologies and legislation?
- In Europe, many of the financial products that are offered have their own credit scoring process linked to the application. This causes inefficiency and delay in applications for products. Does * think of data partnerships with banks to optimize such processes?
  - In these kinds of markets where credit scoring agencies are not common like in the U.S., how do you increase the utility of your product?
- What are the primary regulatory, data ownership, privacy and validity challenges you have been posed while developing your product?
  - How does the European legislation framework impede the operations or strategic goals?

Appendix C – TDWI Big Data Maturity Model Criteria Description

This appendix describes the individual criteria which are used to determine the maturity of the dimensions within the TDWI Big Data Maturity Model.

Organization

- **Leadership** - To what extent the established leadership and management supports the Big Data analytics program. Indicates what level of leadership is exhibited by employees and managers on the subject of Big Data.
- **Funding** - To what extent the funding currently inhibits or supports a successful Big Data analytics program at the enterprise. The long-term level at which future funding is secured.
- **Strategy** - The interest spent in strategically including the Big Data analytics program in the formal strategic vision of the enterprise. Indicates if this strategy is in aligned with the overall vision of the enterprise or individual departments. Also determines if concrete actions can be derived company-wide, top-down and bottom-up from this strategy.
- **Culture** - The extent at which a data-driven and analytics culture is present at the company. Indicates if innovation is inhibited or stimulated by the business culture present.
- **Value** - Indicates how business, operational and organizational value is derived from the Big Data analytics program. Also indicates the awareness of value extraction. Represents how value is measured, prioritized and converted to tangible outcomes.

Infrastructure

- **Development** - The level of the developmental operations throughout the enterprise in terms of the Big Data analytics program. Describes how the processes are streamlined to innovate.
- **Technologies** - The rate and scale at which new innovative Big Data technology is used in support of the advanced analytics program such as Hadoop or NoSQL databases.
- **Architecture** - To what extent the infrastructure supports all parts of the company and potential users. How advanced and coherent the architecture is in support of Big Data initiatives.
- **Integration** - The level and detail at which the Big Data technology is integrated into the existing environments throughout the enterprise.
- **Scope** - How infrastructural resources are allocated in terms of prioritizing Big Data projects. The ease at which the scope is interchanged from small scale (departmental) to large scale (company-wide/external collaborative)

Data Management

- **Variety, Volume, Velocity** - The range of different challenging Big Data stored and used in terms of the 3 Vs. Indicates scale, difference in types of data and real-time challenges.
- **Processing** - Indicates the capacity and rate of the enterprise at which the organization can process the data (parallel) without grave errors.
- **Storage** - The ease at which local and centralized storage of data takes place. Indicates the capacity in solving storage issues when they occur.
- **Quality** - Indicates the level of data quality of the stored and enhanced data within the organization. Determines the accuracy and level of detail in which data is stored.
- **Access** - The level of access to the data or information by relevant and capable departments and employees which can use it to come to insights or create value.

Analytics

- **Skills** - The level of individual analytical skills of the engineers and data scientists or analysts potentially involved with the Big Data analytics program and beyond.
- **Mindset** - The mindset of the individuals in terms of advancement and innovation regarding Big Data analytics. Strongly linked with organizational culture.
- **Techniques** - The actuality of Big Data analytics techniques used by the company and its departments in order to generate value.
- **Applications** - The level and speed at which the Big Data analytics program produces concrete tangible applications which generates value throughout the enterprise.
- **Delivery Methods** - The level of the process and structure which dictates the way analytics are delivered throughout the organization.

Governance

- **Policies** - The existence, coherency and upholding of the policies in support of the company's Big Data analytics program.
- **Structure** - The level at which a sound, transparent and present governance structure is implemented to guarantee compliance to standards.
- **Compliance** - Determines the rate of compliance within the Big Data governance program and the validation of compliance.
- **Stewardship** - The extent at which stewardship exists at the organization or department and is delegated to certain functions. Data stewardship focuses on tactical coordination and implementation compliant to governance.
- **Security and Privacy** - Indicates the level at which security and privacy is safeguarded. For example, by delegation to certain security and privacy officers within the Big Data analytics program. They hold the processes and applications to legislation and legal frameworks to test for failure or endurance.

