7 Zweites Moment: Singularität und Vergleichbarkeit

Momente der Datafizierung – das haben wir im vorigen Kapitel gesehen – zeichnen sich dadurch aus, dass in ihnen Geschenk und Gegengeschenk zusammenfallen: Indem Nutzerinnen das Geschenk des Unternehmens annehmen, liefern sie gleichzeitig das Gegengeschenk der persönlichen Daten. Während ich im letzten Kapitel die Beziehung zwischen Unternehmen und Nutzerinnen diskutiert habe, geht es jetzt um die spezifische Qualität von Personendaten als Gabe und Ware zugleich – d.h. um die Frage, wie Personendaten zugleich unentfremdbar und entfremdet sein können beziehungsweise singuläre Handlungen von Individuen und zugleich generische, vergleichbare Verhaltensweisen von Nutzerinnen. Im Anschluss diskutiere ich zwei Varianten der Datafizierung, welche die Singularität von digitalen Handlungen »technisch ignorieren« und diese so, wenn nicht zum Verschwinden bringt, sie zumindest unsichtbar macht. Die erste Variante der Datafizierung ist genuin digital. In ihr werden Verhaltensweisen bereits als digitale, verrechenbare Daten »geboren« (7.2). Eine zweite Variante macht Verhaltensweisen »after the fact« zu Daten, wie ich am Beispiel der Vektorisierung zeige (7.3). Die Gemeinsamkeit besteht darin, dass Verhaltensweisen in beiden Fällen zu Daten und dadurch *vergleichbar* gemacht werden.

7.1 Singuläre Verhaltensweisen

Personendaten verweisen als Spuren auf situierte Verhaltensweisen und Bedeutungen, welche die Handelnden damit verbinden. Was auch immer diese Bedeutungen und Kontexte sind: Damit Personendaten als Daten funktionieren können – d.h. unterschiedliche Verhaltensweisen soweit gleich machen, dass sie vergleichbar sind – muss von diesen Bedeutungen abstrahiert werden. Einerseits sollen Daten »echt« sein, d.h. durch authentische Verhaltens-

weisen authentischer Nutzerinnen zustande kommen. Andererseits werden Bedeutungen und Sinninvestitionen von Nutzerinnen weitgehend ignoriert. Das Wissen, dass Verhaltensweisen authentischer Ausdruck des Selbst oder der sozialen Beziehungen der Nutzerinnen sind, reicht aus. Welche Bedeutungen die Nutzerinnen konkret damit verbinden, ist irrelevant. Wichtiger ist, dass damit gerechnet werden kann.

»Schöne Daten«

Eine Variante, an Memberdaten zu kommen, besteht darin, Member direkt danach zu fragen. Das hat Earlybird vergeblich versucht: Wenn ich mich als Member auf der Webseite anmelde und die Einstellungen meines Profils anklicke, komme ich auf eine Seite, auf der ich Earlybird meine Interessen direkt angeben kann: Zu vier Überkategorien (zum Beispiel: »Lifestyle«) existieren jeweils mehrere Unterkategorien. Für die Überkategorie »Lifestyle« sind das »Reisen«, »Sport«, »Theater/Kultur«, »Fashion«. Diese kann ich jeweils mit einem Häkchen als »interessiert mich« markieren. Zu Earlybirds Bedauern taten das nur sehr wenige Member – trotz teurer »Kommunikationsmaßnahmen«.

Auf Simons Anregung hin – oder in seinen Worten: seinen »Predigten« –, traf Earlybird eine »strategische Entscheidung«: Es sollen möglichst keine weiteren Ressourcen in die Erhebung von »expliziten Daten« gesteckt werden. Solche expliziten Daten, die Interessen und Präferenzen direkt abfragen, sind für Earlybird zu »teuer«, weil sie mit aufwendigen »Kommunikationsmaßnahmen« verbunden sind, um Member aufzufordern und zu motivieren, ihr Profil auf der Webseite auszufüllen. »Von alleine« gehe keiner auf diese Profilseite, um »das anzukreuzen«, sagt Beni.[1] Beni erläutert, was mit »impliziten Daten« gemeint ist: Implizit heiße, Informationen ließen sich aus dem »Verhalten eines Members« oder daraus, was »er uns Preis gibt«, durch »Analysen ableiten«.

In der Literatur zu Empfehlungssystemen, auf die mich Dani hinweist, finde ich weitere Hinweise zur Unterscheidung von impliziten und expliziten Daten. Gemäß Michael Ekstrand et al. (2011, S. 129 ff.) sind explizite Daten von den Nutzerinnen eines Empfehlungssystems explizite geäußerte Präferenzen – zum Beispiel wie gut jemand einen Film auf einer Skala von 1 bis 5 bewertet. Im Kontrast dazu: »implicit›ratings‹ are inferred by the system from observable user activity, such as purchases or clicks« (Ekstrand et al., 2011, S. 129).[2] Der Nachteil von expliziten Daten liege darin, dass oftmals eine Diskrepanz zwischen geäußerten Präferenzen und dem, was die Nutzerinnen tatsächlich

mögen, bestehe: »a discrepancy between what the users say and what they do« (Ekstrand et al., 2011, S. 130). Ekstrand & Willemsen (2016) legen dar, dass in der Entwicklung von Empfehlungssystemen ein behavioristisches Paradigma dominant sei, das sich nicht mehr auf Umfragen oder ethnografische Analysen verlasse, um Nutzungsweisen zu analysieren, sondern das »tatsächliche« Nutzerverhalten bevorzuge – »ignoring [stated] preference when it disagrees with behavior« (2016: 221).[3]

Die Unterscheidung von expliziten und impliziten Daten ist für Earlybird aber nicht in erster Linie als technische Unterscheidung zentral. In Earlybirds Praxis und dem, was Simon im Rahmen des Datenexperiments (siehe Kapitel 5) »schöne Daten« nennt, verschwimmt die Unterscheidung. Die »schönen Daten« müssen explizit erfragt werden, doch steckt in ihnen mehr als die bloßen Informationen darüber, wohin die Teilnehmerinnen reisen und was sie dort tun möchten. Es sind – in Benis Verständnis – auch implizite Daten, weil sie auf latente Sinngehalte und Potenziale verweisen.

Was macht Daten zu »schönen Daten«?

In den Daten-Diskussionen bei Earlybird kristallisieren sich drei Gründe heraus, wieso der betreffende Datensatz »schön« ist. Die drei Aspekte der Schönheit von Daten entsprechen je einem Moment der Datafizierung.

Erstens fallen Freitextantworten nicht automatisch als Nebenprodukt des Gebrauchs von digitalen Infrastrukturen an: Die Member müssen auch hier über spezifische »Kommunikationsmaßnahmen« zur Teilnahme und Preisgabe von Informationen motiviert werden. Im Fall des Wettbewerbs hat das quasi »zufällig« geklappt – ohne technische Vorrichtungen zur »Qualitätssicherung«, wie Simon sagt. Das wirft für Earlybird die Frage auf, wie man diesen Erfolg replizieren könnte und welche »Anreize« sie Membern bieten müssen. Wie im letzten Kapitel beschrieben, positionieren sie die Chance auf einen Gewinn als Motivator und Tauschgegenstand. Für Simon ist klar, dass es sich bei den Wettbewerbsantworten um explizite Daten handelt. Es sei schwierig, aber möglich, »explizite Präferenzen« zu erfragen – wie die Wettbewerbsantworten zeigen sogar in »unfassbarer Qualität«. Diese unfassbare Qualität verweist im ersten Moment darauf, dass Member unter den richtigen Umständen bereit sind, etwas von sich preiszugeben (siehe Kapitel 6). Die meisten Teilnehmerinnen gaben ausführliche Wettbewerbsantworten, obwohl das gar nicht nötig gewesen wäre, um am Wettbewerb teilzunehmen. Niemand ge-

be auf der Seite von Earlybird seine Präferenzen an. Hier hätten die Teilnehmerinnen aber mit Begeisterung Antworten gegeben, so Simon. Der Wettbewerb machte aus unmotivierten Membern motivierte Teilnehmerinnen, die Daten im Austausch gegen eine Gewinnchance preisgeben.

