Die Versprechungen datengetriebener Prognostik

Arne Brusis

Seminararbeit im Interdisziplinären Lehrangebot
des Instituts für Informatik

Leitung: Prof. Hans-Gert Gräbe, Ken Pierre Kleemann

`http://bis.informatik.uni-leipzig.de/de/Lehre/Graebe/Inter`

Leipzig, 09.04.2018

Inhaltsverzeichnis

1 Einleitung

Sobald sich ein Nutzer im digitalen Raum bewegt, das heißt die Informations- und Kommunikationsnetze des Internet gebraucht, um Bücher zu lesen, Musik zu hören Fotos und Videos zu schauen oder einen Telefonanruf zu tätigen, werden Daten über ihn gespeichert und verarbeitet. Außerdem sind viele Bestandteile des Alltags Teil des Internets der Dinge, in dem sich die Welt selbst beobachtet und über unzählige Sensoren die Welt vermessen.

Die Informationskonzerne versuchen aus dieser exponentiell wachsenden Datenmenge Muster und Korrelationen zu erkennen, um daraufhin Prognosen für die Zukunft zu erstellen: Wann bricht die Grippewelle dieses Jahr aus? Welches Marktprodukt wird sich am besten verkaufen? Wo wird da nächste Verbrechen stattfinden?

Können die Informationskonzerne die selbst ins fantastische getriebenen Erwartungen erfüllen, oder sind es leere Versprechungen?

Um diese Versprechungen der datengetriebenen Prognostik verstehen zu können, werden zuerst die Begriffe, Big Data, Data Mining und schließlich Predictive Analytics erläutert. Sie sind als einzelne, aufeinander aufbauende Schritte zu verstehen, die für den Modellbildungsprozess der Prognose elementar sind.

Im weiteren Teil der Seminararbeit wird die datengetriebene Prognostik anhand von zwei ausführlichen Beispielen erläutert, die zu Letzt in Hinblick auf gesellschaftlichen Auswirkungen diskutiert werden.

2 Datengetriebene Prognostik

Viele von uns tragen ein Mobiltelefon in der Hosentasche, das permanent Daten über uns erhebt, diese weiterleitet und auf Servern großer privatwirtschaftlicher Firmen, wie Facebook und Google speichert. Diese beiden sind aber nur die bekanntesten von vielen tausenden anderen Unternehmen, die ihren Umsatz damit machen Informationen zu generieren, miteinander zu verknüpfen und weiterzuverkaufen. Einige dieser Unternehmen gehen noch einen Schritt weiter. Sie behaupten mehr über uns zu wissen als wir selbst, weil sie in unsere Zukunft sehen könnten. Sie erstellen datengetriebene Prognosen über zukünftige Ereignisse. Um die Methoden und Prozesse dahinter verstehen zu können, müssen wir die Gestalt der großen Datenmengen verstehen, die Analysemethoden und Algorithmen kennen, sowie die Modellierungsprozesse, die zum Erstellen einer datengetriebenen Prognose nötig sind, nachvollziehen.

2.1 Big Data

Der Begriff Big Data bezieht sich auf den Umfang der Daten, die in der heutigen digitalen Welt generiert und verarbeitet werden. Das Datensammeln geschieht sowohl in unserem beruflichen als auch im privaten Umfeld: Unternehmensdaten, Finanztransaktionsdaten, Kundendaten, Social Media Daten bis hin zu Daten, die unseren Gesundheitszustand beschreiben.

Sucht man nach einer Definition für den Begriff Daten, so stößt man schnell auf folgende Vorstellungen: Mit Daten bezeichnet man „eine Ansammlung von Zeichen mit der dazugehörigen Syntax“ (Duhigg, 2012, S. 37). Die Begriffe Zeichen und Daten werden dann als Fundament einer Wissenspyramide untergeordnet, an deren Spitze die Begriffe Information und Wissen stehen.