Appendix D – Big Data Maturity Dimensions Assessment NeoBank

Organization Criteria	Grade
Leadership	[redacted]
Funding	[redacted]
Strategy	[redacted]
Culture	[redacted]
Value	[redacted]
Total Score	[redacted]

Infrastructure Criteria	Grade
Development	[redacted]
Technologies	[redacted]
Architecture	[redacted]
Integration	[redacted]
Scope	[redacted]
Total Score	[redacted]

Data Management Criteria	Grade
Variety, Volume, Velocity	[redacted]
Processing	[redacted]
Storage	[redacted]
Quality	[redacted]
Access	[redacted]
Total Score	[redacted]

Analytics Criteria	Grade
Skills	[redacted]
Mindset	[redacted]
Techniques	[redacted]
Applications	[redacted]
Delivery Methods	[redacted]
Total Score	[redacted]

Governance Criteria	Grade
Policies	[redacted]
Structure	[redacted]
Compliance	[redacted]
Stewardship	[redacted]
Security and Privacy	[redacted]
Total Score	[redacted]

Appendix E – Big Data Maturity Dimensions Assessment DDA

Organization Criteria	Grade
Leadership	[redacted]
Funding	[redacted]
Strategy	[redacted]
Culture	[redacted]
Value	[redacted]
Total Score	[redacted]

Infrastructure Criteria	Grade
Development	[redacted]
Technologies	[redacted]
Architecture	[redacted]
Integration	[redacted]
Scope	[redacted]
Total Score	[redacted]

Data Management Criteria	Grade
Variety, Volume, Velocity	[redacted]
Processing	[redacted]
Storage	[redacted]
Quality	[redacted]
Access	[redacted]
Total Score	[redacted]

Analytics Criteria	Grade
Skills	[redacted]
Mindset	[redacted]
Techniques	[redacted]
Applications	[redacted]
Delivery Methods	[redacted]
Total Score	[redacted]

Governance Criteria	Grade
Policies	[redacted]
Structure	[redacted]
Compliance	[redacted]
Stewardship	[redacted]
Security and Privacy	[redacted]
Total Score	[redacted]

Appendix F – Capital Requirements Regulation and Directive IV

The Basel (III) Accords specify worldwide rules on capital measurement and capital standards to sharpen risk management, regulation and supervision within the financial sector. Reform takes place in risk inciting remuneration policies, financial instruments, resecuritization and disclosure of securitization risks. In overall, the measures taken by the financial authorities are risk aversive in the light of economic crises of the past decades. The Capital Requirements Directives IV (CRR/CRD IV) is a legislative package containing rules based on the Basel Accords, effective in the European Union. This European Legislation package treats many financial areas in order to create order and a safer financial system. In specific for our thesis, it also offers standards for banks to organize and compute their risk-related capital buffers in a transparent and prudent manner. The standards require the quality and the amount of capital held by banks to improve. Additional buffers such as the capital conservation buffer and countercyclical buffer (depending on nation) are implemented to deal with periods of economic stress (European Commission, 2016).

