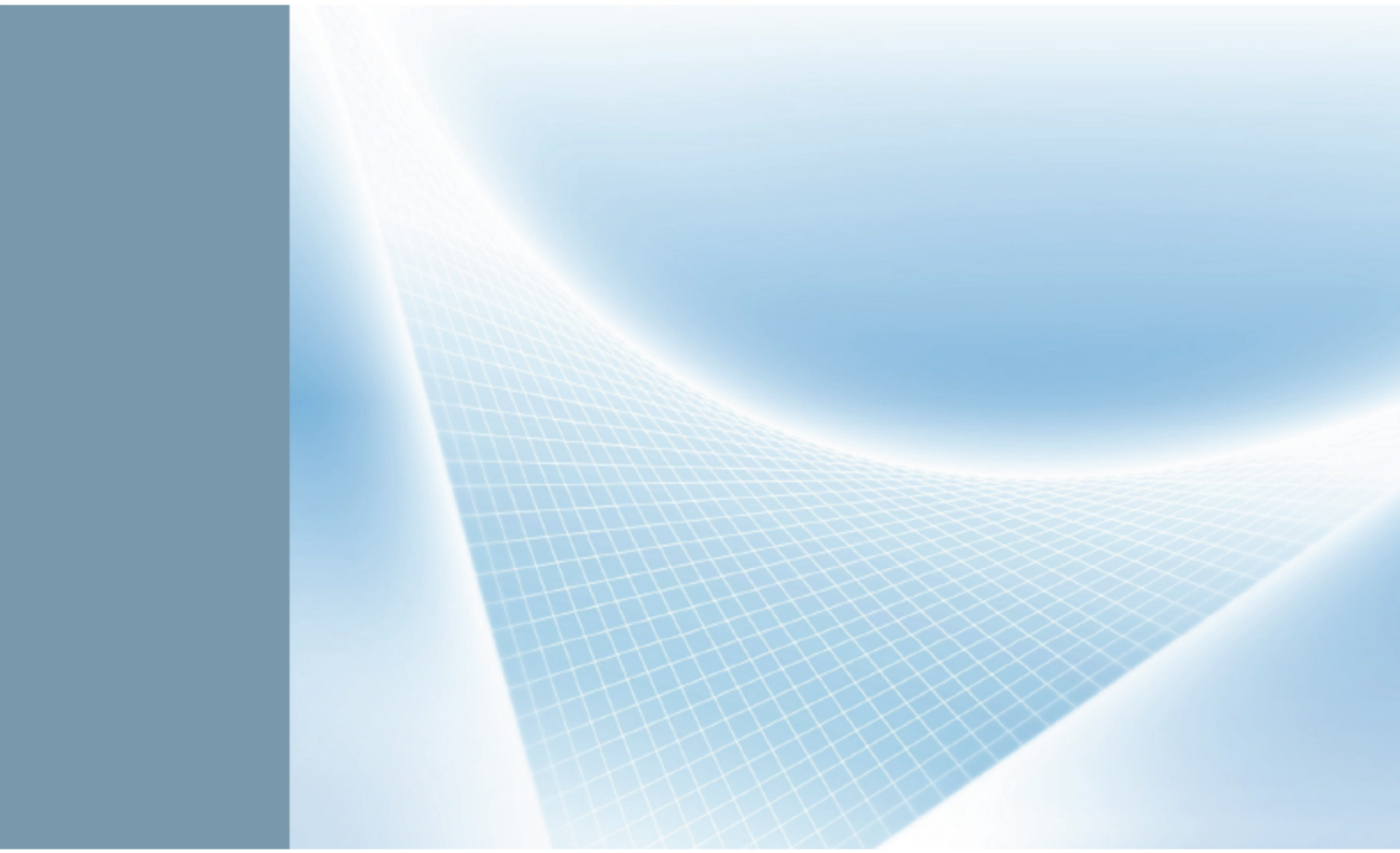
Leo Brecht, Birgit Stelzer



Big Data im Kontext von Industrie 4.0

Eine Technologievorausschau anhand IT-gestützter bibliometrischer Analyse und Szenariotechnik

Silke Burmeister, Johannes Lang, Niklas Bayrle, Murat Catalkaya, Birgit Stelzer, Edgar Schiebel

ITOP-Schriftenreihe Nr. 7
ISSN (Print) 2199-4919
ISSN (Online) 2511-1698

Projektmitarbeiter

Wissenschaftlicher Direktor
Prof. Dr. Leo Brecht, Universität Ulm

Wissenschaftliche Projektleiterin
Dr. Birgit Stelzer

Studentische Mitarbeiter
Silke Burmeister, B. Sc.
Johannes Lang, B. Sc.
Niklas Bayrle, B. Sc.
Murat Catalkaya, B. Sc.

Kooperationspartner:
Dr. Edgar Schiebel, Austrian Institute of Technology, Wien

Impressum

Herausgeber

Prof. Dr. Leo Brecht, Institutsdirektor und Dr. Birgit Stelzer, Wissenschaftliche Mitarbeiterin
Institut für Technologie- und Prozessmanagement

Universität Ulm
Helmholtzstraße 22
D-89081 Ulm

Fon +49 (0) 7 31 50-3 23 00
Fax +49 (0) 7 31 50-3 23 09

E-Mail: leo.brecht@uni-ulm.de
Internet: www.uni-ulm.de/mawi/itop.html

Druck

Copyright © by Prof. Dr. Leo Brecht, Universität Ulm.
Diese Publikation ist urheberrechtlich geschützt.

Ulm, September 2016. Printed in Germany.

Inhaltsverzeichnis

Impressum I
Inhaltsverzeichnis II
Abbildungsverzeichnis IV
Tabellenverzeichnis VI
Abkürzungsverzeichnis 1
1 Executive Summary 2
2 Industrie 4.0 – Anwendung von Big-Data-Technologien 6
2.1 Big Data - eine Begriffsbestimmung 8
2.2 Big Data im Zusammenhang mit Industrie 4.0 10
3 Forschungsmethodischer Rahmen 16
3.1 Bibliometrie 16
3.2 Szenariotechnik 17
4 Bibliometrische Analyse zu Big Data in der Industrie 4.0 19
4.1 Einbettung des Technologiefeldes in die Dimensionen der Wissenschaft 19
4.1.1 Hotspots 20
4.1.2 Die Dimensionen der Wissenschaft 22
4.2 Netzwerkanalysen 28
4.2.1 Netzwerk der Autoren 29
4.2.2 Netzwerk der Organisationen 36
4.2.3 Netzwerk der Schlagwörter 51
4.3 Wissensbasen und Forschungsfronten 54
4.3.1 Wissensbasen: Landkarte der Co-Zitationen 54
4.3.2 Forschungsfronten: Gebirgslandschaft der bibliografisch gekoppelten Publikationen 62
4.4 Forschungsportfolio: Positionierung der Forschungsfronten in der Wissenschaft 72
5 Szenariotechnik zu Big Data in der Industrie 4.0 75
5.1 Definition des Untersuchungsgegenstandes 75
5.2 Einflussbereiche und Treiber 79
5.3 Deskriptoren und Projektionen 85
5.4 Alternativenbündelung 94
5.5 Beschreibung der Szenarien 97
5.5.1 Trendszenario: Visionen verwirklichen 97
5.5.2 Enthusiastisches Szenario: Bytes im Freudentaumel 98

5.5.3 Horrorszenario: Datenkrieg 98
5.6 Störereignisse und deren Wirkung auf die Szenarien 99
6 Ableitung robuster Technologien und Implikationen für das strategische Management 106
6.1 Emergente Technologien ableiten 106
6.2 Robuste Technologien identifizieren und den Einfluss auf jedes Szenario darstellen 108
6.3 Implikationen für das strategische Management ableiten und Maßnahmen bestimmen .. 109
7 Fazit und Ausblick 111
8 Anhang 112
Literaturverzeichnis 118

Abbildungsverzeichnis

Abbildung 1: Suchanfragen zum Begriff „Industrie 4.0“ auf der Suchmaschine Google 7
Abbildung 2: Entwicklung des generierten Datenvolumens bis 2020 in Exabyte 8
Abbildung 3: Begriffsbestimmung Big Data gemäß BITKOM 9
Abbildung 4: Umfrage zum Nutzen von Big Data für verschiedene Geschäftsbereiche 2012 11
Abbildung 5: Anbieter von Big Data-Lösungen weltweit im Jahr 2013 (in Millionen US-Dollar) 12
Abbildung 6: Umsatz mit Big Data-Lösungen in Deutschland von 2013 bis 2016 (in Milliarden Euro) 13
Abbildung 7: Zukünftig wichtige Standorte für Technologieunternehmen 14
Abbildung 8: Ranking der wertvollsten digitalen Start-ups bis März 2015 (in Milliarden Dollar)..... 14
Abbildung 9: Hype Cycle für emergente Technologien aus 2014 der Firma Gartner Inc..... 20
Abbildung 10: Anzahl der Publikationen in den wichtigsten Fachzeitschriften und Konferenzen 21
Abbildung 11: Im Rahmen der Bibliometrie untersuchte wissenschaftliche Dimensionen..... 22
Abbildung 12: Anzahl der Publikationen pro Jahr 23
Abbildung 13: Anzahl der Suchanfragen in Google nach „Big Data“ von 2004 bis 2015 24
Abbildung 14: Geographische Verteilung der Publikationen, in % der Gesamtheit 26
Abbildung 15: Anzahl der veröffentlichten Artikel nach Organisationstyp 27
Abbildung 16: Anzahl der Artikel, sortiert nach dem Publikationskanal..... 28
Abbildung 17: Autorennetzwerk in Bezug auf das Schlagwort Big Data 30
Abbildung 18: Komponente um Prof. Jinjun Chen (Nr. 1) 31
Abbildung 19: Komponente um Yan Li (Nr. 2) 32
Abbildung 20: Komponente um Prof. Yi-Ke Guo (Nr. 3) 33
Abbildung 21: Komponente um PhD Yu Cao (Nr. 4) 34
Abbildung 22: Komponente um Prof. Volker Markl (Nr. 5) 35
Abbildung 23: Netzwerk der Organisationen mit den fünf größten Clustern 37
Abbildung 24: Teilkomponente um die University of Technology in Sidney, Australien (Nr. 1) 38
Abbildung 25: Teilkomponente um die University of Southern California (USC) in Los Angeles, USA (Nr. 2) 39
Abbildung 26: Teilkomponente um die Carnegie Mellon University in Pittsburgh, USA (Nr. 3) 41
Abbildung 27: Anzahl der Publikationen je Fachrichtung und gewichteter Anteil der Fachrichtungen 45
Abbildung 28: Netzwerk der Fachrichtungen 46
Abbildung 29: Netzwerk ausgewählter Kategorien mit übergeordneten Clustern 47
Abbildung 30: Die 25 wichtigsten Schlagwörter aus dem Datensatz 52
Abbildung 31: Netzwerk der wichtigsten Schlagwörter 53
Abbildung 32: Erklärung der Co-Zitation durch drei ausgewählte Publikationen 55
Abbildung 33: Wissensbasen - Dichteverteilung von co-zitierten Referenzen 56
Abbildung 34: Teilwissensbasis „Big Data - theoretischer Hintergrund“ 58
Abbildung 35: Teilwissensbasis „Sozialwissenschaften“ 60
Abbildung 36: Schematische Darstellung einer Forschungsfront 63
Abbildung 37: Forschungsfronten aus BibTechMon4 64
Abbildung 38: 3D Darstellung der Forschungsfronten 66
Abbildung 39: 3D Darstellung der Forschungsfronten, Vergrößerung um eine Zehnerpotenz 67
Abbildung 40: Modellierung verschiedener Datenquellen in einem Unternehmensnetzwerk 72
Abbildung 41: Forschungsportfolio 73
Abbildung 42: Semantisches Netz zu Big Data mit zehn Entitäten 76
Abbildung 43: Einflussbereiche und Treiber 80

Abbildung 44: System-Grid für den Begriff Big Data 83
Abbildung 45: Projektionen für die Deskriptoren des Treibers Datenschutz 86
Abbildung 46: Projektionen für die Deskriptoren des Treibers Gesetzgebung 87
Abbildung 47: Projektionen für die Deskriptoren des Treibers Effizienz 88
Abbildung 48: Projektionen für die Deskriptoren des Treibers wirtschaftlichen Erfolg 89
Abbildung 49: Projektionen für die Deskriptoren des Treibers Konnektivität 91
Abbildung 50: Projektionen für die Deskriptoren des Treibers Know-how 92
Abbildung 51: Projektionen für die Deskriptoren des Treibers IT-Sicherheit 93
Abbildung 52: Technologien der Anwendungsfelder von Big Data 107

Tabellenverzeichnis

Tabelle 1: Nutzen von Big-Data-Technologien in verschiedenen unternehmerischen Bereichen 10
Tabelle 2: Anzahl der veröffentlichten Artikel sortiert nach dem Herkunftsland 25
Tabelle 3: Wichtigste Ergebnisse aus der Analyse der Autoren 43
Tabelle 4: Wichtigste Ergebnisse aus der Analyse der Organisationen 44
Tabelle 5: Übersicht über ausgewählte Kategorien im Kontext *Industrie 4.0* 49
Tabelle 6: 40 am häufigsten verwendeten Web of Science Kategorien nach Publikationsanzahl 50
Tabelle 7: Wissensbasen zum Suchbegriff Big Data 57
Tabelle 8: Teilwissensbasis „Big Data - theoretischer Hintergrund“ 58
Tabelle 9: Teilwissensbasis "Sozialwissenschaften" 61
Tabelle 10: Forschungsfronten des Suchbegriffes „Big Data“ 65
Tabelle 11: Zeitliche Entwicklung der Publikationen in den Forschungsfronten ab dem Jahr 2011 68
Tabelle 12: Wichtige Forschungsfragen im Supply Chain Management 70
Tabelle 13: Einflussmatrix der Einflussfaktoren/Treiber 81
Tabelle 14: Relevanzmatrix der Treiber/Einflussfaktoren 82
Tabelle 15: Ranking der Treiber entsprechend der Relevanzsumme 82
Tabelle 16: Einflussbereiche und Treiber für die Szenarienentwicklung 84
Tabelle 17: Deskriptoren für den Einflussbereich *Gesellschaft* 86
Tabelle 18: Deskriptoren für den Einflussbereich *Wirtschaft* 88
Tabelle 19: Deskriptoren für den Einflussbereich *Technik* den Treibern zugeordnet 90
Tabelle 20: Übersicht über Treiber, den dazugehörigen Deskriptoren und deren Projektionen 95
Tabelle 21: Alternativenbündelung der Projektionen 96
Tabelle 22: „Datenkrieg“ mit Störereignis „globaler Stromausfall“ 100
Tabelle 23: „Bytes im Freudentaumel“ mit Störereignis „globaler Stromausfall“ 101
Tabelle 24: „Visionen verwirklichen“ mit Störereignis „globaler Stromausfall“ 102
Tabelle 25: „Datenkrieg“ mit Störereignis „künstliche Intelligenz“ 103
Tabelle 26: „Bytes im Freudentaumel“ mit Störereignis „künstliche Intelligenz“ 104
Tabelle 27: „Visionen verwirklichen“ mit Störereignis „künstliche Intelligenz“ 105
Tabelle 28: Strategische Roadmap für Big Data 108
Tabelle 29: Projektionskatalog für den Einflussbereich Gesellschaft 113
Tabelle 30: Projektionskatalog für den Einflussbereich Wirtschaft 114
Tabelle 31: Projektionskatalog für den Einflussbereich Technik 117

Abkürzungsverzeichnis

ACATECH	Akademie der Technikwissenschaften e.V.
AI	Artificial Intelligence
BD	Big Data
BCG	Boston Consulting Group
BI	Business Intelligence
BMBF	Bundesministerium für Bildung und Forschung
bzw.	beziehungsweise
CAD	Computer Aided Diagnostics
CE ProMM	Center of Excellence Produktmanagement, Universität Ulm
CEO	Chief Executive Officer
CGC	Conference on Cloud and Green Computing
CPS	Cyber-Physical Systems
CRM	Customer Relationship Management
CS	Computer Science
DFG	Deutschen Forschungsgemeinschaft
DIMA	Database Systems and Information Management
dpa	Deutsche Presseagentur
etc.	et cetera
F&E	Forschung und Entwicklung
ff.	fortfolgende
Hrsg.	Herausgeber
IEEE	Institute of Electrical and Electronics Engineers
IKT	Informations- und Kommunikationstechnologie
IT	Informationstechnik
i. V. m.	in Verbindung mit
KIT	Karlsruher Institut für Technologie
Nr.	Nr.
PESTEL	political, economic, social, technological ecological, legal
S.	Seite
SCM	Supply Chain Management
sog.	sogenannt(e, er, es)
TAB	Büro für Technikfolgen-Abschätzung beim deutschen Bundestag
TACC	Texas Advanced Computing Center
TIM	Technologie- und Innovationsmanagement
TU	Technische Universität
USC	University of Southern California
usw.	und so weiter
v.	von
vgl.	vergleiche
WoS	Web of Science
z. B.	zum Beispiel

1 Executive Summary

"The greatest shortcoming of the human race is our inability to understand the exponential function." - Albert A. Bartlett

Besonders wenn die Frage nach der Verwertbarkeit von großen, unstrukturierten Daten im Rahmen der aktuellen vierten industriellen Revolution – *Industrie 4.0* – thematisiert wird, muss ein Grundverständnis bei den Entscheidern vorhanden sein. Ziel in der aktuellen *Industrie 4.0* - Bewegung ist es, die deutsche Industrie vorzubereiten auf die technischen Herausforderungen und Möglichkeiten des 21. Jahrhunderts. Gemeint ist damit die Verschmelzung der physikalischen Welt mit der virtuellen Welt. Durch das Internet der Dinge kann eine Fabrik und ihr gesamtes Produktionsumfeld zu einer intelligenten Umgebung vernetzt werden. *Cyber-physical Systems* umfassen dabei „intelligente Maschinen, Lagersysteme und Betriebsmittel, die digital entwickelt wurden und durchgängig von der Eingangslogistik über die Produktion, das Marketing und die Ausgangslogistik bis zum *Service* mittels IKT verzahnt sind."[1]

Die Informationen stecken hierbei in den, von Maschinen und einzelnen Hardwarekomponenten produzierten Daten – *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." - Peter Sondergaard[2]

Daraus ergeben sich eine Vielzahl von Chancen und auch Risiken, denen mit Hilfe von Handlungsempfehlungen im strategischen Management begegnet werden muss. Diese werden in diesem Forschungsbericht mit den Methoden der **Bibliometrie** und der **Szenariotechnik** konkretisiert.

Hot Spots

Aus den **bibliometrischen Analysen** lassen sich die Schlagworte und damit auch die Themenbereiche *Hadoop*, *MapReduce*, *Machine Learning* und *Cloud Computing* als zentral identifizieren. Auf diesen Schlagwörtern basieren daher nicht nur viele internationale Veröffentlichungen; sie zeigen, dass sich das Thema Big Data im Zusammenhang mit der *Industrie 4.0* in der Zukunft in die Richtung der Entscheidungsunterstützung entwickelt.

Die wichtigste Konferenz fand 2013 in Santa Clara, Kalifornien statt, ausgerichtet von der IEEE. In diesem Zusammenhang wurden die meisten Artikel in den Jahren bis 2015 veröffentlicht. Alle anderen Fachzeitschriften oder Konferenzen veröffentlichen in deutlich kleinerem Rahmen.

Die Empfehlung aus den Analysen ist eine datenbasierte Unternehmensführung. Dies kann z. B. durch eine Ergänzung von *Operational Intelligence* Lösungen anhand einer Echtzeitübermittlung von wichtigen Informationen der Lieferkette erreicht werden. Auch Algorithmen gestützte Vorhersagen für Geschäfts-, Liefer- und Beschaffungsprozesse (*Predictive Analytics*) und Maschinen (*Predictive Maintanance*) können

[1] ACATECH (2011), S. 6
[2] Zitat von Sondergaard aus Gartner (2013)

zu einem Wettbewerbsvorteil beitragen. Im Bereich des *Machine Learnings* werden autonome Technologien, welche es den Maschinen oder Computersystemen ermöglichen, Probleme selbstständig zu lösen, in der Zukunft an Relevanz gewinnen.

Autorennetzwerke

Die wichtigsten Autoren sind Associate Prof. Jinjun Chen, Yan Li, Prof. Yi-Ke Guo und Prof. Volker Markl. Darüber hinaus weisen eine Vielzahl von kleineren Netzwerken untereinander Verknüpfungen durch Forschungsaktivitäten auf. Zusätzlich wird die Vielschichtigkeit innerhalb des Themenfeldes *Big Data* durch die Abstände der einzelnen Knoten und gesamten Komponenten zueinander verdeutlicht. Damit bestätigt sich die bisher getroffene Vermutung, dass *Big Data* ein **interdisziplinäres Forschungsfeld** ist. Dies wird auch durch die Breite der Forschungsthemen *MapReduce* Optimierung, Analytiktechnologien für die translationale Medizin und neuartige Informationssysteme für Wirtschaftsunternehmen unterstrichen.

Forschungsfronten

Der überwiegende Teil der für die *Industrie 4.0* relevanten Forschungsfronten findet sich in der BCG-Matrix im Bereich der „Question Marks“ und an der Grenze zu den „Stars“ wieder. Dies legt nahe, dass in der Zukunft Themen wie selbstlernende Algorithmen oder aber der Einsatz von Big-Data-Technologien im Marketing eine bedeutende Rolle spielen werden.

Die Popularität der theoretischen Forschung am Übergang von den „Stars“ zu „Cash Cows“ reduziert sich minimal. Dies ist auf die stagnierende Zahl der Erkenntnisse seit 2013 zurückzuführen. Dasselbe gilt für die Forschungsfront „*Big Data* Möglichkeiten für die Wissenschaft und Industrie“, in der generelle Einsatzgebiete von *Big-Data-Technologien* identifiziert wurden.

Szenarien für die Zukunft von *Big Data*

Ausgehend von der bibliometrischen Analyse wurden mit den Einflüßen aus Makro- und Mikroumwelt drei Szenarien entwickelt. „**Bytes im Freudentaumel**“ beschreibt die wachsende Verwendung von digitalen Technologien im alltäglichen und Geschäftsleben. Außerdem ist die Akzeptanz in der Bevölkerung am höchsten im Vergleich zu den anderen beiden Szenarien. Die Technologien für Datensammlung und –verarbeitung werden sehr gut verstanden, was auf ein hohes Maß an technologiegetriebenen Ausbildungen hinweist.

Im Gegensatz dazu wird der „**Datenkrieg**“ von Skepsis und Misstrauen gegenüber den Datenmengen und Technologien geprägt. Dazu gehört auch eine Ablehnung der Bevölkerung was die Nutzung moderner Technik angeht. Dies ist meist zurückzuführen auf eine geringe Akzeptanz aufgrund mangelnden Verständnisses. Zusätzlich entwickeln die Algorithmen eine Intelligenz, die in Konkurrenz zu den Menschen steht, was eine weitere Separation hervorruft.

Moderat hingegen ist das Szenario „**Visionen verwirklichen**“, in dem Menschen und Maschinen Hand in Hand arbeiten. Auf Seiten der Menschen ist man sich der Herausforderung bewusst ein gesundes Mittelmaß an eigenständig denkender Algorithmen gewähren zu lassen, sodass diese dem Menschen in größtmöglicher Weise assistieren aber keine unkontrollierte Eigendynamik entwickeln können. In diesem

Szenario ist eine maximale Nutzung der Kompetenzen beider Seiten möglich und generiert so gesteigerte Produktivität und Zufriedenheit.

Robuste Technologien

Innerhalb des Szenario-Transfers werden robuste Technologien werden, die über alle Szenarien Bestand haben. Hierfür wird die Unternehmensperspektive eingenommen, zukünftige Marktsegmente gebildet und mögliche Produktfelder abgeleitet. Zu den robusten Technologien zählen die Basis-Softwarepakete in Form von **Programmiersprachen** (Java, C-Familie, Python) und **Infrastrukturen** (*Hadoop*, *MapReduce*) sowie **kundenspezifische Lösungen**, zumeist im Bereich der Analyse von großen Datenmengen.

Zusätzlich kann im Rahmen einer technologischen Roadmap formuliert werden, dass in den nächsten 5 Jahren die Entwicklung der drei Bereiche „**Datenbank-Systeme**“, „**Verarbeitung großer Datenmengen**“ und „**Echtzeitanalysen**“ zunehmen wird, um als Unternehmen konkurrenzfähig zu bleiben.

Wichtiger Hinweis:

Zur Technologievorausschau sind in der ITOP-Schriftenreihe bereits erschienen:

1. Technologievorausschau – eine Bestandsaufnahme der organisationalen Umsetzung in Unternehmen (ITOP-Schriftenreihe Nr. 2)
2. Future Shopping (ITOP-Schriftenreihe Nr. 4)
3. „3D-Druck“ (ITOP-Schriftenreihe Nr. 5)

Diese können ebenfalls über das ITOP bezogen werden:

http://www.uni-ulm.de/mawi/itop/itop-schriftenreihe.html

Bei Rückfragen zu diesem Bericht oder sonstigen Fragen wenden Sie sich bitte an:

Universität Ulm
iTOP Institut für Technologie- und Prozessmanagement
Frau Anette Lesle

Helmholtzstraße 22
89081 Ulm
Tel.: +49 (0) 7 31/50-3 23 01
anette.lesle@uni-ulm.de

2 Industrie 4.0 – Anwendung von Big-Data-Technologien

Der Begriff *Industrie 4.0* ist in der deutschen Wirtschaft eine sehr stark diskutierte Entwicklung im Rahmen des technischen Fortschrittes. Das Bundesministerium für Bildung und Forschung hat den Begriff folgendermaßen definiert: „Das Zukunftsbild *Industrie 4.0* zeigt eine mögliche Zukunft der industriellen Produktion aus der Perspektive des Jahres 2025 und blickt zurück auf die „Vergangenheit" des Jahres 2012"[3].

Ziel ist es, die deutsche Industrie vorzubereiten auf die technischen Herausforderungen und Möglichkeiten des 21. Jahrhunderts. Um ein klares Bild der vierten industriellen Revolution zu erlangen, werden zunächst die vier Entwicklungsschritte bis zum heutigen Tage aufgezeigt.

Ende des 18. Jahrhunderts begann mit dem ersten mechanischen Webstuhl die erste industrielle Revolution. In dieser Phase wurde Wasser- und Dampfkraft für den Betrieb mechanischer Produktionsanlagen verwendet. Mit dem Beginn des 20. Jahrhunderts wurde dann zusätzlich elektrische Energie für den Betrieb verwendet. So sind elektrisch angetriebene Fließbänder charakteristisch für die zweite industrielle Revolution. Ab 1960 wurde die Automatisierung der Produktion vor allem durch IT und Elektronik vorangetrieben. So wurde 1969 die erste speicherprogrammierbare Steuerung in Betrieb genommen. Die vierte industrielle Revolution lässt sich ab dem heutigen Tag mit der Nutzung von *Cyber-Physical Systems* (CPS) beschreiben.[4]

Gemeint ist damit die Verschmelzung der physikalischen Welt mit der virtuellen Welt. Durch das Internet der Dinge kann eine Fabrik und ihr gesamtes Produktionsumfeld zu einer intelligenten Umgebung vernetzt werden. CPS umfassen dabei „intelligente Maschinen, Lagersysteme und Betriebsmittel, die digital entwickelt wurden und durchgängig von der Eingangslogistik über die Produktion, das Marketing und die Ausgangslogistik bis zum *Service* mittels IKT verzahnt sind."[5]

Dabei werden verschiedene Ziele wie zum Beispiel die individualisierte Produktion verfolgt. Es können individuelle Produkte in kleinen Mengen zu günstigen Preisen produziert werden.

Ein weiteres Ziel ist die Vernetzung von Unternehmen, wodurch effizientere Prozesse erreicht werden. Daraus resultieren Wettbewerbsvorteile aufgrund von flexiblen, anhand von Datenanalysen optimierten Wertschöpfungsnetzwerke. Darüber hinaus wird der Arbeitskomfort durch intelligente Assistenzsysteme gesteigert.

Als Chance zu sehen ist die Entwicklung neuer Geschäftsfelder, die innerhalb des Forschungsberichts aufgezeigt werden.

Hauptaugenmerk dieser Ausarbeitung liegt auf dem Einsatz von *Big-Data-Technologien* im Zusammenhang mit dem Begriff *Industrie 4.0*. Um was es sich bei *Big Data* handelt, wird in Kapitel 2.1 erläutert.

[3] Vgl. BMBF (2014), S. 6
[4] Vgl. ACATECH (2011), S. 17
[5] Vgl. ACATECH (2011), S. 6

Der Begriff *Industrie 4.0* wurde vor allem im deutschsprachigen Raum geprägt. International wird die Entwicklung als *Smart Production*, *Smart Manufacturing*, *Smart Factory* und *Advanced Manufacturing* bezeichnet.[6] International agierende Unternehmen mit Hauptsitz in Deutschland verbreiten allerdings den Begriff *Industrie 4.0* immer mehr.[7]

Die starke Bindung zwischen dem Begriff und der deutschsprachigen Länder lässt sich auch nach einem Blick auf die Analyse Suchanfragen auf Google bestätigen (vgl. Abbildung 1). Es ist offensichtlich, dass dieses Themengebiet erst seit Ende 2012 im Internet gesucht wird und das Interesse seitdem beständig zunimmt. Des Weiteren kann angemerkt werden, dass dieser Begriff nur in Deutschland und Österreich vermehrt gesucht wurde.

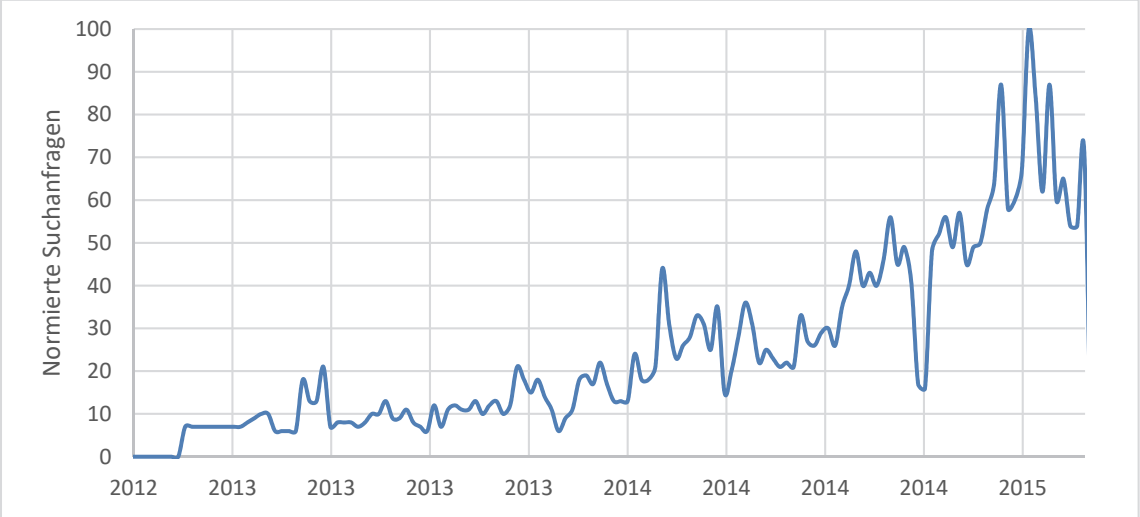


Abbildung 1: Suchanfragen zum Begriff „Industrie 4.0“ auf der Suchmaschine Google[8]

[6] Vgl. ACATECH (2011), S. 71
[7] Vgl. BMBF (2013), S. 9
[8] Vgl. Google (2015b)

2.1 Big Data - eine Begriffsbestimmung

Um den Begriff *Big Data* zu verstehen, muss zunächst die Exponentialfunktion ins Gedächtnis gerufen werden, da diese fundamental für das Verständnis ist. In Abbildung 2 ist zu sehen, dass das jährlich generierte digitale Datenvolumen exponentiell anwächst, was sich besonders eindrücklich im Jahr 2015 zeigt.

"The greatest shortcoming of the human race is our inability to understand the exponential function."
- Albert A. Bartlett

Im Jahr 2015 wurden mehr Daten generiert als in fast allen bisherigen Jahren zusammen. Folglich werden Technologien benötigt, die diese enormen Datenmengen verwalten können. Genau hier setzt *Big Data* an. Als wichtige Entwicklung ist hier, die exponentiell steigende Rechenkapazität zu nennen, welche sich alle 18 Monate verdoppelt und somit eine Grundlage für die technologischen Lösungen bildet, die aus den Anforderungen von *Big Data* resultieren. Ein weiterer Punkt ist die Datenvielfalt, da in Zukunft nicht nur strukturierte, sondern auch unstrukturierte Daten ausgewertet werden können.

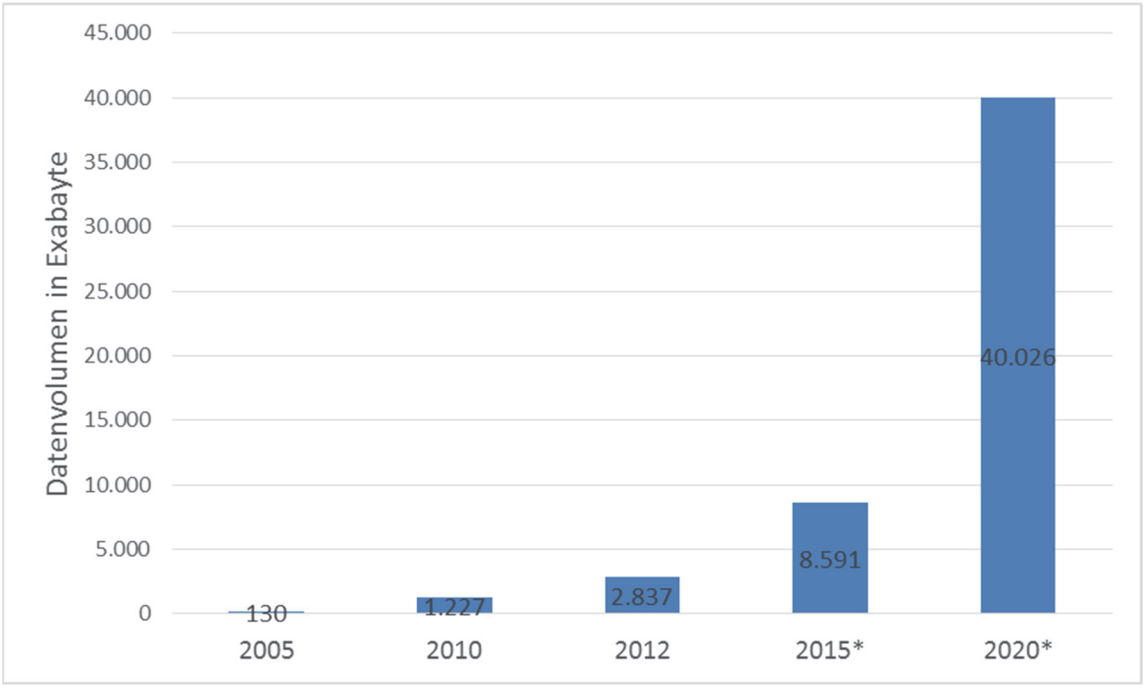


Abbildung 2: Entwicklung des generierten Datenvolumens bis 2020 in Exabyte[9]

Die neue Datenvielfalt resultiert beispielsweise aus Maschinen- und Sensordaten in der industriellen Produktion oder von fertigen Produkten. Außerdem können aus sozialen Netzwerken, wie Facebook und Twitter, Daten einbezogen werden um z. B. eine genauere Marktsegmentierung vornehmen oder Trends frühzeitig erkennen zu können. Diese Daten müssen nicht unbedingt in strukturierten Datenbanken vorliegen, sondern können auch einzelne Dokumente sein.

Ein weiterer wichtiger Aspekt ist die Analyse dieser Daten nach Mustern, Zusammenhängen, Bedeutungen und Vorhersagen. Ein praktisches Beispiel handelt hierbei von der sogenannten *Predictive Maintenance* des Logistikunternehmens DB Schenker, welches in Zusammenarbeit mit IT-

[9] Vgl. EMC (2015)

Systems ein *Big Data* System für die vorrausschauende Wartung implementierte. Für die Betreiber von Lokomotiven können Defekte aufgrund resultierender Ausfallzeiten sehr kostspielig sein. Daher generiert das neue System aus verschiedenen Daten schon im Voraus mögliche Schwachstellen durch die ein Ausfall eintreten könnte. Diese Informationen werden automatisch an die entsprechende Stelle weitergeleitet, um so proaktiv einen Ausfall zu beheben.[10] Das Beispiel zeigt zudem, dass Daten in der Zukunft verlässlich in Echtzeit ausgewertet werden müssen. Die technischen Herausforderungen hierbei sind nicht nur die schnelle Dateienformatierung und –konvertierung zwischen den unterschiedlichsten Dateitypen, sondern auch verlässliche und vertrauenswürdige Ergebnisse zu erhalten. Die Dimensionen des Begriffes *Big Data* sind in Abbildung 3 grafisch dargestellt.

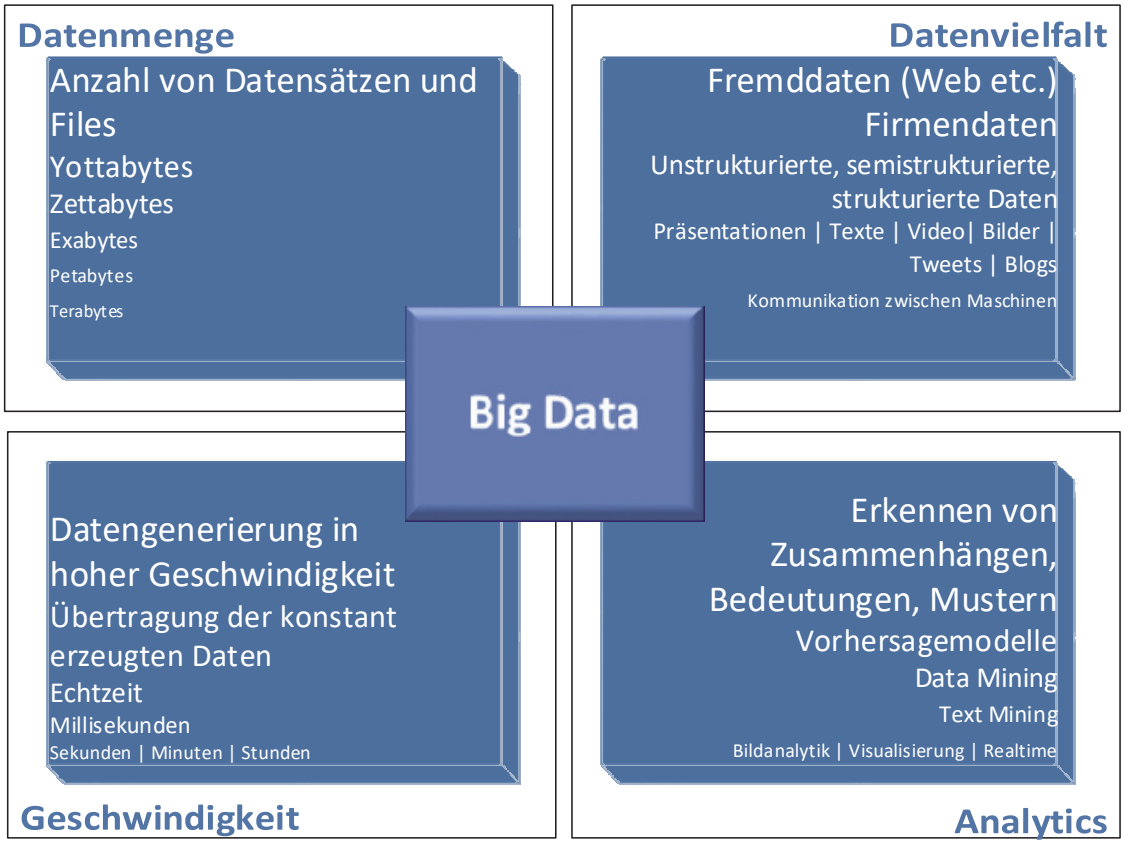


Abbildung 3: Begriffsbestimmung Big Data gemäß BITKOM[11]

Eine kurze und passende Definition liefert unter anderem das Beratungsunternehmen Gartner Inc.: „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.“[12], wodurch ersichtlich wird, dass Entscheidungen aufgrund von Datenanalysen werden immer wichtiger werden.

