

HUMBOLDT-UNIVERSITÄT ZU BERLIN

INSTITUT FÜR PHILOSOPHIE



Proseminar: Das Netz und die Digitale Revolution

Leitung: Prof. Dr. Olaf Müller

Modul: Theoretische Philosophie

Sommersemester 2015

Das Datenmonopol

ÖKONOMISCHE TENDENZEN VON BIG DATA

Christopher Olk

Weichselstraße 57, 12045 Berlin

olkchris@hu-berlin.de, Matrikelnummer 567900

Bachelor of Arts Philosophie und Volkswirtschaftslehre, 2. Semester

Inhalt

Einleitung	2
I. DATEN ZU WISSEN	
1. Wie Big wird Big Data?	4
2. Der linguistic turn 2.0	5
3. Korrelationen – Das Ende der Theorie?	7
4. Skepsis und „Daterminismus“?	8
5. Die exponentielle Wertsteigerung von Daten	8
5.1 Daten machen einander wertvoller	9
5.2 Ein Metadatum macht alle Daten wertvoller	10
II. DATEN ZU DATEN	
1. Ein natürliches Monopol auf Daten	12
1.2 Der Reinvestment-Effekt	12
1.3 Der Übernahme-Effekt	13
2. Einwände und gegenläufige Effekte	16
3.1 Das „geniale Startup“	17
3.2 Die „Chance der Alten“	17
3.3 Staatliche Eingriffe	18
3. Der Unterschied zwischen Daten und Kapital	19
III. WISSEN ZU GELD	
1. Silicon Vall Street	21
2. Geduld zahlt sich aus	22
3. Monetarisierung	23
IV. DATEN ZU ALLEN	
1. Datenschutz?	26
2. Free the Data – eine nützliche Utopie	28
Fazit	30
Literatur	32

Ich werde von meinem iPhone geweckt, dessen GPS-Modul auf dem Weg in die Bibliothek unablässig sendet. 90% der Recherche für diesen Text erfolgt via Google, während Facebook im Tab daneben geöffnet ist und ich über Spotify Musik höre. Gegen Mittag suche ich mit Maps etwas zu essen, währenddessen schreibe ich in WhatsApp und Tinder Nachrichten. Gerade sagt mein Windows-Kalender mir, dass meine Mutter nächste Woche Geburtstag hat, ich bestelle ihr auf Amazon ein Buch.

Einleitung

Big Data bezeichnet nicht nur eine methodische Revolution der Natur- und Gesellschaftswissenschaften. Auch eine umfassende philosophische Reflektion des Phänomens setzt ein. So beschreibt Melanie Swan eine ontologische, epistemische und ethische „Philosophy of Big Data“, die vor allem drei Ziele verfolgt: Neue Methoden auf ein wissenschaftstheoretisches Fundament zu stellen, gesellschaftliche Veränderungen zu analysieren und schließlich eine gute und gerechte „Mensch-Daten-Beziehung“ zu entwerfen (Swan 2014).

Einzig die politische Ökonomik scheint das Feld weitgehend den Big Data-Akteuren selbst zu überlassen. Namen wie Eric Schmidt, Larry Page oder Jeff Bezos dominieren weniger die Feuilletons als die Wirtschaftskolumnen (und Bestsellerlisten) auf beiden Seiten des Atlantiks. Am Pazifik hingegen wird weniger eine offene Debatte über die ökonomischen Implikationen der neuen Technologien geführt als vielmehr schlicht Geld verdient. Wo immer man sich sonst gerne als Avantgarde der Wirtschaftswissenschaft versteht, scheint es an einer öffentlich entwickelten Ökonomik der Daten zu mangeln, die diesen Namen wirklich verdient¹.

Die Informationsökonomik untersucht ganz generell die Bedeutung von (unvollständigen) Informationen für wirtschaftliche Entscheidungsprozesse. Die Medienökonomik kennt eine Reihe von Netzwerkeffekten, aber bisher ohne einen Transfer zu Big Data zu leisten. Ein weiterer Ansatz nimmt vor allem die Technologie hinter der massenhaften Speicherung und Verarbeitung von Daten in den Fokus. Schon 2001 lieferte Hal Varian, der spätere Chefökonom von Google, mit seiner Vorlesung „Economics of Information Technology“ das Standardwerk zu einem Ökonomikzweig, dessen *raison d’*

¹ Auf einem wissenschaftlich fruchtbaren Boden außerhalb Kaliforniens scheint höchstens das Media Lab des MIT unter Alex Pentland zu stehen – allerdings ein Informatiker und Soziologe, kein Ökonom.

être der 451research-Analyst Matt Aslett so formuliert: „We believe [what matters] is not the nature of the data but the economics of storing and processing the data“. (Aslett 2015)

Ich halte es ganz im Gegenteil für essentiell, nach genau dieser „Natur der Daten“ zu fragen und nach ökonomischen Prinzipien, denen große Informationsmengen tendenziell folgen, zu suchen. Eine erste exemplarische Idee, die bewusst etwas naive Skizze eines einzigen solchen Prinzips - das soll diese Seminararbeit bieten.

In einem ersten Schritt werde ich zwei Prämissen aufstellen. Erstens: Für zukünftige Analysetechniken werden fast alle Daten verwertbar sein. Zweitens: Bei linearer Vergrößerung eines Daten-Inputs vergrößert sich der Informations-Output exponentiell. Daraus werde ich im nächsten Schritt folgern: Es gibt auf freien Märkten eine natürliche Tendenz zur Entstehung eines Monopols auf große Datenmengen. Im letzten Schritt werde ich mögliche Konsequenzen aus dieser These diskutieren und einen wirtschafts- und gesellschaftspolitischen Ausblick wagen.

Nun ist die Idee, dass etwa Google ein natürliches Monopol darstelle, fast so alt wie das Unternehmen selbst. Zwar beschränken sich sowohl das US-Verfahren von 2013 als auch die aktuelle EU-Untersuchung dabei auf die Frage, ob Google seine Marktmacht auch tatsächlich missbrauche. Es wird aber auch grundsätzlich diskutiert, inwieweit Google und auch Facebook natürliche Monopole darstellen, also ein Marktversagen herbeigeführt haben (in Deutschland vertritt etwa FAZ-Redakteur Patrick Bernau genau diese Ansicht). Zwar senkt das Internet einerseits Transaktionskosten und fördert damit den Wettbewerb (vgl. Haucap et al. 2013). Andererseits ist aber spätestens seit dem Aufstieg von Microsoft eigentlich unstrittig, dass der IT-Markt generell einer natürlichen Tendenz zur Monopolbildung unterliegt. Die zwei wesentlichen Ursachen dafür sind, zusammengefasst, Skaleneffekte (hohe Fix- und niedrige Grenzkosten wie bei Google) sowie Netzwerkeffekte (jeder neue User² vergrößert den Nutzen aller anderen User wie bei Facebook). Zwar werde ich insbesondere auf den zweiten Mechanismus noch zurückkommen, eigentlich will ich aber noch einen Schritt weiter gehen. Es geht mir weder um die kontingente Frage, ob einer der Internetgiganten tatsächlich ein Monopol innehat, noch will ich die Formeln der klassischen Medienökonomik weiter ausreizen. Vielmehr sollen die gegenwärtigen Praktiken von Google, Facebook und vielen anderen als Ausgangspunkt dienen, um zukünftige Techniken der Datenakquise und -analyse zu skizzieren und auf allgemeine „datenökonomische“ Prinzipien hin zu untersuchen.

² Als Kompromiss zwischen gendergerechter Sprache und Lesbarkeit werde ich in der vorliegenden Arbeit personifizierende Substantivierungen männlich, unbestimmte Personen/ Berufsbezeichnungen etc. im Singular abwechselnd männlich und weiblich, im Plural jedoch nur männlich benennen.

I. DATEN ZU WISSEN

Wie Big wird Big Data?

Die Menge der digital gespeicherten Daten wächst jedes Jahr um 40% (IDC 2015). So verdoppelt sie sich ungefähr alle zwei Jahre. Vielleicht ist diese Wachstumsrate sehr viel wesentlicher für das Verständnis des Phänomens Big Data als die absoluten Zahlen (2013: 4,4 Trillionen Gigabyte, 2020 werden es wohl zehnmal so viele sein). 85% dieser Daten - Tendenz steigend - sind im Besitz von Unternehmen, deren vorrangiges Ziel die eigene Gewinnmaximierung ist. Sie ordnen, analysieren und interpretieren diese riesige Daten-„Maische“, um daraus erst im nächsten Schritt echtes Wissen zu „destillieren“, das letztlich in Gewinne verwandelt werden muss. Die zwei limitierenden Faktoren sind dabei der verfügbare Speicherplatz und die Fähigkeit, Wissen aus den Daten zu gewinnen. Die Kosten für Speicherplatz sinken seit Jahrzehnten exponentiell³, dieser Trend wird sich höchstwahrscheinlich noch einige Zeit fortsetzen (vgl. Klausnitzer 2013). Die Verarbeitung der riesigen Datenmengen, die Gewinnung von „Smart Data aus Big Data“ (ebd.) ist gegenwärtig der weitaus kritischere Faktor.

Melanie Swan zählt *volume*, *velocity*, *variety*, *veracity*, und *variability* zu den definierenden Attributen von Big Data (Swan 2014). Zweifellos stellt dabei immer öfter die *variety*, die Verschiedenartigkeit von Daten, die zentrale Herausforderung für ihre Verwertung dar. Tatsächlich ist sie so groß, dass 2012 gerade einmal 3% aller digitalen Daten geordnet („getagged“) und nur 0,5% analysiert, also überhaupt nutzbar gemacht werden konnten (Klausnitzer 2013).⁴ Der erste Schritt ist nämlich immer das Ordnen und „Übersetzen“ verschiedenartiger Daten in eine gemeinsame „Sprache“ (Dafür sind *Metadaten* der entscheidende Faktor – dazu später mehr.) Erst danach können überhaupt Fragen gestellt, Vergleiche vorgenommen oder Muster erkannt werden. Dabei findet dann eine Vielzahl von Techniken Anwendung - von klassischer Statistik bis zu komplexen Mustererkennungs-Algorithmien. Software und auch neue Hardware-Ansätze werden von Unternehmen und auch von nichtkommerziellen Gemeinschaften wie etwa der OpenSource-Community ständig weiterentwickelt. Für eine

³Grafik: <https://intelligence.org/wp-content/uploads/2014/04/Hard-drive-storage-cost.png>

⁴ eine mögliche Erklärung dafür, dass hier nicht mehr investiert wird, in Teil III der Arbeit

ökonomische Prognose bedeutet das: Auch ohne übermäßigen Optimismus lässt sich annehmen, dass die Fähigkeiten zur Datenverarbeitung noch einige Zeit lang annähernd linear wachsen (und die Kosten dafür entsprechend sinken) werden.

Gleichzeitig wird, parallel zum Speicherplatz, die *volume* der gesammelten Daten weiterhin stark exponentiell steigen. Eine der ältesten Fragen in der Debatte um Big Data ist, ob *bigger* tatsächlich immer *better* ist. Überzeugendste Antwort: Es kommt darauf an. Erstens eben auf technische Beschränkungen, zweitens darauf, wonach (und wie konkret) man einen Datensatz fragt (vgl. Bollier 2010), denn davon hängt letztlich seine Qualität für den Fragenden ab.

