Im Schatten von Big Data?

Die Sozialwissenschaften im Wandel

Jochen Hirschle

Beitrag zur Veranstaltung »Empirische Forschung über geschlossene Gesellschaften« der Sektion Methoden der empirischen Sozialforschung

Einleitung: Algorithmen und Theorien

Unter Informatiker/-innen, die an Übersetzungsalgorithmen für Microsoft arbeiteten, kursierte angeblich eine Zeitlang der Scherz, dass mit jedem Linguisten, der das Übersetzungsteam verlässt, die Übersetzungen besser würden (Mayer-Schönberger, Cukier 2013: 142). Ob dieser Scherz zu den Legenden zählt, die für Data Science Bücher erfunden wurde, oder aus der Realität stammt, ist nicht überliefert. Wahr ist allerdings, dass Übersetzungen bei Microsoft, IBM oder Google heute als Mathematikprobleme aufgefasst werden, und dass Informatiker/-innen, die über keinerlei sprachliche oder sprachwissenschaftliche Spezialkenntnisse verfügen, Algorithmen schreiben, die beliebige Sätze in beliebige Sprachen übersetzen, solange man sie nur mit hinreichend großen Mengen bereits übersetzter Texte füttert.

Chris Andersons Artikel *The End of Theory* (2008), der, wie der Titel impliziert, das Ende des Zeitalters der Theorien einläuten will, ist aus dieser Perspektive nicht mehr als ideologischer Überbau gängiger Praxis. In einer Anleitung einer Text-Mining Anwendung in Python[1], wird zum Beispiel im Detail der Code eines Algorithmus beschrieben, mit dessen Hilfe sich das Geschlecht von Vornamen bestimmen lässt (Bird, Klein, Loper 2009: 245f.). Dazu bedarf es keiner kulturellen Vorbildung, sondern nur eines Trainingsdatensatzes, der eine hinreichend große Menge an Vornamen mit Geschlechterzuordnung (männlich/weiblich) enthält, und eines circa 10-zeiligen Codes, der eine Programmfunktion generiert, die durch die verschiedenen Buchstaben jedes Namens iteriert, deren Position bestimmt, und dann mit Hilfe eines naiven Bayes-Klassifikators prüft, welche Buchstaben an welchen Positionen mit welcher Genauigkeit das Geschlecht prognostizieren. Danach kann der Algorithmus auf eine Liste weiterer Namen angesetzt werden, die er mit einer gewissen Fehlertoleranz den Kategorien männlich/weiblich zuordnet.

[1] *Python* ist eine unter Data Scientists weitverbreitete Programmiersprache, die insbesondere zur Bearbeitung prozessgenerierter Daten eingesetzt wird (vgl. Grus 2016).

Dieses Prinzip lässt sich auf eine Vielzahl anderer Anwendungen übertragen. Die Frage, ob die als Prädiktoren gewonnenen Informationen theoretisch sinnvoll mit den zu errechnenden Zielinformationen zusammenhängen, ist dabei Nebensache. Deshalb handelt es sich streng genommen auch um keine induktive Vorgehensweise. Es geht überhaupt nicht um die Erzeugung von Theorien. Es geht lediglich um die Aufdeckung nutzbarer Korrelationen. Eine Theorie oder eine allzu präzise Hypothese – etwa die über mögliche Eigenschaften von Vornamen, die Auskunft über das Geschlecht geben – würde die Algorithmen nur unnötig einschränken und die Errechnung besserer Lösungen, die im Rahmen der Theorie nicht vorgesehen sind, verhindern. Im Übrigen soll der Algorithmus die Kriterien, nach denen er verfährt, selbstlernend, mit jedem (in Echtzeit) eintreffenden weiteren Trainingsdatensatz verbessern. Eine statische Theorie ist in einem solchen dynamischen System als Endprodukt nicht vorgesehen.

Die Daten zum Sprechen bringen

Bis vor wenigen Dekaden waren auswertbare Informationen über das Verhalten von Individuen noch Mangelware. Sie zu erfassen war mit hohem Aufwand verbunden, mit der Durchführung von Umfragen, in denen Fragebögen ausgedruckt, gefaltet, kuvertiert und postalisch verschickt oder Hausbesuche unternommen werden mussten, der Durchführung von Experimenten in Laboren und Beobachtungen im Feld. Unter diesen Bedingungen musste Forschung zielgenau sein, sie musste die hohen Kosten der Erhebung im Rahmen halten und mit möglichst wenigen Informationen auskommen, um die Vielzahl an Vermutungen über Zusammenhänge auf den Prüfstand zu stellen.

