

Datenbewusstsein

Wo, wie und wozu werden Daten gesammelt und verarbeitet? – Exploration von Empfehlungsdiensten

Ergänzende Informationen für Lehrkräfte zu diesem Unterrichtsmodul

1 Motivation

Im Alltag interagieren wir tagtäglich mit datengetriebenen digitalen Artefakten. Ein datengetriebenes digitales Artefakt erhebt in der Regel während der Interaktion mit diesem – also der Nutzung dessen – persönliche Daten und verarbeitet diese für verschiedene Zwecke, wie etwa für das personalisierte Anzeigen von Produkten oder für standortbezogene Dienste. Die Daten werden entweder explizit erhoben, also durch Nutzer:innen aktiv bereitgestellt (etwa durch die Eingabe von Profilinginformationen auf einer Social Media Plattform) oder implizit bei der Interaktion erhoben, etwa durch Beobachtung der Nutzer:innen (z.B. die Klicks auf Ergebnisse in einer Suchmaschine oder das Verhalten beim Durchsehen eines News Feeds auf einer Social Media Plattform) oder durch Verarbeitung erhobener Daten generiert. Die Verarbeitung und Verwendung dieser explizit und implizit erhobenen Daten hat aus Sicht der Nutzer:innen meist den primären Zweck, ein Feature des datengetriebenen digitalen Artefakts zu nutzen, aus Sicht der Anbietenden bzw. der Entwickler des datengetriebenen digitalen Artefakts werden auch weitere sekundäre Zwecke verfolgt, um etwa darüber hinaus Untersuchungen durchführen oder das datengetriebene digitale Artefakt weiterzuentwickeln. Damit greifen wir bereits mehrere Facetten des Konzepts Datenbewusstsein auf.¹ Nähere Informationen zu dem fachdidaktischen Konzept Datenbewusstsein für den allgemeinbildenden Informatikunterricht finden Sie auf folgender Überblicksseite: <https://unterrichtsmaterial-ddi.cs.upb.de/Datenbewusstsein>

Ein Beispiel für ein datengetriebenes digitales Artefakt ist ein Empfehlungsdienst. Empfehlungsdienste (engl.: Recommender Systems) lassen sich in vielen verschiedenen Kontexten wiederfinden, wie z.B. Streamingdienste (Netflix, Amazon Prime Video, Spotify), Online-Shops, Google Suchergebnisse uvm. – näheres dazu wird im Abschnitt 2.2. Empfehlungsdienste werden in der Regel eingesetzt, um den Nutzer:innen eine Auswahl an Produkten (Items) vorzuschlagen bzw. zu empfehlen. Diese sollten potenziell für die jeweilige Nutzerin/den Nutzer interessant sein. Dies könnten etwa Filme bei einem Streamingdienst sein. Kaum einer ist sich dabei bewusst, welche verschiedenen Daten explizit und implizit erhoben werden, wie genau und wozu diese verarbeitet und verwendet werden (könnten). In diesem Unterrichtsmodul werden diese Fragen aufgegriffen, untersucht, aufgeklärt und im Alltagskontext der Lernenden reflektiert sowie bewertet.

In diesen ergänzenden Informationen werden verschiedene Hintergrundinformationen gegeben, die für Lehrkräfte beim Unterrichten dieses Moduls hilfreich sein können.

Inhaltsverzeichnis

1	Motivation	1
2	Hintergrundwissen zu Empfehlungsdiensten	2
2.1	Empfehlungsdienste im Allgemeinen (in engl.: Recommender System)	2
2.2	Empfehlungsdienste bei Streamingdiensten	2
2.3	Erhebung von Daten bei der Nutzung von Streamingdiensten	3
3	Empfehlungsdienst in diesem Unterrichtsmodul	4

¹ vgl. etwa: Höper, L. & Schulte, C., (2021). Datenbewusstsein: Aufmerksamkeit für die eigenen Daten. In: Humbert, L. (Hrsg.), INFOS 2021 – 19. GI-Fachtagung Informatik und Schule. Gesellschaft für Informatik, Bonn. (S. 73-82). DOI: 10.18420/infos2021_f235

