Concept Drift in Recommender Systems: Building a MLOps Pipeline that takes Concept Drift into account

**Bachelorarbeit**

im Studiengang  
Wirtschaftsinformatik und digitale Medien

vorgelegt von

**Sebastian Sätzler**Matr.-Nr.: 37635

am 18. Juli 2022   
an der Hochschule der Medien Stuttgart

Erstprüfer/in: Prof. Dr. Jan Kirenz   
Zweitprüfer/in: Prof. Dr. Hendrik Meth

# Ehrenwörtliche Erklärung

„Hiermit versichere ich, Sebastian Sätzler, ehrenwörtlich, dass ich die vorliegende Bachelorarbeit (bzw. Masterarbeit) mit dem Titel: „Concept Drift in Recommender Systems: Building a MLOps Pipeline that takes Concept Drift into account“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§ 26 Abs. 2 Bachelor-SPO (6 Semester), § 24 Abs. 2 Bachelor-SPO (7 Semester), § 23 Abs. 2 Master-SPO (3 Semester) bzw. § 19 Abs. 2 Master-SPO (4 Semester und berufsbegleitend) der HdM) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.“

Aichschieß, 30.4.2022,



# Kurzfassung

Gegenstand der hier vorgestellten Arbeit ist eine Dokumentvorlage für Abschlussarbeiten und andere wissenschaftliche Arbeiten (z.B. Bachelorarbeiten, Masterarbeiten, Diplomarbeiten und Studienarbeiten) an der Hochschule der Medien (HdM) Stuttgart. Die Dokumentvorlage basiert auf den Richtlinien zur Erstellung von Abschlussarbeiten in der Fakultät Information und Kommunikation, sie ist aber ohne weiteres über den Fachbereich hinaus innerhalb und außerhalb unserer Hochschule nutzbar und kann für eine Vielzahl wissenschaftlicher Arbeiten und Berichte verwendet werden. Die Dokumentvorlage stellt ein Angebot dar, das von den Studierenden genutzt werden kann, dessen Verwendung an der HdM aber nicht verpflichtend ist. Die existierenden Regelungen zu Abschlussarbeiten bleiben dabei unberührt. Diese Dokumentvorlage wurde zur Verwendung im Textverarbeitungssystem Microsoft Word erstellt. Die hier vorliegende Arbeit ist selbst mit dieser Dokumentvorlage geschrieben und kann in formaler Hinsicht als Muster für die Abfassung von wissenschaftlichen Arbeiten verwendet werden. Auf diese Weise lässt sich die Einhaltung der für wissenschaftliche Arbeiten geltenden Formatvorgaben weitgehend automatisieren, wodurch sich die Qualität der wissenschaftlichen Arbeiten hinsichtlich formaler Kriterien erhöht und sich der Beratungsaufwand verringert.

**Schlagwörter**: Dokumentvorlage, wissenschaftliche Arbeit, Bachelorarbeit, Masterarbeit, Diplomarbeit, Hochschule, Textverarbeitungssystem, Microsoft Word

# Abstract

A style sheet for theses (e.g., master theses, bachelor theses, diploma theses) is being presented. This style sheet may be used by any student, its utilisation, however, is not obligatory in our university. The style sheet is designed for the Microsoft Word text processing system. This document itself is written by using the developed style sheet and can be used as a template for the production of theses. In this way, the fulfilment of the existing formatting guidelines will be automated to a large extent, thus raising the quality of theses with respect to formal criteria as well as reducing the need for training and consulting.

**Keywords:** Concept Drift, MLOps, Recommender System, Machine Learning

# Table of Contents

Ehrenwörtliche Erklärung 2

Kurzfassung 3

Abstract 3

Table of Contents 4

Table of Figures 5

List of Tables 5

List of Abbreviations 6

1 Introduction 7

2 Motivation and Structure 9

2.1 Environment 9

2.2 Artifact 10

2.3 Procedure 12

2.4 Challenges 13

3 Literature Review 15

3.1 Design Science Research 15

3.2 Recommender Systems 18

4 Conclusion and Outlook 19

Anhang A: Beispiele für die Gliederung von Abschlussarbeiten 20

A.1 Literaturarbeiten 20

A.2 Systementwicklungen 20

Anhang B: Formatvorlagen 21

Index 22

# Table of Figures

Figure 1: procedure of artifact development 13

# List of Tables

Table 1: Artifact specification table 11

Table 2 12

# List of Abbreviations

ML Machine Learning

GPU Graphics Processing Unit

CPU Central Processing Unit

RS Recommender System

CD Concept Drift

TFX TensorFlow Extended

TFRS TensorFlow Recommenders

DCN Deep & Cross Network

DD Data Drift

STEM Academic disciplines of science, technology, engineering, mathematics

PoC Proof of Concept

IS Information System

SotA State of the Art

AI Artificial Intelligence

API Application Programming Interface

CF Collaborative Filtering

CB Content-Based Recommender System

TF-IDF Term Frequency-Inverse Document Frequency

MF Matrix Factorization

DL Deep Learning

NN Neural Network

# Introduction

## Preface

Over the last two decades Machine Learning (ML) has become one of the fastest growing technical fields with an estimated growth of 21% in 2022 compared to its previous year (Rimol, 2021).

It managed to secure a position as one of the top fields in computer science for scientific research as well as enterprise adoption. ML combines concepts of linear algebra and statistics and applies them to large datasets to find patterns and generalizations in the data, which can be used to make predictions or classifications.

Leveraging these complex algorithms with the computational power of modern GPUs and CPUs, ML has seen application in a large variety of sectors ranging from medicine for diagnostics, to transportation for self-driving cars and e-commerce for shopping cart optimization (Choy et al., 2018). The latter sector employs so called Recommender Systems (RS) with the goal of suggesting products that coincide with the taste of the customer. With the advent of e-commerce, RSs have gained increasing interest from academia and especially the enterprise sector (Singh, Choudhury, Dey, & Pramanik, 2021). RSs serve a major role for large tech corporations in engaging, retaining and enticing the user-base of their platform (Jannach & Zanker, 2022). Netflix for example uses its own RS to suggest its users what movies they might be interested in. In order to incentivize research in the field of Recommender Systems, Netflix introduced the Netflix Prize in 2007: A dataset was made public with users and their movie ratings. The goal was to create a RS that would beat Netflix’ own RS at that time. For this challenge a prize pool of $1.000.000 was written out (Bennett, Lanning, & others, 2007). To this day datasets of movie ratings remain a popular way to benchmark RSs.

