Fraunhofer FOKUS

FRAUNHOFER-INSTITUT FÜR OFFENE KOMMUNIKATIONSSYSTEME FOKUS

BIG DATA – UNGEHOBENE SCHÄTZE ODER DIGITALER ALBTRAUM

Klaus-Peter Eckert, Lutz Henckel, Petra Hoepner

Kompetenzzentrum Öffentliche IT

IMPRESSUM

Autoren:
Klaus-Peter Eckert, Lutz Henckel, Petra Hoepner

Gestaltung:
Reiko Kammer

Herausgeber:
Kompetenzzentrum Öffentliche IT
Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS
Kaiserin-Augusta-Allee 31, 10589 Berlin
Telefon: +49-30-3463-7173
Telefax: +49-30-3463-99-7173
info@oeffentliche-it.de
www.oeffentliche-it.de
www.fokus.fraunhofer.de

1. Auflage März 2014

VORWORT

Open Data, Linked Data, Big Data, Smart Data, Datability, Data Analytics oder auch eine Kombination wie Big Linked Open Data Analytics – die schnellen Begriffskonjunkturen zeigen auf, wie wichtig heute Daten geworden sind. Dabei waren Daten doch schon zu Beginn der informationstechnischen Revolution integraler Bestandteil der elektronischen Datenverarbeitung. Was ist passiert, dass sich das Thema Daten einer so großen Aufmerksamkeit erfreut?

Die zunächst geringe Leistungsfähigkeit von Computern war der Grund dafür, dass Daten über Datentypen des Computers repräsentiert wurden. Mit der weiteren Entwicklung der Hardware und der Integration immer neuer Zwischenebenen begann die Emanzipation: digitale Daten lösten sich immer mehr von ihren Trägersystemen und wurden zu selbständigen Objekten. Die modernen Systeme und Virtualisierungstechniken von heute erlauben eine immer weitergehende Abstrahierung von der eigentlichen Infrastruktur und ermöglichen die Verarbeitung sowohl in lokalen Systemen als auch in der Cloud.

Digitale Daten sind in erster Linie Zeichen und technische Kodierungen, die verarbeitet, gespeichert und übertragen werden können. Ein gemeinsames Verständnis davon, was diese Daten repräsentieren, sorgt dafür, dass sie als Informationen interpretiert werden können. Im nächsten Schritt kann dann aus den Informationen Wissen extrahiert werden. So steht die Auswertung von vorhandenen und neu generierten Daten immer mehr im Vordergrund, eröffnet neue Geschäftsmodelle, genauere Analysen und stimmigere Voraussagen. Neue technische Lösungen zur verteilten Verarbeitung von nunmehr immer größeren Datenmengen weckten neue Hoffnungen. Seit 2011 versucht die Industrie dieses unter dem Begriff Big Data als Sammelbegriff für die neuen Möglichkeiten zu beschreiben. Grundlegend baut die Idee von Big Data dabei darauf auf, dass große und polystrukturierte Datenmengen an mehreren Orten verteilt abgelegt und auch verteilt ausgewertet werden können.

Den industriellen Heilsversprechen stehen aber auch Risiken gegenüber. Dabei muss es gar nicht mal um missbräuchliche oder gar kriminelle Nutzung der Daten gehen: Auch politische, ideologische und interessensgeleitete Schlussfolgerungen und Empfehlungen kann es auch auf der Basis großer Datenmengen geben. Die Auswertung von Daten bleibt abhängig von Theorien, Modellen und intelligenter Interpretation der Ergebnisse. Aspekte wie Datenintegrität und die Qualität der Daten sind dafür eine entscheidende Voraussetzung; können aus schlechten Daten doch keine qualitativ hochwertigen Ergebnisse errechnet werden. Es gilt: »Müll rein – Müll raus«. In jedem Fall muss bei der Sammlung und anschließenden Auswertung der Schutz der Privatsphäre des Einzelnen gewährleistet werden. Dabei bedeutet informationelle Selbstbestimmung nicht nur ein individuelles Schutzrecht. Der Besitz enormer Datenmengen in der Hand weniger Unternehmen birgt Gefahren für die gesamte Gesellschaft.

Bevor die unbestreitbaren Möglichkeiten durch den Einsatz von Big-Data-Technologien auch in der öffentlichen Verwaltung ausgeschöpft werden können, bedarf es noch wichtiger Entwicklungen. So müssen mandantenfähige Angebote entwickelt werden, sodass kostengünstig ad-hoc Auswertungen vorgenommen werden können. Hierbei werden auch Fragen von Standardisierung und Interoperabilität berührt sowie gesetzliche Regelungen für den Umgang der Daten im öffentlichen und privaten Raum. Diese müssen verstärkt untersucht werden. Erst dann lassen sich aus dem Meer der Daten Hinweise auf Prozessoptimierungen und mögliche Entwicklungen, insbesondere aber politische Steuerungsinformationen fischen.

Ich wünsche Ihnen eine interessante Lektüre.

Jens Fromm

Leiter Kompetenzzentrum Öffentliche IT

UNTER ÖFFENTLICHER IT VERSTEHT MAN INFORMATIONSTECHNOLOGIEN, DIE IN EINEM ÖFFENTLICHEN RAUM DURCH DIE GESAMTGESELLSCHAFTLICHE RELEVANZ UNTER BESONDERER BERÜCKSICHTIGUNG DER STAATLICHEN VERANTWORTUNG STEHEN.

INHALTSVERZEICHNIS

1. BIG DATA – EINE DEFINITION

Der Begriff Big Data wurde in der heute verwendeten Bedeutung im Jahr 2011 durch das amerikanische Marktforschungsunternehmen Gartner Inc. eingeführt.[1] Big Data bezeichnet danach Methoden und Technologien für die hochskalierbare Erfassung, Aufbereitung, Speicherung und Analyse strukturierter und unstrukturierter Daten. Big Data ermöglicht den Aufbau einer Wertschöpfungskette, die mit den Daten beginnt und über Informationen bis hin zu Wissen reicht. Dies kann dabei helfen, die Planung, Steuerung und Optimierung von Prozessen in Wirtschaft, Verwaltung und Zivilgesellschaft zu verbessern.

Der Einsatz von Big-Data-Technologien ist jedoch auch mit einer Vielzahl von Risiken verbunden, die sich aus missbräuchlicher Nutzung, fehlenden rechtlichen Rahmenbedingungen und unzulässigen Schlussfolgerungen aus dem generierten Wissen ergeben.

Das vorliegende White Paper erläutert die technischen Grundlagen von Big Data, beleuchtet die Potenziale dieser Technologien und weist auf die vorhandenen Risiken hin. Zusammenfassend werden Anwendungsfelder und Handlungshinweise für den Einsatz von Big Data in der öffentlichen IT beschrieben.

»Big Data stellt eine Chance dar, existierende Daten zum Nutzen der Allgemeinheit auszuwerten.«

1.1 GRUNDLAGEN

Big Data beschäftigt sich mit der intelligenten Auswertung großer Mengen digitaler Daten. Mit intelligenter Auswertung ist hier gemeint, dass Daten aus unterschiedlichsten Quellen gesammelt, aufbereitet, zusammengeführt, analysiert und zu Informationen veredelt werden. Dies erhöht das Wissen der Auswerter signifikant und ermöglicht es, strategische Entscheidungen auf breiter Wissensgrundlage zu treffen. Aus Bereichen wie Wirtschaftsförderung, Energiewende, Sozialhilfe, Bildung, Verkehr oder öffentliche Sicherheit liegen Beispiele für den erfolgreichen Einsatz von Big Data vor.[2,3] Bei großen Mengen kann es sich um das in heutigen Datenzentren typischerweise vorhandene Datenvolumen von Tera- und Petabytes handeln. Es kann sich aber auch um das tausend- oder millionenfache Datenvolumen von Exa- und Zettabytes handeln, das die Gesamtmenge der heute überwiegend in sozialen Medien im Internet vorhandenen digitalen Daten beschreibt. Die Bandbreite der betrachteten Datenvolumen umfasst den Bereich von neun Zehnerpotenzen. Benötigt man beispielsweise zur Auswertung eines Terabyte Daten eine Sekunde, so braucht man für die Auswertung eines Zettabyte zweieinhalb Jahre. Damit wird klar, dass man zur Verarbeitung dieser Datenmengen auf das betrachtete Datenvolumen zugeschnittene Werkzeuge benötigt.

