I Daten

Joy's Law:
Most of the smartest people work for someone else.

1 Datenexplosion

Erst mit der Erfassung, Verarbeitung und Verwendung von Unmengen von Daten können das **Internet der Dinge und Dienste**, **Cyber Physische Systeme** (CPS) und **Soziale Medien** ihre Stärken entfalten.

Unter **Internet der Dinge (Internet of Things, IoT**) versteht man die Verbindung von IT-Systemen mit weltweiten Subsystemen, Prozessen, Objekten, Lieferanten sowie Kunden. All diese verbundenen Einheiten kommunizieren über das Internet miteinander und mit dem Menschen (Ashton (2009) oder Rheingold (2000)). **Internet der Dinge und Dienste (Internet of Things and Services, IoTaS)** erweitert den Begriff IoT um die angebotenen Dienstleistungen zur Unterstützung der Teilnehmer.

Nach Baheti und Gill (2011) oder Broy (2010) ist ein **Cyber Physisches System (Cyber Physical System, CPS**), ein Verbund informatischer, softwaretechnischer Komponenten mit mechanischen und elektronischen Teilen insbesondere Sensoren und Aktoren, die über eine Dateninfrastruktur wie das Internet, kommunizieren, Maßnahmen einleiten und sich und andere steuern können. Insbesondere werden im CPS eingebettete Systeme mit IP-Adressen ausgestattet und über das Internet in der Regel drahtlos miteinander verbunden. CPS gilt als technologischer Enabler von **Industrie 4.0** (▶ Kap. 8). Das neue Internetprotokoll IPv6 stellt über 340 Sextillionen ($2^{128} \approx 3{,}4 \times 10^{38}$) unterschiedliche Adressen für Computer, Smartphones, Maschinen, Transportbehälter und weitere intelligente Dinge zur Verfügung und schafft damit die kommunikationstechnischen Voraussetzungen für CPS.

Eingebettete Systeme (Embedded Systems) sind in Dingen eingebaute Minicomputer, die in der Lage sind, über Sensoren Daten wie z. B. die Temperatur zu erheben. Mit Hilfe von Programmen werden diese Daten verarbeitet, um daraus Maßnahmen einzuleiten. Eingebettete Systeme sind durch das Zusammenwirken zwischen Mechanik, Elektronik, Software und Hardware geprägt. Embedded Systems sind zentrale Bausteine von Cyber Physischen Sytemen sowie **intelligent vernetzten Dingen** (**Smart Connected Things**) (▶ Kap. 4). Eingebettete Systeme sind in den letzten Jahren billiger und in ihrer Bauteilgröße kleiner geworden. Oftmals gewinnen sie die für den Betrieb erforderliche Energie direkt aus der Umgebung, indem sie z. B. Licht oder Vibrationen in Energie umwandeln. Eine gute Einführung in eingebettete Systeme gibt Bens et al. (2010). Unter **Ubiquitäres Computing** oder **Pervasive Computing** versteht man etwas Ähnliches wie Embedded Systems allerdings mit dem Schwerpunkt auf Alltagsgegenstände, in die drahtlos vernetzte Minicomputer und Sensoren eingebaut sind, die autonome computerbasierte Dienste bereitstellen.

Machine-to-Machine Comunication (M2M) bezeichnet den Datenaustausch zwischen Endgeräten. Dieser Datenaustausch muss nicht wie bei Internet of Things über das Internet erfolgen. Eigentlich existiert M2M seit Bestehen der Automatisierungstechnik, auch wenn sich der Name erst später etabliert hat. Bis jetzt hat sich kein Standard für M2M durchgesetzt. Eine gute Einführung in das Thema M2M wird in Glanz und Büsgen (2013) gegeben.

