Datum: 17. November 2017

Big Data – der Beginn eines goldenen Zeitalters für die Arbeitsmarktforschung? Ein Gespräch mit IAB-Forscherin Frauke Kreuter

Martin Schludi

Big Data! So lautet das neue Zauberwort auch in der Arbeitsmarktforschung. Doch was hat es damit eigentlich auf sich? Was bringt das Ganze? Und was bedeutet es für den Datenschutz und die Aus- und Weiterbildung unserer Forscherinnen und Forscher? Frauke Kreuter, Leiterin des Kompetenzzentrums Empirische Methoden am IAB und Professorin für Statistik und Methoden der empirischen Sozialforschung an der Universität Mannheim, steht im Interview für das IAB-Forum Rede und Antwort.

In letzter Zeit liest man immer häufiger das Schlagwort „Big Data“, insbesondere im Zusammenhang mit den sozialen Medien. Was verbirgt sich hinter diesem Begriff?

Der Begriff wird in der Tat viel verwendet, aber nicht alle verstehen dasselbe darunter. Eines

steht fest: Es geht bei Big Data gar nicht so sehr um die Größe selbst, denn große Datenmengen gibt es in der Physik, der Astronomie, den Geowissenschaften et cetera schon seit langem. Was typischerweise unter Big Data verstanden wird, sind große unstrukturierte Datenmengen, die als Nebenprodukt anderer Aktivitäten entstehen, also nicht für diesen Zweck gesammelt wurden beziehungsweise für die keine eigene Messreihe aufgesetzt wurde. So entstehen zum Beispiel in den sozialen Netzwerken Texte und Bilder in digitaler Form, die jetzt als Daten vorliegen und von uns ausgewertet werden können.

> ***"Unter Big Data werden typischerweise große unstrukturierte Datenmengen verstanden, die als Nebenprodukt anderer Aktivitäten entstehen."***

Die Bundesagentur für Arbeit (BA) arbeitet schon seit vielen Jahren mit riesigen Datensätzen. Inwiefern unterscheiden sich Big Data von diesen administrativen Datensätzen?

Administrative Datensätze entstehen auch als Nebenprodukt anderer Aktivitäten, in diesem Fall der Sozialversicherungsmeldungen. Deshalb werden zum Beispiel in den USA administrative Daten oft auch unter dem Stichwort "Big Data" diskutiert. Wenn zwischen Big Data und administrativen Daten unterschieden wird, dann häufig weil man mit Big Data mindestens drei Eigenschaften verbindet, von denen manche auf administrative Daten nicht ganz zutreffen: *Volume*, also große Datenmengen, haben wir bei den administrativen Daten der BA auch. *Velocity*, also ständig neu anfallende Daten, ist bei administrativen Daten bisher nicht immer gegeben. Meldungen erfolgen in der Regel nicht in Echtzeit, wie das zum Beispiel bei Twitter der Fall ist. Hier erweitert sich der Datensatz permanent mit jedem neuen Tweet. *Variety*, also eine Vielfalt an Datentypen, ist auch nicht unbedingt geben. In der Regel sind die Daten bereits kodiert, es gibt deutlich weniger Freitext, selten Bilder oder Audiofiles – wobei hier Unterschiede zwischen den verschiedenen administrativen Quellen bestehen. Eine vierte Eigenschaft, *veracity*, also die Wahrhaftigkeit der Daten, ist etwas, was bei allen Datentypen eine Rolle spielt. Gerade derzeit wird ja viel von Fake News und Bots gesprochen. Das heißt: Bei Twitter wie bei allen anderen Daten muss man sehr genau schauen, was man da eigentlich misst.

Bei Big Data handelt es sich also häufig um sehr unstrukturierte Daten. Was bedeutet das für die Qualität dieser Daten?

Ob Daten strukturiert oder unstrukturiert vorliegen, sagt zunächst nichts über die Qualität.

Datenqualität kann allerdings nicht unabhängig von der jeweiligen Fragestellung diskutiert werden. Ganz vereinfacht kann man zwischen zwei Typen von Problemen unterscheiden: fehlende Daten für bestimmte Personengruppen, Ereignisse, oder Zeitpunkte, also die Frage "Wie gut sind die Daten für allgemeine Aussagen geeignet?" und die Frage "Messen die Daten eigentlich das, was einen interessiert?".

Können Sie uns ein Beispiel dafür geben?