Die Technologien in dem Zusammenhang, im weiteren Verlauf des Berichts als *Big-Data-Technologien* bezeichnet, sind vor allem der *MapReduce* Algorithmus zur schnellen und unkomplizierten Datenverarbeitung sowie das *Hadoop* Framework. Sie dienen als Fundament und werden von unzähligen Tools als Infrastrukturgrundlage verwendet ohne namentlich explizit erwähnt zu werden. In diesem Atemzug müssen auch die Algorithmen und somit die Programmiersprachen erwähnt

[10] Vgl. BITKOM (2014), S. 32
[11] Vgl. BITKOM (2012)
[12] Vgl. Gartner (2013)

werden, mit denen diese Infrastrukturen gebaut werden. Dazu gehören neben Java und der C-Familie auch weniger verbreitete Sprachen, wie z. B. die noch junge Sprache Python, welche im Bereich der Analysetools immer wichtiger wird.

2.2 Big Data im Zusammenhang mit Industrie 4.0

Im vorherigen Kapitel wurde der Begriff *Big Data* bestimmt. Im Anschluss werden Zusammenhänge mit dem Themengebiet *Industrie 4.0* aufgezeigt.

Unternehmerische Bereiche	Potentieller Nutzen durch Big Data
Marketing & Vertrieb	- Kostenreduzierung in Marketing und Vertrieb - Erhöhung des Umsatzes bei Verkaufsvorgängen - Markt- und Wettbewerberanalysen - Erhöhung von Point of Sales Umsätzen - Management von Kundenabwanderungen
Forschung und Produktentwicklung	- Echtzeitauswertung von komplexen Daten aus wissenschaftlichen Experimenten - Produktneuentwicklungen und -verbesserungen - Social-Media-Trendanalysen für neue Produktideen - Erprobung neuer Medikamente in der Pharmaentwicklung - Verbesserung der Kosteneffizienz im Gesundheitswesen und Fernüberwachung kritischer Parameter bei Patienten
Produktion, Service und Support	- Produktionsoptimierung mit Maschinen- und Sensordaten - Produktionsplanung bei Saisonartikeln in Abhängigkeit von vielfältigen Faktoren - Früherkennung von Produktproblemen im Service durch den Einsatz von Diagnosedaten - Vorausschauende Instandhaltung von Maschinen und Anlagen
Distribution und Logistik	- Optimierung von Lieferketten (Supply Chain Management) - Nutzung von Telematik zur Verminderung von Stillstandzeiten bei LKW-Transporten - Optimierung bei der Mauterhebung
Finanz- und Risiko-Controlling	- Echtzeitreaktionen auf Geschäftsinformationen - Simulationen, Vorhersagen und Szenarienbildung - Manipulationsprävention und Betrugserkennung - Risikocontrolling in Echtzeit - Erkennen von Kreditrisikofaktoren

Tabelle 1: Nutzen von Big-Data-Technologien in verschiedenen unternehmerischen Bereichen[13]

Zunächst kann festgehalten werden, dass *Big-Data-Technologien* ein fundamentaler Baustein zur Umsetzung von der Zukunftsvision *Industrie 4.0* sind. In Tabelle 1 sind alle unternehmerischen Bereiche aufgeführt, auf die *Big-Data-Technologien* Einfluss nehmen können. Hier kann auch das große Potential für die Entwicklung im Zug des Zukunftsszenarios *Industrie 4.0* gesehen werden.

[13] Vgl. BITKOM (2012), S. 34–43

Bei einer Umfrage durch das Magazin Computerwoche aus dem Jahr 2012 wurden verschiedene Unternehmen befragt, welche Geschäftsbereiche den höchsten Business-Nutzen aus *Big Data* ziehen können und welche Themen relevant sind. Dabei wird deutlich, dass vor allem im Bereich des Controllings ein sehr großer Nutzen gesehen wird (vgl. Abbildung 4). Auffällig ist auch, dass jeglicher Nutzen im Bereich der Optimierung angesiedelt ist. Das kann dadurch erklärt werden, dass *Big Data* Lösungen aufgrund der Auswertung von einer Vielzahl von Daten in nahezu Echtzeit bessere Entscheidungen im Sinne von aktuellen Entwicklungen auf Kunden- oder Marktsicht führen.

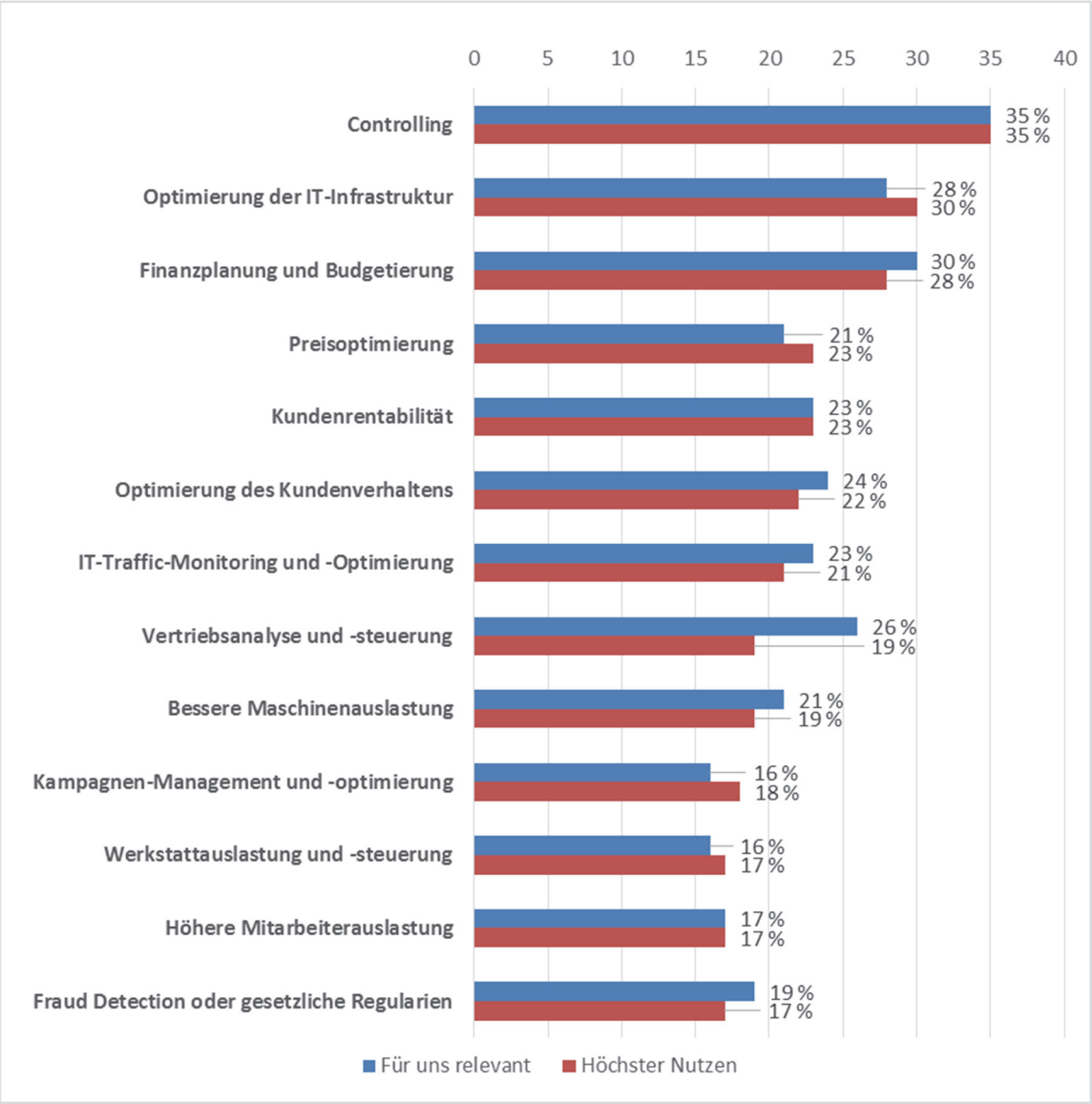


Abbildung 4: Umfrage zum Nutzen von Big Data für verschiedene Geschäftsbereiche 2012[14]

In diesem Abschnitt wird auf die größten Anbieter von *Big Data* Lösungen eingegangen. Die umsatzstärksten Unternehmen sind in Abbildung 5 aufgelistet, wobei IT-Unternehmen wie zum Beispiel IBM, HP, Dell und SAP die größten Umsätze stellen. Jedoch gibt es eine Vielzahl von Anbietern, die neu in den Markt kommen und nicht in dieser Liste auftauchen. Welche Lösungen beispielsweise IBM anbietet wird genauer in Kapitel 4.3.2.2 betrachtet.

[14] Vgl. Computerwoche (2012)

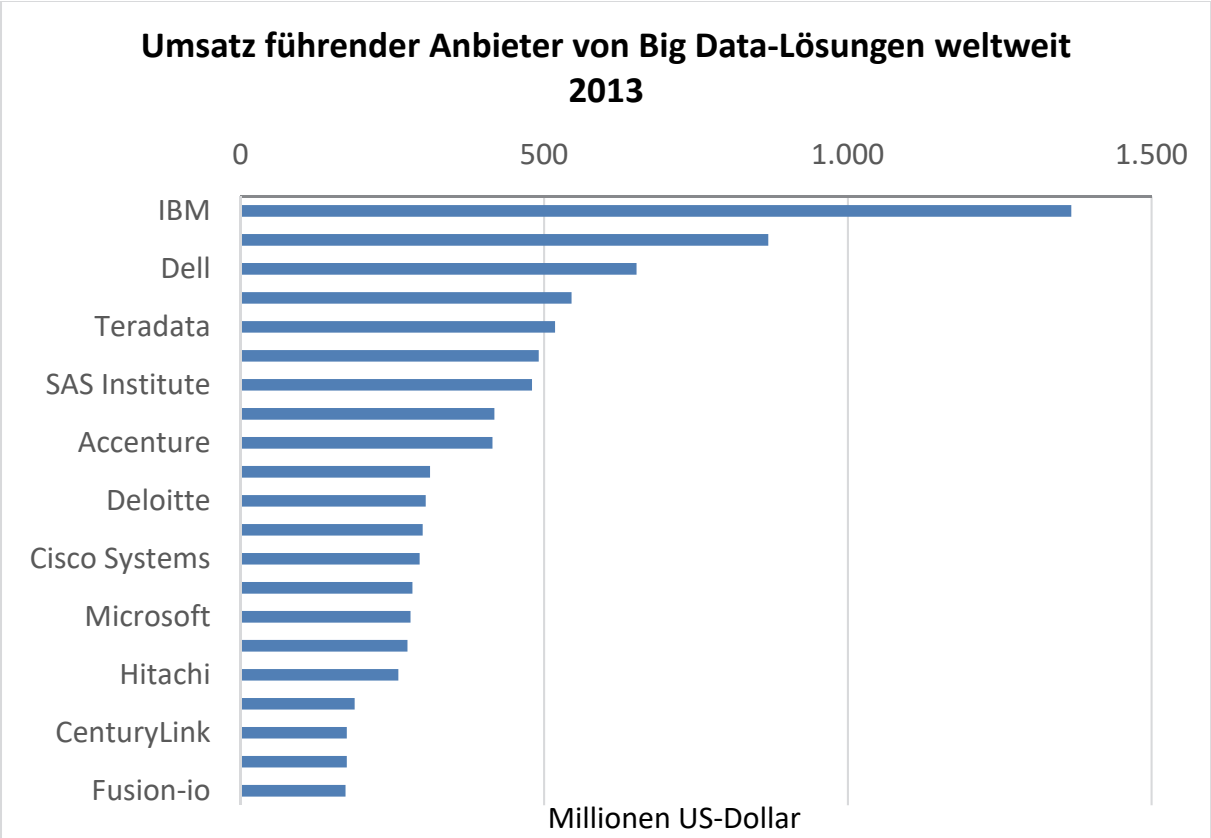


Abbildung 5: Anbieter von Big Data-Lösungen weltweit im Jahr 2013 (in Millionen US-Dollar)[15]

Wichtig für die Bewertung von *Big Data* Anwendungen im Zusammenhang mit *Industrie 4.0* ist auch eine Betrachtung der Umsätze durch *Big Data* Lösungen in Deutschland. Die dazugehörige Umsatzentwicklung ist in Abbildung 5 aufgeführt. Deutlich wird hier, dass von einem großen Wachstum ausgegangen wird und die Entwicklung erst ihren Anfang beschreibt. In Zukunft werden in sehr vielen Branchen und Unternehmensbereichen *Big Data* Lösungen eingesetzt werden.

[15] Vgl. Wikibon (2013)

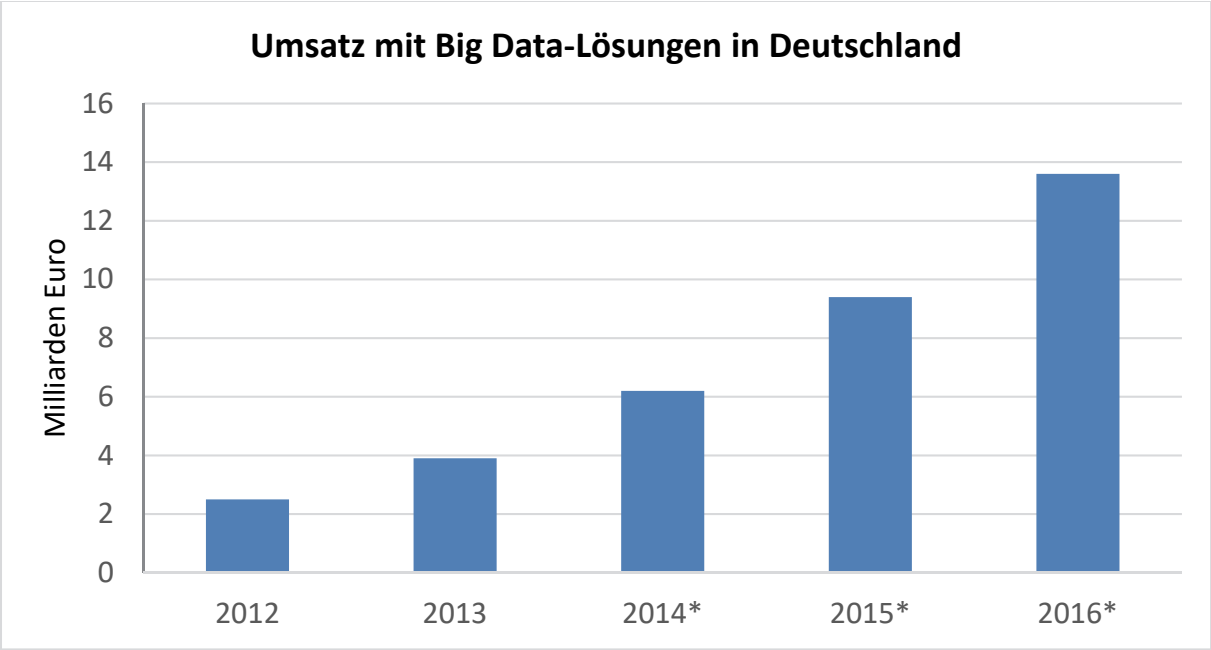


Abbildung 6: Umsatz mit Big Data-Lösungen in Deutschland von 2013 bis 2016 (in Milliarden Euro) [16]

Eine offene Frage bezüglich dem Einsatz von *Big Data* Lösungen betrifft vor allem den Nutzen personenbezogener Daten im Bereich des *Data Minings*. Darunter fallen Lösungen rund um den Vertrieb und das Marketing. Wie können *Social Media* Daten genutzt werden, damit sie im Einklang mit dem deutschen Datenschutz stehen? Welche *Big Data* Verfahren können dazu grundsätzlich angewendet werden?

Diese Fragen rund um den Datenschutz sind vor allem im Zuge der NSA Veröffentlichungen durch Edward Snowden in den Vordergrund getreten.

> *„Ich will nicht in einer Welt leben, in der alles, was ich sage, alles was ich mache, der Name jedes Gesprächspartners, jeder Ausdruck von Kreativität, Liebe oder Freundschaft, aufgezeichnet wird."*
>
> - Edward Snowden

Die Öffentlichkeit ist in punkto Datenschutz sehr sensibel und könnte die Entwicklung von manchen *Big Data* Lösungen nachhaltig beeinflussen, wenn die Datenvielfalt als restriktiv und nicht als Chance verstanden wird.[17] Die Betrachtung erfolgt hier sehr kritisch, obwohl viele aktuelle *Big Data* Anwendungen mit den deutschen Datenschutzgesetzen vereinbar sind.

In den USA hingegen stehen die Chancen durch *Big-Data-Technologien* im Vordergrund. Bei einer Umfrage unter Führungskräften international agierender Unternehmen zeigte sich ein klares Bild, wonach sich die zukünftigen Technologieunternehmen vor allem in den USA und dort speziell im Silicon Valley ansiedeln werden (vgl. Abbildung 7). Vorreiter auf dem Gebiet der Datenanalysen wie Google Inc., Apple Inc., Oracle und andere sind dort bereits beheimatet.

[16] Vgl. BITKOM (2014)
[17] Vgl. Schrage (2012)

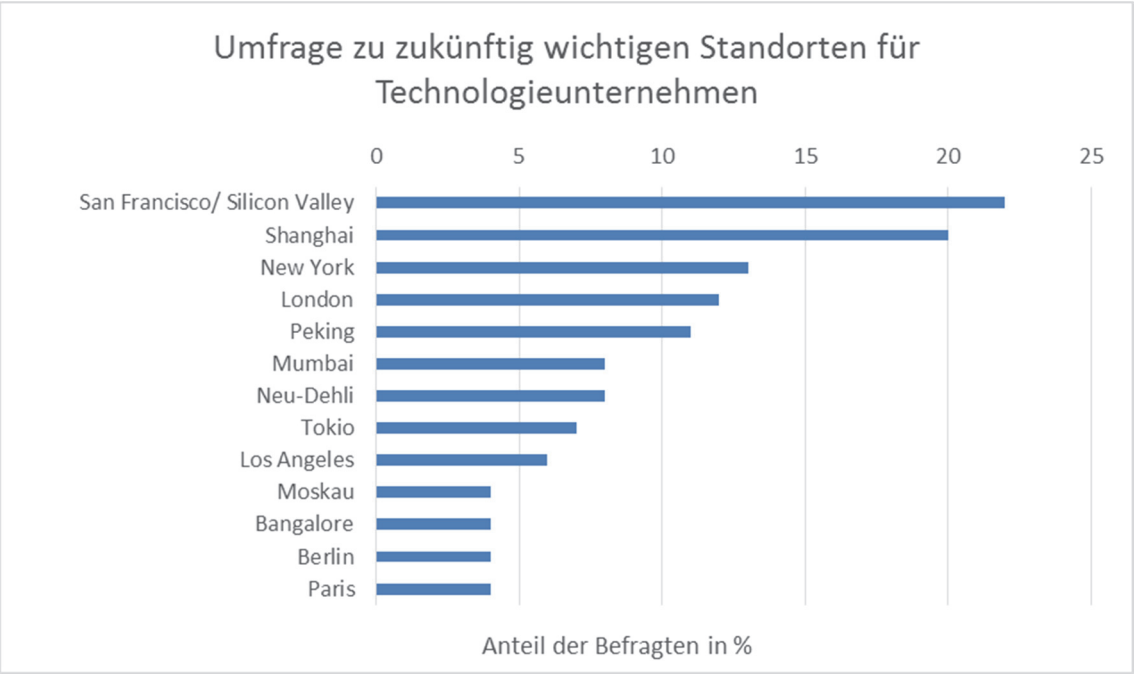


Abbildung 7: Zukünftig wichtige Standorte für Technologieunternehmen[18]

In der Liste der wertvollsten Start-Ups des Jahres 2015 tauchen zum größten Teil nur amerikanische Start-Ups auf und primär Start-Ups aus dem Silicon Valley (vgl. Abbildung 8).

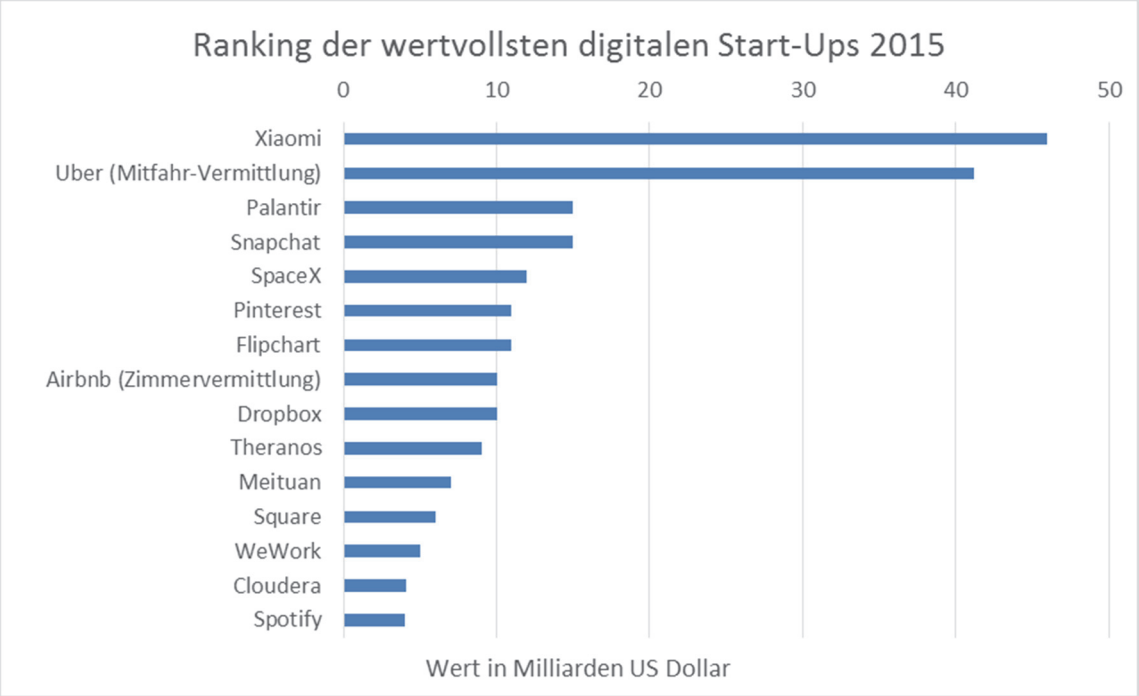


Abbildung 8: Ranking der wertvollsten digitalen Start-ups bis März 2015 (in Milliarden Dollar)[19]

Es ist zu erkennen, dass wenige Start-Ups aus Asien kommen und Spotify das einzige europäische Unternehmen in dieser Liste ist. Der überwiegende Teil der in Abbildung 8 genannten Start-Ups

[18] Vgl. EY (2013)
[19] Vgl. Wall Street Journal (2015)

benötigen für ihr Geschäftsmodell *Big-Data-Technologien*. Die Veränderungen, die dadurch auf die Firmen mit Geschäftsmodellen ohne *Big-Data-Technologien* zukommen, werden in dieser Forschungsarbeit betrachtet.

Die Zukunft in der IT-Branche entwickelt sich Dank der exponentiell ansteigenden Rechenkapazität und Datenmenge rasend schnell fort. Die Lösungen im *Big Data* Bereich werden viele Arbeitsgebiete verändern. Um die möglichen Veränderungen zu betrachten wird in diesem Forschungsbericht eine Technologievorrauschau erstellt. Dabei wird ein Datensatz mit wissenschaftlichen Veröffentlichungen zum Begriff *Big Data* untersucht. Folgende Fragen werden beantwortet: Welche Wissenschaftler und Universitäten spielen herausragende Rollen? Welche Themengebiete werden erforscht? In welchen Ländern wird veröffentlicht?

Gerade deutsche Unternehmen mit Geschäftsmodellen ohne *Big-Data-Technologien* müssen darauf gefasst sein, dass sie Konkurrenz von branchenfremden Start-Ups bekommen. Das Silicon Valley ist Meister im Erkennen von Schwächen in den Geschäftsmodellen etablierter Unternehmen. „Es greift voller Beißlust an - und trägt meistens den Sieg davon.“[20]

[20] Vgl. Keese (2014), S. 107

3 Forschungsmethodischer Rahmen

3.1 Bibliometrie

Die bibliometrische Analyse untersucht bibliografische Informationen unter Verwendung von mathematischen und statistischen Methoden.[21] Die Datengrundlage im vorliegenden Fall sind wissenschaftliche Publikationen. Mit Hilfe der Bibliometrie können wissenschaftliche Publikationsleistungen sowie Forschungsergebnisse und aktuelle Entwicklungen in einem bestimmten Themenfeld analysiert werden.[22] Fragestellungen, die in diesem Zusammenhang untersucht werden, sind zum Beispiel, welche Forschungsschwerpunkte ein bestimmtes Thema aufweist, welche Experten an den Themen arbeiten oder wie sich Forschungsschwerpunkte in Publikationen im Zeitverlauf entwickeln. Da sich aus den Beobachtungen technologische Trends ableiten lassen, kommt der Bibliometrie im Bereich der Technologievorausschau eine besondere Bedeutung zu.[23]

Die bibliometrische Analyse erfolgt mit Hilfe des Programmes BibTechMon des Austrian Institute of Technology (AIT). Diese Software ermöglicht es, mit Hilfe des oben genannten Datensatzes Informationen über die wissenschaftlichen Veröffentlichungen im Zusammenhang mit *Big Data* zu generieren und Netzwerke für die jeweiligen Schwerpunkte der Analysen zu erstellen.

Folgende wesentlichen Ergebnisse aus der bibliometrischen Analyse werden in der vorliegenden Arbeit dargestellt.[24]

- *Deskriptive Statistiken*

Deskriptive Statistiken dienen dazu, einen (quantitativen) Überblick in das Forschungsgebiet zu bekommen. Hierzu gehören u. a. Informationen zu zentralen Autoren und Organisationen. Zentrale Autoren bzw. Organisationen sind solche, die häufig zitiert werden oder die selbst viel publiziert haben.

- *Netzwerk der Autoren und Organisationen*

Auf Basis der Ko-Autorenschaft visualisieren diese Netzwerke die Ähnlichkeit von Autoren und im gleichen Sinne auch von Organisationen. Der Mehrwert der Netzwerkdarstellung im Vergleich zur deskriptiven Statistik liegt in der Darstellung von Beziehungen (Kanten, hier als Ko-Autorenschaften zu interpretieren) zwischen den Knoten (Autoren oder Organisationen). Auf diese Weise kann analysiert werden, wie stark isoliert oder vernetzt geforscht wird. Für die folgenden Ergebnisse gilt die Annahme, dass Zitationen einen Hinweis darauf liefern, dass Artikel in einem thematisch inhaltlichem Zusammenhang stehen.

- *Wissensbasen*

Sind die Knoten im Netzwerk keine Autoren, sondern Referenzen, können grundlegende Forschungsthemen entdeckt werden, welche die intellektuellen Wurzeln repräsentieren. Eine

[21] Vgl. Havemann (2009), S. 7
[22] Vgl. Ball et al. (2005), S. 15
[23] Vgl. Havemann (2009), S. 7
[24] Vgl. Stelzer et al. (2015)

Anhäufung von Artikeln, welche ko-zitiert sind, nennt man demzufolge Wissensbasen. Zwei Artikel sind ko-zitiert, wenn diese Artikel beide in einem dritten Artikel referenziert werden.[25]

- *Forschungsfronten*

Sind die Knoten im Netzwerk Artikel, so können Forschungsfronten aufgedeckt werden, welche emergente Forschungsthemen repräsentieren. Zwei Artikel sind bibliographisch gekoppelt, wenn beide Artikel einen dritten referenzieren. Je mehr Artikel gemeinsam zitiert werden, desto stärker ist die Kopplung. Eine Anhäufung bibliografischer Artikel bilden eine Forschungsfront und enthalten die neuesten Publikationen zu einem Forschungsthema.[26]

- *Forschungsportfolio*

Die Forschungsfronten werden auf eine Zeitachse gelegt und in Anlehnung an die BCG-Matrix angeordnet.[27] Der Mehrwert des Forschungsportfolios liegt in der Darstellung der Forschungsthemen über die Zeit. Es kann abgeleitet werden, welche Themen in den letzten Jahren stark an Bedeutung gewonnen haben, welche Themen erst gerade aufgekommen sind und welche Themen schon viel beforscht sind und womöglich bald in die Unternehmenspraxis abwandern.

3.2 Szenariotechnik

Die Szenariotechnik ermöglicht grundsätzlich das Visionieren unterschiedlicher Zukünfte für einen Betrachtungsgegenstand. Dieser Betrachtungsgegenstand wird als System definiert. Es gilt, die Einflüsse auf dieses System zu bestimmen. Diese werden detailliert beschrieben (Deskriptoren) und in die Zukunft projektiert (Projektionen). Die Projektion erfolgt dabei nicht (nur) trendspezifisch im Sinne eines möglichen Trendverlaufs, sondern in Form von Alternativen. Das heißt, es resultieren mehrere zukünftige Zustände (Szenarien), welche intern möglichst homogen und stabil sind und sich voneinander unterscheiden.

Für die Ableitung der Szenarien gibt es eine Vielzahl an Vorgehensweisen. Für den Zweck der Technologievorausschau verwenden wir[28] im vorliegenden Bericht die Vorgehensweise nach Stelzer et al. (2015). Dort erfolgt die Technologievorausschau anhand von acht Schritten.[29]

1. Systemische Beschreibung des Analyseobjekts
2. Identifizierung und Strukturierung von Einflussbereichen, Einflussfaktoren und Deskriptoren und deren Vernetzung
3. Ableiten von Projektionen und Bestimmung der Anzahl an Szenarien
4. Ableiten von Rohszenarien durch die Bildung konsistenter Projektionsbündel
5. Beschreibung der Szenarien

[25] Vgl. hierzu den Aufsatz von Small (1973)
[26] Vgl. hierzu den Aufsatz von Kessler (1963)
[27] Vgl. BCG (2015). Die x-Achse spiegelt den relativen Anteil der Artikel der Forschungsfront zu allen Artikeln des betrachteten Themas. Die y-Achse entspricht dem relativen Wachstum (Anzahl der Publikationen in Periode 2 – Anzahl der Pubilkationen in Periode 1)/Publikationen in Periode 1). Die Kreisgröße ist die Anzahl der Publikationen im Gesamtzeitraum.
[28] Die Szenariotechnik wurde im Workshop-Format durchgeführt. Beteiligte Personen waren die Autoren dieses Berichts.
[29] Vgl. für die Methodenkombination aus Bibliometrie und Szenariotechnik Stelzer et al. (2015)

6. Identifizierung von Störereignissen und Bestimmung deren Einflusses auf die Szenarien
7. Identifizierung emergenter Technologien und Bestimmung robuster Technologien, die über alle Szenarien Bestand haben
8. Ableitung von Handlungsempfehlungen.

Wichtig ist, dass an die Szenariobildung (Schritt 5) bzw. die Störereignisanalyse die Identifizierung emergenter bzw. robuster Technologien anschließt. Außerdem sollen die Erkenntnisse aus der Technologievorausschau für das strategische Management aufbereitet werden.

In den nachfolgenden Kapiteln werden zunächst die Ergebnisse der bibliometrischen Analyse vorgestellt. Anschließend werden auf Basis der oben dargestellten Vorgehensweise drei Szenarien für den Betrachtungsgegenstand “Future Shopping” entwickelt, emergente bzw. robuste Technologie bestimmt und Handlungsempfehlungen für ein fiktives Einzelhandelsunternehmen abgeleitet.

4 Bibliometrische Analyse zu Big Data in der Industrie 4.0

Zusammenfassung

Aus den bibliometrischen Analysen lassen sich die Schlagworte und damit auch die Themenbereiche *Hadoop*, *MapReduce*, *Machine Learning* und *Cloud Computing* als zentral identifizieren. Auf diesen Schlagwörtern basieren daher nicht nur viele internationale Veröffentlichungen; sie zeigen, dass sich das Thema *Big Data* im Zusammenhang mit der *Industrie 4.0* in der Zukunft in die Richtung der Entscheidungsunterstützung entwickelt.

Die Empfehlung aus den Analysen ist eine datenbasierte Unternehmensführung. Dies kann z. B. durch eine Ergänzung von Operational Intelligence Lösungen anhand einer Echtzeitübermittlung von wichtigen Informationen der Lieferkette erreicht werden. Auch Algorithmen gestützte Vorhersagen für Geschäfts-, Liefer- und Beschaffungsprozesse (Predictive Analytics) und Maschinen (Predictive Maintanance) können zu einem Wettbewerbsvorteil beitragen. Im Bereich des *Machine Learnings* werden autonome Technologien, welche es den Maschinen oder Computersystemen ermöglichen, Probleme selbstständig zu lösen, in der Zukunft an Relevanz gewinnen.

Ziel der nachfolgend dargestellten bibliometrischen Analyse ist es, Schwerpunktthemen der wissenschaftlichen Veröffentlichungen rund um das Themengebiet *Big Data* zu identifizieren. Diese werden eingeteilt in die klassische Forschung und Forschungsfelder mit hohem zukünftigem Potential. Weiter soll beantwortet werden, welche dieser Themen wirtschaftlich umsetzbar sind und wie viel Zeit dazu benötigt wird. Dies kann als Technologievorrauschau angesehen werden, welche in der Ergründung von emergenten Technologien resultiert. Die bibliometrische Analyse generiert ausgehend von wissenschaftlichen Veröffentlichungen quantitative Aussagen über ein Forschungsgebiet.[30] Darüber hinaus geben statistische Methoden Informationen über z. B. Autoren, Forschungsdisziplinen und zukünftige Technologien. Der Vorausschauhorizont beträgt fünf Jahre und soll dem strategischen Managements eines Unternehmens dabei helfen wettbewerbsfähig zu bleiben und Risiken zu minimieren.

4.1 Einbettung des Technologiefeldes in die Dimensionen der Wissenschaft

In diesem Kapitel werden die wichtigsten und aktuellsten Forschungsfelder im Bereich *Big Data* und insbesondere im Kontext von *Industrie 4.0* identifiziert. Dies beinhaltet die Identifikation von relevanten Instituten und Personen, die wissenschaftliche Arbeit auf diesem Gebiet leisten. Darüber hinaus wird die Entwicklung der Forschungslandschaft in den vergangenen Jahren beleuchtet. Dazu werden die Publikationen im Zusammenhang mit *Big Data* systematisch geordnet und aufbereitet. Damit kann z. B. die zeitliche Entwicklung der Publikationen genauer betrachtet werden oder mithilfe der Information über das Herkunftsland der jeweiligen Publikation herausgefunden werden, welche Länder besonders viel auf dem Gebiet *Big Data* forschen. Außerdem kann aus dem Datensatz herausgelesen werden, welche Organisationstypen (z. B. Universitäten, Unternehmen) in der Forschungslandschaft am Stärksten vertreten sind.

[30] Vgl. Van Raan (2003)

4.1.1 Hotspots

Als Hotspot wird der forschungsintensivste Zeitraum aller Publikationen oder die am häufigsten mit dem Schlagwort *Big Data* in Verbindung gebrachte Fachzeitschrift bezeichnet. Für die Identifizierung dieser Hotspots wird der Zeitraum 1974 bis 2015, entsprechend des Publikationsdatensatzes, betrachtet.

Die Technologieberatungsfirma Gartner Inc. veröffentlicht seit vielen Jahren jährlich den sogenannten „Hype Cycle for Emerging Technologies“, der aufgrund der verschiedenen Phasen der öffentlichen Aufmerksamkeit Rückschlüsse auf die Entwicklungsreife einer neu eingeführten oder entdeckten Technologie zu ziehen. Abbildung 9 zeigt diesen Hype Cycle für emergente Technologien für das Jahr 2014.

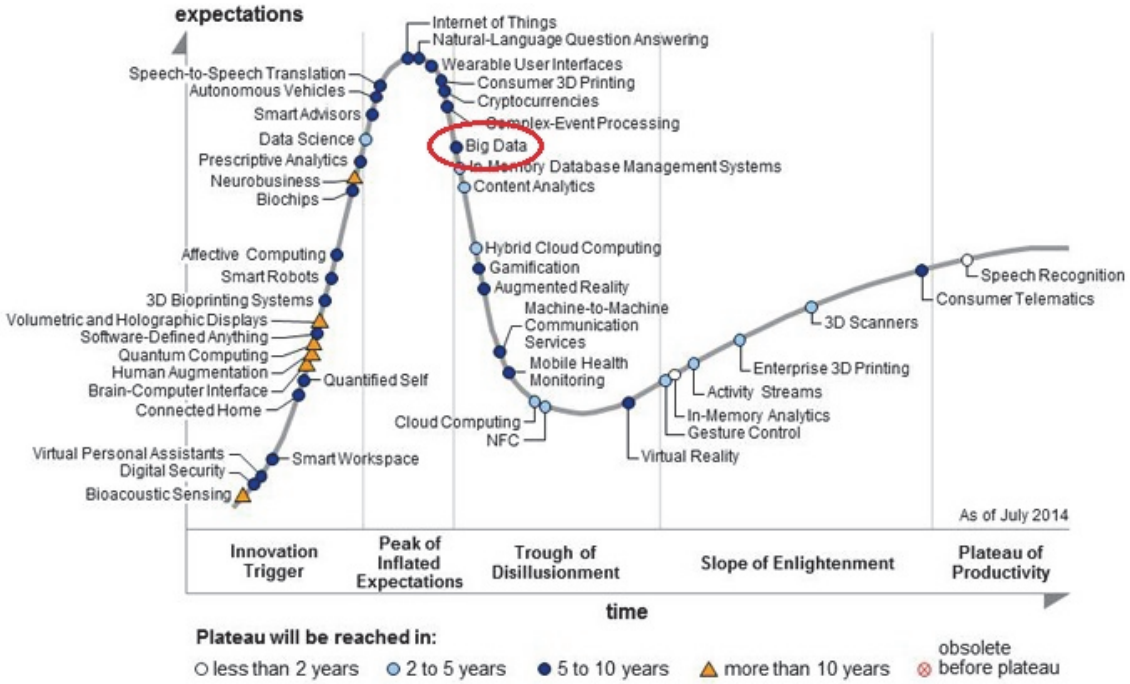


Abbildung 9: Hype Cycle für emergente Technologien aus 2014 der Firma Gartner Inc.[31]

Bewertet wird für jede Technologie der Grad der Erwartungen anhand des Entwicklungsstadiums. Deshalb befinden sich Technologien wie zum Beispiel der Quantencomputer noch in der „Innovation Trigger“-Phase. In Abbildung 9 ist insbesondere die Technologie *Big Data* gekennzeichnet. Diese befindet sich laut Gartner schon über dem Berg der „überzogenen“ Erwartungen und auf dem Weg in das Tal der Desillusionierung. Das heißt, dass sich eventuelle Risiken und Schwierigkeiten bei der Umsetzung der Technologie nun zunehmend bemerkbar machen. Deshalb sind Technologien, wie zum Beispiel das *Cloud Computing*, bereits tief in diesem Tal der Desillusionierung, da datenschutzrechtliche Bedenken und die Sicherheit immer mehr in den Fokus der Öffentlichkeit treten. Die Art des Datenpunktes im Falle *Big Data* bedeutet, dass die Analysten von Gartner in 5 bis 10 Jahren damit rechnen, dass die Technologie „reif“ sein wird.