»Für alle Anwendungsbereiche sind passende Konzepte und Werkzeuge für den Einsatz von Big-Data-Analysen erforderlich.«

Daten, das sind vermehrt die von uns wissentlich und unwissentlich bei der Nutzung sozialer Netzwerke und beim Surfen im Internet hinterlassenen Spuren unserer Netzaktivitäten oder die bei der Nutzung mobiler Endgeräte erfassten Bewegungsdaten. Dazu gehören auch Daten eingebetteter Systeme und Sensoren aus Bereichen wie Medizin, Logistik, Verkehr und Energie sowie Produktions- und Maschinendaten. Ebenfalls gehören dazu die Ergebnisse klassischer kommerzieller, technischer und wissenschaftlicher Datenverarbeitung, die nicht nur durch hohes Datenvolumen, sondern auch durch hohe Datenraten charakterisiert sein können. Es handelt sich dabei sowohl um allgemein und öffentlich verfügbare Massendaten oder Datenströme als auch um nur in internen Netzen zugreifbare private Daten von Bürgern, Unternehmen und Behörden.

[1] Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, http://www.gartner.com/newsroom/id/1731916
[2] BITKOM (2012), Leitfaden Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte, http://www.bitkom.org/de/publikationen/38337_73446.aspx
[3] TechAmerica Foundation, http://www.techamericafoundation.org/category/big-data-commission

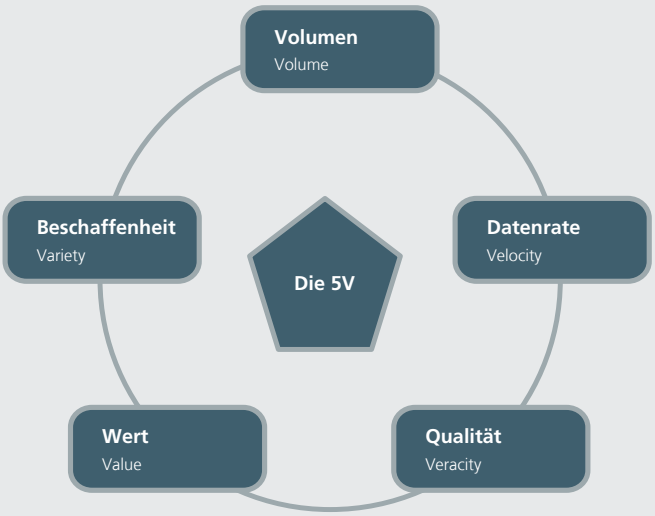

Abbildung 1:
V-Aspekte von Big Data

Neben Datenvolumen und Datenraten werden Big Data noch durch Attribute wie Beschaffenheit, Qualität und geschäftlicher Wert der betrachteten Daten charakterisiert. Im Englischen spricht man hier von volume, velocity, variety, veracity und value, den in Abbildung 1 dargestellten V-Aspekten von Big Data. In den nachfolgenden Kapiteln werden diese Attribute näher erläutert.

1.2 ZIELE UND KONZEPTE

Big-Data-Technologien unterstützen den Umgang mit großen Datenmengen. Dazu gehören die Speicherung der Daten, das Filtern und logische Zusammenführen von Daten aus verschiedenen Quellen (Datenmodellierung), die Bereitstellung von Verfahren zur Analyse der Daten (Informationsgewinnung), die Möglichkeit zur Formulierung dedizierter Anfragen und letztendlich die verständliche Darstellung der Antworten auf diese Anfragen (Wissensextraktion). Wo befindet sich hier aber das Neue gegenüber klassischen Verfahren und Prozessen zur Speicherung und Analyse digitaler Daten, wie klassischen Datenbanken, informationsintegrierenden Datenbanken (Data Warehouses) und Verfahren zur strategischen Auswertung von Unternehmensdaten (Business Intelligence)?

Aus dem skizzierten Spektrum der Datenvolumen über mehrere Zehnerpotenzen lässt sich einerseits ableiten, dass der Übergang von klassischen Analyseverfahren zu Big-Data-Konzepten ein fließender ist und damit eine scharfe Trennung der Konzepte unmöglich ist. Andererseits ergeben sich aus den anderen vier für Big Data charakteristischen Attributen weitere Unterscheidungskriterien. So können beispielsweise Sensoren einen kontinuierlichen Datenfluss liefern, der zeitnah zur Optimierung von Prozessen ausgewertet werden soll. Für solche Datenströme sind Übertragungsrate, Datenmenge und Reaktionszeit (Latenz) bis zur Beendigung der Auswertung wichtige Kenngrößen. Die zugehörigen Daten müssen unter Verwendung geeigneter Technologien zumindest bis zum Abschluss des Analysevorgangs zwischengespeichert werden. Im Allgemeinen kann man davon ausgehen, dass Big Data zumindest durch ein hohes Volumen oder aber durch eine hohe Änderungsrate charakterisiert ist.

»Big-Data-Technologien setzen hohes Datenvolumen oder Datenraten voraus.«

Die Beschaffenheit der betrachteten Daten stellt eine weitere Herausforderung dar. Sogenannte unstrukturierte Daten wie Texte, Bilder oder Videos können syntaktisch und semantisch analysiert und zur Beantwortung von Fragen mit herangezogen werden.[4] In vielen Big-Data-Anwendungen werden unstrukturierte Daten zusammen mit strukturierten und teilstrukturierten Daten betrachtet. Man spricht dabei auch von polystrukturierten Daten. Unstrukturierte Daten müssen speziell annotiert und mit Metadaten versehen werden (Verschlagwortung), um sie für weitere Auswertungen einbeziehen zu können. Hier kann man sich Verfahren zur Bilderkennung oder zur Verarbeitung natürlicher Sprache vorstellen. Semistrukturierte Daten enthalten entweder bereits beschreibende Elemente, wie man sie von XML- oder JSON-Objekten kennt. Es kann sich aber auch um eine Kombination von strukturierten und unstrukturierten Daten handeln, wie sie beispielweise bei E-Mails anzutreffen sind. Absender, Empfänger und Betreffzeile sind strukturell vorgegeben. Inhalt und Anhänge stellen dagegen unstrukturierte Daten dar. Die Inhalte von Tabellen oder Datenbanken werden als

[4] Die syntaktische Analyse bezieht sich auf die Überprüfung der formalen Korrektheit; bei Texten beispielsweise auf die Identifikation grammatischer Bestandteile. Die semantische Analyse bezieht sich auf das inhaltliche Verständnis. Texte können nach Schlagworten, Bilder nach dargestellten Inhalten untersucht werden.

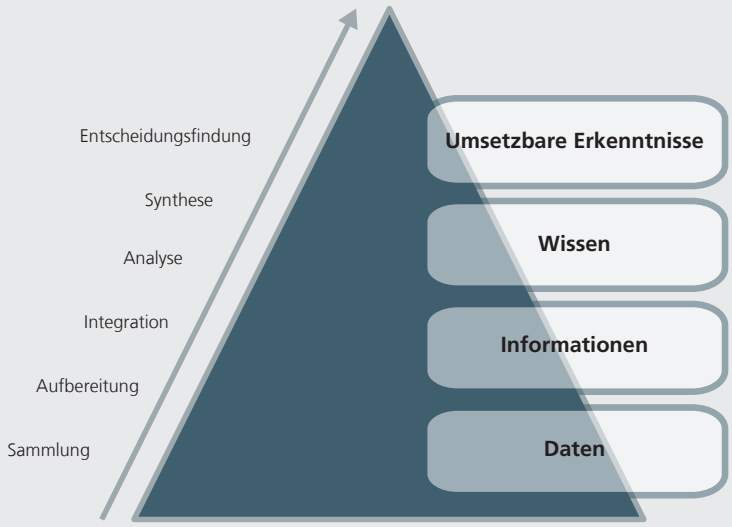

Abbildung 2: Prozessschritte bei der Anwendung von Analysetechniken

strukturierte Daten bezeichnet. Je weniger Struktur die betrachteten Daten aufweisen, desto mehr bietet sich der Einsatz von Big-Data-Analyseverfahren zu ihrer Auswertung an. Oder anders formuliert: In Ergänzung zu klassischen Auswertungsverfahren unterstützt Big Data auch die Analyse unstrukturierter Daten.