3.1	Explizite und implizite Bewertungen	4
3.2	Genutzte Ratingdaten und Aufbereitung dieser für das Unterrichtsmodul	4
3.3	Jupyter Notebooks im Allgemeinen.....	5
3.4	Vorbereitetes Jupyter Notebook	5
3.5	k-Nearest-Neighbor Algorithmus zur Erstellung eines Modells:.....	5
3.6	User-Movie-Tabelle als hilfreiche Tabelle zur Ermittlung von Empfehlungen.....	6
Zusatzmaterialien und weitere Fakten.....		6
	Bild einer Startseite von Netflix	6
	Weiterführende Hinweise	7

2 Hintergrundwissen zu Empfehlungsdiensten

2.1 Empfehlungsdienste im Allgemeinen (in engl.: Recommender System)

Ein Empfehlungsdienst verfolgt das Ziel die Menge aller vorhandenen Items (z.B. Filme, Musiktitel, Shopping-Produkte, ...) auf eine Vorauswahl (Empfehlungen) einzuschränken, um den Nutzer:innen bei der Entscheidungsfindung zu unterstützen. Dem Nutzer/der Nutzerin sollten also nicht alle Items angezeigt werden, sondern nur eine Auswahl an Items, für die sich der Nutzer potenziell interessieren könnte, um eine Informationsüberflutung zu umgehen. Die Anbietenden des Dienstes zielt damit auf eine Gewinnmaximierung ab, indem der Nutzer/die Nutzerin „neue und interessante“ Items „entdeckt“. Dadurch werden die Nutzer:innen zu längeren und häufigeren Zugriffen (Steigerung der Nutzungszeit) angeregt, wodurch sie mehr Daten hinterlassen und womöglich der Umsatz durch Käufe oder Werbungen gesteigert werden kann.

Im Wesentlichen gibt es inhaltsbasierte (content-based), kollaborative (collaborative) und hybride Methoden zum Filtern der Items. Beim kollaborativen Filtern werden ähnliche Nutzer:innen identifiziert, um dann Empfehlungen basierend auf deren Daten (bspw. Filmbewertungen) zu ermitteln (hier etwa: Mittelwerte der Bewertungen der ähnlichen Nutzer:innen). Beim inhaltsbasierten Filtern werden Daten herangezogen, welche inhaltliche Informationen über die Produkte enthalten bzw. zumindest operationalisieren (z.B. Tags, Genres, Wortvorkommen in Textbeschreibungen). Das hybride Filtern verbindet verschiedene Methoden des kollaborativen und inhaltsbasierten Filterns – i.d.R. nacheinander.

2.2 Empfehlungsdienste bei Streamingdiensten

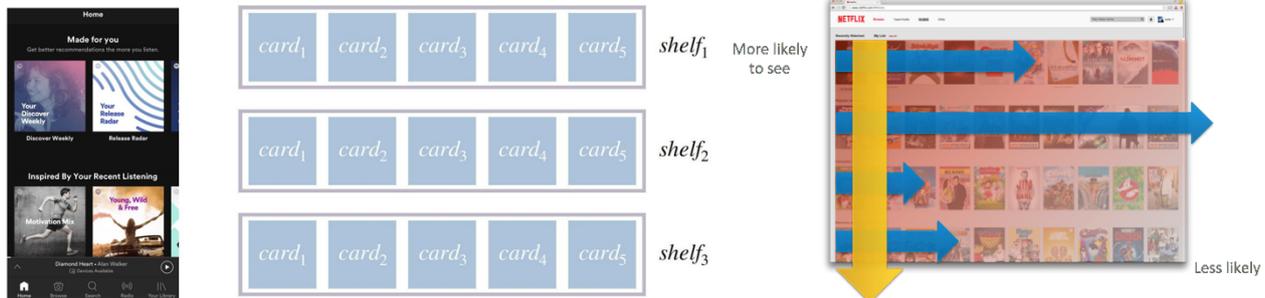
Zwei Zitate zu Empfehlungsdiensten bei Netflix und Spotify:

