Die soziale Dynamik neuer Technologien Big Data als Herausforderung für die Gesellschaftswissenschaften

Markus Konrad und Jonas Wiedner

Summary: Machine learning methods and the usage of cloud services currently enter the phase of broad application in the social sciences. This creates exciting new opportunities for researchers. But does that also mean better research and more reliable results? Not necessarily. New technical means are not simply neutral tools – they form research interests and come along with new limitations and problems.

Kurz gefasst: Aktuell treten neue Verfahren im Bereich des maschinellen Lernens sowie zahlreiche Cloud-Dienste in die Phase der breiten Anwendung in den Sozialwissenschaften ein. Damit einher gehen aufregende neue Möglichkeiten. Bedeutet das bessere Forschung und mehr verlässliches Wissen? Nicht unbedingt. Denn neue technische Mittel sind nicht einfach neutrale Werkzeuge. Sie formen Erkenntnisinteressen und bringen neue Grenzen und Probleme mit sich, derer sich Forschende bewusst sein sollten.

Seit ihrer Entstehung im 19. Jahrhundert bewegt sich die wissenschaftliche Erforschung der Gesellschaft stets nahe an der Grenze des technologisch Machbaren. Heute antiquierte Geräte wie die mechanischen Tabelliermaschinen der frühen amerikanischen Volkszählungen, die Lochkarten der 1950er-Jahre oder die Magnetbänder, auf denen bis in die 1980er-Jahre statistische Daten gespeichert wurden, legen davon Zeugnis ab. Es verwundert daher nicht, dass auch die jüngsten Innovationen in der Datenverarbeitung ihren Niederschlag in neuen Analysestrategien, neuen Daten und Datenerhebungsmethoden von SozialwissenschaftlerInnen finden. Zwei Trends sind hier besonders hervorzuheben: die Verwendung von Cloud-Diensten und der Einsatz von Methoden des maschinellen Lernens (ML).

Cloud-Dienste sind Anwendungen, die von Anbietern online zur Verfügung gestellt werden und durch die WissenschaftlerInnen von rechen- oder datenintensiven Anwendungen profitieren können, ohne selbst die nötige komplexe Infrastruktur aufbauen zu müssen. Häufig werden Cloud-Dienste verwendet, um bestehende Forschungsdaten anzureichern: Aus einem Datensatz von Straßenadressen lassen sich die GPS-Koordinaten ermitteln, die für räumliche Analysen benötigt werden; ein Korpus mit Plenardebatten in unterschiedlichen Sprachen lässt sich in eine einheitliche Sprache übersetzen und so weiter. Die ForscherInnen senden Aufträge an den Cloud-Dienst und erhalten das berechnete Ergebnis zurück. Durch das Automatisieren dieser Abfragen lassen sich enorme Datenmengen in kurzer Zeit verarbeiten. ML-Methoden beruhen auf statistischen Modellen, die anhand von Trainingsdaten Muster erkennen können und, auf einen neuen Datensatz angewendet, datenbasierte Vorhersagen liefern. Eine verbreitete Anwendung ist die Klassifizierung von Texten. So könnten die übersetzten Plenarbeiträge automatisch als emotional oder sachlich im Ton kategorisiert werden, wenn ein ML-Modell vorher mit Beispieldaten darauf trainiert wurde. Viele Cloud-Dienste basieren auf solchen ML-Modellen.

Für WissenschaftlerInnen sind ML-Methoden und Cloud-Angebote aufregend und schaffen neue Möglichkeiten. Bedeutet das also bessere Forschung und am Ende mehr verlässliches Wissen? Gegen diese Perspektive argumentieren wir im Folgenden, dass neue technische Mittel nicht einfach neutrale Werkzeuge sind. Vielmehr formen sie selbst Erkenntnisinteressen: Wenn es möglich wird, bislang untypische Arten von Daten zu analysieren – also beispielsweise Social-Media-Beiträge oder Bewegungsdaten von Smartphone-NutzerInnen –, werden auch verstärkt Fragen untersucht, die sich auf solche Daten beziehen. Und auch die vermeintlich technischen Aspekte einer Wissenschaft sind bestimmten sozialen Dynamiken unterworfen – mit durchaus problematischen Auswirkungen auf die Belastbarkeit von Forschungsergebnissen. Es sind diese nicht technischen Folgen neuer Forschungstechnologien, die wir hier diskutieren wollen. Wissenschaftstheoretiker wie Thomas Kuhn, Imre Lakatos oder Paul Feyerabend haben herausgearbeitet, dass Wissenschaft kein kontinuierlicher Prozess in Richtung größerer Erkenntnis ist. Nach dem einflussreichen Modell Kuhns verläuft wissenschaftlicher Fortschritt in Zyklen: Auf eine radikale Innovation mit neuen Interessen, Prämissen und Methoden folgt eine Phase der „Normalwissenschaft", bis eine „Krise" mit unbewältigbaren Problemen diese Ansätze wiederum fragwürdig erscheinen lässt und erst radikale Innovation neue Wege eröffnet.