Zweitens verweist die Begeisterung der Teilnehmerinnen auf einen weiteren Aspekt der Schönheit: Die Teilnehmerinnen waren »intrinsisch motiviert«, wie Max sagt. Was meint er damit? Die Quasi-Umfragen, die sie auf ihrer Profilseite durchgeführt hatten, waren für die Member bedeutungslos: Sich für ein Unternehmen in Konsumkategorien »einzureihen«, ist keine Tätigkeit, die Jugendliche für sich ausüben. Sich darüber Gedanken zu machen, wohin man reisen möchte und was man dort alles für tolle Dinge tun wird hingegen schon, wie Earlybird spekuliert. Die »Schönheit« von schönen Daten besteht darin, dass die registrierten Verhaltensweisen auf »echte«, für die Nutzerinnen selbst bedeutungsvolle, Handlungen verweisen. Das macht die Freitextantworten aber auch zu singulären, unvergleichbaren Äußerungen, die sich nur unter großem Aufwand maschinell weiterverarbeiten lassen. Wie ich weiter unten zeige (7.3), ersetzen Simon und sein Team den soziale Kontext des Wettbewerbs durch einen abstrakten Vektorraum. Darin erscheinen die verschiedenen Antworten als geometrische Repräsentationen, die sich in Bezug auf ihre Ähnlichkeit und Differenz vergleichen lassen.

Der *dritte* Aspekt der Schönheit besteht darin, dass die Daten in einem weiteren Sinne auf »etwas anderes« verweisen (siehe auch Kapitel 8). Earlybird interessiert sich nicht per se dafür, wohin die Teilnehmerinnen reisen möchten – auch wenn das vielleicht im Interesse eines Partnerunternehmens ist, mit dem der Wettbewerb durchgeführt wurde. Earlybird möchte wissen, welche kategorialen Zugehörigkeiten sich in den Freitextantworten verbergen. Sie sprechen den Daten das Potenzial zu, neue Relationen zwischen Membern und Dingen (beziehungsweise den Deals von Partnerunternehmen und potenziellen Werbepartnern) begründen zu können. Mit den Wettbewerbsdaten sei es möglich, Earlybirds brachliegende Marketingsegmente zu reaktivieren. Bisher habe die Möglichkeit gefehlt, Membern zuverlässig Segmenten zuzuordnen. Aus den ausführlichen und persönlichen Texten der teilnehmenden Member, so Simons Idee, ließe sich die Zugehörigkeit zu den Segmenten ableiten. Die explizit erhobenen Daten könnten auf implizite kategoriale Zugehörigkeiten hinweisen. In diesem Sinn ist es nicht der manifeste Inhalt der Freitextantworten, sondern die latente und kalkulierbare, kategoria-

le Zugehörigkeit, auf welche die Daten hinweisen und die durch eine Analyse manifest gemacht werden können – so die Hoffnung von Earlybird.

Simon – der sich gemäß Beni nicht nur als Mathematikprofessor, sondern auch als Verkäufer sehr gut machen würde – versteht es, das Potenzial dieser Daten zu kommunizieren. Seine Begeisterung wirkt ansteckend: Simon kann Earlybird anhand der Schönheit und des Sinnüberschusses der Wettbewerbsdaten davon überzeugen, ein Datenexperiment zu finanzieren.

Die Sinnüberschüsse der Verhaltensdaten bestehen darin, dass sie als Spuren von Nutzerinnen gelesen werden, die auf subjektiv bedeutsame Handlungen ihrer Urheberinnen verweisen. Earlybirds Problem besteht vor allem darin, den Nutzerinnen eine Infrastruktur für Verhaltensweisen, die sie eigenmotiviert ausüben möchten, anbieten zu können. Mit den Freitextantworten des Wettbewerbs hat das »zufälligerweise« geklappt.

Christian Rudder, Gründer der Datingseite OkCupid, bringt diese Problemlage auf den Punkt: Die Verhaltensweisen im Onlinedating müssen so formalisiert werden, dass Computer sie verstehen können. Gleichzeitig müssen sie den Nutzerinnen aber weiterhin als mehr oder weniger natürliche, »echte» Verhaltensweisen erscheinen:[4]

1 Simon gibt ein anderes Beispiel für die teure Erhebung expliziter Daten. Ein schweizer Einzelhändler hatte einen Brief an alle Mitglieder seines Kundenbindungsprogrammes geschickt, um Geburtstagsdaten abzufragen. Offenbar sei das Geburtsdatum für diesen Retailer wichtig, um »Zielgruppen« zu identifizieren und individuelles »Profiling« zu machen. Das sei eine »teure Variante des Data Cleaning«, könne aber unter Umständen gerechtfertigt sein, erklärt Simon (siehe auch Mützel et al. 2018, S. 122).

2 Siehe auch Thurman & Schifferes (2012, S. 776).

3 Ekstrand/Willemsen (2016) kritisieren diese Praxis, die Nutzerinnen zu »ignorieren«.

4 Die »Echtheit« der Nutzerinnen und ihrer Verhaltensweisen ist auch für Facebook zentral (Bivens, 2017). Am 30. Juni 2012 gab Facebook bekannt, dass der Anteil falscher Profile auf 8.7 Prozent angewachsen sei. Zur Zeit von Facebooks Börsengang am 18. Juni 2012 waren es noch fünf bis sechs Prozent gewesen. In den ersten drei Monaten als börsengehandeltes Unternehmen fiel Facebooks Börsenwert auf knapp die Hälfte der 38 US-Dollar bei Börsengang. Aktuell werden falsche Profile vor allem als politisches Problem behandelt: Als Verbreiter und Verzerrer der öffentlichen Meinung. Für Facebook stellen sie aber ein ökonomisches Problem dar: »Facebook's marketable product is a user base of ›real‹ people that can be targeted with the help of increasingly granular data« (Bivens, 2017, S. 884). »Authentic identity« ist zentraler Bestandteil davon, wie Facebook Wert generiert, wie sie in ihrer IPO-Broschüre festhalten: »Authentic identity is core to the Facebook experience, and we believe that it is central to the future of

> Algorithms don't work well with things that aren't numbers, so when you want a computer to unterstand an idea, you have to convert as much of it as you can into digits. The challenge facing sites and apps is thus to chop and jam the continuum of human experience into little buckets 1, 2, 3, without anyone noticing: to divide some vast, ineffable process – for facebook, friendship, for Reddit, community, for dating sites, love – into pieces a server can handle. At the same time you have to retain as much of the je ne sais quoi of the thing as you can, so the users believe what you're offering represents real life. (Rudder, 2014, S. 13)

Auch Shoshana Zuboff sieht zwischen Formalisierung und »subjectivities« der Nutzerinnen einen Konflikt, wenn sie den Wert von Personendaten in ihrem Verweis auf »subjectivities« sieht. Unternehmen wie Google nehmen aber gegenüber ihren Nutzerinnen eine Position der »formal indifference« ein, die individuelles Verhalten abflacht und auf »bits« reduziert.

> These subjectivities travel a hidden path to aggregation and decontextualization, despite the fact that they are produced as intimate and immediate, tied to individual projects and contexts (Nissenbaum, 2011). Indeed, it is the status of such data as signals of subjectivities that makes them most valuable for advertisers. For Google and other ›big data‹ aggregators, however, the data are merely bits. Subjectivities are converted into objects that repurpose the subjective for commodification. Individual users' meanings are of no interest to Google or other firms in this chain. (Zuboff, 2015, S. 79)

Zuboff vertritt hier eine Position, die in Referenz auf Zelizer als Variante des »hostile worlds«-Arguments verstanden werden kann: Um komplexe und vielschichtige Nutzeraktivitäten zu Datensätzen zu machen, braucht es gewissermaßen den versachlichenden Blick des Markts, um den affektgeladenen, digitalen Handlungen der Nutzerinnen ihr Leben zu entziehen: Der »kalte Blick« der formalen Indifferenz reduziert die warmen Aktivitäten der Nutzerinnen auf ihr Skelett. Demgegenüber hält es Kylie Jarrett (2015) für

---

the web. Our terms of service require you to use your real name and we encourage you to be your true self online, enabling us and Platform developers to provide you with more personalized experiences« (zitiert in: Bivens 2017, S. 885). Auf Facebook den richtigen Namen zu verwenden und auch online das »wahre Ich« zu sein (beziehungsweise von Facebook dazu motiviert zu werden), ist eine Funktionsbedingung für Facebooks Businessmodell. Es beruht auf der Annahme, dass sich in unseren digitalen Verhaltensweisen unser »echtes Selbst« dokumentiert (siehe auch boyd 2014).

notwendig, beides gleichzeitig denken zu können: Unternehmen wie Facebook müssen ihren Nutzerinnen die Ausübung von sinn- und affektgeladenen, digitalen Verhaltensweisen ermöglichen, die für die Nutzerinnen selbst bedeutsam sind. Obwohl diese Sinnüberschüsse weitgehend weggearbeitet und ignoriert werden müssen, um mit den dadurch entstehenden Daten rechnen zu können, sind sie doch elementar dafür, Nutzerinnen zu motivieren. Ansonsten hätten wir es mit einem für Nutzerinnen langweiligen Anklicken von Kästchen und Ausfüllen von Fragebögen zu tun. Wie Earlybird erfahren musste, ist das nicht etwas, was Nutzerinnen begeistert.

