Big Data aus der Sicht eines Data Scientist

von Oliver Bracht, Janosch Schobin

Oliver Bracht im Gespräch mit Janosch Schobin

Soziopolis: Was verstehen Sie unter dem Begriff Big Data?
Oliver Bracht: Lassen Sie mich vorab sagen, dass der Begriff Big Data in unserer Firma gar nicht so wichtig ist. Wir sprechen eher von Data Science, was ich als eine Mischung aus Statistik, Softwareentwicklung und Kommunikation in Form von Datenvisualisierungen beschreiben würde. Im Rahmen von Data Science haben wir es heute jedoch oft mit Big Data zu tun. Klassisch wird Big Data dabei vor allem anhand der Datenmenge, der Aktualisierungsfrequenz und der Struktur der Daten definiert. Big Data beginnt dann im Bereich von einigen Millionen Datenpunkten und bezieht sich häufig auf eine Kombination strukturierter und unstrukturierter Daten. Mit unstrukturierten Daten meine ich Daten, die nicht auf die Verarbeitung durch statistische Verfahren zugeschnitten sind. Ein Beispiel wären Bilder aus Blogs, Tweets und so weiter.
Hadley Wickham, einer der weltweit führenden Entwickler der Statistiksprache R, verwendet folgende Definition: Big Data liegen dann vor, wenn bei ihrer Erfassung die Kognitionszeit geringer ist als die Rechenzeit. Unter der Kognitionszeit versteht er die Zeit, die ein Mensch braucht, um das Analyseverfahren zu entwickeln. Die Rechenzeit ist dann die Zeit, die der Rechner braucht, um die Analyse durchzuführen. Die Umkehrung des Verhältnisses von Rechen- und Kognitionszeit umreißt aus meiner Sicht das praktische Problem ganz gut, mit dem uns Big Data als Data Scientists konfrontiert.
Ein weiterer wichtiger Aspekt ist aus meiner Sicht die Bedeutung der Daten, zum Beispiel für die Entwicklung neuer Geschäftsmodelle oder für die Entscheidungsfindung. Um das strategische Potenzial von Big Data zu erhöhen, werden statistische Modelle entwickelt, die beispielsweise Maschinenausfälle oder Abverkaufsquoten prognostizieren können. Erst diese Verfahren machen Big Data wertvoll, indem sie die relevanten Informationen extrahieren, die sich in der Datenmasse verbergen. Sowohl die Wertschöpfung als auch die Positionierung auf komplexen Märkten hängt heute immer stärker von der Datenverwertung ab. Viele Firmen gehen in diesem Zusammenhang dazu über, datengetrieben – *data driven* – zu operieren.
Wie äußert sich diese Datengetriebenheit? Was verändert das aus Ihrer Sicht?
Datengetriebenheit bedeutet vor allem, dass Entscheidungen heute grundsätzlich stärker evidenzbasiert getroffen werden. Das 'Bauchgefühl' der Manager verliert dagegen an Bedeutung. Lassen Sie mich das an einem Beispiel erläutern: Eine Einzelhandelskette bietet in regelmäßigen Abständen Aktionsware zu günstigen

Preisen an, die für eine bestimmte Zeit in den Märkten erhältlich ist. Diese Aktionsware wird lange im Voraus zentral in großer Stückzahl eingekauft und dann auf die Filialen verteilt. Welche Artikel sollen wann zu welchen Stückzahlen und zu welchem Preis in den Filialen vorgehalten werden? Das ist eine hochkomplexe Fragestellung. Nehmen Sie zum Beispiel den Preis: Es reicht nicht aus, einen optimalen Preispunkt für ein bestimmtes Produkt zu finden. Sie müssen auch berücksichtigen, wie er sich auf das sonstige Sortiment auswirkt und auf das Sortiment der umliegenden Filialen. Das wird schnell unübersichtlich. Bislang werden diese Fragen meist von leitenden Mitarbeitern auf Grundlage einfacher deskriptiver Analysen historischer Daten, vor allem aber auf Basis ihrer Erfahrung beantwortet. Datengetriebene Unternehmen gehen anders vor: Sie kombinieren Daten aus Kassen- und Warenwirtschaftssystemen, ziehen Daten aus der Anzeigenschaltung hinzu und reichern diese internen Daten noch um weitere externe Informationen an. Das ist übrigens auch ein schönes Beispiel für Big Data: eine große Menge sich häufig ändernder Daten aus heterogenen Quellen. Auf dieser Datenbasis können mithilfe komplexer Data-Mining-Algorithmen Modelle erstellt werden, mit deren Hilfe die oben genannten Fragen sehr differenziert zu beantworten sind. Jede Filiale kann so mit ihrem individuell optimalen Sortiment bestückt werden.