“A problem we face is that our catalog contains many more videos than can be displayed on a single page and each member comes with their own unique set of interests. Thus, a general algorithmic challenge becomes how to best tailor each member’s homepage to make it relevant, cover their interests and intents, and still allow for exploration of our catalog.” - Netflix TechBlog (<https://netflixtechblog.com/learning-a-personalized-homepage-aa8ec670359a>)

“Spotify has created engines to control and manage everything from your personal best home screen to carefully chosen and organized playlists like Discover Weekly, and continues to explore new ways to understand music, and why people listen to one song or genre over another. All this is achieved with a combination of different recommender systems.” (<https://www.linkedin.com/pulse/how-spotify-recommender-system-works-daniel-roy-cfa/>)

In diesem Unterrichtsmodul liegt ein Fokus auf Streamingdienste, wodurch vor allem Plattformen wie Netflix und Spotify im Mittelpunkt stehen. Trotzdem ist dieser Markt stetig am Wachsen: Amazon Prime Video, Disney+, Apple Music oder Amazon Music sind nur einige der Wettbewerber. Schon seit dem Aufkommen von modernen Streamingdiensten arbeiten diese an folgender Frage: Wie kann einem Kunden/einer Kundin möglichst maßgeschneiderte (personalisierte) Produktempfehlungen angeboten werden und wie können somit möglichst viele Kunden zum Bezahlen im jeweiligen Preismodell motiviert werden? Für die Berechnung solcher Empfehlungen spielen verschiedene Faktoren eine entscheidende Rolle: Der grundsätzliche Aufbau (Layout) der Plattform, die Möglichkeiten zur Interaktion für den Kunden/die Kundin und beispielsweise auch das soziale Umfeld, welches sich über verschiedene Benutzer:innen hinweg bildet.

Netflix steht mit seinen über 15000 Filmen und Serien² und über 200 Millionen zahlenden Abonnenten weltweit³ vor der Herausforderung, dass es seinen Nutzer:innen nur eine begrenzte Anzahl an Film- und Serientiteln vorschlagen und auf der Startseite anzeigen kann. Ein vergleichbares Bild ergibt sich bei Spotify: mit 70 Millionen Songs⁴ und 365 Millionen monatlich aktive Nutzer:innen⁵ findet sich dort zwar ein anderes Produkt wieder, der Aufbau Plattform-Seite und die Datenbasis sind jedoch ähnlich. Die folgenden Illustrationen verdeutlichen die Ähnlichkeit der beiden Plattformen. Auf der linken Seite befindet sich das Layout von Spotify's Oberfläche, das rechte Bild beschreibt selbigen Aufbau bei Netflix:



Bei diesen Startseiten werden sowohl Reihen als auch Spalten entsprechend an den Nutzer/die Nutzerin angepasst sortiert. Beide Übersichten verdeutlichen, wie wertvoll der Platz auf der Startseite ist und wie relevant damit der Einsatz eines Empfehlungsdienstes wird. Das Geschäftsmodell von Netflix als auch von Spotify ist abonnementbasiert. Dabei gibt es jeweils verschiedene Abo-Pläne, welche im Einzelnen jedoch nicht das Angebot erweitern, sondern z.B. die Anzahl der Nutzenden pro Account verändern. Das Ziel ist es also, die Anzahl der zahlenden Kunden zu maximieren und Abonnements immer wieder zu verlängern. Dies wird dadurch erreicht, möglichst passende Vorschläge für Filme/Musik zu realisieren. Neben Streamingdiensten mit vollem Zugriff auf das komplette Angebot nach Zahlung einer monatlichen Rate gibt es auch andere Geschäftsmodelle, wie etwa bei Amazon Prime Video. Diese Plattform stellt nach einem Abonnement ebenfalls einen Teil seines Film- und Serienangebots zur Verfügung (ca. 13000 Titel). Darüber hinaus gibt es jedoch Medien, welche nach wie vor durch die Zahlung eines einmaligen Betrages freigeschaltet werden müssen (ca. 25000 Titel). Dabei kann es sein, dass die ersten Staffeln einer Serie im Abonnementpreis enthalten sind, ab einer bestimmten Staffel jedoch eine zusätzliche Zahlung fällig wird.