Dieses Modell hilft wenig dabei das Internet der Dinge (internet of things), welches für das große Volumen an Daten verantwortlich ist, zu verstehen. Diese „Dinge“ müssen eine digitale Identität haben, damit wir über sie sprechen und kommunizieren können (Gräbe, 2017, S. 11). So können Daten als formalisierte Informationen betrachtet werden und Informationen sind interpretierte Daten. Beides, Formalisieren und Interpretieren, ist nur in einem speziellen Kontext, dem technischen und sozialen Umfeld, möglich. Es ist also eine „funktionierende Fiktion“, ein gesellschaftlicher Konsens, der sich durch eine Diskussion als gesellschaftliche Normalität herausgebildet hat, nötig (Gräbe, 2017, S. 2).

Um das Phänomen des Big Data beschreiben zu können, müssen drei Dimensionen betrachtet werden (Long, 2015, S. 2):

Zum einen der große Umfang an Daten –*volume*-, die in Datenbanken gespeichert werden. Dabei tritt die Herausforderung auf, dass heute Datenmengen verarbeitet werden müssen, die vor zehn Jahren noch nicht zu bewältigen erschienen.

Weiterhin die Vielfalt und Strukturiertheit der Daten -*variety*-, die sich je nach Quelle sehr stark unterscheiden können. So sind die Messdaten einer privaten Wetterstation, die ihre im Internet zur Verfügung gestellt werden und aus der Temperatur, der Zeit und den Koordinaten bestehen, maximal strukturiert. Der Inhalt einer Website ist dagegen nur semistrukturiert (Long, 2015, S. 6). Die Textdatei des Sourcecodes enthält Bestandteile einer Auszeichnungssprache wie XML und beschreibt sich dadurch selbst. Bilder und Videos hingegen sind letztlich unstrukturiert in dem Sinne, dass sie zunächst bearbeitet werden müssen, damit sie maschinell zu lesen sind.

Die dritte Dimension des Big Data bezieht sich auf die Wachstumsrate und die Geschwindigkeit -*velocity*-, mit der neue Daten erzeugt werden. So verdoppelt sich nach aktuellen Schätzungen die Menge an Daten, die weltweit gespeichert werden, alle zwei Jahre (Mainzer, 2014, S. 232).

Privatwirtschaftliche Firmen oder staatliche Stellen speichern die Daten, die sie Sammeln in einem Data Warehouse. Dies ist eine zentrale Datenbank, die für Analysen optimiert ist. Die Schwierigkeit dabei ist, dass die Daten zu einem sehr unterschiedlichen Grad strukturiert sind. So müssen die Eingangsdaten zuerst strukturiert, normalisiert und in einem passenden Datentyp definiert werden (Long, 2015, S. 14).

2.2 Data Mining

Die erste Frage die sich natürlich sofort ergibt ist: Was wird mit den ganzen Daten, die täglich gesammelt werden, gemacht? Bei der Auswertung von solch umfangreichen Daten reicht ein „Draufschauen“ nicht mehr aus, um einen Erkenntnisgewinn zu erzielen.

Data Mining ist also das Finden vorher unbekannten Mustern in einer Datenbasis, wobei dazu automatische iterative Methoden nötig sind (Kotu & Deshpande, 2015, S. 4). Dabei sind nicht das Erstellen einer deskriptiven Statistik, das Testen einer Hypothese mit experimentellen Daten oder einfache Anfragen an die Datenbank gemeint.

Die Aufgabenstellungen, die sich beim Data Mining Prozess ergeben, können in verschiedene Gruppen unterteilt werden.

Die Klassifikation versucht einen Datenpunkt einer vordefinierten Klasse zuzuordnen, wobei die Ausgabevariablen dabei kategorial oder binär, also eine ja oder nein Entscheidung, sind (Kotu & Deshpande, 2015, S. 11). Beispielsweise versuchen politische Parteien ihre potentiellen Wähler in Kategorien einzuteilen, um sie so besser adressieren zu können. Die Algorithmen, die hinter dieser Methode stehen, sind unter anderem Entscheidungsbäume, gerichtete Bäume, die der Darstellung von Entscheidungsregeln dienen, oder neuronale Netzwerke, die dem menschliche Gehirn nachempfunden sind und während des Lernens synaptische Verbindungen bilden (Mainzer, 2014, S. 152).