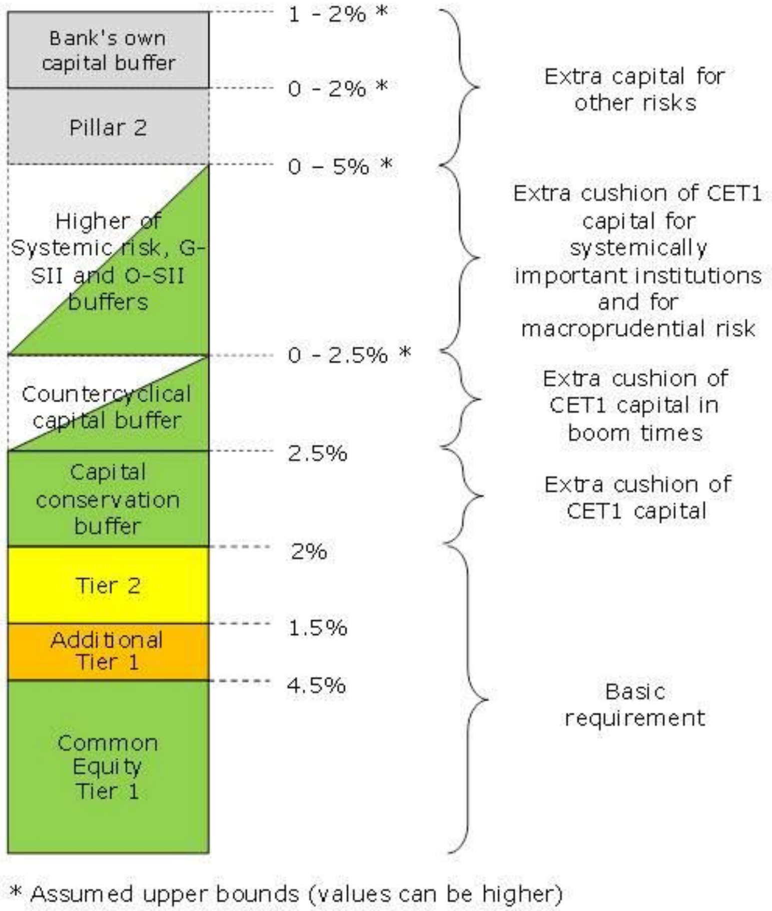

Bank's own capital buffer
Pillar 2
Higher of Systemic risk, G-SII and O-SII buffers
Countercyclical capital buffer
Capital conservation buffer
Tier 2
Additional Tier 1
Common Equity Tier 1
1 - 2% *
0 - 2% *
0 – 5% *
0 - 2.5% *
2.5%
2%
1.5%
4.5%
Extra capital for other risks
Extra cushion of CET1 capital for systemically important institutions and for macroprudential risk
Extra cushion of CET1 capital in boom times
Extra cushion of CET1 capital
Basic requirement
* Assumed upper bounds (values can be higher)

FIGURE 18. EFFECTIVE LEVEL OF REQUIRED REGULATORY CAPITAL (EUROPEAN COMMISSION, 2016)

The European Banking Authority (EBA) enforces this legislation by mandate of the European Union. Member States are lawfully bound to comply with these laws as if part of their national law. The concrete laws which have to be considered during the development phase of credit risk models are the Binding Technical Standards (BTS) of CRD IV. These are legal acts drafted to specify and ensure consistency in certain financial areas such as risk management (European Banking Authority, 2016).

CRD IV was first adopted in 2013, while the package applies as of 1 January 2014. Some of the new provisions are phased-in between 2014 to 2019. The introduction of increased capital requirements in the package is expected to develop as follows in the illustration shown below. It is notable that the requirements are slowly being driven upwards, and that regulation is thus becoming increasingly strict with respect to the buffers maintained by financial institutions.

