Lehre und Forschung

Simon Munzert

Big Data in der Forschung! Big Data in der Lehre?

Ein Vorschlag zur Erweiterung der bestehenden Methodenausbildung

1. Einleitung[1]

Seit einigen Jahren erleben die Sozialwissenschaften eine beinahe kopernikanische Wende: Spektrum und Umfang sozialwissenschaftlicher Daten nehmen rapide zu, Forschung an computerbasierten Methoden zur Klassifikation und Analyse bestehender großer Datenmengen erfährt einen disziplinenübergreifenden Boom. Der vielmals beklagte Mangel an Informationen über individuelles Verhalten oder Institutionen hat sich mittlerweile – zumindest in Teilen – in einen Überfluss verkehrt (King 2011). Dazu ein etwas vereinfachendes, aber doch eindrückliches Beispiel: Die im Rahmen der Deutschen Wahlstudie und des ALLBUS seit 1949 erhobenen Daten lassen sich zusammen problemlos auf einem Gigabyte Speicherplatz ablegen. Über die Social-Media-Plattform Twitter, die zunehmend in den Fokus wissenschaftlichen Interesses gerät, werden, über den Daumen gepeilt, etwa vier Gigabyte Daten generiert – pro Stunde.[2] Dies veranschaulicht nicht unbedingt den rasanten Zuwachs an politikwissenschaftlich relevanten Daten – über den Gehalt manch willkürlich zusammengetragener Tweet-Sammlung für politikwissenschaftliche Forschung lässt sich trefflich streiten –, wohl aber das gewaltige Potenzial neuer Datenquellen, ebenso wie die außerordentlichen technischen Herausforderungen,

1 Dieser Beitrag baut lose auf einem Vortrag auf, der im Mai 2014 auf der Jahrestagung der DVPW-Sektion Methoden unter dem Titel „Automatisierte Datenerhebung im World Wide Web – ein Vorschlag zur Erweiterung der künftigen Methodenausbildung“ gehalten wurde. Ich danke Holger Döring für hilfreiche Kommentare.

2 Die Schätzung basiert auf der Annahme von durchschnittlich 5.700 Tweets pro Sekunde, jeweils mit einer Länge von 140 Zeichen und ca. 200 Bytes Umfang (vgl. https://blog.twitter.com/2013/new-tweets-per-second-record-and-how, Stand: 6.8.2014).

vor der Forscher[3] stehen, die soziales Verhalten im World Wide Web untersuchen möchten.

Was bereits als Ära der *Big Data* aufgefasst wird, ist allerdings mehr als eine Ergänzung von „klein-N"- und „großes-N"-Studien um „sehr-großes-N"-Studien (Monroe 2013: 1). Über die vergangenen Jahre ist ein rasanter Fortschritt in den technischen Möglichkeiten zur Speicherung, Weiterverarbeitung und Analyse großer und dynamischer Datenströme zu beobachten. Dazu hat politikwissenschaftliche Forschung ihren Teil beigetragen.[4] Auch im deutschsprachigen Raum steht *Big Data* mittlerweile auf der politikwissenschaftlichen Forschungsagenda. So adressierte die DVPW-Sektion Methoden im Rahmen der Jahrestagung 2014 unter dem Titel „Big Data? New Data!" Chancen und Herausforderungen dieser Datenrevolution: Dabei wurde unter anderem demonstriert, wie Twitter zur Echtzeitmessung öffentlicher Meinung bei TV-Duellen zwischen Spitzenkandidaten genutzt werden kann (Nyhuis/Faas 2014), welches Potenzial umfangreiche Textmengen in Form von Plenarprotokollen des Deutschen Bundestags bergen, um parlamentarisches Verhalten mit statistischen Mitteln zu beschreiben (Blätte 2014), oder wie Web-2.0-Technologien dabei helfen, einen internetbasierten Katalog politischer Parteien zu erstellen (Döring/Regel 2014).

Trotz dieser Vielfalt an Beiträgen gilt aber: Die technische Expertise zum Umgang mit großen Datenmengen und neuen Datenquellen gehört nicht zum Standard substanziell forschender Politikwissenschaftler. Sie muss entweder mühsam selbst erlernt oder im Rahmen interdisziplinärer Forscherverbünde extern eingeholt werden. Dass sich daran auf absehbare Zeit etwas ändert, zeichnet sich bis jetzt nicht ab: Die gegenwärtige Methodenausbildung an deutschen Hochschulen befähigt die Studierenden der Politikwissenschaft noch nicht zum Umgang mit derartigen Daten – weder technisch noch konzeptionell.

In diesem erörternden Beitrag argumentiere ich für eine Anpassung von Methodenlehrplänen an die skizzierten neuen Entwicklungen des massiven Zuwachses neuer Datenquellen und Analysetechniken. Dazu umreiße ich im folgenden Abschnitt zunächst Chancen und Herausforderungen der Datenrevolution aus Forscherperspektive. Im nächsten Schritt erörtere ich, welche Kompetenzen zur wis-

3 Aus Gründen der Lesbarkeit verwende ich in diesem Beitrag nur die männliche Form. Sämtliche Formulierungen schließen jedoch Frauen ein. Wie die Jahrestagung der DVWP-Sektion Methoden leider ebenfalls gezeigt hat, sind Forscherinnen auf diesem Feld noch unterrepräsentiert.