[31] Vgl. Gartner Inc. (2014)

Als erster Hotspot wurde aufgrund der hohen Publikationsanzahl Associate Prof. Dr. Jinjun Chen von der „Faculty of Engineering and IT“ der Universität in Sidney, Australien, identifiziert. Er hat insgesamt 14 Artikel innerhalb des Datensatzes generell zum Thema *Big Data* und im Speziellen zu außerordentlich datenintensiven Systemen, *Cloud Computing*, sowie *Workflow Management* verfasst. Eine der Publikationen trägt den Titel „SaC-FRAPP: a scalable and cost-effective framework for privacy preservation over big data on cloud“[32]. Prof. Chen ist außerdem Associate Editor mehrerer Fachzeitschriften und Vorsitzender des „IEEE Computer Society’s Technical Committee on Scalable Computing (TCSC)“. Mit seinen bisher über 120 publizierten Artikeln ist er führend auf seinem Forschungsgebiet.

Ein weiterer Hotspot ergibt sich aus der Analyse der Fachzeitschriften oder Konferenzen, die zum Publizieren der wissenschaftlichen Arbeiten von den Autoren genutzt werden. Diese Analyse liefert Informationen darüber, welche Einrichtungen (Fachzeitschriften, Bücher, Tagungen) für das Forschungsgebiet *Big Data* besonders wichtig sind. Abbildung 10 zeigt die sechs am häufigsten verwendeten Zeitschriften oder Konferenzen der insgesamt 1391 Institutionen mit der Anzahl der in ihnen publizierten Artikel.

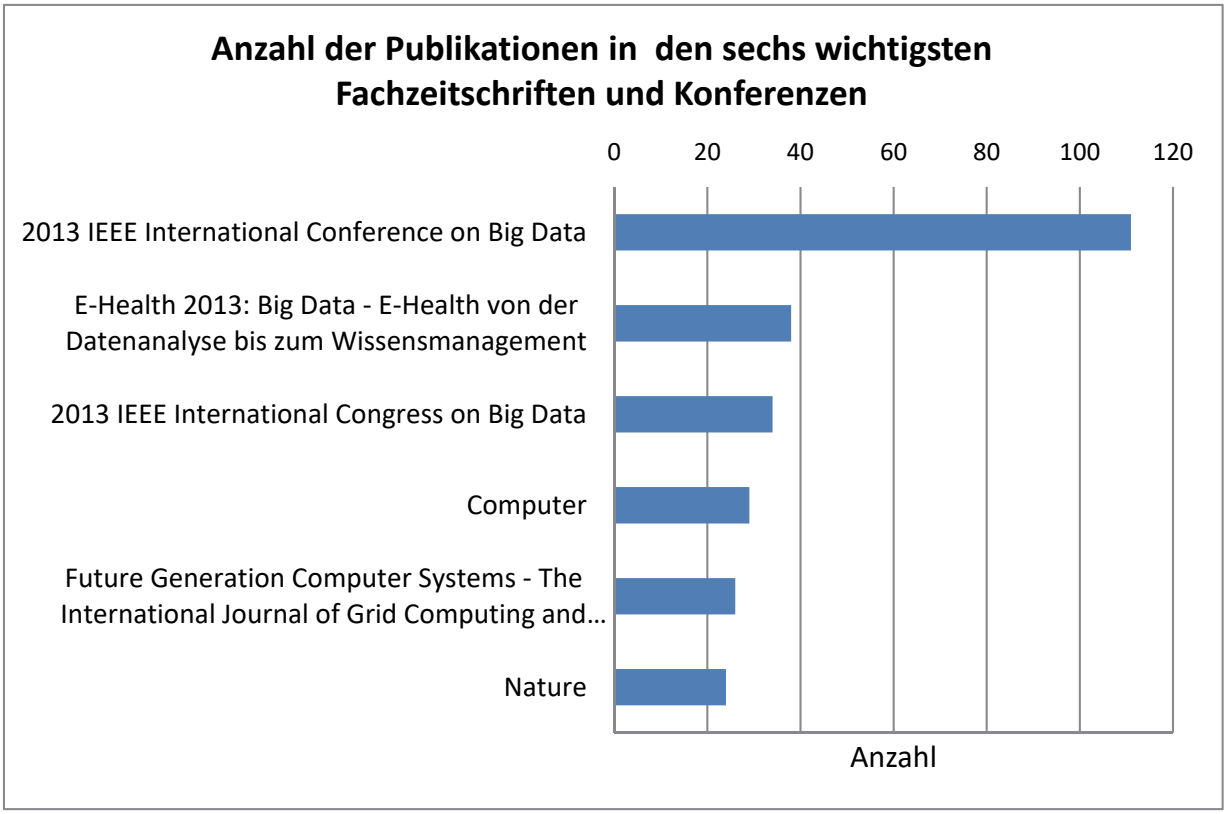


Abbildung 10: Anzahl der Publikationen in den wichtigsten Fachzeitschriften und Konferenzen[33]

Mit 111 Veröffentlichungen ist die „2013 IEEE International Conference on *Big Data*“ mit großem Abstand die wichtigste Konferenz. Die „eHealth“- Konferenz 2013 ist mit 38 Veröffentlichungen an zweiter Stelle zu nennen. Als erste Fachzeitschrift befindet sich das „Computer Journal“ der Oxford University Press auf Platz vier der Rangliste. „NATURE“, die über alle Wissenschaftsbereiche hinweg

[32] Vgl. Zhang et al. (2013)

[33] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

bekannteste Fachzeitschrift, ist mit 24 Veröffentlichungen vertreten. Anhand der breiten Streuung über die verschiedenen Fachbereiche hinweg, kann ein hohes Interesse an der Verwendung von *Big Data* bezogenen Anwendungen diagnostiziert werden.

4.1.2 Die Dimensionen der Wissenschaft



Abbildung 11: Im Rahmen der Bibliometrie untersuchte wissenschaftliche Dimensionen

Nachfolgend soll das Technologiefeld *Big Data* hinsichtlich verschiedenster Eigenschaften (Dimensionen der Wissenschaft) charakterisiert werden. Ziel ist es hier, Publikationskanäle und Organisationstypen zu benennen, sowie die geographische und zeitliche Entwicklung aufzuzeigen. Abbildung 11 zeigt die untersuchten Dimensionen der Wissenschaft, die im Folgenden einzeln beleuchtet werden.

4.1.2.1 Zeitliche Entwicklung der Publikationen

Rückschlüsse auf das Alter des Forschungsgebietes *Big Data* kann mit Hilfe einer Übersicht entsprechend Abbildung 12 gewonnen werden. Diese zeigt die Anzahl der Publikationen pro Jahr, sortiert nach dem Erscheinungsjahr der jeweiligen Artikel. Es wird deutlich, dass das Forschungsfeld noch sehr jung ist, denn in der Zeit zwischen 1974, als der erste Artikel mit dem Schlagwort *Big Data* im Titel erschien und ungefähr 2008 wurden jedes Jahr nur zwischen einem und sechs Artikel veröffentlicht. Die Artikel in diesem Zeitraum beschäftigen sich nur im weitesten Sinne mit denjenigen *Big Data* Anwendungen, um die es in diesem Forschungsbericht gehen soll.[34] Die Rechenkapazitäten und -geschwindigkeiten zu dieser Zeit waren so gering, dass die Algorithmen damals dazu verwendet wurden, mathematische Probleme zu lösen. Das Verständnis hinter dem Begriff *Big Data* hat sich in dem betrachteten Zeithorizont von einem rein mathematischen zu einem holistischen entwickelt.

[34] Vgl. Wainer et al. (1974)

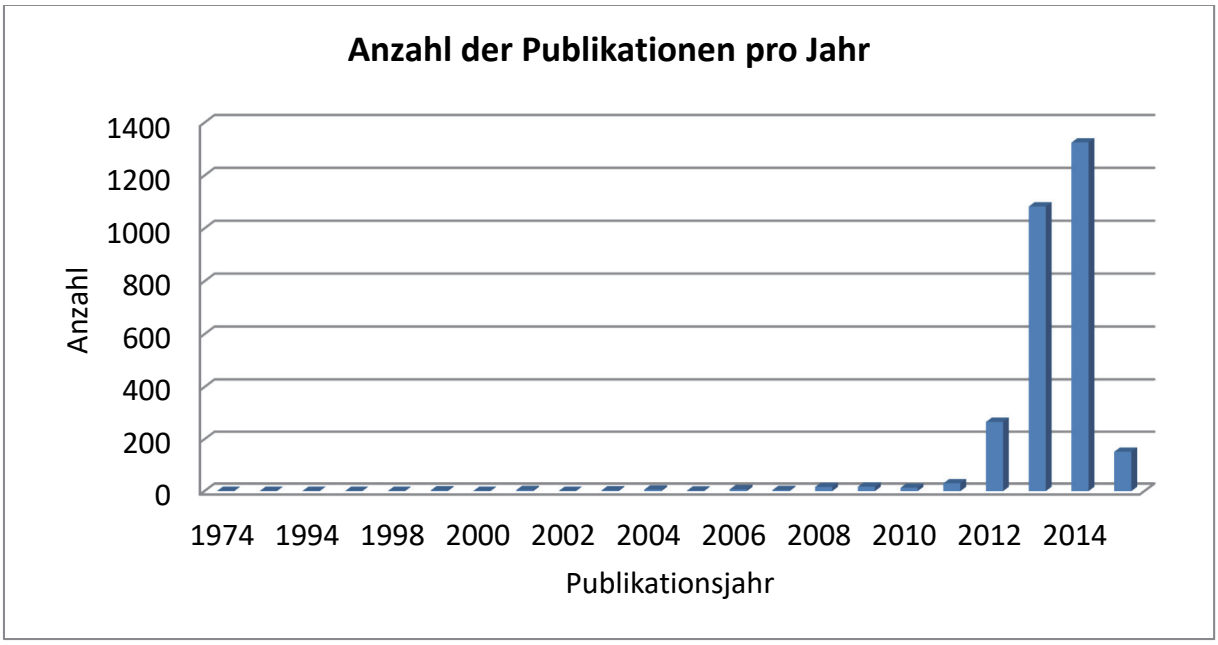


Abbildung 12: Anzahl der Publikationen pro Jahr[35]

Zwar beschäftigen sich die Artikel zwischen 1974 und ungefähr 2008 auch mit großen Datensätzen, aber dies sind homogene, sprich gleichartige Datensätze, während *Big Data* die schnelle Verarbeitung vor allem vielfältiger Datensätze bezeichnet.[36] Allerdings ist aus Abbildung 12 ersichtlich, dass ab ca. 2008 die Relevanz von *Big Data* ein exponentiell ansteigendes Verhalten zeigt, bis die jährliche Anzahl an Artikeln im Jahr 2014 bei 1325 liegt. Das Jahr 2008 spielt bei der Entwicklung eine zentrale Rolle, da in diesem Jahr das *Hadoop Framework* der Firma Apache Software Foundation bereitgestellt wurde. Dieses Programm basiert auf dem *MapReduce* Algorithmus, der bereits 2004 von Google entwickelt wurde[37], aber erst 2008 in einem Journal veröffentlicht wurde. Das *Hadoop Framework* bot damals zum ersten Mal die Möglichkeit, aufwendige Rechenoperationen im Terabyte-Bereich auf verschiedene Systeme zu verteilen, um die benötigte Berechnungszeit drastisch zu verringern.

Neben der bibliometrischen oder statistischen Analyse von wissenschaftlichen Artikeln kann man sich auch anders einen groben Überblick über die Wichtigkeit und deren zeitlichen Verlauf verschaffen. Das Unternehmen Google Inc. bietet dafür zum Beispiel die Möglichkeit, sich den zeitlichen Verlauf der Anzahl der Suchanfragen nach einem (oder mehreren) bestimmten Schlagwörtern anzeigen zu lassen. Abbildung 13 zeigt den Verlauf eines solchen Trends, wobei das Schlagwort *Big Data* ist und die Anzahl der Suchanfragen von Google stets bezogen auf die höchste Zahl in Prozent angegeben wird. Wie zu erwarten war, zeigt sich hier eine sehr große Ähnlichkeit zur Analyse der wissenschaftlichen Artikel. Wie in Abbildung 13 zu sehen ist, bleibt die Häufigkeit der Suchanfragen bis ca. 2011 stets unter 10 Prozent vom Maximum. Danach zeigt sich ein exponentieller Anstieg der Suchanfragen, bis das Maximum ungefähr im Jahr 2015 erreicht wird, was sich auch mit Abbildung 12 deckt.

[35] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[36] Vgl. Gartner (2013)
[37] Vgl. Dean und Ghemawat (2008)

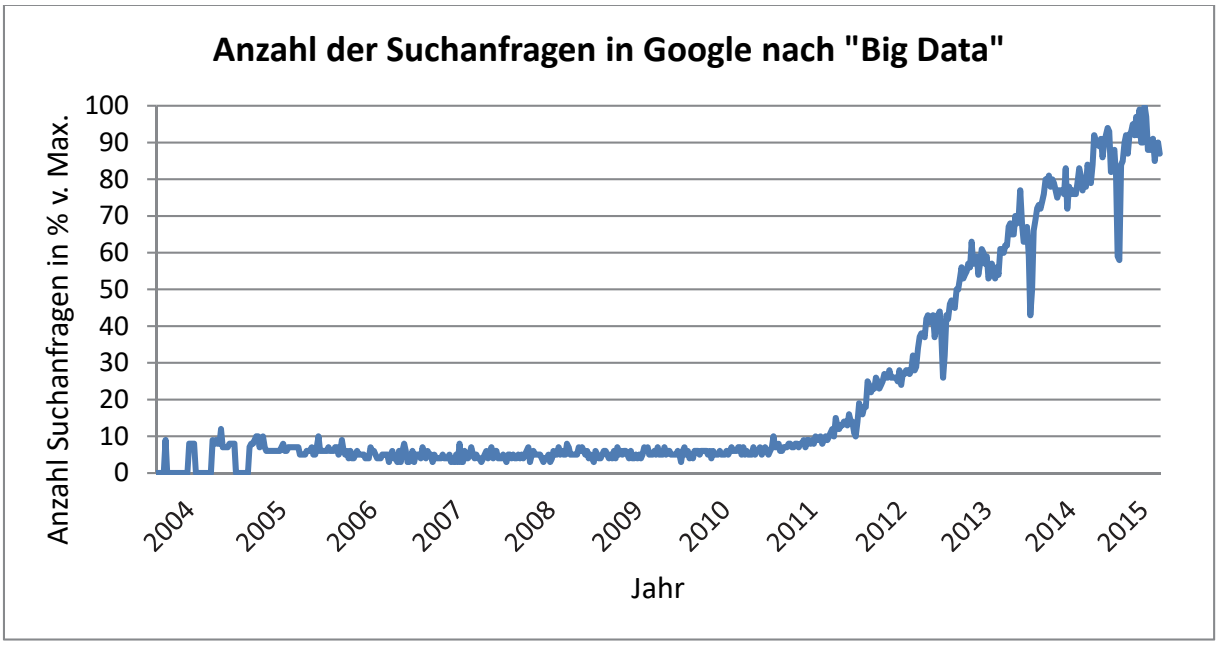


Abbildung 13: Anzahl der Suchanfragen in Google nach „Big Data“ von 2004 bis 2015[38]

4.1.2.2 Geographische Verteilung der Publikationen

In diesem Abschnitt sollen die Publikationen, ähnlich wie in Abschnitt 4.1.2.1 für die zeitliche Entwicklung, nun hinsichtlich ihrer Herkunft charakterisiert werden. Tabelle 2 zeigt die Anzahl der veröffentlichten Artikel, sortiert nach dem Herkunftsland des Instituts (bzw. der Organisation) und in absteigender Reihenfolge der Häufigkeit.

[38] Vgl. Google (2015a)

Land	Anzahl Artikel	Land	Anzahl Artikel
USA	605	Schweden	8
China	147	Norwegen	7
Großbritannien	67	Indien	6
Australien	50	Frankreich	6
Kanada	26	Italien	5
Süd Korea	25	Qatar	5
Singapur	20	Irland	4
Deutschland	17	Rumänien	4
Japan	17	Brasilien	4
Spanien	13	Israel	4
Österreich	11	Zypern	3
Taiwan	11	Kroatien	2
Schweiz	11	Saudi Arabien	2
Niederlande	9	Thailand	2
Polen	9		

Tabelle 2: Anzahl der veröffentlichten Artikel sortiert nach dem Herkunftsland[39]

Es ist ersichtlich, dass die Vereinigten Staaten (605 Publikationen) mit großem Abstand die meisten Beiträge in Form von Publikationen zum Forschungsgebiet *Big Data* beisteuern. Aber auch die asiatischen Staaten wie China (147 Publikationen) veröffentlichen im Vergleich zu den europäischen Ländern wie Großbritannien (67 Publikationen) sehr viel.

Diese Tendenz ist zu erwarten, da die meisten der großen *Big Data* Firmen ihren Sitz im Silicon Valley in den USA haben.[40] Diese Dichte an High-Tech-Firmen spiegelt sich auch in der Forschungslandschaft insbesondere an Universitäten und Forschungseinrichtungen wider. Es bleibt abzuwarten, ob sich dieses Verhältnis bzw. diese Vormachtstellung der Vereinigten Staaten in den nächsten Jahren so fortsetzt, oder ob Europa durch mediale Aufmerksamkeit den Begriff der *Industrie 4.0* international etablieren kann und speziell im Bereich *Big Data* vermehrt in den Aufbau dieses Forschungszweiges investiert. Es wäre auch denkbar, dass gerade die europäischen Unternehmen in *Big Data*-Anwendungen Chancen sehen und somit die Forschungslandschaft in Europa dahingehend verändern. Um einen besseren Überblick zu erhalten, zeigt Abbildung 14 die Anzahl der veröffentlichten Artikel in Prozent der Gesamtheit und nach Kontinenten sortiert.

[39] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

[40] Vgl. Keese (2014)

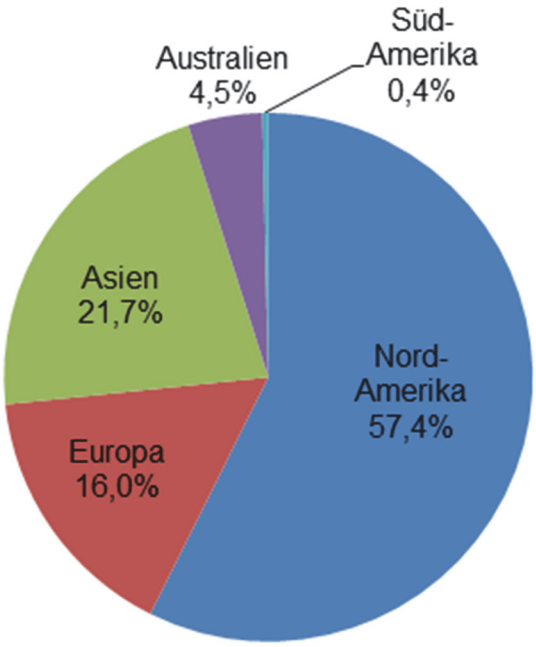


Abbildung 14: Geographische Verteilung der Publikationen, in % der Gesamtheit[41]

Danach kommen über 50 Prozent aller hier untersuchten Publikationen aus Nordamerika. Etwas weniger als ein Viertel aus Asien, gefolgt von Europa mit 16 Prozent und Australien bzw. Südamerika haben kaum einen Anteil an der Forschungsaktivität. Wichtig ist es zu erwähnen, dass Großbritannien dafür verantwortlich ist, dass Europa einen solch großen Anteil besitzt, denn mit 67 Veröffentlichungen liegt das Vereinigte Königreich fast gleichauf mit den USA, wenn man die jeweilige Einwohnerzahl berücksichtigt.

4.1.2.3 Organisationstypen

Nachdem in den beiden vorangegangenen Kapiteln Informationen über die zeitliche Entwicklung der Anzahl der Publikationen, sowie deren Herkunft gesammelt wurden, soll nun die Art der Organisation im Vordergrund stehen. Abbildung 15 zeigt diese Statistik für die Organisationstypen.

[41] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

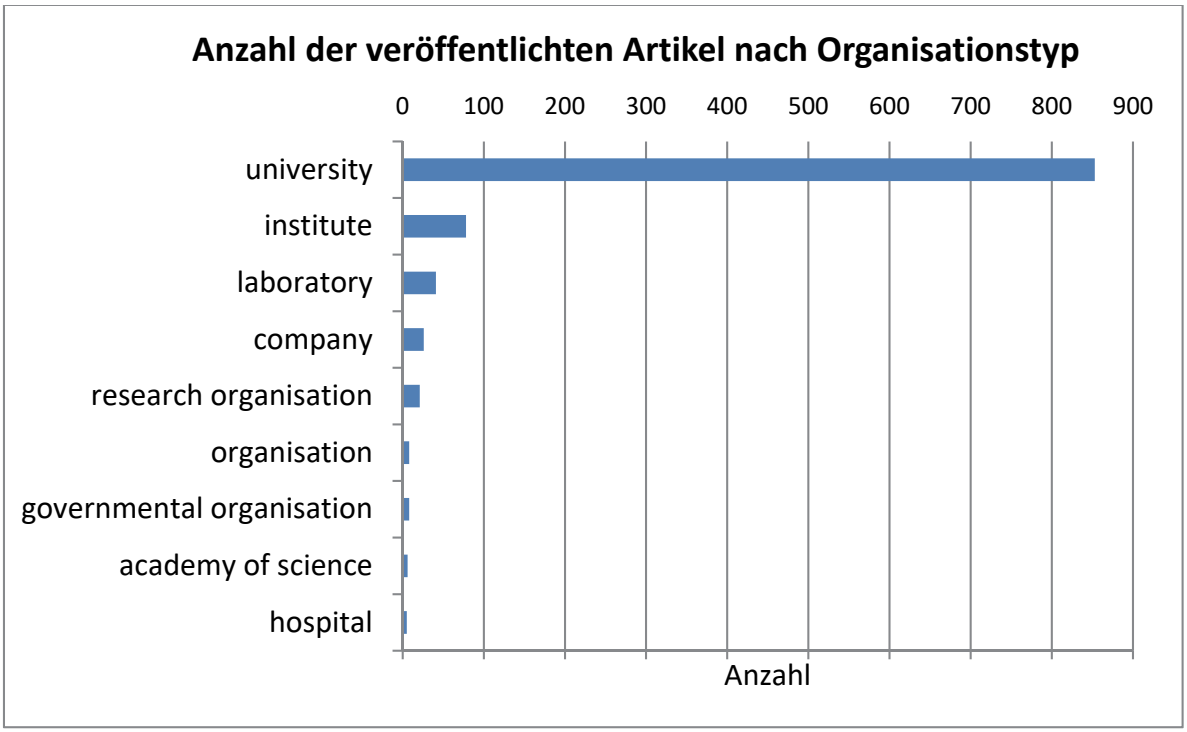


Abbildung 15: Anzahl der veröffentlichten Artikel nach Organisationstyp[42]

Abbildung 15 zeigt, dass die universitären Institute mit sehr großem Abstand (ungefähr 80 Prozent) den größten Anteil haben. Dahinter befinden sich mit etwa sieben Prozent die außeruniversitären Institute, wie etwa in Deutschland die Max-Planck Institute. Erst an vierter Stelle sind mit etwa 2,5 Prozent die Unternehmen zu finden. Dieser Sachverhalt trifft sehr wahrscheinlich auf alle Forschungsgebiete zu, denn an Universitäten wird die Forschungsaktivität hauptsächlich an der Anzahl der veröffentlichten wissenschaftlichen Artikel gemessen. Bei Unternehmen ist jedoch der Erfolg eines Produkts das Maß für den Erfolg der Forschungs- und Entwicklungsabteilungen. Diese sind demnach nicht sehr daran gebunden, wissenschaftliche Artikel zu publizieren, wie dies bei universitären Instituten der Fall ist.

4.1.2.4 Publikationskanäle

In Abbildung 11 wurden die Dimensionen der Wissenschaft dargestellt. Um diese Untersuchung abzuschließen, werden noch die Publikationskanäle aufgezählt, die von den Autoren verwendet wurden, um ihre wissenschaftliche Arbeit der Forschungsgemeinschaft zugänglich zu machen. Wissenschaftliche Arbeiten können nicht nur in sogenannten Journals oder in Form eines Buches publiziert werden, sondern auf Konferenzen in Form eines Vortrags vor Fachpublikum gehalten werden. Diese Vorträge werden gesammelt in einem Tagungsband (engl. Proceedings Paper) publiziert. Abbildung 16 zeigt die Statistik für die verwendeten Publikationskanäle.

[42] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

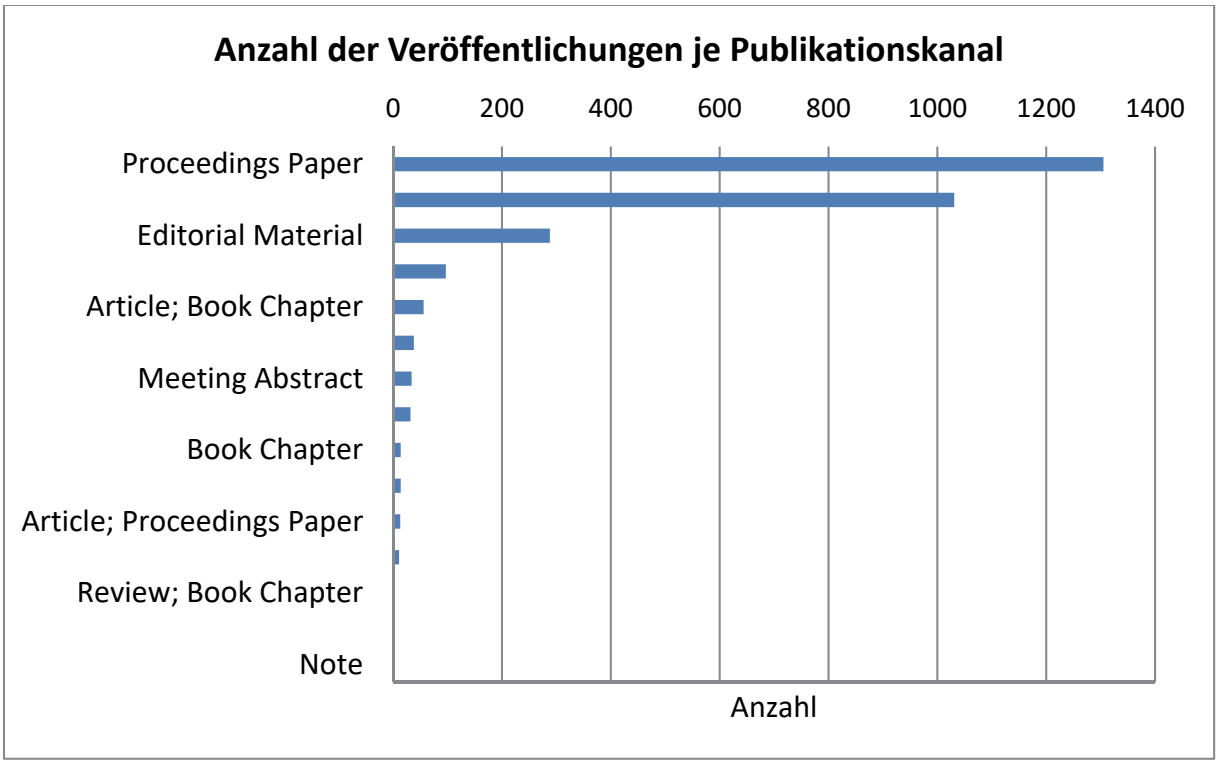


Abbildung 16: Anzahl der Artikel, sortiert nach dem Publikationskanal[43]

Es fällt auf, dass nicht wie erwartet die sogenannten „Articles“, also in Fachzeitschriften veröffentlichte Artikel, sondern die Tagungsbände die Mehrheit der Publikationen darstellen. Diese Proceedings Paper machen 44,4 % aller analysierten Publikationen aus, während an zweiter Stelle mit 35 % die üblichen wissenschaftlichen Artikel in Fachzeitschriften zu finden sind. Dies resultiert aus der Tatsache, dass *Big Data* ein recht junges Forschungsgebiet ist und dadurch die Veröffentlichungen voraussichtlich zeitverzögert publiziert werden. Zusätzlich wurden Konferenzen mit fachfremden Rednern einberufen, um die Interdisziplinarität des Themas zu betonen. Dadurch wurde ein schnellerer, direkter Austausch gewährleistet. Der Anteil der Review Artikel fällt sehr gering aus, denn diese Artikel sind sehr umfangreich und werden als Zusammenfassung eines wissenschaftlichen Schwerpunkts geschrieben. Damit bieten sie aber für Interessierte eine gute Möglichkeit, sich relativ schnell in ein neues Forschungsthema einzuarbeiten.

4.2 Netzwerkanalysen

Im Folgenden werden Autoren, Organisationen, Fachrichtungen und Schlagwörter mit Hilfe von Netzwerken detaillierter analysiert. Hierbei geht es um die Identifizierung der zum einen wichtigsten und zum anderen auffälligsten Zusammenhänge innerhalb der vier Kategorien. Dabei werden besonders die Verknüpfungen unterschiedlicher Netzwerke über gemeinsame Publikationen aufgezeigt.

[43] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

4.2.1 Netzwerk der Autoren

Zusammenfassung

Die wichtigsten Autoren Associate Prof. Jinjun Chen, Yan Li, Prof. Yi-Ke Guo und Prof. Volker Markl, zeichnen sich durch ein breites Netzwerk mit entsprechend 39, 39, 19 und 12 Knoten aus. Darüber hinaus weisen eine Vielzahl von kleineren Netzwerken untereinander Verknüpfungen durch Forschungsaktivitäten auf. Zusätzlich wird die Vielschichtigkeit innerhalb des Themenfeldes *Big Data* durch die Abstände der einzelnen Knoten und gesamten Komponenten zueinander verdeutlicht. Damit bestätigt sich die bisher getroffene Vermutung, dass *Big Data* ein interdisziplinäres Forschungsfeld ist.

Dies wird auch durch die Breite der Forschungsthemen *MapReduce* Optimierung, Analytiktechnologien für die translationale Medizin und neuartige Informationssysteme für Wirtschaftsunternehmen unterstrichen.

Das Autorennetzwerk dient zur Analyse sowie Hervorhebung der wichtigsten Autoren und deren brancheninternen sowie branchenübergreifenden Kooperationen im Zusammenhang mit dem Themenbereich *Big Data*. Das Ziel ist hierbei, basierend auf der Anzahl der Autorenveröffentlichungen sowie des Vernetzungsgrades, große Gruppierungen bzw. Teilnetzwerke (Komponenten) zu erkennen und dadurch die Themen- und Forschungsschwerpunkte einzelner Autoren oder der Autorengruppen zu ergründen. Der Hirschindex, welcher sich auf die Zitationen der Publikationen des jeweiligen Autors bezieht, hilft uns darüber hinaus ebenfalls bedeutende Autoren hervorzuheben. Das heißt je höher der h-Index, desto öfter wurden die Publikationen des Autors in anderen Veröffentlichungen zitiert und impliziert somit einen hohen Bedeutungsgrad.[44]

[44] Vgl. KIT (2014)

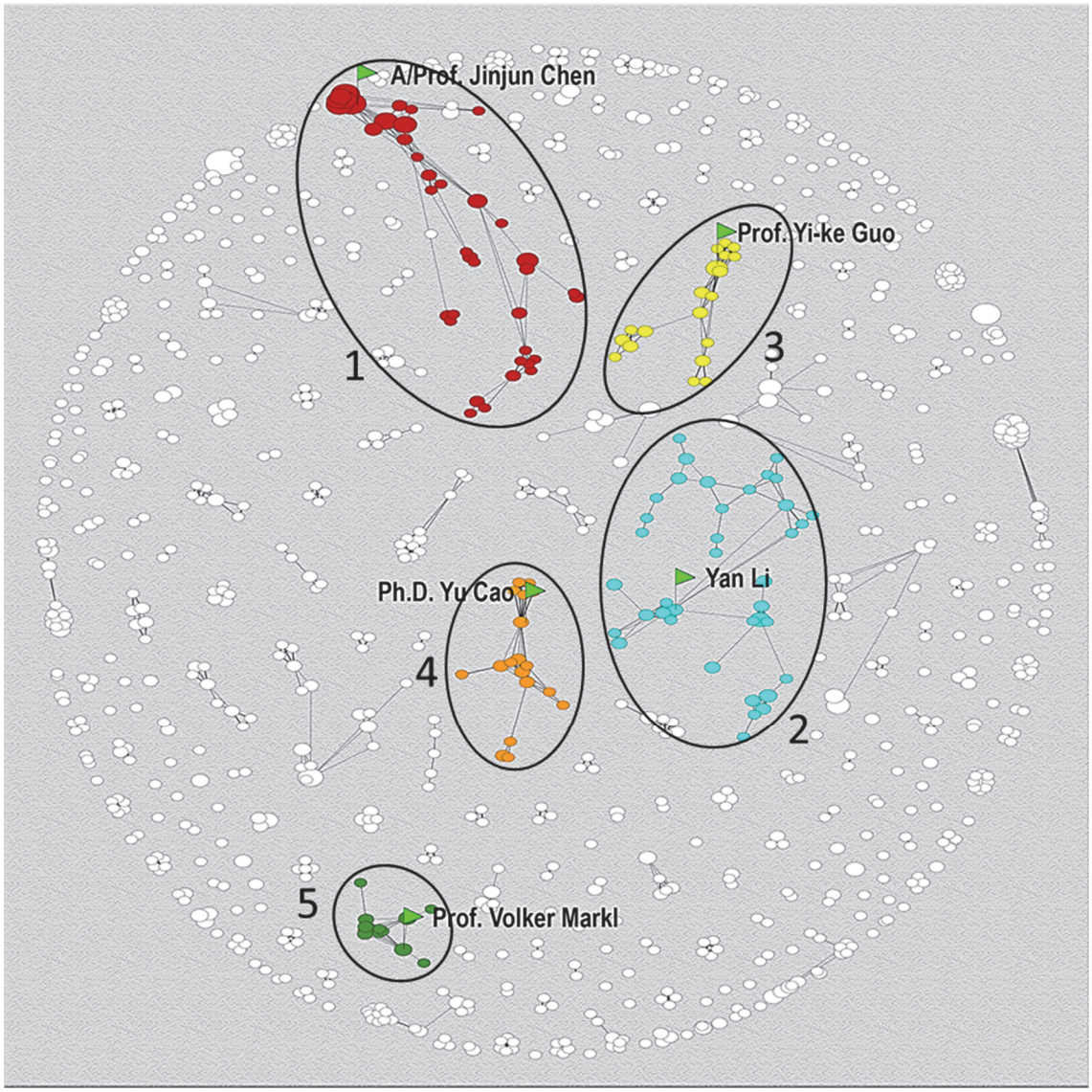


Abbildung 17: Autorennetzwerk in Bezug auf das Schlagwort Big Data[45]

Das Netzwerk der Autoren setzt sich aus vielen Komponenten sowie einzelner und unabhängiger Knoten zusammen. Eine Komponente besteht hierbei aus mindestens zwei Knoten, wobei jeder Knoten mit mindestens einem anderem verbunden ist. Das in Abbildung 17 vorgestellte Autorennetzwerk besteht aus 992 Knoten, die jeweils einen Autor aus dem Datensatz des Web of Science repräsentieren. Es wurden folgende fünf Hauptkomponenten farbig hervorgehoben. Die rote Komponente um Associate Prof. Jinjun Chen (Nr. 1), die türkise Komponente um Yan Li (Nr. 2), die gelbe Komponente um Prof. Yi-ke Guo (Nr. 3), die orangene Komponente um PhD Yu Cao (Nr. 4) und die grüne Komponente um Prof. Volker Markl (Nr. 5). Diese besitzen jeweils die größten Teilnetzwerke und haben mitunter auch die meisten Veröffentlichungen bezüglich *Big Data* herausgebracht. Die größte Komponente bildet dabei das Teilnetzwerk des chinesischen Autors Prof. Jinjun Chen, welches detailliert in Abbildung 18 zu sehen ist.

[45] Vgl. BibTechMon4 i. V .m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

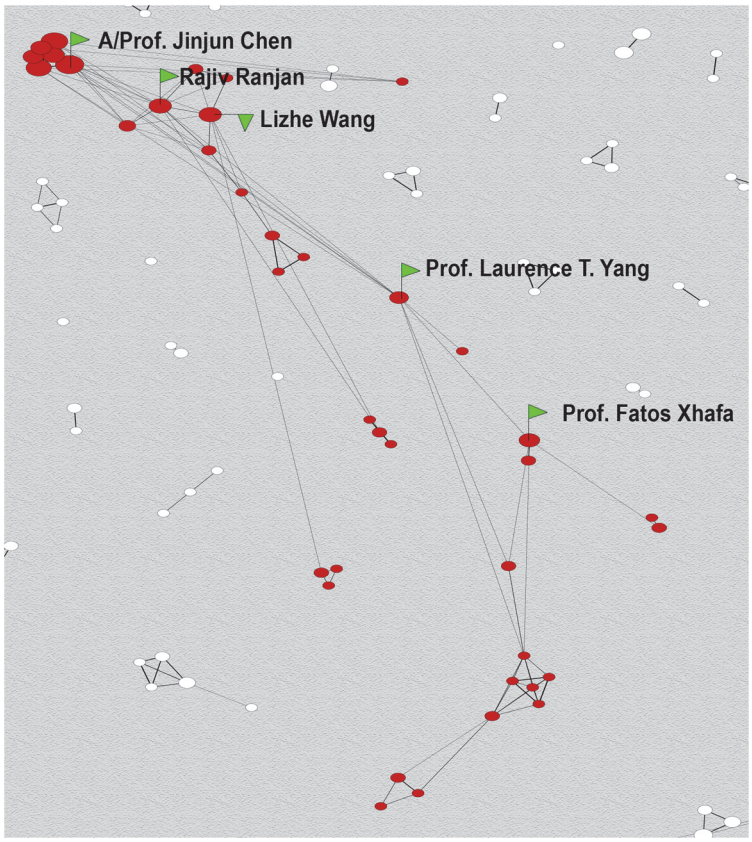


Abbildung 18: Komponente um Prof. Jinjun Chen (Nr. 1)[46]

Prof. Jinjun Chen ist Associate Professor der Faculty of Engineering and IT, sowie der leitende Direktor des "Lab for Cloud Computing and Distributed Systems" an der University of Technology Sidney (UTS) in Australien.[47] Zu Prof. Chen gehört nicht nur das größte Teilnetzwerk in Abbildung 17, er zählt darüber hinaus auch zu den Autoren mit den meisten Veröffentlichungen. Sein Netzwerk besteht aus 38 Autoren, in welchem Prof. Chen eine wesentliche Rolle spielt und dadurch eine hohe Zentralität aufweist. Schwerpunktthema dieses Netzwerks ist dabei, die Verbesserung der *Hadoop* basierten *MapReduce* Methoden anhand von funktionalen sowie sicherheitsspezifischen Erweiterungen. Prof. Chen weist darüber hinaus einen h-index von 26 auf, welches erneut seine bedeutsame Rolle in diesem Netzwerk unterstreicht.[48]

Das zweitgrößte Teilnetzwerk besteht ebenfalls aus 39 Knoten, wobei der Autor Yan Li den Kern dieser Komponente bildet. Aufgrund der mehrfachen Namensbesetzung „Yan Li"[49] wurden mehrere Veröffentlichungen in Verbindung zu unterschiedlichen Fachbereichen in der Datenbank des Web of Science hinterlegt. Nach genauer Recherche beziehen sich viele Artikel auf den Autor Jian Li, welcher Leiter eines Forschungsprogramms an Central Research Institute, Huawei Technologies Inc ist. Li arbeitet in diesem Institut an einer neuen Generation von Datenzentren und Analytik-Technologien sowie Strategien in starker Zusammenarbeit mit Teams aus China, Indien, Kanada, Russland und den USA. Der Schwerpunkt seiner Forschung liegt dabei auf *Big Data* & Analytics Plattformen und Lösungsansätze wie Apache *Hadoop* oder SPARK.