Nutzbare Daten können und müssen durch Vorverarbeitung und Filterung von »Datenmüll« getrennt und auf ihre Qualität zur Beantwortung der gestellten Fragen analysiert werden. Big-Data-Technologien bewegen sich in diesem von Volumen, Geschwindigkeit, Beschaffenheit, Qualität und Wert der Daten aufgespannten Raum. Je stärker die Vielfalt an Strukturen und Interpretationsmöglichkeiten der Daten ist, desto stärker kommen die spezifischen Eigenschaften dieser Technologien zum Tragen und grenzen Big Data von klassischen Ansätzen wie »Business Intelligence« ab.

»In Ergänzung zu klassischen Auswertungsverfahren erlaubt Big Data auch die Analyse unstrukturierter Daten.«

Das Hauptziel beim Einsatz von Big-Data-Technologien besteht in der Extraktion von Wissen aus einer großen Menge von Daten. Diese Aufgabe muss in angemessener Zeit durchführbar sein. Hier spielen neben den spezifischen Techniken, die häufig von einer Verteilung und parallelen Bearbeitung von Teilaufgaben ausgehen, auch moderne IKT-Infrastrukturen wie schnelle Kommunikationsnetze, dedizierte Speichertechnologien und durch Cloud-Computing bereitgestellte Speicherkapazitäten und Rechenleistungen eine wesentliche Rolle. Während »Business Intelligence« zumeist noch unter Nutzung vorhandener IKT-Infrastrukturen genutzt werden kann, setzt Big Data eine spezielle IKT-Unterstützung voraus.

»Das volle Potenzial von Big Data zeigt sich erst in Kombination mit leistungsfähigen IKT-Infrastrukturen.«

1.3 ANALYSE- UND VORHERSAGETECHNIKEN

Die bei Big Data eingesetzten Vorhersagetechniken beruhen weitestgehend auf statistischen Methoden. Für ihre Anwendung ist es erforderlich, die oft aus unterschiedlichen Quellen stammenden Daten in einem ersten Schritt in ein gemeinsames Datenmodell zu transformieren. Sollen beispielsweise aus Wetterdaten verschiedener Wetterstationen die Durchschnittstemperaturen verschiedener Länder zu verschiedenen Jahreszeiten verglichen werden, müssen vorbereitend die Rohdaten den Ländern und Jahreszeiten zugeordnet und dabei die physikalischen Maßeinheiten vereinheitlicht werden; eine Aufgabe, für die fundiertes Domänenwissen erforderlich ist. Anschließend kann man auf diesen Daten mit statistischen Verfahren die Durchschnittstemperaturen ermitteln.

Das Ergebnis dieser Analyse ist zunächst ein Bericht, der in der Vergangenheit liegende Ereignisse zusammenfasst und unter ausgewählten Gesichtspunkten betrachtet. Man hätte sich alternativ auch für nächtliche Tiefsttemperaturen, Niederschlagsmengen oder Sonnenstunden interessieren können. In einer erweiterten Analyse[5] kann nun versucht werden, Zusammenhänge zwischen den gemessenen Werten zu ermitteln. Hängt beispielsweise die nächtliche Tiefsttemperatur vom Grad der

[5] E.S. Levine: Applying Analytics, A Practical Introduction, CRC-Press, 2013, ISBN 978-1466557185.

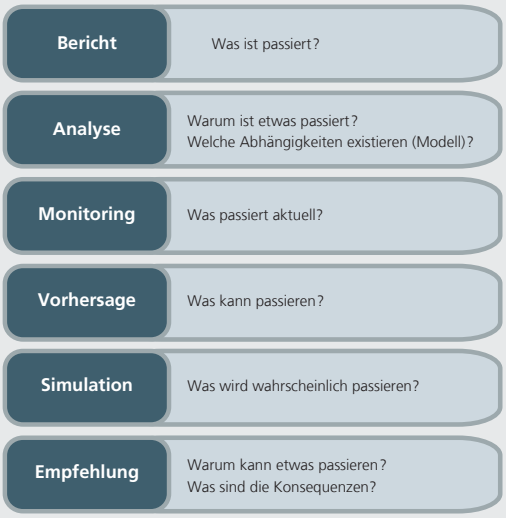

Abbildung 3: Typische Phasen von Datenauswertungen

Bewölkung ab? Das Ergebnis dieser Analyse ist ein Domänenmodell, im Beispiel ein Wettermodell, das Abhängigkeiten zwischen den einzelnen Dimensionen des Datenmodells beschreibt. Mittels dieses Modells können im nächsten Schritt *Vorhersagen* (Prediction) gemacht werden. Vorhersage meint in diesem Zusammenhang nicht notwendig eine Aussage über die Zukunft, sondern die Wahrscheinlichkeit, mit der zwei Ausprägungen gemeinsam vorkommen. In der Statistik stehen verschiedene Regressionsmodelle für solche Vorhersagen zur Verfügung. Wenn zu einer bestimmten Jahreszeit bei klarem Himmel eine Temperatur von x °C gemessen wird, wie hoch ist dann die zu erwartende nächtliche Tiefsttemperatur? Ähnlich wie bei Vorhersagen kann mittels Simulation das erwartete Verhalten des modellierten Systems berechnet werden bzw. die Korrektheit des Modells überprüft werden. Solche Simulationen erlauben die Abbildung komplexer Zusammenhänge, die sich einfacheren Modellierungen entziehen. Im letzten Schritt können aus dem Modell Empfehlungen (Prescriptions) abgeleitet werden, die sich aus Fortschreibungen und Variationen von berechneten Vorhersagen ergeben. Um beim Beispiel Wetter zu bleiben, könnten bei steigernder Wahrscheinlichkeit von Nachtfrösten bestimmte Vorsorgemaßnahmen empfohlen werden. Heutige Klimamodelle zeigen die Notwendigkeit, aber auch die Grenzen dieser Empfehlungen deutlich auf.

»Datenauswertungen bestehen aus folgenden Schritten: Bericht, Analyse, Monitoring, Vorhersage oder Simulation sowie Empfehlung.«

Einfache statistische Methoden stellen den Grundstein für Berichte und Darstellungen in Form von Diagrammen, Cockpits und ähnlichen, aus der Tabellenkalkulation bekannten, Grafiken dar. Darüber hinaus existieren weitere, auch im Umfeld von Big Data einsetzbare, mathematische Methoden. Assoziationsanalysen und Korrelationsanalysen sind Verfahren zum Erkennen und Bestätigen von Abhängigkeiten. Die Merkmalextraktion und die Identifizierung relevanter Einflussgrößen erlauben die Ermittlung wichtiger Ursache-Wirkung-Relationen. Anomalieerkennung unterstützt die Identifizierung von ungewöhnlichen Datensätzen, wie sie beispielsweise bei Störungen in Prozessen und Abläufen auftreten und bei der Erkennung zugehöriger Angriffe hilfreich sind. Regressionsverfahren ermöglichen vorausschauende Vorhersagen (Extrapolations) kontinuierlicher Größen. Clusteranalysen und Klassifikation erlauben die Gruppierung von Objekten aufgrund von Ähnlichkeiten und Prognosen über die Zugehörigkeit unklassifizierter Objekte zu derartigen Gruppen.

Aus dieser Aufzählung wird ersichtlich, dass die Mathematik eine Vielzahl etablierter Verfahren zur Verfügung stellt, die bei Big-Data-Analysen zum Einsatz kommen können. Eine weitere Erkenntnis ist aber auch, dass auf dem Weg vom Bericht zur Empfehlung Experten benötigt werden, die diese Werkzeuge sinnvoll einsetzen können.