Despite the wide use and success of Recommender Systems and Machine Learning in general, it still is a relatively new field with a lot of research opportunity (Jordan & Mitchell, 2015). While Recommender Systems are considered integral to many online-platforms, their precision and accuracy often lack in comparison to other ML fields. This is, among other things, due to the nature of the data that Recommender Systems work with, which is often sparse (Khusro, Ali, & Ullah, 2016). Consequently Recommender Systems are especially susceptible to bad data quality and therefore could profit from comprehensive data curation and monitoring. This lends itself to take a data-centric approach when building, deploying and maintaining a RS, which is one of the subject matters that the field of MLOps sets out to tackle (Miranda, 2021).

MLOps emerged from the paradigm of DevOps and seeks to apply an automated and standardized approach to the lifecycle of ML applications, similar to what DevOps does for conventional Software. MLOps is attuned to the specific needs and problems of Machine Learning, such that its practices vary from those of DevOps, while still sharing the same goal of rapid and frequent deployment of Software (Makinen, Skogstrom, Laaksonen, & Mikkonen, 2021). The effect of data quality on the ML model presupposes that data quality management is an integral part of every MLOps system, since data quality affects all aspects of the machine learning lifecycle (Renggli et al., 2021). Detrimental data to the ML system’s performance can manifest itself in different ways.

One manifestation is concept drift (CD), which describes a changing outcome *y* to a constant input *x* over time (Lu et al., 2018). Real world examples of CD could be changing house prices due to a fluctuating house market, or people changing their taste in movies because of aging or genre trends. Deteriorating RS performance due to CD can directly impact the health of the online platform it is used on, as outputs of RSs are generally reciprocated back to the user experience. For instance, if a movie streaming platform stops recommending appropriate movies to a user because it failed to adapt to the change in taste, the user might stop watching movies on that platform and eventually cancel their subscription. Issues of Concept Drift need to be addressed and mitigated to ensure user-base retention for online services. Additionally, it needs to be incorporated into a MLOps system to benefit from the maintainability, consistency and automation of a unified process.

The product of this work, called an artifact, will be the implementation of a concept drift-aware MLOps pipeline for a RS. CD-awareness meaning, that it possesses the ability to account for potential CD in the data.

This paper serves as a thorough documentation of the design of the artifact, which is based on a comprehensive dissection of scientific literature touching the topics of Recommender Systems, MLOps and Concept Drift. The result will then be qualitatively evaluated and discussed.

This research follows the design science research (DSR) methodology of Alan R. Hevner (Hevner, March, Park, & Ram, 2004).

## Research Methodology

Design science is a research paradigm that emerged as a differentiation to natural science in STEM. Natural science, also referred to as behavioral science, is associated with fields like mathematics, physics, biology and chemistry. Its research methodology follows the objective of uncovering facts and theories about reality. Juxtaposed to the natural science lies the design science. Instead of uncovering rules and theories about the nature of reality, design science sets out to engineer and create artifacts with tools from scientific literature. Design science is predominantly represented in the engineering and computer science fields, where proof of concepts (PoC) and prototypes are the result of a lot of academic works. Both behavioral science and design science have distinguished approaches on how to conduct research.

Design science research contains a set of frameworks and best practices to manage academic work in the design science department. One of the more prominent methodologies is Alan R. Hevner’s “three cycles” of DSR (Hevner et al., 2004). Hevner originally designed his framework to involve the research aspect more closely to the development process of Information Systems (IS) in enterprise environments. It consists of 3 cycles which are closely related to each other and serve to build an artifact. The three cycles are what Hevner argues separates design science from other research paradigms (Hevner, 2007).

The artifact is the eventual product of the academic work using DSR. Since its first publication in 2004, DSR has found application in a wide variety of fields that surpasses conventional engineering and computer science. This means that the term “artifact” has a broad definition and is consequently hard to delimit. Generally, an artifact means anything that emerges from design science research. It could range from a theoretical model that was derived from other academic work, to a physical prototype or a production-ready software system.

The goal of DSR is to create an innovative artifact, which incorporates both theoretical-scientific, as well as the practical-environmental (e.g. business) aspects into its design. The iteration through the three cycles creates a mutual feedback-loop between the artifact and the science and business environment (Hevner et al., 2004). The result is an artifact, which is attuned to the business needs of an enterprise, while also holding scientific value and enriching the academic field with new insights and findings.

In the following, the three cycles will be elaborated in more detail.

1. **The relevance cycle**: The relevance cycle initiates the DSR process. In this cycle all requirements relevant to the artifact and the research are worked out. First, a problem is defined and opportunities and arguments are laid out supporting research to resolve the problem with an artifact. Since DSR has its roots in the enterprise sector, it is vital to map out and contextualize the environment this research takes place in, as it directly influences the design of the artifact. In order to evaluate the artifact, acceptance criteria need to be defined. This way a conclusion can be made whether the artifact succeeded in its goals to improve the environment or not (Hevner & Chatterjee, 2010). In this work the results of the relevance cycle are to be found in the introduction of this paper.
2. **The rigor cycle**: Following the relevance cycle, comes the rigor cycle. While the relevance cycle establishes the requirements for the project, the rigor cycle introduces the methods, drawn from scientific literature, used to create the artifact (Hevner & Chatterjee, 2010). This so called *knowledge base* consists of engineering methods and scientific theories and sets the foundation from which the artifact will be designed and built from. A thorough rigor cycle ensures that the artifact is grounded on state-of-the-art literature from the academic field. This establishes the connection to other scientific contributions and thus sets it apart from routine designs and routine design processes (Hevner & Chatterjee, 2010). The rigor cycle gives the artifact the scientific weight it requires to be acknowledged as an academic contribution, consequently it is vital that the design of the artifact draws sufficiently from the knowledge base of the rigor cycle. The knowledge base is referenced in the State of Research of this paper
3. **The design cycle**: “*The internal design cycle is the heart of any design science research project.*” (Hevner & Chatterjee, 2010) The design cycle is the culmination of the relevance and the rigor cycle. The information acquired from the two prior cycles will now be deployed to design and implement the artifact. Hevner points out that it is not possible to retain both maximum relevance and rigor simultaneously, thus a balance between both need to be struck (Hevner & Chatterjee, 2010). In the context of the design cycle the rigor represents the actual construction of the artifact, meaning the implementation of the knowledge base gained from the rigor cycle. In opposition to the rigor stands the relevance. The relevance represents all the requirements and evaluation criteria that were specified in the relevance cycle. The discrepancy between the relevance and the rigor gets resolved by the artifact, which is the bridge between both and thus constitutes the business and scientific contribution. The design cycle is documented in the Artifact design chapter of this paper.