Auch auf einem für WissenschaftlerInnen alltäglichen Niveau, der Ebene der Forschungsmethoden, sind zyklische Prozesse auszumachen: Am Anfang steht methodische Innovation, manchmal als Reaktion auf ein Forschungsproblem, das sich mit etablierten Hilfsmitteln nicht bearbeiten lässt. Diese Innovationen bauen auf Kenntnissen auf, die im Mainstream der Disziplin kaum vorhanden sind, daher werden sie oft von hochspezialisierten WissenschaftlerInnen oder Fachfremden (zum Beispiel aus der Informatik oder der Mathematik) entwickelt. Entscheidend ist, dass in dieser frühen Phase die Anwendung der neuen Methoden Spezialkenntnisse und eine aufwendige Anpassung an das Forschungsproblem erfordert, sodass der Kreis der NutzerInnen und Anwendungen zunächst stark beschränkt ist. Damit sind diese Methoden allerdings auch besonders erkenntnisreich, ihre Anwendung ist prestigeträchtig. In der Folge wird daher der Zugang immer weiter demokratisiert. Es häufen sich Anwendungen in Fachzeitschriften, es erscheinen Handbücher, und die neuen Techniken werden als vorgefertigte Routinen in verbreitete Statistikprogramme oder Cloud-Dienste integriert.

Häufig kommt es allerdings zu problematischen Entwicklungen, bevor eine methodische Innovation produktiv in das Arsenal des neuen normalwissenschaftlichen Vorgehens eingereiht werden kann. Wenn sich der Kreis der AnwenderInnen über SpezialistInnen hinaus auf die angewandte Forschung erweitert, sinkt notwendigerweise die durchschnittliche technische Kompetenz. Daraus können Probleme erwachsen, etwa wenn wichtige Annahmen der Verfahren ignoriert werden. Ergebnisse wären dann viel eingeschränkter zu interpretieren, als es de facto geschieht. Ein Beispiel für diese Probleme liefern die seit den 1960er-Jahren immer weiter verbreiteten Regressionsmodelle in den Sozialwissenschaften. Diese erlauben es unter bestimmten Bedingungen, Störvariablen konstant zu halten. Die Erwartungen an diese Art der Datenanalyse waren anfangs geradezu utopisch („It is the regression coefficients which give us the laws of science", schrieb etwa der ansonsten durchaus kritische Soziologe Hubert Blalock 1972). Seit den 1980ern hat die methodische und theoretische Forschung verstärkt darauf hingewiesen, wie voraussetzungsreich die Interpretation selbst vermeintlich einfacher Modelle dieser Art ist. Kurz: Unzureichend verstandene Methoden führen zu schlechter Wissenschaft. Nach der Popularisierung einer neuen Technik muss also eine Phase der Aufklärung stattfinden, und die fachwissenschaftliche Community muss sich Standards geben, die die Anwendung neuer Verfahren regulieren.