4 Vergleiche auch das Virtual Issue in *Political Analysis* zu „Big Data in Political Science" sowie das kürzlich erschienene Sonderheft zu „Quantifying Politics Using Online Data" im *Social Science Computer Review*, Jahrgang 32, Ausgabe 2 (2014), ebenso das Symposium „Technology, Data, and Politics", erschienen in *PS: Political Science & Politics*, Jahrgang 46, Ausgabe 2 (2013).

senschaftlichen Nutzung dieser Daten benötigt werden und inwiefern diese bereits jetzt zum Methodenkanon gehören. Danach berichte ich aus Lehrveranstaltungen, die die Methodenausbildung an der Universität Konstanz in den letzten Jahren mit Blick auf web-basierte Datenerhebung erweitert haben. Abschließend präsentiere ich einen Vorschlag, wie die künftige Methodenausbildung an politikwissenschaftlichen Fachbereichen und Instituten in den nächsten Jahren erweitert werden könnte, um zwei Anforderungen gleichermaßen zu erfüllen: die Vermittlung neuer technischer Kompetenzen einerseits, bei gleichzeitiger Einbindung in die bestehenden, straffen Studienpläne andererseits. Zusammengefasst lautet meine Argumentationslinie wie folgt:

1. Neue Datenformen und Analysemethoden bergen ein immenses Potenzial für politikwissenschaftliche Forschung.
2. Angehende Forscher müssen frühzeitig zum Umgang mit neuen Datenquellen befähigt werden.
3. Die Heterogenität und Masse an neuen Datenformen erfordert einen routinierten Umgang mit skriptbasierter Statistiksoftware, außerdem methodische und statistische Kenntnisse, wie sie grundlegend bereits vermittelt werden.
4. Technisches Know-how muss in der Methodenausbildung gestärkt werden, insbesondere durch verpflichtende Software-Kurse und ergänzende Kurse zu Informationstechnologien (Web-Scraping-Techniken, Datenbankmanagement, Data- und Textmining).

2. Chancen und Herausforderungen der Datenrevolution

Um Chancen und Herausforderungen von *Big Data* für die Profession zu erörtern, scheint es eigentlich zunächst sinnvoll, sich um eine exakte Definition des Gegenstands zu bemühen. Das ist jedoch, so vielfältig wie der *Big-Data*-Begriff inner- und außerhalb der Forschung genutzt wird, weder sonderlich einfach noch hilfreich. Anstelle eines weiteren Versuchs der Begriffsfassung greife ich deshalb im Folgenden auf einen bestehenden Vorschlag zurück, die relevanten Entwicklungen der letzten Jahre zu bündeln, und illustriere Chancen und Herausforderungen dieser Entwicklungen anhand bemerkenswerter jüngerer Forschungsbeiträge. Monroe (2013) fasst, in Anlehnung an Laneys (2001) „Three Vs of Big Data", Chancen und Herausforderungen von *Big Data* für die Politikwissenschaft in fünf „Vs" zusammen: Volumen, Verbreitungsgeschwindigkeit, Vielfalt, Verknüpfung und Validität.[5] Diese Differenzierung ist nützlich, weil sie verdeutlicht, dass die Entwicklun-

5 Im Original: "[V]olume, velocity, variety, vinculation, and validity" (Monroe 2013: 1).

gen der letzten Jahre mehr umfassen als die plötzliche Verfügbarkeit von Datenmengen ungekannten Ausmaßes, und auf bislang zu wenig beachtete Probleme hinweist.

a) Chancen in den „fünf Vs"

Die Massennutzung sozialer Medien zur Kommunikation, Netzwerkpflege und Meinungsäußerung und das damit verbundene Daten*volumen* birgt aus politikwissenschaftlicher Sicht gewaltiges Potenzial – sie bietet eine wertvolle, in vielen Fällen nie dagewesene Grundlage für eine empirische Mikrofundierung existierender Theorien politischen Handelns, aber auch für die Entwicklung neuer Methoden zur Messung politischer Einstellungen. Darüber hinaus erlauben stark gestiegene Rechnerkapazitäten, existierende Datenquellen zusammenzuführen und gemeinsam auszuwerten. Neuere umfragebasierte Wahlvorhersageansätze, wie sie zuletzt bei der US-Präsidentschaftswahl angewandt wurden, sind nicht zuletzt deshalb so erfolgreich, weil sie auf Massen an veröffentlichten Meinungsumfragen zurückgreifen und diese effizient bündeln (Silver 2012; Linzer 2013).

Daten aus sozialen Netzwerken, der Nutzung von Kreditkarten oder Smartphones, Suchmaschinenabfragen und vielen weiteren Diensten werden heute automatisiert und in Echtzeit erfasst und zugänglich gemacht – in einer zuvor ungekannten *Verbreitungsgeschwindigkeit*. Solche Datenströme erlauben die Analyse von Ereignissen, die mit klassischen Erhebungsmethoden kaum zu erfassen sind, zum Beispiel spontan ausbrechende Aufstände oder Reaktionen auf Katastrophen, und können Dynamiken politischer Phänomene besser erfassen als beispielsweise Befragungen, die öffentliche Meinung vergleichsweise statisch abbilden. So nutzt Mellon (2013) den Service Google Trends zur Messung von Issue-Salienz, die sonst nur in einem groben zeitlichen Raster und beschränkt auf wenige thematische Felder mit Most-Important-Problem-Batterien zu messen ist. Shaw und Hill (2014) analysieren Paradaten aus über 600 Online-Wikis, um Organisationsdynamiken zu beobachten, wie sie bereits 1915 in Robert Michels „ehernem Gesetz der Oligarchie" postuliert wurden. King/Pan/Roberts (2013) schließlich erheben wiederholt Inhalte aus Millionen Social-Media-Nachrichten auf chinesischen Plattformen und spüren damit die der Zensur durch die Behörden wesentlich zugrundeliegende Logik auf, nämlich die Unterdrückung von Nachrichten, die Mobilisierungspotenzial bergen. Diese Nachrichten müssen schnell und wiederholt abgerufen werden, um die Spuren der Zensur aufzudecken.