[46] Vgl. BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[47] Vgl. IEEE (2013)
[48] Vgl. Google Scholar (2015a)
[49] Alternativ auch „Jian Li"

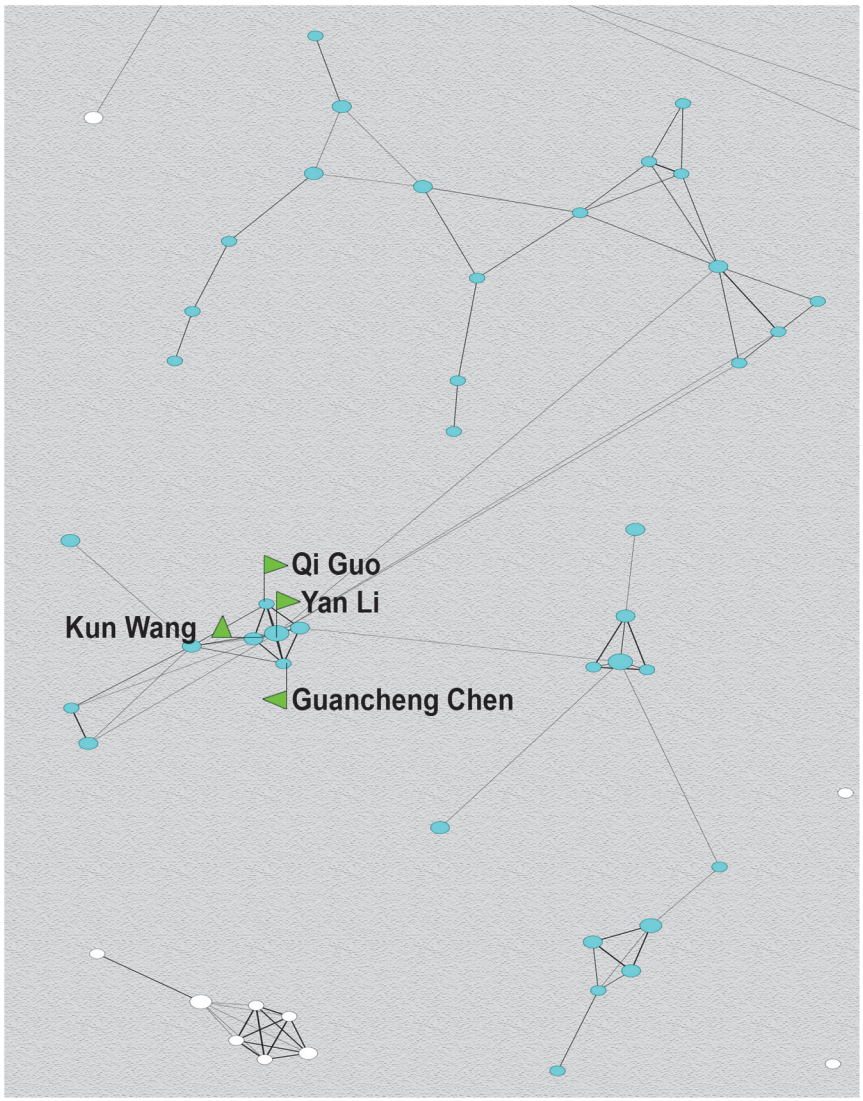


Abbildung 19: Komponente um Yan Li (Nr. 2)[50]

Kernthemen seiner Veröffentlichungen sind jedoch die Einführung eines universellen Systems, welches die verschiedenen Datenvarianten vereinheitlichen soll sowie eine Performanceoptimierung von *MapReduce* Anwendungen. Jian Li weist zudem eine starke Kooperation mit Qi Guo, Kun Wang und Guancheng Chen aus. Dies wird ebenfalls anhand der Kooperationsdichte von ca. 0,45 verdeutlicht. Der Schwerpunkt der Autoren bezieht sich wie bei Yan Li auch auf die Performanceoptimierung von *MapReduce* Anwendungen. Möglich war dies durch die gemeinsame Zusammenarbeit am IBM Research Laboratory in Beijing, China. Darüber hinaus werden die Publikationen von Jian Li auch verstärkt in den Bereichen der Informatik, Elektrotechnik sowie Telekommunikation wieder aufgegriffen. Des Weiteren Besitz Jian Li ein Hirschindex von 13, welches ihm in seinem Forschungsgebiet eine relativ moderate Bedeutung zuweist.[51]

[50] Vgl. BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[51] Vgl. Google Scholar (2015c)

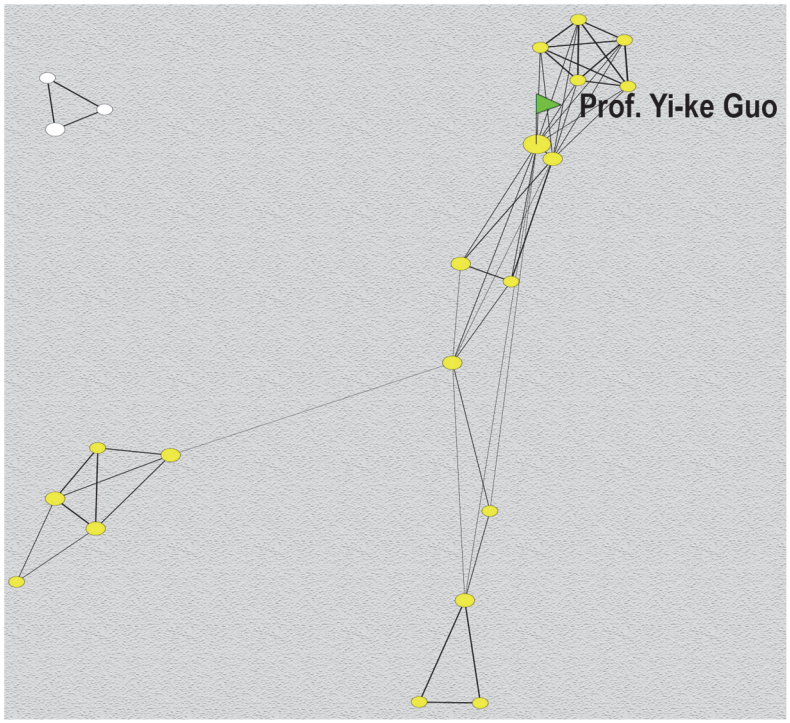


Abbildung 20: Komponente um Prof. Yi-Ke Guo (Nr. 3)[52]

Die Netzwerkkomponente um den Autor Prof. Yi-Ke Guo ist in Abbildung 20 zu sehen. Yi-Ke Guo ist Professor am Department of Computing des Imperial College in London sowie Gründer des Data Science Instituts im selbigen College.[53] Seine Forschungsgebiete sind das *Computing Science* und der Datenaustausch sowie Analytiktechnologien für die translationale Medizin. Prof. Guo bildet somit eine Schnittstelle zwischen den Bereichen der Informatik und der Medizin. Seine Veröffentlichungen finden ebenfalls großen Anklang für viele weitere Forschungen auch in den Gebieten der Biochemie bzw. Biotechnologie. Die Publikationen Prof. Guos fokussieren sich auf das Datenmanagement von Cloud basierten Plattformen, welche die von den Gesundheitssensoren anfallenden Messdaten sammeln, oder auch auf einen neuen Ansatz bzw. ein Rahmenkonstrukt, die für einen elastischen Datengewinnungsalgorithmus dienen. Prof. Guo besitzt mit einem Hirschindex von 51 im Vergleich zu vielen hier aufgeführten Autoren den höchsten Indexwert, welches seine wichtige Rolle erneut hervorhebt.[54]

[52] Vgl. BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[53] Imperial College London (2015)
[54] Vgl. Google Scholar (2015b)

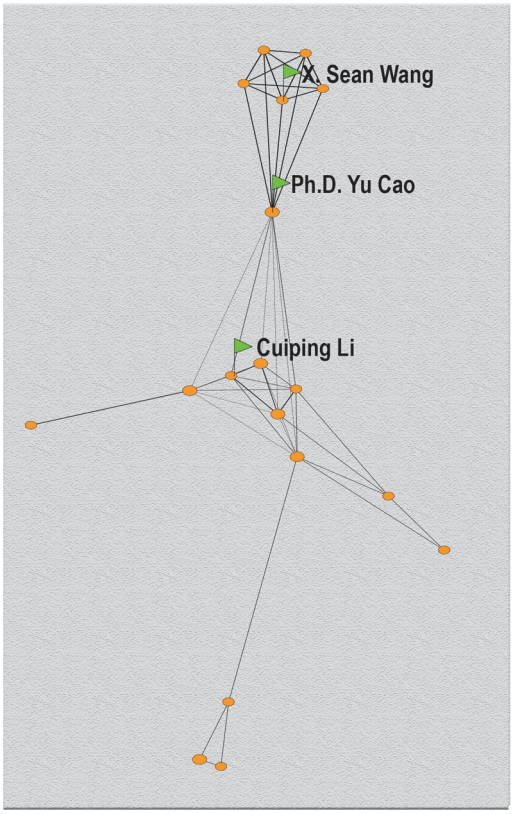


Abbildung 21: Komponente um PhD Yu Cao (Nr. 4)[55]

Abbildung 21 zeigt uns die Teilkomponente um den Autor Yu Cao. Sie besteht aus 38 Knoten und bildet daher die viert größte Komponente.

Yu Cao ist technischer Direktor am EMC Laboratory, an diesem arbeitet er als Data Scientist, Solution Architect und technischer Leiter für kundenspezifische *Big Data* Lösungen (vgl. Strata+*Hadoop* World). Er bildet ebenfalls ein Bindeglied zwischen der School of Computer Science, Fudan University, Shanghai, China und der Renmin University China, Beijing, China, mit denen er gemeinsam einige Publikationen herausgebracht hat. Kernthemen seiner Forschung bilden die Kostenreduzierung der *Hadoop* basierten Algorithmen sowie die Optimierung von Auslastungsanfragen der *Hadoop*-Datenbanken. Aufgrund fehlender Informationen konnte leider der Hirschindex von Yu Cao nicht identifiziert werden.

Die letzte Komponente besteht aus dem Autorennetzwerk um den Prof. Volker Markl. Herr Markl ist Professor und Vorsitzender der Database Systems and Information Management (DIMA) Group an der Technischen Universität Berlin (TU Berlin). Er ist des Weiteren Sprecher des Data Analytics und Cloud Labs der TU Berlin, Sprecher der Forschergruppe Stratosphere der Deutschen Forschungsgemeinschaft (DFG) und Direktor des Berlin Big Data Center, welches die wissenschaftlichen Disziplinen des skalierbaren Datenmanagements und des maschinellen Lernens verbindet.[56] Stratosphere ist dabei ein weltweit führendes Open-Source-System zur skalierbaren Datenanalyse und bildet dadurch eine wichtige Grundlage zum Thema *Big Data*.

[55] Vgl. BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[56] Vgl. TU Berlin (2013)

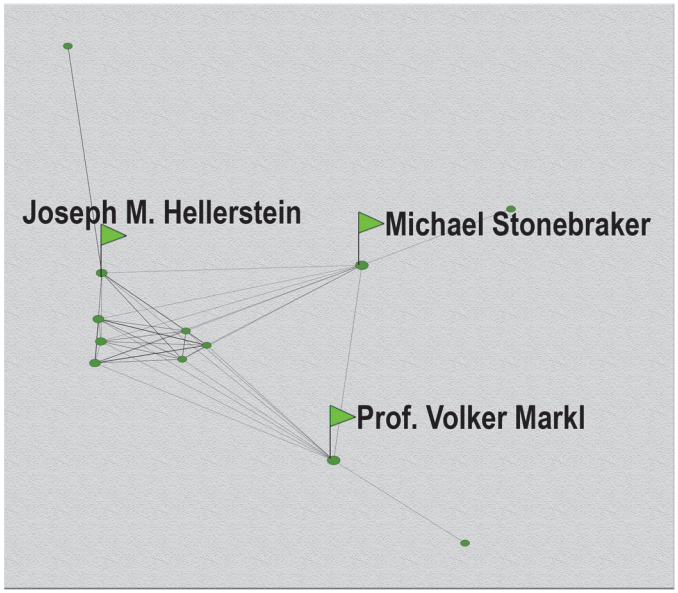


Abbildung 22: Komponente um Prof. Volker Markl (Nr. 5)[57]

Wie in Abbildung 22 zu erkennen ist, hat Prof. Markl eine verknüpfende Rolle in der Komponente. Seine Themengebiete handeln von Forschungsinformationssystemen, die die anfallenden *Big Data* Informationen fördern und die dadurch gewonnen Erfahrungen für das Beschreiben und Gestalten neuer und innovativer Informationssysteme in der Wirtschaft und Verwaltung anwenden. Prof. Markl kooperiert bei seinen Veröffentlichungen mit Universitäten wie der Technischen Universität München oder der Universität Münster in Bereichen der Informatik und Telekommunikation. Prof. Markl besitzt darüber hinaus ein h-index von 25 und zählt somit zu den beachtungswürdigen Autoren seiner Fachrichtung.[58]

Das Autorennetzwerk aus Abbildung 17 verdeutlich nochmals die große Anzahl an Autoren sowie Veröffentlichungen in Bezug auf den Begriff *Big Data*. Die Schwierigkeit liegt hierbei jedoch darin, den Datensatz kritisch zu hinterfragen und evtl. außer Acht gelassenen Autoren, die Vielfältigkeit des Begriffs *Big Data* und die damit in Verbindung stehende Suche nach richtigen Schlüsselbegriffen mit einzubeziehen. Wichtige Autoren im Bereich *Big Data* sind beispielsweise Jeffrey Dean und Doug Cutting, die die Gründung bzw. Weiterentwicklung von *Big Data* erst ermöglicht haben. Dean und Cutting sind jedoch nicht auf Anhieb in unserem Netzwerk zu finden, weshalb im folgenden Abschnitt diese Autoren in kürze vorgestellt werden.

Jeffrey Dean ist aktuell Google Senior Fellow in der Knowledge Group. Er hat seinen PhD im Bereich der Computer Science an der Universtiy of Washington erlangt und ist mit Sanjay Ghemawat der Mitentwickler des *MapReduce* Programmiermodells.[59] *MapReduce* ist ein Programmiermodell mit dem sich Software zur Verarbeitung großer Mengen unstrukturierter Daten (*Big Data*) entwickeln lässt, die entweder parallel auf einem Rechner-Cluster oder auf einzelnen Computern läuft (vgl. Kapitel 4.1.2).[60]

[57] Vgl. BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[58] Vgl. Google Scholar (2015d)
[59] Vgl. Google Research (2015)
[60] Vgl. Rouse (2014)

Dean legte somit den Grundstein für die Verarbeitung sowie Auswertung von riesigen Datenmengen innerhalb kürzester Zeit. Die im Jahre 2004 veröffentlichte Publikation „Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters“, setze dadurch neue Maßstäbe.

Ein weiterer bedeutsamer Entwickler und Autor ist Doug Cutting. Cutting ist Befürworter und Gründer von Open-Source Search Technologien. Neben den beiden Open-Source Search Technologien wie Lucene und Nutch ist er der Gründer des *Hadoop* Frameworks. *Hadoop* ist ein *MapReduce* basierter Software-Framework, mit dessen Hilfe die Nutzer rechenintensive Prozesse mit hohem Datenvolumen auf Server-Clustern abarbeiten und somit Anwendungen mit komplexen Computing-Aufgaben auf Tausende von Rechnerknoten verteilen um dadurch Datenvolumina im Petabyte-Bereich verarbeiten zu können.[61] Es stellt sozusagen die Implementierung des *MapReduce*-Modells als Open Source Anwendung dar.

4.2.2 Netzwerk der Organisationen

In diesem Abschnitt sollen die verschiedenen Organisationen, die hinter den analysierten Publikationen stehen, näher beleuchtet und charakterisiert werden. Mit den erhaltenen Netzwerken sollen so Rückschlüsse darauf gezogen werden, ob und wie stark einzelne / verschiedene Organisationen bei ihren Veröffentlichungen zusammenarbeiten und ob es eventuell Kooperationen gibt. Des Weiteren lassen sich durch die wissenschaftliche Ausrichtung der einzelnen Organisationen Informationen gewinnen, in welchen Forschungsgebieten *Big Data* eine Rolle spielt. Um die Publikationen zu kategorisieren, wird mithilfe von BibTechMon4 die Co-Publikation zweier (oder mehrerer) Organisationen ermittelt.[62] Abbildung 23 zeigt das erhaltene Netzwerk der Organisationen mit den größten Clustern farbig markiert und benannt nach dem bedeutendsten Institut.

[61] Vgl. Computerwoche (2012)

[62] Die Größe der Knoten gibt somit an, wie oft eine Organisation publiziert hat. Die Dicke der Verbindungslinien zwischen zwei Knoten gibt an wie oft zwei Organisationen zusammen publizieren. Man erhält so, durch geeignete Wahl der einstellbaren Parameter (einer Art Empfindlichkeit), eine Landkarte (Netzwerk) mit im Idealfall großen Clustern, die auf eine starke Verbindung der jeweiligen Organisationen hindeutet.

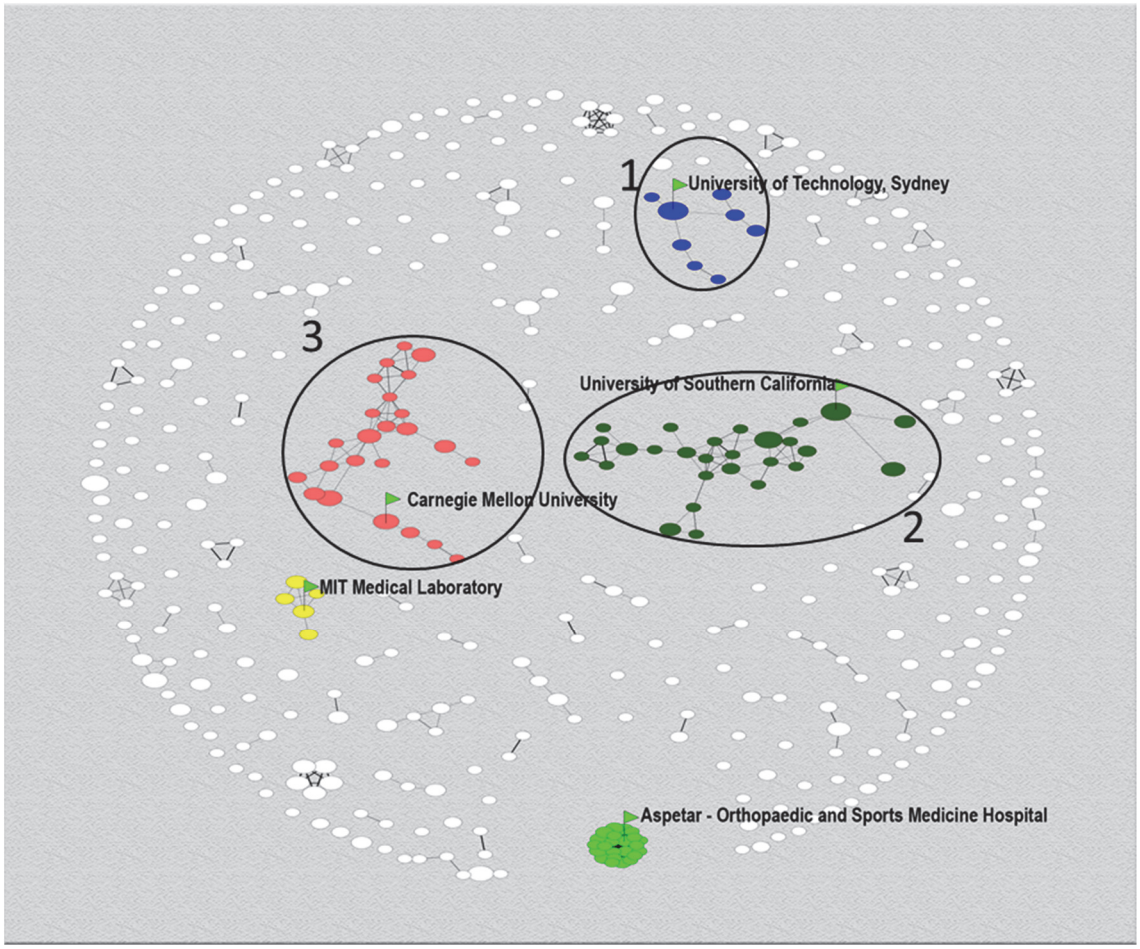


Abbildung 23: Netzwerk der Organisationen mit den fünf größten Clustern[63]

In Abbildung 23 können die einzelnen Organisationen klar voneinander abgegrenzt werden. Dies deutet darauf hin, dass keinerlei Kooperationen zwischen den Organisationen im Bereich der Publikationen stattfinden. Die fehlenden Kooperationen resultieren darin, dass die Organisationen in Abbildung 23 an den Rand des Bildes gedrängt werden. Vereinzelt gibt es einige Organisationen, die sehr nahe beieinanderliegen und teils symmetrisch angeordnet sind.[64] Die wichtigsten drei Organisationen sind in Abbildung 23 farbig markiert und beziffert. Detailliert beschrieben werden die drei universitären Komponenten mit der blauen Komponente um die University of Technology in Sidney, Australien (Nr. 1), der dunkelgrünen Komponente um die University of Southern California in den Vereinigten Staaten (Nr. 2) und der roten Komponente um die Carnegie Mellon University in Pittsburgh, ebenfalls in den Vereinigten Staaten (Nr. 3).

[63] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015;
438 einzelne Organisationen; MIT Medical Laboratory des MIT in Cambridge, USA und Aspetar - Orthopaedic and Sports Medicine Hospital in Doka, Katar haben thematisch keinerlei gewinnbringende Beziehung zum Thema *Industrie 4.0*, daher werden diese Organisationen nicht weiter berücksichtigt.

[64] Web of Science unterscheidet zwischen der Bezeichnung „Institut“ und „Inst.“, was dazu führt, dass Doppelnennungen nicht beachtet werden. Daraus resultiert eine höhere Anzahl von einzelnen Knoten, als real vorhanden.

Teilnetzwerk 1 – University of Technology, Sidney

Das erste Teilnetzwerk ist das der University of Technology in Sidney, Australien (siehe Abbildung 23, Nr. 1). Abbildung 24 zeigt das Teilnetzwerk um dieses Institut und man sieht, dass es im Vergleich zu den anderen Teilnetzwerken, ein sehr kleines ist, mit insgesamt acht Organisationen.

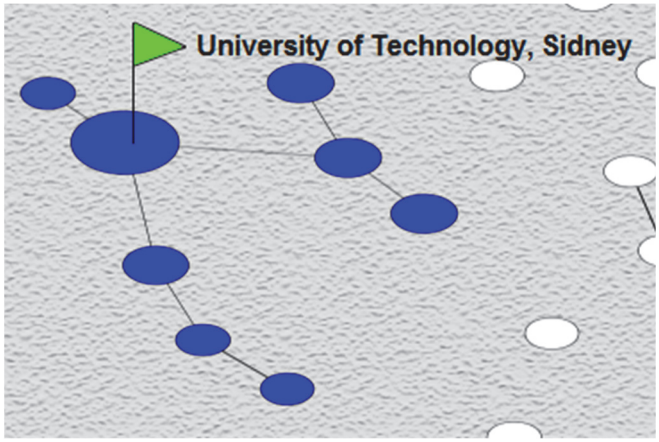


Abbildung 24: Teilkomponente um die University of Technology in Sidney, Australien (Nr. 1)[65]

In Abbildung 24 ist zu erkennen, dass die University of Technology in Sidney direkt mit drei anderen Einrichtungen und indirekt mit weiteren vier Einrichtungen verknüpft.[66] Neben der University of Technology in Sidney sind das Institute for Remote Sensing & Digital Earth der Chinese Academy of Sience in Beijing, China, sowie die Huazhong University of Science & Technology, School of Computational Science & Technology in Wuhan, China an dem Teilnetzwerk beteiligt. Das primäre Forschungsgebiet in diesem Netzwerk ist die Entwicklung verschiedener *Big Data* Frameworks mit den Zielen, größte Datenmengen auf verteilten Clustern in der Cloud zu speichern und dabei die Anonymität zu wahren.[67]

Prof. Jinjun Chen von der Fakultät für Ingenieurswesen und IT der University of Technology in Sidney ist hierbei führend in der oben genannten Forschung und er hat an nahezu allen Publikationen mitgearbeitet, die im Teilnetzwerk in Abbildung 24 dargestellt sind.[68] Es zeigt sich, dass die oben genannte Chinese Academy of Science in Beijing, China ihren Teil leistet, indem dieses Institut für die Entwicklung von speziellen Algorithmen und Plattformen zuständig ist, die Grundlage sind, für die Arbeit, die Prof. Jinjun Chen in Sidney leistet.[69] Man sieht hier, dass dieses Teilnetzwerk mit ihren Organisationen weltweit vernetzt ist und hier scheinbar eine Nachhaltige Kooperation stattfindet. Bei einem solch komplexen Forschungsgebiet ist dies auch zwingend notwendig, da für die Entwicklung vollständiger Systeme im Bereich *Big Data* und Cloud Computing stets eine Vielfalt an Expertenwissen vonnöten ist. Der Hauptakteur Herr Prof. Jinjun Chen wird im Kapitel zu den Autorennetzwerken noch eine zentrale Rolle einnehmen.

[65] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Komponente mit 19 Publikationen
[66] Die Stärke der Kanten gibt an, dass die Forschungsaktivitäten nicht sehr stark sind.
[67] Vgl. Zhang et al. (2013)
[68] Vgl. CGC (2013)
[69] Vgl. Zhao et al. (2014)

Teilnetzwerk 2 – University of Southern California, Los Angeles

Das nächste Teilnetzwerk um die University of Southern California (USC) in Los Angeles, USA (siehe Abbildung 23, Nr. 2) ist deutlich größer als die Komponente in Abbildung 24. In Abbildung 25 lässt sich erkennen, dass Organisationen aus allen Teilen der Welt mit der USC verknüpft sind, der überwiegende Teil dennoch Organisationen aus den Vereinigten Staaten sind.

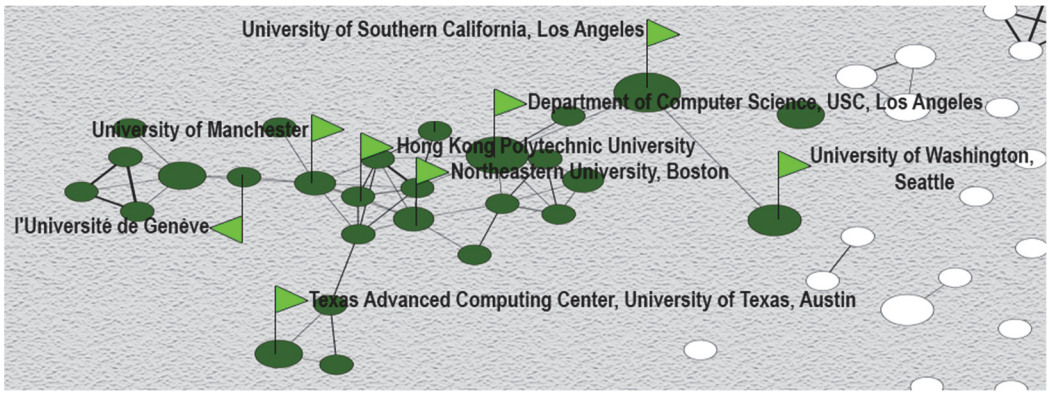


Abbildung 25: Teilkomponente um die University of Southern California (USC) in Los Angeles, USA (Nr. 2)[70]

Man sieht, dass vom Zentrum des Netzwerks Organisationsarme abgehen, die somit thematisch zusammenhängen und auf dem gleichen Forschungsgebiet im Zusammenhang mit *Big Data* arbeiten.

Die USC in Los Angeles beschäftigt sich im Bereich *Big Data* hauptsächlich mit sogenannten Smart Grids, also der intelligenten Steuerung der Stromversorgung bzw. generell der Intelligenz von Netzwerken.[71] Hierbei haben vor allem *Big Data* Anwendungen und deren Erforschung in diesem Zusammenhang großes Potential, da riesige Mengen an Daten, aus unterschiedlichen Quellen in unterschiedlichen Formaten schnell verarbeitet und in Anweisungen überführt werden müssen.

Prof. Viktor Prasanna ist Leiter des Instituts für „Electrical Engineering and Computer Science“ an der USC und verantwortlich für die Forschung im *Big Data* Bereich. Insbesondere beschäftigt er sich mit Hochleistungscomputerberechnungen in parallel bzw. verteilten Systemen.

Die University of Washington in Seattle, USA wird in Abbildung 25 durch die Fahne ganz rechts angezeigt und ist nur eine relativ schwach gebundene Organisation. Das Institut für Electrical Engineering dieser Universität beschäftigt sich nur im weitesten Sinne mit *Big Data*, da es um die Erforschung von Kommunikationschips für Sensoren geht. In diesem Bereich kommen auch *Big Data* Anwendungen zum Tragen und im Kontext von *Industrie 4.0* ist dieser Forschungszweig essentiell, da nur durch die Kommunikation von sehr vielen Sensoren, z.B. in Produktionsanlagen, eine Auswertung dieser Daten im Sinne von *Industrie 4.0* möglich ist.[72]

Die University of Texas at Austin in Texas, USA ist ebenfalls ein kleineres Teilorganisationsnetzwerk und wird durch die Fahne unten links in Abbildung 25 angezeigt. Ganz konkret geht es um das TACC, das Texas Advanced Computing Center, das eine Unterabteilung besitzt, die sich mit Data Mining und Statistics beschäftigt. Dieses Forschungsfeld gehört auch zum Themenbereich *Big Data*, aber nicht im

[70] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Komponente mit 46 Publikationen
[71] Vgl. Simmhan (2013)
[72] Vgl. Rudell et al. (2014)

engeren Sinne. Der Gruppenleiter Dr. Weijia Xu erforscht grundlegende analytische Methoden, um große Datenmengen zu analysieren.[73] Seine Forschung steht somit eine Stufe vor den *Big Data* Anwendungen, wie *MapReduce* oder *Hadoop*. Nichts desto trotz finden sich auch Forschungsprojekte, bei denen klare *Big Data* Algorithmen zum Tragen kommen.

Die Universität Genf ist ein Verbindungsglied zwischen der USC als Zentrum des Netzwerks und einem Teilnetzwerk ganz links in Abbildung 25. Das dort ansässige Institut für Informatik (Centre Universitaire d'Informatique), geleitet von Prof. Stéphane Marchand-Maillet, befasst sich mit den Themen *Machine Learning* und *Data Mining*. Diese grundlegenden mathematischen und informatischen Prinzipien sind essentiell für den Anspruch der industriellen Vernetzung wie im Rahmen der *Industrie 4.0* Definition bereits erwähnt (vgl. Kapitel 2). So befasst sich *Data Mining* mit der Erfassung und Kategorisierung von großen Datenmengen, während *Machine Learning* zum Ziel hat, Maschinen soweit intelligent zu programmieren, dass sie eigenständig Handlungen durchführen und bewerten können. Das kann z. B. durch den Einsatz von Sensoren geschehen und über geeignete Algorithmen, die es der Maschine erlaubt auf bestimmte Situationen zu reagieren und dieses Verhalten zu „lernen". Damit ist auch dieser Zweig des Netzwerks um die USC wie schon die University of Washington in Seattle ein sehr anwendungsnahes Institut. Die theoretischen Rahmenbedingungen werden also durch Kooperationen mit der USC geschaffen.

Der Organisationszweig in der Mitte von Abbildung 25 beinhaltet die University Manchester, die Hong Kong Polytechnic University, sowie die Northeastern University in Boston. Diese Universitäten forschen alle im Bereich *Big Data*, jedoch in relativ unterschiedlichen Gebieten. Diese reichen von *Cloud Computing* bis hin zur Anwendung von *Big Data* Algorithmen in der Meteorologie. Aufgrund dieser Vielfalt soll hier nicht näher auf diesen Netzwerkzweig eingegangen werden. Jedoch zeigt das Netzwerk in Abbildung 25, dass eine relativ starke Verbindung zum theoretischen Grundlagenwissen an der USC besteht. Somit ist für viele der angrenzenden verschiedenen Bereiche der Wissenschaft die Grundlagenforschung essentiell.

Teilnetzwerk 3 – Carnegie Mellon University, Pittsburgh

Als drittes und letztes Teilnetzwerk soll ein Netzwerk untersucht werden, welches viele bekannte Eliteuniversitäten der Vereinigeten Staaten beherbergt (siehe Abbildung 23, Nr. 3). Abbildung 26 zeigt dieses Teilnetzwerk mit den wichtigsten Akteuren im Detail, sowie ihre Verbindungen durch gemeinsame Publikationen.[74]

[73] Vgl. IEEE (2013)

[74] Stärke der Verbindungslinie (Kante) zwischen zwei Knoten gibt die Intensität der Forschungsaktivitäten an. Je dicker die Kante, desto intensiver die Zusammenarbeit. Je größer der einzelne Knoten, desto höher ist die Forschungsaktivität.

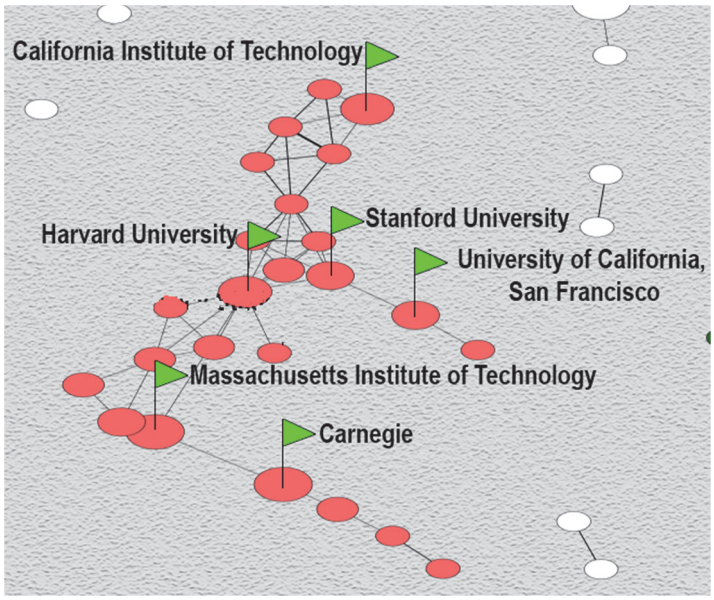


Abbildung 26: Teilkomponente um die Carnegie Mellon University in Pittsburgh, USA (Nr. 3)[75]

Die bekannte Harvard University in Cambridge in den Vereinigten Staaten hat mit der Forschung im Bereich *Big Data* und insbesondere im Kontext von *Industrie 4.0* keine große Bedeutung. Die bibliometrische Analyse in Abbildung 26 zeigt, dass die Publikationen der Harvard University, im biologischen Umfeld liegen. Das heißt, dass hier *Big Data* Anwendungen zum Einsatz kommen, um z. B. bei der Entschlüsselung der DNA im Zusammenhang mit Stammzellenforschung zu helfen.[76] Prof. Winston Hide ist Mitautor des genannten Artikels und forscht in Harvard am Department of Biostatistics. Er versucht, *Big Data* Anwendungen zu verwenden, um die Genomsequenzierung zu beschleunigen und die Ausbreitung von Krankheiten mittels Datenanalyse vorherzusagen.[77] Sehr eng damit verbunden ist auch die Stanford University in Kalifornien.[78] Wichtig ist hier das Cherry Lab der Stanford Universität, mit Prof. J. Michael Cherry als Leiter. Dieses Institut erforscht das Bereitstellen von komplexen Datenbanken, um einen schnellen Zugriff auf gespeicherte Daten im Kontext DNA und Healthcare sicherzustellen. Ähnlich wie in Harvard kommen hier *Big Data* Anwendungen zum Tragen, um die riesigen Datenbanken schnell und effizient zu durchsuchen.

Die University of California in San Francisco befindet sich im Netzwerk in Abbildung 26 auf der rechten Seite und beschäftigt sich mit Biostatistic. Neben vielen anderen Forschern auf diesem Gebiet, sei hier das Krogan Lab, benannt nach dem Institutsleiter Prof. Nevan Krogan, erwähnt, das sich mit der Erstellung und Analyse von großen „Landkarten“ der Genetik beschäftigt. Damit sollen Viren wie HIV besser verstanden und bekämpft werden können.[79]

Ebenfalls auf dem Gebiet der Biologie / Genforschung sind die Carnegie Institution of Science des Department of Plant Biology der Stanford Universität, sowie das MIT (Massachusetts Institute of Technology) tätig. Diese befassen sich mit der Genomforschung von Pflanzen und sind hier deshalb nicht weiter im Fokus.