Jarrett schlägt vor, die Tätigkeiten von Nutzerinnen digitaler Plattformen analog zur Reproduktionsarbeit zu verstehen, um einen Antagonismus zwischen Markt und Intimität, Ware und Geschenk sowie Produktion und Reproduktion zu vermeiden. Sie beschreibt im Anschluss an Leopoldina Fortunati (1995) ein Zwei-Phasen-Modell der Werterzeugung für Social Media. So wie der männliche Arbeiter auf weibliche Reproduktionsarbeit angewiesen ist, um seine eigene Arbeitskraft als Ware für den Markt reproduzieren zu können, ist auch der Werterzeugungsprozess von Personendaten auf nichtkommodifizierte Arbeit angewiesen. Wenn das Verhaltensrepertoire für Nutzerinnen bedeutungslos ist oder sich niemand auf der Seite aufhält, kommt die Zirkulation neuer Inhalte und die Produktion von Personendaten ins Stocken. Verhaltensweisen auf Facebook haben gleichzeitig »use-value« für die Nutzerinnen und »exchange-value« für die Plattform, wie es Jarrett formuliert:

> We ›like‹ things first and foremost because we like them, and it is this use-value that produces the impetus to use and continue to use the site; that produces the instantiated capacity to generate user data. Thus, Facebook can only convert the ›labor-power‹ of user experience (living labor) into the commodified form of user data (labor-time) *after* its experience as inalienable use-value by the user. […]
> ›Liking‹ a friend's status update continues to manifest an inalienable and affectively powerful social relationship, or even asserts a political statement. Thus, while the generation of user data on Facebook is implicated in the capitalist valorization process, it cannot accurately be described as an inherently

> exploitative or wholly commodified process. (Jarrett, 2014, S. 20f., Hervorhebung im Original)[5]

Der »Trick« von Personendaten besteht also gerade darin, dass sie beides gleichzeitig können: Sie verweisen als Spuren immer auf »mehr«, auf etwas, das außerhalb ihrer selbst steht, d.h. die Sinninvestitionen oder der »use-value« der Nutzerinnen selbst. Dieses Mehr ist Gegenstand einer ausführlichen Datenkritik, die einen Reduktionismus der Datafizierung bemängelt (Gitelman, 2013; Puschmann & Burgess, 2014) aber darauf hinweist, dass Daten diesen Kontext immer irgendwie mittragen (Seaver 2015; Loukissas 2019; Leonelli 2019, siehe auch: Kapitel 2). Dieses Mehr begründet den Wert der Daten und motiviert Nutzerinnen zur weiteren Nutzung, muss aber zeitweise »ignoriert« werden, um diesen Wert zum Vorschein zu bringen.

Während die Benutzeroberfläche für Nutzerinnen bedeutungsvolle Verhaltens- und Kommunikationsoptionen zur Verfügung stellt, werden auf der technischen Hinterbühne die sozialen Kontexte und subjektiven Bedeutungen aus den Verhaltensdaten weggearbeitet, um sie zu einer »entfremdeten« Ware oder Ressource zu machen. Die Sinnüberschüsse beziehungsweise die »subjectivities« der Nutzerinnen sind zentral, um weitere Verhaltensweisen zu motivieren, verhindern aber deren Vergleichbarkeit (Heintz, 2010) beziehungsweise Kommensurabilität (Espeland & Stevens, 1998). Grundsätzlich gilt, dass die Kontinuität des gelebten Alltags und gelebter Identitäten in diskrete Kategorien und Handlungsweisen übersetzt oder als solche erst geschaffen werden müssen (Alaimo & Kallinikos, 2017), damit Verhaltensweisen datafiziert oder alternativ kommodifiziert werden können.

Im Folgenden beschreibe ich zwei Varianten, wie diese Vergleichbarkeit technisch hergestellt wird: Die erste besteht darin, Verhaltensweisen und ihre Registrierung über »encoding« (Alaimo & Kallinikos, 2017, 2016) zu vereinen. Die zweite Variante stellt Vergleichbarkeit her, *nachdem* Verhaltensweisen registriert wurden. Beide Varianten »kommodifizieren« Verhaltenswei-

---

5 Hier ließe sich ein ganzer Forschungszusammenhang anfügen, der sich mit der Frage von Medienkonsum als Arbeit (Smythe, 1977) beziehungsweise der Nutzung von Web-2.0-Angeboten und Social Media als Arbeit befasst (Terranova, 2000; Fuchs, 2014; Ekbia & Nardi, 2017). Von Bedeutung wäre insbesondere die Frage, wie Social-Media-Plattformen den Austausch ihrer Nutzerinnen als unentfremdete Arbeit und Tätigkeit instrumentieren und motivieren, um den Motor von Verhaltens- und Datengenerierung am Laufen zu halten.

sen: Sie lösen sie aus ihrem bestehenden sozialen Kontext heraus, befreien sie von subjektiven Bedeutungen und machen sie zu einer Ressource für Analysen (bei Alaimo & Kallinikos 2017: »computation«) und Vergleiche, um neue Relationen abzuleiten (siehe dazu Kapitel 8).[6]

7.2 Encoding

Mit dem Begriff »Encoding« bezeichnen Alaimo & Kallinikos (2017, 2016) die digitale Standardisierung von Verhaltensweisen, durch die Nutzerinnen oder Dinge (zum Beispiel Beiträge auf Social Media) vergleichbar gemacht werden.

Auf digitalen Benutzeroberflächen ist jede Handlungsoption vorgegeben: Im Code einer Webseite, einer App oder eines Streamingdiensts ist im Detail bestimmt, welche »Aktionen« die Nutzerinnen ausführen können. Nutzerinnen »interagieren« mit »Objekten« und erzeugen dadurch »Relationen«. Aus welchen Einheiten, Aktionen und Relationen die digitale Welt besteht, muss dementsprechend im Voraus von Programmiererinnen und User-Experience-Designerinnen festgelegt werden. Bevor ich also überhaupt etwas auf Facebook tun kann, muss Facebook Entscheidungen darüber treffen, was erwünschte Handlungen sind – zum Beispiel »like«, aber nicht »dislike«. Diese erwünschten Handlungen müssen dann in der Sprache der Benutzeroberfläche (»blauer Daumen hoch«) und in der Sprache der Datenbank (User X likes Object Y) artikuliert und in ein formales Modell von Userverhalten übersetzt werden: Der »like« wird darin als Handlung definiert, die User und bestimmte erlaubte Objekte – Kommentare, Posts und Brands, aber

6 Jens-Erik Mai (2016) unterscheidet das Überwachungsmodell und das Capture-Modell (beruhend auf Agre 1994): Das Überwachungsmodell betrachtet Daten als getreue Wiedergaben des Beobachteten. Das Capture-Modell geht davon aus, dass technologische Apparaturen nicht nur beobachten und wiedergeben, sondern das Beobachtete auch verändern (Mai, 2016, S. 198). Im ersten Fall werden bereits bestehende, von der Beobachtungsapparatur relativ unabhängige Phänomene bloß registriert. Im zweiten Fall sind die Beobachtungsapparaturen gleichzeitig Infrastrukturen, die das zu registrierende Verhalten überhaupt erst ermöglichen: Sie stellen Benutzeroberflächen zur Verfügung, in denen die Nutzerinnen nach vorgefertigten »grammars of action« handeln können. Aktivität wird dabei so restrukturiert, dass sie mit ihrer formalen Repräsentation übereinstimmt (Agre, 1994, S. 105-107). Oder weniger zugespitzt: Die Welt und ihre Repräsentation entwickeln sich Hand in Hand (Berg, 1997, S. 409-410).

nicht andere Nutzerinnen – in eine »like«-Relation setzt.[7] Das bedeutet, ich kann auf Facebook nur im Rahmen der vorprogrammierten Möglichkeiten handeln. Der Vorstrukturierung von Aktivitäten auf der Ebene der Nutzeroberfläche entspricht eine Modellierung dieser Aktivitäten in der Datenbank: Objekte wie Nutzerinnen, Posts oder Produkte sind über »actions« wie »like« oder »share« verbunden, die in der Datenbank Relationen zwischen den Objekten erzeugen.[8]

Die mythologisierende Rede des Sammelns von Daten basiert auf diesem Prozess der Infrastrukturierung von alltäglichen sozialen Verhaltensweisen (Alaimo & Kallinikos, 2019; Gerlitz & Helmond, 2013): Sobald Nutzerinnen die instrumentierten Verhaltensweisen wie den »like« als legitime Verhaltensweisen akzeptieren und ausführen, erscheinen solche Aktivitäten nicht

7 Nur weil der »like« als vorprogrammierte, digitale Verhaltensweise zur Verfügung steht, heißt das aber noch nicht, dass »liking« auch tatsächlich eine für die Nutzerinnen bedeutsame Verhaltensweise ist, die sie von sich aus ausführen: »The people who engage in the articulated activity are somehow induced to organize their actions so that they are readily ›parsable‹ in terms of the grammar« (Agre, 1994, S. 110).

