Aktueller Begriff
Big Data

Jüngste Enthüllungen um internationale Datenspionage haben den Blick auch auf die unter dem Stichwort „Big Data“ bekannt gewordenen neuen Möglichkeiten im Umgang mit großen Datenmengen gelenkt. Dabei geht es nicht um eine einzelne neue Technologie. Vielmehr bezeichnet Big Data ein Bündel neu entwickelter Methoden und Technologien, die die Erfassung, Speicherung und Analyse eines großen und beliebig erweiterbaren Volumens unterschiedlich strukturierter Daten ermöglicht. Für die IT-Branche wie auch die Anwender in Wirtschaft, Wissenschaft oder öffentlicher Verwaltung ist Big Data daher zum großen Innovationsthema der Informationstechnik geworden.

Daten sind heute im Wesentlichen durch drei Charakteristika gekennzeichnet, die ihren englischen Bezeichnungen zufolge als die „drei Vs“ bezeichnet werden. Dabei handelt es sich zum einen um die **Datenmenge (*Volume)***, die durch die fortschreitende Digitalisierung praktisch aller Bereiche des modernen Lebens in unvorstellbar großen Quantitäten produziert wird und sich etwa alle zwei Jahre verdoppelt. So wurden Schätzungen zufolge in diesem Jahr (2013) bereits über 2 Trilliarden Bytes an Daten weltweit gespeichert – was auf iPads gespeichert und gestapelt eine 21.000 km lange Mauer ergäbe. Ein weiteres Charakteristikum heutigen Datenverkehrs ist seine **Geschwindigkeit *(Velocity)***: Während früher Daten in bestimmten Abständen anfielen, die es erlaubten, sie nach und nach zu verarbeiten, ist man heute aufgrund von Vernetzung und elektronischer Kommunikation dem Datenfluss ununterbrochen ausgesetzt. Um sie nutzen zu können, müssen die einlaufenden Informationen immer schneller oder sogar in „Echtzeit“ aufgenommen und analysiert werden. Das dritte wichtige Merkmal ist die **unterschiedliche Beschaffenheit *(Variety)*** der heute in so vielfältig und komplex strukturierten Quellen wie z.B. sozialen Netzwerken, Fotos, Videos, MP3-Dateien, Blogs, Suchmaschinen, Tweets, Emails, Internet-Telefonie, Musikstreaming oder Sensoren „intelligenter Geräte“ vorkommenden Daten. Besonders interessant z.B. für Werbung, Marketing oder auch Wahlkämpfe sind dabei subjektive Äußerungen in Text- oder Wortbeiträgen aller Art, die Stimmungen oder Meinungen ausdrücken. Um letztere maschinenlesbar zu machen, werden Programme benötigt, die wertende Aussagen über Produkte, Marken u.ä. oder sogar Emotionen erkennbar machen, was technisch besonders herausfordernd ist.

Die wirtschaftliche Bedeutung von Daten wird inzwischen als so groß angesehen, dass diese neben Arbeitskraft, Ressourcen und Kapital als „**vierter Produktionsfaktor**“ angesehen werden. Denn der Wert von Erkenntnissen, die durch Auswertung vorhandener Daten gewonnen werden können, gilt als potentiell gewaltig. So versprechen sich **Unternehmen** unter anderem verbesserte Marketingmethoden oder auch neue Produktentwicklungen durch genauere Informationen über das Informations- und Konsumverhalten ihrer Kunden sowie Kostenersparnisse durch optimierte Logistikprozesse. Auch für die **öffentliche Verwaltung** sind die neuen Möglichkeiten interessant, wie erste Erfahrungen zeigen - z. B. beim Verkehrsmanagement in **Stockholm**. Dort konnten durch die Integration von Wetter- und Verkehrsdaten (Unfall- und Staumeldungen, Videos usw.), Verkehrsaufkommen und Emissionen um 20% und Fahrzeiten um 50% gesenkt werden. Auch der amerikanische Präsident Obama baute bei seinem Wahlkampf 2012 auf Big Data und beschäftigte in seinem

Nr. 37/13 (06. November 2013)

Team fast 50 Datenanalytiker. Ihnen gelang es, mit Hilfe detaillierter Datenanalyse aus vielen Quellen die Wahlkampagne erheblich zu effektivieren, indem sie auf die Bundesstaaten und Zielgruppen konzentriert wurde, die – mit den für sie in Inhalt und Form jeweils passenden Botschaften – am ehesten überzeugt werden könnten. Auch die **wissenschaftliche Forschung** baut zunehmend auf die neuen Methoden der Datenanalyse. So stützen erste Erfahrungen mit Big Data-Anwendungen auf **medizinischem Gebiet** die Vision einer nicht mehr reaktiven, sondern präventiven und personalisierten Medizin, die durch die genaue Kenntnis individueller Risikofaktoren, subjektiver Befindlichkeiten und möglicher Nebenwirkungen verabreichter Medikamente möglich werden würde. Nach Schätzungen des McKinsey Global Institute wären durch den Einsatz von Big Data allein im US-amerikanischen Gesundheitswesen Effizienz- und Qualitätssteigerungen im Wert von ca. 222 Mrd. € und für den gesamten **öffentlichen Sektor in Europa** von jährlich 250 Mrd. € möglich.