Die *Verknüpfung* unterschiedlicher Datenquellen birgt großes Potenzial für sozialwissenschaftliche Forschung. Menschliches Verhalten findet immer in einem

sozialen Umfeld statt: Politische Einstellungen bilden sich in der Familie und Netzwerken, Wahlentscheidungen werden nach einem Austausch mit Freunden und Bekannten getroffen, Konflikte zwischen Gruppen oder Staaten ausgetragen. Soziale Netzwerke bilden Bekanntschaften unmittelbar ab und liefern dabei deutlich vollständigere Bilder als beispielsweise berichtete Freundschaftsnetzwerke in standardisierten Befragungen. Dabei veröffentlichte Informationen sind vielfältig nutzbar. Barberá (im Erscheinen) entwickelt eine Methode, um aus Twitter-Follower-Netzwerken, die viele Hunderttausend Nutzer umfassen, Idealpunkte unterschiedlicher politischer Akteure zu schätzen. Darüber hinaus haben kommerzielle und nichtkommerzielle Organisationen das Potenzial der Verknüpfung von Daten erkannt, um beispielsweise Käuferprofile anzulegen oder zielgenau Spenden zu akquirieren. Auch solche Daten sind bereits politikwissenschaftlich ausgebeutet worden: So greift Bonica (2013) auf einen Datensatz zurück, der über 100 Millionen Spendeneingänge im Zuge US-amerikanischer Wahlen aufführt, um Positionen für Kongressmitglieder, die Exekutive, Interessengruppen und Wähler in einem ideologischen Raum zu schätzen.

Die *Vielfalt* neuer Datenquellen spiegelt sich in den bereits zitierten Arbeiten. Ein häufiges Merkmal neu erschlossener Datenquellen ist die zweckentfremdete Nutzung zur Messung menschlichen Verhaltens, wie bereits in den beiden vorigen Beispielen beschrieben. Eine weitere Anwendung, die existierende Daten kreativ ausnutzt, präsentieren Kuhn und Weidmann (im Erscheinen): Sie nutzen hochauflösende Satellitendaten zur Lichtstärkemessung (Chen/Nordhaus 2011; Henderson/Storeygard/Weil 2011), um – in Verbindung mit geokodierten Daten über ethnische Siedlungsmuster – ein Proxy-Maß für Ungleichheit innerhalb ethnischer Gruppen zu entwickeln. So lassen sich Untersuchungen auf geografische Gebiete ausweiten, für die keine anderen reliablen Datenquellen vorliegen.

b) Herausforderungen

Gleichzeitig bergen große, sich rapide erneuernde und unstrukturierte Datenmengen eine Reihe technischer und konzeptioneller Herausforderungen. Auf das Speicherproblem wurde bereits hingewiesen – klassische Datenmanagement- und Analysestrategien (Datensatz einlesen, aufbereiten, analysieren) stoßen schnell an ihre Grenzen, wenn die Masse der Daten die Kapazität des Arbeitsspeichers überschreitet. Darüber hinaus erfordern unstrukturierte Datenquellen (z. B. HTML-Code) oder Datenaustauschstandards, die in den Sozialwissenschaften noch wenig bekannt sind (z. B. XML und JSON), spezielles Wissen zur Datenextraktion; textbasierte Ressourcen außerdem Erfahrung im Umgang mit Textmining-Techniken.

Aus einem *anything goes* der Datenerhebung, nach dem praktisch jede verfügbare Information, die hinsichtlich einer politikwissenschaftlichen Fragestellung relevant erscheint, auch erhoben und genutzt werden kann, resultiert die Frage nach der *Validität* neuer Daten. Sozialwissenschaftler haben sich über die letzten Jahrzehnte Fähigkeiten angeeignet, um vertraute Datenquellen qualitativ beurteilen zu können. Selektions- und Messfehlerproblematiken in Umfragedaten, aber auch beispielsweise in populären ländervergleichenden Datensätzen sind bekannt und können unter Umständen mit geeigneten statistischen Mitteln angegangen werden. Im Umgang mit neuen Datenquellen ist diese Erfahrung noch nicht vorhanden. Wen oder was repräsentieren aus Twitter-Daten extrahierte und aggregierte Meinungen? Wie vollständig und sauber aufbereitet sind umfassende, von Dritten angelegte Datensätze? Verzerren unbeobachtete konfundierende Faktoren die Schätzungen auf Basis von Messungen zur Lichtemission aus Satellitendaten? Es bedarf deshalb sowohl starker Theorien als auch rigoroser Validierungsanstrengungen, um die Nutzung von häufig nicht-reaktiven Messungen menschlichen Verhaltens als Proxy für interessierende politikwissenschaftliche Konzepte zu rechtfertigen. Geschieht dies nicht, beraubt sich sozialwissenschaftliche Forschung selbst entscheidenden Erkenntnissen aus den vergangenen Jahrzehnten. Als das US-Magazin Literary Digest 1936 versuchte, mithilfe einer großangelegten Umfrage unter ihren Lesern den Gewinner der anstehenden Präsidentschaftswahl vorherzusagen, beteiligten sich etwa 2,3 Millionen Leser – heute würde man vermutlich *Big-Data*-Analyse dazu sagen. Trotzdem lag die Prognose – ein haushoher Sieg des Republikaners Landon – fatal daneben, was in erster Linie auf die Auswahl eines Teils der Befragungsteilnehmer auf Basis von Telefonregistern zurückgeführt wurde, denn: Telefonanschlüsse fanden sich damals primär in reichen, republikanischen Haushalten (Squire 1988). Deutlicher kann kaum veranschaulicht werden, dass die alleinige Stärke neuer Datenquellen nicht in ihrem schieren Umfang liegen kann. Zur sinnvollen Nutzung solcher Daten haben die empirischen Sozialwissenschaften deshalb qua Erfahrung einiges beizutragen.[6]