Cyber Physical Systems (**CPS**) tragen insbesondere durch die rasant ansteigende Anzahl **intelligenter vernetzter Dinge**, aber auch die unterschiedlichsten **digitalen Plattformen** dazu bei, dass die weltweit verfügbaren Datenbestände weiterhin explosionsartig wachsen. Nach Velten und Janata (2012) sind Cloud Systeme, Software as a Service, Sensoren, Social Media, Mobile Devices und Location based Services die Haupttreiber für die Datenexplosion.

Eine **Cloud** ist eine virtualisierte IT-Ressource für Speicherung, Analyse und Verwaltung von Daten oder für die Bereitstellung von Diensten, Rechenleistung oder IT-Anwendungen. Die Cloud wird durch einen Serviceanbieter verwaltet und in der Regel werden die Cloud-Dienste über einen Internetzugang bereitgestellt. Cloud-Computing ermöglicht eine bedarfsgerechte Bereitstellung von Daten und Services über das Internet. Cloud-Lösungen gewähren häufig höhere IT-Sicherheit als firmengewachsene IT-Strukturen. Eine hohe Qualität der Cloud ist gegeben, wenn eine hohe stabile Verfügbarkeit, eine hohe IT-Sicherheit und entsprechende Schnittstellen gegeben sind. Cloud-Techniken befähigen zur orts- und zeitunabhängigen Informationsbereitstellung. Neue Methoden wie Big Data oder Advanced Analytics bedienen sich häufig vorteilhafter Cloud-Architekturen (Schmidt und Möhring (2013)).

Soziale Medien (Social Media) sind digitale Medien – meist Internetplattformen, die es Nutzern ermöglichen, sich untereinander auszutauschen, sich anderen mitzuteilen und Inhalte zu erstellen sowie diese zu teilen. Typische Vertreter sozialer Medien sind Facebook, Twitter, Instagram, Wikipedia, Youtube, Second Life, Snapchat aber auch berufliche Netzwerke wie LinkedIn oder XING.

Soziale Medien, Internetplattformen und Suchmaschinen führen über ihre User genaue Profile, die User teilweise sehr detailliert in Bezug auf Vorlieben, Kaufverhalten, Interessenbereiche, Einstellungen, Werte, Tagesablauf, Freunde, Termine, Aufenthaltsorte u. a. beschreiben. Dieses Wissen wird für ortsabhängige, situative und personalisierte Werbung sowie Erstellung von personalisierten Produkt- oder Serviceangeboten verwendet. Darüber hinaus werden Suchergebnisse, Postings oder Bewertungen zu politischen, religiösen oder gesellschaftlichen Themen so gefiltert, dass dem User das präsentiert wird, was seinem Profil, seiner Meinung und seiner Einstellung entspricht, genannt **Filter Bubble** (Pariser (2011)).

Die Beherrschung von Datenmanagement, Big Data sowie Advanced Analytics (▶ Kap. 2 und 3) sind Grundvoraussetzungen, um die verfügbaren rasant wachsenden Datenbestände auch sinnvoll nutzen zu können. Die Services und Anwendungen von Cyber Physischen Systemen oder des Internets der Dinge und Dienstleistungen profitieren wiederum genau von diesen rasant wachsenden Datenbeständen.