Ein anderer Aspekt, der sich aus meiner Sicht total verändert, ist das Verhältnis der Unternehmen zu ihrer technischen Infrastruktur. Analysiert werden ja zunehmend nicht nur Daten, die Menschen in ihrer Interaktion mit digitalen Systemen produzieren, sondern auch und gerade Daten, die Maschinen durch Schnittstellen ihrer Sensorik mit digitalen Systemen generieren. Die Stichworte „Industrie 4.0“ und „Internet of Things“ gehören in diesen Kontext. Big Data ist dafür ebenfalls relevant, denn eine einzige Maschine ist heute oft mit hunderten Sensoren ausgestattet, die im Millisekundentakt Daten produzieren.

In einem unserer Projekte haben wir beispielsweise die von der Sensorik einer Maschine generierten Daten in ein Modell eingespeist, das dann die aktuelle Ausfallwahrscheinlichkeit der Maschine prognostiziert. Die Techniker, die das Gerät nutzen, bekommen automatisch in Echtzeit eine Einschätzung der aktuellen Ausfallwahrscheinlichkeit. Die Prognose setzt die gemessenen Sensordaten gewissermaßen in den Kontext der Geschichte nicht nur dieser, sondern aller baugleichen Maschinen. Wir installieren also eine Art Entscheidungsunterstützung. Diese kleine Veränderung hat enorme Konsequenzen. Mithilfe der Ausfallwahrscheinlichkeit kann ein Techniker die Maschine warten, bevor sie kaputtgeht. Das ist einerseits für Industrieunternehmen sehr attraktiv, weil sich Ausfallzeiten – und damit Kosten – einsparen lassen. Anderseits bedeutet es, dass die genaue Kenntnis einer spezifischen Maschine an Bedeutung verliert. Es schiebt sich ein Algorithmus als Medium zwischen den Techniker und die Maschine. Die Erfahrung des Technikers, der die Maschine betreut, wird so tendenziell weniger wichtig. Ersetzbar wird er dadurch – das ist ja häufig die große Angst – meiner Einschätzung nach aber nicht.

Warum nicht? Frey und Osborne[1] etwa schätzen den Anteil der heute durch digitale Automatisierung bedrohten Arbeitsplätze in den USA auf über ein Drittel. Technologisch produzierte Arbeitslosigkeit ist wieder ein großes Thema.

Dass sich die Arbeitswelt durch zunehmende Automatisierung ganz allgemein und Big Data im Besonderen verändert, will ich gar nicht in Abrede stellen. Das Phänomen ist aber nicht neu. Technische Innovation hat immer auch den Wegfall von Arbeitsplätzen zur Folge. Was aber bisher nicht weggefallen ist und meiner Einschätzung nach auch in Zukunft nicht wegfallen wird, ist die Arbeit selbst. Wegfallende Arbeitsplätze werden mindestens mittelfristig durch neue, andersartige Arbeitsplätze ersetzt. Letztere erfordern oft ein höheres Bildungsniveau. Darauf müssen wir uns einstellen.

