

Mit dem Schlagwort ‚Big Data‘ werden umfangreiche Datenbestände bezeichnet, die mit herkömmlichen Softwarewerkzeugen kaum mehr zu bewältigen sind. Die Daten sind meistens unstrukturiert und stammen aus unterschiedlichen Quellen: Mitteilungen (Postings) aus sozialen Netzen, E-Mails, elektronische Archive mit Multimedia-Inhalten, Anfragen aus Suchmaschinen, Dokumentsammlungen aus Content Management Systemen, Sensordaten beliebiger Art, Kursentwicklungen von Börsenplätzen, Daten aus Verkehrsströmen, Satellitenbilder, Messdaten von den Geräten des Haushalts (Smart Meter); Bestell-, Kauf- und Bezahlvorgänge elektronischer Shops, Daten aus eHealth-Anwendungen, Aufzeichnungen von Monitoring-Systemen etc.

Für den Begriff Big Data gibt es noch keine verbindliche Definition, doch die meisten Datenspezialisten berufen sich auf mindestens drei V's: Volume (umfangreicher Datenbestand), Variety (Vielfalt von Datenformaten; strukturierte, semi-strukturierte und unstrukturierte Daten; siehe Abb. 2.1) und Velocity (hohe Verarbeitungsgeschwindigkeit, ev. Echtzeitverarbeitung).

Im IT-Glossar der Gartner Group findet sich folgende Definition:

Big Data: ‚Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making‘¹.

Die Gartner Group geht soweit, Big Data als Informationskapital oder Vermögenswert (information asset) des Unternehmens zu betrachten. Tatsächlich müssen Unternehmen und Organisationen entscheidungsrelevantes Wissen generieren, um

¹Gartner Group, IT Glossary – Big Data; <http://www.gartner.com/it-glossary/big-data/>, abgerufen am 24. Juli 2017.

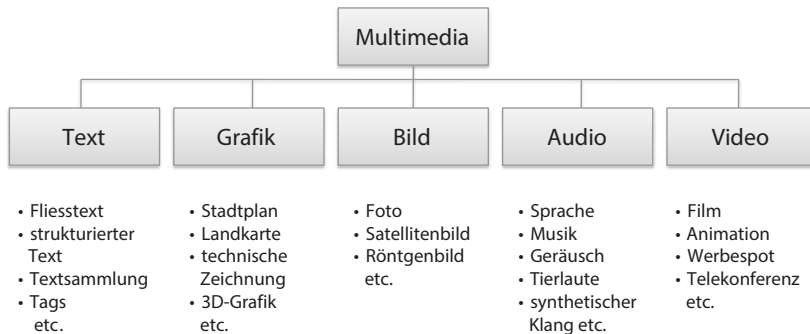


Abb. 2.1 Vielfalt der Quellen bei Big Data. (Quelle: Meier und Kaufmann 2016)

überleben zu können. Dabei setzen sie neben den eigenen Informationssystemen vermehrt auf die Vielfalt der Ressourcen im Web, um ökonomische, ökologische und gesellschaftliche Entwicklungen im Markt besser antizipieren zu können.

Big Data ist nicht nur eine Herausforderung für profitorientierte Unternehmen im elektronischen Geschäft sondern auch für das Aufgabenspektrum von Regierungen, öffentlichen Verwaltungen, NGO's (Non Governmental Organizations) und NPO's (Non Profit Organizations).

Als Beispiel seien die Programme für Smart City oder Ubiquitous City erwähnt, d. h. die Nutzung von Big Data Technologien in Städten und Agglomerationen. Ziel dabei ist, den sozialen und ökologischen Lebensraum nachhaltig zu entwickeln. Dazu zählen z. B. Projekte zur Verbesserung der Mobilität, Nutzung intelligenter Systeme für Wasser- und Energieversorgung, Förderung sozialer Netzwerke, Erweiterung politischer Partizipation, Ausbau von Entrepreneurship, Schutz der Umwelt oder Erhöhung von Sicherheit und Lebensqualität.

Zusammengefasst geht es bei Big Data Anwendungen um die Beherrschung der folgenden drei V's:

- **Volume:** Der Datenbestand ist umfangreich und liegt im Tera- bis Zettabytebereich (Megabyte = 10^6 Byte, Gigabyte = 10^9 Byte, Terabyte = 10^{12} Byte, Petabyte = 10^{15} Byte, Exabyte = 10^{18} Byte, Zettabyte = 10^{21} Byte).
- **Variety:** Unter Vielfalt versteht man bei Big Data die Speicherung von strukturierten, semi-strukturierten und unstrukturierten Multimedia Daten (Text, Grafik, Bilder, Audio und Video; vgl. Abb. 2.1).
- **Velocity:** Der Begriff bedeutet Geschwindigkeit und verlangt, dass Datenströme (Data Streams, vgl. Kap. 7) in Echtzeit ausgewertet und analysiert werden können.

In der Definition der Gartner Group wird von Informationskapital oder Vermögenswert gesprochen. Aus diesem Grunde fügen einige Experten ein weiteres V zur Definition von Big Data hinzu:

- **Value:** Big Data Anwendung sollen den Unternehmenswert steigern. Investitionen in Personal und technischer Infrastruktur werden dort gemacht, wo eine Hebelwirkung besteht resp. ein Mehrwert generiert werden kann.

Viele Angebote für NoSQL-Datenbanken sind Open Source. Zudem zeichnen sich diese Technologien aus, dass sie mit kostengünstiger Hardware auskommen und leicht skalierbar sind. Auf der anderen Seite besteht ein Engpass bei gut ausgebildetem Personal, denn das Berufsbild des Data Scientist (siehe Kap. 9) ist erst im Entstehen. Entsprechende Ausbildungsangebote sind in Diskussion oder in der Pilotphase.

Ein letztes V soll die Diskussion zur Begriffsbildung von Big Data abrunden:

- **Veracity:** Da viele Daten vage oder ungenau sind, müssen spezifische Algorithmen zur Bewertung der Aussagekraft resp. zur Qualitätseinschätzung der Resultate verwendet werden. Umfangreiche Datenbestände garantieren nicht per se bessere Auswertungsqualität.

Veracity bedeutet in der deutschen Übersetzung Aufrichtigkeit oder Wahrhaftigkeit. Im Zusammenhang mit Big Data wird damit ausgedrückt, dass Datenbestände in unterschiedlicher Datenqualität vorliegen und dass bei Auswertungen dies berücksichtigt werden muss. Neben statistischen Verfahren existieren unscharfe Methoden des Soft Computing, die einem Resultat oder einer Aussage einen Wahrheitswert zwischen 0 (falsch) und 1 (wahr) zuordnen (vgl. Fuzzy Datenbanken in der Forschungsliteratur resp. die Buchreihe Fuzzy Management Methods des Springer Verlags von Meier et al. 2017).

Werkzeuge der digitalen Wirtschaft: Big Data, NoSQL &
Co.

Eine Einführung in relationale und nicht-relationale
Datenbanken

Meier, A.

2018, IX, 60 S. 20 Abb., Softcover

ISBN: 978-3-658-20336-8