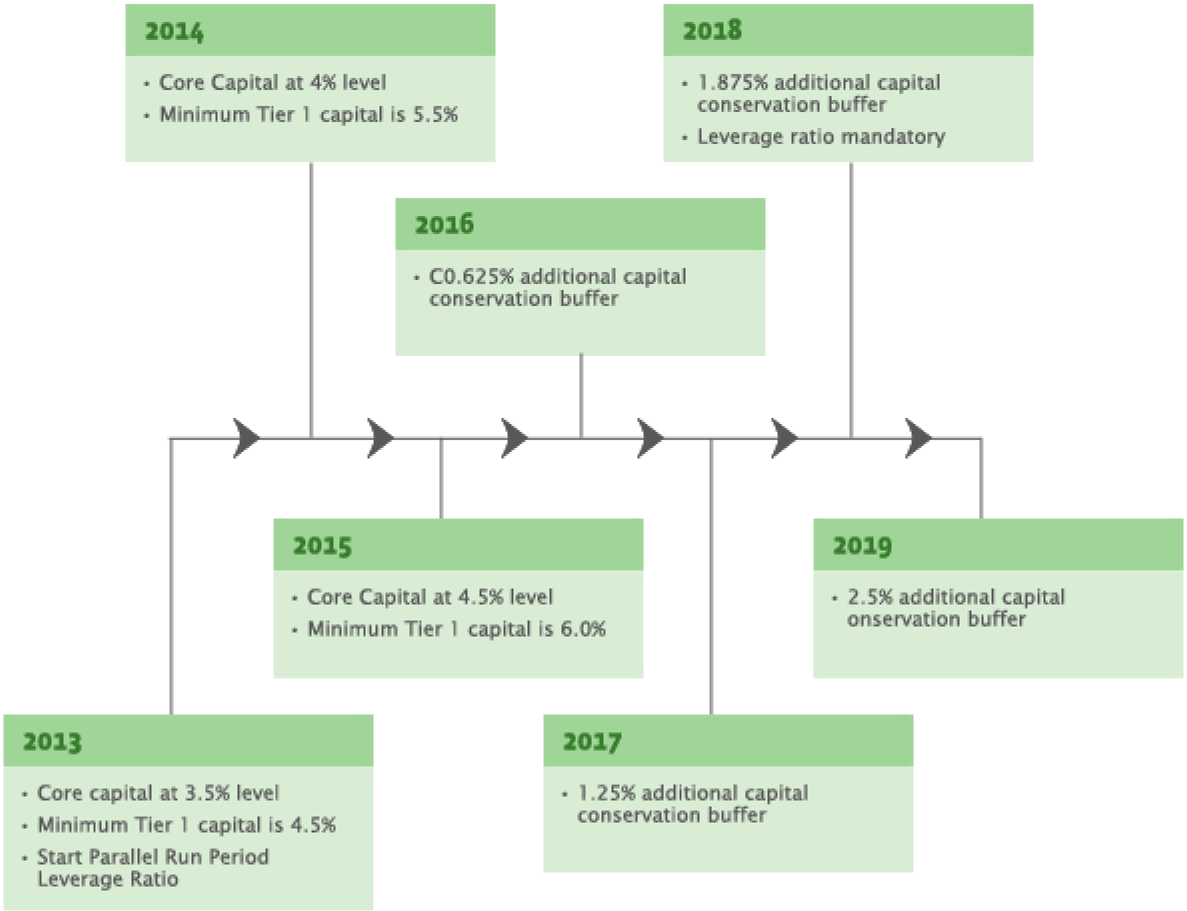
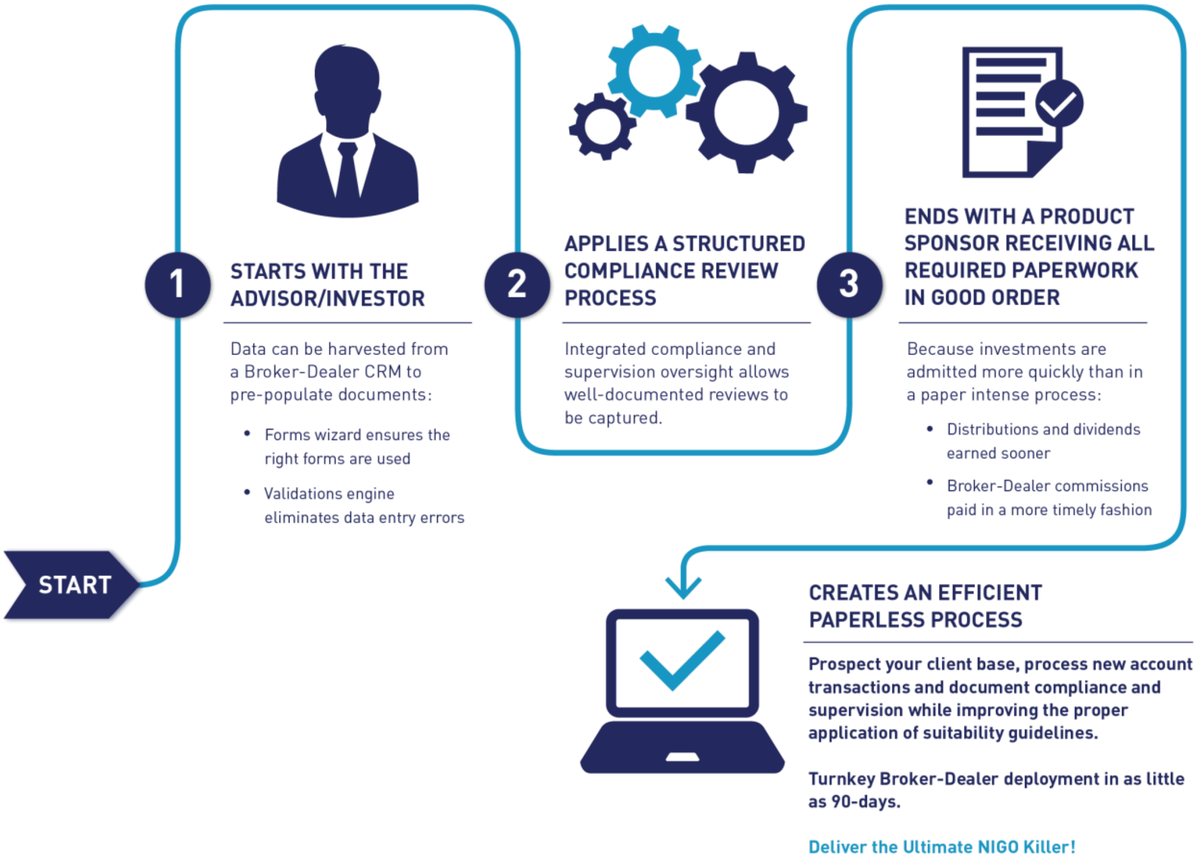