»Die Mathematik bietet eine Vielzahl statistischer Verfahren für die Big-Data-Analyse. Expertenwissen ist jedoch erforderlich.«

Ein weiteres Anwendungsfeld für analytische Verfahren besteht in der Ableitung von Modellen aus vorliegenden Daten. So kann man sich beispielsweise fragen, ob bestimmte Wetterparameter dazu führen, dass gewisse Phänomene zu beobachten sind. Man messe beispielsweise die Verteilungen von Temperatur, Luftdruck, Windstärke usw. im Atlantik und untersuche, ob es gewisse Konstellationen gibt, die das Auftreten von Wirbelstürmen in Florida oder den Grad des Abschmelzens der Polkappen beeinflussen. Ein solches Vorgehen, bei dem unklare

DER SINNVOLLE UND KOSTENEFFIZIENTE EINSATZ IST FÜR KLEINERE UNTERNEHMEN UND KOMMUNALE BEHÖRDEN IM EINZELFALL ZU PRÜFEN.

Wirkungszusammenhänge untersucht werden, wird als explorative Datenanalyse bezeichnet. Die umfassenden, neuen Möglichkeiten der explorativen Datenanalyse stellen nach Ansicht einiger Experten einen der wesentlichen Fortschritte beim Einsatz von Big-Data-Technologien dar.

»Explorative Datenanalyse dient der Herleitung von Modellen aus vorhandenen Datensätzen. Die Bewertung ist jedoch ohne Domänenwissen nicht möglich.«

1.4 WERKZEUGE UND TECHNIKEN

Big Data wird häufig mit Begriffen wie »hadoop«, »mapreduce«, »hbase«, »pig«, »in-memory-computing«, oder »NoSQL«-Datenbanken verbunden. Eine ausführliche Diskussion dieser Techniken würde den Rahmen dieses Dokuments deutlich sprengen und ist zudem für den technisch Interessierten in der einschlägigen Fachliteratur nachlesbar.[6] Interessant für das technische Verständnis von Big Data ist jedoch die zeitliche und thematische Einordnung dieser Werkzeuge und Techniken.

Seitdem es digitalisierte Texte gibt, besteht die Anforderung, in diesen Texten nach Begriffen und Themen suchen zu können. Bibliothekare und Anbieter von Fachliteratur müssen dazu die Texte verschlagworten. Die so annotierten Texte werden über Indexe effizient durchsuchbar gemacht. Für diese Aufgabe wurden als Open Source Lösungen durch die »Apache Software Foundation« Frameworks wie »Lucene«[7] (1997) oder »Solr«[8] bereitgestellt. Die Verschlagwortung von Texten im Web ist insbesondere bei der Einbeziehung sozialer Netzwerke ein wesentlicher Schritt für die initiale Datenveredelung für Big Data. Die Entwickler von Lucene gingen nun einen Schritt weiter und beschäftigten sich mit der Analyse dieser annotierten Daten bzw. allgemein mit der Analyse von Daten mittels Techniken des verteilten und parallelen Rechnens. Das Ergebnis dieser Arbeiten ist die Apache-Open-Source-Implementierung »hadoop«[9] (2008), eine Implementierung des von Google entwickelten »Map-Reduce«-Verfahrens[10] zur Suche in großen Datenmengen. Ergänzt wird das Suchverfahren durch das Apache-Open-Source-»Data-Warehouse«-Projekt »hive«[11] zur Datenanalyse, das von Facebook entwickelt und von Amazon erweitert wurde.

Zusammenfassend stellen sich die genannten Techniken für Big Data demzufolge als Apache-Open-Source-Projekte dar, deren Entwicklung von Unternehmen wie Google, Amazon und Facebook getrieben wird. Damit ist jedoch ein Nutzerkreis von Big-Data-Technologien identifiziert, der aufgrund seiner technischen Möglichkeiten und Anforderungen klassische Unternehmen und Verwaltungen um Größenordnungen über-

»Zur Analyse hochvolumiger Datenströme werden spezielle Speichertechniken und Datenbanken benötigt.«

[6] Es existieren zahlreiche Publikationen über Big-Data-Techniken. Alle namhaften Anbieter dieser Techniken haben einschlägige White Paper veröffentlicht. Eine Zusammenfassung findet man beispielsweise bei O'Reilly (Pete Warden: Big Data Glossary – A Guide to the New Generation of Data Tools) als eBook unter http://shop.oreilly.com/product/0636920022466.do oder bei Apress (S.Mohanty et al.: Big Data Imperatives, 2013, ISBN 978-1430248729) als eBook unter http://www.apress.com/9781430248729

[7] Apache Lucene, http://lucene.apache.org/

[8] Apache Solr, http://lucene.apache.org/solr/

[9] Apache Hadoop, http://hadoop.apache.org/

[10] Map-Reduce-Verfahren basieren auf der Verteilung und Analyse der Daten auf unterschiedlichen Rechnern. Die Rechner ermitteln parallel Teilergebnisse und führen diese anschließend zu einem Gesamtergebnis zusammen.

[11] Apache Hive, http://hive.apache.org/

INFRASTRUKTUREN FÜR BIG DATA SETZEN ENTWEDER DIE NUTZUNG EIGENER IKT-RESSOURCEN ODER VON VERTRAUENSWÜRDIGEN ANBIETERN BEREITGESTELLTE CLOUD-DIENSTE VORAUS.

ragt. Inwieweit die entwickelten Techniken trotzdem nutzbringend einsetzbar sind, muss im Einzelfall entschieden werden.

Die bislang betrachteten Verfahren eignen sich insbesondere zur Analyse großer Datenmengen (volume), wobei historisch gesehen ein Schwerpunkt auf annotierte, textuelle Dokumente gelegt wird. Diese Situation ist typischerweise in sozialen Netzwerken gegeben, wobei global betrachtet durch die unterschiedlichen menschlichen Sprachen eine deutliche Komplexitätssteigerung zu erkennen ist.

Das zweite Merkmal von Big Data ist die Betrachtung von Datenströmen (velocity), wobei ebenfalls große Datenmengen anfallen können. Die eigentliche Herausforderung besteht jedoch in der hohen Datenrate, der Flüchtigkeit der betrachteten Daten und den quasi Realzeitanforderungen an deren Auswertung. Die bereits erwähnten, parallel arbeitenden »Map-Reduce«-Verfahren stellen hier keine vollständig befriedigende Lösung dar. Stattdessen bieten sich die Zusammenführung und temporäre Speicherung der Daten an. Zu diesem Zwecke wurden »In-Memory«-Techniken[12] entwickelt. Die Daten werden dabei in einem extrem schnell zugreifbaren Speicher gehalten und dort bis zum Abschluss ihrer Auswertung gespeichert. Große Hauptspeicher und viele Rechenkerne sind eine Hardwarevoraussetzung für den Einsatz derartiger Technologien.

Weitere Eigenschaften von Big Data sind die Unstrukturiertheit, Unvollständigkeit und Widersprüchlichkeit der analysierten Daten. Diese Eigenschaften machen den Einsatz klassischer Datenbanken oft unmöglich, da diese zumeist hohe Anforderungen an Konsistenz und transaktionales Verhalten stellen. Stattdessen reicht bei Big-Data-Techniken die Unterstützung einfacher Strukturen ohne strenge Konsistenzanforderungen aus. Das schnelle Lesen und Schreiben von Daten steht im Vordergrund, wie es bei häufigen Datenänderungen und auswertenden Zugriffen erforderlich ist. Diesen Anforderungen genügen die NoSQL-Datenbanken[13], die zunehmend in Big-Data- und Realzeit-Anwendungen genutzt werden. Ein typisches Beispiel stellt die Apache-Open-Source-Lösung »HBase«[14] dar, die auf von Google entwickelten Konzepten aufbaut.