Die Regression hat zum Ziel, einem Datenpunkt einen nummerischen Wert zuzuordnen. Die Ausgabevariable ist dem entsprechend oftmals ein Wahrscheinlichkeitswert für eine bestimmte Eingabevariable. So versuchen Staaten die Arbeitslosenzahlen oder das Wirtschaftswachstum des nächsten Jahres im Voraus zu bestimmen, damit die mit großem Vorlauf politische Maßnahmen ergreifen können. Die lineare Regression ist der wohl am leichtesten zugänglichen Algorithmus zur Regression. Darüber hinaus gibt es noch komplexere Regressionsverfahren, die auch mehrere unabhängige Eingabevariablen bearbeiten können (Kotu & Deshpande, 2015, S. 11).

Die Clusteranalyse ist ein Prozess, der natürliche Gruppierungen innerhalb einer Datenbasis finden soll. Eine Firma kann aus Kundendaten, wie Transaktionsdaten, Webdaten und Kundenanrufen den eigenen Markt segmentieren, das heißt die Kunden anhand unterschiedlicher Gesichtspunkte, wie Alter, Einkommen oder Geschlecht, unterteilen. Danach können sie durch geeignete Marketinginstrumente angesprochen werden. Das wichtigste Verfahren zum Clustering ist der K-Means-Algorithmus. Dabei werden die Orte einer bestimmten Anzahl von Clusterzentren per Zufall bestimmt. Danach wird durch Iteration der Abstand der Datenpunkte zu den Clusterzentren minimiert, indem

diese Verschoben werden (MacGregor, 2014, S. 275). Da die Startpunkte der Analyse schon durch das Verfahren zufällig gewählt sind, muss nach der Anwendung noch untersucht werden, warum die Daten überhaupt Cluster bilden und wie diese zu Generalisieren sind (Kotu & Deshpande, 2015, S. 9).

Die Assoziationsanalyse findet Zusammenhänge und Abhängigkeiten in der Datenbasis, die einer Form von Regeln wie „Aus A folgt B“ folgen. Sie kann dafür genutzt werden, Nutzern eines Onlineshops nach einem getätigten Kauf weitere Artikel zu empfehlen. Typisch für diesen Fall ist eine Regel wie: „Kunden, die Artikel A gekauft haben, kaufen auch Artikel B“. Um solche Assoziationsregeln zu finden, kann der Apriori-Algorithmus genutzt werden. Er findet dafür in einer Datenbasis Itemmengen, die besonders häufig vorkommen, und erzeugt im zweiten Schritt Assoziationsregeln (MacGregor, 2014, S. 251).

Die Ausreißer-Erkennung identifiziert die Datenpunkte einer Datenbasis, die signifikant von den anderen Datenpunkten abweichen. So kann ein Kreditkartenanbieter betrügerische Transaktionen aufdecken, indem er beispielsweise eine Transaktion erkennt, deren Geldbetrag bedeutend von den anderen, über die selbe Kreditkarte abgewickelten, Transaktionen abweicht. Dafür sind Ausreißertests nötig, die Datenpunkte aussortieren, die sehr stark vom Erwartungswert abweichen und so aus dem Streubereich herausfallen.

2.3 Predictive Analytics

Die datengetriebene Prognostik ist ein Bereich des Data Mining, bei dem auf Basis von Datenmodellen eine Voraussage getroffen wird, wie sich eine Situation in Zukunft entwickeln könnte, indem die Wahrscheinlichkeit für ein zukünftiges Ereignis bestimmt wird (MacGregor, 2014, S. 25). Im englischsprachigen Raum hat sich der Begriff „Predictive Analytics“ etabliert, wobei teilweise auch von „Predictive Analysis“ gesprochen wird.

Um am Schluss bei der Eingabe von mehreren Variablen als Ausgabe einen Wahrscheinlichkeitswert, und somit eine Prognose über die Zukunft zu erhalten, ist ein aufwändiger Modellbildungsprozess nötig. Es kann grob zwischen Modellen unterschieden werden, die entweder überwacht oder nicht überwacht mit Hilfe von Data Mining Methoden lernen. Mit nicht überwachten bzw. ungesteuerten Modellen wird versucht, Muster zu finden, die bisher unbekannt waren. Dabei gibt es natürlich keine Ausgabevariablen die vorhergesagt werden könnten (Kotu & Deshpande, 2015, S. 8). Diese Modelle können nur ein erster Schritt sein, um später ein überwachtes Modell zu erstellen, das Vorhersagen treffen kann.