Das Besondere bei Big Data-Analysen ist vor allem die neue Qualität der Ergebnisse aus der **Kombination bisher nicht aufeinander bezogener Daten**. In der Regel sind dies Bestandsdaten, die zu 85% bislang technisch nicht ausgewertet werden konnten. Zu den **technischen Voraussetzungen** für Big Data-Analysen gehören vor allem die zwei Neuentwicklungen *MapReduce* und *Hadoop.* Letzteres ist eine Open Source-Software und Plattform, die aus einem Forschungsprojekt der Firma Yahoo hervorging und mittlerweile faktisch als Standard-Anwendung im Big Data-Bereich gilt. *Hadoop* ermöglicht es, schnell und dezentral große Datenmengen zu speichern und parallel zu bearbeiten. Dies wird durch ein Verteilsystem von Datenspeichern erreicht, durch das jeder Nutzer mit Netzanschluss große Datenmengen auf Gruppen oder Cluster von Rechnern verteilen und später wieder schnell auf sie zugreifen kann. Die eigentliche mathematische Analyse der Daten erfolgt dann durch den ursprünglich von Google entwickelten Algorithmus *MapReduce*, der sehr große Datenmengen parallel bearbeiten kann, indem sie zerlegt und auf zahlreichen Rechnern verteilt werden. Damit wird auch deutlich, dass Big Data-Analysen aufgrund der großen Datenmengen in der Regel nicht ohne dezentrale Speicherorte *(Clouds)* möglich sind. Auf der Basis der beiden frei erhältlichen zentralen Elemente von Big Data-Technologien wurden zwischenzeitlich diverse Erweiterungen und Werkzeuge entwickelt, die auch als externe Software-Dienstleistung angeboten werden, so dass die Nutzung finanziell und organisatorisch einer größeren Anzahl von Unternehmen möglich wird.

Neben den unbestritten großen Potentialen von Big Data für Wirtschaft, Wissenschaft und Gesellschaft werden in der zunehmend intensiver geführten Debatte über die neuen Möglichkeiten auch **kritische Stimmen** laut. Denn gerade die Nutzung der für Big Data besonders interessanten personenbezogenen Daten kollidiert mit zentralen **europäischen datenschutzrechtlichen Prinzipien,** wie dem Recht auf informationelle Selbstbestimmung, dem Schutz personenbezogener Daten und der Zweckbindung von erhobenen Daten, kodifiziert in der Europäischen Grundrechtecharta und dem Bundesdatenschutzgesetz. Auch eine Pseudonymisierung oder Anonymisierung von Daten ist hier nur von begrenztem Nutzen, weil die für Big Data typische Kombination vieler Datensätze häufig eine **De-Anonymisierung** ermöglicht. Einige Beobachter richten zudem den Blick auf die möglichen Auswirkungen auf unser **wissenschaftliches Weltbild**, in dem die Ergründung und die Wichtigkeit kausaler Zusammenhänge nun zunehmend durch statistische Korrelationen abgelöst werden könnte. Und schließlich bleibt zu fragen, wo in einer Welt, in der Entscheidungen zunehmend von datenverarbeitenden Maschinen dominiert werden, die **menschliche Urteilsfähigkeit** oder auch Intuition ihren Platz finden kann. Denn diese könnte manchmal auch nahelegen, bei bestimmten Entscheidungen eben gerade nicht der Datenlage zu folgen.

Quellen:

– S. Heuer. Kleine Daten, große Wirkung. Digitalkompakt Nr.6. Landesanstalt für Medien NRW 2013.
– K. Cukier / V. Mayer-Schönberger. The Rise of Big Data. In: Foreign Affairs 5/6 2013, S. 28 – 40.
– T. Weichert. Big Data und Datenschutz. Unabh. Landeszentrum für Datenschutz Schl.-Holst.: 3-2013.

Verfasserin: Sabine Horvath – Fachbereich WD 10, Kultur, Medien und Sport