DSR is a non-linear process. With progression of the research project, the relevance, rigor, and design cycle can change as new insights are garnered. Through iterative cycles Hevner’s DSR methodology accounts for the often unpredictable nature of the artifact creation process. Should either parts of the relevance, rigor or design fall out of line with the current state of the project, it needs to be updated by reiteration (Hevner et al., 2004).

Once the research is conducted and the artifact is created it itself becomes part of the knowledge base, whose insights can now be used for other research projects.

# Current Environment & State of Research

## Environment

Machine Learning serves great value to businesses. In 2017, Netflix for instance claimed an estimated saving of $1 billion through their use of RSs (Columbus, 2017). Since 2017, ML algorithms became more sophisticated and hardware more powerful to make Artificial Intelligence (AI[[1]](#footnote-1)) operations more efficient, effective and in return more profitable. Gartner calculated the revenue of the AI software market to be over $51 billion in the year 2021 with a prediction to surpass $60 billion by the end of 2022 (Rimol, 2021). Open-source software, ML cloud services and an active community make AI more accessible to a wide variety of businesses. These developments make a growing number of institutions consider optimizing, augmenting, or even reinventing their current operations with ML. McKinsey’s “The state of AI in 2021” reported that 56% of their surveyed businesses have adopted ML and AI in at least one of their business functions. An increase of 6% compared to the preceding year (Chui, Hall, Singla, & Sukharevsky, 2021). It’s apparent that ML and AI receive increasing interest in the enterprise sector.

According to a survey conducted by Refinitiv, out of 447 international institutions that use ML, only 46% have deployed AI in multiple areas and are core to its business, whereas 44% deployed ML in pockets, while the remaining 10% were still prototyping and investing in its infrastructure (Baker, 2019). This indicates that a majority of enterprises, while considering or using ML for their businesses, struggle to embed it into their existing infrastructure. This observation is also supported by Algorithmia’s 2020 report on enterprise machine learning, which uncovers that 55% of companies “*actively developing machine learning lifecycles or […] beginning their machine learning journey*” (Algorithmia) have yet to deploy a machine learning model. This report highlights that a lot of the main difficulties of ML lie in its operational aspect, such as reproducibility, versioning of models and scaling of the ML system. This leads to “*unreasonably long roads to deployment*” (Algorithmia) and impedes evolving the ML system to higher levels of maturity (Algorithmia). Another Refinitiv study also identified lacking data quality to remain the biggest challenge for ML and data science (Refinitiv, 2020).

The aforementioned McKinsey whitepaper made the observation that companies most successful with AI were employing advanced operation procedures such as MLOps, as well as putting greater effort into mitigating and reacting to “*AI-related-risks*”, such as concept drift. This applies to the operation of RSs as well. Operationalizing the ML lifecycle and ensuring high model quality through data quality assurance are essential to a RS in a productive environment. A data-centric approach to RSs combined with the application of MLOps practices should prevent or alleviate unsatisfactory RS performance while ensuring maintainability and reproducibility to an mostly automated ML process. Especially the phenomenon of concept drift needs to be accounted for as it is an inevitable occurrence for a majority of real world data. A key challenge for RSs is its often unpredictable nature and sudden appearance, which can be detrimental to business operation. Shift in the data can be subliminal as it is not tangible in most cases and therefore can go unnoticed by Data Scientists and ML engineers. Consequently the effects of CD can first become apparent through degradation of model performance. Scalable and automated ways to account for change in ML data still are in their infancy and therefore require further research and development to engineer solutions. In the long run a state-of-the-art ML infrastructure will make RS operations more profitable for businesses through labor reduction and performance improvements.

This introduction poses the baseline for the relevance cycle of Hevner’s design science research method. The collected insights from the whitepapers serve to map out the environment and establish the motivation for this research. The acceptance criteria for the later artifact will be in part derived from the findings and motivation in this chapter, which will be elaborated in the following chapter.

Having established the environment and motivation, the following research question is formulated for this work: “*What can a MLOps pipeline for a recommender system, that takes concept drift into account, look like?*”

## Recommender Systems

### Overview

Online platforms like Amazon, Netflix and Facebook sit upon massive amounts of data garnered from their user-base through interactions on their websites and applications. The internet enabled firms to leverage information about customers in a large-scale, centralized and automated fashion. In the past, in order to collect insights about customers, polls had to be conducted on sight of a physical store. The results were costly polling ventures of small sample sizes over a limited time period. Nowadays large quantities of high quality data is streamed uninterruptedly into data lakes where they are stored. This unprecedented amount of data enables data science and machine learning to thrive. While many large corporations live in abundance of user data, it still remains a main challenge for them to gain useful insight from the data (Cai & Zhu, 2015).

Recommender Systems use the given data to uncover preferences of end-users to then tailor the service to their likings. The most known examples are shopping basket recommendations on e-commerce sites like Amazon or movie and video recommendations of streaming platforms like Netflix and YouTube. As a generalization, the term *item* is used to denote any object in a set of items that can be recommended to a user, such as a product or a movie. The recipient of an item is referred to as the *user* or *query*. The user-item relationship makes up the foundation of any recommender system, as RSs work under the assumption that there exists dependencies between user- and item-centric activities.