Wonach werden wir also fragen? Welche Gestalt werden die wichtigsten Daten der mittelfristigen Zukunft annehmen? Die Slogans *reality mining* (Pentland 2008) oder *Internet der Dinge* (Weiser 1991) deuten es schon seit Längerem an, ich wage eine Spekulation: Nach den Geo- und Bewegungsdaten, die etwa dank *Maps* und *Streetview* gegenwärtig den wohl größten Hype erleben⁵, und den gerade erst aufkommenden Biodaten, etwa aus „Quantified-Self“-Apps oder vernetzter Kleidung, könnte die unerschöpfliche Menge an Text- und Bilddaten aus sozialen Netzwerken die nächste Generation von Big Data einläuten und bisher unvorstellbare soziologische, psychologische und verhaltensökonomische Erkenntnisse liefern.⁶

Der *linguistic turn* 2.0

„Wenn der Status eines Nutzers mehr positive als negative Wörter enthält, erhöht sich die Zahl der positiven Wörter in seinem Freundeskreis um 7%“ (Moorstedt 2013) – Facebook-Erkenntnisse wie diese sind in ihrer Banalität symptomatisch für die gegenwärtigen Analysefähigkeiten der Sozialwissenschaften⁷. Doch deren Wandel zu *Digital Humanities* hat längst eingesetzt. Die Naturwissenschaften haben nicht länger das methodische Monopol auf die Akquise und Analyse großer Datenmengen, die „digitale Vermessung der Kultur“ (Lauer 2013) hat begonnen – so bietet beispielsweise das prominente Google-Projekt *Books* nie dagewesene Chancen für die vergleichende Literaturwissenschaft. Nur ein zentraler Graben zwischen den traditionell mathematisch agierenden Natur- und den philologisch geprägten Geisteswissenschaften, so scheint es, muss noch überwunden

⁵ Schon heute lässt sich die Position eines Smartphone-Trägers zu einem beliebigen Zeitpunkt mit einer Wahrscheinlichkeit von durchschnittlich 90%, mindestens aber 80% vorhersagen (Barabasi 2011).

⁶ Übrigens lassen sich diese drei exemplarischen Datenarten sehr anschaulich aufeinander beziehen. Beispiel: Wo hält sich eine Person auf und wie hoch ist ihr Blutdruck, während sie schreibt „Ehrlich, Schatz, ich war die ganze Nacht im Büro“?

⁷ Darunter insbesondere die Verhaltensökonomie mit ihrem immensen ökonomischen Potenzial

werden: Die Quantifizierung von Sprache, die Algorithmen einen Text erst zugänglich macht, ist zugleich die neueste Herausforderung und eine alte Problemstellung. Die Frage, um die schon Wittgensteins Sprachphilosophie, spätestens aber die analytische und logische Philosophie seit dem *linguistic turn* kreisen, ließe sich womöglich so aktualisieren: Könnte Sprache auf binäre, durch Computer analysierbare Information reduziert werden? Dabei geht es mir weniger um die Entwicklung sprachbasierter Benutzeroberflächen als vielmehr um die Möglichkeiten, (synthetische, nicht-triviale) Erkenntnisse aus gespeicherten Texten und Tonaufnahmen zu gewinnen.

Ohne mich in die Tiefen der sprachphilosophischen Diskussion stürzen zu wollen, halte ich es für sehr wahrscheinlich, dass die sprach- und textbasierte Datenverarbeitung mittelfristig immense Fortschritte machen und eine bedeutende Rolle in allen Bereichen der Geisteswissenschaft und Datenwirtschaft übernehmen wird. Vorläufer dieser Entwicklung sind schon heute Programme wie das Natural Language Toolkit (vgl. nltk.org) oder auch die Programmiersprache *Python* - vor allem müssen ja aber alle Anbieter von (kostenlosen, also datensammelnden) sozialen Netzwerken, Mail- und Messengerdiensten fest daran glauben, dass es in absehbarer Zeit möglich ist, den allergrößten, aus Texten bestehenden Teil ihres Datenschatzes überhaupt zu heben (dazu mehr in III).

Die linguistische Analyse wird durch Tools wie Gesichts- und Stimmerkennung sowie die Verarbeitung präziser Geo- und Biodaten ergänzt werden. Mit über fünf Milliarden Smartphone-Nutzern (Pentland 2012) wird schon heute ein beträchtlicher Teil der menschlichen Kommunikation digital gespeichert – frei nach Goethe: „Der Worte sind genug gewechselt, lasst mich auch endlich Daten sehn!“. Diese Masse an sozialen Daten wird ein neuartiges Forschungsfeld ermöglichen, das sich schon jetzt als das Erkennen von „human behaviour modeling patterns“ (ebd.) skizzieren lässt. In einem visionären Beitrag für die FAZ erläutern Thomas Hofmann (ETH Zürich/Google) und Bernhard Schölkopf (Max-Planck-Institut für Intelligente Systeme): „Komplexe Modelle werden automatisch aus Daten gelernt, ihr Ziel ist nicht mehr die Verständlichkeit, sondern nur noch die Tauglichkeit für Voraussagen. (...) Sie versuchen, komplexe Prozesse wie die Handlungen von Menschen zwar nicht verständlich, aber doch statistisch vorhersagbar zu machen.“ (Hofmann et al. 2015)

Ganz unabhängig von geistesphilosophischen Fragen wie der nach mentaler Verursachung oder freiem Willen zeichnet sich ab, dass immer mehr unbewusste und chaotisch erscheinende Verhaltensmuster in Wahrheit Regeln oder Muster erkennen lassen⁸ – und solche Zusammenhänge werden einfach umso besser erkennbar, je mehr Daten es dazu gibt.

⁸ Vgl. hierzu etwa Barabási 2011

Korrelationen – das Ende der Theorie?

Dabei durchforsten Wissenschaftler zunehmend Datensätze, ohne überhaupt eine konkrete Ausgangsfrage zu stellen. Und sie finden immer öfter Korrelationen ohne erkennbare Kausalität. Das bewog Wired-Chefredakteur Chris Andersen schon 2009 dazu, das „Ende der Theorie“ zu verkünden. Seine These: Wer genug Daten betrachten könne, brauche keine Theorien, er müsse noch nicht einmal Kausalzusammenhänge erklären können. Die Korrelationen reichten aus, um Erkenntnisse zu gewinnen und auch Vorhersagen treffen zu können. (vgl. Andersen 2009). Er wurde von einer Welle der Empörung getroffen, letztlich hatte er aber nur eine uralte epistemologische Diskussion um Induktion und Deduktion wieder angefacht. Für eine vorrangig ökonomische Betrachtung sind allerdings zwei ganz andere Aspekte relevant. Zum einen ein wichtiger Punkt, den Hal Varian in Erinnerung ruft: „Der Königsweg, um Kausalität herauszufinden, sind Experimente“ (Varian 2013). Und wenn Akteure wie Google oder Facebook etwas nach Belieben durchführen können, dann eine Vielzahl von sozialen Experimenten.

Zum anderen zeigt sich die wissenschaftliche Praxis ohnehin unbeeindruckt von dieser Diskussion und forscht, oft interdisziplinär, an korrelationsbasierten Vorhersagen. So erklärt Gary King: “Wenn wir genug Daten haben, können wir (...) selbst dann Modelle entwickeln, wenn wir die entsprechenden Gesetze gar nicht kennen.“ (Bollier 2010) - ein Beispiel: Wer Temperatur, Luftdruck und Feuchtigkeit an jedem Ort der Erde kennt, kann perfekte Wettervorhersagen treffen, ohne je etwas von Strömungs- oder Strahlungslehre gehört zu haben. Weit interessanter als Wettervorhersagen sind jedoch, insbesondere für Unternehmen in Dienstleistungsgesellschaften, Vorhersagen über menschliches Verhalten. Es spricht nichts dagegen, dass solche anhand von Korrelationen aus den riesigen Mengen der Kommunikationsdaten getroffen werden können. Und das wesentliche Merkmal solcher Forschungen wird sein, dass viel mehr verschiedenartige Daten von Bedeutung sein werden als für jedes andere mithilfe von Big Data erforschte Phänomen überhaupt – ökonomisch formuliert: Auf dem wichtigsten Datenmarkt der Zukunft werden so gut wie alle Daten handelbar sein. Außerdem ist nichts weiter nötig als eine möglichst große Menge an Daten. Ignoriert man limitierende Faktoren wie Speicherplatz und begrenzte Analysefähigkeiten, gibt es keinen Grund, warum *bigger* nicht auch *better* sein sollte.⁹

⁹ Das wohl prominenteste Beispiel für eine simple Big Data-Anwendung stammt aus *Moneyball* (der Film von 2011 basiert auf *Moneyball. The Art of Winning an Unfair Game* von Michael Lewis): Ein Baseball-Scout der alten Schule lehnt einen Spieler mit den Worten ab: „Er hat eine hässliche Freundin. (...) Hässliche Freundin bedeutet: Kein Selbstvertrauen!“ (Lewis 2008).

Skepsis und: „Determinismus“?

Je tiefere Blicke man gerade in sensible Phänomene wie menschliches Verhalten wirft, desto eher droht man an eine Grenze zu stoßen, die von einer philosophischen Debatte um die Vorhersagbarkeit der Welt bewacht wird – was im Übrigen nicht mit dem Determinismus gleichzusetzen ist: „The world could be highly predictable, in some senses, and yet not deterministic; and it could be deterministic yet highly unpredictable, as many studies of chaos (...) show.“ (Hoefer 2015) Auch wenn anti-deterministische Philosophen bei weitem nicht die einzigen Skeptiker der Big Data-Euphorie sind - solange nichts notwendigerweise und prinzipiell dagegen spricht, bin ich geneigt im Folgenden einfach anzunehmen, dass die Entwicklung neuer, insbesondere linguistischer Technologien und eine noch unvorstellbaren Zunahme der Menge an verfügbaren Daten früher oder später alle Zweifel ausräumen wird.

Die exponentielle Wertsteigerung von Daten

Ich treffe keine Aussage über die Geschwindigkeit dieser Entwicklung, will aber drei Trends festhalten:

1. Die Kosten für Speicherplatz sinken weiterhin stark exponentiell (in gleichem Maße steigt die Menge der gespeicherten Daten).
2. Die Fähigkeit zur Verarbeitung komplexer Datensätze, ausgedrückt als der durchschnittliche Anteil von nutzbaren Daten an einer beliebigen Datenmenge, wächst annähernd linear.¹⁰
3. Verschiedenartigste Daten werden für die Vorhersage von Verhalten nützlich sein. Dies ist die erste Prämisse meiner Kernthese.

Meine zweite und wichtigere Prämisse ist im Grunde nichts anderes als das zentrale Prinzip, nach dem Lernen, Folgern und Forschen funktionieren: Das exponentielle Wachstum von Wissen. Neue Information wird mit alter Information verknüpft und schafft so einen weiteren Informationszuwachs.

Die oft zitierte Moral des Films: Die systematische Analyse von historischen Spielerdaten ist effektiver als das Bauchgefühl des Scouts. Um in diesem Bild zu bleiben: Ich stelle mir eine nächste Generation von Big Data vor, die in der Lage sein wird, das Selbstvertrauen von Spielern wie auch die Schönheit ihrer Partnerinnen anhand von Social-Media-Daten zu quantifizieren und auf Korrelationen zu überprüfen, um die bestehenden Leistungsdaten der Spieler subtil zu ergänzen. Erst dann wird der Scout wirklich ersetzt werden können.

¹⁰ Der wohl am schwierigsten vorherzusagende Trend. Später werde ich der Einfachheit halber eine konstante Rate von Datenverarbeitung annehmen, was allerdings tatsächlich einer Entwicklung im exponentiellen Tempo des Datenmengenwachstums entspräche – dieses wird aber sicherlich nicht ewig gehalten werden können.

Dieser ganz alltägliche Netzwerkeffekt kann, übertragen auf Daten, als schwache Tendenz auftreten, aber auch sehr starke Ausprägungen annehmen. Es lassen sich dabei noch einmal zwei Strukturen unterscheiden, die sich auf folgende Formeln bringen lassen:

1. Daten machen einander wertvoller

Das Prinzip lässt sich gut anhand des Bücherkaufs aus dem Intro verdeutlichen. Angenommen, Amazon liegen folgende (bereits sehr weit verarbeitete) Daten vor:

- X hat das Buch „Mehr Licht“ gekauft
- Y hat zuletzt die Bücher „Zur Farbenlehre“ und „Goethe contra Newton“ gekauft
- Y hat nächste Woche Geburtstag

Angenommen, Amazon erhält Zugriff auf ein einfaches Datum (etwa aus Facebook):

- Y ist die Mutter von X

Nun kann Amazon schließen, dass X mit einer sehr hohen Wahrscheinlichkeit das Buch „Mehr Licht“ als Geschenk für seine Mutter gekauft hat¹¹. Der Vorgang lässt sich so interpretieren: Das vierte Datum hat die anderen drei Daten erst in einen sinnvollen Zusammenhang gebracht, also wertvoller gemacht. Ökonomisch ausgedrückt: Ein Input von 3 Informationseinheiten ergab einen Output von 3 Einheiten, ein Input von 4 Einheiten ergab hingegen einen Output von 5 Einheiten. Ein anderes Beispiel:

Input:

- Z geht von ihrer Wohnung zur U-Bahn-Station „Hermannplatz“ (Smartphone-GPS-Daten)
- Z sagt „Ich bin im *Milch und Zucker*, Oranienstraße“ (Whatsapp-Sprachnachrichtendaten)
- Z hat außerdem noch Geld am Kottbusser Tor abgehoben (Überwachungskamera-Daten)

Zusätzlicher Output durch den dritten Inputfaktor:

- Z ist vom Hermannplatz mit der U8 zum Kottbusser Tor gefahren
- Z ist anschließend von dort zum Café *Milch und Zucker* gegangen

¹¹ Jetzt wird z.B. der Kaufvorschlag für „The Rediscovery of Colour“ nicht auf der Amazon-Startseite von X erscheinen, sondern auf der von Y – falls aber Y das Buch nicht kauft, wird X pünktlich zum nächsten Geburtstag von Y Werbung dafür unterbreitet werden.

