Barbara Engels

Big-Data-Analyse - Ein Einstieg für Ökonomen

Big Data ist eines der spannendsten Mysterien der Gegenwart. Jeder redet davon, keiner weiß, wie es genau geht, alle geben vor, es zu tun – so hat es der Ökonom Dan Ariely mal formuliert. Dieser Kurzbericht gibt einen Überblick darüber, inwiefern Big-Data-Analysen in Wirtschaftswissenschaft und Politikberatung bereits eingesetzt werden und welche Nutzungspotenziale es gibt.

Big Data ist gekennzeichnet durch eine Reihe von Superlativen, die sich auf fünf Vs reduzieren lassen: Volume, Velocity, Variety, Veracity und Value(z.B. Marr, 2015). Die Analyse von Big Data erfordert die Speicherung und Verarbeitung von riesigen Datenmengen (Volume), wobei die Datenrate immer höher wird (Velocity). Es gibt eine Vielzahl von unterschiedlichen Datenquellen und –formen (Variety). Veracity ist besonders für Wissenschaftler ein sine qua non: Um eine hohe Qualität und Vertrauenswürdigkeit der Daten zu gewährleisten, müssen besondere Informationsextraktionsverfahren angewandt werden; die Datenaufbereitung ist aufwendig. Der Value der Daten bezeichnet ihre Wertigkeit im Sinne der Vorhersagbarkeit. Etwa 90 Prozent der digital produzierten Daten sind unstrukturiert (etwa Vijayan, 2016). Interaktionen im Web, Transaktionsdaten und Daten von Sensoren sind zwar Big und Fast Data, aber nicht gleich Smart Data – Letzteres erfordert eine sorgfältige Aufbereitung und Analyse.

Die Datenquellen, die derzeit für ökonomische Big-Data-Analysen genutzt werden, sind nicht so vielfältig, wie sie sein könnten, wie eine Untersuchung von 58 europäischen Big-Data-Initiativen mit 111 Datenquellen zeigt. 35 Prozent der von den Initiativen genutzten Datenquellen sind administrativer Art, 22 Prozent sind Statistikbüros, 15 Prozent basieren auf Umfragen. Neuere Datenquellen wie Sensoren (13 Prozent), Konsumentendaten (9 Prozent) und soziale Medien (7 Prozent) spielen noch eine eher untergeordnete Rolle (Poel et al., 2015, 22). Generell steigt die Zahl der in ökonomischen Analysen verwendeten Quellen, außerdem wird es immer beliebter, verschiedene Datensätze zu kombinieren (data linking).

Die Hälfte der in den Big-Data-Projekten verwendeten Datensätze ist offen oder halb-offen, das heißt die Nutzung erfordert eine Registrierung. Open Data spielt also eine zentrale Rolle. Open Data ist ein Konzept, bei dem maschinenlesbare und strukturierte Informationen, besonders aus der Verwaltung, durch

den Einsatz offener Nutzungsrechte frei verwendet, weiterverarbeitet und verbreitet werden können. Geoinformationsdaten, Umwelt-, Transport- und Haushaltsdaten, beispielsweise verfügbar über das europäische Datenportal www.europeandataportal.eu oder das deutsche Pendant www.govdata.de, dienen zahlreichen Analysezwecken. Big Data ist in diesem Fall vor allem auf Volume bezogen.

Daten aus dem Internet, die nicht so strukturiert und gebündelt in sogenannten Dumps oder über ein Application Programming Interface (API) vorliegen wie Open Data, fordern hingegen die Veracity heraus. Liest man beispielsweise über sogenannte Crawler Daten von Webseiten aus (das so genannte Spidern), oft auch iterativ, wird die Datenaufbereitung zentral, da oft auch falsche Informationen extrahiert werden.

Big-Data-Analysen bedeuten nicht unbedingt, dass völlig neue Arten der Analyse angewandt werden. Bisher hat sich die Forschung im Bereich Big Data vor allem auf deskriptive Analysen fokussiert (mehr als 70 Prozent laut Poel et al., 2015, 28). Trendanalysen sind laut dieser Studie die zweithäufigste Analyseart. Neuere Formen wie Text und Sentiment Mining spielen noch eine untergeordnete Rolle. Text Mining extrahiert analysierbare Informationen aus Text (z.B. Worthäufigkeiten), Sentiment Mining fängt Stimmungen zu einem Thema ein und kategorisiert Inhalt etwa in positiv oder negativ oder auf einer Werteskala. Das ermöglicht beispielsweise die Analyse der tatsächlichen Reaktion der Bevölkerung auf eine bestimmte politische Maßnahme, denn online äußern sich Menschen unter Umständen ehrlicher als in einer direkten Umfrage (Ceron/Curini/Iacus/Porro, 2014).

Allerdings fängt etwa die Analyse von Twitter-Daten nur einen geringen Teil der Bevölkerung und damit nur bestimmte Stakeholder ein. Sie ist außerdem aufwendig und fehleranfällig: Die informationsextrahierenden Algorithmen sind nicht in der Lage, Sarkasmus oder einen kulturellen Kontext zu verstehen oder Tippfehler zu übergehen. In der Anwendung ist Text Mining sehr vielfältig. Beispielsweise kann damit untersucht werden, wie Kundenrezensionen in Onlineshops die Absatzzahlen beeinflussen (zum Beispiel Ghose/Ipeirotis, 2010), oder inwiefern Annahmen von Privatpersonen über bestimmte zukünftige wirtschaftliche Entwicklungen einen Effekt auf reale Größen wie Zinsen haben (zum Beispiel Meinusch/Tillmann, 2015).