The main task of a RS is to correctly identify these dependencies and use them to match appropriate items and queries (Aggarwal, 2016). In order to reduce information overload on websites with video and product catalogues, RSs are applied to identify items that will interest the user (Alyari & Jafari Navimipour, 2018). Aggarwal (Aggarwal, 2016) lists four aspects that define a good recommendation:

1. *Relevance*: The most fundamental objective of a RS is to recommend items that are relevant and interesting to the individual user.
2. *Novelty*: The RS should introduce items to the users that they weren’t aware of before. It has been shown that recommender systems can negatively impact the sales diversity of e-commerce, when only popular items are recommended to users (Fleder & Hosanagar, 2007).
3. *Serendipity*: Similar to the point above, RSs should have the ability to surprise users with somewhat unexpected but relevant items. While novelty focuses on the quantitative aspect of recommending items that aren’t on the radar of the user, serendipity considers the qualitative-psychological impact of a surprise in the recommendation.
4. *Increasing recommendation diversity*: This last point envisions that RSs should ensure a diverse set of items in a recommendation feed. This should be done to not fatigue the user with items of similar categories. An example would be the recommendation of movies from a variety of different genres.

Throughout the years, different approaches have been developed to optimize recommendations and overcome obstacles in this field. These recommendation systems are broadly categorized in content-based RS (CB), collaborative filtering (CF), knowledge-based RS and hybrid RS (Aggarwal, 2016). In the following, each category will be introduced to provide an overview of the RS landscape. Collaborative filtering will be elaborated in more detail, as it is a key concept for the RS implemented in the artifact. Figure 2 shows the structure and an overview of the recommenders covered in this paper. Highlighted in red is the type of recommender system that will be used in this this research project, which is a deep and cross network.



Figure 1: structure and overview of RSs derived from Aggarwal (2016)

### Data

The data used by RSs often differ from those of other Machine Learning tasks, as it is often sparse. Sparsity refers to very few user-item interactions relative to the size of the whole dataset. Assuming one would construct a matrix of an e-commerce website, wherein represents the entirety of its user-base and represents all the items in the product catalogue. Every time a customer buys or inspects a product, an interaction between the user and the product is denoted in . Since customers only interact with a fraction of the offered products in an e-commerce web shop, there are very few entries in the 2-dimensional matrix, hence it is sparse. This is a fundamental problem that RSs face and try to overcome.

The value of an interaction between item and user can be represented in different ways depending on the context of the recommendation. Is an interaction defined as a review with a rating between 1 and 5 stars, it will be marked with an integer value in a range of 1 to 5. These rating types allow the user to quantifiably express their like or dislike of an item. These are called interval-based ratings (Aggarwal, 2016). Another popular rating type is the unary rating system. These are used on social media platforms like Instagram and Twitter. These ratings are encoded in a binary fashion, wherein an interaction (e.g. like) will be conventionally denoted with the value 1 and abstinence will be denoted by the value 0 (Aggarwal, 2016).

Both the like-feature as well as the 5-star rating system are called explicit ratings, as they are actively and consciously given by a user. While explicit ratings generally provide reliable information about an users opinion, it remains a challenge to entice users to rate. This in return makes the data more susceptible to shilling attacks, where item ratings get skewed by coordinated bombardments of overly positive or negative ratings by external sources (Khusro et al., 2016).

In opposition to explicit ratings, lie the implicit ratings. These rating systems derive a rating from the user through their behavior. An examples could be the watch time of a movie. This example works under the assumption that people that don’t close a movie are engaged with it. The drawback of an implicit rating is its ambiguity, as online behavior needs to be interpreted first and then translated into a rating. User behavior can be interpreted wrongly and skew the data. Continuing the aforementioned movie example, a user who slept during the movie out of boredom, would be misinterpreted as a user who was engaged throughout the whole movie, since they didn’t close it. An implicit rating system has the advantage that it is not reliant on the user to give feedback, which generally results in more ratings. This is a significant factor considering sparsity is a main challenge for RSs (Su & Khoshgoftaar, 2009).

In a recommender dataset the value of a user-item interaction is referred to as a *sparse feature*. They can be augmented with so-called *dense features*. Whereas sparse features only occur through user-item interaction (e.g. a rating), dense features are specific attributes that describe the item or user at hand, comparable to features in other ML tasks.

Data understanding plays a vital part in the success of recommender systems. There are many variables ranging from its rating system over its features and sparsity that determine, which recommender system is the most suitable for the given task (Aggarwal, 2016, p. 128).

### Content-based recommender systems

Content-based methods make recommendations based on a user's past interactions with items. These users are referred to as objects and contain information from their history of item-interactions (e.g. movie watches, ratings, etc.) (Lü et al., 2012). These items have descriptive attributes, named content, which form the basis of which these recommender systems derive their recommendation decisions from (Aggarwal, 2016). These contents are used to picture the preferences of a user. In a movie recommendation setting such a content could be the genre of a movie. Depending on what contents are found in an object, certain assumptions can be made about a user's likings. Other items can then be assigned to the user based on a similarity measure between the object and other item objects. Alternatively new items can be suggested by weighing the content by their frequency in an object. Popular methods for content-based filtering are information retrieval algorithms like TF-IDF (term frequency-inverse document frequency) (Lü et al., 2012). One characteristic of CB is that both the item and the user are represented in the same feature space, in order to measure similarities between an user object and all item objects.

Item recommendations can be done independently from other users, as it is only of interest how the individual user is positioned to the items in the vector space, which is one of the advantages it has over CF. This means that CB can be applied in environments with relatively small user-bases. Since CB relies on a user history to make recommendations, it has challenges finding the right recommendations for new users, since there isn’t enough content to map out the user (Aggarwal, 2016). Conversely, making recommendations for new items is relatively easy, since the item objects only need to be provided with appropriate content. The success of content-based filtering is highly determined by the quality of attributes in these objects. These features need to be cautiously engineered. Since the predictions are calculated on the basis of the user’s and item’s attributes, it is not possible for these algorithms to uncover hidden variables in the data, like for instance some CF algorithms could. In general, the performance of CBs are very reliant on how the data has been engineered by data and ML engineers.

### Collaborative filtering

CF has been the baseline for RSs for years. In September 2009, the BellKor team was awarded with the grand Netflix prize, which it has won by applying a CF model on the competition dataset (Koren, 2009). As the name implies, CF uses the "collaborative" power of the user-item interactions in the dataset (Aggarwal, 2016). Instead of solely relying on item and user descriptions, CF uses the composition of the dataset to determine the output of a recommendation. Colloquially, one would say that the recommendation is driven by a "joint effort" of every user-item interaction. Unlike CB, recommendations can be made without relying on hand-engineered features (Google LLC, 2021).