Wenn WhatsApp nicht nur die GPS- und Nachrichtendaten kannte, sondern außerdem Zugriff auf die Daten der Kamera am Bankomaten hätte, (Vergrößerung des Inputs von 2 auf 3) könnte es die Lücken in der Route von Z schließen und so eine Vergrößerung des Outputs von 3 auf 5 erzielen.

Die Liste ließe sich auf eine Reihe von Beispielen aus allen Bereichen ausdehnen. Sehr viel größere Ausmaße nimmt der Effekt etwa bei allen Lernprozessen an: Das Erlernen einer Sprache ist ein Paradebeispiel für exponentiellen Wissenszuwachs (das der Netzstruktur von Synapsen entspricht) beim Menschen - und, wie gesagt, höchstwahrscheinlich mittelfristig auch bei Maschinen.¹²

1. Ein Metadatum macht alle Daten wertvoller

Um bei Sprachen zu bleiben: Ein „Schlüsselwort“ erschließt den Sinn eines Satzes oder gar eines ganzen Textes. Bezogen auf Daten entspricht das einem Metadatum¹³, das der „Schlüssel“ zum Verständnis vor allem von sehr verschiedenartigen Daten sein kann.

Bezogen auf das Positionsdatenbeispiel von Z: Will man GPS-, Sprach- und Kameradaten zueinander in Beziehung setzen, muss man zusätzlich zur reinen Information auch das Format der Information kennen (Bild? Ton? GPS?) und wie diese verknüpfbar sind (z.B. anhand der GPS-Positionen der Kamera und des Cafés) – genau das sind Metadaten.

Ein besseres Beispiel ist die Identifizierbarkeit einer Person. Oft gibt ein entscheidendes Metadatum den Ausschlag für die Identifizierung eines Individuums, die es erst möglich macht, eine Menge verschiedener Daten zu einem personenbezogenen Datensatz zu ordnen.

In vielen Fällen ist gerade bei großen und verschiedenartigen Datenmengen völlig unklar, welche der Daten Metadaten sein könnten – eine interessante Konsequenz: Vom Standpunkt einer Analytistin, die noch nichts über die Struktur eines Datensatzes weiß, ist die statistische Wahrscheinlichkeit für jedes einzelne Datum, das entscheidende Metadatum zu sein, dann genau gleich groß. Das kann zu einem Paradoxon führen: Je größer eine Menge an Daten, desto aufwendiger ist es, sie zu ordnen – ein einfacher linearer Zusammenhang. Doch gleichzeitig steigt mit der Größe auch die Wahrscheinlichkeit, den entscheidenden Schlüssel zu erhalten, der das Ordnen schlagartig vereinfacht. Der Nutzen dieses

¹²Beispiel: Ein Algorithmus durchforstet eine altertümliche Übersetzung der Bibel. Sobald er die Bedeutung des Wortes „gebären“ erfährt, kann er verstehen, dass „er erkannte sie“ keine kognitive, sondern eine sexuelle Handlung ist. Es kommt zu einem Schneeballeffekt (diese Wendung taucht häufig auf) – das System erkennt, dass in nur in einem einzigen Satz „erkannte“ weit entfernt von „und zeugte“ auftaucht, dafür aber gefolgt von „dass er der Messias war“. Von da an kann der Algorithmus es z.B. für wahrscheinlicher halten, dass ein Hilfsverb vor einem dass-Satz eine kognitive Bedeutung hat - usw. usw.

¹³ Für einige aufschlussreiche und tiefgründige Gedanken zu Metadaten vgl. Baecker 2013

Hinweises hängt wiederum von der Größe des zu ordnenden Datensatzes ab, weshalb der Effekt letztlich eine exponentiell wachsende Gesamtbilanz aufweist.

Ein vergleichbares Phänomen wird in der klassischen Netzwerktheorie beschrieben: Robert Metcalfe geht davon aus, dass jeder Punkt eines Netzwerkes eine Verbindung zu allen anderen Punkten herstellt und sie damit wertvoller macht, entsprechend beschreibt Metcalfe's Law ein quadratisches Wachstum¹⁴ (vgl. Briscoe et al. 2006). Das Prinzip ist bei Daten das gleiche, nur machen wohl nicht alle Daten sich gegenseitig wertvoller, sondern entweder ein Teil einen anderen oder aber ein Metadatum die Gesamtheit aller anderen Daten. Außerdem sind natürlich alle Abstufungen dazwischen denkbar. Ich bin (mangels Daten...) nicht in der Lage, mich auf eine genaue Wachstumsfunktion festzulegen. Für eine grobe Modellierung des Effekts würde ich aber in Anlehnung an Metcalfe von quadratischem Nutzenwachstum eines konstanten Teils der gesamten Daten ausgehen: Wenn also beispielsweise ein Zehntel eines Datensatzes in der Lage ist, sich mit einem Zehntel aller Daten eines anderen Satzes zu verknüpfen (was einem Exponenten von 2, also quadratischem Wachstum entspräche) und die beiden Datensätze zusammengeführt werden (der Effekt also erst eintritt), lässt sich das Wachstum des Nutzens mit folgender Funktion darstellen: $U=d+(0,1d)^2$. U steht dabei für den ökonomischen Nutzen (Utility), d gibt die Menge der genutzten Daten (oder die Zahl gleich großer Datensätze) an. Es ließe sich lange darüber diskutieren, ob mein Zehntel annähernd gerechtfertigt ist, ob der Anteil an verknüpfbaren Daten viel höher liegt und etwa in Texten sogar gegen 1 tendiert, oder ob er in Wahrheit viel niedriger liegt, etwa weil ein Teil der Daten redundant sein könnte. Das Zehntel ist aber ohnehin vollkommen willkürlich gewählt und soll nur der Veranschaulichung dienen.

Mit dem direkten Netzwerkeffekt lassen sich reale, mächtige und über lange Zeit stabile Monopole der Kommunikationstechnologie erklären: Der auf den Erfinder des Telefons Alexander Bell zurückgehende Telefonkonzern konnte im Wesentlichen über 100 Jahre lang monopolartige Strukturen in den USA aufrechterhalten.¹⁵ Auch Microsoft hat mit *Windows* bis heute ein de-facto-Monopol auf Betriebssysteme. Die Erklärung ist für beide Phänomene die gleiche: Der Wert eines Telefons oder von Betriebssystemen steigt für einen Nutzer exponentiell mit der Anzahl der anderen Nutzer des Telefons oder des gleichen Systems. Wenn Information per se die gleiche Struktur aufweist – was heißt das dann für die massenhaft gespeicherte Information im Besitz von Unternehmen?

¹⁴ Der tatsächliche Netzwerkeffekt in der Kommunikationstechnologie entspricht allerdings wahrscheinlich doch eher einem logarithmischen oder „n-log n“-Wachstum, einfach weil nicht jeder mit jedem kommunizieren will (vgl. Briscoe et al. 2006).

¹⁵ von der Patentierung des Telefons 1876 bis zur gezielten Zerschlagung des Konzerns durch das US-Justizministerium 1984.

II. DATEN ZU DATEN

Ein natürliches Monopol auf Daten

Für eine Datenmenge, deren Nutzung mehreren Akteuren möglich ist und deren Nutzen exponentiell zu ihrer Größe steigt, bildet sich auf einem Markt von frei handelbaren Daten ein natürliches Monopol, weil ein Akteur die Datenmenge effizienter nutzen kann als mehrere Akteure je einen Teil von ihr.

Unternehmen haben dabei zwei Möglichkeiten, ihren Datenbesitz zu vergrößern: Sie können selbstständig Daten sammeln, etwa durch das Anbieten eines Dienstes mit Nutzerdatenspeicherung, oder Daten von anderen Unternehmen erwerben. In beiden Fällen hat dasjenige Unternehmen, das bereits über den größten Datenbesitz am Markt verfügt, einen Wettbewerbsvorteil.

1. Das Erlangen von Daten (Reinvestment-Effekt)

Die Fähigkeit eines Unternehmens, Daten aus einer Datenquelle zu gewinnen, lässt sich in hohem Maße anhand von bereits erlangten Daten aus derselben Quelle verbessern. Die Attraktivität (und damit das Datensammelpotential) einer Suchmaschine, eines sozialen Netzwerkes oder eines Nachrichtenservices hängt zunehmend davon ab, wie gut der betreffende Anbieter seine Nutzer kennt – individuell, aber auch in ihrer Gesamtheit. Ein exponentielles Wachstum setzt ein: „Reichtum an Daten führt zu besseren Online-Diensten und zu mehr Nutzern, was den Reichtum weiter erhöht.“ (Hofmann et al. 2015)

So verbessert Google die Qualität seiner Suchergebnisse permanent durch eine simple Nutzerfeedback-Schleife: Sowohl anhand des Online-Verhaltens einer großen Menge an Nutzern generell als auch anhand von individuellem Verhalten des jeweiligen Nutzers wird jede Suchanfrage neu optimiert. Auch Facebook ist für viele seiner User mittlerweile weder das ästhetisch ansprechendste noch das angesagteste soziale Netzwerk¹⁶, doch je mehr Verbindungen, Nachrichten, Fotos und Likes eine Nutzerin bereits dort integriert hat, desto größer erschien ihr der Verlust bei einem Wechsel in ein anderes Netzwerk. Gleichzeitig optimiert auch Facebook beispielsweise seinen individuellen News-Feed mit exponentiell zunehmendem Erfolg. Andere Seiten haben eine Vielzahl von Strukturen entwickelt, die zunächst Daten gewinnen und sich dann mit diesen Daten weiter

¹⁶ In den USA, aber auch in Deutschland lehnen zunehmend junge Nutzer Facebook ab, weil sie die gemeinsame Nutzung mit der Generation ihrer Eltern ablehnen.

verbessern lassen: Von den Bewertungs- und Belohnungssystemen der Online-Shops und Vergleichsportalen über die Beobachtung des Klickverhaltens auf News- und Entertainmentseiten bis hin zu einer Vielzahl von Cookies, Links und Add-Ons.

Diese Technik bestimmt letztlich das Überleben aller Webfirmen, und zwar umso mehr, je bedeutsamer das Datensammeln für ihr Geschäftsmodell ist. Dabei ist allein steigende Attraktivität angesichts einer steigenden Zahl innovativer Start-Ups noch lange kein Garant für eine Monopolstellung - sie erklärt aber, warum Daten sich jeden Tag ganz von selbst mehr konzentrieren. In einer Zukunft, die von einer sehr viel umfangreicheren Erfassung von verhaltensrelevanten Daten bestimmt sein könnte, wird sich dieser Effekt aber noch massiv verstärken. Wer in der Lage ist, präzise Sprache zu verstehen, Verhaltensvorhersagen zu treffen und sein Geschäftsmodell durch neue verhaltensökonomische Theorien zu aktualisieren, wird wiederum Daten auf einer komplett neuen Stufe an Effizienz sammeln können (und abgesehen davon möglicherweise beginnen, riesige Gewinne einzufahren – dazu in III.) Bisher spricht aber alles dafür, dass der Return eines Reinvestments bei keiner Kapitalart so groß ist wie bei Daten. Ökonomisch noch sehr viel offensichtlicher ist jedoch der zweite Mechanismus der Datenkonzentration:

2. Der Erwerb von Daten (Übernahme-Effekt)

Angenommen, nützliche Daten werden auf einem freien Markt mit gut informierten Akteuren gehandelt. Steht eine Menge an Daten zum Verkauf, erhält immer die Meistbietende den Zuschlag. Der Betrag, den eine potentielle Käuferin in Daten investieren kann, spiegelt im Großen und Ganzen den Gewinn wieder, den sie durch die (wie auch immer geartete) Nutzung der Daten erwartet.