Man könnte auf die Idee kommen, die Poppersche Wissenschaftstheorie als Resultat dieser (materiellen) Rahmenbedingungen zu betrachten. Wie man sich erinnert, lehnt Popper die induktive Methode mit dem Argument ab, dass Beobachten ein aktiver Vorgang sei, der niemals theorielos erfolgen könne. Jede Beobachtung, so Popper, sei ein aktiver Messvorgang, der bestimmte Informationen auswählt und andere ausschließt (Popper 1935/1989). Beobachten heißt: Informationen gemäß dem Raster einer nur impliziten (wie im Alltag) oder expliziten Hypothese (wie in der Wissenschaft) selektieren.

Das Experiment ist der Prototyp dieser Vorgehensweise. Es besteht aus einem künstlichen Setup, das einzig dem Zweck dient, gezielt jene Informationen auszuwählen und zu messen, die für den Test der Hypothese notwendig sind. Die Survey-Forschung, die sich in der zweiten Hälfte des 20. Jahrhundert herausgebildet hat, ist eine abgewandelte Form dieses Ansatzes: das Fragebogeninventar stellt das Messinstrument dar, das die Konzepte der Hypothesen, die es zu testen gilt, selektiv erhebt.

Weder die Durchführung eines Experiments noch die Erhebung von Umfragedaten sind ohne Vorannahmen sinnvoll. Wenn die Erhebung bereits weniger Daten mit so hohem Aufwand verbunden ist, müssen Informationen gezielt zum Zweck der Prüfung einer Theorie gesammelt werden. Endzweck jedes empirischen Tests sind nicht statistische Auswertungen, sondern Rückschlüsse auf den Wahrheitsgehalt einer Theorie, die anschließend – unabhängig von weiteren empirischen Daten – für Prognosen über Zusammenhänge in der realen Welt genutzt werden können.

Big Data ist im Vergleich ein Nebenprodukt menschlicher Handlungen oder maschineller Vorgänge – digitale Spuren des Lebens, die nicht gezielt erhoben werden. Wer eine Seite im Internet aufruft, einen Begriff in eine Suchmaschine eingibt, ein Produkt bestellt oder Zeitung liest oder auch nur die SIM-Card in ein Smartphone einsetzt und aktiviert, sendet vielfältige Informationen, die registriert und in Datenbanken hinterlegt und abgespeichert werden. In einem solchen umfassenden Registrierungs-

system sind Daten keine Mangelware mehr. Sie stehen im Überfluss zur Verfügung, immer und überall, ohne dass sie jemand bewusst erhoben hätte.

Das sind die Rahmenbedingungen, unter denen empirische Arbeit im Big Data Zeitalter betrieben wird: Myriaden digitalisierter Informationen, die mehr oder weniger strukturiert auf die Datenbanken der Server strömen.

Von nun an geht es um Techniken der Datenauswertung. Es geht um Nutzbarmachung, um Ansätze, mit denen sich aus gesammelten Informationen verwertbare Erkenntnisse ableiten lassen. Daten sind Ausgangspunkt und nicht mehr Durchgangsstation der Forschung. Sie müssen zum Sprechen gebracht werden.

Auch wenn Analysen ohne Vorannahmen, wie Popper richtig feststellt, unmöglich sind, kann man diese Vorannahmen doch auf ein Minimum reduzieren. Algorithmen, wie der oben skizzierte, zur Bestimmung des Geschlechts von Vornamen, lösen diese Aufgabe, indem sie eine Vielzahl naheliegender, aus dem verfügbaren Datenmaterial ableitbarer Thesen generieren – zum Beispiel, die These, dass der Buchstabe *a* an der letzten Stelle eines Vornamens aussagekräftig im Hinblick auf das Geschlecht von Vornamen ist. Diese Annahmen werden im gleichen Moment aufgestellt und überprüft und dann verworfen oder angenommen. Algorithmen bilden temporäre Hypothesen aus, die notwendig sind, um ein empirisches Testverfahren zu füttern.