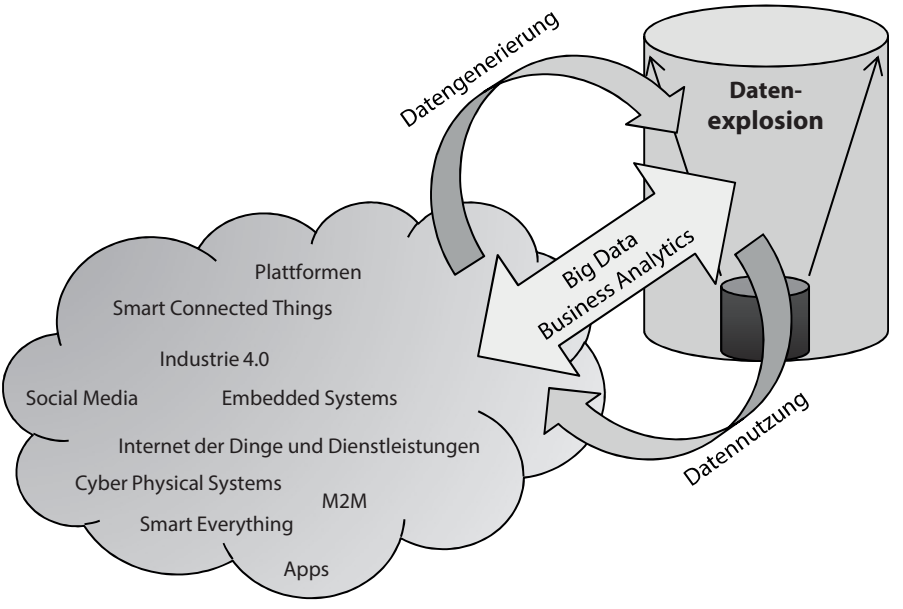

Abb. 1.1: Spirale des Datenbestandswachstums

Es ist anzunehmen, dass wir gerade am Beginn einer sich noch verstärkenden Datenbestandswachstumsspirale, basierend auf dem Kreislauf

- Datengenerierung durch IoTaS sowie CPS inkl. Plattformen, Social Media, Apps, Smart Everything, Industrie 4.0, Smart Connected Things, Embedded Systems, Machine-to-Machine Communication (M2M),
- Transformation der Daten durch Big Data und Business Analytics,
- Nutzung der Daten für Services, Anwendungen und Entscheidungen in den Bereichen IoTaS sowie CPS inkl. Plattformen, Social Media, Apps, Smart Everything, Industrie 4.0, Smart Connected Things, Embedded Systems, M2M, …

stehen. Im vorliegenden Buch werden dieser Kreislauf und seine Elemente, bezogen auf die Wertschöpfung und hier insbesondere die Produktion von Wirtschaftsgütern, diskutiert.

Seit der Jahrtausendwende ist die explosionsartige Entwicklung des Datenbestandes festzustellen (Jodlbauer (2016a)). Im Jahr 1997 schätzte der angesehene Informatiker Mike Lesk (1997), dass die Menschheit etwa 12.000 Petabyte (1 Petabyte $= 10^{15}$ Byte) an Daten geschaffen und in Bibliotheken, Museen, digitalen Speichermedien, Archiven bis hin zu privaten Datenablagen gespeichert hat. Im Jahr 2015 (in diesem Zeitraum hat sich das Smartphone weltweit verbreitet und viele Services im Bereich IoTaS sowie CPS am Markt etabliert) sind über 8.500 Exabyte (1 Exabyte $= 10^{18}$ Byte) neue digitale Daten pro Jahr vor allem im Internet zu verzeichnen. In anderen Worten: Im Jahr 2015 wurden in

nur zwölf Stunden gleich viele digitale Daten neu geschaffen wie die gesamte Menschheit von der Urzeit bis zum Jahr 1997 in Summe unter Berücksichtigung aller Medien (Buch, Bilder, digitale Datenträger etc.) an Daten generiert hat. Heute produzieren wir in nur zwei Stunden digitale Daten, deren Umfang höher ist als der Datenumfang aller jemals gedruckten Bücher. Wir leben heute in einer Welt, in der sich die verfügbaren Daten etwa alle zwei Jahre verdoppeln. Der Vergleich des Datenwachstums mit dem Bevölkerungswachstum sowie der Entwicklung des Buchbestandes untermauert die rasante Entwicklung der Datenbestände.