2.3 Erhebung von Daten bei der Nutzung von Streamingdiensten

Das Nutzungsverhalten der Nutzer:innen von Streamdiensten ist essenziell, um das Geschäftsmodell und damit den Umsatz stetig zu optimieren. Dabei spielen grundsätzlich jegliche Arten von Interaktionen der Nutzer:innen eine Rolle. Dies fängt bei einfachen Feedbackmechanismen an, wie etwa der "Gefällt mir"-Button, und geht weiter zum Nutzungsverhalten über angeschauten Filme und angehörte Musik. Dies ist nur der Anfang von dem, wie Streamingdienste Daten erheben und verarbeiten. Ein Ziel ist es, einen digitalen Doppelgänger eines jeden Nutzenden zu konstruieren. Auf Basis dessen können dann Empfehlungen und weitere Verarbeitungen und Verwendungen dessen erfolgen.

Ein detailliertes Beispiel der Datenverwendung ist die sogenannte "Completion Rate" (deutsch: Abschlussrate)⁶ einer Serie. Dabei kann folgende Frage beantwortet werden: "Wie viele Nutzer:innen, die Serie XY angefangen haben zu schauen, schauten diese bis zu Staffel 3?". Fällt die Antwort auf diese Frage z.B. auf 30% aus, öffnet sich ein weiterer Pool an Fragen: "An welchem Punkt haben die meisten Nutzer:innen die Serie abgebrochen?" und "Wie lange und mit wie großen Pausen zwischen den Folgen haben die 30% die Serie bis Staffel 3 angesehen?". Wenn nun ein gewisser Anteil eine Serie zu Ende geschaut hat, liegt es nahe, dass der Serienproduzent (wie inzwischen etwa Netflix) eine weitere Staffel produzieren wird – und im anderen Fall eventuell eine neue Serie produziert, die das Feedback der Nutzer:innen aufgreift, die die Serie nicht zu Ende gesehen haben.

² <https://www.comparitech.com/blog/vpn-privacy/netflix-statistics-facts-figures/>

³ <https://de.statista.com/statistik/daten/studie/196642/umfrage/abonnenten-von-netflix-quartalszahlen/>

⁴ <https://de.statista.com/statistik/daten/studie/378806/umfrage/anzahl-der-verfuegbaren-songs-auf-spotify/>

⁵ <https://de.statista.com/infografik/13769/monatlich-aktive-nutzer-und-zahlende-abonnenten-von-spotify-weltweit/>

⁶ <https://neilpatel.com/blog/how-netflix-uses-analytics/>

Generell gilt: Bei allen Streamingdiensten werden zunächst alle verfügbaren (persönliche) Daten explizit und implizit erhoben und zu einem gegebenen Zeitpunkt verarbeitet, um das Geschäftsmodell weiter zu optimieren. Darüber hinaus können potenziell auch einzelne Daten an Dritte weitergegeben werden (z.B. Facebook oder andere Werbepartner).⁷ Es ist jedoch wichtig anzumerken, dass gerade bei Netflix die monetäre Einnahmequelle auf den Abonnements und nicht auf der Weitergabe von Daten basiert. Diese Entscheidung wurde getroffen, um einen möglich Abgang von Nutzer:innen zu verhindern.⁸

Empfehlungsdienste finden sich heute in fast allen bekannten Online-Plattformen wieder. So nutzt Google solche Systeme beispielsweise bei der Google Suche zum Anzeigen der Suchergebnisse oder auf der Plattform YouTube zum Erzeugen einer personalisierten Startseite bzw. geben personalisierter Videoempfehlungen. Um jeweils zum Nutzer/zur Nutzerin passende Produkte vorzuschlagen, nutzt auch Amazon solchen Modellen. Auch die Bibliothek des KIT in Karlsruhe arbeitet mit ähnlichen Verfahren auf Basis dessen Literaturempfehlungen ausgegeben werden können.

Dabei spielen im Grunde immer ähnliche Daten eine Rolle. Auf der einen Seite die verschiedenen Produkte, welche auf der Plattform angeboten werden (Filme, Videos, Bücher, ...). Auf der anderen Seite stehen die Nutzer:innen und ermöglichen die Erhebung und Generierung wertvoller Daten durch die Interaktion mit den jeweiligen digitalen Artefakten (Schreiben von Rezensionen, Ansicht von Produkten, Verbindungen zu anderen Nutzer:innen, ...).

