

Big Data: mehr, schnell, gut genug¹ Gerhard Klocker

Illegal Conversions ist eines der Probleme, mit dem sich die Stadtverwaltung von New York herumzuschlagen hat.² Vermietete Wohnungen werden in so viele kleine Einheiten aufgeteilt, dass bis zu zehnmal mehr Mieter aufgenommen werden (können), als für die übliche Wohnung vorgesehen wäre. Diese Praxis ist nicht erlaubt und führt u. a. zu erhöhtem Brandrisiko. Aber wie kann die Stadtverwaltung damit umgehen?

In den vergangenen Jahren gab es jährlich an die 25.000 Anzeigen von *Illegal Conversions*, aber nur 200 Inspektoren, die ihnen nachgehen konnten. Das waren zwar erfahrene Experten, aber das Problem der erhöhten Brandgefahr konnte nicht gelöst werden. Es schien keine probate Methode zu geben.

New York setzte ein Team von Datenwissenschaftlern auf die Problematik an. Diese entwickelten ein System, in welches die Datenbestände unterschiedlicher Behörden eingespeist wurden: von den Grundstücken New Yorks, Informationen über Art und Baujahr des Gebäudes über Rückstände bei der Bezahlung der Grundsteuern oder Zahlungsverzug bei den Stromrechnungen bis hin zu Informationen über Notarztbesuche, Verbrechensraten und Rattenbefall.

Danach wurde nach Mustern und Korrelationen gesucht, um daraus abzuleiten, welchen Anzeigen am dringlichsten nachzugehen war. Das System funktionierte: Die Inspektoren konnten ihre Effizienz verfünffachen. Während sie früher selbst eingeschätzt hatten, welchen Anzeigen sie nachgingen und es damit in 13 Prozent der Fälle zu Räumungsanordnungen kam, erreichten sie jetzt aufgrund der *Big Data-Vorhersagen* eine Quote von 70 Prozent. Die Arbeit wurde für die Inspektoren befriedigender, weil sie sich auf die wirklich schweren Fälle konzentrieren können und damit verstärkt großes Unheil abwenden können.

Ein weiteres anschauliches Beispiel liefert der *H1N1-Grippevirus*, der im Jahr 2009 neu entdeckt wurde und sich schnell ausbreitete.³ Die Gesundheitsbehörden warnten bereits vor der Gefahr einer möglichen Pandemie. Für Ärzte wurde eine Meldepflicht für neue Grippefälle eingeführt. Seitens der Gesundheitsbehörden ging es darum, Informationen über das Fortschreiten der Epidemie zu erhalten. Die Daten waren allerdings, bis sie genutzt werden konnten, immer schon veraltet, weil einerseits die meisten Menschen nicht sofort zum Arzt gingen, wenn sie sich krank fühlten, und andererseits auch die Übermittlung der Meldungen seine Zeit benötigte. Es kam zu ein bis zwei Wochen Zeitverzögerung, was bei einer sich rasch ausbreitenden Epidemie eine Ewigkeit ist und die Gesundheitsbehörde nur sehr beeinträchtigt handlungsfähig machte. Die Lösung brachte Google. Google war aufgrund seines Datenmaterials im Stande, Aussagen über die Ausbreitung des *H1N1-Virus* zu machen. Die Google-Suchmaschine erhielt täglich drei Milliarden entsprechender Anfragen, aufgrund dessen Rückschlüsse auf die Ausbreitung gemacht werden konnten. Es ging jedoch nicht ausschließlich um die inhaltliche Qualität der Suchbegriffe, z. B. Begriffe wie „Medikamente gegen Husten und Fieber“ usw., sondern das System suchte Korrelationen zwischen der Häufigkeit bestimmter Suchbegriffe und der Ausbreitung der Grippewelle. Damit konnte Google genauso gut wie die US-amerikanische

Gesundheitsbehörde die Ausbreitung feststellen, aber eben nicht mit zwei Wochen Verspätung, sondern unmittelbar.

„*Big Data – Die Revolution, die unsere Leben verändern wird*“ heißt das aktuelle Buch von Mayer-Schönberger und Cukier. In ihm finden sich viele Beispiele für das, was unter *Big Data* verstanden wird. Sie kommen zu dem Schluss, dass wir mit *Big Data* am Anfang grundlegender Umwälzungen stehen, die sich auf das Selbstverständnis und die Organisation unserer Gesellschaft auswirken werden.