[75] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Komponente mit 35 Publikationen
[76] Vgl. Howe (2008)
[77] Vgl. Hide (2015)
[78] Vgl. Howe (2008)
[79] Vgl. Bushman (2013)

Mehr mit *Big Data* im Kontext von *Industrie 4.0* verbunden, ist das „California Institute of Technology", kurz CALTECH. Diese Organisation ist in Abbildung 26 ganz oben zu erkennen. Das CALTECH beschäftigt sich, ähnlich wie das MIT seit jeher mit neuester Technologie in den Ingenieurs- und Naturwissenschaften. Deshalb finden sich unter den analysierten Publikationen in diesem Teilnetzwerk einige des CALTECH, die für *Big Data* im Kontext von *Industrie 4.0* von Bedeutung sind. Diese Organisationen befassen sich mit *Machine Learning*, allerdings im Umfeld von Astronomie, weshalb sie am NASA Jet Propulsion Laboratory (JPL) der CALTECH Universität angesiedelt sind. Die Gruppe um Dr. Kiri Wagstaff der Machine Learning and Instrument Autonomy (MLIA) Group des JPL beschäftigt sich mit „data mining, knowledge discovery, pattern recognition, and automated classification and clustering"[80]. Wie das Zitat über den Institutionsnamen schon sagt, wird hier im Bereich Machine Learning geforscht. In der von Dr. Kiri Wagstaff durchgeführten Arbeit wird der Frage nachgegangen, in welchen Gebieten *Big Data* Anwendungen wichtig sein könnten, so zum Beispiel in der Luftfahrt, Infrastrukturplanung oder der Charakterisierung von Menschenmassen und deren Bewegung. In der Publikation „Machine learning for science and society" aus dem Jahre 2014 stellt sie all die verschiedenen Bereiche vor, in denen *Big Data* eine Rolle spielen kann und wo *Big Data* Anwendungen Probleme lösen, die bisher als unlösbar galten.[81] Natürlich spielt hier *Machine Learning* eine große Rolle, sodass man dieses Forschungsfeld auch in anderen Teilnetzwerken und bei anderen wichtigen Forschern wiederfindet, wie z. B. Prof. Marchand-Maillet der Universität Genf.[82]

[80] Vgl. MLIA (2015)
[81] Vgl. Rudin und Wagstaff (2014)
[82] Vgl. Marchand-Maillet (2015)

In Tabelle 3 und Tabelle 4 sind die wesentlichen Erkenntnisse der Autoren und Organisationsanalyse zusammengefasst.

Nr.	Autor	Themengebiete	Institut
1	Prof. Jinjun Chen	Verbesserung der *Hadoop* basierten *MapReduce* Methoden anhand von funktionalen sowie sicherheitsspezifischen Erweiterungen	Faculty of Engineering and IT, University of Technology Sidney (UTS) in Australien Direktor des "Lab for Cloud Computing and Distributed Systems"
2	Yan Li	*Big Data* & Analytics Plattformen und Lösungsansätze wie Apache *Hadoop* oder SPARK	Leiter eines Forschungsprogramms am Central Research Institute, Huawei Technologies Inc, Shenzjen, China
3	Prof. Yi.ke Guo	*Computing Science* und der Datenaustausch sowie Analytiktechnologien für die translationale Medizin	Professor am Department of Computing des Imperial College, London
4	Ph.D. Yu Cao	Data Scientist, Solution Architect und technischer Leiter für kundenspezifische *Big Data* Lösungen	technischer Direktor am EMC Laboratory, Tsinghua Science Park, Beijing, China
5	Prof. Volker Markl	Forschungsinformationssystemen, die *Big Data* Informationen und dadurch gewonnen Erfahrungen in der Wirtschaft und Verwaltung anwenden	- Professor der Database Systems and Information Management (DIMA) Group, TU Berlin - Sprecher Data Analytics und Cloud Lab, TU Berlin - Sprecher Forschergruppe Stratosphere, DFG - Direktor des Berlin Big Data Center

Tabelle 3: Wichtigste Ergebnisse aus der Analyse der Autoren

Nr.	Institut	Autor	Themengebiete
1	University of Technology, Sidney, Australia	Prof. Jinjun Chen	Ingenieurswesen und IT; *Big Data* Frameworks zur Speicherung großer Datenmengen in Clustern in der Cloud
2	University of Southern California (USC), Los Angeles, USA	Prof. Viktor Prasanna	*Smart Grids* Hochleistungscomputerberechnungen in parallel bzw. verteilten Systemen
2	University of Washington, Seattle, USA		Electrical Engineering: Kommunikationschips für Sensoren
2	University of Texas, Austin, USA Texas Advanced Computing Center (TACC)	Dr. Weijia Xu	*Data Mining* and Statistics Grundlegende analytische Methoden, um große Datenmengen zu analysieren
3	Carnegie Mellon University, Pittsburgh, USA	Tom M. Mitchell	*Machine Learning*
3	Harvard University, Cambridge, Great Britain	Prof. Winston Hide	*Big Data* Entschlüsselungen der DNA
3	Stanford University, Stanford, USA Cherry Lab	Prof. J. Michael Cherry	Bereitstellen von komplexen Datenbanken für Biostatistics, DNA
3	University of Califonria, San Francisco, USA	Prof. Nevan Krogan	Erstellung und Analyse von großen „Landkarten“ der Genetik z. B. für HIV

Tabelle 4: Wichtigste Ergebnisse aus der Analyse der Organisationen

Netzwerk der Fachrichtungen

Ein interessanter Blick auf den Datensatz *Big Data* bietet die Analyse der Fachrichtungen. Die Analyse ermöglicht uns herauszufinden, in welchen Fachrichtungen die jeweiligen Veröffentlichungen publiziert wurden. Beispielsweise kann eine Fachrichtung die Physik oder die Informatik sein.

Wie wir in den bisherigen Kapiteln gesehen haben, ist die wissenschaftliche Auseinandersetzung mit dem Begriff *Big Data* relativ jung. Aus diesem Grund ist es interessant zu sehen, in welchen wissenschaftlichen Gebieten im Zusammenhang mit *Big Data* publiziert wird. Die Datenbank von Web of Science teilt die Wissenschaft in verschiedene Fachrichtungen ein. Unser Datensatz besteht aus 2939 Publikationen, diese können entweder in einer oder mehreren Fachrichtungen veröffentlicht worden sein. Dies bedeutet, dass zum Beispiel der Artikel „Exploring managers' intention to use business intelligence: the role of motivations“[83] in den Fachrichtungen Computer Science und Engineering veröffentlich wurde. Daraus folgt, dass die 2939 Publikationen, nachdem sie allen wissenschaftlichen Richtungen zugeordnet wurden, 4563 Einträge in den verschiedenen Fachrichtungen darstellen.

In der Datenbank Web of Science liegen 153 verschiedene Fachrichtungen vor[84], die 2939 Publikationen aus unserem Datensatz wurden in 121 Fachgebieten davon veröffentlicht. Daraus kann abgeleitet werden, dass *Big Data* nicht nur in Bereichen eine Rolle spielt, in denen man es erwartet

[83] Vgl. Chang et al. (2014)
[84] Vgl. Reuters (2012)

wie zum Beispiel der Informatik, Physik oder Telekommunikation, sondern auch in der Biotechnologie und der Medizin, wie in Kapitel 4.2.2 beschrieben.

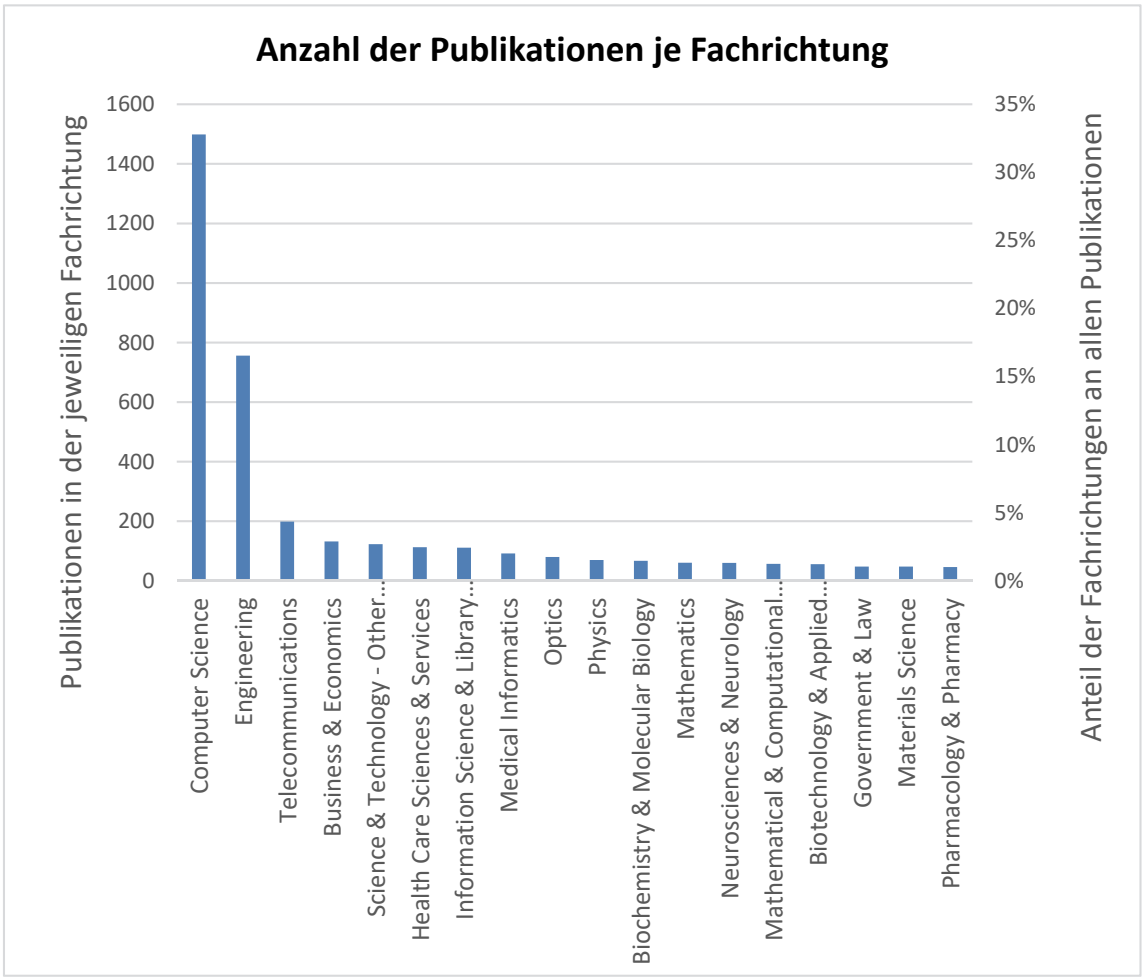


Abbildung 27: Anzahl der Publikationen je Fachrichtung und gewichteter Anteil der Fachrichtungen[85]

Wie in Abbildung 27 zu sehen ist, werden die meisten Publikationen der Fachrichtung *Computer Science* und anschließend der zweitwichtigsten Kategorie *Engineering*, zugeordnet. Darüber hinaus ist anzumerken, dass die weiteren Fachrichtungen insgesamt nicht so viele Veröffentlichungen beinhalten, wie die beiden oben genannten. Die in der Abbildung 27 aufgeführten Fachrichtungen sind diejenigen, die gewichtet jeweils über 1% am gesamten Volumen der Publikationen verfügen.

Interessante Kategorien für die *Industrie 4.0* basieren hier hauptsächlich auf den am meisten beachteten Fachrichtungen. Darunter fallen vor allem die Bereiche *Computer Science*, *Engineering*, *Telecommunications*, *Business & Economics* und *Science & Technology*.

[85] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

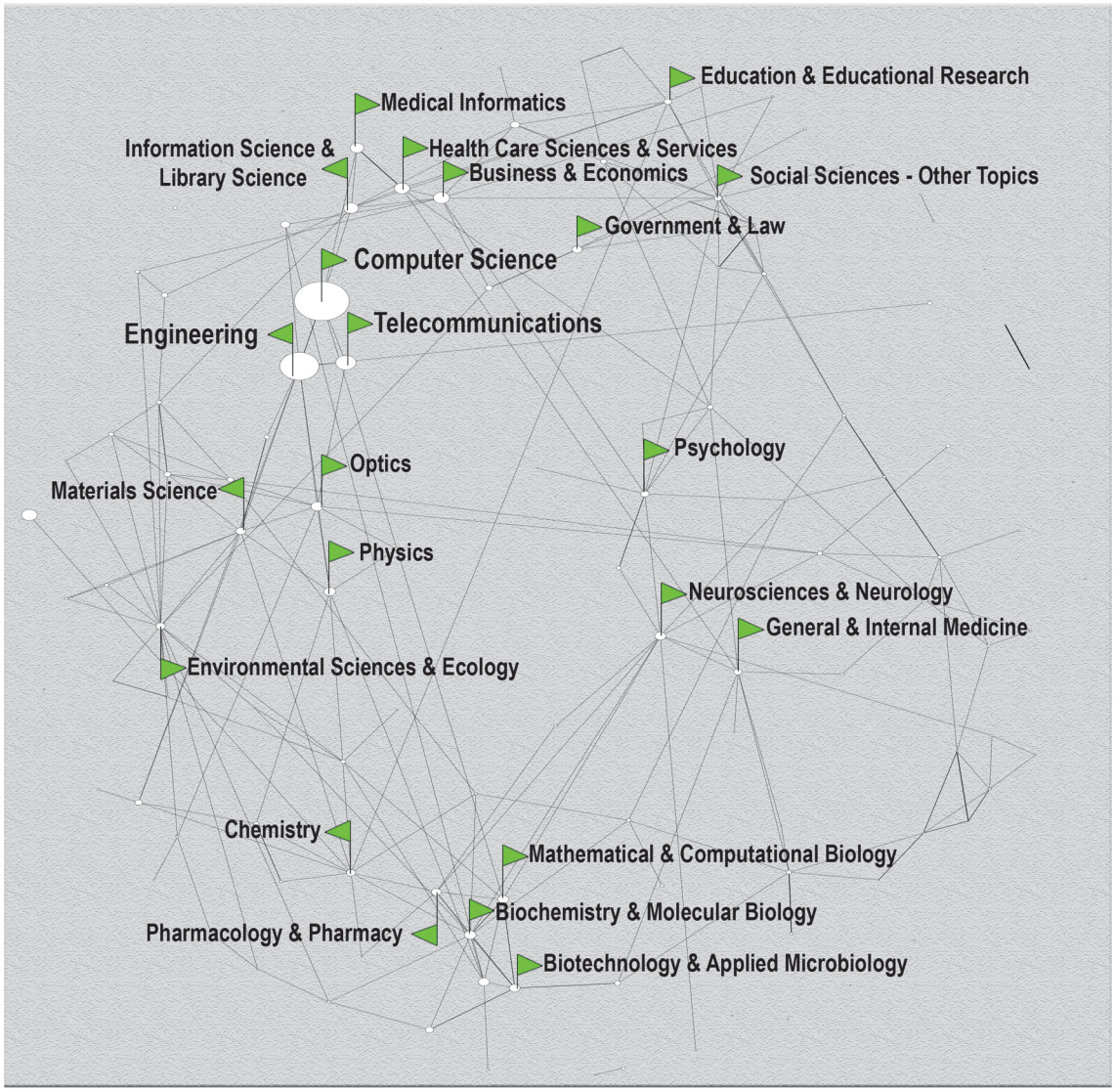


Abbildung 28: Netzwerk der Fachrichtungen[86]

In Abbildung 28 ist das Netzwerk der Fachrichtungen dargestellt. *Computer Science* ist hierbei der größte Knoten und weißt die höchste Zentralität auf. Diese Fachrichtung ist eng mit dem Bereich *Engineering* verbunden, welcher wiederum stark mit dem Forschungsgebiet der *Telecommunications* verbunden ist. Diese drei Fachrichtungen haben einen gewichteten Anteil von 53,8 % an allen Fachrichtungen und sind somit die wichtigsten Kategorien. Des Weiteren kann im Netzwerk gesehen werden, dass *Business & Economics* ein wichtiger Knoten ist, welcher mit vielen anderen Knoten eine gute Vernetzung aufweist. Das bedeutet, dass die Publikation in diesem Bereich auch andere Gebiete wie zum Beispiel *Information Science & Library Science* oder auch *Health Care Science & Services* behandeln. Naturwissenschaftliche Fachgebiete sind in den Randgebieten des Netzwerkes zu erkennen. Dazu gehören Chemie, Biotechnologie und Pharmazie. Andere Fachrichtungen rund um die Soziologie und Sozialwissenschaften sind oben rechts im Netzwerk angesiedelt, diese beinhalten auch Publikationen, die sich mit dem Marketing beschäftigen.

[86] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Netzwerk mit 112 Knoten und 309 Kanten. Die Kantendicke wurde mithilfe des Jaccard-Index berechnet.

Daraus folgt, dass die interessanteren Gebiete für die Entwicklung bezüglich der *Industrie 4.0* eher im oberen Teil des Netzwerkes angesiedelt sind. Diese Fachrichtungen sind sehr hoch vernetzt und zeigen somit, dass *Big-Data-Technologien* (vgl. Kapitel 2.1) in vielen Bereichen zur Anwendung kommen. Daraus kann gefolgert werden, dass diese Technologien für viele Branchen bereits wichtig sind und zunehmend wichtiger werden.

Ein detaillierter Blick in die verschiedenen Kategorien des Datensatzes wird mit Hilfe eines Netzwerkes möglich, das in Abbildung 29 dargestellt ist.

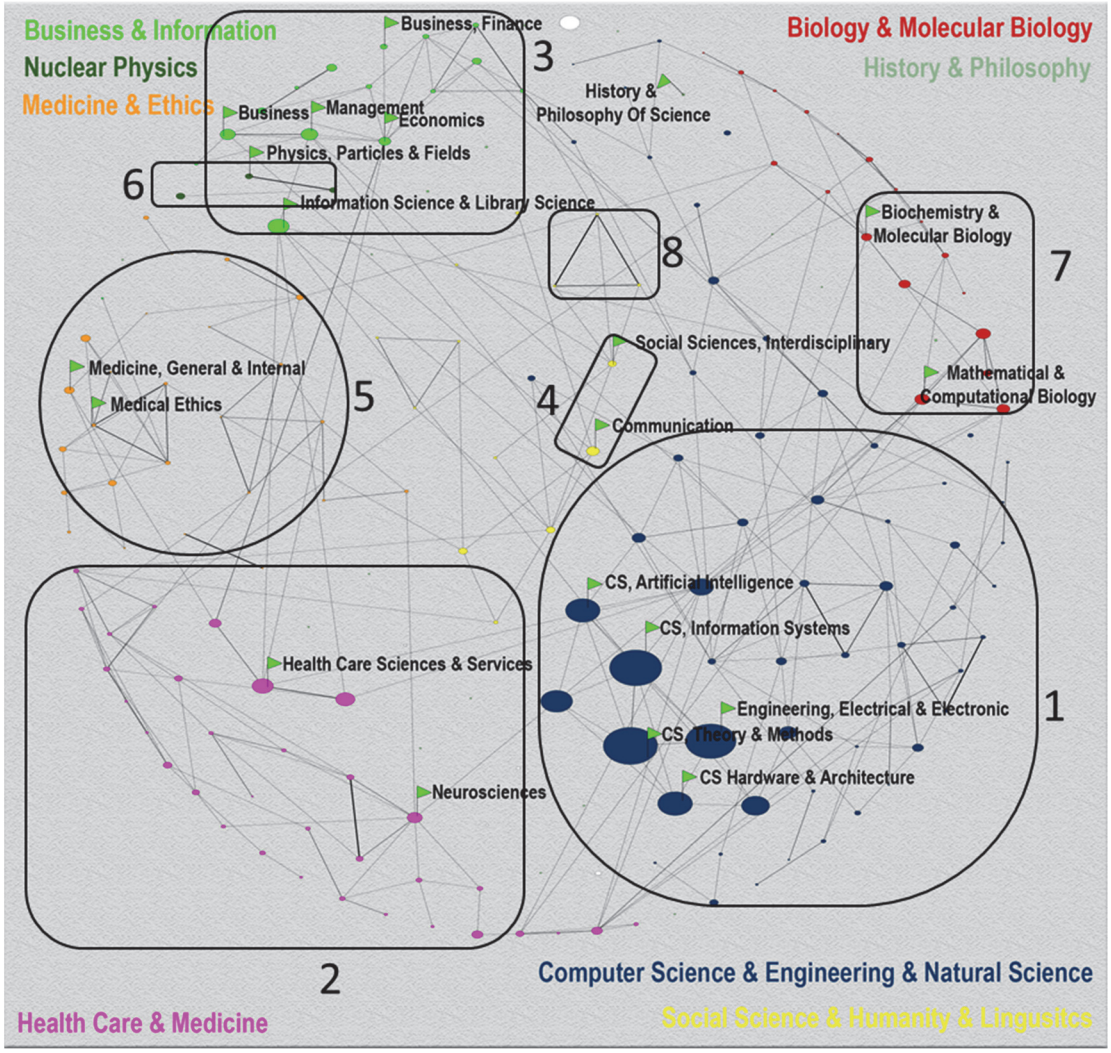


Abbildung 29: Netzwerk ausgewählter Kategorien mit übergeordneten Clustern[87]

Wie in Abbildung 29 zu sehen ist, gibt es mehrere Überbegriffe für die Web of Science Kategorien. Die dunkelblaue Überkategorie (Nr. 1) ist dabei die Bedeutsamste, da in den dazugehörigen Kategorien die meisten Publikationen des Datensatzes veröffentlicht wurden. Der Vorteil dieses Netzwerkes ist, dass die Fachrichtung *Computer Science* aufgeteilt wird. Das Netzwerk zeigt, dass es Veröffentlichungen gibt, die sich primär mit der benötigten Hardware beschäftigen oder andere mit der *Artifical Intelligence* (AI), was übersetzt künstliche Intelligenz bedeutet. Solche AI-Systeme sind selbstlernende

[87] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Netzwerk mit 189 Knoten und 298 sichtbare von 546 Kanten

Systeme, welche in den letzten Jahren starke Fortschritte gemacht haben. Darunter fällt zum Beispiel der Supercomputer Watson von IBM, welcher in Zukunft unter anderem Ärzten bei Operationen oder Beurteilungen von Symptomen unterstützen wird, was an den Verbindungen zu den rosafarbenen Knoten (Nr. 2) sichtbar ist.[88] Andere Beispiele sind die selbstfahrenden Fahrzeuge von Google oder Apples persönlicher Assistent Siri.[89] Darüber hinaus wird die Aktivitäten in der Fachrichtung *Engineering* primär der Kategorie Elektronik zugeordnet.

Neben den beiden bereits beschriebenen Kategorien sind auch die hellgrün eingefärbten Knoten (Nr. 3) der Kategorie *Business & Information* essentiell. Hierbei werden vorrangig Themen wie die Informationswissenschaften, Management, Business, Finance, Economics und Transporttechnologien betrachtet. Aus diesem Grund wird der Bereich in Kapitel 4.3 im Rahmen der Wissensbasen noch einmal aufgegriffen. Als wichtiger Artikel sei an dieser Stelle dennoch „*Big Data*, Analytics and the Path From Insights to Value“ erwähnt. Dieser behandelt vor allem den Wert von in Unternehmen gesammelten Daten und den daraus resultierenden Informationen.[90] Dieser Artikel kann auch als Ratgeber für Unternehmen gesehen werden, da er die Vorteile von *Big Data* Analysetools aufzeigt. Unternehmen, die Entscheidungen basierend auf einer großen Datenmenge treffen, schneiden in allen Bereichen besser ab als Konkurrenten ohne diese Informationstechnologien.

[88] Vgl. Brynjolfsson und McAfee (2014)
[89] Vgl. Lohr (2012)
[90] Vgl. LaValle et al. (2011)

Die weiteren Kategorien werden mit Hilfe von Tabelle 5 beschrieben.

Farbliche Kennzeichnung in Abbildung 29	Kategorie	Kurzbeschreibung
Gelb (Nr. 4)	Social Science & Humanity & Linguistics	- Betrachtungen bzgl. Sozialwissenschaften und Soziologie - „Critical questions for *Big Data*“ beschäftigt sich mit den Auswirkungen von *Big Data*-Technolgien auf die Gesellschaft[91]
Orange (Nr. 5)	Medicine & Ethics	- rechtliche und ethische Auseinandersetzung mit *Big Data* im Rahmen der Medizin und des Gesundheitswesen
Dunkelgrün (Nr. 6)	Nuclear Physics	- keine relevanten Artikel für den Kontext *Industrie 4.0*
Rot (Nr. 7)	Biology & Molecular Biology	- sehr günstig und schnell Sequenzierung von Zellen und DNA Strängen mit Hilfe von *Big Data* Anwendungen[92]
Grau (Nr. 8)	History & Philosophy	- keine relevanten Artikel für den Kontext *Industrie 4.0*

Tabelle 5: Übersicht über ausgewählte Kategorien im Kontext *Industrie 4.0*

Zum Abschluss dieses Kapitels werden in Tabelle 6 die 40 am häufigsten verwendeten Web of Science Kategorien aufgeführt. Erklärend bedeutet dies für die Kategorie Management, dass 68 Artikel des Datensatzes dieser Kategorie zugeordnet werden können. Auch aus Tabelle 6 wird ein weiteres Mal klar, dass *Big-Data-Technologien* in sehr vielen verschiedenen Wissenschaftsbereichen aber auch Industriebranchen eingesetzt werden können.

[91] Vgl. Boyd und Crawford (2012)
[92] Vgl. Schadt et al. (2010)

Web of Science Kategorie	Häufig-keit	Web of Science Kategorie	Häufigkeit
Computer Science, Theory & Methods	743	Biochemical Research Methods	43
Computer Science, Information Systems	688	Communication	42
Engineering, Electrical & Electronic	650	Statistics & Probability	42
Computer Science, Hardware & Architecture	306	Health Policy & Services	35
Computer Science, Artificial Intelligence	302	Economics	34
Computer Science, Software Engineering	256	Genetics & Heredity	34
Telecommunications	199	Pharmacology & Pharmacy	31
Computer Science, Interdisciplinary Applications	148	Physics, Applied	31
Health Care Sciences & Services	112	Chemistry, Multidisciplinary	29
Information Science & Library Science	111	Engineering, Industrial	29
Multidisciplinary Sciences	109	Environmental Sciences	29
Medical Informatics	92	Biochemistry & Molecular Biology	28
Optics	80	Automation & Control Systems	27
Management	68	Medicine, General & Internal	25
Business	57	Political Science	25
Mathematical & Computational Biology	57	Remote Sensing	25
Biotechnology & Applied Microbiology	56	Engineering, Mechanical	23
Neurosciences	55	Engineering, Multidisciplinary	23
Operations Research & Management Science	45	Law	23
Materials Science, Multidisciplinary	44	Mathematics, Applied	23

Tabelle 6: 40 am häufigsten verwendeten Web of Science Kategorien nach Publikationsanzahl[93]

[93] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

4.2.3 Netzwerk der Schlagwörter

Zusammenfassung

Im Rahmen einer Textanalyse innerhalb aller Publikationen sowie im Netzwerk aller Schlagwörter fällt erwartungsgemäß das Schlagwort *Big Data* am häufigsten, vor *Cloud Computing, MapReduce* und *Hadoop*, bzw. hat eine zentrale Position inne.

Trotzdem kann die Erwartung nach einer eindeutigen Definition des Begriffs *Big Data* nicht erfüllt werden. Vielmehr wurde deutlich, dass der Begriff mit Hilfe einer Vielzahl anderer Begriffe verknüpft ist, was sich in der Breite der Fachgebiete (vgl. Abbildung 29) widerspiegelt.

Die Analyse von Schlagwörtern ist von äußerstem Interesse, da eine schnelle Übersicht über häufig genannte Themen möglich ist. So kann mit einem Blick die aktuelle Verteilung der Kernthemen erkannt werden. Bei Schlagwörtern handelt es sich um charakterisierende Wörter für einzelne Publikationen. Beispiele können *Big Data*, *Cloud Computing*, *Social Media* etc. sein. Dadurch können Forschungsgebiete identifiziert werden, welche für das Hauptthema dieser Arbeite *Industrie 4.0* wichtig sind.

Die 25 wichtigsten Schlagwörter sind in Abbildung 30 in Form eines Kuchendiagramms dargestellt. Dabei stellt die Größe des Kreissegmentes die Häufigkeit des Schlagwortes dar.[94]

[94] Bei der Analyse der Schlagwörter wurden nur jene verwendet, welche auch von den Autoren selbst vorgeschlagen wurden. Dabei wurde der Datensatz so bereinigt, dass jedes Schlagwort mehr als einmal vorkommen muss. Trotz dieser Maßnahme führte unser Datensatz von 2939 Publikationen zu 974 Schlagwörtern.

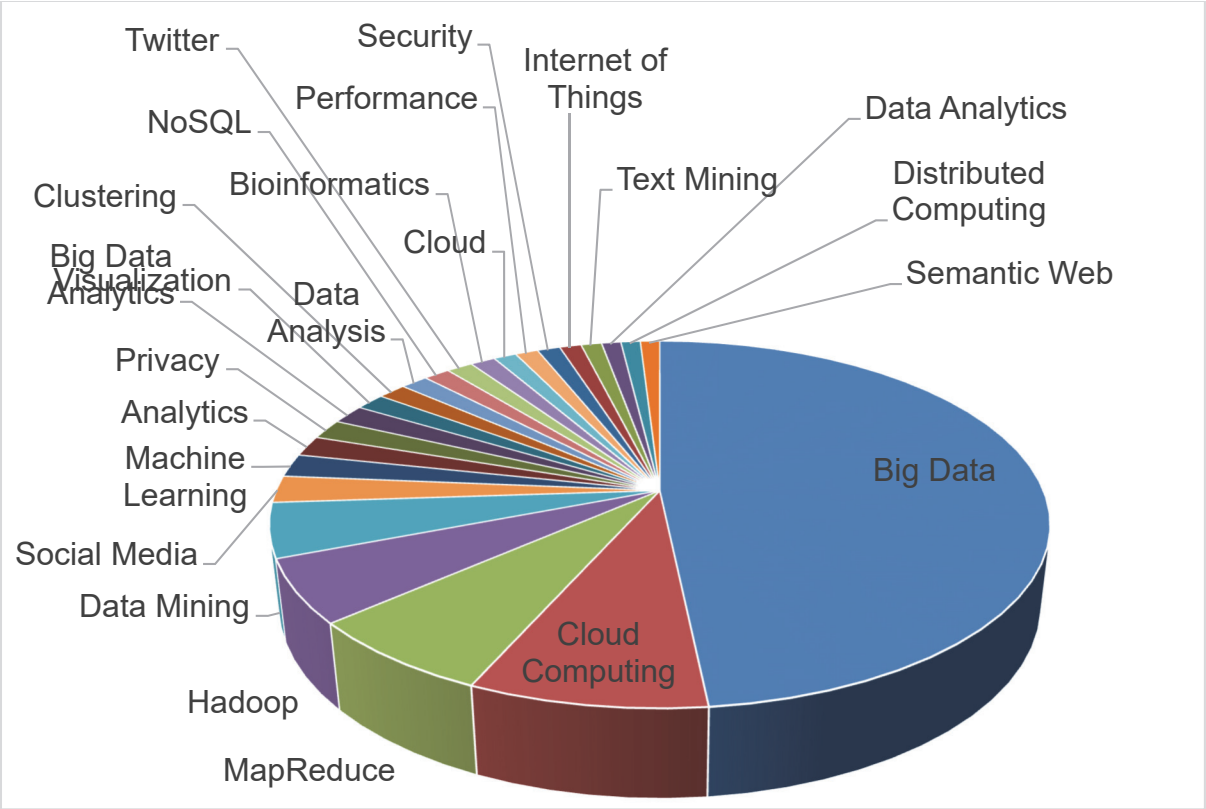


Abbildung 30: Die 25 wichtigsten Schlagwörter aus dem Datensatz[95]

Wie bereits intuitiv erwartet, wird das Schlagwort *Big Data* am Häufigsten verwendet und ist in Abbildung 31 dementsprechend als zentraler Knoten zu erkennen. *Cloud Computing* ist das zweithäufigste Schlagwort. Was dieses bedeutet wird genauer in Kapitel 4.3.1.1 erklärt. Zwei weitere wichtige Schlagwörter sind *MapReduce* und *Hadoop*, welche wie bereits erwähnt fundamentale *Big-Data-Technologien* darstellen und auch technologisch eng mit einander verbunden sind. Diese wurden schon im Kapitel 4.2.1 Netzwerk der Autoren vorgestellt und in Kapitel 4.3.1.1 genauer erklärt.

Ein weiteres wichtiges Schlagwort ist *Data Mining.* Dabei handelt es sich um "automatische Auswertung großer Datenmengen zur Bestimmung bestimmter Regelmäßigkeiten, Gesetzmäßigkeiten und verborgener Zusammenhänge"[96]. Die Tätigkeiten des *Data Mining* sind für industrielle Anwendungen äußerst interessant. Es kann zum Beispiel eingesetzt werden, „um transaktionale Daten zu analysieren und daraus Geschäftsrisiken und -opportunitäten frühzeitig zu erkennen"[97].

Machine Learning (vgl. Kapitel 4.3.1.1) soll hier nur kurz beschrieben werden. "Der Begriff *Machine Learning* beschreibt eine Vielzahl von Anwendungen und Methoden, in denen Computerprogramme selbstständig neues Wissen erwerben. Üblicherweise geschieht dies durch statistische oder logische Analysen gegebener Daten sowie durch die Anwendung rechenintensiver Algorithmen, um optimale Lösungen zu gestellten Aufgaben zu finden"[98]. Aus dieser Beschreibung geht hervor, dass diese Technologie für den Einsatz in der *Industrie 4.0* sehr wichtig sein kann.

[95] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[96] Vgl. DUDEN (2013)
[97] Vgl. BITKOM (2014), S. 61
[98] Vgl. BITKOM (2014), S. 66

Die Verknüpfungen zwischen den einzelnen Schlagwörtern werden im Folgenden mit Hilfe eines Netzwerkes dargestellt.

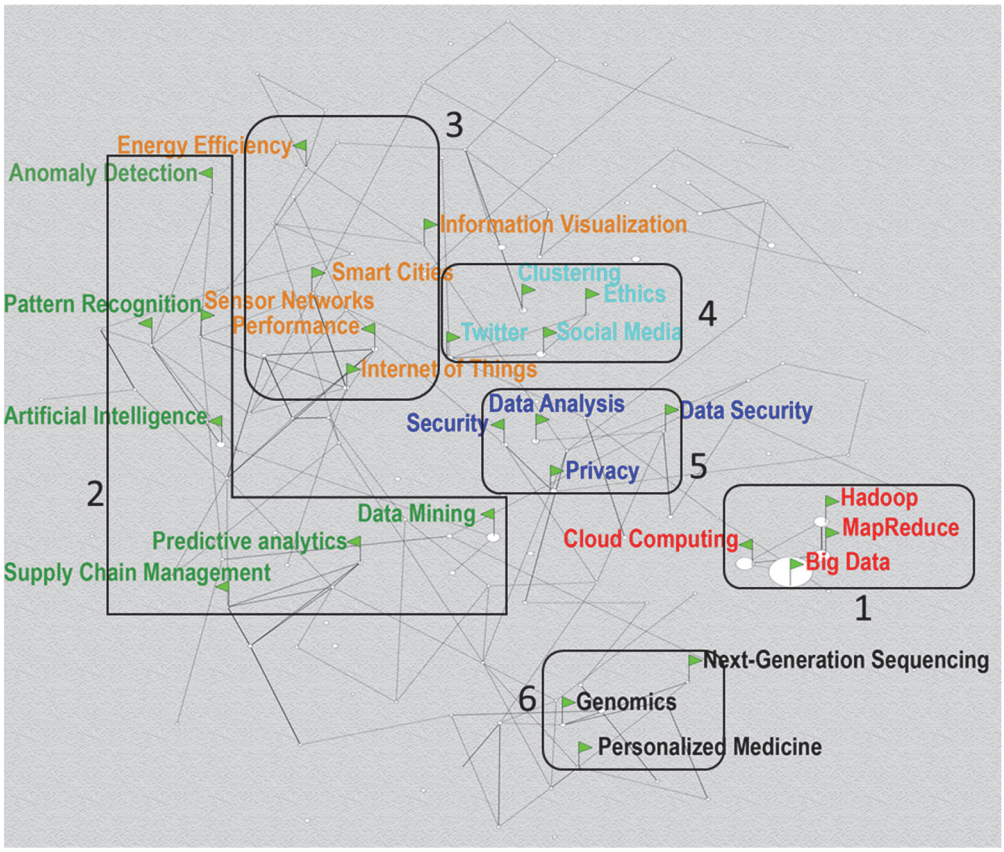


Abbildung 31: Netzwerk der wichtigsten Schlagwörter[99]

In Abbildung 31 sind die Schlagwörter erneut farblich kodiert und entsprechend der zugehörigen Fahnen beziffert. Die einflussreichste Einheit, deutlich zu sehen an der Größe der Knoten, ist die rote Teilkomponente (Nr. 1) um die Schlagwörter *Cloud Computing*, *Hadoop*, *MapReduce* und *Big Data*. Die grün hinterlegten Schlagwörter (Nr. 2) beziehen sich auf die Analyse von Daten im Zusammenhang mit Supply Chain Management. Allgemeiner formuliert mit der Analyse von Daten um Regelmäßigkeiten, Muster und Zusammenhänge zu erkennen. Aus diesem Grund sind diese Schlagwörter eng mit der orange gefärbten Teilkomponente (Nr. 3) verbunden. Bei dieser dreht es sich um das Internet der Dinge und Sensornetzwerke. Beispielsweise sind diese Technologien wichtig für Smart Cities oder auch für eine erhöhte Energieeffizienz. Die türkis hinterlegten Schlagwörter (Nr. 4) beziehen sich alle auf die Analyse von Social Media Daten. Interessant ist hier vor allem, dass Twitter ein eigenes Schlagwort bekommen hat. Bei den dazugehörigen Publikationen versucht man User in Gruppen zu clustern. Damit können beispielsweise politische Trends besser verstanden werden.[100] Bei den blauen Schlagwörtern (Nr. 5) handelt es sich vor allem um Datenschutz, Datensicherheit und Datenanalyse. In einer Veröffentlichung geht es um *Big Data* Anwendungen im Gesundheitswesen. Hier gilt umso mehr Daten von Patienten zur Verfügung stehen umso besser können Zusammenhänge erkannt werden und somit allen Patienten besser geholfen werden. Jedoch gibt es hier auch einige Probleme beim

[99] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; 134 Knoten; 141 von 890 Kanten sichtbar. Clusterung bei 974 Schlagwörter nicht gewinnbringend, daher manuelle Reduzierung auf alle Schlagwörter, die mindestens in fünf Artikeln genannt wurden.

[100] Vgl. Ituski et al. (2013)

Datenschutz und der Datensicherheit, welche in der Zukunft aber lösbar sind.[101] In den schwarz dargestellten Schlagwörtern (Nr. 6) werden primär Themen der personalisierten Medizin und Genomsequenzierung betrachtet. Durch den Einsatz von *Big-Data-Technologien* können weite Teile der DNA-Sequenzierung automatisiert werden, was erhebliche Effektivitätssteigerungen garantiert. Eine Sequenzierung benötigt weniger Zeit als zuvor, was dazu führt, dass in einer Zeiteinheit wesentlich mehr Sequenzierungen durchgeführt werden können als ohne den Einsatz von *Big-Data-Technologien*.[102]

4.3 Wissensbasen und Forschungsfronten

Wissensbasen und Forschungsfronten sollen einen Blick in die Vergangenheit und einen in die Zukunft des Themenfelds *Big Data* aufzeigen. Die Wissensbasen beschreiben dabei die Forschungsgebiete, aus welchen sich die Themengebiete von *Big Data* entwickelt haben. Die Forschungsfronten beantworten die Frage, in welchen Forschungsgebieten momentan am meisten publiziert wird und zeigen somit, was die Wissenschaft gerade am meisten beschäftigt.

Dies führt zu einem erweiterten Verständnis des bisher geschaffenen Bildes zum Themengebiet *Big Data*.

