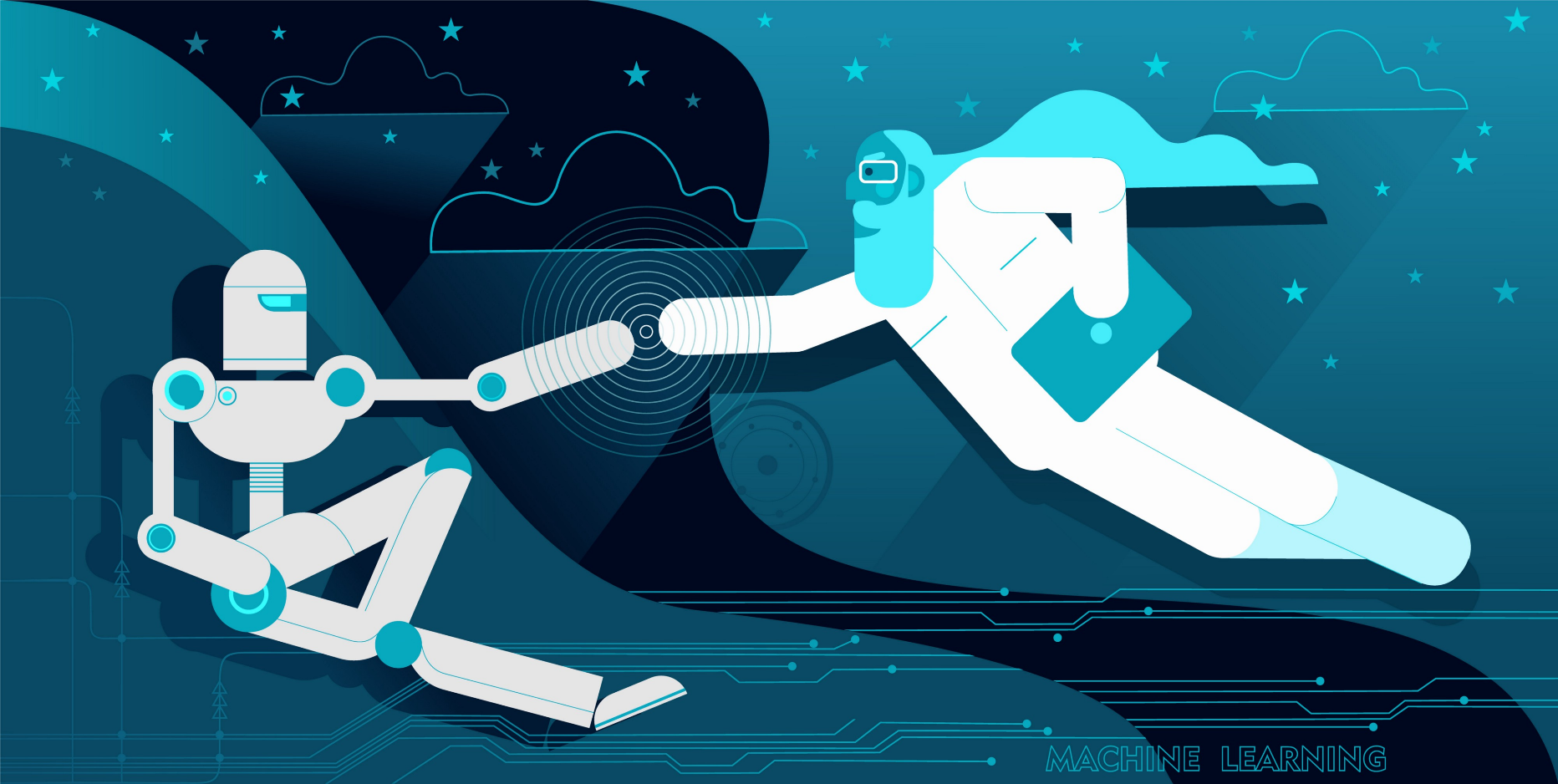
Data driven business development

- Loan data

Eirik Duesund Helland



Executive summary

*Executive summary - A brief summary (1-2 pages) of the key findings in the report.*

*------------------------------------------Introduction Paragraph ---------------------------------------------------------------------*

*------------------------------------------problem statement Paragraph -----------------------------------------------------------*

*------------------------------------------overview of the work performed Paragraph --------------------------------------------*

*-----------------------------------Overview of the key findings and recommendations ------------------------------------------*

Abbreviations

Eg.