**Memory-based filtering**

CF can be divided into *memory-based* methods and *model-based* methods. The prior uses neighborhood based algorithms to make recommendations. These methods belong to the earliest CF practices, where ratings are predicted based on similar users or items (Aggarwal, 2016). "Memory-based" relates to the fact, that the dataset is loaded into memory, in order to make calculations. The dataset is then actively used to compute the right recommendation through nearest neighbor methods. For that reason the data in memory-based CF can be described as part of the model, as recommendations are determined by the arrangement of datapoints within the vector space.

Memory-based methods can either take a *user-based* or *item-based* approach. In user-based CF, rating predictions are provided by like-minded users. Similar user groups are identified using nearest neighbor techniques, then the unobserved ratings are filled in by the values of the peer group. This could be done by calculating a weighted average of all ratings from that group (Aggarwal, 2016). In item-based CF a set of items a user has rated most similar to an unrated item are retrieved with a neighborhood algorithm. In item-based CF a set of items, already rated by a user, are used to determine the value of an unspecified item. Via neighborhood algorithms items are determined that resemble the unrated item the closest. Like with user-based CF, the rating is derived from the neighboring entities, only that instead of using the values from similar users, the values from similar items are chosen (Aggarwal, 2016).

The advantage of a memory-based method is its interpretability, as ratings are directly taken from other users or items. Memory-based algorithms don’t work well with sparse data, as there might not be sufficient items or users similar to each other (Aggarwal, 2016).

**Model-based filtering**

Unlike memory-based approaches, model-based methods have a training phase. The goal is to create a function (the model) that approximates the results of an item-user interaction, which is conventionally achieved through optimization of a loss function with gradient descent algorithms (Koren, Bell, & Volinsky, 2009). Two of the most prominent model-based CF techniques are matrix factorization (MF) and deep learning (DL).

The former has been the standard in the field of RS and was also utilized by the winning team of the Netflix prize in 2009 (Koren, 2009). MF is foremost a dimensionality reduction technique, with which the user-item matrix gets decomposed into *low-rank latent factors*. These latent factors are the components that make up the *latent factor model* (LFM), which is a low dimensional representation of the initial matrix. LFM work under the assumption that there are underlying latent variables in the data. Latent variables are not directly observable but can be inferred through mathematical computation like for example MF (Loehlin & Beaujean, 2017).

### Deep Neural Network

### Deep & Cross Networks

### TensorFlow Recommenders

## MLOps

### Overview

### Concept Drift

### TensorFlow Extended

# Goal & Specification

## Artifact

The goal of the DSR methodology is the design and creation of an artifact. This particular work sets out to create an automated machine learning pipeline for a RS, the main research focus being the implementation of CD-awareness into the pipeline. For evaluation purposes, a specification sheet will be generated on which the artifact will be measured against. The artifact will be compared with the features listed in the specification. Based on the specifications, the final product will be analyzed, hurdles during the design and development will be identified, and the success of this research will be determined. The specifications are structured in *base specifications* and *research specifications*. While both types of requirements are integral to the whole research project, the research requirements have a direct connection to the research question at hand. During the DSR process both requirement types receive equal prioritization as both requirement types make up the whole artifact. A table with all specifications can be seen in *table 1*.

The base requirement list consists of creating a base pipeline (without explicit CD-awareness) and implementing the recommender system. The base MLOps pipeline will be realized with TensorFlow Extended (TFX), a package with various tools to orchestrate and monitor the ML lifecycle (Baylor et al., 2017). As for the RS a Deep & Cross Network (DCN) will be implemented with the TensorFlow Recommenders Python API (TFRS). MovieLens 25M will be used as the dataset, a collection of 25 million movie scores with 62.000 movies and 162.000 users (Harper & Konstan, 2016).

Table 1: Artifact specification table

|  |  |  |
| --- | --- | --- |
| **Specification type** | **Specification** | **Description** |
| Base specification | Recommender system | The artifact has deep & cross network RS made with TFRS. |
|  | Base pipeline | The artifact has a TFX pipeline for the RS. |
| Research specification | CD detection | The artifact has the ability to detect CD. |
|  | CD understanding | The artifact has the ability to extract information out of CD. |
|  | CD adaptation | The artifact has the ability to react to CD. |
|  | Integrated pipeline | The final artifact is one pipeline that can be executed in one go |
|  | Automated pipeline | The pipeline can finish without human intervention |

The research specifications have three criteria related to concept drift, which are taken from the 2018 paper “Learning under Concept Drift: A Review” (Lu et al., 2018):

*Concept drift detection* specifies the technical ability for the artifact to register concept drift in the data. An example would be the implementation of a concept drift detection algorithm.

*Concept drift understanding* specifies the ability of the artifact to retrieve additional information from the concept drift. This criteria serves to give concept drift more tangibility and helps data scientists identify the source of concept drift. The most rudimentary implementation of this specification would be a timestamp for every occurring CD. A timestamp for CD is also the minimum requirement to realize an automated pipeline, as in order to initiate concept drift adaptation the system needs to know when CD occurs.

*Concept drift adaptation* specifies an action that accounts for concept drift, like starting a retraining of the model.

Each of these components amount to what we define in this paper as CD-awareness. All three specifications will be explained in further detail in the literature review.

Besides the concept drift specifications, there are also pipeline conditions in the research specifications that envision the artifact to be one unified and automated process.

These specifications serve as an aide during the design and development of the artifact. Rather than evaluating the artifact based on how many specifications it technically fulfilled, the whole artifact will be qualitatively evaluated using the specification sheet as a guideline. In the scope of this work, one solution to CD will be implemented that is derived from the literature. It is not a comparison and evaluation of various implementations of CD-awareness. This work does not entail a quantitative evaluation of the artifact, as the focus lies on a qualitative analysis of the prototype. Lastly, the concept of Data Drift (DD) is not subject of this work and thus only will be touched in the context of Concept Drift.

## Procedure

Throughout this project various tools are used to track, organize and document this work. As part of this research, a GitHub repository was created, which contains both the thesis Word document and the software artifact.[[2]](#footnote-2) GitHub and Git are used for version control and enable work on different systems. It also serves to make the development of this research transparent and traceable. Internally, a GitHub Project Kanban board is used for project management. There, the project is broken down into individual tasks, all of which have their progress tracked. This way it is possible to gain an overview of the current status of this project and plan future development steps.