6 Aus Platzgründen nur angedeutet werden sollen weitere Herausforderungen wie beispielsweise forschungsethische Fragen, die mittlerweile rege diskutiert, aber noch nicht beantwortet sind. Dazu gehört die Frage, inwiefern Individualdaten aus sozialen Netzwerken für Forschungszwecke genutzt werden können, ohne die Nutzer darüber aufzuklären.

3. Welche Kompetenzen sollten vermittelt werden?

Um die künftige Generation Politikwissenschaftler frühzeitig zu befähigen, relevanten Forschungsfragen mit modernen Mitteln nachzugehen und damit die Chancen zu nutzen, die neuartige Datenquellen mit sich bringen, müssen die dafür notwendigen technischen Kompetenzen frühzeitig vermittelt werden. Teilweise sind diese bereits Bestandteil der methodischen Ausbildung (wie die klassische Methodenlehre und einführende Statistik), teilweise nicht. Im Folgenden illustriere ich Letztere.

a) Web-Technologien

Das *Big-Data*-Phänomen ist, insbesondere für die Sozialwissenschaften, auch und in erster Linie ein Webdatenphänomen. Gleichzeitig liegen entsprechende Daten häufig unstrukturiert vor. Viele Inhalte von Interesse sind nicht für wissenschaftliche Analysen aufbereitet, sondern richten sich an Webnutzer, wie etwa Blogs, Artikel in Online-Medien, Wikipedia-Einträge, Forenbeiträge, Präsenzen politischer Institutionen etc. Um Informationen aus Webseiten mittels Web Scraping oder API-Anbindung zu extrahieren, sind Grundkenntnisse der darunter liegenden Web-Technologien (HTTP, HTML, XML, JSON, AJAX, XPath und weitere) nötig. Die Nutzung dieser Technologien war essenziell für die oben beschriebene Forschung von Barberá (im Erscheinen), King/Pan/Roberts (2013), Mellon (2013), Shaw und Hill (2014). Es ist grundsätzlich davon auszugehen, dass soziale und politische Phänomene im Web weiter an Bedeutung gewinnen.

Grundwissen über diese Grundlagen der Web-Infrastruktur ist nicht nur aus technischen Gründen, sondern auch für substanziell interessierte Forscher wichtig. Es ist grundlegend ist für das Verständnis davon, wie Daten „zustande kommen", an welcher Stelle es zu möglichen Selektivitätsproblemen oder zu Ungenauigkeiten in der Erfassung und Aufbereitung solcher Daten kommen kann. Außerdem hilft technisches Verständnis bei der Beurteilung forschungsethischer Fragen, die sich aus den neuen Möglichkeiten der Datenerhebung ergeben, wie etwa: Ist Webseitenbetreibern der automatisierte Download von Ressourcen auf ihren Servern zuzumuten, die in erster Linie lediglich zur Betrachtung im Browser aufbereitet wurden? Welche Serverbelastung ist den Betreibern von Webseiten zumutbar, von denen automatisiert Daten erhoben werden sollen? Ist es ethisch vertretbar, mit menschengenerierten Daten Forschung zu betreiben, ohne dass explizit Einverständnis dazu gegeben wurde?

b) Datenbank-Technologien

Forschung mit umfangreichen Datenmengen, wie sie insbesondere im Zusammenhang mit web-basierten Untersuchungen entstehen, ist mit der klassischen „ein Forschungsproblem, ein Datensatz"-Logik oft nicht zu bewältigen. Zur Organisation und Analyse großer Datenstrukturen sind relationale Datenbanken ein nützliches, oft sogar notwendiges Mittel. Die Erstellung und Abfrage solcher Datenbanken zum Beispiel mittels SQL ist deshalb eine weitere wichtige Kompetenz, die es zu vermitteln gilt.

c) Machine-Learning-Techniken

Machine-Learning-Verfahren haben in den letzten Jahren große Aufmerksamkeit erfahren, da sie besonders gut dafür geeignet sind, in großen Datenmengen interessierende Muster zu erkennen, Dimensionalität zu reduzieren oder Daten in vorgegebene (oder nicht vorgegebene) Klassen zu gruppieren. Insbesondere, wenn es um die Klassifikation großer Textkorpora geht, deren schiere Masse nicht mehr vollständig per Hand kodiert werden kann, werden mittlerweile automatische und teilautomatische Klassifikations- und Skalierungsverfahren von Politikwissenschaftlern angewendet und auch weiterentwickelt (Laver/Benoit/Garry 2003; Slapin/Proksch 2008; Grimmer 2010; Hopkins/King 2010; für einen Überblick: Grimmer/Stewart 2013).