ML – Machine Learning

List of figures

[Figure 1 - Molka, T., Gilani, W. & Zeng, X.-J. (2013). Dotted Chart and Control-Flow Analysis for a Loan Application Process. Business Process Management Workshops, Tallinn, Estonia, pp. 219-220. Berlin, Heidelberg: Springer. 10](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602781)

[Figure 2 – System Architecture Flow Chart. 12](#_Toc74602782)

[Figure 3 – System Architecture part 1. 13](#_Toc74602783)

[Figure 4 – System Architecture part 2. 13](#_Toc74602784)

[Figure 5 – Features and Missing Values. 14](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602785)

[Figure 6 – Histogram showing the Data’s Distribution. 14](#_Toc74602786)

[Figure 7 – Boxplot of Data. 15](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602787)

[Figure 8 – Numerical Variables plotted vs Loan Status. 15](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602788)

[Figure 9 – Categorical Variables plotted vs Loan Status. 15](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602789)

[Figure 10 – Correlation Matrix. 15](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602790)

[Figure 11 – Pre-Processing script. 16](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602791)

[Figure 12 – Model training script. 17](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602792)

[Figure 13 – Confusion Matrix. 18](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602793)

[Figure 14 – Output from console after modelling. 18](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602794)

[Figure 15 – Imports for the application script. 18](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602795)

[Figure 16 – The rest of the application script. 19](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602796)

[Figure 18 – Declined Loan in Streamlit. 20](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602797)

[Figure 17 – Approved Loan in Streamlit 20](file:///C:\Users\Eirik\NMBU\TIN200\loan_application_processing\Data%20driven%20business%20development%20-%20TIN200.docx#_Toc74602798)

List of Tables

**No table of figures entries found.**

Content

[Introduction 7](#_Toc74602749)

[Background 7](#_Toc74602750)

[Problem statement 8](#_Toc74602751)

[Goals and Objectives 8](#_Toc74602752)

[Limitations 8](#_Toc74602753)

[Theory and key concepts 9](#_Toc74602754)

[Data driven business development 9](#_Toc74602755)

[Robotic process automation 9](#_Toc74602756)

[Machine learning 10](#_Toc74602757)

[Loan Application process 10](#_Toc74602758)

[State-of-the-art analysis 11](#_Toc74602759)

[Summary 11](#_Toc74602760)

[Develeopment 12](#_Toc74602761)

[Design description 12](#_Toc74602762)

[Concept Overview 12](#_Toc74602763)

[System Architecture 12](#_Toc74602764)

[Exploring the data 14](#_Toc74602765)

[system development 16](#_Toc74602766)

[PreProcessing 16](#_Toc74602767)

[Modelling 17](#_Toc74602768)

[Application 18](#_Toc74602769)

[Summary 21](#_Toc74602770)

[Evaluation and results 22](#_Toc74602771)

[Data 22](#_Toc74602772)

[Model 22](#_Toc74602773)

[Implementation 22](#_Toc74602774)

[Business value 22](#_Toc74602775)

[Conclusion and Recommendations 23](#_Toc74602776)

[Further work 23](#_Toc74602777)

[Attachments 24](#_Toc74602778)

[Source code 24](#_Toc74602779)

[References 25](#_Toc74602780)

# Introduction

## Background

Smooth transition to problems statement...

## Problem statement

Applying for a loan is a tedious process that demands that the applicants are interviewed by a loan officer. This takes up unnecessary time and resources of both the customers and the bank. Is applied machine learning a way for the applicants to get instant response on their loan applications, without having to physically interact with the bank? Will this yield a positive result for the bank, freeing resources and doing as good of a job as the loan officers?

### Goals and Objectives

The overall goal of this report is to explore the possibility of automating the approval process of loan applications using machine learning.

The following objectives has been set to achieve this:

* Research
  + Evaluate key concepts to be included in the theoretical framework of this paper.
  + Chart the standard application process of loans.
  + Research the existing field of Robot Process Automation.
  + Identify state-of-art solutions in Robot Process Automation. Analyse their evaluations to find elements of interest for this project.
  + Explore similar data sets to the one used in this project to consider further implementation of the model.
  + Investigate the business value of the development of similar systems.
* Development
  + Collect data.
  + Explore the data set.
  + Structure the RPA system.
  + Pre-process data.
  + Create model.
  + Create app for automatic processing of new loan data.
  + Deploy app.
* Evaluation
  + Evaluate data used for training and testing. Evaluate the similarity to other loan data.
  + Evaluate the model for predicting approval.
  + Evaluate the app and its implementations.
  + Evaluate the business value of this system.
  + Compare machine learning vs programming.
* Recommendations and further work

### Limitations

This paper is limited by the data used and the short project scope. The principal of garbage in, garbage out takes full effect in every machine learning project. A model can only be as good as the quality of the data it uses to train. The project spans over three weeks and this will also impact decisions regarding improving the model, and the paper in general.