The overarching process of the artifact creation is structured in underlying phases (Figure 1). The initial phase of this project is dedicated to the set-up of the artifact environment. This phase encompasses every necessary step to build a software environment on which the artifact will run on. The individual tasks would consist of choosing the python version, setting up a work directory and installing needed packages and other software. After setting up the environment, a simple base TFX RS pipeline is implemented. This pipeline serves as the foundation from which the research artifact is built upon. The next phase involves incrementally expanding the baseline with the remaining TFX components to build a more comprehensive pipeline. This step might also uncover potential compatibility issues between TFX components and the TFRS Recommender System.[[3]](#footnote-3) Depending on what components work and don’t work, design changes to the artifact need to be made. In the next phase the base RS will be replaced by a state of the art deep & cross network (DCN) RS. Having implemented a more sophisticated pipeline, an intermediate evaluation of the current pipeline will be done. The objective in this phase is to find out the capabilities of this pipeline in relation to the research question. What components can be leveraged to fulfill the objective of this work? Accomplishing these phases should fulfill the base specifications detailed in the previous chapter. The following phase consists of incorporating CD-awareness into the pipeline. With the knowledge base from scientific papers and information about the environment from the whitepapers, a solution to CD will be engineered. This process will run iteratively and in accordance with the DSR methodology. Each iteration of this phase concludes with an evaluation, which will determine whether another iteration is initiated. Once the artifact reaches a state where no additional iterations are done, a final evaluation of the artifact will be made, which concludes this research project.

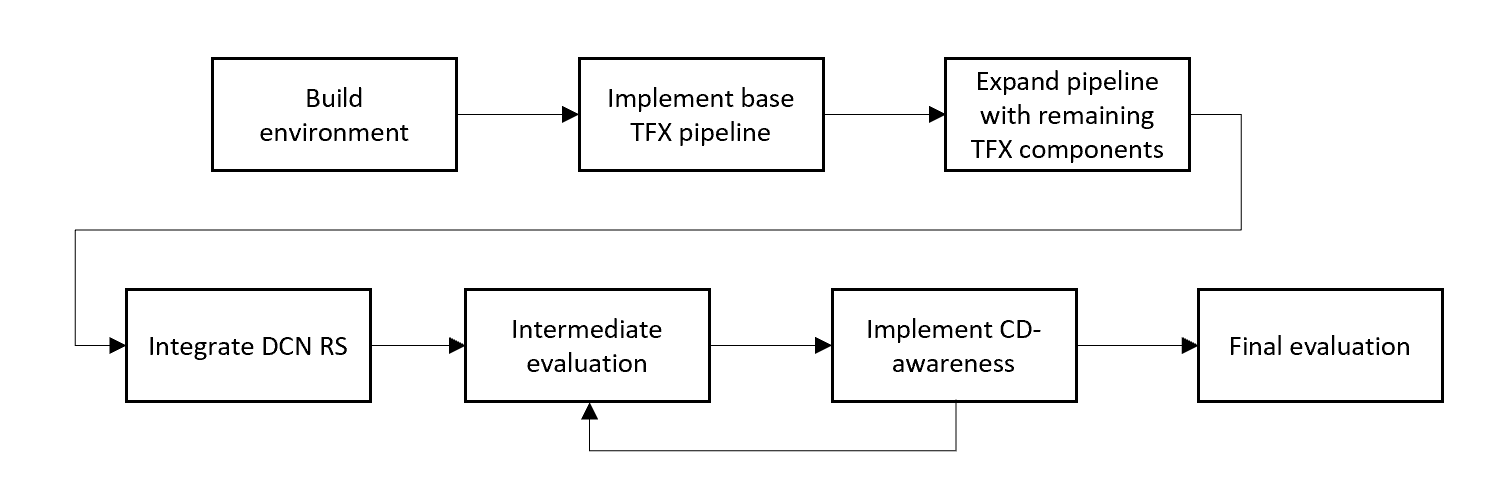


Figure 2: procedure of artifact development

## Challenges

This work is not done in association with a company, despite being a study directed towards businesses exploring a potential solution to CD in an AutoML scenario. Consequently this research can't be conducted in an existing enterprise infrastructure. This entails the challenge that, first, an environment needs to be built approximating a productive business setting. For instance, instead of using real world data from a data lake or data warehouse, this research needs to revert to publicly available datasets. Simulating a productive MLOps environment means that there is a discrepancy between this work's environment and an organization's environment, which might impact applicability of the artifact.

The open-ended nature of this work’s research question is another source of challenge that is inherent to DSR projects. In Hevner’s DSR methodology it is only specified what problem the final artifact should solve, not how it should look like, as it is part of the process to come up with an innovative design. This room for interpretation of what the artifact should look like makes it more difficult to quantify the artifact in the end. For this reason a specification has been created, which should make the idea of the artifact more concrete. Even with a specification sheet, the challenge remains on how to meaningfully evaluate a ML pipeline. Unlike ML algorithms, there is no reference pipeline to measure the artifact against, assuming that using quantitative metrics like processing time would make sense for this work in the first place. Another aspect that hinders quantifiability of such an artifact is the phenomena of CD itself. As mentioned in the environment chapter, CD is a burden for businesses because it lacks tangibility and is therefore hard notice, let alone isolate and quantitatively measure. This project needs to deal with these limitations and find a way to evaluate the artifact in a value adding manner.

Another hurdle is a lack of scientific studies and official documentation on this specific subject, as the operationalization of ML systems only received mainstream attention in the last few years with the advent of terms like AutoML and MLOps, or tools like TFX. TFX had its first release in May 2019 (Google LLC), whereas TFRS had its initial release in fall of 2020 (Google LLC). Only in end of March 2022 has Google released its first TFX documentation using TFRS (Google LLC, 2022). As of conducting this project, more fundamental documentation is planned for both tools (Wei, 2022). It would be defensible to say that both tools still are in their infancy, going by their recent releases, their ever expanding documentation and active development. Besides the novelty of these fields, a lot of research and development in ML systems goes unnoticed by the academic community, as they are conducted behind closed doors and used as proprietary software. In 2019, Naumov, et al. stated the following about the current environment of RS research: “*Although recommendation and personalization systems still drive much practical success of deep learning within industry today, these networks continue to receive little attention in the academic community*” (Naumov et al., 2019)

# Design & Development

# Evaluation

# Conclusion

## Design Science Research

## Recommender Systems

### Overview

# Conclusion and Outlook

Mit der hier vorgelegten Dokumentvorlage steht ein Werkzeug zur Verfügung, das es möglich macht, auf relativ einfache Weise die Textgestaltung für Abschlussarbeiten und andere wissenschaftliche Arbeiten zu automatisieren. Dies spart Arbeitszeit, Beratungsaufwand und erhöht zugleich die Qualität des Ergebnisses hinsichtlich formaler Kriterien.