Die Angst, Arbeitsplätze seien durch Big Data bedroht, wird unterfüttert durch die Vorstellung, dass – etwas überspitzt formuliert – algorithmisch gesteuerte Roboter die Menschheit unterwerfen. *The rise of the robots* von Martin Ford[2] ist das wohl populärste Beispiel für diesen Topos. „Maschinelle Intelligenz" übersteigt schon heute in bestimmten Bereichen die Leistungsfähigkeit von Menschen. Es besteht aber aus meiner Sicht ein fundamentaler Unterschied zu menschlicher Intelligenz. Menschliche Akteure können im Gegensatz zu Algorithmen unbekannte und unerwartete Kontextinformationen interpretieren, die ihr Handeln beeinflussen. Sie kennen den Sinn ihrer Handlungen. Das macht in der Regel keinen Unterschied, hat ganz plötzlich aber eine enorme Bedeutung. Nehmen Sie die spektakuläre Notlandung im Hudson River im Jahr 2009. Nach dem Start waren bei einem Passagierflugzeug in geringer Höhe beide Triebwerke durch Vogelschlag ausgefallen. Der Pilot entschied sich für eine Notwasserung mitten in New York. Kein algorithmischer Autopilot hätte diese Entscheidung so getroffen. Kontextinterpretierende und sinnverstehende Algorithmen sehe ich in weiter Ferne, vielleicht sind sie sogar ein Widerspruch in sich. Vollautomatische, algorithmisch gesteuerte Systeme sind deshalb nur in Umwelten sinnvoll, die wenig kontextuelle Interpretation erfordern. Viel häufiger wird es jedoch erforderlich sein – und so verstehe ich unsere Aufgabe als Data Scientists vornehmlich –, dass wir Menschen möglichst transparente und nachvollziehbare algorithmisch gesteuerte Entscheidungshilfen an die Hand geben.

Sie sagen immer wieder Data Scientist und nicht Statistiker oder Methodiker. Warum? Ist das nur eine Sprachmode?

Keineswegs. Der Beruf eines Data Scientist – übriges ein Berufsfeld, das auch Soziolog_innen gute Chancen bietet – unterscheidet sich von dem eines Statistikers erheblich. Statistik und statistische Programmierung sind natürlich der Kern unserer Tätigkeit. Das Repertoire der Verfahren, die wir als Data Scientists verwenden, geht dabei sogar weit über das hinaus, was in der Statistikausbildung einer Fachdisziplin gelehrt wird. Aber Statistik allein ist nicht genug. Kehren wir zu dem Beispiel mit den Einzelhandelsketten zurück. Um Data Science in den Tagesablauf von Unternehmen einzubetten, muss die Analyse in die bestehende IT-Landschaft integriert werden. Es reicht nicht aus, Ergebnisberichte in PowerPoint zu produzieren. Die automatisierte Kommunikation mit anderen technischen Systemen ist somit ein zweites zentrales Aufgabenfeld, das traditionell in den Bereich der Informatik fällt. Schließlich geht es darum, die Ergebnisse der Analysen handlungsrelevant zu machen. Das bedeutet, dass wir die uninterpretierten Daten in Informationen verwandeln müssen, die bei ihren Nutzern auf Akzeptanz treffen.

Dazu gehören verschiedene Abwägungen: Sehr komplizierte statistische Modelle

lassen sich oft nur schwer kommunizieren. Gibt es ein einfacheres Modell, das bei geringen Einbußen in seiner Vorhersagekraft besser verständlich ist? Wie kommunizieren wir die Erkenntnisse aus den implementierten Modellen? Besonders visuelle Lösungen spielen dabei eine wichtige Rolle. Das ist eher etwas, mit dem sich Grafikdesigner befassen. Wir müssen nicht einfach nur gute Modelle für Daten finden, sondern sie visuell so zum Sprechen bringen, dass jeder im Unternehmen, der sich mit ihnen befassen muss, sie versteht. Wenn Sie so wollen, ist ein Data Scientist auch ein Datendiplomat, der dafür Sorge trägt, dass die Datengetriebenheit auch auf die operative Ebene übertragbar ist.

Lassen Sie uns noch einmal etwas abstrakter auf die Datengetriebenheit zurückkommen: Datengetriebenheit kann als eine Art algorithmisch gestützter Empirismus verstanden werden. Man könnte zu der Einschätzung gelangen, dass sie mit einer Entwertung von theoretischem Wissen und ganz allgemein von Theorie, soziologischer und anderweitiger, einhergeht. Wer datengetrieben Sachverhalte analysiert, fragt sich nicht, warum dieser oder jener Algorithmus etwas so gut prognostiziert. Sind für Sie die Kausalitäten nicht von Belang?