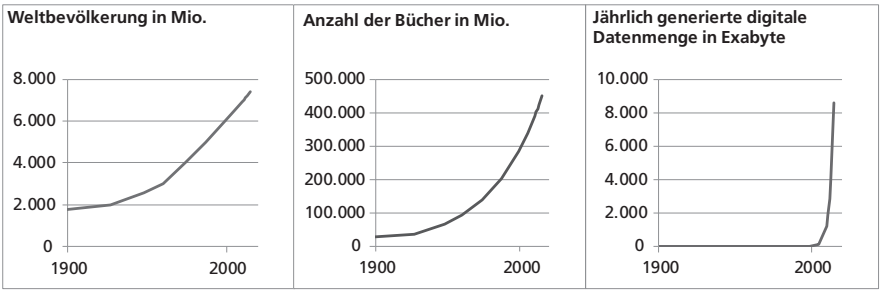

Abb. 1.2: Entwicklung der Weltbevölkerung, des Buchbestandes und der digitalen Datenmenge

Die Weltbevölkerung ist in gut 100 Jahren von unter zwei Milliarden auf knapp 8 Milliarden angewachsen – dies entspricht einem durchschnittlichen jährlichen Zuwachs von knapp 1,5 %. Die Anzahl der weltweiten Bücher hat sich von geschätzten 20 Millionen auf fast eine halbe Milliarde in einem Jahrhundert entwickelt – damit durchschnittlich etwa 10 % jährlicher Zuwachs. Die neu generierten digitalen Daten sind dahingegen explodiert: Von 130 Exabyte im Jahr 2005 auf 8.500 Exabyte, also 8,5 Zetabyte (1 Zetabyte = 10^{21} Byte) , im Jahr 2015, das entspricht einer jährlichen durchschnittlichen Zuwachsrate von 50 %. Laut Jodlbauer (2016a) schätzen Experten, dass bereits ab 2020 das Datenvolumen 100 Zetabyte überschreiten wird. Bemerkenswert ist übrigens die Nutzung der Daten. Experten schätzen, dass maximal 3 % der weltweit vorhandenen Daten konkret analysiert, genutzt oder verwertet werden. Es existiert damit ein enormes ungenutztes Potenzial in den Daten.

Die Welt der Daten ist durch vier Megatrends (Wrobel et al. (2014)) charakterisiert:

- Digitale Konvergenz
- Ubiquitäre intelligente Systeme
- Nutzererzeugte Inhalte
- Verknüpfte Daten

Digitale Konvergenz beschreibt das Zusammenwachsen unterschiedlichster Bereiche, Ebenen und Systeme. Das gleichzeitige Handhaben von Text, Bild, Musik und Video im

Medienbereich ist ebenfalls ein Aspekt der digitalen Konvergenz wie die allgegenwärtige Nutzung des gleichen Computers oder des gleichen Smartphones im Beruf, in der Freizeit, beim Einkaufen, beim Spielen usw. Manche Autoren sprechen auch von Entgrenzung (Jodlbauer (2016a)).

Gleichzeitig kann in Ergänzung zu Wrobel et al. (2014) eine **Digitale Divergenz** festgestellt werden. **Filter Bubble** bezeichnet eine Technologie mit der einem Internetnutzer vorzugsweise jene Inhalte gezeigt werden, die er bzw. sie gerne sehen möchte. Studien belegen das mit modernen Psychometrieverfahren kombiniert mit Big Data und Advanced Analytics wenige Postings, Suchabfragen und Likes genügen, um mit hoher Wahrscheinlichkeit Hautfarbe, sexuelle Orientierung, politische Einstellung, Religionszugehörigkeit, Intelligenz, Alkoholkonsumverhalten, Familienstand u. v. a. korrekt bestimmen zu können (Bachrach et al. (2012)). Bei einer solchen personalisierten Suche werden persönliche Einstellungen, Werte, Überzeugungen usw. gestärkt und Argumente gegen die eigene Meinung und Sichtweise unterdrückt. Gesellschaftlich kann dies zur Polarisierung und Entzweiung führen.