2014
• Core Capital at 4% level
• Minimum Tier 1 capital is 5.5%
2018
• 1.875% additional capital conservation buffer
• Leverage ratio mandatory
2016
• C0.625% additional capital conservation buffer
2015
• Core Capital at 4.5% level
• Minimum Tier 1 capital is 6.0%
2019
• 2.5% additional capital onservation buffer
2013
• Core capital at 3.5% level
• Minimum Tier 1 capital is 4.5%
• Start Parallel Run Period Leverage Ratio
2017
• 1.25% additional capital conservation buffer

FIGURE 19. TRANSITION OF REQUIRED CAPITAL BUFFERS BASED ON CRD IV (FINANCIAL MARKET LAWYERS, 2016)

Appendix G – Straight Through Processing

Financial firms use Straight Through Processing in order to shorten the lead time on transaction processing without human involvement. In essence, STP is a process technology that allows digitally shared transaction processing. By adapting this, manual redundant reentry is eliminated in the process which has already been completed once in the source. This also implies that for instance, elements of a loan application only have to be entered once, where after it can be reused. STP based means that direct and complete automatic execution takes place without human intervention and human error. Two important conditions to implement STP are: (a) The various automatized information systems must be integrated in such a way that the activities of the operations process can be performed correctly. (b) The formal process definition must be aligned with the process logic behind the system. The process logic directs the process operations in the system. This would eventually result in shorter lead times and less errors (lean). Like in the image shown below, this process has been initially adapted in the trading environment, where capital market and payment transactions are transitioned to a digital environment (Investopedia, 2016). An example in trading is given in the illustration below.


START
1
STARTS WITH THE ADVISOR/INVESTOR
2
APPLIES A STRUCTURED COMPLIANCE REVIEW PROCESS
3
ENDS WITH A PRODUCT SPONSOR RECEIVING ALL REQUIRED PAPERWORK IN GOOD ORDER
CREATES AN EFFICIENT PAPERLESS PROCESS

Data can be harvested from a Broker-Dealer CRM to pre-populate documents:

- Forms wizard ensures the right forms are used
- Validations engine eliminates data entry errors

Integrated compliance and supervision oversight allows well-documented reviews to be captured.

Because investments are admitted more quickly than in a paper intense process:

- Distributions and dividends earned sooner
- Broker-Dealer commissions paid in a more timely fashion

Prospect your client base, process new account transactions and document compliance and supervision while improving the proper application of suitability guidelines.

Turnkey Broker-Dealer deployment in as little as 90-days.