8 Kent (2012) hebt die vielen kontingenten Entscheidungen hervor, die in die Gestaltung von Datenbanken eingehen: Wie wird die soziale Welt in »entities«, »relationships« oder »attributes« formalisiert? In seinem Buch »Data and Reality – A timeless perspective on perceiving and managing information in our imprecise world« beschreibt William Kent, welche Probleme sich bei der »representation of information in computers« ergeben (Kent, 2012, S. 28). Selbst bei so »einfachen« Dingen wie dem Wareninventar, Personaldateien oder Bankkonten müssen zahlreiche Fragen beantwortet werden, die trivial zu sein scheinen: Was ist »ein« Ding? Wie viele Dinge sind es? Was ist es? Für wie lange? Kent zeigt an ganz alltäglichen Beispielen, wie voraussetzungsreich es ist, die Einheit und Differenz der Dinge zu bestimmen. Zum Beispiel: Wie ist damit umzugehen, wenn es mehrere Kopien des gleichen Buchs in der Bibliothek gibt? Immer wieder betont Kent die »arbitrariness« der Entscheidungen, die Programmierer bei der Modellierung der chaotischen und kontinuierlich verlaufenden Realität treffen müssen. Graeme Simsion, Autor von »Data Modeling Essentials« (2007) und »Data Modeling: Theory and Practice« (2013), schreibt im Vorwort zur Neuauflage von Kents »Data and Reality«: »William Kent uses the word [arbitrary] throughout the book […] to characterize some of the most important decisions that data modelers make. The boundaries of an entity are arbitrary, our selection of entity types is arbitrary, the distinction between entities, attributes, and relationships is arbitrary« (Kent, 2012, S. 13). Leider habe sich daran kaum etwas geändert: Die grundlegenden Probleme seien immer noch die gleichen. Es werden zwar neue Formalismen entwickelt, doch der Fokus liege auf dem Vergleich der verschiedenen formalisierten Modelle und nicht auf den grundlegenderen Fragen, die Kent aufgeworfen hat.

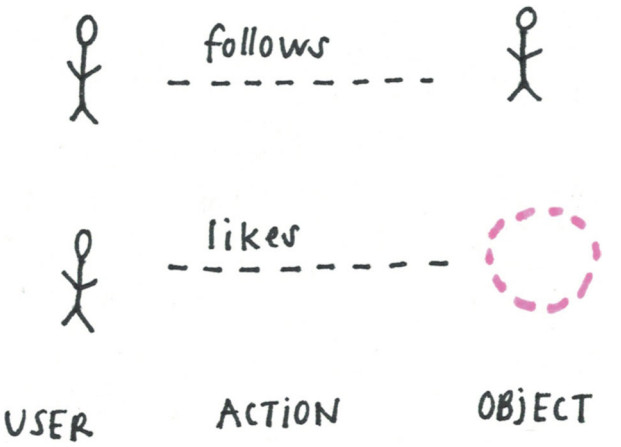

Abb. 3: Datenrelationen (nach: Alaimo & Kallinikos 2016, S. 81)

mehr als »Erfindung«, sondern können »entdeckt« und beobachtet werden, als würden sie natürlich auftauchen.[9]

> Encoding does not record transactions, or simple online behavior (e.g., time spent on Web pages or clickthrough rates); it does not record prior facts, which it then places online, nor does it categorize existing social activities (we do not usually ›follow‹ friends offline). Rather, encoding creates the actions that users are invited to perform and records the performance of such actions into distinct data fields. In this regard, it establishes the terms of user

9 Auch historische Beispiele verdeutlichen, dass sich Daten nicht einfach auffinden lassen, sondern erzeugt werden müssen, um die »Welt der digitalen Computer« und die »Welt der Menschen« zu koppeln (Gugerli, 2018, S. 50). So schrieben Ridenour & Brown (1953, S. 80): »appropriate input and output equipment to couple the world of the digital computer to the world of men often does not exist«. Um beispielsweise einen Computer für die Buchhaltung verwenden zu können, müsse erst ein »tiefes Verständnis« der Aktivität der Buchhaltung vorliegen und Mittel und Wege zur Verfügung stehen, um dem Computer die relevanten Informationen zu übergeben. Es handelt sich also um ein Problem der Formalisierung von Tätigkeiten, die in diskrete Teiltätigkeiten zerlegt werden müssen. Ein Problem der »Formatierung«, wie Gugerli schreibt: »Das Formatieren von Daten war mithin die conditio sine qua non, um unterschiedlichste Handlungsfelder an die Fähigkeiten des Rechners anzupassen« (Gugerli, 2018, S. 49-59).

> platform participation and involvement through the structuring of the user interface. (Alaimo & Kallinikos, 2017, S. 177)[10]

Christina Alaimo und Jannis Kallinikos stellen fest, dass sich digitale Handlungen und ihre Beobachtung verdichten. Am Beispiel von Social Media (Alaimo & Kallinikos, 2016, 2019) und Webshops (Alaimo & Kallinikos, 2017) zeigen sie, wie Handlung und Beobachtung immer näher zusammenrücken und in digitalen Infrastrukturen zusammenfallen. Carolin Gerlitz betont, dass die Aktivität und die Erfassung dieser Aktivität gleichzeitig als zwei Seiten einer Medaille stattfinden:

> Friending, following, liking, commenting, sharing or favoriting allows users to act in prestructured form in the front end whilst at the same time producing equally prestructured data points in the back end. (Gerlitz, 2017, S. 242)

Encoding ermöglicht die Gleichzeitigkeit subjektiv bedeutsamer Verhaltensweisen auf der Nutzeroberfläche und einer »formal indifference« in der Datenbank. Ich kann auf ironische Art das Buch »Harry Potter und der Feuerkelch« auf Facebook »liken«, um einem befreundeten Harry-Potter-Fan, mit dem ich tags zuvor über die kulturelle Relevanz von Harry Potter gestritten hatte, ein Signal zu senden. Dadurch entsteht eine formale Relation in der Datenbank, die nicht zwischen meinem »ironischen« und einem ernst gemeinten Like unterscheiden kann. Differenzen in der Gebrauchsweise des Likes lassen sich so einebnen. Die Bedeutung, die Facebook mit einem Like verbindet, muss nicht dieselbe sein, welche User damit verbinden. Facebook mag den Like als positive emotionale Reaktion definieren. Die User müssen diese Deutung aber keineswegs teilen. Ein Like kann ironisch, als Lesezeichen oder als Geschenk in einer reziproken Like-Ökonomie (Romele & Severo, 2016) vergeben werden, wie es zwischen Jugendlichen üblich ist.[11]

---

10 Gerade in Bezug auf Social-Media-Plattformen wie Facebook ist diese Reorganisation menschlicher Aktivität besonders evident: Die von Facebook vorgegebenen Verhaltensangebote wie »friending« oder »liking« erscheinen zwar wie Alltagsaktivitäten, sind in ihrer Anwendung aber grundsätzlich nicht vorgefundene und bloß registrierte, sondern von Facebook erzeugte Aktivitäten. Wie danah boyd (2006) am Beispiel von Myspace zeigt, entstehen dabei ganz neue Handlungsprobleme, wenn beispielsweise Social-Media-Nutzerinnen entscheiden müssen, welche ihrer Freunde zuoberst in ihrer Freundesliste erscheinen.

11 Paßmann & Gerlitz (2014) beschreiben beispielhaft, wie die Like-Funktion auf Twitter erst von findigen Nutzerinnen erfunden und später von Twitter integriert wurde. Sie

Die Stärke und der Wert des Likes als formalisierte »action« bestehen gerade darin, dass auf der Seite der User interpretative Flexibilität möglich ist und auf der Seite der Datenbank die Handlung des Likens gleichzeitig so standardisiert ist, dass unterschiedlichste User über ihre Likes vergleichbar werden: »Defining an individual user as an aggregation of likes immediately renders the individual qua likes commensurable to other individuals qua likes« (Alaimo & Kallinikos, 2017, S. 179).