Die käufliche Datenmenge ist für alle potentiellen Käufer gleich groß, doch jeder von ihnen könnte die neuen Daten mit seinem schon vorhandenen Datenbesitz verknüpfen. Wenn der Nutzen von Daten aber exponentiell zu ihrer Menge steigt, heißt das: Diejenige Käuferin, die bereits über den größten Datenbesitz verfügt, erwartet, den größten Nutzen aus den neuen Daten zu ziehen. Damit kann sie das beste Gebot auf die käufliche Datenmenge abgeben und wird den Zuschlag erhalten (womit sich ihr Datenbesitz erneut vergrößert und sie für den nächsten Datenkauf in eine noch günstigere Position versetzt).

Das prominenteste Beispiel für eine Transaktion dieser Art: Im Februar 2013 übernahm Facebook für 19 Milliarden Dollar den Nachrichtendienst WhatsApp. Dabei handelte es sich zwar nicht um einen reinen Kauf bestehender Daten, WhatsApp soll auch weiterhin sammeln (mittlerweile wurde es in Facebook integriert). Interessant ist aber: WhatsApp hatte noch nicht einmal eine eigene erkennbare

Datennutzungsstruktur entwickelt. Die Gründer scheinen es also von Anfang an auf den Verkauf der gesammelten Daten angelegt zu haben¹⁷. Andere Beispiele sind der Kauf der Videoplattform Vine durch Twitter (2013) oder, schon 2006, die Google-Übernahme von Youtube.

Für die Übernahme von Daten spielt es im Kapitalismus auch nur eine untergeordnete Rolle, ob eine Käuferin tatsächlich über mehr Finanzkapital verfügt als ihre Wettbewerber, ihre Gewinnerwartung muss nur die Höhe der Finanzierungszinsen übersteigen.¹⁸ Natürlich sind Ausnahmen denkbar: Unternehmen, die kurzfristig außergewöhnlich hohe Investitionen tätigen, um sich mittelfristig in eine günstigere Marktposition zu manövrieren (vgl. Abschnitt: „Das geniale Startup“). Auf lange Sicht verhindert das jedoch keine Monopoltendenz.

Eine entscheidende Frage ist, ob überhaupt eine ausreichend große Zahl an Wettbewerbern die zum Verkauf stehenden Daten mit ihrem schon vorhandenen Datenbesitz verknüpfen kann. Das ist eine Annahme, die sich auf meine anfängliche Darstellung der Entwicklung stützt, die ich für wahrscheinlich halte: Die wichtigsten Daten der Zukunft werden extrem gut vernetzbar sein, weil es vor allem um soziale Informationen, ergänzt durch geografische und biologische Daten geht.¹⁹ Aus diesen werden sich extrem gute Vorhersagen für menschliches Verhalten treffen lassen, die für alle kapitalistischen Akteure in einer Dienstleistungsgesellschaft von zentralem Interesse sind. Deshalb halte ich die Tendenz zur Monopolbildung, selbst wenn sie nur für diese Art von sehr gut verknüpfbaren Daten gelten sollte, insgesamt für höchst relevant.

Eine grobe Skizze

Ich will nur der Anschaulichkeit halber eine sehr, sehr grobe Skizze davon zeichnen, wie ich mir das Wachstum der monopolbildenden Kräfte vorstelle. Dazu übernehme ich für den Datenverknüpfungseffekt die Idee des Zehntels aus Teil II, was sich als $u = d + (0,1d)^2$ darstellen lässt. Dabei steht d für Daten, den Informationsinput, und u für Nutzen, den ich mit Informationsoutput gleichsetze.

¹⁷ Wie der größte Teil der datensammelnden Startups seine Datensätze irgendwann an größere Konkurrenten verkaufen muss, wird in „Die Datenfresser“ (Kurz et al. 2011) anschaulich beschrieben

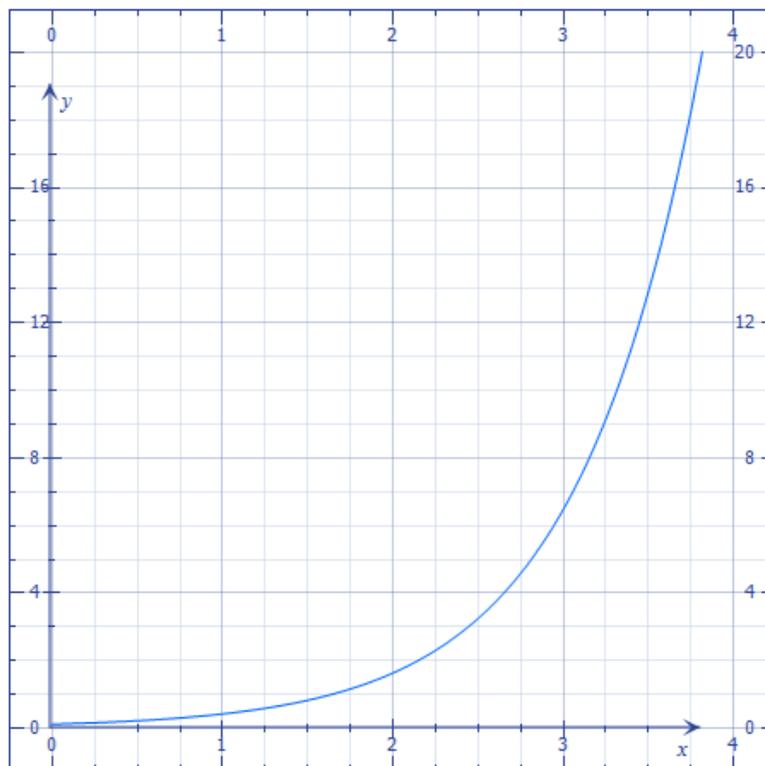
¹⁸ In die Praxis wird die Käuferin mit den höchsten Gewinnerwartungen – und damit den niedrigsten relativen Kosten – häufig auch über die günstigsten Finanzierungsbedingungen verfügen bzw. solche aushandeln können.

¹⁹ Denkbar wäre sogar eine Art digitale *lingua franca* für Verhalten, in die verschiedenartige Daten übersetzt werden könnten. Durch die Etablierung solcher gemeinsamer Standards würden die informationstechnischen Transaktionskosten für Datenübernahmen – ein durchaus kritischer Faktor – drastisch sinken.

Ich bin in Teil I. davon ausgegangen, dass das Wachstum der Analysemöglichkeiten zunächst mit den exponentiell wachsenden Datenmengen mithalten und wahrscheinlich im Verhältnis sogar noch linear wachsen kann, bevor es irgendwann von den explodierenden Datenmengen abgehängt wird. Der Einfachheit halber werde ich hier aber, relativ willkürlich, eine konstante Verarbeitungsrate von 10% annehmen (was zwanzigmal über der gegenwärtigen Rate, siehe S.4, liegt). Der Dateninput d abhängig von der Zeit t ist die einzige Größe, die halbwegs plausibel prognostiziert werden kann: Sie wächst insgesamt ziemlich genau mit $D_0^{0,4t}$ (wegen der Wachstumsrate von 40% pro Jahr, vgl. Teil I.) wobei D_0 den gegenwärtig gespeicherten rund 4 Trillionen Gigabyte Daten entspricht, wovon dann aber nur konstant 10% verarbeitet werden. Daraus ergibt sich ein Wachstum von $d = 4^{0,4t}/10$.

Angenommen, ein einzelner Akteur könnte all diese Daten in seinen Besitz bringen und analysieren – dann würde sich sein Output an nützlichen Informationen langfristig wie folgt entwickeln:

$$u = 0,1 \left(4^{0,4t} + \left(\frac{4^{0,4t}}{10} \right)^2 \right)$$



(y-Achse: Zeit t, x-Achse: Datenmenge d)

Einwände und gegenläufige Effekte

Eine naheliegende Kritik an meiner These könnte darauf abzielen, dass ich den Datenverwertungstechniken eine so viel geringere Aufmerksamkeit schenke als der Datenmenge. Tatsächlich nehme ich an, dass alle Marktteilnehmer über annähernd die gleichen Algorithmen, Programmiersprachen und nicht zuletzt die gleiche Menge an menschlicher Erfahrung und Talent verfügen. Zwar werden diese Techniken sich extrem weiterentwickeln: Ich bin ja zu Beginn von einer ungefähr linearen Steigerung des Anteils nutzbarer Daten an der exponentiell wachsenden Datenmenge ausgegangen. Das würde bedeuten, dass die Menge der nutzbaren Daten noch schneller wächst als die der gesamten „Roh-“Daten. Somit würde die Wachstumskurve noch sehr viel steiler verlaufen als in meiner Skizze (auf S.16), für die ich ja nur eine konstante Rate angenommen habe.²⁰

All das hat aber keine wesentliche Bedeutung für die Monopoltendenz, denn die Unterschiede zwischen den entsprechenden Wettbewerbern dürften bei der technischen Entwicklung insgesamt sehr viel geringer ausfallen als bei der Größe der Datenmengen. Dafür gibt es eine Reihe guter Gründe:

1. Know-How wird traditionell als eines der beweglichsten Güter betrachtet. Auch die komplexesten Algorithmen erfordern sehr viel weniger Speicherplatz als die Daten, auf die sie angewendet werden. Deshalb und aufgrund zunehmender Etablierung gemeinsamer Standards, etwa für Datenformate, sind Kooperationen zwischen Unternehmen relativ einfach.
2. Die personelle Fluktuation und damit der Transfer von Wissen ist in der Datenbranche kein unerheblicher Faktor. Die hellsten Köpfe der Datenwirtschaft lassen sich heute sogar, meist jeweils nur für ein Projekt, über die Plattform *Kaggle* von den verschiedensten Unternehmen rekrutieren. Außerdem werden Informatiker und *Data-Scientists* weltweit an vielen unabhängigen Standorten und Universitäten ausgebildet.
3. Ein großer Teil der innovativsten Software wird heute in offenen Gemeinschaften wie der OpenSource-Community entwickelt. Damit sind sie für jeden und oft sogar kostenlos zugänglich.

²⁰ Übrigens ist langfristig auch ein ganz anderes Szenario denkbar, nämlich wenn die Datenverarbeitung irgendwann nicht mehr mit den explodierenden Datenmengen Schritt halten können sollte und die Menge der nutzbaren Daten langfristig nicht exponentiell, sondern z.B. nur noch logarithmisch oder linear wächst. Das Wachstum wie auch die Monopoltendenz würde dadurch aber höchstens verlangsamt, nicht komplett ausgeschaltet.

4. Die Kosten für Prozessoren wie auch für Speicherplatz folgen normalen Marktmechanismen, es gibt hier keine Tendenz zu Monopolen.

Vor allem aber waren zumindest bisher große Datenmengen erst die Bedingung für die Entwicklung von Verarbeitungstechniken: Über die innovativsten Algorithmen verfügen heute höchstwahrscheinlich Google oder Facebook, nicht Moodle oder der Webmaster der HU-Homepage. Sollte sich dieser Trend fortsetzen, würde sich die Monopoltendenz nicht abschwächen, sondern noch verstärken.

Das „geniale Startup“

Trotzdem ist natürlich ein Szenario vorstellbar, in dem ein Akteur, auch mit geringem Datenbesitz, eine so innovative Analysetechnik entwickelt, dass er einen größeren Nutzen aus einem Datensatz ziehen könnte als Wettbewerber, die über mehr Daten verfügen. Dadurch wäre er in der Lage, eine große Investition in das Sammeln von neuen Daten zu tätigen oder höhere Gebote auf Datensätze abgeben zu können als seine größeren Wettbewerber.