3 Empfehlungsdienst in diesem Unterrichtsmodul

3.1 Explizite und implizite Bewertungen

Bewertungen für Produkte, wie bspw. Filme, können explizit oder implizit vorliegen. Explizit sind Bewertungen dann, wenn der Nutzer/die Nutzerin das Produkt direkt beurteilt, bspw. über ein Gefällt-mir-Button oder eine Sternebewertung. Dadurch gibt der Nutzer/die Nutzerin i.d.R. seine Meinung von dem Produkt bzw. sein Interesse an dem Produkt aktiv zum Ausdruck. Implizite Bewertungen werden nicht von dem Nutzer/der Nutzerin direkt angegeben. Das bedeutet, dass bestimmte Daten erhoben, generiert und verarbeitet werden, welche bspw. als Operationalisierung für das Interesse an dem Produkt dienen können. Beispiele für implizite Bewertungen sind: Hat der Nutzer/die Nutzerin das Produkt gekauft? Hat er oder sie den Film vollständig geschaut oder früher beendet? Wurde der Film mehrmals geschaut? Wurde sich das Produkt gemerkt (Merklisten)?

In dem Datensatz, welcher im Unterricht eingesetzt wird (Beschreiben siehe Abschnitt 3.2), wurden neben den expliziten Bewertungen zusätzlich implizite Beurteilungen generiert, um die beiden Konzepte zu veranschaulichen. Dabei sind implizite Filmbewertungen im FilmDATENSATZ die binäre Antwort auf die Frage, ob Nutzer:innen einen Film zu Ende geschaut hat oder nicht. Dieses neue Attribut wurde künstlich, jedoch auf Basis der vorhandenen Bewertungen angelegt. Mit einer Wahrscheinlichkeit von 85% wurden Bewertungen mit mehr als vier Sternen auf den Status "1" (Film zu Ende angeschaut) gesetzt. Liegt die Bewertung unter vier Sterne fand dies nur in 40% der Fälle statt. Alle übrigen Bewertungen erhielten den Status "0" (Film nicht zu Ende angeschaut). Im Anschluss wurden 25% der expliziten Bewertungen entfernt, um die Relevanz der impliziten Bewertungen darzustellen.

3.2 Genutzte Ratingdaten und Aufbereitung dieser für das Unterrichtsmodul

In diesem Unterrichtsmodul ziehen wir reale Nutzungsdaten von Nutzer:innen der Plattform MovieLens (movielens.org) heran. Auf der Plattform angemeldete Nutzer:innen können dort u.a. Filme bewerten und Filmempfehlungen bekommen. Es ist also ein Empfehlungsdienst eingebettet. Die Betreiber haben Bewertungsdaten öffentlich zugänglich gemacht⁹. Für das Unterrichtsmodul haben wir diese Daten aus Performancegründen verkleinert, sodass wir lediglich ca. 50000 Bewertungen von ca. 5000 Usern zu insgesamt ca. 600 Filmen nutzen.

Die Filme, die bewertet werden können, wurden manuell nach einer subjektiven Einschätzung des Bekanntheitsgrades sowie unter Einbezug von IMDB-Hitlisten ausgewählt. Filme, welche unter den möglichen Empfehlungen erscheinen, haben eines Mindestanzahl an Bewertungen erhalten. Die Nutzer:innen in dem Datensatz wurden so ausgewählt, dass

⁷z.B. <https://www.spotify.com/de/legal/privacy-policy/>

⁸https://geschaeftsmodell-workshop.de/geschaeftsmodell/beispiele/netflix-business-model#Das_Business_Model_Canvas_von_Netflix_als_Powerpoint_und_PDF

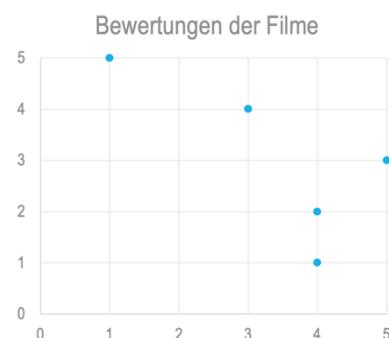
⁹ Referenz zum Projekt: <https://dl.acm.org/doi/10.1145/2827872>; Daten: <https://grouplens.org/datasets/movielens/>