Was ist Big Data?

Big Data ist, wenn Daten sprechen lernen. Das ist so neu auch nicht, geht es doch immer darum, aus der Fülle der „nichtssagenden“ Daten, die uns zur Verfügung stehen, Erkenntnisse und Einsichten zu gewinnen und einen Sinn zu entwickeln.

Die große Veränderung ist die Quantität, die Menge an Daten, die zur Verfügung steht. Da ist im großen etwas möglich, was im kleinen Maßstab nicht funktioniert. Die Quantität der Daten erlaubt, zu neuen Erkenntnissen zu gelangen.

Die Kosten für Speicherplatz von Daten sinken, die Qualität der Analysemethoden steigt. Die Geschwindigkeit der Verarbeitung steigt ebenfalls. Der Umfang der gesammelten Daten wird weiterhin sprunghaft wachsen – und mit diesen Daten lässt sich vieles machen.

Mehr, unscharf, gut genug⁴

„**Mehr**“

Dadurch, dass uns mehr und mehr Daten zur Verfügung stehen, lässt sich auch mehr aus ihnen generieren. 2011 wurden weltweit erstmals 1,8 Zettabyte an Daten produziert. Auch wenn man sich 1,8 Zettabyte wahrscheinlich

nicht vorstellen kann, beeindruckend ist die Zahl dennoch und vor allem die Prognosen, dass sich diese Datenmenge alle zwei Jahre verdoppelt wird.⁵ Bisher war es nicht möglich, über (fast) alle Daten zu spezifischen aber auch unterschiedlichen Themen zu verfügen. Wir halfen und helfen uns mit Stichproben (Samples). Und schließen daraus wieder zurück auf die Grundgesamtheit. Weil wir mehr Daten haben ($N = \text{alle}$), kann man Details erkennen, die vorher nicht erkennbar waren. In diesem Fall bringt uns das „Mehr“ an Daten neue Einblicke, die mit der Stichprobenmethode nicht möglich sind.

„Unschärfe“

Wenn wir sehr viele oder annähernd alle Daten haben, dann können diese auch unschärfer sein, d. h. weniger exakt oder von unterschiedlicher Qualität. *„Interessantes Wissen liegt nicht nur in strukturierten Daten, sondern auch in vielen anderen Datenquellen, die semi- oder unstrukturierte Daten liefern.“*⁶ Das können Maschinendaten aus der Produktion, Sensordaten in Gebäuden oder der Umwelt, das World Wide Web oder Social Media-Daten aus Facebook, Twitter oder Blogs u. a. m. sein.

Im beschriebenen Beispiel der *Illegal Conversions* wurden sehr unterschiedliche Datenbestände miteinander in Beziehung gesetzt. Die Unschärfe dieser Daten wird bewusst in Kauf genommen, weil es oft darum geht, eine generelle Richtung oder ein Muster erkennen zu können. *„Wir geben die Exaktheit nicht gänzlich auf, sondern nur unsere Versessenheit darauf. Was wir an Genauigkeit auf der Mikroebene verlieren, gewinnen wir an Erkenntnis auf der Makroebene.“*⁷

„Gut genug“ bzw. Korrelation

Die dritte zentrale Eigenschaft von Big Data ist die Chance, Muster und Korrelationen zu entdecken. Mit Big Data lässt sich nicht feststellen, *warum* etwas passiert, sondern *was* passiert. Solche Korrelationen helfen uns, den Zusammenhang verschiedener Phänomene zu erkennen und zu beschreiben, und sie erlauben

Vorhersagen für die Zukunft zu treffen. Es geht nicht um Sicherheit, sondern es geht um Wahrscheinlichkeit – und das kann sehr wertvoll sein. Folgendes Beispiel unterstreicht die Bedeutung von Korrelationsanalysen: Der kanadischen Computerwissenschaftlerin Carolyn McGregor ist es damit gelungen, 24 Stunden früher, als es bisher möglich war, zu erkennen, wenn frühgeborene Kinder Infektionen bekommen. *„Das hängt damit zusammen, dass sie in den Vitalwerten eines Frühgeborenen ein gewisses Muster ablesen kann, das auf eine beginnende Infektion hinweist. Sie sammelte dafür 1.200 Datenpunkte pro Sekunde. Sie kann zum Schluss zwar nicht sagen, warum die Babies erkranken, aber um eine Infektion zu bekämpfen reicht das Wissen, dass sie wahrscheinlich erkranken werden, völlig aus.“*⁸