Tendenziell nehmen Unterschiede in der Innovationsgeschwindigkeit in jungen Branchen wie der Datenverarbeitung aber mit der Zeit durch die oben beschriebenen Effekte eher ab, was die Wahrscheinlichkeit für das Auftreten solch „genialer Startups“ verringert – aber natürlich nicht ausschließt. Die Qualität von Verarbeitungstechniken ist schließlich der Hauptgrund, warum sich real noch keine extremen Datenmonopole gebildet haben: Es gibt einfach noch keine Super-Algorithmen, die mit vielen verschiedenartigen Daten fertigwerden. Und nur wenn ein einzelner Akteur über längere Zeit einen solchen Algorithmus als einziger verwenden könnte, würde der Mechanismus des Datenmonopols wirklich nachhaltig ausgehebelt (allerdings würde dann stattdessen ein, wahrscheinlich zeitlich begrenztes, „Algorithmenmonopol“ entstehen). Dann läge aber nach klassisch-ökonomischer Auffassung kein Marktversagen vor, sondern lediglich die beeindruckende Leistung eines einzelnen Akteurs.²¹

²¹ Vgl. in diesem Zusammenhang Welter 2013. Ein kurzer Exkurs: Sehr viel besser als das Vorpreschen einzelner Datenverarbeiter kann ich mir eine Datensammel-Entwicklung vorstellen, die heutige Dienste geradezu lächerlich ineffizient anmuten lässt. Heißester Anwärter wäre m.E. eine neuen Generation sozialer Netzwerke, die auf extrem realistisch simulierten und personalisierten Spielen basiert (der „Virtual Reality“ – Marktführer Oculus Rift wurde kürzlich von Facebook gekauft...). Die Erfolge von „Social Games“ wie *Second Life*, das Facebook-eigene *Farmville* oder auch das Fantasy-Rollenspiel *World of Warcraft* mit seinem enormen Suchtpotential geben Hinweise darauf, wie eine Umwelt geschaffen werden könnte, aus der sich ein Vielfaches der Daten aus heutigen Sozialen Netzwerken gewinnen ließe. Wo sonst ließen sich besser individuelle Entscheidungen aufzeichnen und gezielter verhaltensökonomische Experimente durchführen als in komplexen, beliebig anpassbaren Spielwelten? Und anhand von bestehenden Nutzerdaten sowie permanentem In-Game-Feedback ließe sich jedem Spieler (mit exponentiell zunehmender Perfektion) das „Spiel seiner Träume“ maßschneidern. Es

Die „Chance der Alten“

Ein anderer gegenläufiger Effekt lässt sich am besten als „zukünftige Knappheit älterer Daten“ beschreiben. Wegen der exponentiell wachsenden Menge an speicherbaren Daten werden in Zukunft Daten umso knapper erscheinen, je älter sie sind. Für viele Mustererkennungsprozesse ist der beobachtete Zeitraum ein kritischer Faktor, und für Forscher, die solche Studien in Zukunft beginnen, können einige historische Daten möglicherweise nützlicher sein als eine große Menge neuerer Daten. Dieser Effekt kann die Regel vom exponentiellen Wachstum des Datenwertes aushebeln. Und Unternehmen, die heute schon Daten sammeln, könnten einen relativen Vorteil gegenüber zukünftigen Datensammlern haben, auch wenn diese schon einen größeren Datenschatz angehäuft haben. Das würde einer Konzentrierung von Daten entgegenwirken.

Staatliche Eingriffe

Weder das deutsche oder europäische, noch das US-amerikanische oder überhaupt irgendein Wettbewerbs- oder Kartellrecht greifen bei Datenmonopolen. Kartellgerichtlich untersucht werden können Unternehmen nämlich nur, wenn ihre Umsätze – nicht etwa ihr Börsenwert, der wohl durchaus vom Datenbesitz abhängt (vgl. S.25) – über einer „Aufgreifschwelle“ liegen, die z.B. WhatsApp bei der Übernahme durch Facebook gar nicht erreichte. Auch wenn die deutsche Monopolkommission, die das Bundeskartellamt berät, in einem Sondergutachten vom Juni 2015 entsprechende Änderungsvorschläge gemacht hat²² – es bleibt ungeklärt, nach welcher Methode Daten, deren Nutzung erst in ferner Zukunft erwartet wird, ein fiktiver Unternehmenswert zugeordnet werden könnte.

Auch die wichtigste Form von staatlicher Intervention fällt m.E. bei Daten weg: Eine progressive Daten-Steuer würde zwar effektiv Monopole verhindern. Ich halte sie aber nicht nur für praktisch undurchführbar, sondern möglicherweise auch unvereinbar mit den demokratischen Grundrechten auf unabhängige Information und Kommunikation.²³ Generell zahlen große Konzerne, die umfangreiche Möglichkeiten zum Offshoring (gerade ihrer Datensparte) haben, eher weniger reguläre Steuern als kleinere Unternehmen. Es gibt also weltweit keine staatlichen Eingriffe in den Markt, die

würde mich nicht wundern, wenn der Monopolist der Zukunft nicht Google oder Facebook, sondern der Anbieter solcher Spiele wäre – Suchverhalten mag eine passable Datenquelle sein, *Suchtverhalten* aber ist unschlagbar.

²² So soll die EU-Kommission auch Zusammenschlüsse untersuchen dürfen, in denen ein einzelner – nicht notwendigerweise umsatzgenerierender - Vermögenswert 5 Mrd. € übersteigt. (vgl. Monopolkommission 2015)

²³ In Ungarn entwarf die rechtsnationale Regierung (im Zuge der quasi-Gleichschaltung öffentlicher Medien) im Oktober 2014 eine solche Daten-Steuer. Nach massiven Protesten von Opposition und Bürgern waren die Pläne in wenigen Tagen vom Tisch (vgl. <http://uk.reuters.com/article/2014/10/22/uk-hungary-internet-tax-idUKKCN0IB0RI20141022>).

direkt oder indirekt die Konzentrationstendenz von Daten in relevantem Maße beeinträchtigen würden.

Noch einige Worte zu einer zentralen Annahme: Meine Herangehensweise ist in dem Sinne mainstream-ökonomisch, dass sie einen freien Daten- und Kapitalmarkt mit rationalen, gut informierten Akteuren voraussetzt. Paradoxerweise verstärkt sich der beschriebene Effekt aber umso mehr, je weniger die (neo)klassischen Fiktionen freier Markt und *homo oeconomicus* sich als gerechtfertigt herausstellen: Die Analyse von Verhaltensdaten schafft einen umso größeren Wissensvorsprung und damit ökonomisches Potential, je mehr das tatsächliche Verhalten der Menschen und Märkte von dem abweicht, was die bestehenden Modelle der klassischen Ökonomik ohnehin nahelegen würden. Dafür, diese Vereinfachung zunächst zu übernehmen, spricht allerdings: Wenn ich schon die Anhänger relativ „marktoptimistischer“ Prämissen überzeugen kann, dass ein Marktversagen droht - dann nur umso mehr Verhaltensökonomien und solche, die sich ohnehin als egalitaristisch oder linksliberal verstehen.

Der Unterschied zwischen Daten und (anderem) Kapital

Auf den ersten Blick mag es aussehen, als könne der Zinseszinsseffekt jegliches Kapital, egal ob Bargeld, Immobilien oder Produktionsgüter, in exponentieller Weise vergrößern. Warum das tatsächlich nicht geschieht, ist eine der grundlegendsten ökonomischen Fragen. Vereinfachend lässt sich festhalten, dass vor allem die Knappheit von Ressourcen, die Sättigung von Absatzmärkten und die nicht-regelmäßig ablaufende Veränderung der Produktivität verschiedener Kapitalarten (z.B. durch neue Erfindungen) dafür sorgt, dass nicht ewig in den gleichen Kapitalstock reinvestiert werden kann, sondern Kapital sich immer wieder einen neuen, nämlich idealerweise den effizientesten Weg sucht. Warum nun sollen sich ausgerechnet Daten anders verhalten?

Ich kann diese Frage in der vorliegenden Betrachtung leider nicht ausführlich klären. Nur ein kurzer Transfer der Faktoren, die das Potential eines Kapitalstocks limitieren, auf Daten: Die „Ressource“ Information ist nicht in gleicher Weise einer natürlichen Begrenztheit unterworfen wie materielle Rohstoffe und auch eine Sättigung des Datenmarktes (gleichsam des menschlichen Wissenshungers) ist schwer vorstellbar: Der Satz „Zu viel Kapital tötet das Kapital“ (Piketty 2013) gilt bei Daten einfach nicht, Information mindert niemals den Wert anderer Information – sie erweist sie höchstens als falsch, was wiederum nichts anderes als einen Zugewinn an (besserer) Information darstellt.

Einzig die Frage nach der Qualität einer Art von Daten in einem Nutzen generierenden, „produktionstechnischen“ Sinne ist auf lange Sicht schwer abzuschätzen. Im Zusammenhang mit den im ersten Teil der Arbeit beschriebenen Entwicklungen würde ich gerne einfach annehmen, dass die Nutzenfunktion insbesondere der verhaltensrelevanten Daten zunächst wachsen und dann über einen gewissen Zeitraum stabil bleiben wird. In Anbetracht der Bedeutung von Verhaltensvorhersagen für alle Bereiche der Wirtschaft halte ich das für einigermaßen gerechtfertigt.

Ob Daten nun terminologisch als eine Art von Kapital, als Ressource oder gar als Produkt verstanden werden sollten, spielt vorerst keine Rolle. Daten können zu jedem dieser Dinge werden – die Frage ist nur, wann das jeweils geschieht. So sind endlose Datenzuwächse ja kein Selbstzweck, das letzte Ziel von Unternehmen ist die Maximierung des eigenen Gewinns. Wann also werden Daten zu Geld?

III. DATEN ZU GELD

„To make money, you've got to predict two things. What is going to happen and what people think is going to happen“

– Hal Varian, Google-Chefökonom (Varian 2013)

Ich werde von meinem M-Phone geweckt, dessen GPS-Modul auf dem Weg in die Bibliothek unablässig sendet. 90% der Recherche für diesen Text erfolgt via M-Search, während M-Network im Tab daneben geöffnet ist und ich über M-Streaming Musik höre. Gegen Mittag suche ich mit M-Maps etwas zu essen – genau wie ein Mädchen auf der anderen Seite der Bibliothek. Angesichts der historischen Daten aus M-Dating und unserer Interessen auf M-Network schlägt M-Maps uns – scheinbar zufällig – das gleiche Café (Starbucks) vor. Zur gleichen Zeit passt M-Trading seine Investitionen in Verlobungsring-Edelmetalle, in Starbucks-Aktien sowie in Soja-Derivate um winzige Bruchteile eines Cents an (weil die Wahrscheinlichkeit, dass jeder von uns seinen Kaffee mit Sojamilch bestellt, basierend auf historischen M-Messenger-Konversationen ausreichend hoch eingeschätzt wird). Durch diese Anpassungen an statistische Wahrscheinlichkeiten verdient das Monopol M vielleicht insgesamt nur Tausendstel von einem Cent, aber genau so etwas tut es nun in jeder Minute milliardenfach und mit zunehmender Perfektion.

Seit Jahren wird diskutiert, ob Google und Facebook natürliche Monopole sind. Wer dafür argumentiert, verweist auf Netzwerk- und Skaleneffekte²⁴. Die Gegenseite beruft sich dann vor allem darauf, dass es doch genug Konkurrenten gebe, die (Online-)Anzeigen verkaufen – und das sei schließlich der relevante Markt²⁵.

²⁴ vgl. etwa Bernau 2015 in der FAZ. Wie gesagt: Diese klassischen Modelle können weder den Erfolg von Google und Facebook befriedigend erklären, worauf Google auch zu Recht hinweist (vgl. Varian 2008), noch lässt ihr Horizont Platz für die drastischen Veränderungen des Marktes durch Big Data.

²⁵ Zu diesem Schluss kommen etwa die US-Wettbewerbsjuristen Manne und Wright in einem 70-Seiten-Beitrag für das Harvard Law Journal (vgl. Manne et al. 2011). Die Umwandlung von Google in die Holding Alphabet im Oktober 2015 beweist, dass man in Mountain View alles sein will, aber keine Suchmaschine, von einem Anzeigenverkäufer ganz zu

Ich bin ganz im Gegenteil davon überzeugt, dass die Daten nur einen Bruchteil ihres Potentials entfalten, wenn Sie zu Werbezwecken genutzt werden – Personalisierung hin, manipulative Ansätze her. Natürlich ist Werbung ein einfacher Weg, Daten zu Geld zu machen – aber der Effizienteste?