Ein überwachtes Modell berechnet die Wahrscheinlichkeit einer Ausgabevariable auf Basis einer Menge von Eingabevariablen. Dafür ist eine Datenbasis nötig, in der die Werte sowohl der Eingabevariablen, als auch der Ausgabevariable bekannt sind. Diese Trainingsdaten wurden vorher in einem Datawarehouse gesammelt und können bei Bedarf mit nützlichen Datenquellen ergänzt werden. Durch die Analysemethoden, die im vorherigen Abschnitt vorgestellt wurden, kann das Modell erstellt werden. Dabei werden Beziehungen zwischen Eingabe- und Ausgabevariablen generalisiert, um daraufhin Prognosen zu erstellen (Kotu & Deshpande, 2015, S. 8).

Dieser Modellbildungsprozess ist in eine breiter angelegte Schrittfolge eingebettet, die durchlaufen wird, damit ein solches Prognosemodell erfolgreich entsteht. Der erste Schritt ist, dass das angestrebte Projekt definiert wird, indem eine Zielstellung erarbeitet wird und die Datensätze bestimmt werden, die dafür benötigt werden. Danach werden die entsprechenden Daten zusammengetragen, passend strukturiert und gegebenenfalls durch externe Datenquellen ergänzt. Als dritte Schritt folgt die Datenanalyse, wobei mit unüberwachten Methoden, wie oben schon erläutert, Muster in den Daten gefunden werden sollen. Diese Erkenntnisse und Hypothesen werden danach durch eine statistische Analyse auf ihre Validität überprüft. Der fünfte Schritt ist dann der beschriebene Modellbildungsprozess, bei dem mit Hilfe von Trainingsdaten und Analysemethoden ein Prognosemodell entwickelt wird. Dieses Modell kann dann mit neuen, unbekannten

Eingabevariablen versorgt werden und gibt Prognosen für zukünftige Ereignisse in Form von Wahrscheinlichkeiten aus. Während der Nutzungsdauer des Prognosemodells können eingetretene Ereignisse im Nachhinein ausgewertet werden und damit das Modell angepasst und korrigiert werden (Long, 2015, S. 28).

3 Beispiele

Prognosen, die auf Basis einer gesammelten Datenmenge erstellt werden, kommen in finden in vielen Bereichen der Industrie eine Anwendung (MacGregor, 2014, S. 35). In der Versandbranche wird die Rücksendewahrscheinlichkeit eines Artikels bestimmt, damit der Lagerplatz optimal genutzt werden kann und nicht zu viele Artikel auf Vorrat gehalten werden müssen.

Im Bankenwesen wird die Ausfallwahrscheinlichkeit eines Kundenkredits berechnet, um das Risiko des Kreditinstituts über die große Anzahl ihrer Kunden hinweg abschätzen und kontrollieren zu können.

In der Versicherungsbranche können durch Analysemodelle auffällige Datenpunkte isoliert werden und so ein Verdacht auf Versicherungsbetrug geäußert werden.

Produktionsstätten kontrollieren permanent während des Produktionsprozesses die Wahrscheinlichkeit dafür, dass eine Maschine ihre Fertigungsgenauigkeit verlieren und ermitteln den optimalen Zeitpunkt eine Produktionsanlage neu zu kalibrieren.

Die öffentlichen Sicherheitsapparate treffen Vorhersagen darüber, an welchem Ort ein Verbrechen besonders wahrscheinlich ist und koordinieren dementsprechend ihre Polizeibeamten.

Außerdem wird in der Medizin mit Hilfe von Daten über Vorerkrankungen und Erkrankungen in der Familie eine Prognose zum Gesundheitszustand des Patienten erstellt, um so Vorsorgeuntersuchungen optimal zu planen.

In den beiden folgenden Abschnitten werden zwei Beispiele ausführlich erläutern und drauf folgend im vierten Kapitel der Arbeit diskutiert.