d) Schlüsselkompetenz Programmierkenntnisse

Web-basierte Daten sind in der Regel nicht statisch und abgeschlossen (wie beispielsweise Umfragedaten), sondern dynamisch, häufig unstrukturiert und voller Ungenauigkeiten. Datenerhebung und -bereinigung werden so zu repetitiven Prozessen, die am besten automatisiert werden sollten. Außerdem bieten sich bei der Auswertung solcher Daten häufig Visualisierungs- und statistische Analysemethoden an, die noch nicht über standardisierte Befehle abrufbar sind und eigenständige Programmierarbeit erfordern. Eine Schlüsselkompetenz im wissenschaftlichen Umgang mit web-basierten Datenquellen ist deshalb der versierte Umgang mit einer Programmiersprache, die solche Prozesse ermöglicht und den Datenerhebungs- und Aufbereitungsprozess somit reproduzierbar macht. Dazu gehören unter anderem Python und R.[7]

7 Kommerzielle Alternativen wie SPSS oder Stata sind hierfür trotz zugehöriger Skriptsprachen ungeeignet, da sie im Wesentlichen auf der Ein-Datensatz-Logik aufbauen und nicht für heterogene, kleinteilige Datenstrukturen gedacht sind.

Die Wahl der Software, deren Nutzung gelehrt wird, ist eine Grundsatzentscheidung. Die Vermittlung von Programmierkenntnissen und der Umgang mit Statistiksoftware ist kein Selbstzweck. Sie soll auf eigene Forschungsarbeiten vorbereiten und ist darüber hinaus eine wichtige Schlüsselqualifikation am Arbeitsmarkt. Weiterführende Methodenkurse sollten auf erworbenen Softwarekenntnissen aufbauen können. Das Erlernen einer Programmiersprache, insbesondere der ersten, verschlingt enorme zeitliche Ressourcen (Norvig 2001). Es ist Studierenden deshalb – ars longa, vita brevis – eigentlich nicht zumutbar, im Laufe ihres Studiums mehrere Programmiersprachen zu lernen. Stattdessen sollte der Fokus auf substanziellen und methodischen Themen liegen, die in direktem Bezug zur Disziplin stehen. Welche Statistiksoftware also lehren?

An deutschen politikwissenschaftlichen Fachbereichen und Instituten, die empirische Methoden im Lehrplan fordern, werden im Wesentlichen drei verschiedene Software-Suiten verwendet: SPSS, Stata und R. Während SPSS und Stata als eher einsteigerfreundlich und deshalb besonders lehrtauglich angesehen sind, zeichnet sich R dadurch aus, kostenlos verfügbar zu sein und als Open-Source-Software beständig erweitert zu werden. Mit Blick auf klassische Datenmanagement- und Analyseaufgaben ist die Wahl der Statistiksoftware mitunter eine Glaubensfrage. Für die Arbeit mit neuen, oft web-basierten Datenquellen ist das nicht der Fall. Für R frei verfügbare Erweiterungen wie das XML- oder das RCurl-Paket ermöglichen das Sammeln unstrukturierter (z. B. von klassischen HTML-Webseiten) oder strukturierter Daten (z. B. mittels APIs, also Schnittstellen, die von Web Services zur Verfügung gestellt werden), die Aufbereitung, Visualisierung, Analyse und Ankopplung an Datenbanksysteme. R fungiert als Programmierumgebung, die für jede erdenkliche Aufgabe genutzt werden kann. Damit geht die Funktionalität weit über die vorhandenen kommerziellen Statistikprogramme hinaus. Im selbständigen Umgang mit *Big Data* im Sinne von komplexen Datenstrukturen und unaufgeräumten Datenquellen ist dies nicht nur ein Vorteil, sondern in der Regel Voraussetzung.

e) Weshalb man technische Big-Data-Kompetenzen nicht alleine Informatikern überlassen sollte

Kann die politikwissenschaftliche Profession die technischen Details nicht denjenigen überlassen, die sich professionell damit auseinandersetzen? King (2014: 166) argumentiert, dass zukünftige Forschungspraxis in den Sozialwissenschaften zunehmend auf interdisziplinäre Zusammenarbeit aufbauen wird, da Wissen und Fähigkeiten im Umgang mit neuen Datenquellen vom Einzelnen nicht mehr umfassend gelernt werden können. Diese Prognose erscheint plausibel. Sollten wir uns deshalb

auf die Computerwissenschaften verlassen, wenn es um die Erschließung neuer sozialwissenschaftlicher Datenquellen in großem Umfang geht? Nein – genauso wenig, wie man sich auf Statistiker bei der Datenauswertung verlässt. Die Bearbeitung von Forschungsgegenständen setzt grundlegendes Verständnis darüber voraus, wie Daten generiert werden. Im Umgang mit Umfragedaten oder Länderstatistiken haben wir Politikwissenschaftler das gelernt. Bei der Erschließung neuartiger Datenquellen besteht die Chance, frühzeitig an der Entwicklung adäquater Analysemethoden mitzuwirken und die über viele Jahre erworbene Expertise im Umgang mit menschengenerierten Daten einzubringen, insbesondere die Erfahrung mit Selektions- und Messproblemen, die man in web-basierten Datensätzen verstärkt vermuten kann. Die Sensibilität für und die Kompetenz zum Umgang mit solchen Problemen ist eine Stärke sozialwissenschaftlicher gegenüber rein technischen Disziplinen, die in interdisziplinäre Zusammenarbeit eingebracht werden muss. Dafür braucht es jedoch technisches Grundverständnis. Und das muss bereits möglichst früh vermittelt werden.

