

### Lernt programmieren! Big Data in der sozialwissenschaftlichen Lehre

Munzert, Simon

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

#### Empfohlene Zitierung / Suggested Citation:

Munzert, S. (2016). Lernt programmieren! Big Data in der sozialwissenschaftlichen Lehre. *Soziopolis: Gesellschaft beobachten*. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-80515-8>

#### Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:

<https://creativecommons.org/licenses/by/4.0/deed.de>

#### Terms of use:

This document is made available under a CC BY Licence (Attribution). For more information see:

<https://creativecommons.org/licenses/by/4.0>

Simon Munzert | Essay | 09.11.2016

## Lernt programmieren!

### Big Data in der sozialwissenschaftlichen Lehre

In der Kognitionspsychologie kennt man die 10.000-Stunden-Regel. Sie besagt, dass ein immenser Zeitaufwand (im Schnitt ebenjene 10.000 Stunden; manchmal ist auch von zehn Jahren die Rede) und kontinuierliche Übung nötig sind, um zu Expert\_innen in einer kognitiv anspruchsvollen Tätigkeit zu werden.<sup>1</sup> Forschung in jeder Disziplin, also auch in den Sozialwissenschaften, muss als kognitiv anspruchsvoll bezeichnet werden. Für die Ausbildung zum sozialwissenschaftlichen Arbeiten sind in der Tat derzeit etwa zehn Jahre vorgesehen, wenn man drei bis vier Jahre für das BA-Studium, zwei Jahre für das MA-Studium und drei bis fünf Jahre für die Promotion rechnet. Ist also alles in Butter bei der Ausbildung neuer Expert\_innen der Sozialwissenschaften?

### Neue Chancen, neue Anforderungen

Die Anforderungen an empirisch forschende Sozialwissenschaftler\_innen haben in den letzten Jahren stark zugenommen. Wir müssen nicht die Erledigung einer eng umgrenzten Aufgabe perfektionieren, etwa den Sprung über eine möglichst hohe Hürde oder das Toreschießen. Stattdessen geht es um die Erklärung sozialen Verhaltens, und das war schon immer ein weites Feld. Der Themenschwerpunkt zu Big Data, in dessen Rahmen dieser Beitrag erscheint, verdeutlicht, wie vielfältig die Möglichkeiten sozialwissenschaftlicher Forschung geworden sind. Wollen Forschende diese Möglichkeiten aber ausschöpfen, benötigen sie Kompetenzen, die bislang nicht Teil ihrer Ausbildung waren. Dazu zählen:

1.

die Sammlung umfangreicher Datenmengen mithilfe von Datenbanken, Rechenclustern oder Sampling-Techniken,

2.

der Umgang mit neuen Datenformen, z.B. unstrukturierten Daten von Webseiten, Geodaten oder Textdaten,

3.

die Verknüpfung unterschiedlicher Daten miteinander, z.B. administrativer Daten, Bewegungsdaten oder Umfragedaten,

4.

die Auswertung massiver Datenmengen und das Aufspüren substanziiell bedeutsamer Information in großem Datenrauschen und schließlich

5.

der forschungsethisch verantwortungsvolle Umgang mit potenziell sensitiven und kompromittierenden Informationen.

Freilich klingt das eher nach der Grundausbildung angehender Informatiker\_innen als nach einem sozialwissenschaftlichen Curriculum. Darüber hinaus bleibt das bestehende Standardrepertoire der Methodenausbildung an sozialwissenschaftlichen Instituten – also die Einführung in Methoden der empirischen Sozialforschung, Statistik und Forschungsdesign – natürlich weiterhin relevant.

Sollte man die technischen Details im Umgang mit neuartigen, komplexen Daten nicht besser ausgebildeten Informatiker\_innen überlassen? Ich denke nein – genauso wenig, wie man sich ausschließlich auf Statistiker\_innen bei der Datenauswertung verlassen darf. Seriöse sozialwissenschaftliche Forschung setzt ein grundlegendes Verständnis davon voraus, wie Daten generiert werden. Im Umgang mit Umfragedaten haben wir das gelernt. In der Sensibilität für und der Kompetenz im Umgang mit klassischen Mess- und Selektionsproblemen humangenerierter Daten liegt sogar eine Stärke sozialwissenschaftlicher gegenüber rein technischen Disziplinen, die in interdisziplinäre Zusammenarbeit eingebracht werden muss. Dafür braucht es jedoch ein technisches Grundverständnis, das bereits möglichst frühzeitig zu vermitteln ist.