3.1 Einzelhändler Target

Die meisten Menschen kaufen nicht alle Produkte, die sie brauchen bei nur einem Einzelhändler. Für Hygieneartikel suchen sie einen Drogeriemarkt auf, für Spielsachen ein Spielwarengeschäft und nur wenn sie bestimmte Artikel, wie Reinigungsmittel, neue Socken oder einen großen Vorrat an Toilettenpapier, die sie mit der Einzelhandelskette Target verbinden, kaufen möchten, besuchen sie eine der vielen Filialen (Duhigg, 2012). Dabei hat Target eigentlich den Vorteil, dass man fast alles dort kaufen kann. Trotzdem können sie ihre Kunden, mit klassischen Werbekampagnen, nicht stärker an sich binden, weil das Kaufverhalten den Kunden so tief verwurzelt ist, dass es nur schwer zu beeinflussen ist.

Allein drei Schlüsselmomente im Leben eines Menschen führen dazu, dass sie ihre Gewohnheiten ändern und in diesen Lebensphasen stärker zu beeinflussen sind (Long, 2015,

S. 22): Einerseits kaufen Kunden viele neue Produkte, wenn sie heiraten wollen, oder gerade geheiratet haben. Des Weiteren führt eine Scheidung dazu, dass sie andere Produkte als Sonst kaufen und ihr Kaufverhalten ändern. Der für den Einzelhändler aber lukrativster Moment, so haben es die Statistiker und Datenanalysten von Target herausgefunden, ist die Schwangerschaft in einer Familie. Sobald diese bekannt ist, kauft das Paar Produkte, dass es bisher noch nie benötigt hat, wie spezielle Medikamente oder Möbel, und empfindet dabei eine gewisse zeitliche Dringlichkeit. Ist das Kind erst einmal geboren, haben die jungen Eltern nicht mehr so viel Zeit wie vorher, ihre Einkäufe zu erledigen und möchten am liebsten alle Produkte in einem Laden erwerben. Für einen Einzelhändler ist es also besonders attraktiv, Menschen in dieser Lebensphase an sich zu binden.

Da in Amerika Geburtseinträge in der Regel öffentlich einsehbar sind, werden die jungen Eltern nach der Geburt ihres Kindes mit Werbung förmlich überschwemmt (Duhigg, 2012). Target hat es sich deshalb als Ziel gesetzt, schon vorher zu wissen, ob seine Kundinnen ein Kind erwarten.

Dazu kann die Firma auf Daten zurückgreifen, die sie über Jahre hinweg von ihren Kunden gesammelt hat. Jeder Kunde, der in einer der viele Filialen einen Kauf tätigt, bekommt eine „Guest ID“ Nummer zugewiesen, mit der er eindeutig zu identifizieren ist. Diese Nummer wird mit Kreditkartendaten, die bei der Bezahlung angefallen sind, und Informationen, die der Kunde beim Besuch der Website hinterlassen hat, verknüpft.

Darüber hinaus ist es in den Vereinigten Staaten möglich, demografische Daten, wie Alter und Wohnort, von privaten Anbietern zuzukaufen (Duhigg, 2012). So ist es für Target möglich, eine umfangreiche Kundendatenbank aufzubauen, deren Informationen weit darüber hinaus gehen, welcher Kunde, welchen Artikel, wann gekauft hat.

Diese Datenbank enthält Kundendaten, die mehr als zehn Jahre zurückreichen. In dieser Zeit sind Kundinnen schwanger geworden und haben ihr Kinde geboren. Alle Produkte, die sie in dieser Zeit bei Target erworben haben, sind bekannt. Target konnte also mit Data Mining Analyseverfahren Indikatoren bestimmen, die Einfluss darauf haben, ob eine Frau schwanger ist oder nicht. Das sind einerseits demografische Daten, wie Alter, Geschlecht und Einkommen, aber auch das Einkaufsverhalten in der frühen Phase der Schwangerschaft. Mit diesen Trainingsdaten konnte ein Modell gebildet werden, das die Wahrscheinlichkeit dafür angibt, dass eine Kundin schwanger geworden ist.

Das entwickelte Vorhersagemodell wird dafür genutzt, die Schwangerschaft einer Kundin möglichst früh zu bemerken und daraufhin individualisierten Rabattgutscheine zu ihr nach Hause zu schicken (Duhigg, 2012). So soll sie Produkte wie Windeln und

Nahrungsergänzungsmittel bei Target kaufen und schlussendlich dauerhaft an den Einzelhändler gebunden werden.