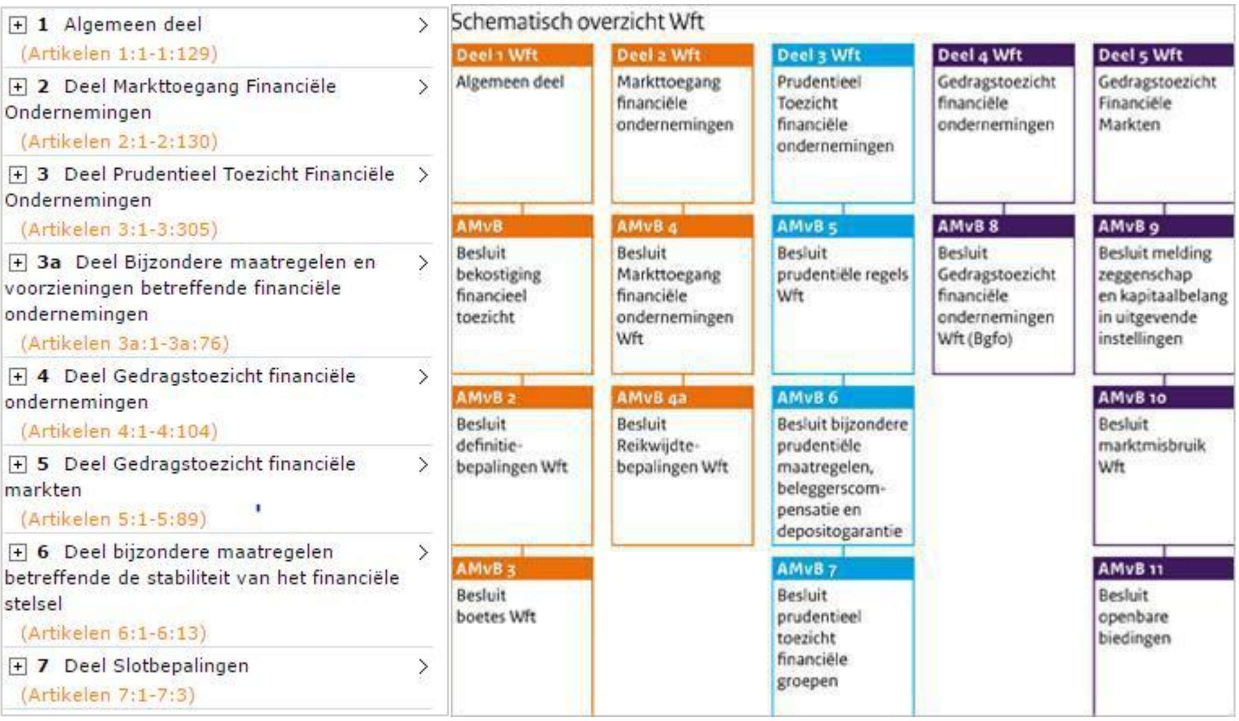
Deliver the Ultimate NIGO Killer!

FIGURE 20. EXAMPLE OF A STP-BASED PROCESS IN TRADING (DOCUPACE, 2016)

Appendix H – Wet op het Financieel Toezicht

The “Wet op het financieel toezicht” (Wft) was implemented in January 2007 and regulates the supervision on the Dutch financial sector. Its goal is to reform financial market legislation into a coherent, transparent and focused whole. The law simplifies many subjects in finance and describes the collaboration between the Dutch National Bank (DNB) and the Dutch Market Authorities (AFM). General Part One of the law forms the basis of this legal framework and elaborates on the authority and legislative power of the supervising organs. The DNB is responsible for the prudential supervision of banks to guarantee financial stability. The AFM as mentioned earlier, conducts oversight on how banks behave in the market. They supervise the trust and entry into the Dutch financial market. They also pose a set of rules of conduct that financial institutes need to comply with while offering their products and services. These rules serve to prevent exploitation of consumers, to promote financial market accessibility, sustain the free market processes and ensure confidence in the financial market (Autoriteit Financiële Markten, 2016). There is a lot of overlap between the issues treated by the two authorities, but the law mandates their relative power in relation to their areas of authority.