Die Datenbank wird damit zum zentralen Produktionsmittel, das Nutzerinnen (und Objekte) als datafizierte Relationen von Nutzerinnen und Objekten herstellt, welche standardisiert und vergleichbar sind und sich für weitere Bearbeitungsschritte anbieten:

> By capturing consumer activities ubiquitously and in minute detail, databases become repositories of complex consumer lives by turning behavior into abstract aggregates of individualized and individualizing data points. Once consumption has been dematerialized and been made available as coded, standardized and manipulable data, there are no more limits to the construction of difference, to classification, and to social sorting. (Zwick & Denegri Knott, 2009, S. 222)

Die Vorstrukturierung möglicher Verhaltensweisen immunisiert die Datenproduktion gegen die subjektiven Deutungen der Nutzerinnen und ermöglicht damit die Produktivität von Daten, die nun für unterschiedlichste Zwecke genutzt werden können: Zum Beispiel für die Analyse der Plattformaktivitäten (z.B. an welcher Stelle verlassen User regelmäßig die Plattform), die Optimierung der Benutzeroberfläche (siehe Holson 2009 für Marissa Mayers »41 Shades of Blue«-Anekdote) oder die Berechnung von Scores und Affinitäten/Interessen, auf deren Basis Werbung angezeigt oder Empfehlungen ausgesprochen werden können (siehe Kapitel 8 und 10 zur Art und Weise, wie diese Daten weiterverarbeitet werden können).

Bisher bin ich davon ausgegangen, dass Nutzerinnen sich immer schon in Datafizierungsinfrastrukturen befinden. Das nachfolgende Beispiel von Earlybird zeigt, dass der Formalisierungsprozess der Nutzerinnen und ihrer Verhaltensweisen schon früher einsetzt. In Bezug auf die Nutzerinnen zeige ich im Folgenden, wie sie durch verschiedene »infrastrukturelle Quantensprünge« von Personen zu Membern und dann zu »Nutzerinnen« werden,

---

zeigen, dass für unterschiedliche Nutzergemeinschaften der »Twitter-Fav« ganz unterschiedliche Bedeutungen haben kann.

denen encodierte Verhaltensweisen zur Verfügung stehen (siehe auch Kapitel 10.4).[12]

Eintreten ins »Earlybird-Universum«

Personen begeben sich in Dateninfrastrukturen hinein oder werden in sie hineingezogen. Dort durchlaufen sie (im Falle von Earlybird) verschiedene Kategorien: Erst das Eröffnen eines Jugendkontos macht aus normalen Jugendlichen »Earlybird-Member«. Kommen sie aus dem bezugsberechtigten Alter heraus oder künden sie ihr Konto, werden sie von aktiven Membern zu passiven Datenbankobjekten. Member, welche die App herunterladen, sich anmelden und tätig werden, indem sie Deals anschauen, liken oder bookmarken, werden zu Nutzerinnen. Wer die App genügend oft benutzt, kann zu einem »engaged user« werden. Wer genügend Informationen mitteilt, kann im Vergleich mit anderen einer bestimmten Kategorie zugeordnet werden (siehe Kapitel 10.4). Die verschiedenen Jugendlichen werden durch infrastrukturelle Siebe[13] geschüttet, so dass genügend Homogenisierung erreicht werden kann, um Differenzen zwischen den Jugendlichen beobachten zu können. Die Metapher des Siebs ist aber auch trügerisch: Ob die Jugendlichen ein Sieb passieren oder nicht, ist möglicherweise weniger von tatsächlichen Eigenschaften oder Verhaltensweisen abhängig als vielmehr davon, was für die Datenbank sichtbar ist.

Das Sieb der Banken: Member

Die Datenbank von Earlybird umfasst mehr als 200 000 Personen in der Deutschschweiz. Sobald ein Jugendlicher ein Jugendkonto bei einer teilnehmenden Bank abschließt, übermittelt die Bank Personendaten an Earlybird: Die Person wird zu einem Earlybird-Member, sobald Name, Adresse und Geburtstag von der Datenbank, dem Excel-File oder der Liste der Bank in die Datenbank von Earlybird wandert. Die Kundengewinnung ist Sache der Banken: Sie übernehmen das Marketing für ihre Jugendkonten, wobei der Verweis auf Earlybirds Geschenke ein zentrales Argument ist, wie die Jugendwerbung verschiedener Banken nahelegt. Die Überweisung neuer Member an Earlybird geschieht zunehmend reibungslos: Viele Banken verfügen über Protokolle mit Earlybird, die den Prozess automatisieren.

12 Der Begriff der infrastrukturellen Quantensprünge ist an Zerubavel (1996) angelehnt, der die Überbrückung kategorialer Grenzen als »mental quantum leaps« bezeichnet.

Das Sieb des Alters: Aktive und passive Member

Sobald Jugendliche in der Datenbank von Earlybird angekommen sind, werden sie zu Membern. Earlybird sendet ihnen per Post und Email (zu Beginn und dann periodisch) Hinweise auf aktuelle Angebote, den Link zur Webseite und zur Installation der App. Zusätzlich erhalten sie jedes Jahr eine Kundenkarte, welche sie zum Bezug vergünstigter Angebote bei den verschiedenen Partnerunternehmen berechtigt. Wer sein Konto auflöst oder aufgrund des Alters (die Grenze ist je nach Bank verschieden und liegt zwischen 26 und 30 Jahren) die Berechtigung verliert, verbleibt zwar in der Datenbank, erhält aber in einem spezifischen Statusfeld den Eintrag: »passiv«. Damit endet auch die Berechtigung zur »Aktivität«: Der Bezug von Earlybirds Rabattangeboten oder die Teilnahme an Wettbewerben ist den berechtigten »aktiven« Membern vorbehalten. Der Wert im Statusfeld aktiv/passiv entscheidet über die aktuelle kategoriale Zugehörigkeit im Earlybird-Universum.

Das Sieb des Logins: User und Nicht-User

Es können zwar sowohl Member als auch Nicht-Member die App herunterladen, doch wird die volle Funktionalität nur freigeschaltet, wenn die im Loginprozess eingegebene Telefonnummer in der Datenbank vorhanden und nicht mit dem Passiv-Flag im Statusfeld versehen ist. Das ist zumindest die Idealvorstellung von Earlybird. Die Gestaltung des Loginprozesses erweist sich aber als nicht so einfach. In einer Arbeitssitzung zum Loginprozess warnte Sabina: Die Zuordnung über die Telefonnummer funktioniere nicht, wenn sie ein »Datenghetto« hätten. Sie fügte an: »Und das haben wir!«. Es stellte sich heraus, dass eine der Banken verlangt habe, zusätzliche Telefonnummern in den Memberdatensatz aufzunehmen. Seither gebe es Probleme mit diesem Daten-»Güsel«: Member seien doppelt vorhanden. Bei manchen seien falsche Nummern angegeben.

Das Sieb der encodierten Verhaltensweisen

Der Loginprozess und die Telefonnummern sind für Earlybird von zentraler Bedeutung, weil die Telefonnummer als »unique identifier« der User dient. Schaut sich eine Nutzerin Deals oder Partnerunternehmen an, vergibt Likes oder nimmt an Wettbewerben teil, soll dies als Tätigkeit dieser spezifischen Nutzerin registriert werden. Dies funktioniert nur, wenn sie eindeutig identifiziert werden kann.

Wie das Tracking-Schema der App zeigt (siehe Abbildung 4), haben Member in der App oder auf der Webseite zahlreiche Möglichkeiten mit verschiedenen Objekten wie Deals, Notifikationen oder Wettbewerben zu interagieren. Tun sie das, entsteht in der Datenbank eine »view«-, »like«- oder »use«-Relation zwischen der spezifischen Nutzerin und beispielsweise einem Deal, den sich die Nutzerin angesehen, gelikt oder eingelöst hat. Die Handlungsoptionen und ihre »Bedeutungen« sind formal im Tracking-Schema festgelegt (siehe Abbildung unten oder sehr zugänglich in Bezug auf Datenmodelle bei Kent 2012), das darüber Auskunft gibt, wie das Verhalten der Nutzerinnen in der Datenbank abgelegt wird.

Als ich das Tracking-Schema fotografierte, arbeitete Earlybird gerade an einer zweiten Version der App. Dani, Junior Data Scientist bei Earlybird Digital, bemerkte eine Unstimmgkeit in der letzten Version der App: Wer einen Partner likt, likt automatisch auch alle Deals, die dieser Partner anbietet. In einem Konzeptpapier für ein Empfehlungssystem führt er weiter aus:

> Because the system does not track interaction with the deal, but with its partner, all the preferences expressed for a partner are applied to all its deals. This is unfortunate, because it is not clear that the user actually would have expressed also, e. g. a like for another deal of the same partner. […] The data collection in the new version of the app will track the preference for a deal and not for its partner.

Dieses Problem soll in der nächsten App-Version gelöst werden, so dass eine Deal-Like nicht mehr automatisch als Partner-Like interpretiert wird.[14]

Die formale »Bedeutung« von digitalen Verhaltensweisen ist in der Datenbank festgelegt. In Diskussionen um die Interpretation von Partner-Likes oder auch einfachen Likes zeigen sich aber auch bei Earlybird »interpretative Flexibilität« beziehungsweise eifrige Diskussionen darüber, was ein »Herzchen« bedeutet: Bedeutet ein Like für einen Partner, dass die Nutzerin alle seine Deals mag? Ist ein Like wirklich Ausdruck einer Präferenz oder eher eine Art Lesezeichen, um später etwas wieder aufzufinden? Diese Frage kann problemlos offen bleiben.