Es stimmt schon, dass diese Methode die Gesetze der Wissenschaft nicht außer Kraft setzt. Algorithmen bieten keine Alternative, um umfassende Theorien induktiv aus Daten abzuleiten. In dieser Hinsicht bleibt alles beim Alten: Die Genese von Theorien und deren Überprüfung kann nicht automatisiert werden. Das Problem für die Wissenschaft besteht allerdings darin, dass Theorien unter den beschriebenen Rahmenbedingungen obsolet werden. Wenn sich funktionale Erkenntnisse zielgenau ohne umfassendere theoretische Vorannahmen aus den Daten ableiten lassen, wofür sollte man dann überhaupt an der Aufstellung von Theoriegebäuden festhalten (Anderson 2008)?

Theorien sind, grob gesagt, vereinfachte Vorstellungen über Ausschnitte der Welt und ihre Zusammenhänge. Über den reinen Erkenntniswert hinaus, lassen sich aus ihnen mit einer gewissen Genauigkeit Prognosen über Vorgänge in der Welt ableiten. In den exakten Wissenschaften ist die Genauigkeit solcher Theorien für Prognosen schon immer deutlich höher gewesen als in den Sozialwissenschaften. Trotzdem waren Theorien für die Vorhersage sozialen Verhaltens lange Zeit unverzichtbar.

Eine Klassentheorie postuliert, dass die materiellen Lebensbedingungen darüber entscheiden, welche Einstellungen und Ideen Personen über die Welt ausbilden, die sich in konkreten Praktiken, etwa dem Wahlverhalten – „Arbeiter wählen SPD“, „Unternehmer wählen FDP“ – niederschlagen. In der Markt- und Meinungsforschung wurden solche und davon abgeleitete Ansätze wie Schicht- und später Milieumodelle lange Zeit zur Bestimmung von Zielgruppen herangezogen. Messbare Indikatoren wie Einkommen, Bildung oder Wohnort wurden als Indikatoren für die Zugehörigkeit zu bestimmten sozialen Gruppe herangezogen und die eingruppierten Haushalte dann mit Produkthinweisen versorgt, von denen man annahm, dass sie die Mitglieder dieser Gruppe präferieren. Diese Vorgehensweise ist, auch wenn sie rein ökonomische Ziele verfolgt, theoretisch orientiert: sowohl die Kriterien der Zuordnung zu einem Milieu als auch die daraus abgeleiteten Maßnahmen verdanken sich bestimmten theoretischen Grundannahmen über soziale Milieus und ihrer Überprüfung mittels deduktiver Verfahren.

Mit der Ausbreitung von Big Data haben sich hingegen ganz andere Verfahren durchgesetzt. Wer bei Amazon einkauft, wird keinem Milieu, geschweige denn einer Klassenfraktion zugeordnet. Ein Algorithmus vergleicht das Einkaufsverhalten einzelner Personen mit anderen Personen mit ähnlichen Einkaufsbiographien und leitet daraus Produktempfehlungen ab. Dabei geht er, grob gesagt, von der banalen Annahme aus, dass ein Konsument A, der in der Vergangenheit eine Reihe gleicher Produkte

wie ein Konsument B gekauft hat, sich mit hoher Wahrscheinlichkeit auch für Produkte interessiert, die Konsument B aber nicht A gekauft hat. Um Bücher zu verkaufen, genügt es zu wissen, dass Kund/-innen, die ein Buch von Thomas Bernhard bestellt haben auch Bücher von Max Frisch oder Robert Musil lesen. Die Frage, ob sich dahinter ein bestimmtes Sozialisationsmilieu oder eine Klasse verbirgt, die das Kräftefeld eines sozialen Raums bevölkert (Bourdieu 1996), ist für Amazon unerheblich.

Auch wenn die quantitative Sozialforschung durch die zunehmende Verbreitung von Sekundärdaten sich in den letzten Jahrzehnten in die Richtung einer Variablensoziologie entwickelt hat, in der Theorien allenfalls eine Nebenrolle spielen, sollte man den wissenschaftstheoretischen Sprung von einer theoriearmen Hypothesenforschung zur „Agnostischen Theorie" der Data Sciences nicht unterschätzen (Anderson 2008). Selbst Variablensoziolog/-innen kommen nicht umhin, einzugestehen, dass empirische Variablen nur Operationalisierungen theoretischer Konzepte sind (Esser 1996). Sie müssen Hypothesen in Form von Indikatoren operationalisieren und die Ergebnisse empirischer Analysen auf die Ebene der Hypothesen rückübersetzen.