4. Erfahrungen aus der Lehre

Im Folgenden möchte ich kurz eigene Erfahrungen zusammenfassen, die ich in der Lehre web-basierter Datenerhebungstechniken seit 2012 an der Universität Konstanz gemacht habe. Die Kurse legten drei Schwerpunkte; erstens: theoretische Überlegungen zu Aspekten der Auswahl von web-basierten Daten sowie Herausforderungen und Möglichkeiten, wie diesen mit bereits vorhandenen Mitteln begegnet werden kann; zweitens: Vermittlung von Grundkenntnissen über Web-Technologien; drittens: Befähigung zur Nutzung von R als Scraping- und Textmining-Software im Forschungsprozess. Tabelle 1 berichtet einen generischen Syllabus zu diesen Veranstaltungen. Da keine fortgeschrittenen technischen oder statistischen Kenntnisse vorausgesetzt wurden, standen die Kurse für Studierende aller Programme (BA, MA und Promotionsprogramm) offen und wurden auch dementsprechend besucht.

Aus dem ersten Kurs heraus entstand das Vorhaben, benötigte Grundlagen in einem Lehrbuch zusammenzufassen, um insbesondere Forschern ohne tiefere Programmier- und Webkenntnisse die Möglichkeit zu geben, sich diese Techniken selbständig und systematisch anzueignen (Munzert u. a. im Erscheinen). Trotz des beträchtlichen Umfangs des verfassten Lehrbuches ließ sich der Stoff mit wenigen Kürzungen im Laufe eines Semesters vermitteln – auch deshalb, weil mit R konsistent auf eine einzige Software für alle Arbeitsschritte gesetzt wurde, was die zeitlichen Kosten, die mit dem Erlernen von Programmiersprachen einhergehen,

minimierte. Die Struktur des Buches deckt sich weitgehend mit dem in Tabelle 1 skizzierten Kursplan und hält darüber hinaus noch einige Fallstudien bereit, anhand derer die Studierenden beispielhafte Anwendungen nachvollziehen konnten.

Tabelle 1: Skizze eines Kursplans zur web-basierten Datenerhebung mit R

Sitzung	***Block***	***Thema***
1	Einführung	Was tun mit web-basierten Daten?
2		Theoretische Überlegungen zur Datenauswahl
3	Technologien des World Wide Web	HTML und XML
4		XPath
5		HTTP
6		JSON und APIs
7		Datenbankmanagement mit SQL
8	Web Scraping mit R	R als Scraping-Software
9		Reguläre Ausdrücke
10		Anwendungen I: Statische Webseiten
11		Anwendungen II: Arbeit mit APIs
12		Anwendungen II: Dynamische Webseiten, AJAX
13		Scraping-Etikette und Organisation eines Datenerhebungsprojekts

Das Feedback der Studierenden im Rahmen der Lehrevaluation war bei allen Veranstaltungen ähnlich: Es wurde ein im Vergleich zu anderen Seminaren überdurchschnittlicher Arbeitsaufwand berichtet, gleichzeitig aber auch ein großes Interesse am Thema und der Wunsch, die erlernten Techniken in eigenen Arbeiten weiter anzuwenden. Als Ersatz für die klassische Hausarbeit wurde die schriftliche Prüfungsleistung im letzten unterrichteten Kurs in zwei kürzere Blog-Artikel für ein wissenschaftlich ausgerichtetes Fachpublikum geändert. So hatten die Studierenden die Möglichkeit, die erlernten Techniken in zwei Beispielen anzuwenden und in einem „datenjournalistischen" Stil über ihre Ergebnisse zu berichten. Einige Studierende nutzten die erworbenen Fähigkeiten, um im Rahmen ihrer Abschlussarbeiten innovativen Fragestellungen nachzugehen: Eine dieser Arbeiten beschäftigte sich mit dem Einfluss der Einführung der City-Maut-Zone in London auf das Immobilienpreisniveau – und damit möglicherweise nichtintendierte Konsequenzen lokaler Policies. Dazu wurden mehrere zehntausend Einträge von einer Online-Immobilienplattform erhoben, geolokalisiert und mit weiteren Informationen aufbe-

reitet. In einer anderen Arbeit wurden Klick-Statistiken von ausgewählten Wikipedia-Artikeln verwendet, um die Salienz des Energiewende-Issues fein aufgelöst über die Zeit messbar zu machen.

Das bisherige Fazit zu den unterrichteten Veranstaltungen fällt positiv aus: Die heterogene Zusammensetzung des Kurses aus BA-, MA- und Promotionsstudierenden stellte keinen Nachteil dar, da viele der vermittelten Techniken für alle Teilnehmer Neuland waren. Eine größere Herausforderung waren die unterschiedlichen R-Vorkenntnisse. Für Teilnehmer, die ohne jegliche Programmierkenntnisse in R oder Stata begonnen hatten, war die Lernkurve sehr steil und gerade zu Beginn des Kurses der Arbeitsaufwand enorm, was einige zum Aufgeben bewegte. Die Aneignung von Programmierkenntnissen stellt sicherlich die größte Hürde und gleichzeitig, wie oben dargelegt, eine entscheidende Kompetenz im Umgang mit neuen Datenquellen dar. Die aus dem Kurs heraus entstandenen Arbeiten waren durchweg kreativ und machten die Freude am Entdecken und Aufbereiten eigener Datenquellen deutlich. Gerade weil ein primär auf Datenerhebung ausgerichteter Kurs keine besonderen Vorkenntnisse bezüglich statistischer Analysefähigkeiten voraussetzt, ist die Materie aus meiner Sicht sehr gut dazu geeignet, um Studierende an selbständiges empirisches Arbeiten heranzuführen, da hier wesentliche Arbeitsschritte wie die Datensatzaufbereitung und deskriptive Exploration von Daten wieder und wieder geübt werden können. Deshalb bietet sich ein Kurs dieser Art an, um Studierende früh im Studium sowohl mit einer Programmiersprache als auch dem empirischen Arbeiten vertraut zu machen, ohne nennenswerte statistische Vorkenntnisse vorauszusetzen.