The actuation of the law replaced eight previous statutes each focused on one branch of the financial market. The current law consists of various parts that treat specific topics in the financial sector. The law is frequently updated through the year as there is a minimal yearly change cycle. In summary, the most recent version of the Wft (De Nederlandse Overheid, 2016) consists of the following. A general part (1), many parts (2) (3) (3a) (4) on financial firms plus the activities and services of financial firms, parts (5) (6) on the general market and a remaining part (7). The DNB has jurisdiction on part three, prudential supervision on financial firms. The AFM uses all other parts of the Wft law. Some specific areas are shown in the overview below.



⊞ **1** Algemeen deel >
(Artikelen 1:1-1:129)

⊞ **2** Deel Markttoegang Financiële Ondernemingen >
(Artikelen 2:1-2:130)

⊞ **3** Deel Prudentieel Toezicht Financiële Ondernemingen >
(Artikelen 3:1-3:305)

⊞ **3a** Deel Bijzondere maatregelen en voorzieningen betreffende financiële ondernemingen >
(Artikelen 3a:1-3a:76)

⊞ **4** Deel Gedragstoezicht financiële ondernemingen >
(Artikelen 4:1-4:104)

⊞ **5** Deel Gedragstoezicht financiële markten >
(Artikelen 5:1-5:89)

⊞ **6** Deel bijzondere maatregelen betreffende de stabiliteit van het financiële stelsel >
(Artikelen 6:1-6:13)

⊞ **7** Deel Slotbepalingen >
(Artikelen 7:1-7:3)

Schematisch overzicht Wft

Deel 1 Wft	Deel 2 Wft	Deel 3 Wft	Deel 4 Wft	Deel 5 Wft
Algemeen deel	Markttoegang financiële ondernemingen	Prudentieel Toezicht financiële ondernemingen	Gedragstoezicht financiële ondernemingen	Gedragstoezicht Financiële Markten
AMvB Besluit bekostiging financieel toezicht	AMvB 4 Besluit Markttoegang financiële ondernemingen Wft	AMvB 5 Besluit prudentiële regels Wft	AMvB 8 Besluit Gedragstoezicht financiële ondernemingen Wft (Bgfo)	AMvB 9 Besluit melding zeggenschap en kapitaalbelang in uitgevende instellingen
AMvB 2 Besluit definitie-bepalingen Wft	AMvB 4a Besluit Reikwijdte-bepalingen Wft	AMvB 6 Besluit bijzondere prudentiële maatregelen, beleggerscompensatie en depositogarantie		AMvB 10 Besluit marktmisbruik Wft
AMvB 3 Besluit boetes Wft		AMvB 7 Besluit prudentieel toezicht financiële groepen		AMvB 11 Besluit openbare biedingen

FIGURE 21. AN OVERVIEW OF THE WFT (DE NEDERLANDSE OVERHEID, 2016) (RIJKSOVERHEID, 2016)

Appendix I – Improvement (ML) Process for Credit Scoring

------------------------------------ Access Restricted Due to Confidentiality ------------------------------------

FIGURE 22. MACHINE LEARNING IN CREDIT SCORING

Appendix J – Machine Learning Techniques

This appendix explains some of the most widely applied machine learning techniques in de industry. The below list is known for its use in bankruptcy prediction models (Hafiz, et al., 2015). The best choice of algorithms differs over time for every unique scenario. This can be dependent on the size and dimensionality of the training set, but also if data shows traces of linearity. Qualities are weighed against each other such as computational efficiency, predictive accuracy, speed and comprehensibility. Therefore, the experience and knowledge of the data scientist is of utmost importance.