Silicon Vall Street

Forscher der Boston University und der Warwick Business School beobachteten über acht Jahre lang wöchentlich die Häufigkeit US-amerikanischer Google-Suchanfragen nach 98 ökonomisch relevanten Begriffen (wie „unemployment“, „housing“, „revenue“), die *Google Trends* öffentlich bereitstellt, und nutzten die Daten zum Handel mit der Dow-Jones-Indexaktie (vgl. Preis et al. 2013). Sie fanden heraus: Mit einer ausschließlich auf diesem Suchverhalten basierenden Anlagestrategie hätten sie in diesem Zeitraum eine Rendite von 60% (im Vergleich: 16% mit einer *buy and hold*- , 33% mit einer *random trading*-Strategie) erzielt, anhand einzelner Begriffe²⁶ sogar bis zu 326% – und das als externe Börsen-Amateure, die einmal pro Woche ein öffentlich zugängliches Google-Tool nutzten, um eine einzige Aktie zu handeln. Das ist noch kein Fuß in der Tür der Schatzkammer, das ist höchstens ein kurzer Blick durchs Schlüsselloch. Es erfordert keine allzu große Fantasie, sich vorzustellen, was Google selbst schon heute allein mit Börsenspekulationen verdienen kann. Neben den Suchanfragen stünden Google-Tradern einerseits unvorstellbare Mengen an exklusiven Bewegungsdaten, Mailinhalten, Kalenderdaten etc. pp. zur Verfügung, dazu kommen dann noch umfangreichen Möglichkeiten zur Beeinflussung der Realität (wie im Café-Szenario, das so fantastisch also gar nicht ist). Ganz offensichtlich lassen sich aus einer ausreichend großen Datenmenge schon mit den simpelsten Algorithmen relativ präzise Vorhersagen treffen, und in einem globalisierten Finanzsystem lässt sich ein solcher Wissensvorsprung so effektiv zu Geld machen wie nie zuvor.²⁷ Wenn also alle großen Datenkonzerne allein durch den Börsenhandel schon heute ein Vielfaches ihrer durch Werbung generierten Einnahmen erzielen könnten - warum tun sie es ganz offensichtlich nicht?²⁸

schweigen. Die Wahl der elementarsten Struktur von Sprache als Name ist angesichts der Entwicklung, die ich in Teil I annehme, nicht ganz uninteressant.

²⁶Der ertragreichste Suchbegriff war „debt“ – die Forscher erkannten die Finanzkrise von 2008 früher als viele andere.

²⁷ Allein der bereits automatisierte High-Speed-Handel, der an der Wall Street über 70% der Börsenumsätze ausmacht (vgl. Klausnitzer 2013) macht die direkte, blitzschnelle Anwendung jeder guten Vorhersage extrem lukrativ.

²⁸ Interessanterweise nimmt der Erfolg der Google Trend- basierten Anlagestrategie über den beobachteten Zeitraum bereits ab. Das könnte durchaus ein Hinweis darauf sein, dass gegen Ende des Forschungszeitraums bereits „echte“ Investor begannen, solche Datenanalyse in ihre Strategien zu integrieren, und den Wissensvorsprung der Forscher damit sukzessive ausglich. (vgl. Preis et al., 2013).

Geduld zahlt sich aus

Im Gegenteil scheinen die bekanntesten dieser Unternehmen das ökonomische Potential ihres Datenkapitals in der Öffentlichkeit systematisch herunterzuspielen. So spricht Google-Chefökonom Hal Varian davon, dass Google „nicht die besseren Zutaten, sondern das bessere Rezept“, also nicht einen größeren Datenschatz, sondern lediglich einen besseren Algorithmus als seine Mitbewerber habe (Varian 2008). Parallel dazu verbrauchen dann auch harmlose bis gemeinnützige Projekte wie Googles FluTrends (das Grippewellen im Schnitt zwei Wochen vor dem CDC erkennt, vgl. Loukides 2010) oder Schlagzeilen wie „Facebook kann Liebes-Aus vorhersagen“²⁹ auch einen weitaus größeren Teil der knappen Ressource öffentliche Aufmerksamkeit als etwa das Börsenexperiment der Bostoner Forscher.

Da die Opportunitätskosten für den Wechsel der Suchmaschine oder der Sozialen Medien sich trotz aller Netzwerkeffekte als relativ niedrig erwiesen haben (man erinnert da gerne an den Niedergang von Yahoo! oder Myspace), wandern solche Firmen immer auf einem schmalen Grat: Zu einen können sie nur gerade so viele Daten sammeln, dass keine Massenflucht durch empörte User initiiert wird.³⁰ Zum anderen sollte aber auch die Anwendung der Daten, etwa zu Werbezwecken, nicht in den Bereich allzu verblüffender oder gar gespenstischer Effizienz gelangen, was den gleichen Effekt auslösen könnte– genau deshalb kann ich mir auch vorstellen, dass personalisierte Werbung von vielen Anbietern schon heute eigentlich besser positioniert werden könnte, als sie es tatsächlich wird³¹. Es geht aber eben niemandem um diesen mittelfristigen Profit, sondern vorerst um einen Kompromiss zwischen der langfristigen Erhaltung von Datensammelstrukturen und gerade genug Gewinn für ein optimales Reinvestment in diese. Würde Google morgen beginnen, massiv und erfolgreich an der Börse Kapital „aufzusaugen“, würde übermorgen kein Mensch mehr Maps und Gmail nutzen³². Nein,

²⁹ <http://www.n24.de/n24/Nachrichten/Netzwelt/d/3760478/facebook-kann-liebesaus-vorhersagen.html>

³⁰ Tatsächlich steht die Frage im Raum, warum nicht schon die Enthüllungen Edward Snowdens einen solchen Exodus auslösten. Mir scheint, User empfinden ungefähr folgendermaßen: Die kommerziellen Datensammler sind selbst Opfer des staatlichen Missbrauchs, und zwar alle in gleicher Weise. Deshalb müsste man, um individuell vor staatlichem Missbrauch sicher zu sein, sich entweder komplett aus dem Digitalen zurückziehen oder aber öffentlich die Nachrichtendienste (wieder) einer strengeren demokratisch-rechtsstaatlichen Kontrolle unterwerfen (die erste Option ist langfristig sicherer, kurzfristig aber definitiv weniger verlockend als die zweite). Würde hingegen massiver Missbrauch durch einzelne Unternehmen bekannt, wäre ein Wechsel zur „sauberen“ Konkurrenz eine vergleichsweise einfache (ökonomisch: kostengünstige) Lösung.

³¹ Amazons notorisch unzulängliche Empfehlungen sind hier das beste Beispiel, vgl. etwa <http://www.huffingtonpost.com/2010/03/12/the-most-ridiculous-amazon496445.html>

³² Ein etwas verschwörungstheoretisch anmutendes Szenario, in dem Datensammler einfach mit bereits aktiven Börsenhändlern kooperieren, halte ich für unwahrscheinlich. So etwas wäre in der Größenordnung etwa eines Google/Goldman Sachs – Konglomerats angesichts recht strenger Insiderhandelskontrollen weder zu

Google und Facebook sind im Wesentlichen noch immer Start-Up-Unternehmen in dem Sinne, dass ihr gegenwärtiges Investment in keinem Verhältnis zum zukünftig erhofften Gewinn steht.

Dazu kommt noch, dass die Kosten sowohl für Rechenleistung als auch für die Ausbildung von Fachkräften ohne explizites Zutun der Datensammler sinken, und zwar höchstwahrscheinlich exponentiell³³. Zusammen mit dem Effekt der „Knappheit älterer Daten“ (vgl. Teil II) spricht das dafür, zunächst Daten zu sammeln und erst später in ihre Verarbeitung zu investieren.

Letztlich verfolgen alle Internet-Unternehmen zu einem gewissen Grad diese Strategie - und gießen damit Wasser auf die Mühlen meiner Idee vom exponentiellen Wachstum des Datenwertes: Wer seinen Profit zugunsten des Datensammelns einschränkt, muss letztlich davon überzeugt sein, dass der Wert von Daten sich stärker vermehren lassen wird als jede Art von (Finanz)kapital. Mit anderen Worten: Dafür, dass Google oder Facebook noch nicht an der Börse aktiv sind, kann es nur eine Erklärung geben. Jeder dieser Akteure weiß, dass sowohl sein aktueller Datenbesitz als auch die Möglichkeiten, diese zu analysieren, nur einen Bruchteil der zukünftigen Kapazitäten ausmachen. Deshalb lohnt es sich, weiter still und leise zu sammeln. Keine Kapitalart, in die sich durch eine Monetarisierung reinvestieren ließe, verspricht so hohe *Returns* wie die natürliche exponentielle Wertsteigerung von Daten.

All das kann nicht wirklich ein Geheimnis sein. Auch die oft astronomisch hoch erscheinenden Aktienkurse der Internetkonzerne spiegeln natürlich den Glauben der Anleger an zukünftige extrem lukrative Geschäftsmodelle oder, je nach Unternehmen, sogar nur an den steigenden Wert der gesammelten Daten wieder. Vergleicht man Unternehmenswert und Umsatz der dotcom-Akteure von 2000 mit ihren Nachfolgern im heutigen Silicon Valley, haben sich die Proportionen nicht wesentlich verändert. Dennoch spricht heute kaum jemand von einer Rückkehr der dotcom-Blase, weil das Vertrauen in den zukünftigen Nutzen der Daten so viel größer ist als noch um die Jahrtausendwende.

34

verheimlichen noch ansatzweise so profitabel wie die solitäre Strategie – wohlgemerkt: in einer fiktiven Verhandlung zwischen Kapital – und Daten“händlern“ säßen letztere den weit längeren Hebel...

³³ Zumindest die exponentielle Tendenz von „Moore’s Law“, wenn auch nicht seine Geschwindigkeit, wird wohl noch einige Jahrzehnte gehalten werden können – vgl. dazu <http://www.heise.de/newsticker/meldung/Intel-Chef-verabschiedet-sich-vom-bisherigen-Moore-s-Law-2751848.html>

³⁴ Der astronomisch hohe Börsenwert einer Marke wie WhatsApp hängt dabei noch nicht einmal mit Erwartungen auf zukünftige Gewinne des Unternehmens zusammen, und schon gar nicht an den Centbeträgen, die der Download der App heute generiert, sondern ausschließlich am Wert seiner gesammelten Daten. Siehe auch: WhatsApp-Übernahme durch Facebook.

Monetarisierung

Die Frage ist, an welchem Tag es sich nicht mehr lohnt, zu warten. Ausführliche Erklärungen hierzu würden den Rahmen der Arbeit sprengen, nur vier vorsichtige Antworten:

1. Wenn die Datenquelle unabhängig von der Monetarisierung sprudelt. Das ist eigentlich nur vorstellbar, wenn zum Beispiel ein Webdienst so unentbehrlich geworden ist, dass auch die öffentliche Erkenntnis, dass sein Betreiber Unmengen an Kapital abzieht, seine Nutzer nicht davon abhalten könnte, mehr und mehr Daten preiszugeben.
2. Wenn genau dieser Punkt ohnehin, auch aus anderen Gründen, erreicht ist. Dazu muss der Strom einer Datenmenge nicht einmal schrumpfen: Sobald das Wachstum einer Datenmenge nicht mehr exponentiell ist (und erwartungsgemäß auch nicht mehr wird), kehren sich die Relationen von Geld- und Datennutzen soweit um, dass eine umgehende Monetarisierung des Datenbestandes Sinn macht.
3. Wenn einem Unternehmen – aus welchen Gründen auch immer – schlicht das Geld ausgeht, es sich oder seine Daten aber nicht an einen (größeren) Konkurrenten verkaufen möchte. Heute sind solche Unternehmen aber meist noch zu klein, um eine echte Datennutzung etwa an der Börse in Angriff zu nehmen. Stattdessen werden oft verzweifelt wirkende Versuche unternommen, optionale Bezahlssysteme zu etablieren - wie zuletzt beispielsweise bei Soundcloud oder Tinder.
4. Nach der kapitalistischen Logik: Wenn der erwartete Gewinn aus der hinreichend effizienten Analyse einer hinreichend großen Menge an Daten maximal ist. Solange der Datennutzen steigt, tritt dieser Fall nicht ein. Sollte der Datennutzen ewig steigen, tritt er genau dann ein, wenn ein Unternehmen sämtliches bestehendes Kapital auf der Erde abziehen kann.