Aktuell treten die neuen Verfahren im Bereich des maschinellen Lernens und die zahlreichen Cloud-Dienste in die Phase der breiten Anwendung ein. Es ist also an der Zeit, dass sich ForscherInnen als AnwenderInnen, aber auch als Scientific Community darüber bewusst werden, welche Grenzen die neuen Technologien haben – und wie der Forschungsprozess so gestaltet werden kann, dass wissenschaftliche Standards nicht durch neue, aufregende, aber unzureichend durchdrungene Verfahren gefährdet werden. Die neuen Möglichkeiten auf Basis maschinellen Lernens und der Datenanreicherung durch Cloud-Dienste weisen nämlich konkrete Risiken auf, die die Qualität sozialwissenschaftlicher Forschung beeinträchtigen können. Die meisten ML-Algorithmen wurden mit dem Ziel entwickelt, möglichst exakte Vorhersagen zu liefern, wenn sie mit neuen Daten gefüttert werden. Dafür wird in Kauf genommen, dass die Modelle so kompliziert sind, dass Menschen sie kaum inhaltlich interpretieren können. In den Sozialwissenschaften werden die komplizierten ML-Modelle daher vor allem beim Berechnen von Zwischenergebnissen genutzt, etwa wenn, wie im obigen Beispiel, Textfragmente hinsichtlich ihrer Gefühlsrichtung klassifiziert werden. Diese Ergebnisse fließen dann in klassische statistische Modelle ein, die inhaltliche Interpretationen gestatten – zum Beispiel in ein Regressionsmodell, das Unterschiede in der Aggressivität von Debattenbeiträgen zwischen verschiedenen Parteienfamilien schätzt. Klassifikationen aus ML-Modellen sind jedoch – wie jede Vorhersage auf Basis eines statistischen Modells – mit Unsicherheit behaftet. Problematisch wird das, wenn, wie es in der Praxis häufig geschieht, diese Unsicherheiten nicht in den weiteren Berechnungen berücksichtigt werden. In diesem Fall scheinen die Schlussfolgerungen aus dem finalen Modell präziser zu sein, als sie es in Wirklichkeit sind. Um wirklich reproduzierbare Schlussfolgerungen zu ziehen, ist es deshalb gerade bei Analysen auf Basis von ML-Techniken wichtig, die statistische Unsicherheit jedes Teilschritts miteinzubeziehen.

Markus Konrad ist als Data Scientist in der IT-Abteilung des WZB zuständig für Datenaufbereitung, -analyse und -visualisierung. Der Schwerpunkt seiner Arbeit liegt auf Data Mining, quantitativer Textanalyse und Analyse von Geodaten; über seine Arbeit schreibt er auch im Data Science Blog. *[Foto: Martina Sander]*

markus.konrad@wzb.eu

Jonas Wiedner ist wissenschaftlicher Mitarbeiter in der Abteilung Migration, Integration, Transnationalisierung. Dort forscht er zur wohnungsbezogenen Mobilität ethnischer Minderheiten in Deutschland. Daneben interessiert er sich für Fragen der sozialen Mobilität und der Arbeitsmarktforschung, vor allem in Bezug auf Menschen mit Migrationshintergrund. *[Foto: Martina Sander]*

jonas.wiedner@wzb.eu

Dass ML-Modelle darüber hinaus prinzipiell nur so gut sein können wie die Daten, mit denen sie trainiert wurden, ist mittlerweile gut erforscht (das Problem wird kurz gefasst in dem Slogan: „Garbage in – Garbage out"). Ein frappierendes Beispiel sind geschlechterstereotype Übersetzungen. Wer die Passage „The doctor sits down. She looks at the patient" per Google Translate, einer Anwendung, die auf ML-Modellen basiert, ins Deutsche übersetzen lässt, wird geschlechterstereotype Algorithmen am Werk sehen. Werden solche Modelle zur Ermittlung von Zwischenergebnissen eingesetzt, übertragen sich deren Verzerrungen womöglich in die Endergebnisse. In einem einfachen Beispiel sind Verzerrungen noch offensichtlich, aber in tatsächlichen Anwendungen ist es deutlich komplizierter, solchen Problemen auf die Schliche zu kommen.