Mehr Daten, Unschärfe bezüglich der Datenqualität und -struktur, blitzschnelle Verarbeitung und vor allem das Entdecken von Korrelationen, das sind die Ingredienzien für kommende Veränderungen. Die Vorhersagen, die aufgrund dieser Korrelationen möglich sind, bilden das Fundament für Big Data und zukünftige Entwicklungen.

Die Risiken

Doch spätestens seit den Veröffentlichungen von Edward Snowden und der NSA-Spionage-Affäre sind wir uns darüber klar, dass Big Data mit vielen Unsicherheiten und großen Gefahren verbunden ist.

Im Kleineren erleben wir derzeit die Diskussion über die intelligenten Stromzähler, die *Smart Meters*. Dabei werden rund um die Uhr, teilweise in Intervallen von nur sechs Sekunden, Daten zum Stromverbrauch aufgezeichnet. Mit diesen Daten lassen sich Korrelationen entdecken, die Rückschlüsse auf den Tagesablauf der Haushalte und ihrer Bewohner, deren Gesundheit und Verhalten bis hin zu illegalen Aktivitäten offenbaren können. Gleichzeitig können uns die *Smart Meter* beim sorgsamsten Umgang mit unserer Energie helfen.

Ein anderes Beispiel ist der *Social Graph*, d. h. die Spur, die wir durch unsere Aktivitäten im Internet hinterlassen und die für Dritte sehr interessant und von Nutzen sein kann. Das Werkzeug ist als Suchfunktion auf Facebook eingerichtet und ermöglicht, die Daten und Beiträge des Freundeskreises oder aber aller Nutzer auf Facebook nach Zusammenhängen zu durchsuchen: „*Personen, die gerne tanzen und in meinem Ort wohnen*“ oder „*alle Fotos von NN, die mit ‚Party‘ verknüpft sind*“.

Wo bleibt der Schutz unserer Privatsphäre? Nur zu sagen: „*Ich erteile meine Zustimmung nicht*“ oder „*Ich bin auf keiner Social Media Plattform dabei*“ wird zu wenig sein, um ausreichend geschützt zu sein.

Mayer-Schönberg vergleicht die *Big Data-Analysen* mit dem Einsatz von Antibiotika. Diese können einerseits lebensrettende Wirkung haben, aber nimmt man zu viel davon, resultiert das in problematischen Resistenzen. „*Der Gesetzgeber muss Unternehmen und sich selbst Fesseln anlegen und festlegen, wer wie Big Data verwenden darf.*“⁹ Hier gibt es noch viel zu tun. Es bleibt zu beobachten, wie schnell wir auf die technologische Entwicklung und die wirtschaftliche Nutzung der Möglichkeiten gesellschaftlich reagieren werden bzw. können.

Was wird kommen

Die Datafizierung wird weiter zunehmen

Immer mehr Dinge unseres täglichen Lebens werden digitalisiert, immer mehr Daten stehen zur Verfügung. Das Fraunhofer-Institut IAIS sieht drei zentrale Chancen für die Wirtschaft:¹⁰

- **Effizientere Unternehmensführung durch bessere Auswertungen und Vorhersagen:** Die Energiebranche kann besser prognostizieren, wann wie viel Strom benötigt wird. Der Einzelhandel kann besser vorhersagen, wann welches Pro-

dukt verkauft wird und wieder nachbestellt werden muss.

- **Massenindividualisierung:** Dadurch, dass während der Bearbeitung einer Anfrage oder Bestellung relevante Informationen über den Kunden vom System „mitgelernt“ werden, können die Dienstleistungen künftig auf die einzelne Person individualisiert werden. Als Beispiele werden nutzungsbezogene Versicherungsprodukte, tägliche Gesundheitsdiagnosen, vorausschauender Service bei Maschinen usw. genannt.
- **Intelligente Produkte:** Durch die enorme Geschwindigkeit der Datenverarbeitung können zum Beispiel Flugzeugturbinenhersteller bereits nach der Landung analysieren, welche Turbinen in die Inspektion müssen. Mittels der direkten Verarbeitung von Sensordaten werden diese Möglichkeiten zunehmen: selbstlernende Thermostate, selbstregulierende Häuser, autonom fahrende Fahrzeuge sind keine Zukunftsmusik mehr.