3.2 Cambridge Analytica

Psychometrie, auch Psychografie genannt, ist „der wissenschaftliche Versuch die Persönlichkeit eines Menschen zu messen (Grassegger & Krogerus, 2018). Dafür ist in der heutigen Psychologie die Ocean-Methode zum Standard geworden. Sie bestimmt jeden Charakterzug eines Menschen anhand von fünf Persönlichkeitsmerkmalen, den Big-Five (Kindler, 2017):

Die Offenheit für neue Erfahrungen (Bin ich aufgeschlossen gegenüber neuem?), die Gewissenhaftigkeit (Wie perfektionistisch bin ich?), die Extravision (Wie gesellig bin ich?), die Verträglichkeit (Wie kooperativ verhalte ich mich?) und der Neurotizismus (Wie verletzlich bin ich?).

Die diagnostische Methode zur Bestimmung dieser Persönlichkeitsdimensionen ist in der Regel ein Fragebogen.

Ab dem Jahr 2008 hat sich an der Cambridge University eine kleine Forschungsgruppe damit beschäftigt, zunächst eine App für Facebook zu programmieren, mit die Nutzer einen Fragebogen beantworten konnten und dadurch ein Persönlichkeitsprofil auf Grundlage der Ocean-Werte erstellt bekamen (Grassegger & Krogerus, 2018). Für die Nutzer dieser App war diese nur eine kleine Spielerei, die Forscher aber haben von ihnen die Erlaubnis zur Nutzung der Daten ihrer Facebook Aktivitäten eingefordert. So konnte die Forschungsgruppe eine sehr große Datenbasis von mehr als einer Million Nutzer zusammengetragen, die sie als Trainingsdaten für ihre Data Mining Algorithmen benutzen konnten. Das Ziel ihrer Forschung war es, aus den Aktivitäten eines Nutzers auf Facebook, Rückschlüsse auf seine Persönlichkeitsmerkmale zu ziehen (Grassegger & Krogerus, 2018).

Als erster Schritt konnten sie, mit nicht überwachten Analyseverfahren, Muster in den Daten erkennen und Eingangsvariablen bestimmen. Zweitens haben sie dann Prognosemodelle entwickelt, die für bestimmte Eingangsvariablen, wie Likes von Marken oder Musik, die Wahrscheinlichkeit für Persönlichkeitsmerkmale in Form der Ausprägung der Ocean-Werte, sowie außerdem auch für Merkmale, wie die sexuelle Orientierung und die politische Einstellung, bestimmten.

Im Jahr 2014 hat dann eine Person aus dem Umfeld der Forschungsgruppe eine private Firma gegründet und die an Forschungszwecke gebundenen Nutzerdaten mit in das Unternehmen genommen. Zusätzlich hat die Person auch die Idee des Ocean-Modells kopiert und ist mit diesem Wissen an die Firma Cambridge Analytica herangetreten.

Cambridge Analytica ist eine Firma, die beratend tätig ist und behauptet demokratische Wahlen zu Gunsten ihrer Kunden beeinflussen zu können (Grassegger & Krogerus, 2018).

Die Vorgehensweise des Unternehmens ist so, dass auf Grundlage der Facebook-Interaktionen eines Nutzers, über diesen ein Persönlichkeitsprofil erstellt wird. Je nachdem, wie die Persönlichkeit des Nutzers ausgeprägt ist, erhält er dann bei Facebook personalisierte Wahlwerbung zu der beauftragenden politischen Partei oder dem politischen Kandidaten, die den Facebook-Nutzer besonders gut ansprechen soll. Auch können diese Erkenntnisse für den klassischen Straßenwahlkampf nutzbar gemacht werden, indem der Wähler direkt an seiner Haustür mit Hilfe eines personalisierten Gesprächsleitfadens überzeugt werden soll (Grassegger & Krogerus, 2018).

Im November 2015 verkündet die Brexit-Kampagne „leave.eu“, dass sie die Firma Cambridge Analytica beauftragt hat ihren Wahlkampf zu unterstützen. Beim Referendum im Juni 2016 wird die Kampagne mit 52 Prozent Zustimmung zum Erfolg (Hunt & Wheeler, 2018).