- **Support Vector Machines (SVM)** – A dividing "hyperplane" is made by using SVM on selected variables of importance of sample firms in order to distinguish between failing and non-failing. This is the boundary that separates classes by as wide a margin as possible. Only variables close to this so called hyperplane are relevant and thus used for classification. This is categorized as a supervised learning model, and used a lot in anomaly detection.
- **Multi-Discriminant Analysis (MDA)** – This method is used to support classification by allowing an analyst to focus on variables that are most relevant for the analysis. In the specific case of this paper, a single value called the Altman Z-Score is calculated to predict the chance of bankruptcy. A sample is taken and classified in failed/non-failed based on a pre-established threshold. Requires rather restrictive assumptions for a real life analysis.
- **Artificial Neural Networks (ANN)** – This AI tool simulates the brain's neural network to make classifications. It is used to estimate functions dependent on large quantities of uncertain and unknown inputs. Machine Learning is employed to train the network before classification is actually done. Over- and undertraining can be detrimental to the prediction models. There are many types and classes of ANNs as they are usually used to approximate the mechanisms of the human mind. Classification is done binary. This algorithm is known to support "Deep Learning".
- **Rough Sets** – This method breaks all variables down into components and groups similar variables in order to build partitions between sample firms. It is a rough approximation of a conventional (original) set with a lower and upper bound.
- **Case Based Reasoning (CBR)** – Cases are stored in a case library on a regular basis and decision rules are made based on these cases for classification. In layman's terms, newly posed problems are solved by offering functional solutions of similar past problems.
- **Iterative Dichotomiser 3 (ID3)** – The most discriminating variables are analyzed between failing and non-failing firms (classes), then a recursive partition is constructed between firms. This is a category of decision tree learning as a decision tree is generated from a data set in iterative fashion. These are easily interpretable and visualized.
- **Genetic Algorithm (GA)** – This AI tool is an optimizing search tool which tracks the global minimum in search space by simulating the principles of Darwin's Evolution Theory. Decision rules are derived and all rules must be satisfied before a decision is made. This search (meta)heuristic drops solutions based on these criteria until the optimal solution is found. It is part of the class of evolutionary algorithms used for evolutionary learning.
- **Active learning** – This technique is categorized as a semi-supervised type of Machine Learning. A learning algorithm iteratively queries the user or other information source in an active fashion to obtain outputs at new data points.

Appendix K – Machine Learning Model Validation Techniques

This appendix highlights a few of the most widely used validation techniques for predictive machine learning models. These techniques are used select the best model and to fine-tune the parameters. Beforehand the proportion of training, validation and test data sets needs to be established. This has effects on the accuracy by which is measured, the representativeness of the model and the confidence interval of a test. Decisions have to be made on how the data is split to deal with potential bias such as class imbalance as the method also influences the model. Kohavi (1995) makes mention of these techniques in his article.

In practice, the most used method to find the optimal complexity is **cross-validation**. This technique avoids overlap in test sets and does this by splitting the main data set into subsets of equal size. Then the subsets are individually in turn taken for testing, while the remainder in each instance is used for training.

The thought process behind this is that it is not possible to compute the bias and variance of models but it is possible to determine the total error. Depending on the data set, the sample is divided into a training and validation set within a certain proportion. The Machine Learning models are trained of different complexities and their error is tested by the validation set which is originally left out of the training. The validation set is introduced once the initial model is finished in order to optimize parameters. In general, as the complexity within a model increases by adding variables and parameters, training error are continuously decreased as the model fits the training data better. The error on the validation set decreases up to a certain level of complexity. At this certain point it stops decreasing, staves off or even increases if there is significant noise. This peak corresponds to the optimal complexity level. The most standard method is the stratified ten-fold cross-validation. The experience by the community has shown that this is the best choice.

To use **Bootstrap** for validation means to produce multiple sample instances from one master sample uniformly. The data set is sampled *with replacemen*t and thus the probability of instances being chosen or not is a fixed constant (0,368 vs 0,632). The bootstrap might also that generates new samples by drawing instances from the original sample with replacement. The error estimate calculated with bootstrap is pessimistic as training data only held 63% of the instances. In general, this is considered the best way to estimate performance of small data sets.

In the **Holdout** method, the data is partitioned into two exclusive subsets called the training set and test set. Common practice is to use two thirds for the training set and the remainder for the test set. The inducer, or learner is fed the training set to enhance the predictive capabilities. The test set is used to validate this. It can be made more reliable by repetition in different subsamples. The holdout estimated accuracy is defined with the following formula.

$$\mathrm{acc}_h = \frac{1}{h} \sum_{(v_i, y_i) \in \mathcal{D}_h} \delta(\mathcal{I}(\mathcal{D}_t, v_i), y_i) \ ,$$