Die Risiken von Big Data sind groß und werden uns noch sehr beschäftigen

Datensicherung und Datenschutz wurden schon genannt. Eine Gefahr liegt in der Beurteilung von Menschen aufgrund von individualisierten *Big Data-Vorhersagen*: So könnten Krankenversicherer mittels *Big Data-Analyse* erkennen, wer aufgrund eines Gendefekts wahrscheinlich bestimmte Krankheiten bekommen wird. Damit könnten sie beschließen, die betreffende Person nicht oder nur mit sehr hoher Prämie zu versichern. Es wird viel Bewusstseinsbildung, aber auch Auflagen, gesetzliche Regelungen bis hin zu Verboten benötigen, um Big Data kontrollierbar zu halten.

Kreativität, Expertise und Erfahrung werden weiterhin erfolgskritisch bleiben

Die Formen von Fachwissen und Erfahrung, die sich auf die langjährige Ansammlung von Wissen berufen, aber

nicht bereit sind, eingefahrene Denkmuster zu überwinden, werden unter Druck geraten. Big Data und Expertise verbunden mit Kreativität werden die Quellen für den Erfindergeist der Zukunft sein. Das Illegal Conversions-Beispiel ist bezeichnend: Die *Big Data-Analysen* kamen zu besseren Ergebnissen als die erfahrenen Experten. Die Instruktoren konnten umlernen. Es gelang ihnen, die neuen Erkenntnisse zu nutzen und die datenbasierten Entscheidungen mit ihrem Know-How zu verbinden. In der Verbindung von Kreativität und Expertise mit den *Big Data-Analysen* entsteht das neue Wissen und die Anwendungsoptionen.

Mayer-Schönberger/Cukier sprechen von einer gesellschaftlichen Revolution, die mit der verstärkten Nutzung von Big Data einhergehen wird. „*Der Boden unter unseren Füßen ist in Bewegung. Alte Gewissheiten, die schon immer Gültigkeit zu haben schienen, werden infrage gestellt ... Eine Weltsicht, die auf Kausalität beruht, wie wir glaubten, wird von der Überlegenheit der Korrelationen bedroht.*“¹¹ Kausalität reicht nicht mehr aus. Auch die Auseinandersetzung mit Big Data zeigt, dass es um das Erkennen von Mustern, um Korrelationen geht und dass wir verstärkt kausales Denken mit relationalem Denken verbinden müssen.

Literatur

- 1 In Anlehnung an Mayer-Schönberger/Viktor/Cukier, Kenneth: Big Data, Die Revolution, die unser Leben verändern wird, 2013, 20.
- 2 vgl. ebenda, 234ff.
- 3 vgl. ebenda, 7ff.
- 4 vgl. ebenda, 20ff.
- 5 Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAI: Big Data – Vorsprung durch Wissen, Innovationspotenziale, 2012, 2, http://www.iais.fraunhofer.de/fileadmin/user_upload/Abteilungen/KD/uploads_BDA/FraunhoferIAIS_Big-Data-Analyse_Doku.pdf (14.01.2014).
- 6 Grosser, Tim/Bange, Carsten: Big Data – Neue Ansätze für die Analyse poly-strukturierter Daten, 2012, 20; in CeBIT Guide Business Intelligence 2012, 20 – 23, http://www.isis-specials.de/profile_pdf/1b989_ed3_bi0112.pdf (14.01.2014).
- 7 Mayer-Schönberger/Cukier, 2013, 22.
- 8 vgl. <http://derstandard.at/1379293481682/Big-Data-Duerfen-nicht-Minority-Report-werden> (10.01.2014).
- 9 vgl. <http://derstandard.at/1379293481682/Big-Data-Duerfen-nicht-Minority-Report-werden> (10.01.2014).
- 10 IAI, 2012, 7.
- 11 Mayer-Schönberger/Cukier, 2013, 239.