13 Zum Sieben als Metapher für eine Anthropologie der Algorithmen siehe Kockelman 2013; Maurer 2013.

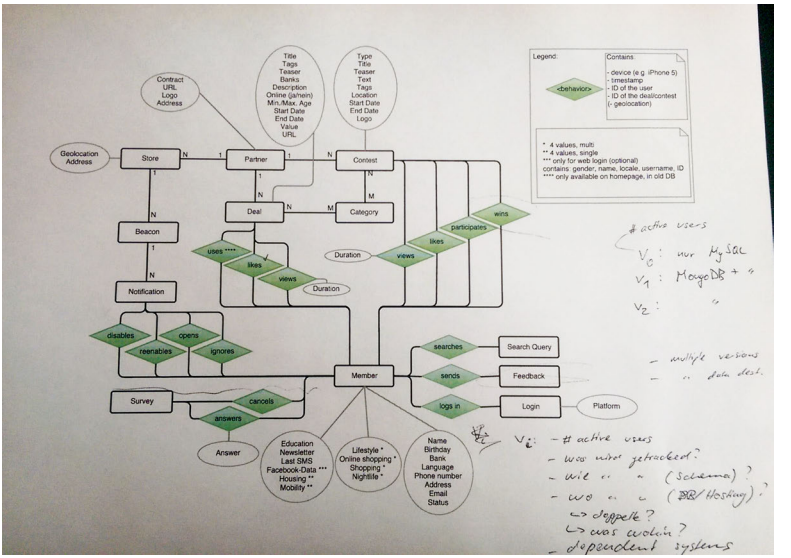

Abb. 4: Das Tracking-Schema der Earlybird-App

7.3 Vektorisierung

Nicht alle Verhaltensweisen von Nutzerinnen werden auf eine so strukturierte Weise registriert, wie Alaimo und Kallinikos (2017) es unter dem Begriff »encoding« beschreiben. Sie argumentieren, dass sich unstrukturiere Daten wie Texte, Bilder, Videos oder Audio grundlegend davon unterscheiden (siehe Kitchin 2014 im Allgemeinen und Bechmann & Bowker 2019; Buolamwini & Gebru 2018; Crawford & Paglen 2019 zu Bildern; zu Musik siehe weiter unten). Dieser »user generated content« (Beer & Burrows, 2007; Bruns, 2008; Ritzer & Jurgenson, 2010; Ekbia & Nardi, 2017) bildet die Kristallisationspunkte der Nutzerpartizipation in sozialen Netzwerken wie Facebook, Tumblr, Twitter, etc., ist aber nicht »encodiert«:

14 Christian Sandvig (2014) zeigt, wie Facebook diese Art von Uneindeutigkeit sogar zu nutzen weiß, um ein »like-recycling« zu betreiben. Ein Like für einen Beitrag einer bestimmten Quelle wie beispielsweise das Vice-Magazine wurde von Facebook als Like aller zukünftigen Beiträge von Vice interpretiert. Oder ein Like eines Kommentars zu einem Artikel erzeugte eine Relation zwischen Nutzerin und Artikel, obwohl der Like für den Kommentar abgegeben wurde.

> It is important, however, to distinguish between the content, say, of the uploading or posting (what users generate as content) and the very act of uploading or posting that content (social data). (Alaimo & Kallinikos, 2017, S. 177)

Während durch »encoding« die Verhaltensweisen der Nutzerinnen schon immer »maschinenlesbar« und so gesehen vergleichbar beziehungsweise kommensurabel sind, müssen unstrukturierte, von Nutzerinnen generierte Inhalte erst maschinenlesbar gemacht werden.[15] Wie im vielzitierten und auch in meinem Feld verwendeten Handbuch *Data Science for Business. What You Need to Know about Data Mining and Data-Analytic Thinking* von Provost & Fawcett (2013) sind gerade über Freitextfelder erfasste Textdaten fehleranfällig und »verschmutzt«. Deshalb müssen eine Vielzahl von datenbereinigenden Arbeitsschritten unternommen werden (zum Beispiel das Entfernen von irrelevanten »stopwords« oder »stemming«, d.h. die Reduktion der verschiedenen Wortformen auf ihre Stammform). Am Ende des Aufbereitungsprozesses stehen Daten in »a nice format, like something with columns: name | event | year | gender | event time« (Schutt & O'Neil, 2013, S. 41) (siehe zu »messy data« und zum Prozess der Datenaufbereitung auch: Mützel et al. 2018).

Das grundlegende Problem unstrukturierter Daten besteht darin, diese kontinuierlichen Phänomene diskret zu machen: Wie Gitelman & Jackson (2013) argumentieren, ist der Prozess der Imagination von Daten immer auch ein Kategorisierungsprozess, der aus einer amorphen Realität Formen definiert. Datafizierung heißt, sich die Welt aus Datenpunkten bestehend vorstellen zu können. Das setzt voraus, die Kontinuierlichkeit der Welt in diskontinuierliche Einheiten zu zerschneiden und Ordnungen zu konstruieren (Siehe auch Lury et al. 2012).

Die Datafizierung von Musik ist ein eindrückliches Beispiel, welches diese datafizierenden Ordnungsleistungen verdeutlicht, wenn Musikstücke in immer kleinere Einheiten zerlegt werden. Für das Unternehmen The Echo Nest besteht ein durchschnittliches Musikstück aus ungefähr 2000 »events« (Prey, 2016, S. 33). Unter anderem identifiziert The Echo Nest für jedes Lied in seiner Datenbank »musically relevant elements that occur sequenced in time« (Jehan & DesRoches, 2014, S. 2). Die im Alltag intuitiv verständliche

15 Das Beispiel des »livecoding« (Swift et al., 2014; McLean, 2017) aus der digitalen Kunst unterläuft diese Unterscheidung.

Einheit des Liedes wird aufgebrochen, um das Lied als Datenpunkte neu zu versammeln. Der Song *Never Gonna Give You Up* von Rick Astley hat zehn Sections, 397 Beats und 935 Segmente. Für jedes Segment wird Klangfarbe, Tonhöhe und Lautstärke ausgewiesen. The Echo Nest versteht Musikstücke als Daten: Jeder Song besteht aus einer bestimmten Anzahl und Arten von Events mit bestimmten Eigenschaften. Die Gesamtheit von Klängen, die ein Lied ausmachen, wird dabei in eine neue Ordnung gebracht. Daran anschließend lassen sich »musikalisch ähnliche« Lieder identifizieren und Nutzerinnen empfehlen. Lieder in der Datenbank lassen sich auch auf der Basis abgeleiteter Eigenschaften wie dem »danceability score« vergleichen – »the higher the value, the easier it is to dance to this song« (Lamere, [o.D.]b). Die Aufsplittung der Stücke ermöglicht aber auch Manipulationen wie z.B. eine automatisierte Neuanordnung der Elemente: »The Eternal Jukebox« macht aus endlichen Musikstücken nicht endende, indem ähnliche Segmente des Songs identifiziert und neu zusammengesetzt werden (Lamere, [o.D.]a). An den Übergängen zwischen den Segmenten »springt« der Song zu ursprünglich nicht vorgesehenen, aber ähnlichen, Stellen.[16]

Eine weit verbreitete Methode, um Texte – seien es Blogbeiträge, Statusmeldungen oder ganze Bücher – zu datafizieren, ist die Vektorisierung (Mackenzie, 2017; Rieder, 2020). Das Verfahren der Vektorisierung wurde im computerwissenschaftlichen Forschungsfeld des »information retrieval« entwickelt und maßgeblich von Gerard Salton et al. (1975) geprägt (Rieder, 2020, K. 5). Rieder beschreibt Vektorisierung als Methode, wie Texte in eine »intermediary form« gebracht werden können, um die statistische Verarbeitung zu ermöglichen. Vektorisierung bildet die Basis für viele Techniken

16 Das Beispiel der Musik weist auch auf den Unterschied zwischen Digitalisierung und Datafizierung hin. Spätestens seit dem Aufkommen der CD ist Musik digital. Von datafizierter Musik zu sprechen, wäre an diesem Punkt aber nicht angebracht. Musik als datenförmig zu verstehen und ihr bloßes Vorliegen in einem digitalen Format sind zwei unterschiedliche Dinge. Die Imagination von Dingen als Daten geht einher mit dem Wunsch bzw. der Notwendigkeit von Datenanalyse und Datenmanipulation. Dementsprechend wäre genauer zu untersuchen, ab wann von datafizierter Musik die Rede sein kann und wo deren Ursprünge liegen – zum Beispiel in der musikindustriellen Praxis des Masterings von Aufnahmen und der Manipulation von Tiefen und Höhen (Milner, 2019) oder der Erfindung von Kompressionsverfahren zur effizienteren Übermittlung von Telefongesprächen, wie Sterne (2012) in seiner Analyse des Audioformates MP3 darlegt.

des »machine learning« und des »natural language processing« (Mackenzie, 2017).