Eine weitere, immer häufiger genutzte Form der Big-Data-Analyse ist das sogenannte Nowcasting. Nowcasting nutzt Informationen, die früher oder häufiger zur Verfügung stehen als die eigentliche Variable des Interesses. Es liefert damit eine sehr zeitnahe Prognose, die als „early estimate“ dienen kann. Google-Suchanfragen nach Begriffen wie „Arbeitsamt“ oder „Arbeitslosengeld beantragen“ werden etwa analysiert, um eine zeitnahe Aussage über die Entwicklung des Arbeitsmarktes zu treffen. Diese kann außerdem die zeitverzögert erscheinenden amtlichen Statistiken ergänzen beziehungsweise verifizieren (zum Beispiel Askitas/Zimmermann, 2011; Tuhkuri, 2014).

Generell eignen sich Big-Data-Analysen, um die traditionellen Statistiken zu ergänzen. Beispielsweise können Google-Suchanfragen nach bestimmten Produkten einen Konsumindikator bilden (Schmidt/Vosen, 2011). Standortdaten von Handys zeigen an, wie viele Menschen sich tatsächlich zu einer bestimmten Zeit an einem bestimmten Ort aufhalten (De-Facto-Bevölkerung), was etwa bei der Stadt- und Verkehrsmittelplanung helfen kann (Masso, 2016).

Abstriche müssen bei der Kausalität der Analysen gemacht werden. Oft legen Big-Data-Analysen lediglich Korrelationen offen. Dies ist jedoch in vielen Fällen für Politikberatung ausreichend. Andere Partikularitäten der Big-Data-Analyse sind Selektion und Endogeneität. Je nach Datenquelle wird nur die Online-Bevölkerung erfasst, der so genannte Digital Divide kommt zum Tragen. Generell ist das Online-Verhalten auch nicht notwendigerweise ein adäquater Spiegel des Verhaltens in der Offline-Welt. Kognitive Verzerrungen gibt es etwa bei Google-Suchanfragen, bei der das Auto-Vervollständigen

die Suchanfrage eines eigentlich neutralen Nutzers in eine bestimmte Richtung lenkt (Askitas, 2016). Der Wissenschaftler muss außerdem der Versuchung widerstehen, die Unmengen an vorhandenen Daten wahllos zu analysieren. Er muss relevante von vorhandenen Daten unterscheiden können. Letzteres ist möglicherweise die größte Herausforderung von ökonomischen Big-Data-Analysen.

Literatur

Askitas, Nikos, 2016, Big Data Is a Big Deal But How Much Data Do We Need?, IZA Discussion Paper No. 9988, http://ftp.iza.org/dp9988.pdf [12.11.2016]

Askitas, Nikos / Zimmermann, Klaus F., 2009, Google Econometrics and Unemployment Forecasting, Applied Economics Quarterly, Jg. 55, Nr. 2, S. 107–120, http://ftp.iza.org/dp4201.pdf [11.11.2016]

Askitas, Nikos / Zimmermann, Klaus F., 2009, Google Econometrics and Unemployment Forecasting, Applied Economics Quarterly, Jg. 55, Nr. 2, S. 107–120, http://ftp.iza.org/dp4201.pdf [11.11.2016]

Ceron, Andrea / Curini, Luigu / Iacus, Stefano / Porro, Giuseppe, 2014, Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France, New Media & Society , Jg . 16, Nr. 2, S. 340–358

Ghose, Anindya / Ipeirotis, Panagiotis G., 2010, Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics, IEEE Transactions on Knowledge and Data Engineering, http://ieeexplore.ieee.org/document/5590249/ [9.11.2016]

Masso, Jaan, 2016, The potential of big data for migration research: Internet data and mobile positioning data, Presentation at the expert workshop "Using Big Data to Advance Policy Research", CEPS, Brüssel, 23.5.2016

Marr, Bernard, 2015, Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance, Chichester

Meinusch, Annette / Tillmann, Peter, 2015, Quantitative Easing and Tapering Uncertainty: Evidence from Twitter, MAGKS Papers on Economics, Philipps-Universität Marburg, Faculty of Business Administration and Economics, Department of Economics (Volkswirtschaftliche Abteilung), http://EconPapers.repec.org/RePEc:mar:magkse:201509 [9.11.2016]

Poel, Martijn / Schroeder, Ralph / Treperman, Jérôme / Rubinstein, Mor / Meyer, Eric / Mahieu, Bea / Scholten, Chiel / Svetachova, Marina, 2015, Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking, http://www.data4policy.eu/state-of-the-art-report [11.11.2016]

Schmidt, Torsten / Vosen, Simeon, 2011, Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. Journal of Forecasting, Jg. 30, Nr. 6, S. 565-578

Tuhkuri, Joonas, 2014, Big data: Google Searches Predict Unemployment in Finland, https://www.etla.fi/en/publications/33195/ [17.11.2016]

Vijayan, Jaikumar, 2016, Solving the Unstructured Data Challenge, http://www.cio.com/article/2941015/big-data/solving-the-unstructured-data-challenge.html [17.11.2016]