1.5 TECHNISCHE INFRASTRUKTUREN

Ob sich Big Data durch große Datenmengen oder hohe Datenraten definiert, die Aufbereitung und Analyse der Daten benötigen leistungsfähige IKT-Infrastrukturen. Berücksichtigt man die aus dem Cloud-Computing bekannte Terminologie XaaS (Anything as a Service)[15], wobei der erbrachte Dienst dessen Nutzer befähigt, bestimmte Aufgaben durchzuführen, so lassen sich folgende Varianten unterscheiden:

In der ersten Variante wird die benötigte IKT-Infrastruktur, bestehend aus Speicher, Netz und Rechenleistung, als »Infrastructure as a Service« eingekauft, um damit ein eigenes Big-Data-System zu implementieren. Die für die Analyse benötigten Daten können dann entweder als Rohdaten oder als bereits aufbereitete und integrierte Daten direkt in das System einbezogen oder als »Data as a Service« von Dritten bezogen werden. In der zweiten Variante werden die Verfahren zur Datenaufbereitung bereits als »Information as a Service« eingekauft, um sie als Plattform für die eigenen Verfahren zur Analyse und Wissensextraktion einzusetzen. Dabei sind jedoch standardisierte Schnittstellen zu den angebotenen Verfahren erforderlich, um Abhängigkeiten zu einzelnen Anbietern zu vermeiden. Schließlich ist eine dritte Variante denkbar, bei der »Analytics as

[12] In-Memory-Verfahren halten die auszuwertenden Daten in einem extrem großen Arbeitsspeicher und vermeiden damit aufwendige Zugriffe auf externe Speicher.
[13] NoSQL-Datenbanken, http://de.wikipedia.org/wiki/NoSQL
[14] Apache HBase, http://hbase.apache.org/
[15] ITWissen, http://www.itwissen.info/definition/lexikon/XaaS-anything-as-a-service-Anything-as-a-Service.html

a Service« einschließlich der zugehörigen Visualisierungsverfahren eingekauft wird, sodass der lokale Einsatz von Big Data ausschließlich aus der Wissensgenerierung besteht.

Diesen theoretischen Möglichkeiten steht jedoch die Frage gegenüber, in welchem Umfang derartige Varianten angeboten werden. Die erste Variante, also die Nutzung externer IKT-Ressourcen, ist unabhängig von Big Data. Ihre Vor- und Nachteile sind in Studien[16] zu Cloud-Computing ausführlich diskutiert worden. Die Aufbereitung von Daten ist domänenspezifisch, sodass es nie universell aufbereitete Daten geben wird. Es ist jedoch technisch vorstellbar und wirtschaftlich sinnvoll, dass es für spezielle Domänen wie Bereiche des öffentlichen Sektors oder die Stadt von morgen dediziert veredelte Daten und zugehörige Analyseverfahren gibt, die zentral und mandantenfähig als Plattform oder noch spezialisierter als Software-Dienst betrieben und angeboten werden. Will man unternehmensspezifische, vertrauliche Daten in die Auswertung einbeziehen, so wird man entweder einen Großteil der benötigten IKT-Infrastruktur selbst oder durch einen vertrauenswürdigen Anbieter betreiben müssen. Alternativ kann man die Daten über einen »Storage as a Service«-Dienst vom Anbieter der Big-Data-Lösungen speichern lassen, wobei ebenfalls eine hohe Vertrauenswürdigkeit des Anbieters dieser Dienste gegeben sein muss.

[16] P. Deussen et al.: Cloud Concepts for the Public Sector in Germany, Fraunhofer FOKUS, 2012, ISBN 978-3-00-038674-9.

2. CHANCEN UND HERAUSFORDERUNGEN

2.1 POTENZIALE

Sind wir heute in der Lage, die Potenziale der großen Datenmengen überhaupt zu erkennen? Sind wir in der Lage, die Fragen zu stellen, die die Analyse vorliegender Daten sinnvoll beantworten könnte? Big-Data-Analysen liefern mathematisch untermauerte Aussagen über Gemeinsamkeiten, Anomalien, Zusammenhänge, Ursache-Wirkung-Beziehungen, Statistiken, Prognosen und Optimierungspotenziale und -strategien. Wir müssen lernen und lehren, diese Potenziale zu erkennen und ihre Auswirkungen für Entscheidungsträger aus Politik und Wirtschaft, aber auch für die Zivilgesellschaft richtig einzuschätzen. Klassische Vorhersagetechniken beschäftigen sich mit Antworten auf die Fragestellungen: Was ist in der Vergangenheit passiert? Warum ist etwas passiert? Was passiert aktuell und was wird wahrscheinlich passieren? Neuere Analyseansätze gehen einen Schritt weiter. Sie geben Antworten auf die Fragen: Warum kann zukünftig etwas passieren und wie sollten wir darauf reagieren? Big Data könnte somit die Beantwortung weitreichender Fragestellungen auf Grundlage einer stetig wachsenden Menge von Daten unterstützen.

Zahlreiche Studien liefern Beispiele für die erfolgreiche Nutzung von Big-Data-Konzepten. Die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD)[17] und das Weltwirtschaftsforum[18] betrachten Anwendungsszenarien aus den Bereichen Bildung, Gesundheit, Landwirtschaft, Finanzwesen, Dienstleistungen, Logistik und öffentlicher Sektor. Für den öffentlichen Sektor werden neben Kosteneinsparungen eine verbesserte Transparenz, die Bereitstellung neuer, personalisierter Dienstleistungen, die zeitnahe Unterstützung von Entscheidungsprozessen und das frühzeitige Erkennen bzw. Vermeiden von Fehlern und Betrug als wesentliche Vorteile herausgestellt. Dabei spielt die kollektive Bereitstellung von Daten durch Bürger, Unternehmen und Behörden, die sogenannte Datenphilanthropie, eine wichtige Rolle. In der von Fraunhofer durchgeführten Innovationspotenzialanalyse[19] sowie in »Best Practice« -Berichten von BITKOM[20] und TechAmerica[21] werden ergänzend die Bereiche Forschung, Produktion, Wetter- und Klimamodelle sowie Cyber Security angesprochen. Fasst man diese Anwendungsbereiche zusammen, decken sie viele Fragestellungen für die Stadt von morgen ab. Die Verbesserung der Lebensumstände der Bürger und der Qualität strategischer Entscheidungen in Verwaltung, Politik und Unternehmen stehen beim Einsatz von Big-Data-Technologien im Vordergrund. Ziel ist eine datengestützte Optimierung vieler Prozesse im öffentlichen Sektor.

»Big Data im öffentlichen Sektor kann höhere Transparenz, personalisierte Dienstleistungen und zeitnahe Entscheidungsprozesse ermöglichen.«

Auf diesem Weg dürfen wir uns von den technischen Möglichkeiten jedoch nicht blenden lassen und müssen stets die Qualität der Daten und der aus ihnen gewonnenen Informationen im Auge behalten. Wie bei jeder Entwicklung werden auch bei Big Data die positiven Nutzungsszenarien und der potenzielle Missbrauch kontrovers diskutiert. Im wirtschaftlichen Umfeld steht insbesondere bei Werbung, Kundenpflege und Produktplanung die folgende Frage im Raum: »Welches Risiko gehe ich ein, wenn meine Konkurrenz Big Data einsetzt und ich nicht?« Auch im öffentlichen Bereich sind Konkurrenzsituationen denkbar. Was passiert, wenn meine Nachbarkommune unter Einsatz

[17] Organisation for Economic Co-operation and Development (OECD), http://www.oecd.org/
[18] World Economic Forum, http://www.weforum.org/
[19] Fraunhofer IAIS (2013), Big Data – Vorsprung durch Wissen, Innovationspotentialanalyse, http://www.bigdata.fraunhofer.de/de/big-data.html
[20] Siehe[2], Seite 5
[21] Siehe[3], Seite 5

BIG DATA MUSS UNTER EINBEZIEHUNG VON RECHTLICHEN UND ETHISCHEN ASPEKTEN DISKUTIERT WERDEN.

von Big Data schnellere, bessere, bedarfsgerechtere Wirtschaftsförderung, Verkehrsplanung und -steuerung, Bildungsmanagement oder einfach werbewirksamere PR-Aktionen durchführt? Sicherlich ist die Einführung von vorhersagenden Analyseverfahren auf großen Datenmengen nicht direkt möglich. Aber zumindest ein schrittweises Vorgehen, beginnend mit dem Einsatz von Vorhersagetechniken auf eigenen Datenbeständen bis hin zur Auswertung großer Datenmengen, sollte auch im öffentlichen Bereich selbstverständlich sein. So macht es durchaus Sinn, sowohl von Verwaltungen, als auch von Unternehmen bereitgestellte offene Daten zu eigenen Zwecken analytisch auszuwerten.