Noch am Tag des Wahlerfolgs von Donald Trump am 9. November 2016 versendet das Unternehmen Cambridge Analytica eine Pressemitteilung in der ihr CEO Alexander James Ashburner Nix mit den Worten zitiert wird: „Wir sind begeistert, dass unser revolutionärer Ansatz der datengetriebenen Kommunikation einen derart grundlegenden Beitrag zum Sieg für Donald Trump leistet.“ (Grassegger & Krogerus, 2018).

4 Diskussion

Betrachtet man das Beispiel des Einzelhändlers Target, so fällt auf, dass die Firma bei dem Aufbau seiner Datenbasis nicht nur auf die Daten zurückgegriffen haben, die sie selbst erhoben haben. Neben Transaktionsdaten aus den Filialen, haben sie auch auf ihrer Website bzw. ihrem Onlineshop Kundendaten erhoben und auch die Daten von Mitgliederaktionen und dem Kundenservice hinzugefügt. Zusätzlich ist es in den USA möglich, Personendaten zu seinen Kunden von externen Anbietern hinzuzukaufen. Globale Datenhändler, wie Acxiom und Experian, bieten Daten aus Grundbucheinträgen, Bonuskarten, Wählerverzeichnissen, Clubmitgliedschaften, Zeitungsabonnements und medizinischen Datenquellen an (Grassegger & Krogerus, 2018).

Erst diese personenbezogenen Daten zu ihren Kunden ermöglichten es der Firma ihr Prognosemodell so auszureifen, dass genaue Vorhersagen möglich waren.

Daraus stellt sich die Frage, ob der Aufbau einer solchen Kundendatenbank, außerhalb der Sphären großer amerikanischer Informationskonzerne, in Deutschland möglich ist. Verfügen auch hier die Einzelhandelskonzerne über eine so detaillierte Kundendatenbank, die sich auch aus anderen Quellen speist, als den Mitgliederprogrammen, bei denen die Kunden ihre Personendaten gegen Bonuspunkte eintauschen, die ein Bruchteil des ausgegebenen Geldes Wert sind?

Dabei drängt sich der Fall der Post Direkt, einer Tochterfirma der Deutschen Post, auf. Sie soll im Rahmen des Bundestagswahlkampfes 2017 mit Adressen gehandelt haben (Wilkens, 2018). Die Parteien CDU und FDP haben dabei aber nur auf anonymisierte Daten zugegriffen. Nur wenn die Daten personalisiert sind, ist der Handel mit ihnen illegal und kann mit einem Bußgeld von bis zu 300.000 Euro geahndet werden (Wilkens, 2018).

Die Firma Post Direkt hat also nur statistische Wahrscheinlichkeiten für einzelne Zustellbezirke dargestellt und Käufern angeboten. Informationen über einzelne Haushalten waren nicht enthalten.

Es scheint also so, dass die Kunden in Deutschland durch eine striktere Datenschutzgesetzgebung besser geschützt sind.

Zum zweiten Beispiel, dem vermeintlichen Datenmissbrauch durch Cambridge Analytics, muss betont werden, dass die Profildaten von mehreren Millionen Nutzern auf legalem Wege erhoben wurden. Der Facebook Gründer Mark Zuckerberg spricht selbst nur von einem „Vertrauensbruch“ (Lee, 2018)

So hat Facebook Wissenschaftlern, die mit den Daten der Nutzer forschen wollten, keine Leitlinie an die Hand gegeben, wie sie mit Nutzerdaten verfahren dürfen. Die Nutzer haben

es, beim Ausfüllen von wissenschaftlichen Fragebögen, weitgehend selbst in der Hand welche Daten sie preisgeben. Dieser Umstand befreit die Wissenschaftler aber nicht von ihrer Verantwortung und gibt ihnen nicht das Recht die Daten frei zu nutzen (Kosinski, Matz, Gosling, Popov, & Stillwell, 2016).

So sollte der Fokus eher auf die Praxis von Facebook gelegt werden, Forschergruppen weitgehend unkontrollierten Zugang zu Nutzerdaten zu ermöglichen und die Ziele und Intentionen dieser tausenden Forschungsgegenstände nicht zu überprüfen.