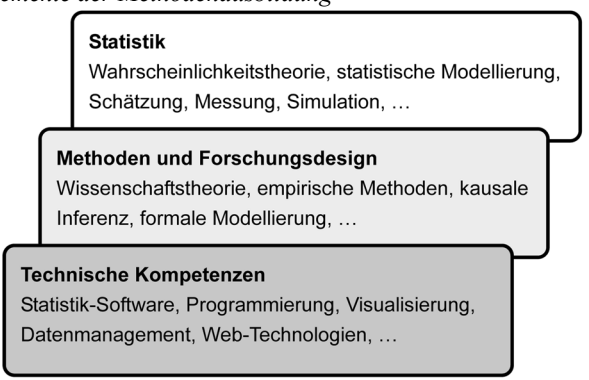
5. Ein Vorschlag zur Erweiterung der bestehenden Methodenausbildung

Was bedeuten die skizzierten Entwicklungen für die bestehende Methoden- und Statistikausbildung? Muss sich die Disziplin in diesem Bereich völlig neu ausrichten? Mit Blick auf die Proliferation neuer Methoden in den Sozialwissenschaften gäbe es jeweils gute Gründe, Themen wie Bayesianische Statistik, Netzwerkanalyse, agentenbasierte Modellierung, Qualitative Comparative Analysis, neue Praktiken der kausalen Inferenz und experimentellen Forschung und viele andere „Trendgebiete“ ins Pflichtprogramm der Methodenausbildung aufzunehmen. Dafür ist in einem zeitlich beschränkten politikwissenschaftlichen Studium allerdings kein Raum und sind die Inhalte zu spezifisch.

Ein Blick in die Prüfungsordnungen früherer Diplom- und Magisterstudiengänge ruft in Erinnerung, dass an vielen deutschen politikwissenschaftlichen Fachbereichen und Instituten verpflichtende Statistikveranstaltungen etwa gegen Ende der

1990er Jahre in die Methodenausbildung aufgenommen wurden – als etwas verspätete Reaktion darauf, dass sich computergestützte Regressionsanalysen in den 1980er Jahren zum Standard in Fachzeitschriften entwickelt hatten. Dieser Trend hält weiter an, und die Methodenentwicklung hat in den letzten Jahren nochmals an Fahrt aufgenommen. Umso wichtiger ist die grundlegende Ausbildung in Methoden, Statistik und Forschungsdesign, wie sie bereits jetzt – meist in Form einer Methoden- und einer Statistikveranstaltung im BA-Studium – Teil des Pflichtprogramms ist.

Abbildung 1: Elemente der Methodenausbildung



Quelle: Eigene Darstellung.

Das Gros der erforderlichen Kompetenzen im Umgang mit neuen Daten und Methoden, die oben aufgeführt wurden, ist vergleichsweise schnell vermittelbar. Für den Umgang mit *Big Data* besteht kein methodologisches oder statistisches Defizit, sondern primär ein technisches. Fähigkeiten wie der routinierte Umgang mit einer skriptbasierten Statistiksoftware, unterschiedlichen Datenformaten, neuerdings aber auch Webtechnologien und Datenbanken bilden den Unterbau für eigenes empirisches Arbeiten und damit die Umsetzung dessen in die Praxis, was bereits in Methoden- und Statistikveranstaltungen unterrichtet wird (vgl. auch Abbildung 1). Dabei ist aus meiner Sicht entscheidend, dass die Bedeutung der Schlüsselkompetenz der Programmierkenntnisse erkannt und im Studium mit entsprechender Aufmerksamkeit bedacht wird. Sobald man routiniert mit einer Software wie Stata oder R umgehen kann, erleichtert dies das Erschließen neuer statistischer Methoden oder technischer Fähigkeiten wie dem Web Scraping enorm. Da R hierbei gerade im Umgang mit neuen Datenquellen deutlich flexibler nutzbar ist und Studierenden

zudem kostenlos zur Verfügung gestellt werden kann, halte ich es für die Software, die bevorzugt werden sollte.

Bisher ist die technische Grundausbildung an politikwissenschaftlichen Fachbereichen und Instituten meist auf einen in der Regel freiwilligen Softwarekurs beschränkt. In empirisch ausgerichteten Studiengängen sollten entsprechende Fähigkeiten jedoch spätestens dann erworben sein, wenn im Studium selbständiges empirisches Arbeiten gefordert wird – am Ende des BA- oder zu Beginn des MA-Studiums. Minimalziel einer technischen Grundausbildung sollte deshalb zumindest ein verpflichtender Kurs sein, der Programmierlogik und eine grundlegende Einführung in Daten- und Webtechnologien vermittelt. Darauf aufbauend können, analog zu anderen vertiefenden Methodenkursen und je nach Kapazität der Fakultät, ergänzende Kurse zu Informationstechnologien (z. B. Web-Scraping-Techniken, Datenbankmanagement, Data- und Textmining) angeboten werden. Der Ressourcenaufwand für eine technische Grundausbildung hält sich so in Grenzen. Gleichzeitig würden von einer technischen Kompetenz, wenn diese bei allen Studierenden vorausgesetzt werden könnte, alle weiteren vertiefenden Methodenkurse profitieren.