Ubiquitäre intelligente Systeme, **Smart Connected Things** und **Cyber Physical Systems** sind intelligent vernetzte Dinge, die mit vielfältiger Sensorik und Aktorik ausgestattet sind und Teil unseres täglichen Lebens geworden sind. Die Sensorik ermöglicht den intelligenten Dingen die Wahrnehmung der Umwelt und die Aktorik das aktive Beeinflussen der Umwelt. Verkehrstechnikanlagen, Autos, Smartphones, bereits viele Haustechnikgeräte oder Haushaltsgeräte und natürlich Maschinen, Werkzeuge und Anlagen der Industrie sind Beispiele dieser intelligenten vernetzten Dinge.

In der Frühphase des Internets stellten wenige Anbieter Inhalte für viele User ins Internet. Heute erstellt de facto jeder User Inhalte für viele User. Dies bewirkt neben der Explosion der Daten auch **nutzererzeugte Daten**. Diese nutzererzeugten Daten beschreiben teilweise sehr detaillierte Aspekte des Lebens (privat als auch beruflich), das Freizeitverhalten, das Konsumverhalten, die Erfahrung mit einem Produkt, die Zufriedenheit mit einer Dienstleistung, den Aufenthaltsort oder andere häufig für Dritte relevante und verwertbare Gegebenheiten.

Erst **verknüpfte Daten** führen in vielen Anwendungen zum erhofften Mehrwert. Felddaten aus dem Produkteinsatz richtig kombiniert mit Wetter-, Standort- oder Daten aus Sozialen Medien können zu einem besseren Verständnis der Vergangenheit, zu treffsichereren Prognosen und zu besseren Entscheidungen führen, als wenn isoliert nur die Felddaten herangezogen werden. Dabei ist die sinnvolle Verknüpfung und damit der semantische Zusammenhang unterschiedlicher Daten zentraler Punkt (Bizer et al. (2009)).

Die vier Megatrends der Daten nach Wrobel et al. (2014) ermöglichen die **digitale Transformation oder Digitalisierung der Wertschöpfung**. Im Gegensatz zum Deutschen unterscheidet das Englische die zwei Bedeutungen von **Digitalisierung**:

- digitization
- digitalization

Digitization meint die Konvertierung eines anlogen Signals (mit kontinuierlichen Werten) in ein digitales Signal (mit diskreten Werten). Digitization entstammt techni-

schen Disziplinen wie der Signalverarbeitung oder Elektrotechnik. Der Begriff **Digitalization** ist den Sozialwissenschaften zuzuordnen und bezieht sich auf die Auswirkungen der IT und Software insbesondere der digitalen Kommunikation und digitaler Medien auf Gesellschaft, Wirtschaft, Politik und weitere Lebensbereiche. Digitalization umfasst insbesondere die sich aus den digitalen Kommunikationsmitteln neu ergebenden Gestaltungsmöglichkeiten des Lebens. In diesem Buch wird der deutsche Begriff Digitalisierung im Sinne von Digitalization verwendet.

2 Big Data

Bereits 1998 ist der Begriff Big Data in der Literatur zu finden (Weiss und Indurkhya (1998)). Big Data ist ein Ansatz, mit dessen Hilfe Daten mit

- hohem Volumen (Volume),
- hoher Geschwindigkeit (Velocity),
- hoher Mannigfaltigkeit (Variety) und
- hoher Unsicherheit (Veracity)

gesammelt, gespeichert, verarbeitet, kommuniziert, ausgewertet, bereitgestellt und zielgerichtet genutzt werden können (Beyer und Laney (2012)). Im Englischen spricht man von den 4 V's (Volume, Velocity, Variety, Veracity). Einige Autoren ergänzen diese durch weitere Merkmale (Fan und Bifet (2012) oder Manyika et al. (2011)):

- Visualisierung (Visualisation)
- Bedeutungswandel (Variability)
- Wert (Value)