Der bisher benutzte Datensatz aus dem Web of Science wird nun um die Referenzen der Publikationen erweitert. Anschließend wird ein Blick auf die Vergangenheit und den damit verbundenen Wissensbasen möglich.

Bei den Forschungsfronten, diese sind Anhäufungen von Artikeln mit aktuell wichtigen Forschungsthemen, werden Publikationen betrachtet, die gemeinsame Quellen in ihrer Referenzliste aufweisen. Ähnliche Quellenverzeichnisse sind ein starker Hinweis, dass diese beiden Publikationen inhaltlich ein ähnliches Themengebiet behandeln. Anhäufungen mit besonders vielen Publikationen behandeln dabei Forschungsgebiete von zunehmender Bedeutung.

4.3.1 Wissensbasen: Landkarte der Co-Zitationen

Sinn und Zweck der Analyse von Wissensbasen ist es, die Quellen der Artikel aus unserem Datensatz zum Themengebiet *Big Data* zu untersuchen und miteinzubeziehen. Einzelne Themenfelder werden hier durch Quellen definiert, welche eine hohe Gemeinsamkeit aufweisen. Diese einzelnen Themengebiete definieren die Teilwissensbasen im Themenfeld *Big Data*. Sie liefern detaillierte Informationen über das zugrundeliegende Themengebiet *Big Data* und betrachten die wissenschaftliche Basis des Forschungsgebietes.

Wie können Quellen bzw. Referenzen systematisch erkannt werden und wie wird ihre Gemeinsamkeit definiert? In den folgenden Abschnitten soll dies an einem Beispiel erklärt werden.

Einer der wichtigsten Forscher im Gebiet *Big Data* ist Jeffrey Dean, welcher für das Unternehmen Google Inc. arbeitet. Mit seiner Veröffentlichung „MapReduce: simplified data processing on large clusters“ stieß er das Tor für neue Anwendungen auf und machte *Big Data* zum großen Thema in allen Wissenschaftsbereichen. Das Programmiermodell *MapReduce* macht die parallele Berechnung über

[101] Vgl. Fabian et al. (2015)
[102] Vgl. Veltman et al. (2013)

sehr große Datenmengen auf Computerclustern möglich *MapReduce* wurde durch das Open Source Framework *Hadoop* für alle denkbaren Anwendungen nutzbar gemacht. Dazu erschien das Buch „Hadoop: the definitive guide: the definitive guide" von Tom White aus dem Jahre 2009. Dieses kann als Anleitung zur Implementation und Benutzung von *MapReduce* gesehen werden. Sowohl die Veröffentlichung von Jeffrey Dean als auch das Buch von Tom White sind Referenzen für die Veröffentlichung „The Family of MapReduce and Large-Scale Data Processing Systems" von Skar et al.[103] Dieser beschäftigt sich mit den Grenzen des Programmiermodells *MapReduce* und wie diese im *Hadoop* Framework und durch andere Methoden über die Jahre verbessert worden sind. Außerdem gibt er einen Überblick über implementierte Systeme.

Am obigen Beispiel der drei Autoren ist zu sehen, dass sich die Gemeinsamkeit zweier Referenzen über die Co-Zitation ermitteln lässt. Dies bedeutet, dass beide als eine Referenz in einer Publikation zu finden sind. Schematisch ist dies in Abbildung 32 dargestellt.

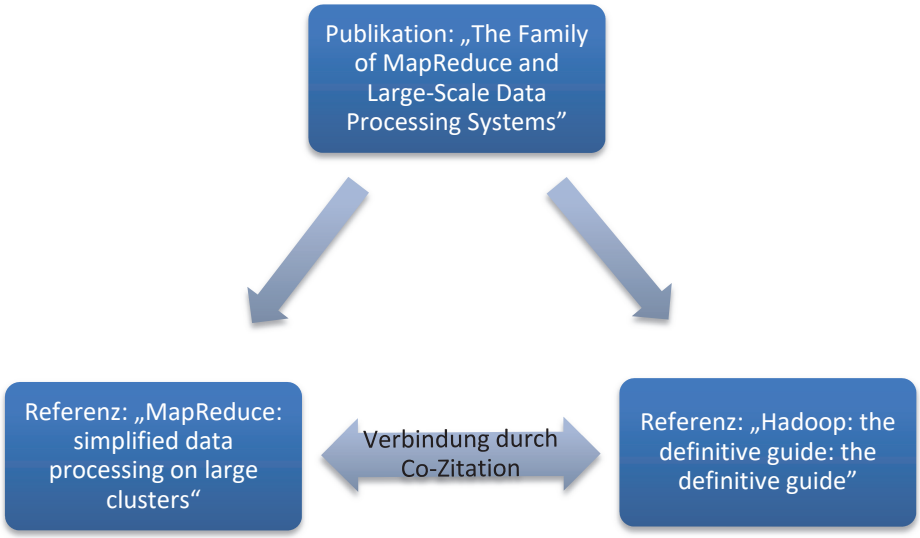


Abbildung 32: Erklärung der Co-Zitation durch drei ausgewählte Publikationen

In einer Wissensbasis liegt eine Anhäufung von co-zitierten Publikationen vor. In dieser Wissensbasis bilden sich Teilwissensbasen heraus, die sich auf bestimmte Forschungsgebiete beziehen. Je mehr Zitate enthalten sind, umso größer ist die Dimension der Teilwissensbasis. Abbildung 33 zeigt das Co-Zitationsnetzwerk mit den Teilwissensbasen.

[103] Vgl. Sakr et al. (2013)

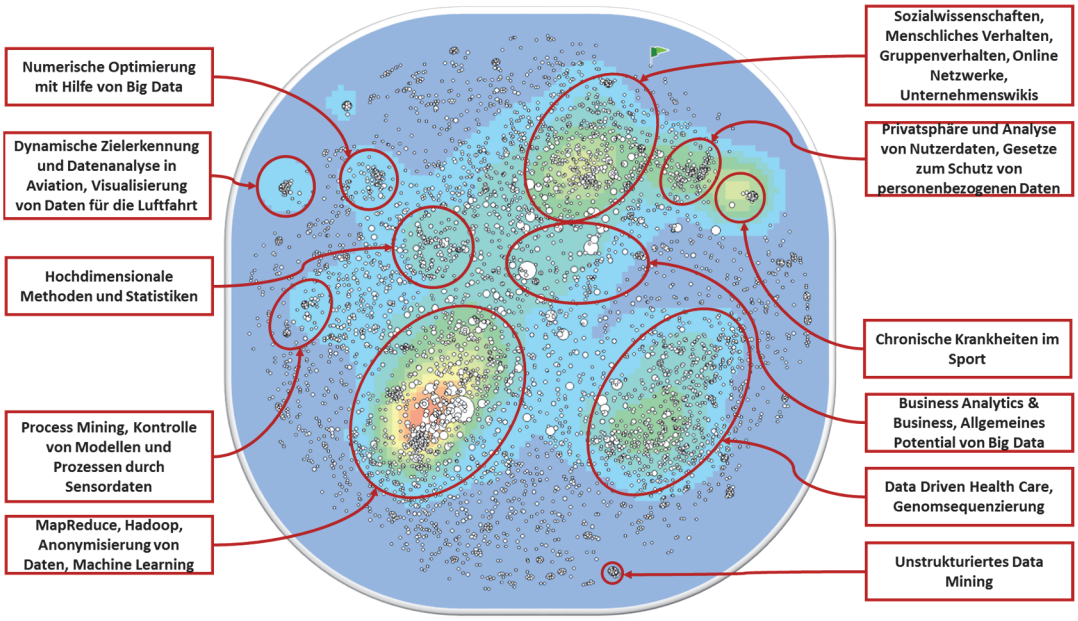


Abbildung 33: Wissensbasen - Dichteverteilung von co-zitierten Referenzen[104]

In der nachfolgenden Tabelle 7 sind die einzelnen Teilwissensbasen mit der dazugehörigen Zahl der Referenzen aufgelistet.

[104] BibTechMon4 i.V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Komponente mit 2939 Artikeln. Die Knoten, hier als weiße Kreise dargestellt, sind die Referenzen. Umso größer der Kreis ist, desto mehr Artikel zitieren diese Referenz. Erklärung zur Farbgebung: blaue Zonen zeigen keine oder niedrige gewichtete Anzahl co-zitierter Referenzen rote Zonen weisen eine hohe gewichtete Anzahl von Referenzen aus, Gewichtung durch die Stärke aller Verbindungen (Jaccardindex der Zahl der Co-Zitationen) je Referenz.

Wissensbasen	Zahl der Referenzen	Anteil an Gesamt-referenzen in %
- *MapReduce* - *Hadoop* - Anonymisierung von Daten - Machine Learning	426	31.03%
- Sozialwissenschaften - Menschliches Verhalten - Gruppenverhalten - Online Netzwerke - Unternehmenswikis	178	12.96%
- Data Driven Health Care - Genomsequenzierung	167	12.16%
- Privatsphäre und Analyse von Nutzerdaten - Gesetze zum Schutz von personenbezogenen Daten	114	8.30%
- Chronische Krankheiten im Sport	101	7.36%
- Hochdimensionale Methoden und Statistiken	89	6.48%
- Numerische Optimierung mit Hilfe von *Big Data*	68	4.95%
- Business Analytics & Business - allgemeines Potential von *Big Data*	67	4.88%
- Dynamische Zielerkennung und Datenanalyse in Aviation - Visualisierung von Daten für die Luftfahrt	63	4.59%
- Process Mining - Kontrolle von Modellen und Prozessen durch Sensordaten	62	4.52%
- Unstrukturiertes Data Mining	38	2.77%

Tabelle 7: Wissensbasen zum Suchbegriff Big Data[105]

Es wurden elf Wissensbasen identifiziert. Dies lässt sich durch den relativ breit aufgestellten Begriff *Big Data* erklären. *Big-Data-Technologien* kommen in sehr vielen verschiedenen Forschungsgebieten zum Einsatz. Die beiden größten Wissensbasen lassen sich in kleinere Teilgebiete unterteilen. Eine sehr starke Wissensbasis bildet sich um die bereits als essentiell identifizierten Begriffe *MapReduce*, *Hadoop* und *Machine Learning* (vgl. Kapitel 4.2.3).

Big Data ist ein relativ junges Forschungsfeld, was auch in dem vorherigen Kapitel 4.1.2. gezeigt wurde. Dort war die zeitliche Entwicklung der Veröffentlichungen zu finden, welche ab dem Jahre 2006 signifikant zunahm. Aus Tabelle 7 lassen sich die wichtigsten Wissensbasen des Forschungsgebiets *Big Data* entnehmen. Im Folgenden sollen vorrangig die Wissensbasen mit Bezug zur *Industrie 4.0* behandelt werden. Anhand der Anzahl der Referenzen ist das Gebiet „*Big Data* – theoretischer Hintergrund“ als wichtigste Wissensbasis zu identifizieren.

[105] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

4.3.1.1 *Teilwissensbasis „Big Data – theoretischer Hintergrund“*

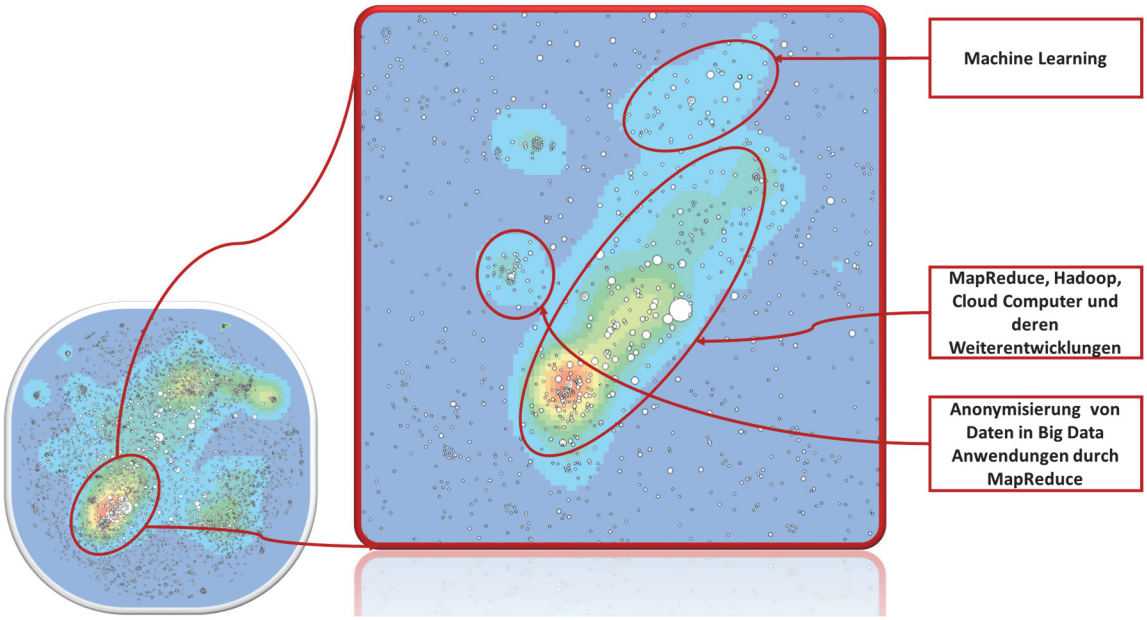


Abbildung 34: Teilwissensbasis „Big Data - theoretischer Hintergrund“[106]

Diese behandelt vor allem den Algorithmus *MapReduce*, das Framework *Hadoop*, *Cloud Computing*, die Anonymisierung von Daten und das *Machine Learning*. Ein detaillierter Blick in die Wissensbasen erlaubt diese Aufschlüsselung, welche in Abbildung 34 zu erkennen ist. Die dazugehörige Zahl der Referenzen pro Themengebiet lässt sich in nachfolgender Tabelle 8 ablesen.

Teilwissensbasis	Nebenbasen	Zahl der Referenzen
Big Data – theoretischer Hintergrund	- *MapReduce* - *Hadoop* - *Cloud Computing* und Weiterentwicklungen	340
	- *Machine Learning*	43
	- Anonymisierung von Daten in *Big Data* Anwendungen durch *MapReduce*	43

Tabelle 8: Teilwissensbasis „Big Data - theoretischer Hintergrund“[107]

Der am häufigsten zitierte Artikel in der Teilwissensbasis um *MapReduce*, *Hadoop*, *Cloud Computing* und deren Weiterentwicklungen lautet „MapReduce: simplified data processing on large clusters“ und wurde von Jeffrey Dean 2008 veröffentlicht. Diese Referenz wurde von 198 anderen Artikeln zitiert. Dieses Modell stellt die Basis für sehr viele *Big Data* Anwendungen dar. Nutzbar gemacht wurde *MapReduce* durch das Open Source Framework *Hadoop* der ehrenamtlich operierender Organisation

[106] BibTechMon4 i.V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Farbgebung siehe 105
[107] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

Apache Software Foundation.[108] Wie das Framework *Hadoop* implementiert werden kann, wird im Buch „Hadoop: the definitive guide: the definitive guide" von Tom White aus dem Jahre 2009 beschrieben. Dieses Buch ist eine sehr wichtige Referenz in der Teilwissensbasis „*Big Data* – theoretischer Hintergrund", da es im Datensatz von 67 Artikeln zitiert wird. Im Zusammenhang mit dem Programmiermodell *MapReduce* und dem dazugehörigen Open Source Framework *Hadoop* wird auch oft der Begriff *Cloud Computing* genannt. Dieser Begriff wird durch das amerikanische National Institute of Standards and Technology folgendermaßen definiert: "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."[109] Übersetzt für unsere Wissensbasis bedeutet dies, dass die *Big-Data-Technologien MapReduce* und *Hadoop* als Enabler für *Cloud Computing* angesehen werden können.[110]

Ein wichtige Nebenwissensbasis der Teilwissensbasis „*Big Data* - theoretischer Hintergrund" ist die „Anonymisierung von Daten in *Big Data* Anwendungen". Der Leitfaden „*Big Data* – Wissen für Entscheider" des Branchenverbands BITKOM aus dem Jahre 2014 fordert, dass "[die] unternehmensweit gültigen Verfahren für Datenschutz, Anonymisierung und Pseudonymisierung […] auf Datenvolumina im Multiterabytes-Bereich und Petabytes-Bereich angepasst werden [müssen]."[111] Dies bedeutet, dass gerade im Anwenderbereich *Industrie 4.0* von *Big-Data-Technologien* über die Anonymisierung von Daten nachgedacht werden muss. Diese Nebenwissensbasis beschäftigt sich mit den Grundlagen der Anonymisierung von persönlichen, unternehmensbezogenen, aber auch maschinenbezogenen Daten mit Hilfe von *Big-Data-Technologien*. Eine wichtige Referenz ist der Artikel "Incognito: Efficient full-domain k-anonymity" von Kristen LeFevre, welcher sich mit der k-Anonymisierungstechnik beschäftigt. Dies bedeutet, dass beispielsweise bei einem personenbezogenen Datensatz der Menge k kein Individuum mehr ausfindig gemacht werden kann.[112] Der theoretische Hintergrund von *Big Data* wird von der Nebenwissensbasis *Machine Learning* abgerundet. Eine wichtige Fragestellung ist, was *Machine Learning* bedeutet. Tom M. Mitchell von der Universität Carnegie Mellon benennt in seinem Artikel über die Disziplin *Machine Learning* die entscheidende Frage wie folgt: "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"[113]. Diese Art maschinellen Lernens ist eine wichtige Voraussetzung für *Big Data* Anwendungen im Kontext *Industrie 4.0*, da z. B. Produktionsdaten einer Maschine in Gigabytegrößen anfallen und ohne maschinelles Lernen nicht gewinnbringend für Unternehmen weiterverarbeitet werden können. Eine Wichtige Referenz ist hier die Veröffentlichung "Support-vector networks" von Corinna Cortes und Vladimir Vapnik aus dem Jahr 1995. Diese stellen ein neuartiges Konzept des Machine Learnings vor, welches von vielen Artikeln in unserem Datensatz zitiert wird.

---

[108] Vgl. Apache (2015)

[109] Vgl. Mell und Grance (2011)

[110] *MapReduce* ist dabei ein Algorithmus, welcher die Verteilung und Verarbeitung von Daten auf einem Servercluster zulässt. Dabei findet eine zweiphasige Verarbeitung statt (Map-/Reducephase). *Hadoop* ist ein Framework, welches den *MapReduce* Algorithmus beinhaltet. Es ist ein Open-Source-Projekt, welches auf der Programmiersprache Java basiert und zum verteilten Speichern von Daten und deren paralleler Verarbeitung auf Standardhardware ermöglicht.

[111] Vgl. BITKOM (2014); Falkenberg et al. (2014), S. 140

[112] Vgl. LeFevre et al. (2005)

[113] Vgl. Mitchell (2006)

4.3.1.2 Teilwissensbasis „Sozialwissenschaften“

Das Netzwerk zur Teilwissensbasis „Sozialwissenschaften“ ist in Abbildung 35 zu sehen. Es werden Veränderungen in Bezug auf das menschliche Verhalten, das Miteinander und die Kommunikation analysiert, die durch den Einsatz von *Big-Data-Technologien* im Alltag beobachtet werden. Hierunter fallen auch Analysen zum Gruppenverhalten, die Analyse von ortsbezogenen Daten, der Einfluss auf Online Communities, sowie die für die *Industrie 4.0* interessanten Unternehmenswikis.

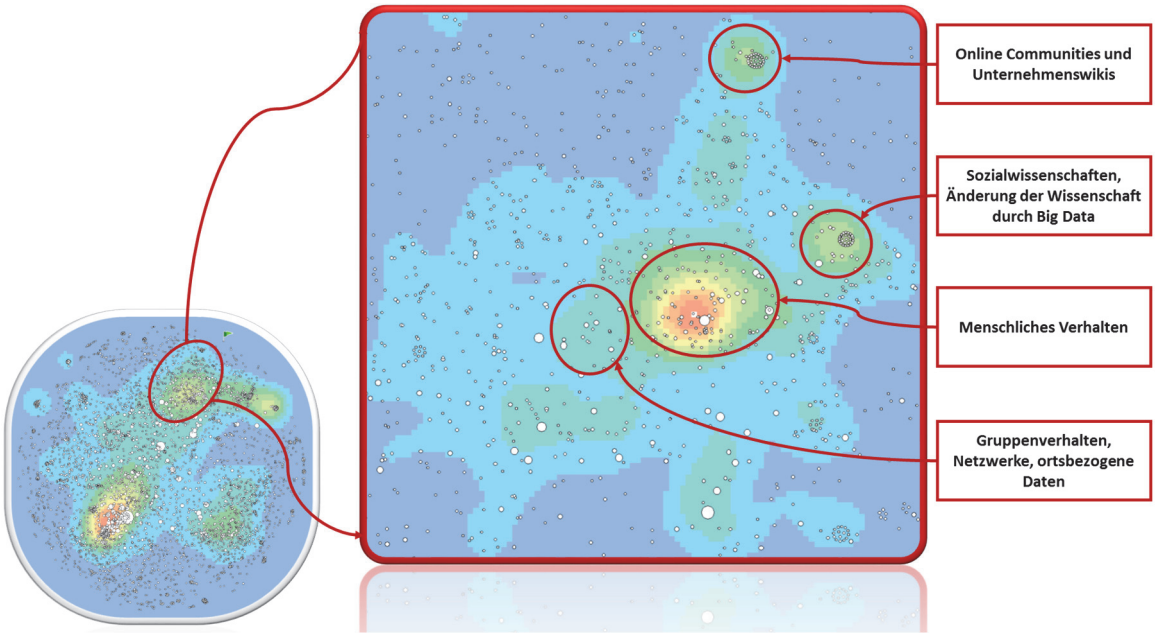


Abbildung 35: Teilwissensbasis „Sozialwissenschaften“[114]

Eine Auflistung der Nebenwissensbasen mit den dazugehörigen Anzahlen an Referenzen liegt in Tabelle 9 vor. Dort ist zu sehen, dass sich bisher am meisten mit dem menschlichen Verhalten beschäftigt wurde. Aber auch die Auswertung von Online Community Daten, so wie Unternehmens-Wikis, spielen eine nicht zu verachtende Rolle. Diese Nebenwissensbasis ist für die *Industrie 4.0* nicht uninteressant, gerade im Bereich des Marketings bietet *Big Data* großes Potential für beispielsweise die Mikrosegmentierung von Märkten, die Erhöhung von Cross-Selling, ortsbezogenes Marketing und Stimmungserkennung.[115]

[114] BibTechMon4 i.V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015
[115] Vgl. Manyika et al. (2011), S. 67

Teilwissensbasis	Nebenwissensbasen	Zahl der Referenzen
Sozialwissenschaften	- Menschliches Verhalten	82
	- Sozialwissenschaften - Änderung der Wissenschaft durch *Big Data*	39
	- Online Communities - Unternehmenswikis	36
	- Gruppenverhalten - Netzwerke - ortsbezogene Daten	21

Tabelle 9: Teilwissensbasis "Sozialwissenschaften"[116]

Eine äußerst wichtige Referenz der Nebenwissensbasis „Menschliches Verhalten“ ist die Publikation „Social Science – Computational Social Science“ von David Lazer et al. aus dem Jahre 2009. Hier wird das Potential von durch die Nutzer generierten Daten betrachtet, diese werden bisher vor allem durch kommerzielle Unternehmen wie Google und Yahoo analysiert und nutzbar gemacht. Gerade das menschliche Verhalten, sowohl individuell betrachtet als auch in der Gruppe, lassen sich durch die benutzerbezogenen Daten weit genauer analysieren, als es bisher möglich war. Dies wird auch Sozialwissenschaften allgemein revolutionieren. Um nochmal eine Brücke zur *Industrie 4.0* zu schlagen soll nun die Nebenwissensbasis „Online Communities und Unternehmenswikis“ betrachtet werden. Die wichtigste Referenz in diesem Themenbereich lautet „Wikinomics and its discontents: a critical analysis of Web 2.0 business manifestos” von Jan van Dijck et al. aus dem Jahr 2009. Diese Publikation beschäftigt sich mit Wikinomics und dessen Nutzen in der Unternehmenswelt. Bei Wikinomics handelt es sich um eine neue Form der Zusammenarbeit. Ziel ist es, dass Menschen selbstorganisiert ohne Hierarchien und starre Organisationsstrukturen gemeinsam an Projekten arbeiten. Diesem Konzept liegt die Annahme zugrunde, dass Menschen in einem kreativen und kommunikativen Umfeld viele Ideen und mögliche Innovationen hervorbringen können. Die Veröffentlichung setzt sich damit kritisch auseinander und zeigt auf, wie zum Beispiel Nutzerwissen von Unternehmen genutzt werden kann.

4.3.1.3 Teilwissensbasis „Business Analytics & Business“

Diese Teilwissensbasis ist kleiner als die vorherigen. Allerdings ist das Themengebiet äußerst interessant für den potentiellen Nutzen von *Big Data* im Rahmen der *Industrie 4.0.* Die wichtigste Referenz ist hier das Buch „Big data: A revolution that will transform how we live, work, and think” von Viktor Mayer-Schöneberger. Dies ist zunächst als populärwissenschaftliches Werk einzuordnen, es richtet sich an einen sehr großen Interessentenkreis, bezieht sich aber in einigen Kapitel direkt auf *Big Data* in Unternehmen. Viktor Mayer-Schöneberger probiert hierbei das Potential von *Big Data* Anwendungen herauszustellen. Beispielsweise im Bereich des *Process Mining*, der *Predictive Maintenance* oder auch direkt im *Business Analytics* Bereich. Das McKinsey Global Institute veröffentlichte im Jahr 2011 durch Chui Manyika et al. das Dokument „Big Data: The Next Frontier for Innovation, Competition, and Productivity”. Diese Publikation bietet einen Überblick über das Potential von *Big Data* in verschiedenen Sektoren der öffentlichen, aber auch privaten Wirtschaft. *Business Intelligence* nimmt hierbei ein großes Themenfeld ein. *Business Intelligence* bezeichnet

[116] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

allgemein Verfahren um systematisch Daten in elektronischer Form zu analysieren. Die Ergebnisse sollen dem Unternehmen helfen bessere strategische als auch operative Entscheidungen zu treffen. *Big-Data-Technologien* bieten hier die Möglichkeit extrem viele verschiedene Daten zu analysieren und zeitlich fast in Echtzeit aufzubereiten. McKinsey schätzt das Potential für die Produktivitätssteigerungen im Bereich der Informationsbereitstellung durch *Big-Data-Technologien* sehr hoch ein.

4.3.1.4 Teilwissensbasis „Process Mining"

Zuerst soll eine Definition von *Process Mining* betrachtet werden. "Process Mining: techniques, tools, and methods to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs commonly available in today's (information) systems."[117] Übersetzt und zusammengefasst bedeutet *Process Mining*, dass durch die Analyse von Daten, die über vorhandene Prozesse vorliegen, eine Verbesserung dieser Prozesse erreicht werden kann. Dabei kann das *Process Mining* im Zuge des Prozessmanagement der Prozesslenkung zugeordnet werden. Innerhalb der Prozesslenkung ergibt sich beispielsweise die Frage, wie die Effektivität und Effizienz eines Prozesses laufend gemessen und analysiert werden kann.[118] Eine wichtige Referenz in diesem Themenbereich ist die Publikation „Conformance checking of processes based on monitoring real behavior" von Anne Rozinat aus dem Jahre 2008. Es wird ein Ansatz betrachtet, der es möglich macht Daten aus einem Prozess mit dem Prozessmodell zu vergleichen. Dabei kann entweder kontrolliert werden, wie gut der Prozess funktioniert, oder ob das angenommene Modell den Prozess bestmöglich beschreibt.

4.3.2 Forschungsfronten: Gebirgslandschaft der bibliografisch gekoppelten Publikationen

Nachdem der Blick in die Vergangenheit zum Thema *Big Data* im vorherigen Kapitel 4.3.1 abgeschlossen wurde, soll in diesem Abschnitt analysiert werden, welche Forschungsthemen aktuell von großer Bedeutung sind und in Zukunft noch an Bedeutung gewinnen werden.

Forschungsfronten sind Themengebiete, welche aktuell vorherrschend sind in der Wissenschaft. Dazu werden die Artikel aus Web of Science zum Thema *Big Data* auf gemeinsame Quellen hin untersucht. Die Ähnlichkeit zweier Artikel wird hier durch die Untersuchung der Quellenverzeichnisse festgelegt. Annahme ist, dass zwei Artikel ähnlich zueinander sind, wenn sie die gleichen Referenzen haben (bibliografisch gekoppelte Artikel, vgl. Abbildung 36). Die Dimension einer Anhäufung ähnlicher Artikel (Forschungsfront) ergibt sich aus der Anzahl der dazugehörigen Artikel.

[117] Vgl. Van Der Aalst et al. (2012), S. 15
[118] Vgl. Brecht (2015), S. 149

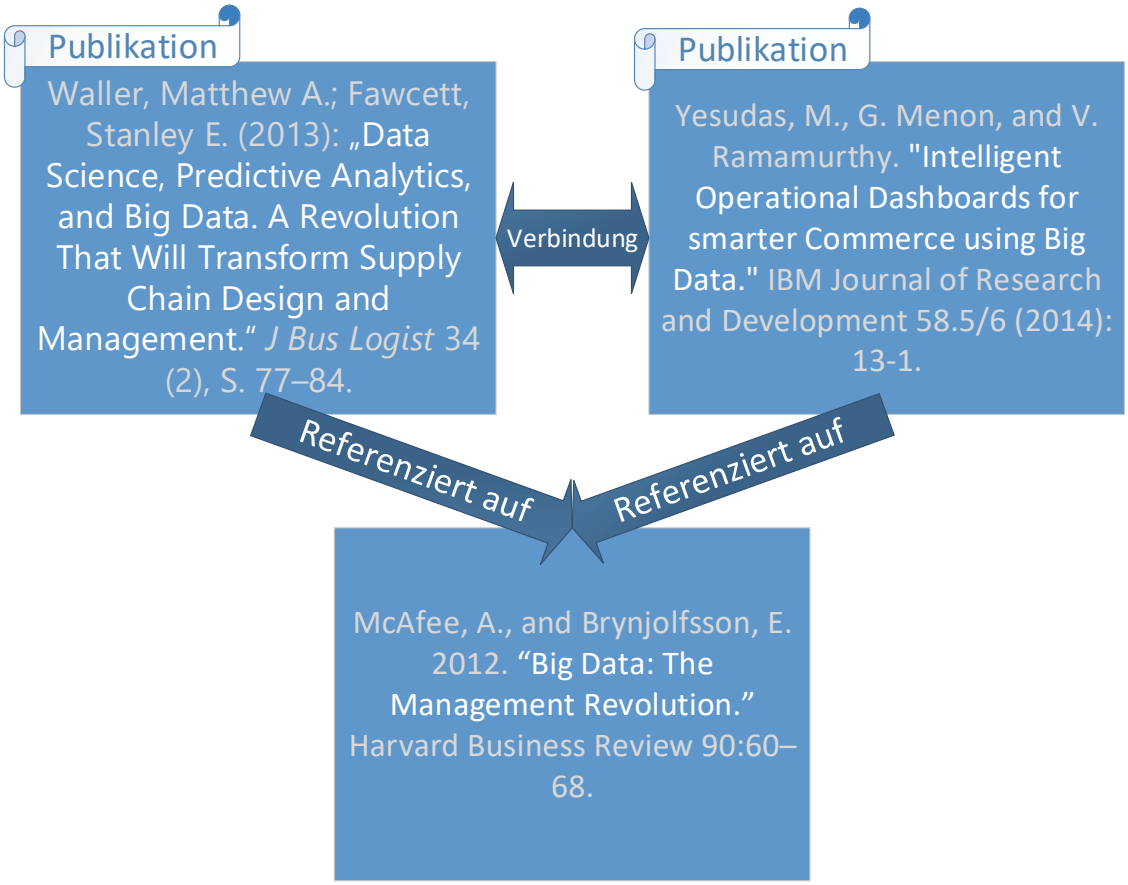


Abbildung 36: Schematische Darstellung einer Forschungsfront[119]

Das Netzwerk, welches in Abbildung 37 zu sehen ist, zeigt zwei große Anhäufungen (dick umkreist), auch Cluster genannt. Die größte Anhäufung (linke Umrandung) beschäftigt sich hierbei mit der theoretischen Forschung (*MapReduce*), mit Parallelisierungstechnik und *Cloud Computing*. Die zweite große Anhäufung (rechte Umrandung) beschäftigt sich mit den Möglichkeiten von *Big Data* in der Wissenschaft und Industrie.

[119] Die beiden oberen Publikationen behandeln inhaltlich ein ähnliches Themengebiet im Rahmen des Supply Chain Managements und referenzieren beide auf die unten aufgeführte Referenz. Sie sind somit durch mindestens diese Referenz gekoppelt und werden durch eine Kante verbunden. Je identischer die Referenzlisten, desto stärker die Kante.

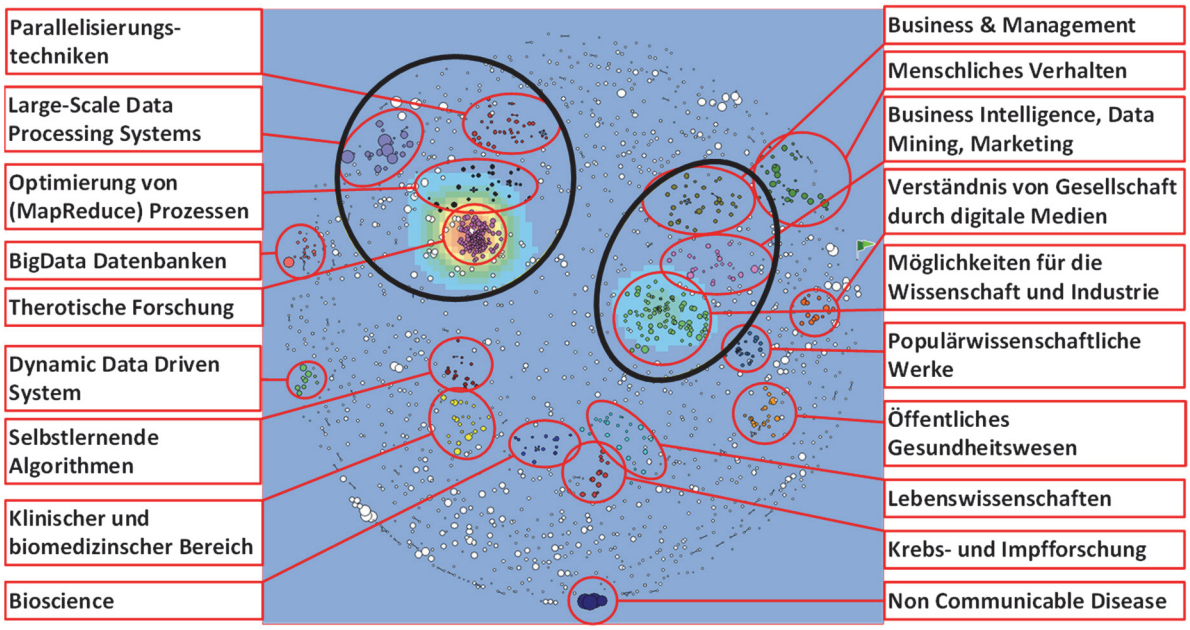


Abbildung 37: Forschungsfronten aus BibTechMon4[120]

Weitere kleinere Forschungsfronten wurden identifiziert und sind Tabelle 10 zusammengefasst. Dabei ist festzustellen, dass *Big Data* in sehr vielen verschiedenen Wissenschaftsgebieten und industriebezogenen Forschungsgebieten vertreten ist.

[120] BibTechMon4 i.V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Netzwerk aus 1847 Knoten, in dem die 200 stärksten Kanten sichtbar sind. Die Knotengröße beschreibt die Anzahl der Referenzen der jeweiligen Publikation.

Forschungsfronten	Anzahl der Publikationen	Anteil in %
MapReduce, Parallelisierung, Cloud - theoretische Forschung	141	23,1%
Big Data - Möglichkeiten für die Wissenschaft und Industrie; Veränderungen durch *Big Data* in der Industrie	90	14,8%
Parallelisierungstechniken und deren Optimierung bzw. Weiterentwicklung	39	6,4%
Business & Management, vor allem Process Mining	37	6,1%
Optimierung von (*MapReduce*) Prozessen und Inkludierung von Social Media	36	5,9%
Menschliches Verhalten ergründen und voraussagen, auch Gruppenverhalten	34	5,6%
Populärwissenschaftliche Werke die sich mit dem potentiellen Nutzen und Eingriff in der Privatsphäre in allen Bereichen in der Zukunft beschäftigen	25	4,1%
Lebenswissenschaften - Biologie, Genforschung, Suchtforschung	25	4,1%
Öffentliches Gesundheitswesen - Analyse von Userdaten, BigData für Epedemievorhersagen	24	3,9%
BigData im klinischen und biomedizinischen Bereich	24	3,9%
Business Intelligence, Data Mining, Analyse von Daten für Marketingzwecke	23	3,8%
Large-Scale Data Processing Systems	22	3,6%
Big Data in Bioscience	18	3,0%
Selbstlernende Algorithmen - Analyse von visuellen, audio und anderen Daten	16	2,6%
Analyse Methoden und Algorithmen für die Auswertung von BigData Datenbanken	15	2,5%
Verständnis von Gesellschaft durch digitale Medien	14	2,3%
BigData für die Krebsforschung und Impfforschung	13	2,1%
Erik Blasch - Dynamic Data Driven System	9	1,5%
Non Communicable Diseases	5	0,8%

Tabelle 10: Forschungsfronten des Suchbegriffes „Big Data“[121]

Der Vergleich von Wissensbasen und Forschungsfronten ermöglicht es an dieser Stelle, die Verknüpfung von Vergangenheit und Zukunft auf Basis der gewonnenen Erkenntnisse herzustellen. Daraus werden grundsätzliche Basisthemen abgeleitet, die für den Einstieg eines Unternehmens in die Thematik *Big Data* unerlässlich sind. Darauf aufbauend kann die Entwicklung einzelner Themenbereiche für die Zukunft prognostiziert werden.

Beim Vergleich der Wissensbasen (vgl. Tabelle 7) mit den Forschungsfronten (vgl. Tabelle 10) ist festzuhalten, dass die inhaltlichen Schwerpunkte gleich sind. Der Anteil der theoretischen Forschung geht in den Forschungsfronten zurück, was darauf zurückzuführen ist, dass der theoretische Unterbau

[121] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015

mit der Zeit mehr und mehr etabliert ist und der Fokus auf die Anwendungsfelder gelegt wird. Eindeutig identifizierbare Anwendungsfelder sind die Parallelisierungstechniken und deren Optimierung. Auch wird das Forschungsgebiet für *Big Data* in der Industrie immer wichtiger. Es kann in drei verschiedene Forschungsfronten unterschieden werden. Erstens in die allgemeinen Möglichkeiten von *Big Data* für die Industrie. Zweitens in *Big Data* für *Business & Management* Anwendungen, wie z. B. mit Hilfe von *Process Mining*. Drittens in *Business Intelligence* Anwendungen, welche in Teilen *Process Mining* verwenden, darüber hinaus aber auch andere Analysen beinhalten. Diese drei Themengebiete sind räumlich sehr nahe im Forschungsfrontennetzwerk (vgl. Abbildung 37).