Gleichzeitig unterstützt die Dokumentvorlage die Lehre in Fächern *wie Arbeits-, Lern- und Präsentationstechniken* und *Vorbereitung auf die Bachelor- bzw. Masterarbeit*,

Die Dokumentvorlage wurde im Sommer 2000 in der vorliegenden grundlegend überarbeiteten Version zum ersten Mal hochschulweit benutzt, nachdem bereits die Vorgängerversion im Jahr 1999 für drei Abschlussarbeiten erfolgreich eingesetzt wurde.

Erweiterungen der Dokumentvorlage und der darin enthaltenen Anleitung zur Erstellung von Abschlussarbeiten mit dem Textsystem Microsoft Word sind auch für die Zukunft geplant. Das Feedback aus der Nutzung der Dokumentvorlage für Abschlussarbeiten an der HdM wird dabei fortgesetzt aufgegriffen und zur Verbesserung genutzt werden.

# Anhang A: Beispiele für die Gliederung von Abschlussarbeiten

Die nachfolgenden Gliederungen stellen lediglich Vorschläge dar, die stets am konkreten Fall überprüft und in der Regel angepasst werden müssen.

## A.1 Literaturarbeiten

1. Überblick (oder: Zusammenfassung, „Executive Summary“, alles Wichtige für den „Manager“ oder Schnellleser)
2. Fragestellung (oder: Ziele, Ausgangspunkt, Motivation)
3. Übersicht über den Stand der Wissenschaft und Technik (Beschreibung der Lösungsansätze, Beispiele etc. in einzelnen Abschnitten)
4. Bewertung der einzelnen untersuchten Ansätze, Beispiele etc., Identifikation von Defiziten
5. Synthese: Erstellung einer Gesamtschau, allgemeine Prinzipien, Beschreibung einer eigenen Sicht auf das Problem, evtl. auch eigene Vorschläge
6. Zusammenfassung (Erklärung des Nutzens), Ausblick

Anhang: eventuell recherchierte Texte, Produktbeschreibungen, etc.

## A.2 Systementwicklungen

1. Überblick (oder: Zusammenfassung, „Executive Summary“, alles Wichtige für den „Manager“ oder Schnellleser)
2. Problemstellung (oder: Ziele, Ausgangspunkt), Vorgesehener Benutzerkreis, Bedürfnisse der Benutzer
3. Stand der Technik (Wie wird das Problem bisher gelöst, wo sind die Defizite)
4. Gewählter Lösungsansatz (allgemeines Prinzip, welche Werkzeuge, z.B. Programmiersprachen werden verwendet)
5. Beschreibung der durchgeführten Arbeiten
6. Ergebnis (z.B. Screenshots mit Erläuterungen)
7. Zusammenfassung (Erklärung des Nutzens), Ausblick

Anhang: evtl. (ausgewählte) Programmbeispiele

Evtl. CD-ROM als Beilage

# Anhang B: Formatvorlagen

Nachfolgend sind die für die Benutzer der Dokumentvorlage wichtigsten Formatvorlagen aufgelistet. Vermerkt ist jeweils auch der Zweck der Formatvorlage und ob es sich um eine neue oder eine modifizierte Formatvorlage handelt.

# Index

Abbildung 25

Abbildungsverzeichnis 18

Abkürzungsverzeichnis 18

Absatz 35, 44

Absatzmarken 39

Abstand zwischen Absätzen 35

Abstract 18

Angebotsorientiertheit 12

Anhänge 19

Anleitungen 13

Ansicht 37

Arbeitserleichterung 12

Aufzählungen 30

Auto-Wiederherstellen 38

Beschriftung 25

Bildschirmabzüge 25

Bindestrich 36

Computerprogramm 29

Datei-Info 39

Dokumentvorlage 13, 44

Ebenen 31

Eidesstattliche Versicherung 18

Einrückungen 30

Fehler 37

Fließtext 24

Formatvorlage 14, 22, 44

Formatvorlagen-Katalog 33

Funktionalitäten 22

Fußnoten 25

Glossar 20

Index 20

Inhalt der Arbeit 19

Inhaltsverzeichnis 18

Keywords 18

Kopfzeile 23

Kurzfassung 18

Leerzeichen 39

Literaturverzeichnis 33

Muster 14, 16

nicht druckbare Zeichen 20, 36

Nummerierungen 30

Qualitätssicherung 12

Quellenangabe 29

Rechtschreibprüfung 25

Rechtschreibung 16, 35, 39

Schlagwörter 18

Schnellspeicheroption 38

Schriftart 24

Seiteneinrichtung 23

Serifenschrift 24

Sicherungen 37

Sichtbarkeit 36

Silbentrennung 25, 36, 39

Speichern 37

Standard (Formatvorlage) 24

Stichwortverzeichnis 20

Tabellen 27

Tabellenüberschrift 28

Tabellenverzeichnis 18

Titelblatt 17

Trennstriche 36

Überschriften 31

Untertitel 17

Vorgaben 13

Vorwort 19

Word-Dokument 14

Word-Dokumentvorlage 14, 33

Zitat 28

References

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Cham: Springer International Publishing.