Ursprünglich hat man nur von den 3 V's (Volume, Velocity und Variety) gesprochen (Douglas (2001) oder Russom (2011)). Ein weiterer Ansatz Big Data zu definieren ist das sogenannte **HACE** Theorem (Wu et al. (2014)). Nach dem HACE-Theorem liegt Big Data vor, wenn es sich um

- **H**uge heterogeneous data (große Datenmengen in unterschiedlichen Strukturen, siehe Volume und Variety)
- **A**utonomous sources (verteilte Datenquellen, siehe Variety)
- **C**omplex (Beschreibung komplexer Sachverhalte, siehe Veracity)
- **E**volving (sich ändernde Aspekte, siehe Variability)

handelt.

Die zentralen Ziele von Big Data sind, die Realität auf Basis umfangreicher Daten besser zu verstehen, datengestützte Aussagen zu formulieren und damit fundierte Entscheidungen zu treffen. Wesentliche Themen von Big Data sind die Beschaffung, Bereitstellung, Bereinigung, Vervollständigung, Zusammenführung, Analyse, Interpretation, Visualisierung und Nutzung von umfangreichen Daten aus unterschiedlichen Quellen.

Die unterschiedlichen Versuche, den Begriff Big Data zu definieren, können in folgender Struktur zusammengefasst werden:

- Technische Dimensionen
  - Volume
  - Velocity
  - Variety
- Qualitative Dimensionen
  - Veracity
  - Variability
- Zieldimension
  - Visualisation
  - Value

Die technischen Dimensionen beschreiben den Umfang, die Bereitstellungs- und Verarbeitungsgeschwindigkeit und die Mannigfaltigkeit der Daten. Technische Entwicklungen bei Hardwarekomponenten (Speicherplatz, Rechenleistung, Übertragungsgeschwindigkeit, u. a.) und verbesserte Methoden im Bereich Advanced Analytics ermöglichen es, den höheren Anforderungen in den technischen Dimensionen gerecht zu werden.

Die qualitativen Dimensionen adressieren die Richtigkeit, Vertrauenswürdigkeit sowie die Gültigkeit der Daten. Nur durch Beziehungswissen (Semantik) sowie Wissen über das Anwendungsfeld (auch kontextbezogenes Wissen oder Domainwissen genannt) können die Anforderungen im Bereich Veracity sowie Variability erfüllt werden. In Kapitel 2.3.1 zur Visualisation wird gezeigt, dass die Transformation von Daten zu Informationen nur gelingen kann, wenn die qualitative Dimension von Big Data beherrscht wird.

Die Zieldimension umfasst die Themen Darstellung der Ergebnisse und Schaffung von Werten durch zielgerichtete Datennutzung. Beide Zieldimensionen unterstützen die Transformation der Informationen zu sinnvollem Wissen, das zu konkreten Entscheidungen oder Handlungen führt.

Einige Autoren (Dinter et al. (2015)) sehen Analytics als zusätzliche Dimension von Big Data. In diesem Buch wird Analytics als zentraler Teil von Business Analytics gesehen und im Kapitel 3.1 eingehend behandelt.

Nach Wrobel et al. (2014) bezeichnet Big Data den Trend zur Verfügbarkeit immer detaillierterer, komplexerer und zeitnäherer Daten, den Wechsel von einer modellgetriebenen zu einer daten- und modellgetriebenen Herangehensweise und die wirtschaftlichen, gesellschaftlichen und persönlichen Potenziale, die sich aus der Nutzung großer Datenbestände ergeben.

Es gibt zahlreiche Werkzeuge wie Hadoop, Rapid Miner, R oder Python, die das Bearbeiten von Daten im Sinne von Big Data in den sieben Dimensionen und Advandced Analytics unterstützen. Der Trend geht hier ganz klar Richtung Open source und Open data.

Einige Autoren sehen Big Data und den Versuch, immer komplexere Sachverhalte datenmäßig zu beschreiben bzw. zu verarbeiten durchaus kritisch und warnen vor