Vektorisierung, wie sie im Beispiel von Earlybird beschrieben ist (siehe unten), löst die subjektiven Bedeutungen der Wettbewerbsantworten auf und gibt ihnen eine neue Bedeutung. Der Kontext des Wettbewerbs, der Wettbewerbsfrage oder der subjektiven Wünsche und Hoffnungen weicht dem mathematischen Kontext des Vektorraums von Wikipedia. Die Bedeutung des Texts besteht so gesehen nicht in seiner subjektiven Interpretation durch die Urheberinnen oder jenen, die den Wettbewerb durchgeführt haben, sondern lässt sich nun mathematisch als Kombination von Vektoren ausdrücken. Die Bedeutung eines Worts liegt nicht mehr darin, was ich oder jemand anderes darunter versteht, sondern welche anderen Wörter sich innerhalb des aufgespannten Vektorraums in der Nähe befinden.

> [Vectorizing data] produces a common space that juxtaposes and mixes complex localized realities. [...] In vector space, identities and differences change in nature. Similarity and belonging no longer rely on resemblance or a common genesis but on measures of proximity or distance. (Mackenzie, 2017, S. 73)

Im Folgenden beschreibe ich, wie Earlybird Digital Wettbewerbsantworten in Vektoren transformiert und so die Grundlage schafft, um Member automatisiert ihren Marketingkategorien zuordnen zu können.

Rechnen mit Text

> Ich möchte am liebsten nach Jordanien und ganz früh am morgen mit dem Pferd durch die Wüste zur antiken Stadt Petra reiten, damit ich vor den Touristenbussen die Stadt im Sonnenaufgang bewundern kann. Auf den Cook Inseln schnorcheln mit Walhaien und einfach die Seele baumeln lassen. In den USA einmal die unglaubliche Atmosphäre am Burning Man Festival erleben.

So (ähnlich) lautet eine der vielen Wettbewerbsantworten, die bei Earlybird eingegangen sind. Simon sieht Freitexte wie diesen als »missing link« mit dem sich ein Problem von Earlybird beheben lässt. Er formuliert das Problem anhand zweier Thesen. Erstens: Im »Earlybird Universum« existieren fünf Member-Typen: Hedonisten, progressive Postmoderne, Traditionelle, Young Performer und Freestyle Actionsportler (sowie eine Rest-Kategorie).[17] Als Ma-

thematiker sei er bei solchen Dingen skeptisch. Er habe aber immer wieder mit Beni darüber gesprochen und musste irgendwann einmal sagen: »so ist wahrscheinlich die Welt«. Simon geht also davon aus, dass es tatsächlich diese verschiedenen Jugendmilieus gibt und dass sie unter den Earlybird-Membern »ein Stück weit« vertreten sind. Zweitens geht er davon aus, dass das Marketing für die einzelnen Typen Kampagnen entwickeln und durchführen könne. Der »missing link« zwischen den Typen und den Kampagnen sei, wie die Member den einzelnen Typen zugeordnet werden können, so dass sie das Marketing mit entsprechenden Kampagnen ansprechen kann.

Er präsentiert eine »verrückte Idee«, wie sich dieses Problem mit künstlicher Intelligenz lösen lasse. Ein Mitarbeiter von ihm sei gerade dabei, die deutsche Wikipedia herunterzuladen. Damit wollen sie »ein multilayer neuronales Netz« trainieren, das »den Kontext von Wörtern« lernt. Jedes Wort in der Wikipedia wird dafür in einen Vektor[18] transformiert, der im Prinzip so viele Dimensionen haben kann wie Wikipedia Artikel hat (also 2.5 Millionen Dimensionen), aber auf einige hundert reduziert wird. Jedes Wort auf Wikipedia wird dann abgebildet in diesem multidimensionen Vektorraum.

Um Wikipedia als Vektorraum zu beschreiben, wird eine Tabelle erstellt, die alle einmaligen (und lemmatisierten) Wörter und alle Artikel von Wikipedia umfasst (siehe Grafik 5). In die Felder der Tabelle wird dann eingetragen, in welchen Artikeln jedes Wort jeweils vorkommt.[19]

Ein Beispiel: Nehmen wir an, Wikipedia verfüge nur über zwei Artikel: einen Artikel über Soziologie und einen Artikel über Mathematik. Wir zählen nun, wie oft das Wort »Mensch« und das Wort »Rechnen« in beiden Artikeln vorkommt: »Mensch« erscheint 12 mal in Soziologie, 4 mal in Mathematik; »Rechnen« erscheint 1 mal in Soziologie, 5 mal in Mathematik. Diese Transformation erlaubt es, mit Wörtern und Texten zu rechnen: Ähnlichkeiten und Differenzen verschiedener Wörter, lassen sich nun quantitativ ausdrücken, indem beispielsweise die Distanz oder der Winkel zwischen den Wörtern gemessen wird. Auch ganze Texte lassen sich als »bag of words« im Vektorraum lokalisieren und mit anderen Wörtern oder Texten quantitativ in Beziehung setzen.

Die Zahlen im Vektor seien »eine Art Codierung« darüber, in welchem Kontext zum Beispiel das Wort »Mensch« über alle Seiten der Wikipedia verwendet wird. Dies erhalten sie für jedes Wort, das auf Wikipedia verwendet wird. Das ist das Resultat des neuronalen Netzes: ein »word2vec«-Modell. Si-

mon kann nun für jedes Wort in einer Wettbewerbsantwort den Vektor auslesen und addieren. Dieser Vektor entspreche dann dem »Kontext in Bezug auf Wikipedia, wo es gelernt wurde«.

In einem nächsten Schritt werden die Beschreibungen der Lifestyle-Segmente ebenfalls in Vektoren transformiert, zum Beispiel zeichnet sich das Segment der »Hedonisten« durch »Unterhaltung«, »Musik«, »Tanzen«, »Club«, etc. aus. Jedes in der Beschreibung vorkommende Wort wird zu einem Vektor. Alle zusammen werden zu einem »Centroiden« addiert, der für das entsprechende Segment steht. Dasselbe macht Earlybird mit jedem Segment und jeder Wettbewerbsantwort. Daraus ergibt sich ein Vektor für »Hedonist« und einer für die Wettbewerbsantwort eines Members (siehe Grafik 6). Das ermögliche nun ein »algorithmisches Mapping zwischen Membern und den Marketing-Persona«. Das könne man dann dem Marketing übergeben, um Kampagnen damit zu machen.

Encoding und Vektorisierung sind zwei Varianten, mit denen Nutzerinnen und ihre Verhaltensweisen vergleichbar gemacht werden. Die Soziologie der Quantifizierung verwendet den Begriff der »commensuration«, um zu benennen, wie qualitative in numerische Differenz verwandelt wird:

17 Bei einem Milieu wisse man nicht wohin damit: das sei der Abfalleimer. Dort gebe es im Text nichts Charakteristisches wie bei den anderen. Sie könnten damit nicht anfangen, weil es kein Wort gebe, das diese Gruppe beschreibe. »Keine Zuordnung möglich« bedeute, dass es in diese Kategorie komme. Wenn der Vektor des User-Inputs relativ weit weg von allem anderen sei, dann gehöre es in diese Kategorie. Simon nimmt aber an, dass sie auch keine Kampagnen für solche Leute designen. Sabina meint, dass diejenigen so verschieden seien, dass einzelne Gruppen darin wieder spezifisch angesprochen werden müssten, z.B. »Straight-Edge«, eine Subkultur, die gar nichts konsumieren wolle.

18 Ein Vektor ist ein mathematisches/geometrisches Konzept. Ein Vektor hat eine Länge und eine Richtung. In einem zweidimensionalen Koordinatensystem – typischerweise als gerader Pfeil abgebildet – beginnt ein Vektor beispielsweise am Nullpunkt (0, 0) und geht zum Punkt (2, 5) (Rieder, 2020, S. 217).

19 Hier gibt es mehrere Möglichkeiten: Salton et al. (1975) zählen, wie oft ein Wort in den jeweiligen Dokumenten vorkommt. Simon und sein Team verwenden den TF-IDF Algorithmus – ein Mass dafür, wie relevant ein Wort in einem Text ist (Gabrilovich & Markovitch, 2007).