## **Schlüsselkompetenz Programmierkenntnisse**

Die obige Auflistung der dringend erforderlichen Kompetenzen im Umgang mit Big Data macht eines klar: Sozialwissenschaftliche Methodenausbildung benötigt einen stärkeren

technischen Fokus. Prozessproduzierte Daten wie Bewegungsprofile, Tweets, Online-Suchverhalten etc. sind in der Regel nicht statisch und abgeschlossen, sondern dynamisch, kumulativ und unstrukturiert. Da sie häufig zu anderen Zwecken als der wissenschaftlichen Auswertung erstellt werden, stecken sie außerdem voller Ungenauigkeiten, Redundanzen und unnützer Information. Die Datenerhebung und -aufbereitung nehmen so noch mehr Kapazitäten im empirischen Forschungsprozess in Anspruch. Gleichzeitig ist es wünschenswert, diese häufig repetitiven Aufgaben zu automatisieren. Außerdem erfordert die Erfassung neuer Datenstrukturen Visualisierungs- und Analysestrategien, die sich mit standardisierten Befehlen in etablierten Analyseprogrammen gar nicht oder nur schwer umsetzen lassen.

Meine Hauptthese lautet daher: Fundierte Programmierkenntnisse sind unabdingbar für die eigenständige, originelle Analyse neuer Daten und die Behandlung damit einhergehender Forschungsfragen. Nur so lässt sich ein tiefergehendes technisches Verständnis der Charakteristiken von Datenquellen und programmiertechnischer Lösungen voraussetzen. Die Vermittlung solcher Programmierkenntnisse darf allerdings nicht auf Kosten der substanziellen sozialwissenschaftlichen Ausbildung gehen, zumal die Studienzeit angesichts gut gefüllter Lehrpläne in sozialwissenschaftlichen BA- und MA-Studiengängen bereits jetzt knapp bemessen ist. Gleichzeitig erfordert die Nutzung effizienter Datenanalysesoftware das Erlernen skriptbasierter Programmierung, was enorme zeitliche Ressourcen verschlingt (hier sind wir wieder bei der 10.000-Stunden-Regel). Liegt darin ein unüberwindbares Hindernis für eine Ausbildung, die den Forschungsnachwuchs für den Umgang mit Big Data wappnet?

An anderer Stelle habe ich einen möglichen Ausweg aus diesem Dilemma skizziert<sup>2</sup> und vorgeschlagen, die Hochschulen sollten sich auf die Vermittlung einer im Idealfall vielfältig einsetzbaren Software beschränken. Das würde den Studierenden ermöglichen, im Verlauf des Studiums aufeinander aufbauende, vertiefte Kenntnisse im Umgang mit der gewählten Software zu erwerben. Auch die Lehrenden könnten von einer – nicht nur innerhalb der Institute, sondern auch über Institute hinweg – einheitlichen Softwareausbildung profitieren, da beispielsweise weiterführende Methodenkurse auf zuvor erworbenem Wissen aufbauen könnten. Als Programm der Wahl bietet sich R an, eine Datenanalyseumgebung, die in den vergangenen Jahren in den Sozialwissenschaften massiv an Popularität gewonnen hat und zudem als Open-Source-Software frei zugänglich ist.<sup>3</sup>