# Theory and key concepts

Present the chapter………….. usually the main objective is to present key concepts or underlying

## Data driven business development

The Economist published an article named “The world’s most valuable resource is no longer oil, but data”, informing us that the five most valuable listed companies in the world are data companies. Companies such as Google and Facebook have a tremendous value, but their most popular services are free (The world’s most valuable resource is no longer oil, but data, 2017). What is their value?

Their value lies in the data they accumulate about you, the consumer. Google and Facebook, revenue-wise, are mainly advertisement companies, and profits from selling ads to other companies (Sweney & Canon, 2021). Their effectiveness as advertisement companies comes from their ability to tailor ads to different users on their platform, therefore increasing the value of each advertisement. This shows that there is value in data.

Other businesses can take part of this discovered value, by utilizing their data to its fullest potential. Data. Data gathered in a business can used internally to help the business to separate themselves in a competitive ecosystem, providing better products and services to their competitors. Data can also be used commercially by selling it to external businesses (Werger et al., 2020).

Data and physical assets have a lot in common. It can be bought and sold. It makes us more efficient and able to win, and keep, customers. Having it, without using it is a waste. It degrades over time and must be maintained. The big differences are that it does not become less when used and it is easily made accessible to everyone (Treder, 2019).

Data should be treated as every other asset. Give it a price, inventory it, maintain it, refine it and increase its value (Treder, 2019).

### Robotic process automation

“RPA is the technological imitation of human worker with the goal of automating structured tasks in a fast and cost efficient manner”(Aguirre & Rodriguez, 2017). RPA is not a physical robot, but a software created to do repetitive operational processes that usually are performed by humans. RPA can be used to automate processes based on structured data and deterministic outcomes. One of the huge advantages of RPA is that it can implemented on top of already existing systems, across platforms. They are also made to not require programming skills to use them (Aguirre & Rodriguez, 2017).

Some benefits of RPA are as following: Reduced costs by automating processes, increasing productivity and less employees. Better customer experience by freeing up resources, giving more time to focus on customers. Lowering risk by removing the highest cause of errors, the humans. RPA can leverage your existing systems, so you don’t need to replace your infrastructure (Mitra, 2018).

## Machine learning

Machine learning is the application of algorithms that makes sense of data. By using self-learning algorithms, we can turn data into something much more valuable; knowledge, by spotting patterns and making predictions. Instead of relying on humans to make complicated models by making rules, ML offers a faster way of capturing knowledge and provides you with the ability to make data-driven choices (Raschka & Mijalili, 2019).

There are three different kind of machine learning. Supervised learning uses labelled data to train up a model, giving direct feedback to make predictions. Unsupervised learning has no labels or feedback, but works by finding hidden structure in the data. Reinforcement learning is a decision process that works by rewarding wanted behaviour. Reinforcement learning knows nothing to start with, so decisions are random, but every time it does something right, its rewarded, and will therefore do it similarly the next time (Raschka & Mijalili, 2019).

## Loan Application process

Control flow diagram of the loan application process
The loan application summarized by a control-flow analysis:

Figure 1 - Molka, T., Gilani, W. & Zeng, X.-J. (2013). Dotted Chart and Control-Flow Analysis for a Loan Application Process. Business Process Management Workshops, Tallinn, Estonia, pp. 219-220. Berlin, Heidelberg: Springer.

* Application is submitted, but not necessarily finished.
* All applications are either internally pre-accepted, declined, cancelled by applicant, or contacted by call.
* Now that all applications are completed by call, the same decision process is initiated, and the application in finalized.
* Now remains approval, registration, and activation.

After this application process, the loan is offered, and a new process starts around the applicants answer (Molka et al., 2013).

The Norwegian newspaper “Finansavisen” published an article last year about the long processing times of loan applications. You could expect to wait up to three weeks for processing in some Norwegian banks, at the time (Parr, 2020).

## State-of-the-art analysis

Some of the recent papers from the Business Process Management conference in 2020 gives an insight to the relevant state of art of the field of study RPA. The proceeding includes papers such as “A Conversational Digital Assistant for Intelligent Process Automation”, “How to Trust a Bot: An RPA User Perspective” and “From Robotic Process Automation to Intelligent Process Automation”(*Business Process Management: Blockchain and Robotic Process Automation Forum*, 2020).