Eine weitere Möglichkeit, die Forschungsfronten darzustellen ist in Abbildung 38 zu sehen. Dort wurde eine dreidimensionale Darstellung des Netzwerkes generiert. Auffallend ist hier, dass die Spitze der Forschungsfront theoretische Forschung um ein Vielfaches höher ist als alle anderen Erhebungen im Netzwerk. Die Höhe der Spitzen und Erhebungen wird definiert durch die Anzahl der Artikel pro Flächeneinheit, gewichtet nach ihrer Zentralität. Abbildung 38 zeigt die beiden größten Forschungsfronten in einer 3D-Darstellung.

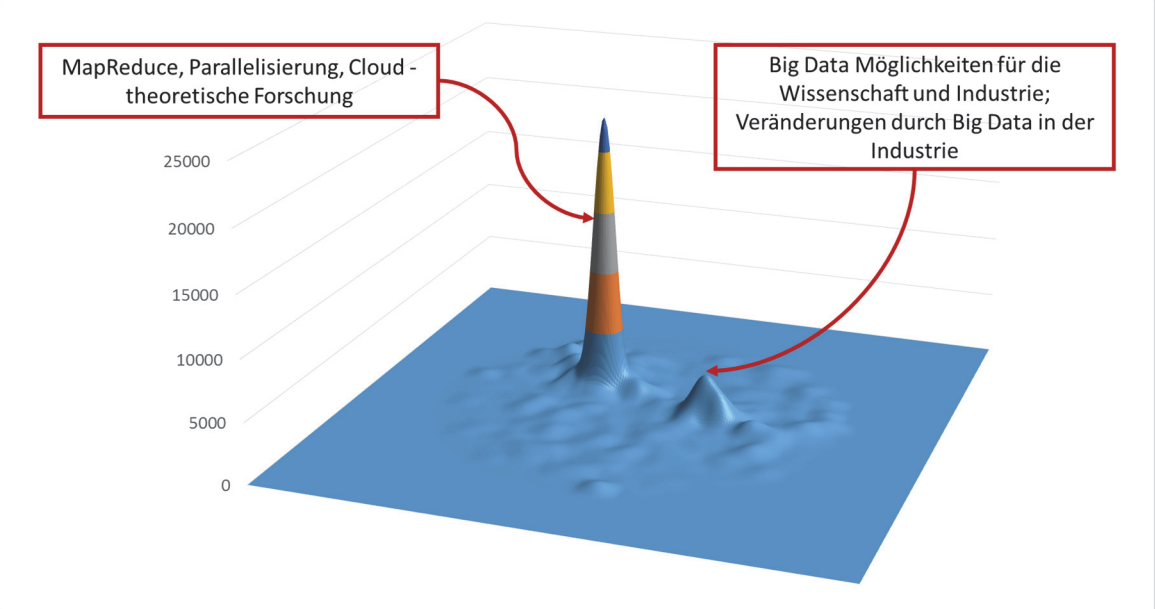


Abbildung 38: 3D Darstellung der Forschungsfronten[122]

Für eine detailliertere Betrachtung der Forschungsfronten ist in Abbildung 39 eine dreidimensionale Darstellung mit höherer Auflösung zu sehen. So konnten weitere wichtige Forschungsfronten identifiziert werden.

[122] Blaue Zonen zeigen keine oder niedrige gewichtete Anzahl gekoppelter Publikationen, rote Zonen weisen eine hohe gewichtete Anzahl von Publikationen aus; Gewichtung durch die Stärke aller Verbindungen je Publikation (Jaccard-Index der Anzahl gemeinsamer Referenzen). Auffällig ist, dass die Spitze „theoretische Forschung“ sehr viel größer ist als alle anderen Spitzen. Die Höhe ergibt sich durch die Anzahl der Artikel pro Flächeneinheit, gewichtet nach ihrer Zentralität.

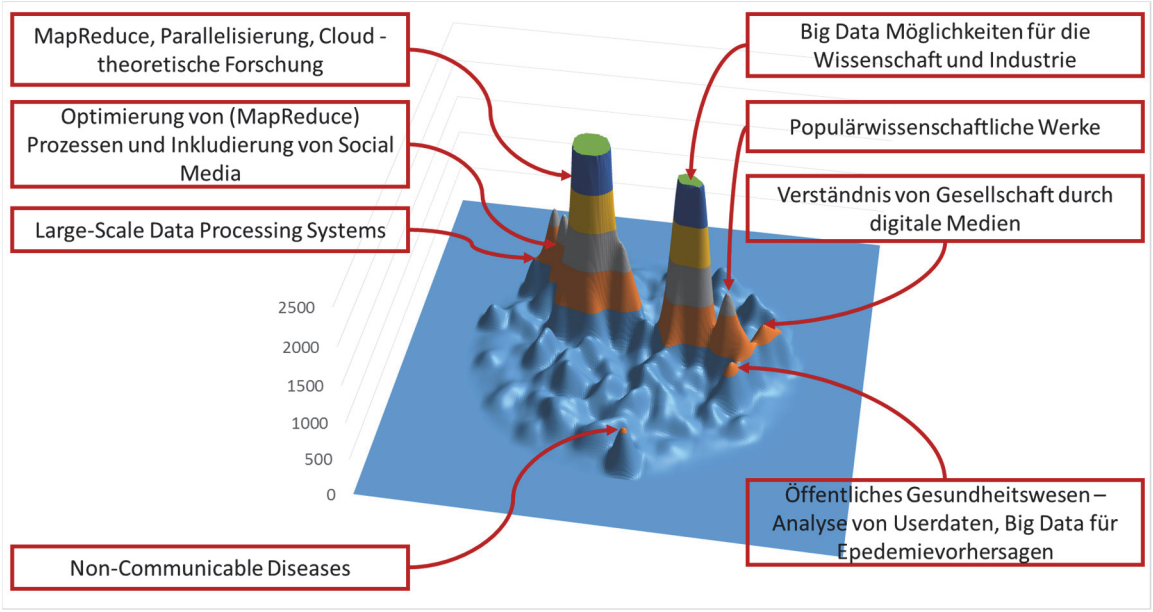


Abbildung 39: 3D Darstellung der Forschungsfronten, Vergrößerung um eine Zehnerpotenz[123]

Im Vergleich zu Abbildung 38 werden in Abbildung 39 sieben zusätzliche Forschungsfronten sichtbar. Während die Erhebungen links in Abbildung 39 erneut die theoretischen Themen ansprechen, finden sich auf der rechten Seite auch Schnittstellen zu den Themen der Populärwissenschaften, des Gesundheitswesens und dem Verständnis der Gesellschaft durch digitale Medien. Innerhalb dieser Fronten befinden sich generelle Veröffentlichungen zum Themengebiet *Industrie 4.0* ohne auf genaue Implementierungen einzugehen.

Die Betrachtung der zeitlichen Entwicklung einzelner Forschungsfronten ist sehr interessant, um die Aktualität der einzelnen Themengebiete zu bewerten. Außerdem können Abnahmen der Forschungsaktivitäten innerhalb einer oder mehrerer Forschungsfronten beobachtet werden. Dies legt eine Sättigung der Forschungsaktivitäten in diesem Thema nahe oder aber ein Abschwächen der Nachfrage an die Forschungseinrichtungen. Dies wird in Tabelle 11 anhand der Publikationsanzahl pro Jahr und Forschungsfront verdeutlicht.

[123] Die maximale Höhe wurde auf 2500 begrenzt, um kleine Forschungsfronten sichtbar zu machen. Vgl. auch Fußnote 122.

Forschungsfronten	2011	2012	2013	2014	2015	Summe
Nicht zugeordnete Publikationen	15	110	417	593	76	1237
Theoretische Forschung		14	72	51	4	141
Big Data - Möglichkeiten für die Wissenschaft und Industrie		4	40	39	7	90
Parallelisierungstechniken		4	18	15	2	39
Business & Management		3	13	19	2	37
Optimierung von (*MapReduce*) Prozessen und Social Media		4	12	20		36
Menschliches Verhalten	1	2	16	13	2	34
Lebenswissenschaften		1	8	8	3	25
Populärwissenschaftliche Werke			6	19		25
Big Data im klinischen und biomedizinischen Bereich		1	6	15	1	24
Öffentliches Gesundheitswesen			4	15	5	24
Business Intelligence, Data Mining, Marketing, Social Networks		2	3	16	2	23
Large-Scale Data Processing Systems		1	14	6		22
Big Data in Bioscience	3		7	8		18
Selbstlernende Algorithmen			4	12		16
Analyse Methoden und Algorithmen für die Auswertung von *Big Data* Datenbanken		1	3	8	2	15
Verständnis von Gesellschaft durch digitale Medien			6	8		14
Big Data für die Krebsforschung und Impfforschung			3	7	2	13
Erik Blasch - Dynamic Data Driven System		1	4	4		9
Non Communicable Diseases			3	2		5

Tabelle 11: Zeitliche Entwicklung der Publikationen in den Forschungsfronten ab dem Jahr 2011[124]

Die theoretische Forschung hat ab dem Jahr 2013 stark abgenommen, wohingegen in der Forschungsfront „Optimierung von (*MapReduce*) Prozessen und *Social Media"* ein stetiges Ansteigen der Publikationen beobachtet wird. Eine ähnliche Entwicklung ist in den Forschungsfronten „*Big Data* im klinischen und biomedizinischen Bereich" und „Öffentliches Gesundheitswesen", sowie „Business & Management" und „Business Intelligence, Data Mining, Marketing, Social Networks" zu beobachten. Diese Themenbereiche werden somit in Zukunft eine wichtige Rolle bei der Ausgestaltung von unternehmerischen und privaten Tätigkeiten in Verbindung mit *Big Data* erfahren. Aus diesem Grund werden die beiden zuletzt genannten Fronten neben der Forschungsfront „Theoretische Forschung" in den folgenden Unterkapiteln detaillierter beschrieben. Darüber hinaus wird die Forschungsfront „*Big Data* – Möglichkeiten für Wissenschaft und Industrie" im Zuge des *Industrie-4.0*-Fokus beschrieben.

[124] BibTechMon4 i. V. m. Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Die zeitliche Entwicklung wird nur bis 2014 für Prognosen verwendet.

4.3.2.1 Forschungsfront „Theoretische Forschung“

In diesem Unterkapitel soll auf die theoretische Forschung eingegangen werden. Darunter fallen vor allem Veröffentlichungen zum Programmiermodell *MapReduce*, zur Parallelisierung und zum *Cloud Computing*. Die Wichtigkeit dieser Forschungsfront hat im Jahr 2014 etwas abgenommen im Vergleich zum Vorjahr. Dennoch fallen in dieses Gebiet die meisten Veröffentlichungen zum Themengebiet *Big Data*. Der Artikel mit höchster Zentralität, also mit der am höchsten übereinstimmenden Referenzliste, heißt „Cloud Parallel Processing of Tandem Mass Spectrometry Based Proteomics Data“ (Mohammed et. al., 2012). Diese Publikation beschäftigt sich mit der Auswertung von Proteomdaten eines Massenspektrometers. Hierbei werden sehr große Datenmengen mit Hilfe des *Hadoop* Frameworks zerlegt, durchsucht und analysiert. Die Rechenkapazität wird mit Hilfe von *Cloud Computing* bereitgestellt. Hier kann man beispielsweise das Potential von *Big Data* Systeme zur Datenaufarbeitung erkennen. Bei dieser Proteinanalyse kann durch *Big Data*-Techniken die benötigte Zeit von beispielsweise „191:51 Minuten auf 5:42 Minuten reduziert werden"[125]

Der Artikel „Artificial Intelligence and Big Data“ von Daniel E. O’Leary aus dem Jahr 2013 beschäftigt sich mit der künstlichen Intelligenz, welche mit Hilfe von *Big-Data-Technologien* eine intelligente Analyse von Daten und deren strukturierter Interpretation ermöglicht. Dies bedeutet, dass in Zukunft in Unternehmen, vermehrt Entscheidungen aufgrund von Daten getroffen werden können und nicht ausschließlich von Menschen.[126]

4.3.2.2 Forschungsfront „Big Data – Möglichkeiten für Wissenschaft und Industrie“

Der Artikel mit der höchsten Zentralität, ist die Veröffentlichung „Data Science, Predictive Analytics, and *Big Data*. A Revolution That Will Transform Supply Chain Design and Management“[127]. In dieser Veröffentlichung wird das Potential von *Big-Data-Technologien* für das Supply Chain Management (SCM) vorgestellt. Analytische Unternehmen, welche aufgrund von Daten Entscheidungen treffen, seien produktiver und profitabler. In Zahlen ausgedrückt waren solche Unternehmen um 5 % produktiver und 6 % profitabler als ihre Wettbewerber.[128] Der Artikel stellt grundsätzliche Ideen zum SCM im Zusammenhang mit *Big Data* vor und fordert andere Wissenschaftler auf, sich an der Forschung zu beteiligen. In Tabelle 12 sind die wichtigsten Forschungsfragen im Supply Chain Management im Zusammenhang mit *Big Data* aufgelistet.

[125] Vgl. Mohammed et al. (2012)
[126] Vgl. O'Leary (2013), S. 99
[127] Vgl. Waller und Fawcett (2013)
[128] Vgl. McAfee und Brynjolfsson (2012), S. 64

Theory	Research question
Transaction cost economics	How does the existence of *Big Data* affect the reduction in internal transaction costs vis-à-vis externaltransaction costs and how is this affecting the size of logistics organizations and the structure of supply chains?
Resource-based view	Can SCM data science be developed as a resource that is valuable, rare, in imitable, and nonsubstitutable?
Contingency theory	How can *big data* and SCM data science be used by logistics managers to meet internal needs and adjust to changes in the supply chain environment?
Resource dependence theory	How does the ability to use *Big Data* for SCM decisions affect a firm's power in comparison with its suppliers or customers?
Agency theory	How does the proliferation of *Big Data* affect the agency costs associated with the use of third party logistics?
Institutional theory	How do differences in freedom of information between countries affect firms operating under these different institutions in terms of their abilities to leverage *Big Data* in the supply chain?

Tabelle 12: Wichtige Forschungsfragen im Supply Chain Management[129]

Der Artikel „Intelligent Operational Dashboards for Smarter Commerce using Big Data" von Yesudas et al. aus dem Jahr 2014 betrachtet *Business Intelligence* Lösungen im Einzelhandel. Bisher basierten diese Lösungen auf statischen Datenmodellen. Das Benutzen von Streaming Daten wurde bei den bisherigen Lösungen nicht in Betracht gezogen. Ziel der hier vorgestellten *Big Data* Lösung ist durch die Benutzung von Streaming Daten eine Operational Intelligence bereit zu stellen. Entscheidungen können dadurch vermehrt durch nahezu Echtzeitdaten getroffen oder unterstützt werden. Die Implementierung in dieser Veröffentlichung erfolgte mit dem IBM Sterling Order Management System.[130]

4.3.2.3 Forschungsfront „Business & Management"

Diese noch relativ kleine Forschungsfront, welche aus 37 Veröffentlichungen besteht (vgl. Tabelle 10), beschäftigt sich mit *Big Data* Anwendungen und Möglichkeiten im Business- und Managementbereich. Ein Schwerpunkt wird hier vor allem auf Process Mining gelegt. Dazu beschreibt die Veröffentlichung „Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining" von van der Aalst aus dem Jahr 2013 ein Process Mining Modell vorgestellt, welches Geschäftsprozesse (basierend auf Ereignisdaten) entdecken, analysieren und verbessern soll. Das in diesem Paper vorgestellte Modell der Process Cubes zerlegt Prozesse in einzelne Würfel, die auch unterschiedliche Dimensionen besitzen können. Diese kleineren Würfel können parallel analysiert werden und liefern später ein Prozessgesamtbild zurück.

Eine weitere erwähnenswerte Veröffentlichung in dieser Forschungsfront beschäftigt sich mit dem Costumer Relationship Management und speziell mit der Abwanderung von Kunden. "Customer event history for churn prediction. How long is long enough?" wurde im Jahr 2012 von Ballings und van den

[129] Vgl. Waller und Fawcett (2013), S. 83
[130] Vgl. Yesudas et al. (2014)

Poel publiziert. Die Abwanderung von Kunden vorherzusagen ist Ziel von vielen *Big Data* Lösungen in diesem Bereich und zählt zu den *Predictive Analytics* Anwendungen. Wichtig ist den Autoren dieses Werkes die Effizienz der Algorithmen. So wird gezeigt, dass mit einer geringeren Datenmenge als bisher ebenso genaue Aussagen über die Kunden getroffen werden können, was eine Effizienzsteigerung für Customer Relationship Departments in Unternehmen verspricht.

4.3.2.4 Forschungsfront „Business Intelligence"

Diese Forschungsfront umfasst mehrere kleinere Themengebiete, die sich vor allem mit Business Intelligence, Data Mining und Analyse von Daten für Marketingzwecke beschäftigen. Der Artikel „Vergleich von Kompetenzanforderungen an Business-Intelligence- und *Big Data*-Spezialisten" (Debortoli et al. 2014) analysiert die Stellenausschreibungen des Online-Portals monster.com, um Informationen darüber zu gewinnen, welche Anforderung an Bewerber im Bereich *Business Intelligence* und *Big Data* gestellt werden. Dies resultiert aus der Diskussion in Praxis und Wissenschaft, welche Unterschiede zwischen *Business Intelligence* (BI) und *Big Data* (BD) vorliegen. Dies ist auch für Unternehmen im Bereich *Industrie 4.0* sehr interessant, um herauszufinden welche Anforderungen an Mitarbeiter in der Zukunft gestellt werden und gestellt werden können. In beiden Bereichen, sowohl in *Big Data* als auch *Business Intelligence*, ist Businesswissen genauso wichtig wie technisches Wissen. Kompetenz bezieht sich im *Business Intelligence* Bereich vor allem auf die Fähigkeiten und Wissen über Produkte kommerzieller Softwareanbieter, während im *Big Data* Bereich mit Kompetenz vorwiegend die Fähigkeiten und das Wissen über Individualsoftware und die Anwendung statistischer Methoden verbunden wird.

Die Datenanalyse für Marketingzwecke ist in dieser Forschungsfront ebenfalls Thema. Eine Publikation ist „Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of *Big Data* Using Latent Dirichlet Allocation" von Tirunillai aus dem Jahr 2014. Ziel des Artikels ist die Identifikation von Schlüsseldimensionen der Kundenzufriedenheit auf Basis der von Internetbenutzern generierten Daten über Produkte. Für den Computermarkt dominieren objektive Dimensionen, beim Schuhmarkt eher subjektive Dimensionen.[131]

Im Bereich der Social Network Analyse ist die Publikation „Semantic Social Network Analysis for the Enterprise" von Chelmis et al. aus dem Jahr 2013 hervorzuheben. Ziel dieser Arbeit ist es, die Unternehmenskommunikationskanäle, wie z. B. Emails, Chats, Diskussionsforum etc. auszuwerten, um auf Probleme, Beschwerden, Meinungen, Trends usw. im Unternehmen aufmerksam zu werden. Dies soll vor allem betrachtet werden, wenn sich Geschäftsprozesse ändern, da es gegen Änderungen häufig Widerstand gibt. Darüber hinaus kann eine solche Rückmeldungskultur dazu führen, dass das Wissensmanagement der Unternehmen mehr Informationen über aktuelle Vorgänge besitzt. So kann die Suche nach unternehmensinternen Experten oder die Identifikation von Trends unterstützt werden. Verbindungen zwischen einzelnen Mitarbeitern, aber auch Abteilungen und Gruppen ist klar identifiziert und abbildbar, was wiederum bestehenden und neuen Mitarbeitern den Wissenstransfer erleichtert.

In Abbildung 40 ist die dynamische und statische Modellierung von verschiedenen Datenquellen nach Chelmis in einem Unternehmen dargestellt. Dieses Modell wurde bereits in Unternehmen

[131] Vgl. Tirunillai und Tellis (2014), S. 463

implementiert. Das Ergebnis dieses Modells ist im rechten Kasten der Grafik dargestellt, dort werden die Empfehlungen aufgrund der Datenanalysen ausgesprochen.

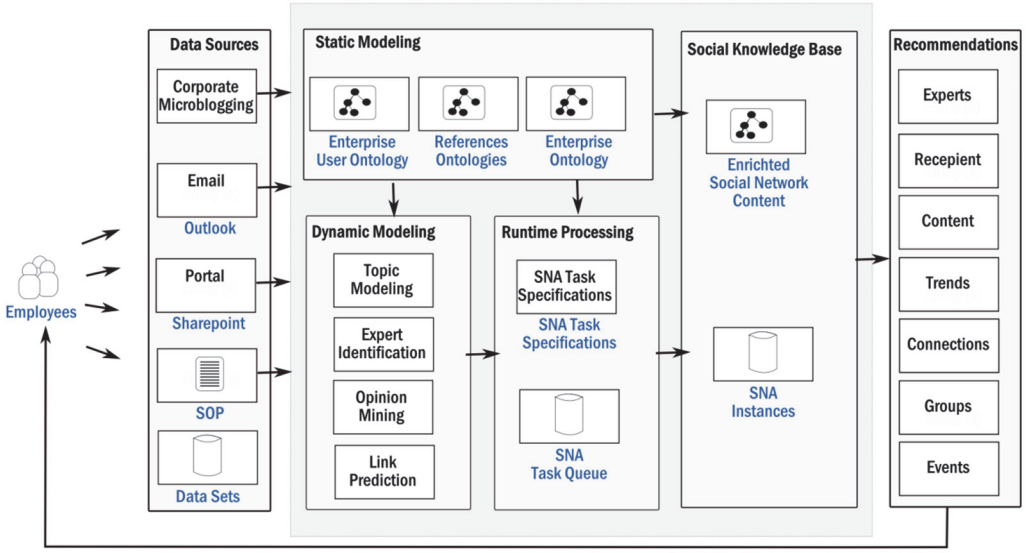


Abbildung 40: Modellierung verschiedener Datenquellen in einem Unternehmensnetzwerk[132]

4.4 Forschungsportfolio: Positionierung der Forschungsfronten in der Wissenschaft

Zusammenfassung

Der überwiegende Teil der für die *Industrie 4.0* relevanten Forschungsfronten findet sich in der BCG-Matrix im Bereich der „Question Marks“ und an der Grenze zu den „Stars“ wieder. Dies legt nahe, dass in der Zukunft Themen wie selbstlernende Algorithmen oder aber der Einsatz von *Big-Data-Technologien* im Marketing eine bedeutende Rolle spielen werden.

Die Popularität der theoretischen Forschung am Übergang von den „Stars“ zu „Cash Cows“ reduziert sich minimal. Dies ist auf die stagnierende Zahl der Erkenntnisse seit 2013 zurückzuführen. Dasselbe gilt für die Forschungsfront *„Big Data* Möglichkeiten für die Wissenschaft und Industrie“, in der generelle Einsatzgebiete von *Big-Data-Technologien* identifiziert wurden.

Das Forschungsportfolio stellt die Forschungsfronten in einen Zusammenhang mit ihrer zeitlichen Entwicklung. Dabei gilt es vor allem die Bedeutung der einzelnen Fronten in der momentanen Wissenschaft zu bewerten, sowie mögliche Veränderungen in der Popularität zu identifizieren.

Die Portfoliodarstellung in Abbildung 41 zeigt die in Kapitel 4.3.2 identifizierten Forschungsfronten eingeordnet in die BCG-Matrix, um das oben genannte Ziel grafisch darzustellen. Dabei wurden lediglich die für die *Industrie 4.0* relevanten Forschungsfronten dargestellt.

[132] Vgl. Chelmis et al. (2013), S. 17

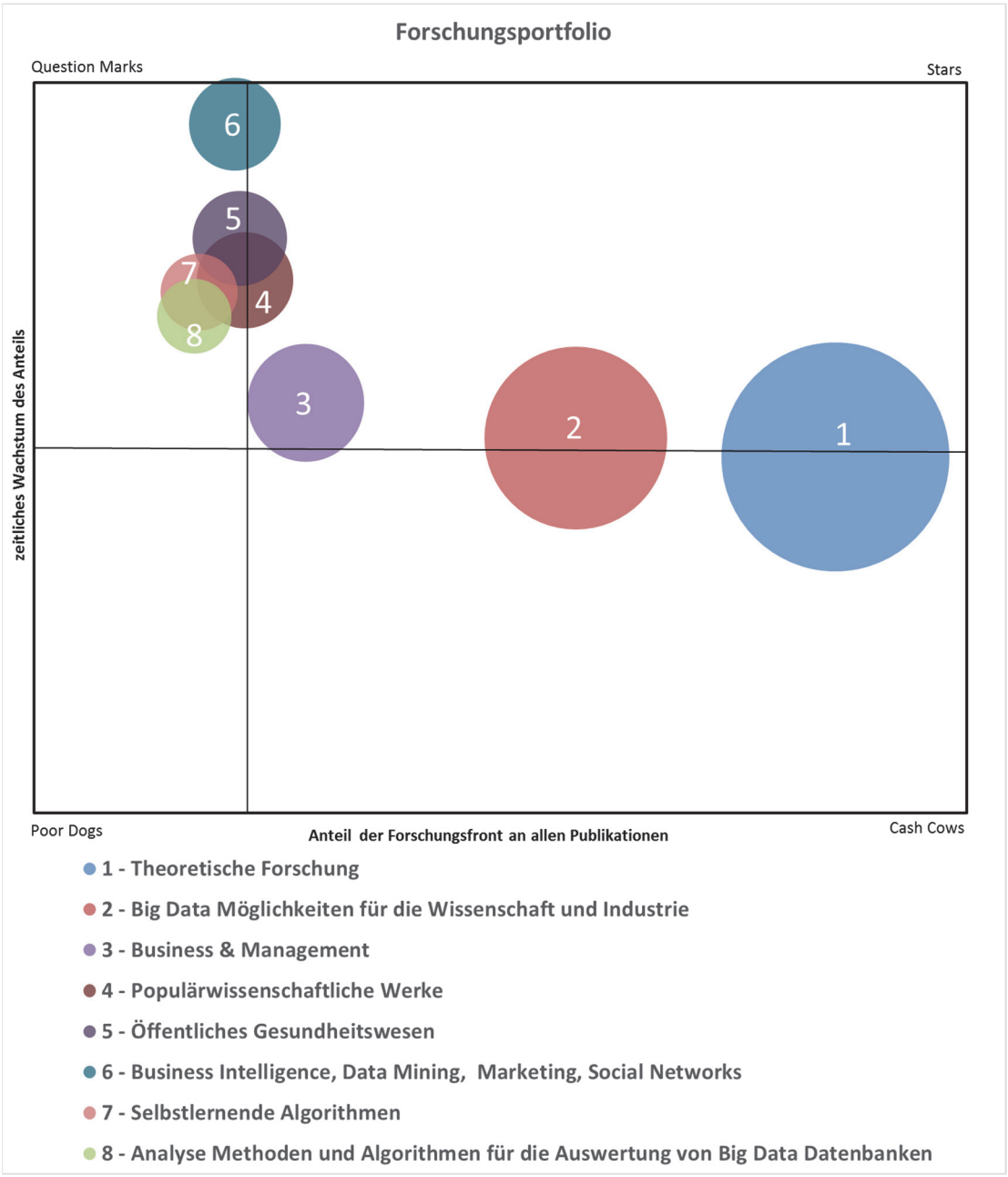


Abbildung 41: Forschungsportfolio[133]

Abbildung 41 zeigt, dass die theoretische Forschung den größten Anteil an Publikationen ausmacht. Allerdings ist das Themengebiet rund um die Möglichkeiten von *Big-Data-Technologien* für die Industrie und Wissenschaft ebenfalls wichtig, was sich an der zweithöchsten Anzahl aller Publikationen zeigt. Zusätzlich befindet es sich am Übergang von „Stars" zu „Cash Cows". Auch zu sehen ist, dass die Wachstumsrate für diese zwei zuvor genannten Forschungsfronten konstant geblieben ist.

[133] Thomson Reuters Corporation (Science), Web of Science Core Collection, Zugriff über Universität Ulm am 13.03.2015; Abszisse: relativer Artikelanteil der Forschungsfront an der Gesamtzahl von 1847 untersuchten Artikeln.; Mantisse: Publikationswachstum der Forschungsfronten innerhalb eines zeitlichen Intervalls. Aufgrund des erst relativ neuen Forschungsfeldes, wurden lediglich die Änderungen zwischen 2013 und 2014 betrachtet.; Kreisgröße: Anzahl der Publikationen der jeweiligen Forschungsfront im gesamten Zeitraum bestimmt.

Forschungsfronten mit sehr hohen Wachstumsraten sind beispielsweise *Business Intelligence*, *Data Mining*, *Marketing*, *Social Networks*, Öffentliches Gesundheitswesen und populärwissenschaftliche Werke. Im Zusammenhang mit *Industrie 4.0* ist auch das Thema der selbstlernenden Algorithmen sehr interessant, diese Forschungsfront verzeichnet ebenfalls eine hohe Wachstumsrate. Daraus kann gefolgert werden, dass viele der dort behandelten *Big-Data-Technologien* und Implementierungen bald den Sprung zu den Stars schaffen und somit unverzichtbar für Unternehmen werden.

Ein eher stagnierendes oder sogar schrumpfendes Thema ist das Thema der *Non-Communicable Diseases*. Andere Themengebiete, die vor allem theoretischer Natur sind, sind auch stagnierend oder sogar schrumpfend. Eines davon wäre die Forschungsfront *Large-Scale Data Processing Systems*, diese ist aber sehr mit der theoretischen Forschung korreliert. Aus diesem Grund kann nicht mit Sicherheit gesagt werden, ob die behandelte Thematik wirklich schrumpfend ist.

5 Szenariotechnik zu Big Data in der Industrie 4.0

Die Szenariotechnik wird im folgenden Abschnitt dazu verwendet, Erfolgspotenziale und Risiken für die Zukunft des Einzelhandels zu prognostizieren. Die Entwicklungsmöglichkeiten des Einzelhandels sollen detailliert im Rahmen mehrerer Szenarien aufgezeigt werden. Im Sinne der Technologievorausschau werden zur Ableitung strategischer Maßnahmen im Vorfeld emergente Technologien abgeleitet. Deren Entwicklungspotenzial wird in den einzelnen Szenarien bewertet (robuste Technologien).

Aus den Zukunftsbildern werden in Kapitel 6 robuste Technologien für drei Unternehmen abgeleitet. Die Technologien zielen darauf ab, ein nachhaltiges Geschäftsmodell für jedwedes Szenario zu gestalten.

Im weiteren Verlauf erfolgt die Durchführung der Szenariotechnik im Kontext der Technologievorausschau und in Kombination mit der Bibliometrie dem Ablauf nach Stelzer et al. (2015).[134]

5.1 Definition des Untersuchungsgegenstandes

Zusammenfassung

Der Begriff *Big Data* wird in zehn Entitäten unterteilt und beschrieben, um so Verknüpfungen zwischen den real existierenden Funktionen und den abstrakten Funktionen aufzuzeigen. Zu den realen Funktionen gehören *Lizenzen*, die *Logistik-*, *Industrie* und *Handelsunternehmen*, sowie die *IT-Dienstleister* und *Forschungsinstitute*. Die abstrakten Funktionen lassen sich beschreiben als die *Analyse großer Datenmengen in Echtzeit*, der *Datenzuwachs (Vielfalt)* und die *Datensicherheit*, sowie die *Effizienzsteigerung*. Diese Entitäten bilden die Verständnis-grundlage für die weiteren Schritte im Rahmen der Szenariotechnik.

Der Begriff *Big Data* wird für viele Phänomene verwendet, wie bereits in Kapitel 2.1 beschrieben ist. *Big Data* lässt sich mit folgendem Zitat prägnant zusammenfassen:

> *"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." - Peter Sondergaard*[135]

Zur besseren Verständlichkeit wird der Begriff in Entitäten zerlegt und die Beziehungen zwischen den selbigen beschrieben. Entitäten werden nach Chen 1976 in einem Entity-Relationship-Model dargestellt.[136] Hierbei werden Entitätstypen, Attribute und Beziehungen in einem semantischen Netz dargestellt (Abbildung 42).

[134] Vgl. Stelzer et al. (2015)
[135] Vgl. Gartner (2013)
[136] Vgl. Chen (1976)

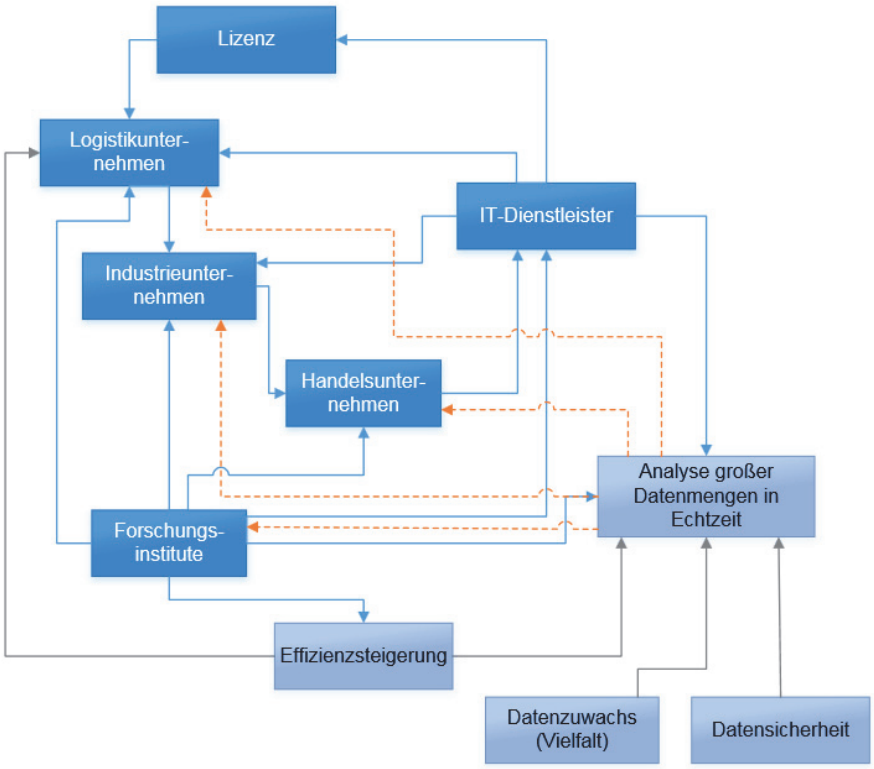


Abbildung 42: Semantisches Netz zu Big Data mit zehn Entitäten[137]

Lizenzen sind für Firmen, die ein oder mehrere Datenanalyse-Tools entwickelt haben eine wichtige Einnahmequelle. Dazu zählen beispielsweise auch Patent- oder Schutzrechtsgebühren, sowie Nutzungsgebühren. Die *Lizenzen* werden als Schnittstelle zwischen den *IT-Dienstleistern* und den *Logistikunternehmen* angesehen werden. In einer erweiterten Betrachtung lassen sich Verknüpfungen zwischen den *Logistik-*, *Industrie-* und *Handelsunternehmen* definieren. Als Attribute finden sich die Reichweite der *Lizenzen* (staatlich oder privat), die Art (Verträge oder Richtlinien) und der Einsatzbereich (Branchenspezifisch oder übergreifend). Der Wertebereich dieser Entität variiert in Aufwand und Komplexität, sodass alle möglichen Ausprägungen an dieser Stelle nicht aufgezählt werden.

Die **Logistikunternehmen** agieren zwischen unterschiedlichen Branchen und haben vielfältige Beziehungen innerhalb des semantischen Netzes. In ihren Aufgabenbereich fallen die Beschaffung, die Lagerung und der Transport von allen Arten von Gütern und Materialien. So kommen sie direkt mit den *IT-Dienstleistern* oder indirekt über *Lizenzen* mit ihnen in Kontakt. Ebenso haben *Forschungsinstitute* einen Einfluss. Die *Effizienzsteigerung* kann entweder durch eine Studie von *Forschungsinstituten* an die *Logistikunternehmen* herangetragen werden oder aber sie entstehen direkt in dem Unternehmen, z. B. durch die Nutzung von Data Analytics Software. Die *Logistikunternehmen* selbst sind darüber hinaus ein wichtiger Partner für die *Industrieunternehmen*, da sie den Fluss von materiellem Input für diese gewährleisten. Die Ausprägung der *Logistikunternehmen* ist innerhalb der Attribute eindeutig definiert.

---

[137] Die durchgezogenen Pfeile stehen für direkte Beziehungen, während die unterbrochenen Pfeile die Wichtigkeit der Rückkopplung zwischen den Unternehmen und den Datenanalysen darstellen. Mit weißer Schrift sind die real existierenden Entitäten gekennzeichnet. Die Entitäten mit schwarzer Schrift sind abstrakter Natur.

Eben jene **Industrieunternehmen** sind als Massenhersteller zu verstehen. Sie versorgen den Markt mit kommerzialisierten Massenprodukten und stellen z. B. im Einzelhandel eine Grundversorgung der Bevölkerung sicher. Daraus ergibt sich eine enge Verbindung zwischen den vorgelagerten *Logistikunternehmen* und den nachfolgenden *Handelsunternehmen*, die später beschrieben werden. *Industrieunternehmen* stellen eine wichtige Schnittstelle in der Bereitstellungskette von Gütern dar. Aus diesem Grund gibt es viele Eigenschaften, die zur Absicherung dieses Ziels dienen. Dazu dienen die wachsende Automatisierbarkeit, eine Gewährleistung der reibungslosen Produktion, Arbeitsteilung, Trennung von Produktion und Erholung, Vereinbarkeit von Familien und Karrieren und eine kontinuierliche technologische Weiterentwicklung. Die Wertebereiche im Bereich Trennung von Produktion und Wohnort reichen von einem Betriebsgelände mit integrierten Wohnhäusern bis hin zu ausgelagerten Produktionsstandorten. Innerhalb des Attributs Arbeitsteilung kann eine Ausprägung eine enge Abstimmung zwischen einzelner Kollegen mit den gleichen Verantwortungsbereichen sein oder aber ein stringenter Schichtplan, der den Arbeitnehmern klare Arbeitszeiten vorgibt.

Als letzte Unternehmenssparte sind die **Handelsunternehmen** zu nennen. Es geht um den gewinnbringenden Vertrieb von Waren, in dem diese Unternehmen Waren direkt einkaufen und ohne Weiterverarbeitung an unterschiedliche Abnehmer verkaufen. Je nach konkretem Produktumfang, werden in Großhandel, Einzelhandel, Außengroß- oder Außeneinzelhandel unterschieden. Ebenso wie zuvor die *Industrieunternehmen* besitzen auch die *Handelsunternehmen* enge Verbindungen zu *Forschungsinstituten* und *IT-Dienstleistern*. Eigenschaften dieser Entität sind hauptsächlich die Kundegruppen, denen die Unternehmen ihre Dienste anbieten. Wie schon anhand der Synonyme zu erkennen ist, werden in In- und Ausland, sowie in Absatzgröße unterschieden. Dabei lässt sich der Wertebereich bereits nahezu identisch zu den Bezeichnungen der Attribute definieren.