Entscheidende Fragen für wissenschaftliche NutzerInnen von Cloud-Diensten lauten also: Welchen Modellen und Diensten kann man für welche Zwecke vertrauen? Mit welchen Daten wurden sie trainiert, welche Verzerrungen existieren, und wie groß ist die statistische Unsicherheit der Ergebnisse? Leider sind Antworten auf diese Fragen schwer zu finden. Gerade Cloud-Dienste sind tendenziell intransparent und werden nicht unabhängig evaluiert. Für WissenschaftlerInnen besteht also die reale Gefahr, sich über den Umweg eines externen Dienstleisters systematische Verzerrungen einzuhandeln, die jedoch weder von den ProduzentInnen noch von den KonsumentInnen der Forschungsergebnisse überblickt werden können. Auch in dieser Hinsicht stehen die neuen Instrumente also in einem Spannungsfeld zum Gebot der Reproduzierbarkeit. Insbesondere in Bezug auf die Nutzung von Cloud-Diensten kommen zu diesen Problemen noch Fragen hinsichtlich des Datenschutzes, der Nachvollziehbarkeit und der Abhängigkeit von Internetkonzernen hinzu. Da Daten auf IT-Systemen von Drittanbietern verarbeitet werden, müssen Forschende hier besondere Sorgfalt in Bezug auf den Datenschutz walten lassen. Setzt man Cloud-Dienste ein, können andere WissenschaftlerInnen die Resultate nur dann replizieren, wenn sie auch die gegebenenfalls teuren Cloud-Dienste nutzen – wenn diese noch auf dieselbe Weise funktionieren.

Um diesen Problemen zu begegnen, muss sich die wissenschaftliche Community ihrer in einem ersten Schritt überhaupt bewusst werden. Damit der Zyklus von Methoden-Hype und Replikations-Kater zumindest abflacht, müssen Beiträge, die auf Machine Learning oder Cloud-Diensten aufbauen, entlang der oben diskutierten Punkte kritisch befragt werden. In der Aus- und Fortbildung von WissenschaftlerInnen sollten auch die Grenzen neuer Methoden verhandelt werden. In einem zweiten Schritt ist es wichtig, dass wissenschaftliche Fachzeitschriften klare Standards für Beiträge auf Basis von Cloud-Computing und ML-Methoden entwickeln. Relativ einfach ist es, das Mitführen von statistischer Unsicherheit aller Zwischenschritte zu verlangen. Am schwierigsten ist es, implizite Verzerrungen in ML-Modellen auszuschließen beziehungsweise abzumildern. Hier ist vor allem weitere eingehende Forschung notwendig. Aus wissenschaftlicher Perspektive wünschenswert wäre es schließlich, wenn Cloud-Anbieter ihre Angebote extern in Bezug auf Verzerrungen evaluieren ließen und ihre Dienste generell transparenter gestalteten. Konkret hieße das, offenzulegen, auf welchen oder auf welcher Art von Trainingsdaten ein ML-Dienst beruht, und statistische Unsicherheitsmaße zu den Ergebnissen mitzuliefern. Um Replikationen zu ermöglichen, wäre außerdem eine Versionskontrolle von Cloud-Diensten nötig: Ein Dienst muss sich zu einem späteren Zeitpunkt in den Zustand zurückversetzen lassen, in dem er sich während der Entstehung der fraglichen Analyse befunden hat. Eine Mindestanforderung wäre ein Änderungsprotokoll, sodass Interessierte zumindest nachvollziehen können, wann und wie sich das Verhalten eines Dienstes geändert hat und ob sich das auf die Replizierbarkeit auswirken kann.

Diese Vorschläge durchzusetzen wird nicht einfach, schließlich baut das Geschäftsmodell der Cloud-Dienstleister auf der exklusiven Verfügung über ihre Datenbestände auf. Dennoch: In Anbetracht der oben skizzierten Problematiken müssen sich SozialwissenschaftlerInnen diesen schwierigen Fragen stellen, wenn die zweifellos großen Versprechen von Cloud-Diensten und maschinellem Lernen ohne böses Erwachen eingelöst werden sollen.

Literatur

Kuhn, Thomas: Die Struktur wissenschaftlicher Revolutionen. Frankfurt am Main: Suhrkamp 1967.

Munafò, Marcus R./Nosek, Brian A./Bishop, Dorothy V.M. et al.: „A Manifesto for Reproducible Science". In: Nature Human Behaviour, 2017, Jg. 1, 0021. DOI: 10.1038/s41562–016–002.

Papakyriakopoulos, Orestis/Hegelich, Simon/Serrano, Juan C.M./Marco, Fabienne: Bias in Word Embeddings. 2020. Online: https://doi.org/10.1145/3351095.3372843 (Stand 18.02.2021).

Salganik, Matthew J.: Bit by Bit. Social Research in the Digital Age. Princeton: Princeton University Press 2019.