The papers explore the use of machine learning combined with RPA interfaces to streamline the experiences of business processes. An example is how “A Conversational Digital Assistant for Intelligent Process Automation” solves the inexperience with the use an RPA with digital assistant, much like the ones we have at home, or in our phones. An example used is the simplification of the loan process. The assistant can help a loan officer without experience with machine learning, to automate the process of approving a loan, just by telling the digital assistant some key information (Rizk et al., 2020).

## Summary

Present the key results from this chapter

# Develeopment

Present the chapter…………..

in the first hand-in the goal is that you structure the following sub-sections in this report

## Design description

### Concept Overview

#### Concept

An application for instant response on loan applications.

#### Details

This app will use a model created using machine learning to decide the outcome of loan application. This will streamline the whole process and be beneficial for both the bank and the customers.

#### Benefits

* Freeing resources for the bank by streamlining a time-consuming process.
* Giving the customers instant feedback on their loan applications.

#### Issues

* Hard to understand how the model decides the outcome.
* Lack of customer service.

### System Architecture

Chart, funnel chart

Description automatically generated

Figure 2 – System Architecture Flow Chart.

Chart, funnel chart

Description automatically generated

Figure – System Architecture part 1.

Chart, funnel chart

Description automatically generated

Figure 4 – System Architecture part 2.

## Exploring the data

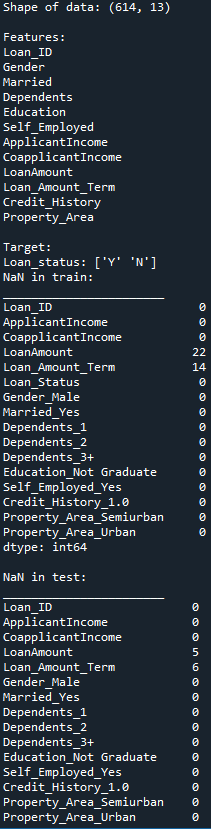
The data used in this project was handed out as part of the TIN200 project material. It contains the information of 614 loan applicants and their loans, and their loans status.

Figure 5 – Features and Missing Values.

Appendix 1, “data\_exploration.py”, is a python script that explores the attributes of the data. The script uses pandas, matplotlib and seaborn to display different characteristics such as missing values, distributions, and correlation.

Figure 5 shows the features the loan dataset contains. It contains information about the applicant’s gender, marriage status, number of dependents, education, employee status, income, co-applicant’s income, amount of loan applied for, terms of payback in months, if they have credit history and their property area.

Figure 5 also shows the number of missing values in the dataset. There are 22 missing values in the LoanAmount colum, and 14 missing values in the Loan\_Amount\_Term column, in the training dataset.

A picture containing table

Description automatically generated

Figure 6 – Histogram showing the Data’s Distribution.

Figure 6 shows the distribution of all the data. We can see that most of the columns are categorical and have only two different values, 0 or 1.

Calendar

Description automatically generated

Chart, bar chart

Description automatically generatedFigure 7 also shows the distribution of the variables in the dataset. The boxplot also gives us an idea of outliers in the dataset.

Figure – Boxplot of Data.

Chart, box and whisker chart

Description automatically generatedFigure 8 shows the distribution between the loan status of each of the categorical variables.

Chart

Description automatically generatedFigure 9 shows the distribution between the loan status of each of the numerical variables.

Figure 8 – Numerical Variables plotted vs Loan Status.

Figure 9 – Categorical Variables plotted vs Loan Status.

Figure 10 shows the correlation between all the variables.

Figure – Correlation Matrix.

## system development

### PreProcessing

Figure 11Text

Description automatically generated shows how the data is imported from csv using Pandas and pre-processed. The target, “Loan Status”, which is either yes or no (Y or N), is mapped to 0 for no, and 1 for yes. The first column, “Loan\_ID”, is removed because it is obviously not important for developing an accurate model. The categorical variables are encoded to multiple binary variables instead because the model does not know the difference between categorical variables. The data is normalized using MinMaxScaler from the scikit-learn library. The missing values are imputed over the median of the other values in each column. The newly pre-processed data is saved as a csv file to be accessed by the model.

Figure – Pre-Processing script.

### Modelling

* Text

  Description automatically generatedPre-processed data is imported.

Figure – Model training script.

* Data is split into y, targets, and X, features.
* The data is split random into 2/3 training data and 1/3 testing data. The testing data is late used to validate the model.
* XGBoostClassifier, or Extreme Gradient Boosting Classifier, is used to make the model.
* The model is fitted with the training data from the earlier split.
* The hyperparameters of the model is tuned using RandomizedSearchCV.
* A confusion matrix is created to show the accuracy of the model in an understandable way.
* The model is trained on all the data, without doing a train/test split.
* The model is exported.