Aus technischer Sicht ist die Zeit reif für den Einsatz von Big Data – was hindert uns noch daran, die Versprechungen des fünften Kondratjewzyklus[22] einzulösen und Information zu einem bestimmenden Wirtschaftsgut unserer Zeit zu machen?

2.2 PRIVATSPHÄRE, RECHTLICHE UND ETHISCHE GRENZEN

Big Data lebt von der Extraktion und Integration von Informationen aus verschiedenen Quellen. In vielen Fällen werden öffentlich verfügbare Daten aus dem Internet mit privaten Daten gemeinsam ausgewertet, um Prognosen und Strategien zu entwickeln oder um eigene Vorhersagen und Modelle zu bestätigen, zu widerlegen oder hinsichtlich optimaler Lösungen miteinander zu vergleichen. Die Auswirkungen für jeden Einzelnen und die Gesellschaft allgemein können jedoch erheblich sein. Aus Online-Shops sind uns die Auswirkungen der Auswertungen von generierten Verhaltensmustern in Form persönlicher Werbung gut bekannt. Die zu eigenen Zwecken genutzten und im Extremfall ohne Kenntnis der Betroffenen veröffentlichten Persönlichkeitsprofile bergen Risiken und Gefahren, wie uns durch die Vorfälle der letzten Monate schmerzhaft bewusst geworden ist. Daher müssen wir sicherstellen, dass der Einsatz von Big Data nicht nur unter technischen Gesichtspunkten diskutiert wird, sondern unter Einbeziehung von rechtlichen und ethischen Aspekten.[23]

Dem Zusammenführen personenbezogener und personenbeziehbarer Daten aus verschiedenen Quellen sind jedoch auch enge rechtliche und ethische Grenzen gesetzt. Die Grundidee der Sammlung und Auswertung von personenbezogenen Daten aus verschiedensten Quellen widerspricht fundamental den Prinzipien der Zweckbindung bei der Erhebung und Speicherung von Daten und der Datensparsamkeit. Bundesdatenschutzgesetz[24], Telemediengesetz[25] aber auch die Europäische Datenschutzrichtlinie[26] definieren einen engen Rahmen für den Einsatz von Big Data im öffentlichen Sektor. Der Bürger muss als Individuum durch Anonymisierung und Pseudonymisierung seiner Daten davor geschützt werden, zum gläsernen Menschen zu werden. Auch seine Identität, Privatsphäre und Reputation müssen gegen Manipulation durch Dritte geschützt werden. Rechtliche und ethische Regeln müssen bereits vor der Einführung von Big Data etabliert werden und die Antworten auf folgende Fragen beinhalten: Welches Recht auf eine digitale Identität besitzen die Bürger im Zeitalter von Big Data? Wie können sie ihre digitale Identität und Reputation im Netz schützen? Wer darf welche Daten über Dritte ins Netz stellen, die dann allgemein für Big-Data-Analysen und speziell zur Definition des digitalen Ichs Verwendung finden?

Die Personenbeziehbarkeit von Daten kann sich mit dem Einsatz von Big-Data-Technologien erhöhen. Auch kann sich der Einsatz explorativer Verfahren zur Erkennung von Korrelationen

[22] Siehe z. B.: Leo A. Nefiodow: Der sechste Kondratieff; Rhein-Sieg Verlag; 2006; ISBN-10: 3980514455.
[23] K. Davis, D.Patterson: The Ethics of Big Data; O'Reilley, 2012, ISBN 978-1-44931-179-7, http://it-ebooks.info/book/1984/
[24] Bundesdatenschutzgesetz, http://www.gesetze-im-internet.de/bdsg_1990/
[25] Telemediengesetz, http://www.gesetze-im-internet.de/tmg/
[26] Europäische Datenschutzrichtlinie, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:DE:NOT

GRUNDLAGE FÜR DIE IDENTIFIKATION STANDARDISIERBARER ASPEKTE VON BIG DATA IST EIN UMFASSENDES, ALLGEMEIN AKZEPTIERTES REFERENZMODELL.

in einer Menge auf den ersten Blick unabhängiger Daten negativ auf den Einzelnen auswirken. Ermittelt eine Versicherungsgesellschaft beispielsweise Attribute (Alter, Geschlecht, Einkommen, Vorlieben, Wohnort usw.) von Personen, aus denen sich ein hohes Risiko bei der Kreditvergabe ableiten lässt, so werden alle Personen verdächtigt, die diese Kriterien erfüllen. Kommt dann noch eine Zahlengläubigkeit, d. h. ein übersteigertes Vertrauen in die ja mit wissenschaftlichen Methoden ermittelten Attribute hinzu, können die betroffenen Personen leicht in die Notwendigkeit geraten, ihre »Unschuld« nachweisen zu müssen. Der »Kontrolleur« geht von einer potenziellen »Schuld« aus und verhält sich entsprechend. Der Rückschluss von Statistiken auf Individuen erlaubt potenziell diskriminierende Eigenschaftszuschreibungen, die mit der Person nichts zu tun haben müssen. Einzelfallgerechtigkeit mittels Statistik sieht sich unlösbaren Komplexitätsproblemen gegenüber.

2.3 STANDARDISIERUNG

Big Data ist eine sehr junge Technologie, sodass noch keine dedizierten Standards existieren. Trotzdem gibt es aktuell Überlegungen aus Standardisierungsorganisationen wie DIN/ISO, NIST, TMF oder OASIS, einzelne Aspekte wie Abfragesprachen, Sicherheitsschnittstellen, Einbettung in Cloud-Infrastrukturen oder Metadatenkonzepte auf ihre Standardisierbarkeit hin zu untersuchen. Belastbare Ergebnisse sind jedoch noch nicht vorhanden.

Ein wesentlicher Bestandteil der Standardisierungsbestrebungen ist die Entwicklung eines Big-Data-Referenzmodells bzw. einer Referenzarchitektur. Dieses Modell beschreibt die funktionalen Komponenten eines Big-Data-Systems und deren Einbettung in geeignete IKT-Infrastrukturen. Zurzeit existiert eine Reihe von durch die genannten Standardisierungsorganisationen und die Industrie entwickelten Modellen, die bis zu einem gewissen Punkt bereits ähnliche Strukturen aufweisen.[27]

2.4 EXPERTEN

Das Fehlen von Experten mit Kenntnissen über die technischen und fachlichen Grundlagen von Big Data wird als ein großes Hemmnis bei der erfolgreichen Einführung gesehen. Allein in den USA fehlen nach einer aktuellen Studie von McKinsey (Mitte 2013) ca. 150.000 Spezialisten zur Entwicklung fachlicher Analysemodelle.[28] Dazu gehören die sogenannten Datenwissenschaftler mit Wissen über technische Zusammenhänge, mathematische Methoden und profunde Kenntnisse der jeweiligen Anwendungsdomäne. Weiterhin fehlen noch 1,5 Millionen geschulte Analysten und Entscheidungsträger, damit diese Modelle genutzt werden können.

Andererseits werden Datenanalyse und Datenaufbereitung, mathematische Methoden, betriebswirtschaftliche Kenntnisse, Datenbankwissen und informationstechnische Infrastrukturen in verschiedenen Studiengängen wie Sozialwissenschaften, Betriebswirtschaft, Informatik, Mathematik usw. längst gelehrt. Es stellt sich daher die Frage, ob eine Umstrukturierung beziehungsweise Ergänzung bestimmter Studienangebote zielführend wäre, um einen Mix an verschiedenen Kompetenzen zu erreichen.