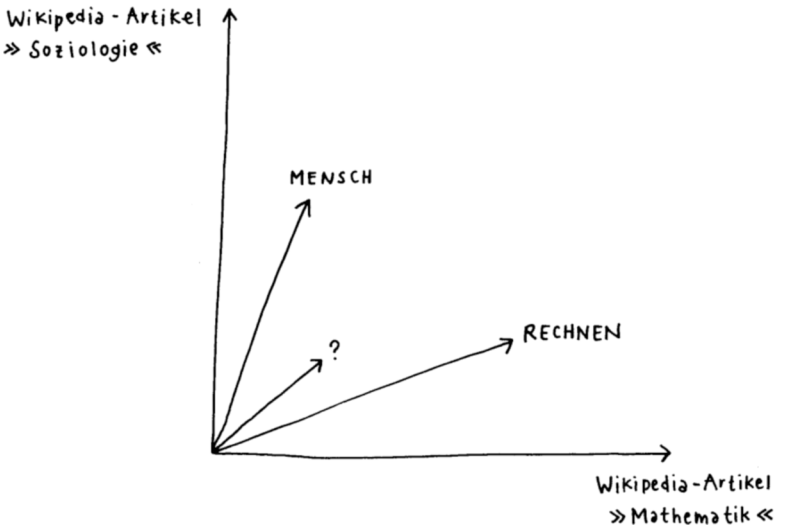

Abb. 5: Beispiel von zwei Vektoren im zweidimensionalen Raum

> Commensuration creates a specific type of relationship among objects. It transforms all difference into quantity. In doing so it unites objects by encompassing them under a shared cognitive system. At the same time, it also distinguishes objects by assigning to each one a precise amount of something that is measurably different from, or equal to, all others. Difference or similarity is expressed as magnitude, as an interval on a metric, a precise matter of more or less. (Espeland & Stevens, 2008, S. 408)

Beide Verfahren reinigen die unterschiedlichen Verhaltensweisen oder Äußerungen von ihren qualitativen Kontextbezügen und subjektiven Sinngehalten. Zahlen oder Daten weisen daher eine geringe »Indexikalität« auf (Heintz, 2010, S. 173). Heintz weist darauf hin, dass dieses »disembedding« die »Anschlussfähigkeit« in kulturell heterogenen Kontexten erleichtert.[20] Die großangelegte ethnografisch-vergleichende Studie »Why We Post«

20 »Um festzustellen, dass Norwegen auf der HDI-Rangliste einen höheren Rang einnimmt als Mexiko und Mexiko einen höheren als Sierra Leone, muss man die Konstruktion des Index kennen, braucht aber nicht zu wissen, wie die Verhältnisse in den Ländern im Einzelnen beschaffen sind. Insofern stellen numerische Darstellungen eine enorme Abstraktions- und Selektionsleistung dar, die die Verständigung vor allem

macht darauf aufmerksam, dass die gleichen Kommunikationsinfrastrukturen (i.e Facebook, Twitter, Whatsapp etc.) in unterschiedlichen kulturellen Kontexten auf unterschiedlichste Weisen verwendet werden (Miller, 2016). Das durch »encoding« etablierte Datenmodell ist für diese Unterschiede aber blind – und braucht davon auch gar nichts zu wissen.

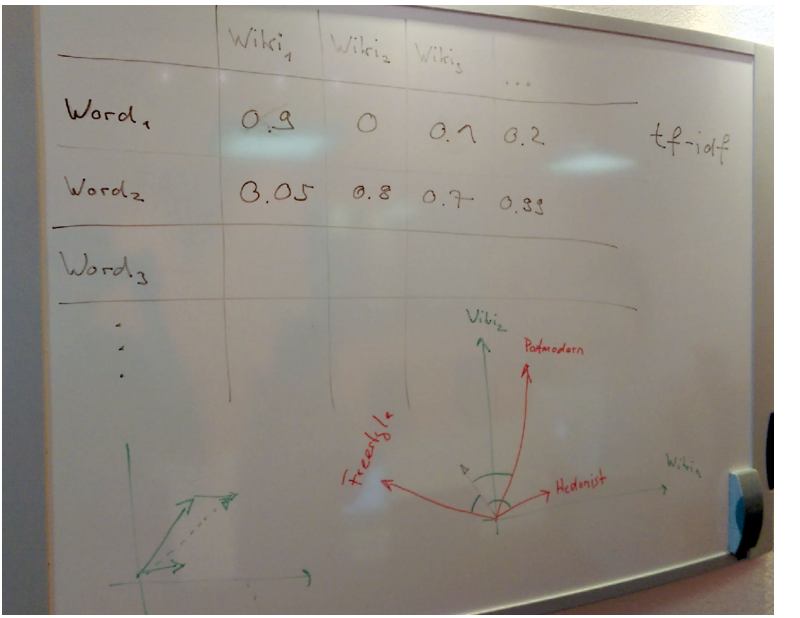

Abb. 6: Simons Darstellung des Wikipedia-Vektorraums

Das Encoding dessen, was Couldry & Mejias (2019a) »life itself« nennen, filtert ü berschüssige B edeutungen, d ie m it H andlungen i m digitalen Alltag verbunden sind. Wie Christina Alaimo und Jannis Kallinikos (Alaimo & Kallinikos, 2019) für Social-Media-Plattformen zeigen, ist die Produktion von Daten »a delicate engineering accomplishment«, das die Handlungen der Nutzerinnen von bedeutungstragenden Kontexten, in welchen diese Handlungen durchgeführt werden, »befreit« und zu digitalen Inskriptionen macht:

---

dann erleichtert, wenn kein gemeinsames kulturelles Hintergrundwissen vorausgesetzt werden kann« (Heintz, 2010, S. 173).

> to dissociate the action users perform from the meaningful contexts in which these actions occur, and treat them as just digital inscriptions, data tokens possible to cross-reference or syndicate, aggregate and combine with other data tokens. (Alaimo & Kallinikos, 2019, S. 304)

»[D]isregarding the ›underlying‹ object« (Charitsis et al., 2018, S. 827) – i.e. die Nutzerin, ihre Wettbewerbsantworten – ist die Bedingung, um aus Daten Wert zu generieren. Sobald encodiert oder vektorisiert wird, geht es nicht mehr darum, wie »people are related in actual life processes«, sondern um ihre Relationen in abstrakten »data spaces« (Arvidsson, 2016, S. 9). Diese Dekontextualisierung öffnet datafizierte Verhaltensweisen für die weitere Verarbeitung und potenzielle, bisher ungeahnte, Verwendungszwecke. Daten und Zahlen sind also nicht nur in kulturell diversen Kontexten anschlussfähig, sondern werden erst dadurch produktiv: Indem sie dekontextualisieren schaffen sie die Möglichkeit, Daten neuen Zwecken zukommen zu lassen und für weitere Verarbeitungsschritte zu öffnen. Verhaltensweisen werden zu »data tokens«, die rekombiniert, aggregiert und an Dritte weitergegeben werden können (Alaimo & Kallinikos, 2019). Solche »Rohdaten« bilden die Grundlage für verschiedene weitere Operationen der Bewertung, der Kategorisierung und des Vergleichs (bei Alaimo & Kallinikos 2017 unspezifisch »computation« genannt). Dabei werden Nutzerinnen oder Gruppen von Nutzerinnen als »audiences« konstruiert und mit neuen, prädiktiven Relationen ausgestattet (Charitsis et al., 2018).

Die Löschung des Kontexts, der die Entstehung der einzelnen Datenpunkte umgibt, ist also nicht als bedauernswerter Umstand oder Fehler zu verstehen (Seaver, 2015). Encoding und Vektorisierung sind Verfahren, die aus Personen und ihren Tätigkeiten Nutzerinnen und Objekte machen, die sich für weitere Operationen der Kategorisierung, der Bewertung und des Vergleichs anbieten: Es ist ein notwendiger Schritt im Prozess, aus »life itself« eine veräußerbare und verarbeitbare Ressource zu machen, die weiteren Transformationsschritten und Verarbeitungsprozessen offen steht. »[T]aking the gift out of the commodity« (Tsing, 2013, S. 21) heißt hier, den sozialen Kontext und die subjektive Bedeutung einer Handlung auszuklammern und nur ihren technischen Kontext in Betracht zu ziehen, so dass zwischen den verschiedenen datafizierten Verhaltensweisen und Nutzerinnen Vergleichbarkeit und neue Relationen hergestellt werden können.

Die sozialen Kontexte und subjektiven Bedeutungen werden durch Encoding und Vektorisierung zwar ignoriert. Seltsamerweise erlaubt gerade diese

Nivellierung datafizierter Verhaltensweisen, dass in einem weiteren Moment der Datafizierung neue Relationen wuchern können. Wie kluge Data Scientists und leistungsfähige Algorithmen solche latenten, vermeintlich immer schon in den Daten steckenden Relationen hervorlocken beziehungsweise produzieren, ist Gegenstand der Kapitel 8 und 10.