Über die Wahl der Schwerpunkte in der Methodenausbildung und auch das Gewicht dieser Ausbildung im gesamten Studienplan lässt sich trefflich streiten. Die Ausrichtung dieser Pläne ist letztlich eine interne Aufgabe der einzelnen Fakultäten. Wenn sich aktive Forschung an einer Fakultät jedoch zunehmend mit Fragestellungen auseinandersetzt, die die Nutzung von neuen Datenformen erfordern, wird die Frage akut, ob und in welcher Form entsprechende Kompetenzen bereits im Bachelor- oder Masterstudium vermittelt werden können. Sicher ist: Der wissenschaftliche Nachwuchs sollte frühzeitig auf die Bearbeitung aktuell relevanter Probleme vorbereitet werden – und das mit modernen Mitteln.

Literatur

Barberá, Pablo, im Erscheinen: Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data, in: Political Analysis.

Blätte, Andreas, 2014: Das parlamentarische Diskursverhalten von Abgeordneten und die Kohäsion der Bundestagsfraktion 1996-2013: Plenarprotokollkorpora als Grundlage von Wordscore und Wordfish. Vortrag bei der Jahrestagung der Sektion Methoden der Politikwissenschaft am 24. Mai an der Universität Duisburg-Essen.

Bonica, Adam, 2013: Mapping the Ideological Marketplace, in: American Journal of Political Science 58 (2), 367-386.

Chen, Xi/Nordhaus, William D., 2011: Using luminosity data as a proxy for economic statistics, in: Proceedings of the National Academy of Sciences 108 (21), 8589-8594.

Döring, Holger/Regel, Sven, 2014: Data collection 2.0: Ein moderner Parteien-Almanach. Vortrag bei der Jahrestagung der Sektion Methoden der Politikwissenschaft am 24. Mai an der Universität Duisburg-Essen.

Grimmer, Justin, 2010: A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases, in: Political Analysis 18 (1), 1-35.

Grimmer, Justin/Stewart, Brandon M., 2013: Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, in: Political Analysis 21 (3), 267-297.

Henderson, Vernon/Storeygard, Adam/Weil, David N., 2011: A Bright Idea for Measuring Economic Growth, in: The American Economic Review 101 (3), 194-199.

Hopkins, Daniel/King, Gary, 2010: A Method of Automated Nonparametric Content Analysis for Social Science, in: American Journal of Political Science 54 (1), 229-247.

King, Gary, 2011: Ensuring the Data-Rich Future of the Social Sciences, in: Science 331 (11), 719-721.

King, Gary/Pan, Jennifer/Roberts, Margaret E., 2013: How censorship in China allows government criticism but silences collective expression, in: American Political Science Review 107 (2), 326-343.

King, Gary, 2014: Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science, in: PS: Political Science & Politics 47 (1), 165-172.

Kuhn, Patrick M./Weidmann, Nils B., im Erscheinen: Unequal We Fight: Between- And Within-Group Inequality And Ethnic Civil War, in: Political Science Research and Methods.

Laney, Doug, 2001: 3D management: Controlling data volume, velocity, and variety, in: Application Delivery Strategies, META Group, Inc., 6. Februar.

Laver, Michael/Benoit, Kenneth/Garry, John, 2003: Extracting Policy Positions from Political Texts Using Words as Data, in: American Political Science Review 97 (2), 311-331.

Linzer, Drew A., 2013: Dynamic Bayesian Forecasting of Presidential Elections in the States, in: Journal of the American Statistical Association 108 (501), 124-134.

Mellon, Jonathan, 2013: Where and When Can We Use Google Trends to Measure Issue Salience?, in: PS: Political Science & Politics 46 (2), 280-290.

Monroe, Burt L., 2013: The Five Vs of Big Data Political Science, in: Political Analysis, Virtual Issue on Big Data in Political Science, 1-9.

Munzert, Simon/Rubba, Christian/Meißner, Peter/Nyhuis, Dominic, im Erscheinen: Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining. Hoboken, NJ.

Norvig, Peter, 2001: Teach Yourself Programming in Ten Years, http://norvig.com/21-days.html (Stand: 05.08.14).

Nyhuis, Dominic/Faas, Thorsten, 2014: Capturing real-time political evaluations using Twitter. Vortrag bei der Jahrestagung der Sektion Methoden der Politikwissenschaft am 23. Mai an der Universität Duisburg-Essen.

Shaw, Aaron/Hill, Benjamin M., 2014: Laboratories of Oligarchy? How the Iron Law Extends to Peer Production, in: Journal of Communication 64 (2), 215-238.

Silver, Nate, 2012: Special Coverage: The 2012 Presidential Election. FiveThirtyEight, http://fivethirtyeight.blogs.nytimes.com/2012/11/06/live-blog-the-2012-presidential-election/ (Stand: 25.07.14).

Slapin, Jonathan B./ Proksch, Sven-Oliver, 2008: A Scaling Model for Estimating Time-Series Party Positions from Texts, in: American Journal of Political Science 52 (3), 705-722.

Squire, Peverill, 1988: Why the 1936 Literary Digest Poll Failed, in: Public Opinion Quarterly 52 (1), 125-133.

Korrespondenzanschrift:

Simon Munzert, M.A.
Universität Konstanz
Fachbereich Politik- und Verwaltungswissenschaft
Postfach D 85
Universitätsstraße 10
78457 Konstanz
E-Mail: simon.munzert@uni-konstanz.de