Die Tatsache, dass die meisten Data-Mining-Verfahren hypothesenfrei sind, bedeutet nicht, dass die inhaltliche Interpretation keine Rolle spielt. Ich glaube, das ist ein weit verbreitetes Missverständnis. Inhaltliche Interpretation ist sehr wohl von großer Bedeutung. In den Daten finden sich oft Artefakte, also Zusammenhänge, die so in der Realität nicht existieren. Sie entstehen, wenn Daten unvollständig oder fehlerhaft sind, und lassen sich nur durch eine inhaltliche Prüfung erkennen. Ich glaube nicht daran, dass Data Mining blind, also ohne Kenntnis der Kontexte, funktioniert. Deshalb arbeiten in unseren Projektteams immer auch Experten aus den jeweils betroffenen Fachbereichen mit, meist durch unsere Auftraggeber gestellt.

Was bedeutet diese Zunahme datengetriebener, algorithmischer Systeme aus Ihrer Sicht für die Soziologie und besonders für die soziologische Ausbildung?

Die tatsächlichen und potenziellen Veränderungen, die algorithmische Systeme mit sich bringen, betreffen sicherlich den Kernbereich der Soziologie. Die derzeitige Diskussion dort bezieht sich vor allem auf Fragen des Datenschutzes und der Privatsphäre. Das sind zweifellos wichtige Aspekte, für die wir als Data Scientists eine besondere Verantwortung tragen. Was mir jedoch fehlt, ist die theoretische Ausdifferenzierung. Die Soziologie kann einen Beitrag dazu leisten, die Diskussion auf eine breitere Basis zu stellen. Was sich aus meiner Sicht dringend verändern muss, ist die Methodenausbildung an den Universitäten. Die ist viel zu stark an die wissenschaftliche Feldforschung angelehnt. Das soll nicht heißen, dass man gleich alles über Bord werfen muss, was bisher gemacht wurde. Die Notwendigkeit der klassischen Feldforschungsinstrumente wird auch durch Big Data nicht verschwinden. Aber Machine-Learning-Verfahren[3] beispielsweise, die der Dreh- und Angelpunkt der hier besprochenen algorithmischen Verfahren sind, kommen in der Methodenausbildung praktisch nicht vor. Man könnte denken, das liegt an den hohen Anschaffungskosten für Machine-Learning-Software. Das Gegenteil ist jedoch der Fall, denn fast jede Software, die für Data Science relevant ist, ist Open-Source-Software. Vor allem R, aber auch Python oder Julia als Analysesoftware und

Hadoop als System der Datenhaltung von Big Data sind kostenlose Open-Source-Programme. Auch die allermeisten Algorithmen sind in diesen Programmen frei zugänglich.
An vielen Fakultäten wird jedoch traditionell mit proprietären (also lizenzpflichtigen) Programmen gearbeitet, vor allem mit SPSS und Stata. Zwar ist ein Trend zur stärkeren Verwendung vor allem von R erkennbar, die Veränderungen verlaufen aber zu langsam. Immerhin ist R unter Nachwuchswissenschaftlern bereits heute weit verbreitet. Etablierte Hochschullehrer sind jedoch selten motiviert, die von ihnen favorisierte Software auszutauschen. Die USA sind uns in dieser Hinsicht weit voraus. Dort ist R in den Hochschulen schon lange Standard, was einer der Gründe ist, weshalb die USA weltweit führend bei der Entwicklung und Operationalisierung algorithmischer Systeme sind. Solche Entscheidungen sind richtungsweisend, und sie sind nicht nur für die Methodenausbildung der Soziologie wichtig. Wenn die Prämisse stimmt, dass wir in einer Welt leben, in der unsere soziale Umwelt vermehrt algorithmisch mitgeneriert wird, dann müssen Soziolog_innen in ihrem Studium lernen, wie die wichtigsten Algorithmen unserer Zeit funktionieren.
Dieser Beitrag ist Teil eines Themenschwerpunkts zu Big Data. Weitere Texte finden Sie hier.

Fußnoten

1 Carl Benedikt Frey / Michael A. Osborne, The future of employment: How susceptible are jobs to computerization, Oxford 2013. LINK

2 Martin Ford, The rise of the robots. Technology and the Threat of a Jobless Future, New York 2015.

3 Unter „Machine Learning“ werden Verfahren zusammengefasst, mit denen Geräte Zusammenhänge weitgehend selbständig erkennen können. Häufig verwendete Algorithmen aus diesem Bereich sind beispielsweise Künstliche Neuronale Netze, Random Forest oder Support Vector Machines.