Wenn ich mich dafür interessiere, wie sich die Gehälter von Männern und Frauen in der Bundesrepublik unterscheiden, dann muss ich sicherstellen, dass ich ausreichend Daten für die gesamte Republik habe. Es ist ja denkbar, dass Gehaltsunterschiede regional oder nach Branche variieren. Wenn ich mich aber dafür interessiere, ob eine neue Kopfschmerztablette für Männer und Frauen gleichermaßen wirkt, ist es möglicherweise weniger wichtig, dass ich Daten aus allen Bundesländern und von Beschäftigten aller Branchen habe – zumindest wenn man davon ausgehen kann, dass die Effekte der Tablette davon unberührt sind. Ob man das misst, was man eigentlich messen möchte, ist gerade bei unstrukturierten Daten eine große Herausforderung. Es gibt tolle Entwicklungen in der Textanalyse, aber je nach Kontext kann es immer noch schwierig sein, einen Algorithmus automatisch entscheiden zu lassen, was ein "like" auf Facebook bedeutet, oder ob ein Satz ernst oder ironisch gemeint war. Einige Sozialwissenschaftler stürzen sich derzeit auf Google-Suchanfragen. Auch hier ist nicht immer klar, warum jemand einen bestimmten Suchbegriff eingibt, und was diese Suche dann im Aggregat bedeutet. Google hatte eine Weile lang großen Erfolg damit, Suchanfragen mit Grippewellen zu korrelieren. Daraus wurde dann ein Index gebildet, um Grippewellen vorherzusagen. Dieser ist dann leider nach ein paar Jahren kein Erfolg mehr, da eine Reihe von Begriffen nur eine Zeit lang hoch korrelierten, inhaltlich aber mit Grippewellen nichts zu tun hatten. Transparenz von Algorithmen und datengenerierenden Mechanismen kann hier helfen.

> ***"Oft entstehen spannende Analysen dann, wenn man verschiedene Datenquellen miteinander verknüpft."***

Lassen sich Big Data auch mit herkömmlichen Daten verknüpfen?

Ja, oft entstehen spannende Analysen dann, wenn man verschiedene Datenquellen miteinander verknüpft. Facebook hat zum Beispiel unglaublich viele Nutzerdaten. Dennoch macht Facebook sehr häufig zusätzliche Umfragen unter seinen Nutzern, weil sie festgestellt haben, dass sie aus dem Verhalten oft nicht genau genug ableiten können, was die Nutzer an

der Plattform mögen oder nicht. In diesem Fall kann eine direkte Verbindung auf Nutzerebene von Big Data und herkömmlichen Daten hergestellt werden. In den meisten Fällen, zum Beispiel bei der Auswertung von Google-Suchbegriffen, ist das nicht so einfach möglich. Die Suchbegriffe stehen in der Regel auf Aggregatebene zur Verfügung, also beispielsweise auf Ebene von Bundesländern oder Städten. Generell sind viele Big-Data-Quellen geokodiert und können über geografische Koordinaten verknüpft werden.

Können Sie auch das anhand eines Beispiels erläutern?

Unter dem Stichwort "Citizen Science" finden sich hier viele Beispiele. Mein derzeitiges Lieblingsprojekt nutzt Fotos, die Besucher von Nationalparks in Afrika machen, und die dazugehörigen Geokoordinaten, um den Bestand von Zebras zu bestimmen. Ideen dieser Art gibt es auch in der Migrationsforschung. So wird zum Beispiel versucht, anhand von Geopositionen bei Twitternachrichten die Bewegungen von Sprachgruppen zu beobachten.

Inwiefern können Big Data einen Mehrwert für die Arbeitsmarktforschung generieren? Teilen Sie die Einschätzung, dass mit Big Data ein goldenes Zeitalter für die Arbeitsmarktforschung anbricht?

Wir haben in der Arbeitsmarktforschung sicher noch nicht alle Datenquellen ausgeschöpft. Die amerikanischen Kollegen hatten zum Beispiel großen Erfolg mit der Auswertung von Twitter-Daten und dem Nachbau eines Indikators für die Arbeitslosenversicherungsstatistik. In Deutschland wird Twitter weit weniger genutzt. Aber die Auswertung von beruflichen Netzwerkplattformen wie Xing und LinkedIn oder die Nutzung der BA-Jobbörse hat gerade erst begonnen, da können wir noch viel lernen. Ob ein goldenes Zeitalter anbricht? Das setzt ja voraus, dass es bisher schlecht stand um die Arbeitsmarktforschung – was man, glaube ich, nicht behaupten kann. Goldenes Zeitalter in dem Sinne, dass wir eine Goldgräberstimmung haben, was man aus Big Data machen könnte und wo Schätze zu heben sind – einschließlich der zu erwartenden Pannen. Auf jeden Fall! (*lacht*)

> ***"Wir haben eine Goldgräberstimmung."***

Wird auch im IAB mit Big Data geforscht?