Fall zwei und drei sind die mit Abstand am wahrscheinlichsten, wobei Fall eins zumindest nicht auszuschließen ist. Das vierte Szenario ist, so grotesk es anmuten mag, das letzte logische Ziel aller Kapitalistinnen. An diesem Punkt ist meine Spekulation nichts weiter als ein Gedankenspiel in einem Bereich, in dem die kapitalistische Logik möglicherweise einfach nicht mehr gelten wird (wie würde sich jemand entscheiden, der alles Kapital auf der Welt besitzen könnte? Was würde er damit tun?).

Unabhängig davon sollte ein ganz anderer Punkt nicht vergessen werden: Die Macht einer Datenmonopolistin würde sich nicht auf weitreichendes Wissen beschränken, sondern auch umfangreiche Möglichkeiten zur Manipulation von individuellen Entscheidungen wie auch gesellschaftlichen Abläufen beinhalten. Soziale Netzwerke, Suchmaschinen und wenige andere

Webseiten bzw. Apps sind schon heute die Fenster, durch die ein wachsender Teil der Menschheit die Welt wahrnimmt. Genau diese Wahrnehmung können die entsprechenden Unternehmen mehr oder weniger nach Belieben beeinflussen, und spätestens an jedem der oben genannten vier Punkte hält prinzipiell nichts (es gibt weder entsprechende Gesetze noch Marktmechanismen) einen kapitalistisch agierenden Akteur davon ab, diese Macht auch einzusetzen – vielleicht der gefährlichste Aspekt, denkt man an die Profitabilität von Kriegen und Krisen aller Art.

So oder so fällt es nicht schwer, sich eine erstrebenswertere Zukunft vorzustellen als eine, die von Datenmonopolen gestaltet wird. Daher werde ich auf eine nähere Untersuchung, unter welchen Bedingungen ein solches Monopol im philosophischen Sinne *schlecht* wäre, verzichten,³⁵ und mich mit einer kurzen Idee dazu begnügen, welche wirtschafts- oder gesellschaftspolitischen Maßnahmen geeignet wären, diese Entwicklung einzuschränken oder gar zu verhindern.

³⁵ Schweren Herzens, da diese hochinteressant und, vergleiche man etwa meine utilitaristisch-ökonomische mit egalitaristischeren oder auch liberaleren Positionen, durchaus divergent ausfallen könnte. Ich vermute, als kleinster gemeinsamer Nenner würde aus allen Perspektiven vor allem die Ungleichheit, die alle Monopole auch in harmloseren Abstufungen verursachen, als eine Bedrohung für individuelle Freiheit, Demokratie sowie den Rechts- und Sozialstaat erscheinen – in Abwandlung des Google-Mottos ließe sich die Schreckensvision eines Datenmonopols also vielleicht auf die Formel „Don't be equal“ bringen...

IV. DATEN ZU ALLEN

Datenschutz?

Schon der Begriff Datenschutz hat für mich etwas Eigentümliches, suggeriert er in seiner Ähnlichkeit zu *Artenschutz* doch, dass die Daten selbst des Schutzes bedürften. Gemeint ist natürlich, dass die Quelle der Daten, nämlich der Mensch als Nutzer, geschützt werden soll, und zwar vor Missbrauch durch Akteure des Staates und auch des Marktes. Auch das ist eigentümlich, weil Daten als Eigentum ihrer Quelle, vergleichbar etwa mit der menschlichen Arbeits- oder Kaufkraft, betrachtet werden. Die Grundidee ist wohl, dass ein Mensch „seine“ Daten, ähnlich wie einen Teil seiner Kaufkraft, im Tausch gegen gewisse Dienste hergibt (oft ersetzen solche Dienste ja vergleichbare kostenpflichtige Angebote). Ganz abgesehen davon, dass dieser Handel dem Daten-Geber in vielen Fällen nicht bewusst sein dürfte, gibt es wesentliche Unterschiede zwischen diesen „Gütern“: Informationen über das Verhalten eines Menschen sind prinzipiell unbegrenzt. Ähnlich wie bei Arbeitskraft gibt es besser und schlechter nutzbare Daten³⁶, aber während Arbeit offenkundig ein endliches Gut ist, kann Wissen über einen Menschen ja sogar noch nach seinem Tod zunehmen. In diesem Kontext sind Verhaltensdaten nicht die Domäne der Ökonomik, sondern der Kommunikationswissenschaften, die erkannt haben: Wir geben permanent Informationen über uns preis, ob wir wollen oder nicht (bei Arbeit sieht das zweifellos anders aus), fraglich ist nur, wer sie aufnimmt. Ich finde: Wenn Daten irgendeiner Entität ähneln, die bereits in politisch-ökonomischen Theorien etabliert ist, dann einem praktisch unbegrenzten Rohstoff wie dem Sonnenlicht: Da jeder es nutzen darf, entwickeln sich historisch unterschiedlich effiziente Techniken, teilweise in hartem Wettbewerb, wovon aber schließlich alle profitieren. Sonne wie auch Informationen wurden schon immer in lebenswichtige Güter umgewandelt – Big Data wäre dann vergleichbar etwa mit der Entwicklung von Photovoltaik. Während letztere massiv subventioniert wurden, spannt man über erstere gleichsam einen Sonnen- bzw. Datenschutz.

Die zentralen Forderungen dieses Datenschutzes lassen sich in drei Kategorien einteilen. Keine davon ist geeignet, die Entwicklung von Datenmonopolen einzuschränken:

³⁶ wobei die Qualität von Arbeit wie auch Informationen immer von der Produktions- bzw. Nutzungsstruktur abhängt, so kann eine Arbeiterin nur hochqualifiziert für eine bestimmte Art von Nutzenproduktion sein, genau wie etwa die GPS-Position eines bekannten Attentäters letztlich nur besser genutzt werden kann als die jedes anderen Menschen

1. Bewahrung des Eigentums an Daten. Wie gesagt: Ich halte es für fragwürdig, sich Daten als persönliches Eigentum vorzustellen. So ist die Forderung nach transparenteren Datensammel-AGBs verständlich, doch von fraglichen Erfolgsaussichten: Jeden Tag nehmen Nutzer die Aufzeichnung „ihrer“ Daten, die ihnen meist auf abstrakte und diffuse Art bewusst ist, in Kauf, wenn der dafür angebotene Dienst attraktiv genug erscheint. Selbst wenn Datensammeln gekennzeichnet würde wie die Gefahren des Rauchens auf Tabakverpackungen – das Potential von Big Data ist so groß, dass Datensammler immer eine Mehrheit der Menschen zur Preisgabe „ihrer“ Daten verlocken werden können. Außerdem wirken AGBs ohne Angabe des Verwendungszweckes (siehe 3.) etwas sinnentleert.
2. „Gerechter“ Umgang mit schon erfassten Daten. Dazu zählen die Forderungen nach Anonymisierung, Offenlegung, Korrektur oder Löschung der „eigenen“ Daten. Das scheitert schlicht daran, dass die schiere Datenmenge Begriffe wie Anonymität oder Personenbezug karikiert. Einerseits machen Vernetzung, Gruppenbezug und multiple Speicherung von Daten es zunehmend unmöglich, die Zugehörigkeit eines Datums zu einer Person zu bestimmen. Andererseits ist in einem umfassenden und rasterbaren Datensatz der Name mithin das unwichtigste Identifikationsmerkmal. Im Übrigen widersprechen sich die Forderungen nach Anonymisierung und Offenlegung der „eigenen“ Daten, für letztere müssen (auch vormals anonyme) Daten ja (re-)personalisiert werden. Die Idee von einem „Recht auf Korrektheit von Daten“ zeugt von besonders kleinem Horizont: Daten sind per se nicht „falsch“, sie können höchstens irrelevant für die gestellte Frage sein. Und je mehr Daten es gibt, desto höher ist die Wahrscheinlichkeit, dass die relevantesten und damit „wahrhaftigsten“ Informationen sich durchsetzen – eine Art Daten-Darwinismus. Die als „inkorrekt“ empfundenen Daten (etwa unangenehme Google-Suchergebnisse zur eigenen Person) enthalten dabei letztlich nur eine weitere wertvolle Information.³⁷
3. Offenlegung von Datenverwertungszwecken und -techniken. Wer in AGBs den Verwendungszweck der Daten im Voraus offenbart sehen will, hat ein zentrales Prinzip von Big

³⁷ Ich gehe davon aus, dass die Forderung nach einem „Recht auf Vergessenwerden“ wohl überwiegend der gleichen Angst vor inkorrekten Informationen entspringt – so oder so zeigt sich schon heute, dass die Durchführung kompletter Löschungen nahezu unmöglich ist.

Data nicht verstanden, denn häufig ist der Reiz großer Datenmengen gerade, dass Forscher und Unternehmen sie analysieren können, ohne konkret zu wissen, wonach sie eigentlich suchen (vgl. Teil I). Andere Datenschützer fordern, die „Algorithmen offenzulegen“ (Noller 2013). Vor dem Hintergrund der Monopolgefahr halte ich das für genau den falschen Weg – so gibt es keinen (nützlichen) Wettbewerb der Verarbeitungsmethoden mehr, sondern nur noch des Datensammelns. Zweifellos ist die Forderung verständlich, weil es Noller vor allem um die Bedeutung von intransparenten Algorithmen für gesellschaftliche Entwicklungen geht. Dieses Problem muss aber durch einen fairen Wettbewerb gelöst werden, wenn man das Potential von Big Data anderweitig noch nutzen will.

Was darüber hinaus alle Vorschläge zur gesetzlichen Regulierung des Datenschutzes hinfällig macht: Jeder relevante Konzern kann sich ihnen durch „Offshoring“ seiner datenverarbeitenden Sparte in Länder ohne Datenschutzgesetze vollständig und ohne nennenswerten Aufwand entziehen, nur kleinere Betriebe und öffentliche Verwaltungen werden getroffen. Ich bin nicht gegen jede Form von Datenschutz. Aber aus einem gewissen historischen (expliziter: historisch-materialistischem) Abstand betrachtet, scheint mir ein makroökonomisches „Update“ der sozialen Marktwirtschaft in das globalisierte Datenzeitalter einfach vielversprechender als einzelne Detail-Regelungen. Es hat sich als langfristig unmöglich erwiesen, stur gegen den Strom des Kapitals zu schwimmen – man kann ihn nur zielführend umleiten. In diesem Fall heißt das: Gleichzeitig die größte Gefahr zu bannen, nämlich die des Datenmonopols, und einen möglichst großen Teil der Gesellschaft am enormen Potential von Big Data teilhaben zu lassen.

Wie bekämpfen Gesellschaften ein klassisches natürliches Monopol? Man öffnet den Markt, man „befreit“ das kritische Gut. Inspiriert davon würde ich gerne eine – erneut: bewusst naive – Skizze im Sinne einer „nützlichen Utopie“ (Piketty 2013) zeichnen.

Free Data – eine „nützliche Utopie“

„Die Situation würde sich erst grundlegend ändern, wenn diese Daten im Firmeneigentum aus den Datensilos (...) entlassen würden. Diese und andere kollektive Datenbestände könnten als Trainingsdaten die Grundlagen neuer Dienstleistungen sein, deren Entwicklung dann nicht in der Hand weniger globaler (amerikanischer) Megakonzerne liegen würde. (...) Die meisten dieser Daten haben die Benutzer niemals willentlich an Internetfirmen übertragen, sie bilden ein kollektives Gut, das aber nicht als Gemeingut verfügbar ist.“

- Thomas Hofmann und Bernhard Schölkopf, FAZ vom 29. Januar 2015 (Hofmann et al. 2015)

„Viel besser wäre es [die Daten] dezentral und untereinander verlinkt in der Cloud zu speichern. Ein gemeinsames Format würde deutlich mehr Menschen die Möglichkeit geben, in den Daten zu stöbern, dabei auf bislang verborgene Verbindungen zu stoßen und so einen Wettbewerb der Ideen in Gang zu setzen.“

- David Weinberger (Weinberger 2014)

Jeder darf so viele Daten sammeln, wie er möchte. Er muss sie nur jederzeit in der Cloud öffentlich zugänglich machen. Mit dieser einfachen Regel ist die Tendenz zu Datenmonopolen auf einen Schlag ausgehebelt. Die Regierungen sind nicht mehr überfordert, eine soziale Marktwirtschaft der Daten ist möglich³⁸. Weiterhin werden alle Akteure Daten sammeln, nun aber auch zunehmend in ihre Verarbeitung investieren. Der Wettbewerb um die beste Analyse und Monetarisierung der Daten wird aufblühen, und dieser unterliegt wieder allen Gesetzen des Marktes sowie bewährten staatlichen Kontrollen (dazu gehören auch zeitlich begrenzte Patente auf Algorithmen). Am stärksten profitieren nicht-kommerzielle, also gemeinnützige Organisationen. Der Erfolg von OpenSource in Verbindung mit einfachem Cloud-Zugriff deutet schon heute an, wie die Zivilgesellschaft Daten nutzen kann. Die größte Sorge des Bundesverfassungsgerichtes zur Vorratsdatenspeicherung BVerfGE 65,43 (auch Big Data S.136) lautete immer: Die Angst vor Aufzeichnung von Verhalten beeinflusst dieses (und gefährdet so die Demokratie). Aber die komplette Offenlegung aller Daten für jeden würde eine so viel größere Transformation der moralischen und rechtlichen Strukturen in Gang setzen, dass die – im Originalkontext Google-Interview geradezu perverse – Vision „If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place“ (Schmidt 2007) in einer tatsächlich erstrebenswerten Weise wiederauferstehen würde.