Die Verfahren, die sich im Zuge der neuen Data Sciences ausbilden, negieren diesen Unterschied gänzlich. Kausalitäts- werden zugunsten von Korrelationsanalysen aufgegeben (Mayer-Schönberer, Cukier 2013). Implizit macht sich auf diese Weise ein Empirismus breit, in dem die Grenzen zwischen theoretischen Konstrukten, Hypothesen und empirischen Indikatoren verschwimmen. Informationen, die nicht erhoben werden oder nicht erhebbar sind, werden als nicht existent betrachtet. Die Hauptaufgabe dieser funktionalen Wissenschaft besteht nicht darin, theoretische Modelle zu entwickeln und auf den Prüfstand zu stellen, sondern darin, empirische Daten in Beziehung zu setzen und daraus konkrete Handlungsmaßnahmen abzuleiten (Pentland 2014).

Anders als in der Wissenschaft, waren in den anderen Subsystemen der Gesellschaft Theorien schon immer nur notwendiges Übel. Unternehmen leiden kaum unter der Tatsache, dass Theorien an Bedeutung verlieren. Ganz im Gegenteil, unter den gegebenen Bedingungen, in denen Daten wichtiger als Theorien sind und die Daten in den Serverkellern privater Anbieter und nicht in den Laboren der Universitäten zusammenfließen, können Unternehmen unabhängig von den Erkenntnissen, die in den Wissenschaftsanstalten produziert werden, operieren. Die theoretische Praxis der Soziolog/-innen wird auf diese Weise, mehr denn je, zum Relikt einer erkenntnisgetriebenen Wissenschaft, die keine Außenwirkung mehr produziert.

Alles halb so schlimm?

Während in England Savage und Burrows bereits 2007 *The Coming Crisis of Empiricial Sociology* prognostiziert haben, scheint die deutsche Soziologie in ihrer wissenschaftlichen Praxis bis heute relativ unberührt von den Folgen der Big Data Revolution zu sein. Woran liegt das? Bestimmt nicht daran, dass die Verarbeitung von Internetdaten bereits selbstverständlicher Teil des methodischen Lehrprogramms der deutschen Soziologie geworden wäre. Aber ist die etablierte empirische Soziologie tatsächlich so unantastbar, so unersetzlich, dass sie um ihren Status nicht zu fürchten bräuchte?

Für diese Hypothese spricht vor allem die Tatsache, dass sich in der quantitativen Sozialforschung mit der repräsentativen Stichprobe ein über die Grenzen der Disziplin hinaus anerkanntes methodisches Instrument zur Erhebung von Informationen über soziale Aggregate herausgebildet hat. Repräsentative Stichproben ermöglichen es, Rückschlüsse auf Grundgesamtheiten zu ziehen, die die Größe der Stichprobe um ein Vielfaches überschreiten. Aus 2.000 Befragten können mit einer vergleichsweise geringen Fehlertoleranz Prognosen über das Wahlverhalten von circa 60 Millionen Deutschen abgelei-

tet werden. 2.000 Befragte aus einer repräsentativen Stichprobe ergeben ein weitaus präziseres Bild der Bevölkerung als zehn Millionen Datensätze aus Facebook. Selbst wenn sich aus einzelnen Facebook-Einträgen die Wahlwahrscheinlichkeit einer Partei mit hinreichend hoher Genauigkeit ableiten ließe, wären die Aussagen dieser Daten im Hinblick auf das tatsächliche Wahlverhalten der Bevölkerung wertlos. Deutsche Facebook-Nutzer/-innen ergeben nur ein äußerst verzerrtes Abbild aller Wahlberechtigten und ihre Präferenzen sagen deshalb wenig über die Präferenzen aller Wahlberechtigten aus.[2] Das gleiche gilt für andere deskriptive Studienziele, zum Beispiel die Eruierung der Entwicklung der Einkommensungleichheit in einer Gesellschaft, die relativen Anteile der Einwohner/-innen eines Landes, die religiös sind oder in die Kirche gehen. Repräsentative Umfragedaten, so aufwändig ihre Erhebung im Vergleich zum Umfang gewonnener Informationen auch sein mag, sind für die zuverlässige Prognose gesellschaftlicher Entwicklungen unentbehrlich. Und das womöglich noch auf lange Zeit hinaus.