»Studiengänge sollten um technische und fachliche Grundlagen von Big Data ergänzt werden.«

[27] BITKOM (2013), Leitfaden für Big Data Projekte, Kap. 8, http://www.bitkom.org/files/documents/LF_big_data2013_web.pdf

[28] McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

3. ANWENDUNGSFELDER IM ÖFFENTLICHEN RAUM

Bei allen unbestreitbaren Risiken dürfen die mit einem verantwortlichen Einsatz von Big Data verbundenen Chancen nicht übersehen werden. Um diese Potenziale zu kommunizieren, ist die Entwicklung und Demonstration von sinnvollen Anwendungen, d. h. erst durch Big Data umsetzbarer Szenarien, ein unverzichtbarer Schritt. Eine Reihe von Studien aus den USA[29], England[30] und Deutschland[31] führen Beispiele für erfolgreich abgeschlossene Big-Data-Projekte an, sodass an dieser Stelle ein Hinweis auf diese Studien ausreichend erscheint. Bei den dort beschriebenen Beispielen ist jedoch zu beachten, dass häufig Analyseverfahren auf »kleine« Datenmengen angewendet wurden. Weitere überzeugende Anwendungsbeispiele wichtiger Handlungsfelder aus dem öffentlichen Raum, wie z. B. Open Government, Smart Energy oder Mobilität sowie intelligente Verkehrs- und Transportsysteme, die Antworten auf neu stellbare Fragen geben, sind jedoch für eine erfolgreiche Etablierung von Big-Data-Konzepten erforderlich.

»Die Versprechungen von Big Data erfüllen sich nur dann, wenn Trends, Beschränkungen, Charakteristika, aber auch Risiken verstanden sind.«

3.1 GESCHÄFTSMODELLE

Die Einführung von Big Data muss sich rechnen. Den Kosten für die Bereitstellung einer Big-Data-Infrastruktur und den Kosten für die Schulung der Mitarbeiter müssen deutliche Kosteneinsparungen und Gewinne bei der Nutzung der Technologien gegenüberstehen. Im kommerziellen Umfeld von Werbung und Internet-Shops ergeben sich die Gewinne aus individuellen Angeboten für Kunden, verkaufsoptimierenden Zusammenstellungen von Angeboten, Optimierung des Produktspektrums, frühzeitiger Identifikation neuer Produktlinien und Ähnlichem.

In der öffentlichen IT ist neben potenziellen Kosteneinsparungen der gesamtgesellschaftliche Nutzen ein Hauptkriterium für Big Data. Die IT-gestützte Auswertung und strategische Analyse großer Datenmengen ist ein Schritt in Richtung auf eine Optimierung von Prozessen im öffentlichen Sektor. Im Rahmen der Transparenzgesetzgebung für die Bürger und die Wirtschaft verfügbar gemachte offene Daten sollten auch innerhalb der Verwaltung analysiert und ausgewertet werden können, wobei, soweit zulässig, eine Verknüpfung mit verwaltungsinternen Daten und verwaltungsexternen, frei verfügbaren Daten erfolgen kann.

»Die im Rahmen der Transparenzgesetzgebung öffentlich verfügbaren Daten sollten auch in der Verwaltung analysiert und ausgewertet werden.«

3.2 BIG DATA FÜR DIE ÖFFENTLICHE VERWALTUNG

Es gibt Beispiele, die zeigen, dass Big-Data-Arbeits- und Entscheidungsprozesse in der öffentlichen Verwaltung verbessern kann.[32] Insbesondere hervorzuheben ist aber der in vielen Big-Data-Studien wiederkehrende Hinweis, zuallererst einen klaren

[29] TechAmerica (2012), Demystifying Big Data, A Practical Guide To Transforming The Business of Government, http://breakinggov.com/documents/demystifying-big-data-a-practical-guide-to-transforming-the-bus/

[30] Chris Yiu (2012), The Big Data Oportunity, Making government faster, smarter and more personal, http://www.policyexchange.org.uk/publications/category/item/the-big-data-opportunity-making-government-faster-smarter-and-more-personal

[31] Siehe[19], Seite 12

[32] Vitako aktuell, Big Data, Enorme Ressourcen, 1-2013, http://www.vitako.de/Publikationen/Documents/Vitako%20aktuell%201-2013.pdf

DIE VERWENDUNG VON BIG-DATA-ANALYSEVERFAHREN ERFORDERT EINE ÖFFNUNG DER FACHVERFAHREN FÜR DATENEXTRAKTION UND -RECHERCHE.

Nutzen für die eigene Einrichtung oder ein konkretes Problem beziehungsweise eine Fragestellung zu identifizieren, bevor in Big-Data-Technologien investiert wird. Im sozialen Bereich, für die Verkehrssteuerung, bei weltweiten Patentrecherchen, bei der Analyse sozialer Netzwerke, um die Stimmung der Bevölkerung zu erkennen, bei der Ermittlung von Lebensmittelketten für die Lebensmittelkontrolle, in der Gesundheitsvorsorge und bei einer Gesetzesfolgenabschätzung ist ein gesamtgesellschaftlicher Nutzen erkennbar.

Gefundene Beispiele betreffen jedoch Behörden von Großstädten, landesweit und international agierende Unternehmen sowie Länder oder EU-weite Institutionen. Für kleinere Kommunen erscheint Big Data überdimensioniert. Hier reichen oftmals die Problembestimmung und herkömmliche Datenanalyse zur Entscheidungsfindung aus.

Hinderlich für Big Data ist auch die Vielzahl von Fachverfahren, die in heutigen Behörden eingesetzt werden, da diese häufig noch gar nicht für die erforderliche Datenrecherche, -extraktion und -analyse vorbereitet sind. Grundsätzlich ist dies jedoch ein Henne-Ei-Problem: Womit fängt man an, mit den Daten oder mit der Fragestellung ?

3.3 BIG DATA FÜR DIE GEZIELTE WIRTSCHAFTSFÖRDERUNG

Um im globalen Wettbewerb auch weiterhin eine starke Stellung einnehmen zu können, ist Deutschland auf Wissen als seine wichtigste Ressource angewiesen. Wissen ermöglicht eine höhere Wertschöpfung durch die Veredelung von importierten Rohstoffen zu Hochtechnologieprodukten. Der Wohlstand des Landes ist bereits heute, aber auch zukünftg, nur dann haltbar, wenn international konkurrenzfähige Hochtechnologieprodukte und entsprechende Dienstleistungen angeboten werden können.

Voraussetzung für die Erhaltung oder das Erreichen einer führenden Rolle auf unterschiedlichen Technologiemärkten ist das frühe Identifizieren und Setzen von Techniktrends. Eine Identifizierung kann mit Hilfe von Big-Data-Analyseverfahren erfolgen, indem zunächst Informationen aus wissenschaftlichen Veröffentlichungen und zu technischen Entwicklungen (z. B. Patentanmeldungen) zusammengeführt und analysiert werden. Sind die Zukunftstrends identifiziert, sollte es eine gezielte staatliche Förderung des gesamten Prozesses geben, der ausgehend von der Identifizierung eines Techniktrends bis hin zum fertigen Hochtechnologieprodukt reicht.

Dieser Prozess schließt auch die wissenschaftliche Erforschung der Technik und die Etablierung von Bildungs- bzw. Weiterbildungsmaßnahmen für Forscher und Ingenieure sowie für zukünftige Mitarbeiter in produzierenden Unternehmen mit ein. Neu gegründete Unternehmen, die in besonderem Maß auf eine Umsetzung von Zukunftstrends in Produkte setzen, sollten daher auch bei der Wirtschaftsförderung eine besondere Berücksichtigung finden. Darüber hinaus sind bereits vorhandene Standortvorteile zu identifizieren, bzw. der zielgerichtete Aufbau von neuen Standorten zu fördern. Die Eignung eines Standortes setzt sich dabei aus unterschiedlichen Faktoren zusammen. Dazu gehört die Existenz von Forschungs- und Bildungseinrichtungen, die sich mit der jeweiligen Technologie bereits auseinandersetzen. Weiterhin stellt sich die Frage, ob Unternehmen in dem jeweiligen Technikumfeld bereits vor Ort sind und ob die Infrastruktur (bzgl. Verkehr, Logistik, Zulieferung, Energie, Entsorgung, etc.) vorhanden ist. Schließlich ist das Know-How der potenziellen Arbeitskräfte für die Forschungs- und Entwicklungseinrichtungen sowie für die entwickelnden und produzierenden Unternehmen in der Region ein wichtiger Standortfaktor.