Algorithmia. *2020 state of enterprise machine learning*. Retrieved from https://info.algorithmia.com/hubfs/2019/Whitepapers/The-State-of-Enterprise-ML-2020/Algorithmia\_2020\_State\_of\_Enterprise\_ML.pdf

Alyari, F., & Jafari Navimipour, N. (2018). Recommender systems. *Kybernetes*, *47*(5), 985–1017. https://doi.org/10.1108/K-06-2017-0196

Baker, T. (2019). *Smarter Humans. Smarter Machines.* Retrieved from Refinitiv website: https://www.refinitiv.com/en/resources/special-report/refinitiv-2019-artificial-intelligence-machine-learning-global-study

Baylor, D., Breck, E., Cheng, H.‑T., Fiedel, N., Foo, C. Y., Haque, Z., . . . Zinkevich, M. (2017). TFX: A TensorFlow-Based Production-Scale Machine Learning Platform. In S. Matwin, S. Yu, & F. Farooq (Eds.), *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1387–1395). New York, NY, USA: ACM. https://doi.org/10.1145/3097983.3098021

Bennett, J., Lanning, S., & others (2007). The netflix prize. In *Proceedings of KDD cup and workshop.* Symposium conducted at the meeting of Citeseer.

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(0), 2. https://doi.org/10.5334/dsj-2015-002

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., . . . Dreyer, K. J. (2018). Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*, *288*(2), 318–328. https://doi.org/10.1148/radiol.2018171820

Chui, M., Hall, B., Singla, A., & Sukharevsky, A. (2021, December 8). *The state of AI in 2021*. Retrieved from McKinsey website: https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

Columbus, L. (2017, July 9). McKinsey's State Of Machine Learning And AI, 2017. *Forbes*. Retrieved from https://www.forbes.com/sites/louiscolumbus/2017/07/09/mckinseys-state-of-machine-learning-and-ai-2017/?sh=63414b1b75b6

Fleder, D. M., & Hosanagar, K. (2007). Recommender systems and their impact on sales diversity. In J. MacKie-Mason, D. Parkes, & P. Resnick (Eds.), *Proceedings of the 8th ACM conference on Electronic commerce - EC '07* (p. 192). New York, New York, USA: ACM Press. https://doi.org/10.1145/1250910.1250939

Google LLC. TFRS Releases. Retrieved from https://github.com/tensorflow/recommenders/releases?page=2

Google LLC. TFX Releases. Retrieved from https://github.com/tensorflow/tfx/releases?page=8

Google LLC (2021). Collaborative Filtering. Retrieved from https://developers.google.com/machine-learning/recommendation/collaborative/basics

Google LLC (2022). Using TensorFlow Recommenders with TFX. Retrieved from https://www.tensorflow.org/recommenders/examples/ranking\_tfx

Harper, F. M., & Konstan, J. A. (2016). The MovieLens Datasets. *ACM Transactions on Interactive Intelligent Systems*, *5*(4), 1–19. https://doi.org/10.1145/2827872

Hevner, March, Park, & Ram (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625

Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*.

Hevner, A., & Chatterjee, S. (Eds.) (2010). *Integrated Series in Information Systems*. *Design Research in Information Systems*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-5653-8

Jannach, D., & Zanker, M. (2022). Value and Impact of Recommender Systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 519–546). New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-2197-4\_14

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Khusro, S., Ali, Z., & Ullah, I. (2016). Recommender Systems: Issues, Challenges, and Research Opportunities. In K. J. Kim & N. Joukov (Eds.), *Lecture Notes in Electrical Engineering. Information Science and Applications (ICISA) 2016* (Vol. 376, pp. 1179–1189). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-0557-2\_112

Koren, Y. (2009). The bellkor solution to the netflix grand prize. *Netflix Prize Documentation*, *81*(2009), 1–10.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models: An introduction to factor, path, and structural equation analysis* (5. ed.). New York: Routledge.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, 1. https://doi.org/10.1109/TKDE.2018.2876857

Lü, L., Medo, M., Yeung, C. H., Zhang, Y.‑C., Zhang, Z.‑K., & Zhou, T. (2012). Recommender systems. *Physics Reports*, *519*(1), 1–49. https://doi.org/10.1016/j.physrep.2012.02.006

Makinen, S., Skogstrom, H., Laaksonen, E., & Mikkonen, T. (2021). Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? In *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)* (pp. 109–112). IEEE. https://doi.org/10.1109/WAIN52551.2021.00024

Miranda, L. J. (2021). Towards data-centric machine learning: a short review. *Ljvmiranda921. Github. Io*.

Naumov, M., Mudigere, D., Shi, H.‑J. M., Huang, J., Sundaraman, N., Park, J., . . . Smelyanskiy, M. (2019, May 31). *Deep Learning Recommendation Model for Personalization and Recommendation Systems*. Retrieved from http://arxiv.org/pdf/1906.00091v1

Refinitiv (2020). *THE RISE OF THE DATA SCIENTIST:: Machine learning models for the future*. Retrieved from https://www.refinitiv.com/en/resources/special-report/refinitiv-2020-artificial-intelligence-machine-learning-global-study

Renggli, C., Rimanic, L., Gürel, N. M., Karlaš, B., Wu, W., & Zhang, C. (2021, February 15). *A Data Quality-Driven View of MLOps*. Retrieved from http://arxiv.org/pdf/2102.07750v1

Rimol, M. (2021, November 22). *Gartner Forecasts Worldwide Artificial Intelligence Software Market to Reach $62 Billion in 2022*. Retrieved from Gartner website: https://www.gartner.com/en/newsroom/press-releases/2021-11-22-gartner-forecasts-worldwide-artificial-intelligence-software-market-to-reach-62-billion-in-2022

Singh, P. K., Choudhury, P., Dey, A. K., & Pramanik, P. K. D. (2021). Recommender systems: an overview, research trends, and future directions. *International Journal of Business and Systems Research*, *15*(1), 14. https://doi.org/10.1504/ijbsr.2021.10033303

Su, X., & Khoshgoftaar, T. M. (2009). A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, *2009*, 1–19. https://doi.org/10.1155/2009/421425

Wei, W. (2022, June 9). Question: examples/best practices for integrating tfrs with tfx #380. Retrieved from https://github.com/tensorflow/recommenders/issues/380#issuecomment-1151934730

1. Throughout this paper ML and AI will be used interchangeably. [↑](#footnote-ref-1)
2. https://github.com/MyPetOctocat/bachelor\_2022 [↑](#footnote-ref-2)
3. During the time of this research, the official documentation regarding TFRS implementation in TFX only uses three out of nine TFX components. (Google LLC (2022) [↑](#footnote-ref-3)