Der k-nearest-neighbor Algorithmus wird in dem vorbereiteten Jupyter Notebook mithilfe der Bibliothek sklearn zur Erstellung eines Modells verwendet. Dieses Modell kann anschließend zum Entscheiden von ähnlichen Nutzer:innen (eigentlich: nächsten Nachbarn) auf Basis von Daten aus z.B. einem Streamingdienst angewendet werden. Die konkrete Funktionsweise des Algorithmus wird im Unterricht nicht im Detail vermittelt, es soll lediglich die Idee der Vorgehensweise verstanden werden. An der Stelle der Modellerzeugung wird bewusst eine Black-Box gesetzt, um die im Rahmen dieser Unterrichtsreihe gesetzten Lernziele zu erreichen und keine Überforderung zu erzeugen.

Beispiel:

In der nebenstehenden Tabelle ist ein Minimalbeispiel gegeben. Es gibt Bewertungsdaten von fünf Nutzer:innen zu zwei Filmen. Anhand dieses Beispiels kann bereits die Suche nach den k nächsten Nachbarn erklärt werden. Gesucht sind zum Beispiel zwei Nutzer:innen, die ähnlich zum markierten User 5 sind. Das sind dann etwa die User 1 und 4, da diese die kleinste Abweichung in ihren Bewertungen der beiden Filme zu User 5 haben. Konkret heißt das, dass die Abstände zwischen der Tabellenzeile von User 5 und denen von User 1 und 4 am kleinsten sind, die Differenz also möglichst klein ist. (Randnotiz: Mathematisch nutzen wir in unserer Umsetzung die euklidische Metrik für die Bestimmung von Abständen.) Dies kann auch in dem nebenstehenden Koordinatensystem visualisiert werden. Die Bewertungen zu Film A entsprechen den Werten auf der x-Achse und zu Film B denen auf der y-Achse. So stellt jeder Punkt im Koordinatensystem einen User da, der beide Filme bewertet hat. Mit dieser Vorgehensweise können zu einem

	Film A	Film B
User 1	5	3
User 2	3	4
User 3	1	5
User 4	4	2
User 5	4	1



	Film A	Film B	Film C
User 1	5	3	4
User 2	3	4	
User 3	1	5	4
User 4	4	2	5
User 5	4	1	

gewählten User die k ähnlichsten Nutzer:innen einfach identifiziert werden. So können im Koordinatensystem etwa beliebig viele Nutzer:innen hinzugefügt werden. Um nun anhand dieser ähnlicher Nutzer:innen zu User 5 für einen dritten Film C herauszufinden, ob dieser empfohlen werden sollte, wird eine Prediction ermittelt. Die Prediction wird etwa durch den Mittelwert der Bewertungen des Films C der ähnlichen Nutzer:innen ermittelt. In dem Beispiel der nebenstehenden Tabelle ist dies dann 4,5 (Mittelwert von 4 und 5). Das heißt, wenn User 5 den Film C schauen und bewerten würde, würde er wahrscheinlich eine Bewertung von 4,5 abgeben. Dem User 5 sollte der Film C also durchaus empfohlen werden. Dieses (hier stark reduzierte) Verfahren ist auf eine große Anzahl von Nutzer:innen und Filmen übertragbar.

3.6 User-Movie-Tabelle als hilfreiche Tabelle zur Ermittlung von Empfehlungen

Die user-movie-Tabelle ist eine Datentabelle, die in diesem Kontext Filmbewertungen (Zellen) von Nutzer:innen (Zeilenweise userIds) zu den jeweiligen Filmen (Spaltenweise Filmtitel) aufführt. Diese Tabelle ist für den Empfehlungsdienst recht zentral, anhand dieser wird beispielsweise das vom k-Nearest-Neighbor Algorithmus ermittelte Modell mit einer aus der Tabelle erstellten sparse-Matrix berechnet. Ähnliche User werden also mithilfe der Abstände zwischen den jeweiligen Zeilen in dieser Tabelle bestimmt.