Wie aber steht es mit Studien, in denen es nicht um Deskriptionen, sondern um den Test von Hypothesen über Zusammenhänge geht – Studien also, die den größten Teil der normalen Wissenschaft der quantitativ orientieren Sozialforschung ausmachen? Auch hier wird oftmals auf die Notwendigkeit der Repräsentativität der zum Test verwendeten Stichproben verwiesen. Ein Blick in die experimentelle Psychologie verdeutlicht indessen, dass sozialwissenschaftliche Repräsentativität in dieser Disziplin keine Rolle spielt. Trotzdem werden Hypothesen empirisch überprüft und Erkenntnisse verallgemeinert, ohne dass der wissenschaftliche Status dieses Faches in Frage gestellt würde.

Offensichtlich ist Repräsentativität im Sinne der Sozialwissenschaften für Hypothesentests nicht zwingend erforderlich. Genau genommen spielt sie nur dann eine Rolle, wenn man aus den Ergebnissen Rückschlüsse auf eine angebbare, räumlich und zeitlich eingrenzbare Grundgesamtheit ziehen möchte, also ein künstliches Gebilde, wie die erwachsene Bevölkerung, die in den Grenzen eines Landes oder einer Region lebt, die Konsumentinnen und Konsumenten eines bestimmten Produkts oder die Mitarbeiterinnen und Mitarbeiter einer Firma. Für Hypothesentests, vor allem wenn es um Hypothesen geht, die auf der Mikroebene angesiedelt sind, ist diese räumlich-zeitliche Eingrenzung nur selten von Interesse. Selbst in den Sozialwissenschaften sind Beiträge, in denen die Ergebnisse solcher Studien explizit auf Grundgesamtheiten hin interpretiert werden, Mangelware.[3]

Das hat vor allem damit zu tun, dass Hypothesen die Vorstellung von Gesetzmäßigkeiten zugrunde liegt (vgl. Popper 1935/1989). Gesetzmäßigkeiten gelten ihrem Wesen nach jedoch nicht nur in einer räumlich abgrenzbaren Stichprobe, etwa der bayerischen, deutschen oder französischen oder taiwanesischen Bevölkerung, sondern auch in anderen Bevölkerungsgruppen und in einzelnen Subgruppen dieser Bevölkerungen. Stellt man Unterschiede zwischen solchen Gruppierungen fest, wird man womöglich nicht nach der Qualität der Stichprobe, sondern nach der Qualität der Hypothese oder der zugrundeliegenden Theorie fragen. Man wird zum Beispiel die Frage stellen, ob Unterschiede im Verhalten zwischen verschiedenen Bevölkerungsgruppen auf andere Ursachen zurückgehen, verschie-

[2] Hinzu kommt, dass es für Außenstehende kaum möglich ist, Stichproben aus Facebook zu ziehen, die ‚repräsentativ' für alle Facebook-Nutzer/-innen sind oder bestimmt nicht alle Facebook-Nutzer/-innen politische Ansichten kommunizieren, aus denen sich mit einer hinreichend hohen Genauigkeit deren Parteipräfenzen ableiten ließen.

[3] Und wenn solche Angaben gemacht werden, dann meist nicht als Qualitätskriterien, sondern als methodische Hinweise, die darlegen sollen, dass die Annahme der Hypothese vorläufig sei, weil sie mit Hilfe einer Stichprobe getestet wurde, die nur einen regional beschränkten Ausschnitt der Bevölkerung einbezieht.