Die Daten über Forschungs- und Bildungseinrichtungen und deren inhaltliche Schwerpunkte liegen staatlichen Behörden ebenso vor wie die zu Unternehmen, Standorten einschließlich existierender Infrastruktur und zum Ausbildungsstand von

DATEN SIND IN DER STADT VON MORGEN EINE GRUNDLAGE FÜR GESTALTUNG UND LEBENSQUALITÄT.

Arbeitskräften. Damit nun eine gezielte Wirtschaftsförderung erfolgen kann, müssen diese Daten aus unterschiedlichen Quellen zusammengeführt und gegebenenfalls mit Big-Data-Analyseverfahren ausgewertet werden, um Prognosen für zukünftige Technologieentwicklungen erstellen zu können.

»Durch den Einsatz von Big-Data-Analyseverfahren kann der technische Vorsprung gesichert und ausgebaut werden.«

3.4 BIG DATA FÜR DIE STADT VON MORGEN

Datenanalysen lassen sich in der Stadt von morgen zur vorausschauenden Planung von Wohngebieten, Erholungsgebieten, Verkehrsverbindungen, Gewerbeansiedlungen, sozialen Einrichtungen usw. verwenden. Daten ermöglichen aber auch Bürgern die einfache Bildung von Interessengemeinschaften, z. B. zur Nachbarschaftshilfe, basierend auf gegenseitigen Kenntnissen über Interessengebiete und die Korrelation von Bedürfnissen mit Fähigkeiten.

Private und öffentliche Verkehrsmittel kommunizieren untereinander und mit ihrer Umgebung, um optimale Verkehrsflüsse zu ermöglichen. Dabei werden Daten übermittelt, die Aufschluss über typische Verhaltensmuster und deren zeitliche Entwicklung geben.

Digitalisierte Geo-Informationen über unter- und überirdisch vorhandene Artefakte erleichtern die Planung und Realisierung neuer Energietrassen. Sensoren messen Luftverschmutzung, Ozonwerte, Wasserstände, Windstärke und weitere Umweltgrößen. Angereichert um Informationen von Bürgern vor Ort lassen sich Menschen bei größeren und kleineren Gefahren gezielt informieren und warnen. Am Beispiel des Erdbebens in Haiti im Jahr 2010 lässt sich zeigen, wie neben Satellitenaufnahmen eine gezielte Auswertung von Bewegungsdaten mobiler Endgeräte und sonstiger Sensordaten zu einer effektiven und bedarfsgerechten Einsatzplanung von Hilfskräften und Gütern sowie der Versorgung der Bevölkerung führte, die ohne entsprechende IT-Unterstützung nicht denkbar gewesen wäre.

Braucht man für alle Szenarien zwangsläufig Big Data ? Die ehrliche Antwort ist »nein«. Werden zukünftig die vorhandenen Daten jedoch unter Einsatz von Big-Data-Analyseverfahren ausgewertet, so können die verschiedenen Nutzungsszenarien schneller, zuverlässiger und in vielen Situationen proaktiv und kostensparend umgesetzt werden.

4. HANDLUNGSHINWEISE

Im Gegensatz zu Unternehmen sollte der Sinn und Zweck von Big-Data-Analysen im öffentlichen Raum der gesamtgesellschaftliche Nutzen sein.

Big Data – Nichts Neues?

Datenspeicherung, Datenanalyse, statistische und mathematische Methoden, Auswertung und Ergebnisbeurteilung sind alles bereits bekannte Teilbereiche von Big Data. Neu sind lediglich die weitaus größeren Datenmengen und die damit einhergehenden Herausforderungen bezüglich Datenspeicherung, Datenübertragung, Datenanalysemethoden sowie Datenverarbeitungskapazitäten.

Privatsphäre schützen

Jüngste Ereignisse haben das Bewusstsein für und die Sorge über Datensammelwut geschürt. Die Verwendung ihrer Daten ist für viele Bürger nicht einschätzbar. Erforderlich sind daher Aufklärung, Methoden, Mechanismen und Maßnahmen, die die Privatsphäre von Bürgern schützen. Dies kann beispielsweise durch Verschlüsselung, Einholen der Zustimmung zur Datenverarbeitung, Nachvollziehbarkeit der Datenübermittlung, Nachweis der Löschung und durch Anonymisierungsverfahren bei komplexen Datenverbünden erfolgen. Gelebte Verhaltensregeln bei gleichzeitiger Einfachheit technischer Möglichkeiten sind zwingend notwendig.

Selbstverpflichtung, Kodex oder Regulierung?

Neben technischen Lösungen müssen auch rechtliche Rahmenbedingungen geschaffen werden, die Eigentumsrechte, Urheberrechte, Nutzungsrechte, die Dauer von Zugriffsberechtigungen auf Daten und vergleichbare, heute noch nicht geregelte Fragestellungen beantworten. Ein verantwortungsvoller Umgang ist bei der Erfassung, Aufbereitung und Analyse von Daten sowie bei dem daraus gewonnenen Wissen einzuhalten. Ob dies durch Selbstverpflichtung, einen Kodex oder auch Regulierung erreicht wird, ist im Dialog zwischen Wirtschaft, Forschung und Politik kontinuierlich zu ermitteln und an die technische Entwicklung anzupassen.

Die Kunst, die richtigen Fragen zu stellen

Es ist ein Trugschluss, große Datenmengen mit besseren Lösungen gleichzusetzen. Immer noch sind die wesentlichen Punkte, die richtigen Fragen zu stellen, die richtigen Daten auszuwählen und die Ergebnisse richtig auszuwerten.

Müll rein, Müll raus

Die Vertrauenswürdigkeit von Daten und die Nachweisbarkeit ihres Ursprungs wird bei vermehrter Zusammenführung von Daten eine immer wichtigere Rolle spielen. Insbesondere wenn viele Daten gesammelt werden, ist das Auftreten von Qualitätsunterschieden wahrscheinlich. Dies ist besonders bei Datenanalysen zu berücksichtigen.

Für kleine Kommunen ungeeignet

Nicht jede Prognose erfordert einen Big-Data-Ansatz. Für kleine Kommunen und Behörden sprengen die Kosten für Infrastruktur und Experten den verfügbaren Finanzrahmen. Herkömmliche Datenanalyse- und Auswertungsmethoden sind gut geeignet, um die hier typischen Datenvolumen zu bearbeiten.

Big-Data-Projekte

Der Nutzen von Big Data hängt in erster Linie von der richtigen Fragestellung, aber auch von der Auswahl der Daten und Werkzeuge sowie den geeigneten Experten ab, die sich mit der Anwendung dieser Technologien auskennen.

Die öffentliche Hand verfügt bereits über eine Vielzahl von qualitativ hochwertiger Daten (etwa die der statistischen Ämter), die bei der Datensuche zunächst berücksichtigt werden sollten.

GEFÖRDERT VOM

Bundesministerium
des Innern

KONTAKT

Jens Fromm
Leiter Kompetenzzentrum Öffentliche IT (ÖFIT)
Tel.: +49 30 3463-7173
Fax: +49 30 3463-99-7173
jens.fromm@fokus.fraunhofer.de

Fraunhofer-Institut für
Offene Kommunikationssysteme FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin

www.fokus.fraunhofer.de
www.oeffentliche-it.de

kre | 1403 (Fotos: Matthias Heyde / Fraunhofer FOKUS)