Ja, ich habe die BA-Jobbörse ja gerade erwähnt, ein Projekt von Christian Hutter setzt sich hiermit seit einiger Zeit auseinander (nähere Informationen dazu finden Sie hier). Andere Projekte sind gerade am Start. Vielleicht ist Ihnen die Marienthal-Studie ein Begriff? Da wurde ja eine ganze Stadt beobachtet und von Hand gemessen und aufgeschrieben, wie sich die

Aktivitäten der Menschen in der Stadt verändert haben, nachdem eine große Fabrik zugemacht hatte. Heute kann man das ganz anders messen, mit unseren Handys haben ja viele von uns bereits ein Messgerät in der Tasche, das man natürlich auch für die Forschung nutzen kann. Zusammen mit dem Bereich von IAB-Forscher Mark Trappmann sind wir gerade dabei, eine App zu konzipieren, die eine moderne Form der Marienthal-Studie erlaubt. Natürlich geht das nur, wenn wir genug Freiwillige finden, die hier mitmachen, und wenn dabei keine zu großen Selektivitäten auftreten. Auch das werden wir untersuchen.

Ist durch die Nutzung von Big Data in der Arbeitsmarktforschung der Datenschutz in Gefahr?

Datenschutz wird am IAB sehr groß geschrieben. Das wird sich im Rahmen von Big Data auch nicht ändern. Spannend ist für uns alle, wie die neue EU-Datenschutzgrundverordnung, die ab Mai 2018 in Kraft tritt, umgesetzt wird. Wichtig ist in meinen Augen, dass alle, die Daten zur Verfügung stellen, ein Grundverständnis darüber haben, was mit ihren Daten passiert, warum es wichtig ist, sie zur Verfügung zu stellen, und welche Risiken es überhaupt gibt. Oft besteht eine diffuse Skepsis, da nicht genug darüber bekannt ist, welche Maßnahmen schon jetzt getroffen werden, um Daten zu schützen. Generell verändert sich aber die Landschaft, wenn immer mehr Daten zur Verfügung stehen, die miteinander verknüpft werden können. Einerseits können wir dank Big Data wertvolle Erkenntnisse zu Tage fördern. Andererseits müssen wir dafür sorgen, dass dies in guter Absicht und im Interesse der Allgemeinheit passiert und der Datenschutz eingehalten wird. Einige Forschungsdatenzentren in Deutschland arbeiten bereits an diesem Themenkomplex – auch hier am IAB. Stefan Bender, der ehemalige Leiter des Forschungsdatenzentrums der BA im IAB, ist Mitherausgeber eines interessanten Buches zu diesem Thema: Privacy, Big Data and the Public Good. Das kann ich allen, die tiefer einsteigen wollen, sehr empfehlen. Diese Themen diskutieren wir übrigens auch auf unserem Podcast #DIGDEEP.

> ***“Wir sind gerade dabei, einen internationalen Weiterbildungsstudiengang in Survey und Data Science aufzubauen.”***

Welche Herausforderungen ergeben sich durch Big Data für die Aus- und Weiterbildung von wissenschaftlichem Personal?

Aufklärung und Sensibilisierung für die Datenschutzthemen und Wahrung der informationellen Selbstbestimmung allgemein liegt mir persönlich sehr am Herzen – genauso wie die Sensibilisierung für die Datenqualität, die wir vorhin schon angesprochen haben. In

der Vergangenheit bestanden Fortbildungen überwiegend im Erlernen von Auswertungstechniken. Das ist immer noch wichtig und hier passiert gerade ganz viel, aber neu hinzu kommt die Notwendigkeit zu wissen, welche Daten sich für welche Fragestellungen eignen, und die datengenerierenden Prozesse zu verstehen. Auch bei der Aufbereitung der Daten, also wie man Bilder oder Texte in Zahlen transformiert, gibt es großen Lernbedarf. Weil das so ist, sind wir gerade dabei, einen internationalen Weiterbildungsstudiengang in Survey und Data Science aufzubauen. Das Projekt wird vom Bundesministerium für Bildung und Forschung gefördert im Rahmen des Bund-Länder-Wettbewerbs „Aufstieg durch Bildung: offene Hochschulen“. Das IAB ist hier ein Partner, der Hauptträger ist die Universität Mannheim, die zusammen mit der Universität Maryland diesen Studiengang in Survey und Data Science organisiert (survey-data-science.net).

Die Fragen stellte Dr. Martin Schludi.