## Lücken in der bisherigen Methodenschulung

Ebenso wichtig wie die Wahl der Analysewerkzeuge ist es aber, eine fundierte technische Ausbildung im Umgang mit Daten endlich fest in den sozialwissenschaftlichen Studienplänen zu verankern. Dass dem noch nicht so ist, die Herausforderung aber in Teilen bereits erkannt wird, hat auch eine Umfrage gezeigt, die ich vor einem Jahr unter Methodendozent\_innen an politikwissenschaftlichen und soziologischen Instituten in Deutschland durchgeführt habe.<sup>4</sup> Während an beachtlichen vierzig Prozent der Fakultäten Softwarekurse (auch begleitend zur Vorlesungen) verpflichtend sind sowie in nochmals knapp vierzig Prozent der Fälle optional angeboten werden, sind nach wie vor SPSS, Stata und Excel am häufigsten in der Lehre im Einsatz. Vor allem SPSS und Excel bieten aber hinsichtlich neuartiger Datenquellen kaum zukunftsfähige Umgebungen, auch wenn sie in der Industrie nach wie vor äußerst populär sind. Der nachhaltigen Förderung des wissenschaftlichen Nachwuchses dienen Kurse zu diesen Programmen in meiner Sicht nur sehr beschränkt. Von einer soliden Ausbildung in fortgeschrittener statistischer Programmierung oder gar einem einheitlichen Softwarestandard kann in jedem Fall hinsichtlich der Ergebnisse der Umfrage noch nicht die Rede sein.

## Auf dem Weg zu einer Neuausrichtung der Methodenausbildung

Es scheint unvermeidlich, die bisherige Methodenausbildung um ein technisches Fundament zu ergänzen. Meine Bestandsaufnahme an politikwissenschaftlichen Instituten hat gezeigt, dass eine grundlegende Statistik- und Methodenausbildung in den meisten Studienplänen fest verankert ist. Wie man zusätzlich eine Grundausbildung in technischen Kompetenzen integrieren könnte, habe ich in der Abbildung skizziert.

Statistik und Modellierung	
<b>BA-Curriculum</b> Wahrscheinlichkeitstheorie, Einführung in die Statistik und statistische Modellierung	<b>MA- bzw. Graduierten-Curriculum</b> Statistische Inferenz, Messung, Simulation, Experimentelle Statistik, Hierarchische Modellierung, Netzwerkanalyse, Text Mining, ...
Methoden und Forschungsdesign	
<b>BA-Curriculum</b> Wissenschaftstheorie, Einführung in die Methoden der empirischen Sozialforschung	<b>MA- bzw. Graduierten-Curriculum</b> Forschungsdesign, kausale Inferenz, formale Modellierung, ...
Technische Grundkompetenzen	
<b>BA-Curriculum</b> Softwarekurs, Implementation wissenschaftlicher Standards (Reproduzierbarkeit, Transparency)	<b>MA- bzw. Graduierten-Curriculum</b> Fortgeschrittene statistische Programmierung, Visualisierung, Web-Technologien, Umgang mit Datenbanken, ...

## Ein Vorschlag zur Neuausrichtung der sozialwissenschaftlichen Methodenausbildung

Da die meisten BA-Studiengänge kaum Spielraum für zusätzliche Pflichtkurse lassen, schlage ich vor, das bestehende Angebot lediglich um einen technischen Kurs zu erweitern, der ausschließlich dem Erlernen einer Statistiksoftware gewidmet ist. Letztere kann auch die Grundlage für alle weiteren methodischen oder forschungspraktisch ausgerichteten Veranstaltungen bilden. In der Regel wird von Studierenden spätestens beim Verfassen der ersten Abschlussarbeit gefordert, selbständig empirisch zu arbeiten. Deshalb sollten ihnen entsprechende Werkzeuge bereits vorher vertraut sein. Das bedeutet auch, dass im Rahmen eines solchen Kurses mögliche Verknüpfungen zwischen Grundprinzipien wissenschaftlichen Arbeitens und deren Implementation in die Arbeit mit Statistikprogrammen vermittelt werden müssen.

Darauf aufbauend wären im MA- und Graduiertenkolloquium weiterführende und spezifischere Kompetenzen zu erlernen, beispielsweise in der Visualisierung von Daten und statistischen Ergebnissen oder im Umgang mit neuen Datenquellen, aber auch Techniken, um große Datenmengen abzulegen. In dieser Studienphase sollte das Problem der Konkurrenz durch andere Pflichtkurse weniger virulent sein, zumal die Studierenden höhere Anreize erhalten, sich in praktischen Forschungskompetenzen weiterzubilden. Wilson identifiziert diese Zeit des Studiums sogar als idealen Zeitpunkt zur Einbettung der technischen Ausbildung.<sup>5</sup>