Zugegeben, der Einsatz von Daten etwa bei der Verbrechensbekämpfung würde ausgehebelt.³⁹ Es können beispielsweise Bewerber Nachteile aus ihrer politischen Überzeugung erwachsen, und was geschieht, wenn „niemand mehr lügen kann“, übersteigt jede Vorstellungskraft. Aber langfristig könnte Open Data in einer demokratischen Zivilgesellschaft einen großen Teil von Verbrechen

³⁸ Linklibertarier und Egalitaristen würden diesem utilitaristischen Ansatz, wenn auch aus anderen Gründen, zustimmen: Wenn Daten als Rohstoff betrachtet werden, darf jeder sie sich aneignen, solange er dadurch nicht die Möglichkeiten anderer, sich in gleicher Weise Daten anzueignen, verletzt.

³⁹ Diese werden aber ohnehin immer überschaubarer, wenn Verbrecher lernen, falsche Muster zu erzeugen. Und ja, auch krimineller Energie stehen die Daten zur Verfügung – aber eben nicht mehr nur der von Google und Facebook...

verhindern, Informationsasymmetrien abbauen und, so weit gehe ich, eine komplett neue, globalisierte (tendenziell utilitaristische) Vorstellung von Ethik begründen.

Natürlich ist die Errichtung eines Open-Data-Regimes alles andere als unkompliziert. Regierungen müssen es schaffen, alle Datensammler zur Offenlegung zu zwingen, und es ist zunächst völlig intransparent, welche Input-Daten einer Output-Information zugrunde liegen (also veröffentlicht werden müssen), insbesondere wenn die Algorithmen dazwischen geheim bleiben sollen. Geschwindigkeit wird in vielen Fällen entscheidend sein, die Echtzeit-Veröffentlichung aller Daten ist weder möglich noch von den jeweiligen Datensammlern gewollt. Außerdem wird es sehr viel schwieriger, sich dem Datensammeln komplett zu entziehen (aber das wird es so oder so). Und möglicherweise wird es sogar eher als „hip“ denn als verdächtig gelten, möglichst wenige Daten preiszugeben. Dirk Baecker beweist hier erneut einige Weitsicht: „Es kann sein, dass Big Data der Moment war, der das System unter seiner eigenen Komplexität zusammenbrechen ließ, so dass sich danach nur noch technikfeindliche Fundamentalistengruppen auf der Erde halten konnten“ (Baecker 2013). Tatsächlich könnte ich mir vorstellen, dass eine Subkultur von „Datenasketen“ versucht, sich dem Datensammeln so weit wie möglich zu entziehen (vgl. dazu auch Kurz et al. 2011). So oder so beginnt eine tiefgreifende Veränderung der Gesellschaft, wenn, überspitzt, jeder alles über jeden weiß. Aber: eine tiefgreifende Veränderung ist ohnehin unvermeidlich – und besser jeder weiß alles über jeden als ein Monopolist weiß alles über jeden.

Ich werde von meinem iPhone geweckt, dessen GPS-Modul auf dem Weg in die Bibliothek unablässig sendet. 90% der Recherche für diesen Text besteht aus der Analyse von öffentlich zugänglichen Daten.

Fazit

Zugegeben: Selbst wenn die Entwicklung eines Datenmonopols tatsächlich ein realistisches Szenario darstellen sollte, ist sie alles andere als das dringendste Problem der globalen Gesellschaft. Sie wird Zeit brauchen, und die Namen der Datensammler von heute sind eher in einem Tweet verlinkt als in Stein gemeißelt (wobei: was hält länger?). Dennoch spricht einfach nichts notwendigerweise dagegen, dass die Techniken aus meinen Prämissen eines Tages entwickelt werden, im Gegenteil deutet einiges darauf hin. Und wenn tatsächlich verknüpfbare Verhaltensdaten die nächste Generation von Big Data darstellen, werden sich weit mächtigere Monopole bilden als heute Google oder Facebook. Es gibt

keine ultimative Lösung für dieses Szenario, aber ein Ansatz, der Daten möglichst vielen Akteuren zugänglich macht, ist vielversprechender als alle bisherigen Datenschutzansätze.

Dabei darf nicht vergessen werden, dass Big Data vor all dem eine riesige Chance für alle Wissenschaften, darunter nicht zuletzt die Philosophie, bedeutet. Wissen ist ein Gut, das sich vergrößert, wenn mehr Menschen an ihm teilhaben. Ökonomisch ausgedrückt: Big Data wird einen enormen Nutzen generieren, die Frage ist vor allem, wie dieser sozialisiert werden und dadurch seinen Nutzen noch vervielfachen kann. Oder, wie Melane Swan schreibt:

“With the aid of big data service, human function could be extended into data interactions that more meaningfully fulfill human emotional and intellectual needs, and produce new levels in the subjectivation possibility for both humans, and data entities.”

- Melanie Swan (Swan 2014)

Die digitale Revolution wird oft mit der industriellen Revolution verglichen. Letztere hat soziale Fragen aufgeworfen, die bis heute nicht beantwortet sind, aber die dreihundert Jahre mit den blutigsten Konflikten in der Geschichte der Menschheit nach sich zogen – nicht zuletzt, weil die ökonomischen und politischen Antworten auf besagte Fragen von ausbeuterischer Ignoranz bis zu radikalsten Ideologien reichen. Wenn die digitale Revolution das Ausmaß von Segen und Fluch der industriellen Revolution vielleicht noch in den Schatten stellt, sollten wir vor allem einen Fehler nicht begehen: Uns zu spät Gedanken machen.

Literatur

- Anderson, Chris: "The end of theory. The data deluge makes the scientific method obsolete". *Wired* 16 No 7. (Im Netz unter http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory).
- Aslett, Matt: „Big Data reconsidered: It’s the economics, stupid“. (Paper aus dem Jahr 2013, im Netz unter <https://451research.com/report-short?entityId=79479&referrer=marketing>).
- Barabasi, Albert-Laszlo: *Bursts. The Hidden Patterns behind everything we do, from your E-Mail to bloody Crusades*. (New York: Plume, 2011).
- Baecker, Dirk: *Metadaten. Eine Annäherung*. Essay in: *Big Data. Das neue Versprechen der Allwissenheit*. (Berlin: Suhrkamp, 2013).
- Bernau, Patrick: "Was tun mit Google?". Kommentar in *Frankfurter Allgemeine Zeitung* (20.4.2015) (im Netz unter <http://www.faz.net/aktuell/wirtschaft/maechtige-internetriesen/das-google-verfahren-und-das-natuerliche-monopol-was-tun-13547280.html>).
- Briscoe, Bob: "Metcalfe’s Law is Wrong". Paper aus dem Jahr 2006 (Im Netz unter <http://spectrum.ieee.org/computing/networks/metcalfes-law-is-wrong>).
- Bollier, David: *The promise and peril of Big Data*. (Washington / DC: The Aspen Institute, 2010).
- Haucap, Justus / Kehder, Christiane: "Suchmaschinen zwischen Wettbewerb und Monopol: Der Fall Google". *Ordnungspolitische Perspektiven* 44 (2013).
- Hofmann, Thomas / Schölkopf, Bernhard: „Vom Monopol auf Daten ist abzuraten“. *Frankfurter Allgemeine Zeitung* vom 29.1.2015 (Im Netz unter <http://people.tuebingen.mpg.de/bs/faz-monopol-auf-daten.pdf>).
- IDC: *The Digital Universe*. Web-Animation, Stand 16.10.2015 (Im Netz unter <http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>)
- Klausnitzer, Rudi: *Das Ende des Zufalls. Wie Big Data uns und unser Leben vorhersagbar macht*. (Salzburg: Ecowin, 2013).
- Kurz, Constanze / Rieger, Frank: *Die Datenfresser. Wie Internetfirmen und Staat sich unsere persönlichen Daten einverleiben und wie wir die Kontrolle darüber zurückerlangen*. (Frankfurt / Main: Fischer, 2011).
- Lauer, Gerhart: *Die Digitale Vermessung der Kultur*. Essay in: *Big Data. Das neue Versprechen der Allwissenheit*. (Berlin: Suhrkamp, 2013).
- Lewis, Michael: *Moneyball. The Art of Winning an Unfair Game*. (New York: Norton&Company, 2004).
- Loukides, Mike: "What is data science?" *O'Reilly Radar* (2.6.2010). (Im Netz unter <http://radar.oreilly.com/2010/06/what-is-data-science.html>).
- Manne, Geoffrey A. / Wright, Joshua D. "Google and the Limits of Antitrust: The Case Against the Case Against Google". *Harvard Journal of Law & Public Policy*, 2013 (Im Netz unter: <http://www.harvard-jlpp.com/wp-content/uploads/2013/10/ManneFinal.pdf>).

Monopolkommission: *Sondergutachten 68. Wettbewerbsrecht: Herausforderung digitale Märkte*, 2015. (Im Netz unter www.monopolkommission.de/index.php/de/gutachten/sondergutachten/sondergutachten-68)

Moorstedt, Tobias: *WWWissenschaft. Ein Gespräch mit Cameron Marlow, dem Haussoziologen von Facebook*. Essay in: *Big Data. Das neue Versprechen der Allwissenheit*. (Berlin: Suhrkamp, 2013).

Pentland, Alexander: „Reality Mining“. Beitrag in *Technology Review by MIT*, 2008. (im Netz unter <http://web.media.mit.edu/~sandy/tr10pdfdownload.pdf>).

Piketty, Thomas: *Das Kapital im 21. Jahrhundert*. (München: C.H.Beck, 2014)

Preis, Tobias / Moat, Helen Susannah / Stanley, H. Eugene: „Can Google predict the stock market?“. *Scientific Reports* **3**, Article number: 1684 (2013). (Im Netz unter <http://www.nature.com/articles/srep01684#introduction>).

Schmidt, Eric / Cohen, Jared: *Die Vernetzung der Welt. Ein Blick in unsere Zukunft*. (Jürgen Neubauer (ed); Reinbek: Rowohlt, 2013).

Swan, Melanie: „Philosophy of Big Data“. (Paper aus dem Jahr 2014, im Netz unter http://www.melanieswan.com/documents/Philosophy_of_Big_Data_SWAN.pdf).

Varian, Hal R. : "Beyond Big Data". (Paper aus dem Jahr 2013).

Varian, Hal R.: *Economics of information technology*. (Manuskript einer Vorlesung aus dem Jahr 2001, überarbeitet 2003).

Varian, Hal R. : "Our Secret Sauce". (Post im Google-Blog aus dem Jahr 2008, im Netz verfügbar unter <https://googleblog.blogspot.de/2008/02/our-secret-sauce.html>).

Welter, Patrick: „Monopol aus eigener Leistung“. Post im FAZ-Blog *Fazit vom 1.4.2013* (im Netz unter <http://blogs.faz.net/fazit/2013/04/01/monopol-aus-eigener-leistung-1349/>).

Weinberger, David: *Die digitale Glaskugel*. Essay in: *Big Data. Das neue Versprechen der Allwissenheit*. (Berlin: Suhrkamp, 2013).

Weiser, M: "The Computer for the 21st Century". *Scientific American* 265(9):66–75 (1991).