denartige Gelegenheiten des Handelns zum Beispiel, Restriktionen oder spezifische Anreize, die im einen Kontext, nicht aber im anderen gegeben sind. In der Praxis der empirischen Sozialforschung steht der Generalisierung von Ergebnissen also häufig ein weitaus gravierenderes Problem entgegen als das der Repräsentativität der Stichprobe: die Verzerrung des ermittelten Zusammenhangs zwischen zwei Variablen durch Drittvariablen. Nicht umsonst setzt man in der Auswertung seit geraumer Zeit auf multivariate Verfahren, die Hypothesen unter kontrollierten Bedingungen auf den Prüfstand stellen, auch wenn sich die Ergebnisse solcher Analysen kaum mehr nachvollziehbar in Relation zu einer angebbaren Grundgesamtheit interpretieren lassen.[4] Man opfert also willfährig die Option der einfachen Verallgemeinerung auf eine Population zugunsten der Erkenntnis, ob der in der Hypothese behauptete Kausaleffekt innerhalb der Stichprobe überhaupt zweifelsfrei auftritt.

Empirische Soziologie oder Computational Social Sciences?

Man wird sich vielleicht die Frage stellen, warum es überhaupt dazu kommen sollte, dass die empirische Soziologie im Schatten von Big Data steht, wie der Titel des Beitrags andeutet. Selbst wenn in Unternehmen das Interesse an Theorien schwindet, muss sich die Wissenschaft davon nicht anstecken lassen. Es gibt genug Hinweise, dass Menschen nicht nur nach ökonomischem Nutzen, sondern auch nach Erkenntnissen um ihrer selbst willen streben.

Warum sollte die Soziologie also nicht die Chancen von Big Data nutzen und dadurch vielleicht sogar eine neue Stufe in der Hierarchie der Wissenschaften erklimmen? Womöglich können mit Social Media Daten sogar die Ideen und Träume der klassischen Soziolog/-innen endlich in die Tat umgesetzt werden. Waren nicht die französischen Gründungsväter, Emile Durkheim, Auguste Comte und Aldolphe Quetelet einer streng empiristischen Ausrichtung verpflichtet? Die Soziologie sollte doch eine positive Wissenschaft sein. Comte stellte sie sich als Sozialphysik vor, Quetelet versuchte aus statistischen Daten mittlere Tendenzen und den Durchschnittsmenschen abzuleiten. Selbst Durkheim sprach von sozialen Kräften, die ein Eigenlebten führten und real wie „kosmische Kräfte“ seien (Durkheim 1897/1997: 359). Er begriff die Soziologie als empirische Disziplin mit einem eigenen Gegenstand und eigenen sozialen Gesetzmäßigkeiten, die es nach dem Vorbild der Naturwissenschaften aufzudecken galt. Warum sollte man also die Möglichkeiten, die durch Big Data entstehen, überhaupt als Bedrohung für die Soziologie wahrnehmen? Warum nicht den Gesetzmäßigkeiten, die in sozialen Netzwerken herrschen, auf die Spur kommen oder mit den Einkaufsdaten von Amazon den Bourdieuschen Raum der Lebensstile im Detail neu zusammensetzen und auf den Prüfstand stellen?

Die Antwort ergibt sich aus zwei sehr praktischen Erwägungen, die Savage und Burrow (2007) bereits angeführt haben und die deshalb an dieser Stelle nur kurz rekapituliert werden:

(1) Zum einen geht es um die Ausbildung und die Grundkompetenzen von Soziolog/-innen. Zwar hat die quantitative Sozialforschung über die Zeit neue Verfahren und Techniken entwickelt und in der soziologischen Lehre umgesetzt, so dass Studierende und Forschende heute Surveydaten wie den *Allbus*, den *World Value Survey* oder den *European Social Survey* mit Hilfe teils komplexer statistischer

[4] Meist kann man solche Ergebnisse (zum Beispiel bei multivariaten Regressionen) nur noch im Rahmen des methodischen Verfahrens, das heißt, unter Rekurs auf die anderen inkludierten Variablen interpretieren – etwa in der Art: unter der Bedingung, dass alle anderen in der Regression berücksichtigten Variablen konstant gehalten werden, hängt das Einkommen mit der Religiosität mit dem Faktor x zusammen.