Aus der europäischen Perspektive wirkt das politische Handeln gegenüber den amerikanischen Informationskonzernen, wie Google und Facebook, oft hilflos.

Dabei wird ab dem 25. Mai 2018 die europäische Datenschutz-Grundverordnung angewendet. Sie sieht bei einem Verstoß die Möglichkeit einer Sanktion in Höhe von maximal vier Prozent des weltweiten Umsatzes vor. Das Einhalten von europäischen Datenschutzstandards, durch weltweit operierende Informationskonzerne, ist bald also nur noch eine Frage des politischen Willens.

5 Fazit

Den Firmen, die mit Hilfe von Big Data und Data Mining, Prognosen erstellen, gelingt es oft auf beeindruckende Weise zukünftige Ereignisse vorauszusagen. Sie können ihre Versprechungen in vielen Anwendungsgebieten halten und ermöglichen dabei einen verblüffenden, kurzen Blick in die Zukunft.

Diese Unternehmen, wie der Softwareunternehmer Stehen Wolfram, haben an sich aber oft den Anspruch eine „neue Art der Wissenschaft“ zu betreiben (Mainzer, 2014, S. 25). So sollen Computerexperimente anstelle von mathematischen Beweisen und Theorien treten. Aber nur solche Theorien und Beweise können die Abhängigkeiten und Muster erklären, die durch prognostische Analyseverfahren gewonnen worden.

Die Versprechung, dass die datengetriebene Prognostik eine „neue Art der Wissenschaft“ ist, kann also nicht eingehalten werden.

6 Literaturverzeichnis

Duhigg, C. (16. Februar 2012). *How Companies Learn Your Secrets*. Abgerufen am 8. April 2018 von The New York Times Magazine: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Gräbe, H.-G. (2017). *Strukturen im digitalen Wandel - WS 17/18, 6. VL.* Abgerufen am 8. April 2018 von OpenOLAT: https://olat.informatik.uni-leipzig.de/auth/RepositoryEntry/76513283/CourseNode/87435489161232/path%3D~~Vorlesungsfolien/0

Grassegger, H., & Krogerus, M. (20. März 2018). *«Ich habe nur gezeigt, dass es die Bombe gibt»*. Abgerufen am 8. April 2018 von Tagesanzeiger: https://www.tagesanzeiger.ch/ausland/europa/Ich-habe-nur-gezeigt-dass-es-die-Bombe-gibt/story/17474918

Hunt, A., & Wheeler, B. (26. März 2018). *Brexit: All you need to know about the UK leaving the EU*. Abgerufen am 8. April 2018 von BBC News: http://www.bbc.com/news/uk-politics-32810887

Kindler, M. (1. März 2017). *OCEAN-Methode*. Abgerufen am 8. April 2018 von Health&Care Management: https://www.hcm-magazin.de/ocean-methode/150/23668/345707

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (März 2016). *Facebook as a research tool*. Abgerufen am 8. April 2018 von American Psychological Association: http://www.apa.org/monitor/2016/03/ce-corner.aspx

Kotu, V., & Deshpande, B. (2015). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner.* Amsterdam ; Heidelberg [u.a.]: Elsevier Morgan Kaufmann.

Lee, D. (22. März 2018). *Facebook's Zuckerberg speaks out over Cambridge Analytica 'breach'*. Abgerufen am 8. April 2018 von BBC News: http://www.bbc.com/news/world-us-canada-43494337

Long, C. (2015). *Data Science and Big Data Analytics : Discovering, Analyzing, Visualizing and Presenting Data / EMC Education Services.* Indianapolis, Ind.: Wiley.

MacGregor, J. (2014). *Predictive Analysis with SAP: The Comprehensive Guide.* Boston: Galileo Press.

Mainzer, K. (2014). *Die Berechnung der Welt: Von der Weltformel zu Big Data.* München: C.H.Beck.

Wilkens, A. (3. April 2018). *Datenschutzbehörde prüft Adresshandel mit Post-Daten im Bundestagswahlkampf*. Abgerufen am 8. April 2018 von heise online: https://www.heise.de/newsticker/meldung/Datenschutzbehoerde-prueft-Adresshandel-mit-Post-Daten-im-Bundestagswahlkampf-4010119.html