Freilich bleibt die Auswahl zu vermittelnder Inhalte in BA- und MA-Studiengängen im Wesentlichen ein Nullsummenspiel. Was soll aus dem Curriculum herausfallen, um der Softwareausbildung Platz zu machen? Diese Frage hängt sicherlich von der inhaltlichen Ausrichtung und Lehrkapazitäten an den einzelnen Instituten ab. Mit Blick auf die technische Grundausbildung hält sich, folgt man meinem Vorschlag, der Ressourcenaufwand allerdings in Grenzen – es müsste lediglich ein einziger zusätzlicher verpflichtender Kurs im BA-Studium untergebracht werden. Womöglich kann eine solche strukturelle Anpassung in der Methodenausbildung auch Kapazitäten an anderer Stelle freisetzen – dann nämlich, wenn substantiell ausgerichtete Lehrangebote auf technisches Vorwissen der Studierenden aufbauen können. So müssten die Studierenden beispielsweise in einem Kurs zu politischer Partizipation, in dem sie selbständig Analysen durchführen sollen, nicht mehr zunächst die dafür notwendige Software erlernen.

Wissenschaftliches Arbeiten ist in erster Linie freies, selbständiges Arbeiten. Als Auszubildende ermöglichen wir Nachwuchsforscher\_innen diese Freiheiten bei der Wahl der Forschungsthemen aber nur, indem wir ihnen ein möglichst solides Training in Techniken und Methoden wissenschaftlichen Arbeitens bieten. Ericsson und seine Kollegen haben

übrigens die von ihnen bekanntgemachte Zehn-Jahres-Regel eingeschränkt: Den erfolgreichen Weg zu\_r Expert\_in könne nur beschreiten, wer zudem kontinuierliches Feedback von Lehrer\_in oder Trainer\_in erhalte. Nur so könnten die ‚Noviz\_innen‘ einer Tätigkeit effizient aus Fehlern lernen und sich dadurch verbessern. Wissenschaftliches Arbeiten sowohl im Studium als auch danach heißt immer auch, Autodidakt\_in zu sein. Die Studiengänge liefern dabei vor allem wichtige Grundlagen und Anleitungen zum weiterführenden Studium neuer Forschungsfelder. Beim Kennenlernen grundlegender Werkzeuge, wozu mittlerweile für Sozialwissenschaftler\_innen eben auch Programmierkenntnisse gehören, sollten wir den wissenschaftlichen Nachwuchs aber nicht alleine lassen – es gibt in der Fülle neuer Spuren menschlichen Verhaltens noch so viel zu entdecken.

*Dieser Beitrag ist Teil eines Themenschwerpunkts zu Big Data. Weitere Texte finden Sie [hier](#).*

## Endnoten

1. Vgl. dazu K. Anders Ericsson / Ralf Th. Krampe / Clemens Tesch-Römer, [The Role of Deliberate Practice in the Acquisition of Expert Performance](#), in: Psychological Review 100 (1993), 3, S. 363–406.
2. Vgl. Simon Munzert, Auf dem Weg zu einer fundierten Softwareausbildung in der Politikwissenschaft, in: Joachim Behnke u.a. (Hrsg.), Big Data: Große Möglichkeiten oder große Probleme?, Baden-Baden 2017 [im Erscheinen].
3. Siehe <https://cran.r-project.org/>.
4. Vgl. erneut Munzert, Softwareausbildung in der Politikwissenschaft.
5. Greg Wilson, Software Carpentry: lessons learned, in: F1000Research (2014), 3.

## Simon Munzert

Dr. Simon Munzert ist wissenschaftlicher Mitarbeiter in der Daten- und Methodeneinheit am Mannheimer Zentrum für Europäische Sozialforschung. Dort leitet er das Social Science Data Lab. Er ist Autor eines Lehrbuchs zur automatisierten Sammlung web-basierter Daten. Seine Forschungsgebiete umfassen die Messung öffentlicher Meinung, politischer Repräsentation und dem Einfluss neuer Medien auf politische Kommunikation.

**Dieser Beitrag wurde redaktionell betreut von Kerstin Völkl.**

**Artikel auf soziopolis.de:**

<https://www.sozopolis.de/lernt-programmieren.html>