A picture containing text, plaque, crowd

Description automatically generated

Chart, bubble chart

Description automatically generated

Figure – Confusion Matrix.

Figure 14 – Output from console after modelling.

Figure 13 and Figure 14 shows the accuracy of the model. The model has an accuracy of 79.7%. 114 of the approved loan have been successfully predicted. 13 of the approved loan have been misclassified as not approved. 27 of the not approved loan has been successfully predicted. 31 of the not approved loan has been misclassified as approved loans.

### Application

Text

Description automatically generated with medium confidenceFigure 15 shows the imports used in the application script.

Figure 15 – Imports for the application script.

Text

Description automatically generated with medium confidenceThe application is created using Streamlit. Streamlit is used to make an easy-to-use interface to interact with the model and to predict the outcome of a loan application. This can be done without having to directly interact with the Python script.

Figure 16 – The rest of the application script.

The inputted data is collected and added to a test dataset which does not contain target labels. This is done so that the inputted data will fit into a DataFrame with the same number of columns as the model takes.

The data is scaled the same way as the data the model is trained on, and eventual missing values are imputed.

The outcome of the application is displayed in the Streamlit app.

Graphical user interface, application, Teams

Description automatically generatedGraphical user interface, application, Teams

Description automatically generated

Figure 18 – Declined Loan in Streamlit.

Figure – Approved Loan in Streamlit

# Evaluation and results

Present the chapter…………..

in the first hand-in the goal is that you structure the following sub-sections in this report

## Data

Example

## Model

Example

## Implementation

Example

## Business value

Example

# Conclusion and Recommendations

Provide the reader with a reminder of project goal

What has the project group done

What is the key results?

What does the project group recommend based on the work done.

## Further work

Imagine you are in charge of project hand-over to a new project group

Provide a section about further work

* Improvements
* Loose threads

# Attachments

## Source code

Key source code

Link to github:

<https://github.com/hellund/loan_application_processing.git>

# References

Aguirre, S. & Rodriguez, A. (2017). *Automation of a Business Process Using Robotic Process Automation (RPA): A Case Study*. Applied Computer Sciences in Engineering, Cartagena, Colombia, pp. 65-71: Springer International Publishing.

*Business Process Management: Blockchain and Robotic Process Automation Forum*. (2020). Business Process Managemen, Seville, Spain. Cham, Switzerland: Springer.

Mitra, M. (2018). *Robotic Process Automation(RPA) and Benefits*. Web: Mantra Labs. Available at: <https://www.mantralabsglobal.com/blog/robotic-process-automationrpa-and-benefits/> (accessed: 10.06).

Molka, T., Gilani, W. & Zeng, X.-J. (2013). *Dotted Chart and Control-Flow Analysis for a Loan Application Process*. Business Process Management Workshops, Tallinn, Estonia, pp. 219-220. Berlin, Heidelberg: Springer.

Parr, O. S. (2020, 1. juli 2020). Travle boligkjøpere stresser bankene. *Finansavisen*. Available at: <https://finansavisen.no/nyheter/bolig/2020/07/01/7540549/rekordaktivt-boligmarked-gir-lanerush-dnb-beklager-lang-ventetid> (accessed: 10.06.2021).

Raschka, S. & Mijalili, V. (2019). *Python Machine Learning*. Third ed. Birmingham, UK: Packt Publishing Ltd.

Rizk, Y., Isahagian, V., Boag, S., Khazaeni, Y., Unuvar, M., Muthusamy, V. & Khalaf, R. (2020). *A Conversational Digital Assistant for*

*Intelligent Process Automation*. Business Process Management: Blockchain and Robotic Process Automation Forum, pp. 85–100: Springer Nature Switzerland AG 2020.

Sweney, M. & Canon, G. (2021, 28. april 2021). Alphabet: revenue soars for Google owner as Covid lockdown boom continues. *The Guardian*. Available at: <https://www.theguardian.com/technology/2021/apr/27/alphabet-google-revenue-quarterly-earnings> (accessed: 10.06.2021).

Treder, M. (2019). *Becoming a data-driven Organisation*

*Unlock the value of data*. 1 ed. Berlin, Heidelberg: Springer Vieweg.

Werger, K., Kenedy, J., Peckham, D., Mather, S., Ginsberg, R., Jooste, A., Robinson, A. & Knappenberger, D. (2020). Data valuation: Understanding the value of your data assets. 12. Available at: <https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Finance/Valuation-Data-Digital.pdf> (accessed: 10.06.2021).

The world’s most valuable resource is no longer oil, but data. (2017, May 6th 2017). *The Economist*. Available at: <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data> (accessed: 10.06.2021).