Zusatzmaterialien und weitere Fakten

Bild einer Startseite von Netflix

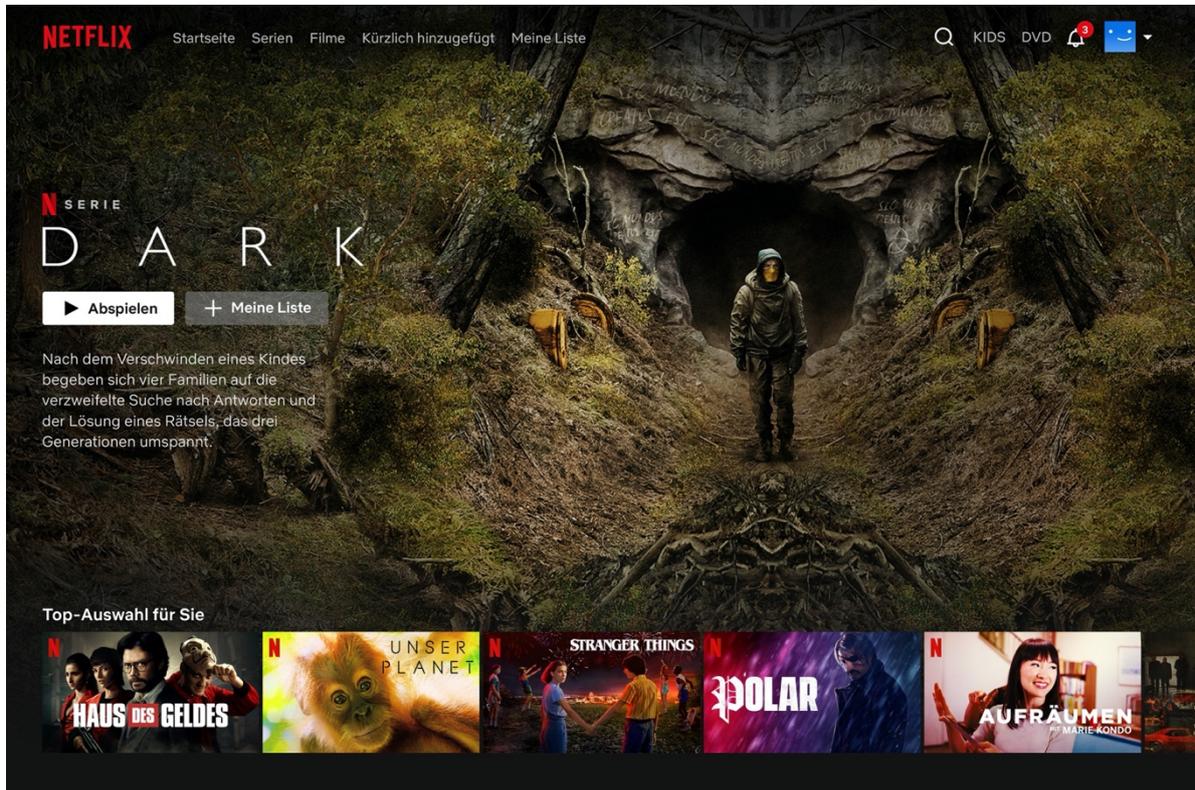


Abbildung 1: Bild von <https://about.netflix.com/de/company-assets>

Die Abbildung der Startseite wird als Motivation und Einstieg für die personalisierte Startseite verwendet. Weitere Informationen hierzu finden sich im Verlaufsplan.

Weiterführende Hinweise

Vorstellung des Konzepts Datenbewusstsein: Höper, L. & Schulte, C., (2021). Datenbewusstsein: Aufmerksamkeit für die eigenen Daten. In: Humbert, L. (Hrsg.), INFOS 2021 – 19. GI-Fachtagung Informatik und Schule. Gesellschaft für Informatik, Bonn. (S. 73-82). DOI: 10.18420/infos2021_f235

Fortbildung zum Unterrichtsmoduls durch: www.prodabi.de

Kontakt Daten bei Fragen, Anregungen und Interesse: Lukas Höper der Didaktik der Informatik an der Universität Paderborn (lukas.hoeper@upb.de).