Verfahren bearbeiten und auswerten können. Allerdings muss man bedenken, dass die gängigen, in der Soziologie verwendeten Statistikprogramme wie *SPSS* oder *Stata,* für Endverbraucher/-innen konzipiert sind. Sie lassen sich mit Hilfe des Menüs ohne besondere Programmierkenntnisse bedienen. Tabellenausgaben, Mittelwertsvergleiche, selbst Regressionen kann man über formschöne Bedienemente zusammenklicken. Zudem werden die Datensätze in unterschiedlichen Formaten, gelabelt und aufbereitet, versehen mit Missing Values und Gewichtungsfaktoren von öffentlichen Instituten wie der GESIS zur Verfügung gestellt. Selbst statistikaverse Forschende können nach einer Einweisung mit wenigen Handgriffen und Klicks einfache Analysen durchführen.

Social Media Daten, die im Internet entstehen, sind im Vergleich dazu weit weniger leicht analytisch konsumierbar. Wer mit Twitter-Daten arbeiten möchte, sollte sich mindestens mit einer Programmiersprache wie Python oder R auskennen, um die für die Suche oder Echtzeitsammlung notwenige API (Application Programming Interface) bedienen zu können. Auch liegen die Daten nach der Erhebung in einem mit Standardstatistiksoftware nicht lesbaren Format vor (JSON). Weitere Module und Codezeilen sind deshalb erforderlich, um die Dateien zu entschlüsseln und die für die Analyse relevanten Informationen in eine auswertbare Datenmatrix zu bringen.

Anschließend stellt sich die Frage, wie man die Inhalte analysiert. Die Auszählung der Häufigkeit des Vorkommens einzelner Wörter mag noch einfach sein und sich mit den gängigen Softwarelösungen zur Inhaltsanalyse durchführen lassen. Komplexere semantische Analysen erfordern dagegen den Einsatz weiterer Zusatzmodule wie *NLTK* oder *Textblob.* Erschwerend hinzukommt, dass sich die Verfahren zur Erhebung und Handhabung solcher Daten derzeit in der Entwicklung befinden. Sie ändern sich stetig, laufend werden neue Programmmodule konzipiert und zur Verfügung gestellt und alte eingemottet. Es ist äußerst unwahrscheinlich, dass sich kurzfristig standardisierte Verfahren entwickeln lassen, mit denen auch Soziolog/-innen mit basalen IT-Kenntnissen eigenständig solche Auswertungen durchführen können.

(2) Zum anderen sind die meisten Social Media Daten, anders als sozialwissenschaftliche Surveydaten, keine öffentlichen Güter. Sie werden weder von Universitäten noch von universitätsnahen Instituten erhoben. Genau genommen können solche Daten überhaupt nicht erhoben werden. Sie entstehen auf den Seiten und Portalen von Unternehmen wie Facebook, Google, Amazon, OKCupid oder Twitter als Nebenprodukte sozial-kommunikativer Tätigkeiten von Menschen und werden dort für die Öffentlichkeit unzugänglich gespeichert. Da solche Unternehmen wirtschaftliche Interessen verfolgen, stellen sie Datenauszüge, wenn überhaupt, nur zu ihren eigenen Bedingungen zur Verfügung. Die Tatsache, dass ein Großteil der wissenschaftlichen Publikationen, die Social Media Daten verwenden, derzeit auf Twitter-Nachrichten beruhen, hängt auch damit zusammen, dass Twitter eines der wenigen Unternehmen ist, das Forschenden überhaupt Zugang zu Stichproben gewährt.

Selbst wenn Soziolog/-innen also über die notwendigen IT-Kompetenzen verfügten, wären sie stets nur Zweitverwerter von Social Media Daten und dabei noch auf die Kooperation der Unternehmen angewiesen, auf deren Servern die Datensätze lagern.

Die methodisch dominierte Disziplin

Wer abschätzen möchte, wie stark die Soziologie als Fach von ihren Methoden abhängt, sollte womöglich einen Blick in die Vergangenheit und in James Colemans Beitrag *Social Theory, Social Research, and a Theory of Action* (1986) werfen. Coleman, der selbst maßgeblich zur mikrotheoretischen Wende in der Soziologie beigetragen hat, geht darin dezidiert auf deren methodischen Ursprünge ein. Er führt aus,

dass makro- und mesosoziologische Zugangsweisen, insbesondere der Strukturfunktionalismus Talcott Parsons', die Soziologie bis in die 1960er Jahre beherrschten. Das änderte sich erst mit der Herausbildung der Umfrageforschung. Und das, wie Coleman betont, ohne ersichtliche theoretische Not – ohne, dass die Soziologie sich in einem jener krisenhaften Zustände befunden hätte, die eine Neuausrichtung unabänderlich gemacht hätten (Kuhn 1976). Vielmehr wurde der makrotheoretischen Zugangsweise durch die Herausbildung der Umfrageforschung allmählich der empirische Nährboden entzogen. Dabei hat sich die Umfrageforschung zunächst ganz unabhängig von der Theoriebildung bzw. sogar gegen die dort vorherrschende Strömung etabliert:

> „[O]ne could say that as social theory was moving to a functionalism that remained at the collectivity level, the main body of empirical research was abandoning analysis of the functioning of collectivities to concentrate on analysis of the behavior of individuals." (Coleman 1986: 1316).

Erst später, als sich die neue Methode als Instrument der Sozialforschung durchgesetzt hatte, schwenkte auch die Theoriebildung um; vorwiegend aus praktisch-methodischen Gründen allerdings: Das Individuum bildet in Umfragen sowohl als Einheit der Informationserhebung als auch als Analyseeinheit den natürlichen Dreh- und Angelpunkt der Analyse. Umfragedaten produzieren, methodisch gesprochen, Varianzen zwischen Befragungspersonen, die es zu erklären gilt. Und so stieg allmählich die Nachfrage nach Mikrotheorien, die in der Lage sind, diese umfragerelevanten Artefakte zu erklären: Unterschiede im beobachteten Verhalten oder in den Einstellungen zwischen Individuen. Makrotheoretische Zugänge, die kollektive Phänomene ins Visier nehmen, gerieten dagegen mehr und mehr aus dem Blickfeld (Coleman 1986: 1315).

Das Beispiel verdeutlicht, wie sehr die Soziologie empirische Wissenschaft ist. Wie andere Disziplinen, wird sie in ihrer Entwicklung von methodischen Erfindungen getrieben. Warum nicht von solchen, die außerhalb der Grenzen der Disziplin ihre Ursprünge haben? Sie können ihr zu neuem Glanz verhelfen oder sie in Krisen stürzen. Den Scherz der Microsoft-Programmierer/-innen über die Rolle der Linguisten für Übersetzungen muss man womöglich nicht allzu ernst nehmen. Die Entwicklungen in den Data Sciences – die Nutzung von Big Data zur Analyse sozialen Verhaltens, die Relativierung von Theorien und die fortschreitende Technisierung und Rationalisierung empirischer Methoden schon. Sie sind schon jetzt zur gängigen Praxis geworden: außerhalb der Soziologie.

Literatur

Anderson, C. 2008: The end of theory. Wired, 23.08.2008, http://www.wired.com/2008/06/pb-theory/ (letzter Aufruf 08.02.2016).

Bird, S. Klein, E., Loper, E. 2009: Natural language processing with Python. Analyzing text with the natural language toolkit. Sebastopol: O'Reilly.

Bourdieu, P. 1996: Die feinen Unterschiede. Kritik der gesellschaftlichen Urteilskraft. Frankfurt am Main: Suhrkamp.

Coleman, J. S. 1986: Social theory, social research, and a theory of action. American Journal of Sociology, 91. Jg., Heft 6, 1309–1335. DOI: https://doi.org/10.1086/228423

Durkheim, É. [1897] 1997: Der Selbstmord. Frankfurt am Main: Suhrkamp.

Esser, H. 1996: What is wrong with 'variable sociology'? European Sociological Review, 12. Jg., Heft 2, 159–166. DOI: https://doi.org/10.1093/oxfordjournals.esr.a018183

Grus, J. 2016: Data Science from Scratch. Sebastopol: O'Reilly.

Kuhn, T. S. 1976: Die Struktur wissenschaftlicher Revolutionen. Frankfurt am Main: Suhrkamp.

Mayer-Schönberger, V., Cukier, K. 2013: Big Data. A revolution that will transform how we live, work and think. London: John Murray.

Pentland, A. 2014: Social Physics. New York: Penguin Press.

Popper, K. [1935] 1989. Logik der Forschung. Tübingen: JCB Mohr.

Savage, M., Burrows, R. 2007. The coming crisis of empiricial sociology. Sociology, 41. Jg., Heft 5, 885–899. DOI: https://doi.org/10.1177/0038038507080443