Bevor die Entitäten mit einem größeren Anteil der analytischen Funktionsbereiche vorgestellt werden, schließt sich der Unternehmensbereich mit der Beschreibung der Entität **IT-Dienstleister**. In dieser Entität handelt es sich um jegliche Unternehmung, die Software zur Datenanalyse bereitstellt und entwickelt. Es ist eine klare Trennung zu den Vertriebsunternehmen zu erkennen, da die *IT-Dienstleister* als Programmierer und Berater für Firmen zur Verfügung stehen. Sie entwickeln somit direkt am Kunden und liefern individuelle Lösungen ebenso wie bereits existierende Software-Pakete. Wie bereits erkennbar, sind die *IT-Dienstleister* im Bereich der Unternehmen der Wissenspool ohne den keine sinnvolle Datenverarbeitung möglich wäre. Dadurch ergeben sich Beziehungen zu allen bisher genannten Entitäten und der *Datenanalyse* selbst. Die Eigenschaften solcher Dienstleister reichen von der Produkttiefe, über die Art und Weise der Kundenintegration, bis hin zur Produktionsstätte. Dabei können Lösungen direkt vor Ort bei und mit dem Kunden entwickelt werden oder sie werden nur installiert.

Aufgrund des wachsenden Interesses an dem Themengebiet *Big Data* und der sich daraus ergebenden Wachstumsmöglichkeiten für Unternehmen und ganze Industrien, beschäftigen sich **Forschungsinstitute** mehr und mehr mit dieser Thematik. In diesem Zusammenhang beinhalten diese auch Erkenntnisse von Unternehmensberatungen, die Forschungsberichte oder Studien zu relevanten Themen veröffentlichen. Hierbei sei dennoch erwähnt, dass eine Quellenüberprüfung grundlegend der erste Schritt bei der Informationsbeschaffung sein sollte. Aufgrund der unvorstellbaren Datenmenge und -vielfalt benötigen bis auf die Entitäten *Lizenz*, *Datenzuwachs* und *Datensicherheit* alle anderen einen Input von den *Forschungsinstituten*. Studien und Forschungsberichte behandeln neben theoretischen Erkenntnissen und der Entwicklung des *Datenwachstums* auch Bereiche der

Wertschöpfung oder der Wiederverwendung von Daten in ganz anderen Zusammenhängen, als ursprünglich angestrebt. Daraus lassen sich eine breite Masse an Eigenschaften und ein unbeschränkter Wertebereich ableiten.

Die letzten vier Entitäten behandeln abstraktere Bereiche. Dazu gehört die treibende Kraft hinter dem Betrachtungsgegenstand; der **Datenzuwachs** und die **steigende Vielfalt an Datenformaten**. Hiermit ist zum einen die Zunahme der speicherfähigen Daten durch Interaktionen im Internet und sozialen Plattformen gemeint und zum anderen die Gänze der unterschiedlichen Dateiformate, unter denen verschiedenste Informationen abgespeichert werden können. Die Begriffe *Datenzuwachs* und *Datenvielfalt* sind also entsprechend der Definition der BITKOM (vgl. Abbildung 3) zu verstehen. Diese Entität hat vorrangig nur Einfluss auf die *Analyse der Daten in Echtzeit*, da mit steigender Menge und Variantenzahl der Formate auch die Serverleistung ansteigen muss. Die Eigenschaften des Datenzuwachses beinhalten neben der Größe der Datenmenge auch den Verfügbarkeitsradius der Daten und die unterschiedlichen Quellen der Daten bezogen auf die Dateiformate. So wurden bereits zu Beginn 2013 die Mengen neu hochgeladener Fotos auf Facebook mit 7 Petabytes pro Monat angegeben.[138]

Daneben spielt auch die **Datensicherheit** eine wichtige Rolle. Ganz eng am Namen geht es um die Sicherung der Daten auf externen ‚offline' Datenträgern aber auch um die Verantwortung der Unternehmen mit den gewonnenen Daten verantwortungsbewusst umzugehen. Auch hier ist die Beziehung zur *Echtzeitanalyse* die Wichtigste. Von der Erhebung der Daten, über die Analyse bis hin zur Verwertung und Verwendung im operativen und strategischen Management. Die Attribute sind somit die Sicherung der Daten, CRM und SRM, wobei die Wertebereiche in jedem Attribut sehr stark auf die Bedürfnisse der Kunden zugeschnitten werden müssen. Nur so kann eine vertrauensvolle Basis geschaffen werden, da neben den Analysetools selbst auch die Daten selbst das Ziel von Angriffen sein können.[139]

Als dritte abstrakte Entität, ist die **Effizienzsteigerung** zu nennen. Inkludiert sind hier Profitabilitätssteigerungen, Verschlankung von Prozessen innerhalb der Supply Chain und weitere intern und extern bemerkbare Verbesserungen finanzieller und nicht-finanzieller Natur.[140] Effizienzsteigerungen werden aus den Analysen und Studien der *Forschungsinstitute* generiert und als Input für die *Logistikunternehmen* und die *Echtzeitanalysen* verwendet. Die Eigenschaften sind hier die betroffenen Bereiche, SBUs oder Divisions der Unternehmung, die Tiefe der Umstrukturierung um eine Effizienzsteigerung zu erzielen, die betroffenen Hierarchiestufen und die zeitliche und finanzielle Ausdehnung. Dabei sind die Wertebereiche sehr breit gefächert, da Effizienzsteigerungen an jeder Stelle innerhalb einer Organisation und zu nahezu jedem beliebigen Zeitpunkt und kleinem wie auch großem Rahmen umgesetzt werden können.

Als Knotenpunkt der abstrakteren Bereiche gilt die **Analyse größerer Datenmengen in Echtzeit**. Hierbei geht es um die IT-gestützte Leistung die aus unterschiedlichsten Quellen erhaltenen Daten nahezu ohne Verzögerung zu verarbeiten und so aufzubereiten, dass eine beliebige, vertrauenswürdige und überprüfbare Entscheidung getroffen werden kann. Dabei steht vor allem die Serverleistung der Rechenzentren im Fokus. Auf die Analysen wirken alle Entitäten direkt oder indirekt

[138] Vgl. Hutter (2013)
[139] Vgl. Tankard (2012)
[140] Vgl. Fawcett und Waller (2014)

ein. Dazu gehören neben den drei abstrakten Entitäten auch explizit die *IT-Dienstleister* und die *Forschungsinstitute*. Die Attribute werden als Serverleistung, Servergeschwindigkeit, Algorithmeneffektivität, Realitätsnähe, Störanfälligkeit und Handhabung/Verständlichkeit betitelt. Neben der Hardware müssen die verwendeten Algorithmen einfach zu verstehen und auch zu modifizieren sein, da das gesamte Geschäftsfeld noch in den Kinderschuhen steckt. Schlussfolgernd unterliegt der Wertebereich vor allem bei den Attributen, bei denen Menschen beteiligt sind einem engen Korridor. So muss eine einfache, unkomplizierte, am besten intuitive Handhabung möglich sein. Noch enger ist der Korridor bei den Anforderungen an die Hardware. Die Leistung der Server muss der Entwicklung nicht nur in Bezug auf die Kapazität folgen können sondern auch in Fragen der Schnelligkeit und Sicherheit. Eine parallele Weiterentwicklung in allen Bereichen ist somit ausdrücklich notwendig, wie mit Moore's Law von Brynjolfsson gezeigt werden kann.[141]

Die branchenübergreifende Relevanz von *Big Data* lässt sich bereits aus den Fachrichtungen in Abbildung 27 ableiten. Darüber hinaus zeigt sich die Verknüpfung zwischen dem Forschungsportfolio und dem semantischen Netz mit Hilfe der Cluster 6 und 7 (Data Mining und selbstlernenden Algorithmen) und der abstrakten Entität **Analyse größerer Datenmengen in Echtzeit**.

5.2 Einflussbereiche und Treiber

Zusammenfassung

Für den Begriff *Big Data* ergeben sich aufgrund der breit gefächerten bibliometrischen Analyse drei langfristig wirkende Makro-Einflussbereiche. Diese sind die *Gesellschaft*, die *Wirtschaft* und die *Technik*. Die Entwicklung dieser Einflussbereiche wird jeweils durch zwei bzw. drei bedeutende Treiber dargestellt. Anhand einer dreistufigen Systemanalyse wurden die Treiber *IT-Sicherheit*, gefolgt von der *Gesetzgebung* und dem *Datenschutz* als Schlüsseltreiber identifiziert.

Nachfolgend werden Einflussbereiche definiert und Treiber abgeleitet, die auf den Betrachtungsgegenstand wirken. Dies dient als Grundlage zur Vorausschau, welche innerhalb der Ableitung der Projektionen (vgl. Kapitel 5.31) durchgeführt wird. Die Einflussbereiche und Treiber wurden anhand einer PESTEL-Analyse[142] abgeleitet. Das Ergebnis zeigt Abbildung 43:

[141] Vgl. Brynjolfsson und McAfee (2014)
[142] Vgl. Burt et al. (2006), Walsh (2005)

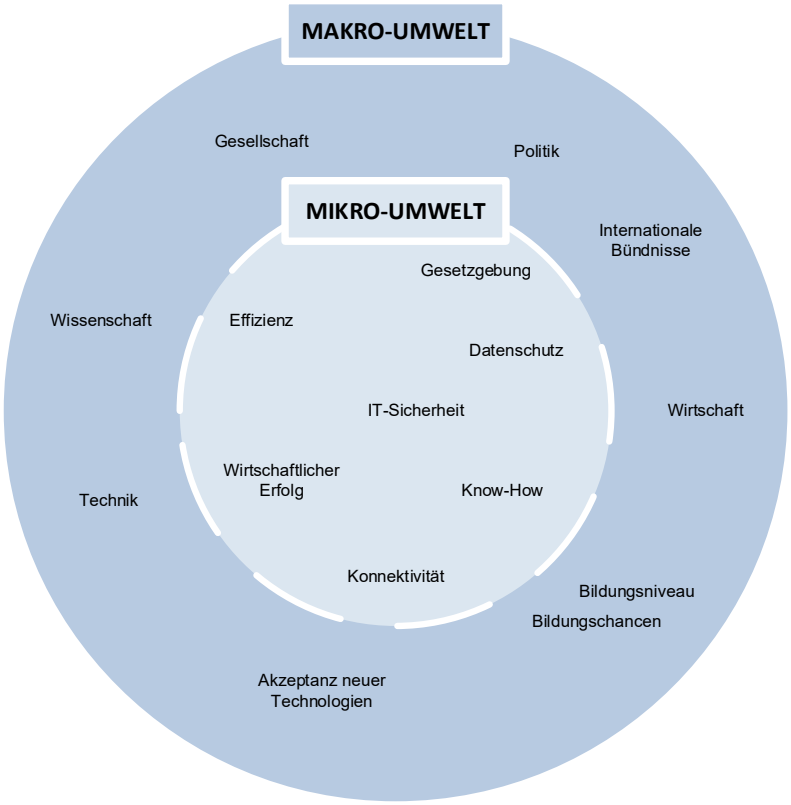


Abbildung 43: Einflussbereiche und Treiber[143]

Die für *Big Data* relevanten Einflussbereiche sind die die *Gesellschaft*, die *Wirtschaft* und die *Technik*. In allen Bereichen treten in der Zukunft größere und kleinere Veränderungen auf. Während die *Gesellschaft* durch den demographischen Wandel und die aktuellen Flüchtlingsströme stark verändert wird, gibt es große Veränderungen in *Wirtschaft* und *Technik* vor allem im Zuge der zunehmenden Digitalisierung und dem Fortschritt von Technologien aufgrund der steigenden globalen Vernetzung, welches für die *Gesetzgebung* Handlungsbedarfe initiiert. Innerhalb dieser Einflussbereiche wurden sieben Treiber identifiziert: *Datenschutz*, *Gesetzgebung*, *Effizienz*, *wirtschaftlicher Erfolg*, *Konnektivität*, *Know-how* und *IT-Sicherheit*.

Im nächsten Schritt wird analysiert, welche Treiber den größten Einfluss auf den Betrachtungsgegenstand bzw. das System haben und welche den geringsten. Dies dient zum tieferen Verständnis von Abhängigkeiten zwischen den Treibern und zu einer möglichen Komplexitätsreduktion. Denn Treiber mit einem sehr geringen Einfluss können im weiteren Verlauf vernachlässigt werden, um die Komplexität der Vorausschau so gering wie möglich zu halten. Methodisch wird dies anhand einer Einflussmatrix durchgeführt. Im paarweisen Vergleich wird die Systemrelevanz der Treiber ermittelt. Tabelle 13 zeigt das Ergebnis.

[143] Einflussbereiche sind Bereiche der Makroumwelt, die Treiber betreffen die Mikroumwelt.

Einfluss auf… / Einfluss von…	Daten-schutz	Gesetz-gebung	Effizienz	wirtschaft-licher Erfolg	Konnek-tivität	Know-how	IT-Sicher-heit	Aktiv-summe
Datenschutz		2	1	2	1	1	2	**9**
Gesetzgebung	3		1	2	1	1	3	**11**
Effizienz	0	0		3	0	0	0	**3**
wirtschaftlicher Erfolg	0	1	1		0	0	0	**2**
Konnektivität	1	2	2	2		2	1	**10**
Know-how	2	0	3	3	1		3	**12**
IT-Sicherheit	3	0	1	3	1	0		**8**
Passivsumme	**9**	**5**	**9**	**15**	**4**	**4**	**9**	**47**

Tabelle 13: Einflussmatrix der Einflussfaktoren/Treiber

Die Aktivsumme gibt Auskunft darüber, wie stark ein Treiber die anderen Treiber beeinflusst. Es zeigt sich, dass der Treiber *Know-how* den größten Einfluss im System hat. Der *Datenschutz* und die *IT-Sicherheit* liegen im Mittelfeld, während die *Effizienz* und der *wirtschaftliche Erfolg* am wenigsten aktiv anderen Treiber beeinflussen. Dieses Ergebnis verwundert nicht, da besonders die beiden zuletzt genannten Treiber eher resultierende Faktoren sind.

Genau diesen Umstand verdeutlicht die Passivsumme. Hier geht es um die Treiber, die am meisten von anderen beeinflusst werden. Wie erwartet liegt der *wirtschaftliche Erfolg* ganz vorne, gefolgt von der *Effizienz*, dem *Datenschutz* und der *IT-Sicherheit*. Stark beeinflusst im Bereich des *Datenschutzes* wird vor allem die Privatsphäre. So wurde früher diese als persönliches Gut verstanden, während in der momentanen Entwicklung viel mehr ein Allgemeingut daraus geworden ist.[144] Das *Know-how*, die *Konnektivität* und die *Gesetzgebung* werden nur sehr eingeschränkt von den anderen hier aufgeführten Treibern beeinflusst. Als gutes Beispiel sei hier die *Gesetzgebung* erwähnt. Da diese von nationalen und globalen Rechtsformen vorgegeben wird, hat eine Unternehmung nur sehr eingeschränkte Beeinflussungsmöglichkeiten.

Ein weiteres Tool zum Verständnis von Abhängigkeiten zwischen den Treibern und ggf. zur Komplexitätsreduktion ist die Relevanzanalyse. Hierbei wird ebenfalls der paarweise Vergleich herangezogen. Allerdings wird nicht nach einer grundsätzlichen Zugehörigkeit gesucht, sondern nach der Priorität eines Treibers im Vergleich zu einem anderen. Ist Treiber i (Zeile) wichtiger als Treiber j (Spalte), so resultiert daraus eine 1, andernfalls eine 0. Aufsummiert ergibt sich die Relevanzsumme der Treiber als Maß für die Prioritätenabfolge. Das Ergebnis ist in Tabelle 14 zu sehen.

[144] Vgl. Cavoukia, Jonas (2012)

Treiber j / Treiber i	Daten-schutz	Gesetz-gebung	Effizienz	wirtschaft-licher Erfolg	Konnek-tivität	Know-how	IT-Sicher-heit	Relevanz-summe
Datenschutz		0	0	0	0	0	0	**0**
Gesetzgebung	1		1	0	0	0	0	**2**
Effizienz	1	0		0	1	0	0	**2**
wirtschaftlicher Erfolg	1	1	1		1	1	1	**6**
Konnektivität	1	1	0	0		0	0	**2**
Know-how	1	1	1	0	1		1	**5**
IT-Sicherheit	1	1	1	0	1	0		**4**

Tabelle 14: Relevanzmatrix der Treiber/Einflussfaktoren

Eindeutig ist der *wirtschaftliche Erfolg* auf der Prioritätsabfolge ganz oben mit sechs Relevanzpunkten, dicht gefolgt vom *Know-how* mit fünf und der *IT-Sicherheit* mit vier Punkten. Die anderen Treiber sind alle auf einer sehr viel niedrigeren Stufe. Zusammengefasst lassen sich die Rankings aus den Ergebnissen der beiden Analysen Tabelle 15 entnehmen. Hierbei ist wichtig zu beachten, dass der erste Platz in den Rankings die höchste Zahl erhält. Nur so kann später im System-Grid eine aussagekräftige Darstellung erreicht werden.

Aktivsumme – Ranking aus Einflussmatrix		Passivsumme – Ranking aus Einflussmatrix		Relevanzsumme – Ranking aus Relevanzmatrix	
7	Know-how	**7**	wirtschaftlicher Erfolg	**7**	wirtschaftlicher Erfolg
6	Gesetzgebung	**4**	Effizienz	**6**	Know-how
5	Konnektivität		IT-Sicherheit	**5**	IT-Sicherheit
4	Datenschutz		Datenschutz	**4**	Gesetzgebung
3	IT-Sicherheit	**3**	Gesetzgebung		Effizienz
2	Effizienz	**1**	Know-how		Konnektivität
1	wirtschaftlicher Erfolg		Konnektivität	**1**	Datenschutz

Tabelle 15: Ranking der Treiber entsprechend der Relevanzsumme

Die Rankings lassen zum einen die Wichtigkeit des *Know-hows* und des *wirtschaftlichen Erfolgs* erkennen, da beide Treiber zweimal unter den ersten Drei rangieren und zum anderen die Berechtigung aller anderen Treiber, da die Summen der beiden Ranglisten Aktivsumme und Relevanz bei den einzelnen Treibern außer dem *Know-how* Werte zwischen sechs und elf liegt.

Als letztes Werkzeug wird das System-Grid angewendet, mit dessen Hilfe auch visuell die beiden zuvor angewendeten Tools verbunden werden. Auf der x-Achse wird der Rangplatz der Aktivsumme aufgetragen, die y-Achse stellt den Rangplatz der Passivsumme dar. Die Größe der Kreise gibt die Relevanz des Treibers auf den Untersuchungsgegenstand an. Dadurch ergibt sich für den *wirtschaftlichen Erfolg* die größte Kreisgröße und für den Datenschutz die Kleinste. Das System-Grid für die Treiber ist in Abbildung 44 abgebildet.

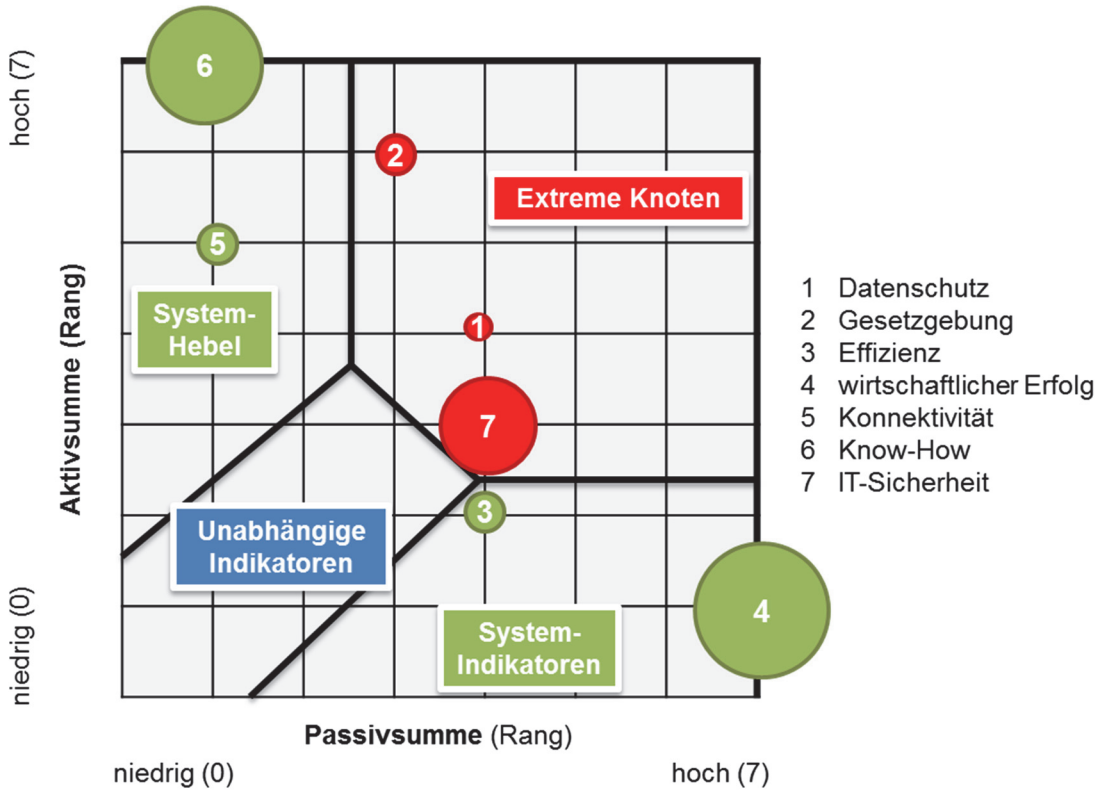


Abbildung 44: System-Grid für den Begriff Big Data

Die drei Treiber mit dem größten Einfluss auf den Untersuchungsgegenstand sind der *wirtschaftliche Erfolg*, das *Know-how* und die *IT-Sicherheit*. Es ist außerdem zu sehen, dass die entsprechenden drei Kreise nicht innerhalb eines der vier Felder des System-Grids auftreten, sondern verteilt sind. Als System-Hebel, welche eine aktive Gestaltung von *Big Data* ermöglichen, sind die *Konnektivität* und das *Know-how* zu identifizieren. Die Gestaltung wird durch das *Know-how* maßgeblich beeinflusst, da nur mit Hilfe einer großen Wissensbasis die aufkommenden Entwicklungen verstanden und sinnvoll in Unternehmen implementiert werden können. Hierbei geht es vor allem um Neuerungen im Softwarebereich, die im Rahmen eines regen Austausches zwischen Forschungsgruppen und Anwendern ohne große Zeitverzögerung über digitale Kanäle oder Messen ausgetauscht werden.

Um frühzeitige Systemveränderungen zu bemerken, müssen die beiden Treiber der *Wirtschaft*, *Effizienz* und *wirtschaftlicher Erfolg*, herangezogen werden. So zeigt sich anhand der Entwicklung von Profitabilität, Kundenzufriedenheit und der getroffenen Entscheidungen am schnellsten, ob die Datenverarbeitung aussagekräftige und vertrauenswürdige Ergebnisse liefert.

Die übrigen drei Treiber (*Datenschutz*, *Gesetzgebung*, *IT-Sicherheit*) im Bereich der extremen Knoten. Sie sind sehr stark in das System *Big Data* eingebunden. Hier zeigt sich eine Tendenz, die sehr verbreitet auch im deutschen Mittelstand und der Bevölkerung verankert ist. Das Sicherheitsbedürfnis ist extrem hoch. Anders als in der globalen Schlagwortanalyse (vgl. Abbildung 30), ist das Thema Sicherheit hier nicht an den Rand gedrängt.[145]

[145] Vgl. Schrage (2012)

Abschließend lassen sich nun alle sieben Treiber für die drei Einflussbereiche bestätigen und fließen in die kommenden Analysen ein. Eine Übersicht ist in Tabelle 16 dargestellt.

Einflussbereiche	Treiber
Gesellschaft	Datenschutz
	Gesetzgebung
Wirtschaft	Effizienz
	wirtschaftlicher Erfolg
Technik	Konnektivität
	Know-how
	IT-Sicherheit

Tabelle 16: Einflussbereiche und Treiber für die Szenarienentwicklung

Die Treiber im Einflussbereich *Gesellschaft* sind vorrangig im Feld der *Gesetzgebung* und des *Datenschutzes* zu sehen. Dies hat den Grund, dass die Digitalisierung in der jetzigen Generation überwiegend als selbstverständlich und alltagsbegleitend angenommen wird. Dazu wurden von der Telekom ein Experiment und eine Umfrage durchgeführt, in denen eine „starke Diskrepanz zwischen den geäußerten Einstellungen der Deutschen zum *Datenschutz* und ihrem tatsächlichen Umgang mit ihren persönlichen Daten" festgestellt wurde. „Dieses Phänomen ist in der wissenschaftlichen Literatur als „Privacy Paradox" bekannt".[146] Daraus können ein unvorsichtiger Umgang und eine fehlende Nachsicht resultieren. Dies ist allerdings im Umgang mit großen Datensätzen, die überwiegend sensible Kundendaten enthalten, nicht zu vertreten, weshalb an dieser Stelle über die Politik und das soziale Verständnis die *Gesellschaft* aufmerksam gemacht werden muss. Des Weiteren gilt es möglichst großflächig auf die Sicherheit und die Handhabung mit Datenmengen einzugehen. So sind Verordnungen im europäischen Raum oder in vergleichbaren Wirtschaftszusammenschlüssen denkbar. Hierfür bedarf es allerdings einer Vorlage, die international ausgearbeitet sein muss, was einen Gebrauch von Data Analytics Anwendungen in weiten Teilen der Welt voraussetzt. Diese Vision wird somit nicht vor einigen Monaten umgesetzt werden können.

Die Ziele von Datenanalysen sind im Bereich der *Wirtschaft* eindeutig die Erhöhung des *wirtschaftlichen Erfolgs* und die Effizienzverbesserung von Prozessen in der Geschäftstätigkeit. Wie aus unterschiedlichen Quellen zu entnehmen, sind die Größenordnungen der Wertgenerierung mit Hilfe von Datenanalysen durchaus bemerkenswert.[147] Dies kann branchenübergreifend als Chance verstanden werden, die bestehenden Prozesse zu überdenken, Data Analytics Systeme zu implementieren und daraus einen *wirtschaftlichen Erfolg* zu generieren. Ein weiterer wichtiger Punkt sind die Arbeitskräfte, die bereits seit drei Jahren am Markt fehlen und durch mangelnde Studiengänge und Ausbildungsplätze in Kürze nicht verfügbar sind.[148] Unternehmen, die aktuell mit stärkerem Interesse als ihre Mitbewerber um Datenanalytiker bemüht sind und Neueinstellungen bzw. Fortbildungen für ihre aktuellen Mitarbeiter vorantreiben, können als Early Mover, vor allem im deutschen Markt verstanden werden. Sie bauen dadurch nicht nur kurzfristig einen Wettbewerbsvorteil auf, sondern können auf Erfahrungswerte und organisches Wachstum der Technologien und des Wissens darüber zurückgreifen. So konnte im nordamerikanischen Raum mit

[146] Vgl. Heilmann und Liegl (2013)
[147] Vgl. BITKOM (2012); Barton und Court (2012)
[148] Vgl. McKinsey Global Institute (2012)

Hilfe einer Studie in 330 öffentlichen Unternehmen gezeigt werden, dass jene mit datenbasierten Entscheidungstools eine 5 % höhere Produktivität und eine um 6 % höhere Profitabilität aufzeigen konnten.[149] Darüber hinaus können auf jeder Stufe der Supply Chain eines Unternehmens Optimierungen mit Hilfe von *Big Data* Tools erreicht werden. Dabei reichen die Anwendungsgebiete vom CRM bis hin zur Lagerlogistik, Fehlerdetektion und den Arbeitsabläufen.[150]

Auch innerhalb der *Technik* ist das *Know-how* ein wichtiger Bestandteil. Daneben ist die zunehmende Konnektivität im privaten, wie im unternehmerischen Bereich ein Bestandteil, der schließlich auch in die IT-Sicherheit mit hineinspielt. Abgesehen von dem Wissen und der Fähigkeit der Mitarbeiter mit den gewonnenen Daten umzugehen, muss auch die Hard- und Software aktuell sein, um die *Datenmengen in Echtzeit* auswerten können. Dazu gehören individuelle Visualisierungen, wie Pie, Area oder Bar Charts, 2D und 3D Graphiken und deskriptive Statistiken. Die *Konnektivität* ermöglicht es hier, dass die Hardware lediglich im Rechenzentrum vorhanden sein muss und auf die Software von unterschiedlichen Punkten aus zugegriffen werden kann. Hierbei ist wichtig, zwischen einem herkömmlichen Rechenzentrum und dem *Cloud Computing*, was viele klassische Speicherformen übertrumpft, indem nicht auf langsame Hardware zugegriffen werden muss, sondern der Zugriff direkt auf den Arbeitsspeicher erfolgt, zu unterscheiden. Dadurch lässt sich eine globale Verbindung zwischen verschiedenen Geschäftsbereichen einer Unternehmung leichter realisieren als bisher, da die gleichen Daten von allen Mitarbeitern abgerufen werden können. Zugleich erfordert dies auch einen sicherheitstechnischen Kraftakt. Auf dem Weg der Übertragung können nicht nur Daten beschädigt oder gar verloren gehen, sie können auch bei nicht ausreichender Sicherung von Dritten eingesehen oder abgerufen werden. Diese Herausforderung gilt es bei der Installation solcher Daten-Intranetze zu lösen.

5.3 Deskriptoren und Projektionen

Zusammenfassung

Die betrachteten Projektionen reichen von steigenden Datenvolumina, über Prozessoptimierung und Preisgestaltung bis hin zu stagnierenden Liberalisierungsgraden im Bereich des Datenschutzes. Für die in Kapitel 5.2 ermittelten Schlüsseltreiber IT-Sicherheit, Gesetzgebung und Datenschutz werden Deskriptoren bestimmt, wie z. B. der Liberalisierungsgrad für den Treiber Datenschutz. Mittels dieser Deskriptoren lassen sich neue, weiter in die Zukunft reichende Projektionen vornehmen anhand derer im weiteren Verlauf Trends fest gemacht werden können.

Nach der Beschreibung der Einflussbereiche und Treiber werden diese nun innerhalb von Deskriptoren quantifiziert, um die Aussagekraft zu erhöhen und die Projektionen in die Zukunft zu ermöglichen. Die Deskriptoren und deren Projektionen werden nachfolgend ausführlich entsprechend der drei Makro-Einflussbereiche *Gesellschaft*, *Wirtschaft* und *Technik* beschrieben. Dieser Projektionskatalog dient als Grundlage für die Alternativenbündelung und die Beschreibung der Szenarien in Kapitel 5.5.

[149] Vgl. McAfee und Brynjolfsson (2012)
[150] Vgl. Baron (2013)

Gesellschaft					
Datenschutz			**Gesetzgebung**		
# herausgebrachten Gesetze	Liberalisierungsgrad	Grad der Transparenz	# herausgebrachten Gesetze	politische Rahmenbedingungen	

Tabelle 17: Deskriptoren für den Einflussbereich *Gesellschaft*

Zur Beschreibung des Treibers *Datenschutz* werden im Folgenden drei Deskriptoren mit je drei Projektionen vorgestellt. Die Gesamtheit der Deskriptoren und ihrer Projektionen ist in Abbildung 45 zu sehen.

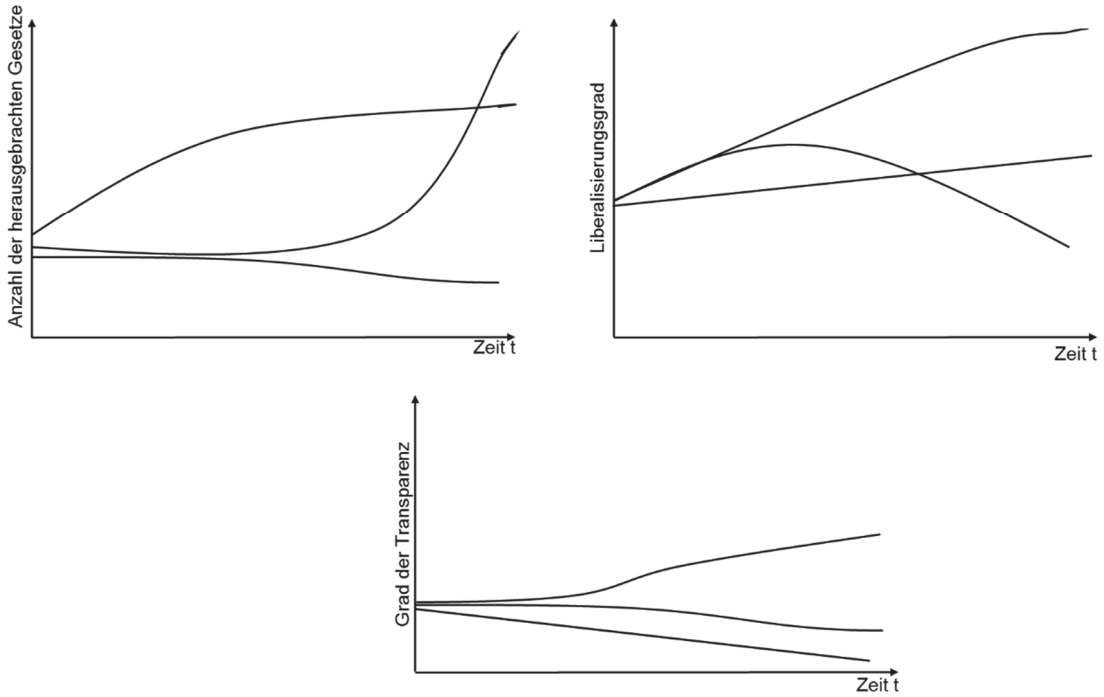


Abbildung 45: Projektionen für die Deskriptoren des Treibers Datenschutz

Die *Anzahl der Gesetze*, die aktiv den *Datenschutz* beeinflussen, können mit zunehmender Sensibilisierung der Bevölkerung mehr oder weniger stark steigen oder aber aufgrund von mangelndem Interesse sinken. Für die zunehmende Sensibilisierung sei an dieser Stelle als Beispiel die Diskussion über die Vorratsdatenspeicherung erwähnt, welche schlussendlich auch der europäische Gerichtshof nicht eindeutig beantworten konnte. So wurde nach den Anschlägen in Madrid 2004 eine Richtlinie zur Datenspeicherung verabschiedet, welche nach Kritiken schließlich 2014 für ungültig erklärt wurde.[151]

[151] Vgl. Voß (2015)

Vor allem der letzte Fall geht einher mit einem steigenden *Liberalisierungsgrad*, also der Bereitschaft von Seiten des Kunden bzw. des Nutzers immer mehr Daten von sich selbst preiszugeben.[152] Dadurch wird das Akzeptanzlevel auf ein Minimum heruntergeschraubt und kaum mehr Kontrolle über die eigenen Daten gewonnen. Allerdings gibt es auch hier die Möglichkeit, dass ein Ereignis den Liberalisierungsgrad sinken lässt und die Nutzer langfristig das Vertrauen in die Datenverwendung verlieren. Kurzfristige Schwankungen wie bei Facebook oder WhatsApp nach den AGB Änderungen bzw. dem Verkauf sind damit nicht gemeint. Sehr wahrscheinlich wird allerdings der gleichbleibende Anstieg des Liberalisierungsgrades in der Zukunft eintreten, da eine kontinuierliche Forderung an den Nutzer nach mehr Transparenz vonstattengeht.

Der *Grad der Transparenz* gibt schließlich die Bereitschaft der Unternehmen an die Art und Weise ihrer Datenakquise offen zu legen und die Weiterverarbeitung oder den Verkauf sichten zu lassen.

Der Treiber *Gesetzgebung* wird durch zwei Deskriptoren ausgestaltet. Zunächst werden die *politischen Rahmenbedingungen* innerhalb eines Landes sowie politische Auswirkungen auf ein Land betrachtet. Anschließend sagt auch die *Anzahl der herausgebrachten Gesetze* etwas über die Notwendigkeit oder den politischen Druck auf ein Land aus. Für die nächsten 20 Jahre lassen sich die in Abbildung 46 dargestellten Projektionen treffen.

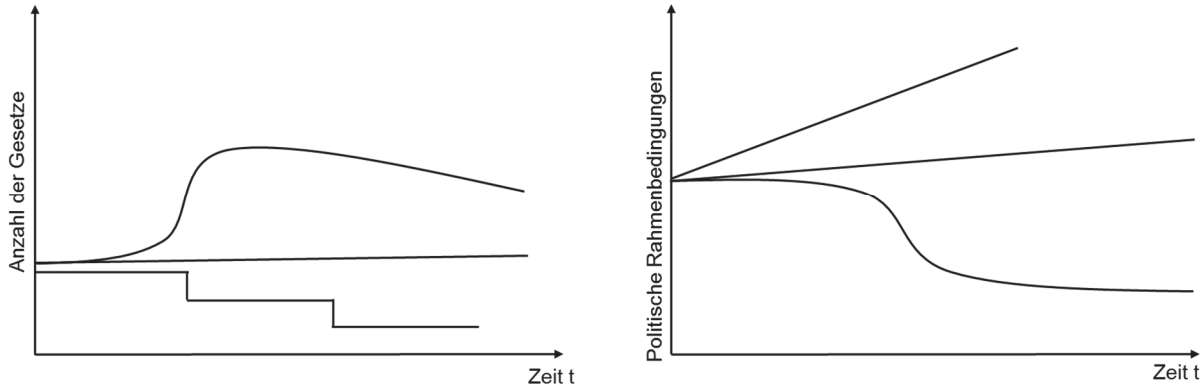


Abbildung 46: Projektionen für die Deskriptoren des Treibers Gesetzgebung

Die *Anzahl der Gesetze* werden in der Zukunft in der ersten Projektion etwa gleichbleiben. Dies lässt eine Prognose auf Basis einer Statistik über die während der letzten 17 Wahlperioden vom Bundestag verabschiedeten Gesetze zu.[153] Die andere Projektion legt ein Schlüsselereignis nahe, woraufhin die Anzahl der Gesetze sprunghaft ansteigt und mit der Zeit kontinuierlich abflacht. Eine weitere Projektion wäre das sprunghafte Abfallen der verabschiedeten Gesetze, wenn durch neue Richtlinien eine weniger starke Regulierung gefordert wird.

Der zweite Deskriptor betrachtet die *politischen Rahmenbedingungen*, die innerhalb eines Staates durch Parteien und außerhalb eines Staates durch bi- oder multilaterale Verträge zustande kommen. Politische Rahmenbedingungen können sich auch in nach außen stabilen Wirtschaftsverbänden ändern, wie die aktuelle Krise der europäischen Union und die Frage nach dem Ausschluss Griechenlands zeigen.[154] Diesen Verlauf verdeutlicht der nach unten einbrechende Graph. Nach einem

[152] Vgl. Heilmann und Liegl (2013)
[153] Vgl. Deutscher Bundestag (2013)
[154] Vgl. Greive (2015)

Ereignis werden die *Rahmenbedingungen* schlechter für den Staat. Die beiden anderen Projektionen zeigen ein gleichbleibendes bzw. leicht positives Bild auf.

Wirtschaft	
Effizienz	**Wirtschaftl. Erfolg**
Prozessoptimierung	Profitabilität
Risikobewertung	Preisgestaltung
Entscheidungen	Kundenorientierung

Tabelle 18: Deskriptoren für den Einflussbereich *Wirtschaft*

Die beiden Treiber *Effizienz* und *wirtschaftlicher Erfolg* werden von je drei Deskriptoren beschrieben, die ihrerseits je drei Projektionen aufweisen. Neben der Übersicht in Tabelle 18 sind die Deskriptor-Ausprägungen in Abbildung 47 und Abbildung 48 zu sehen.

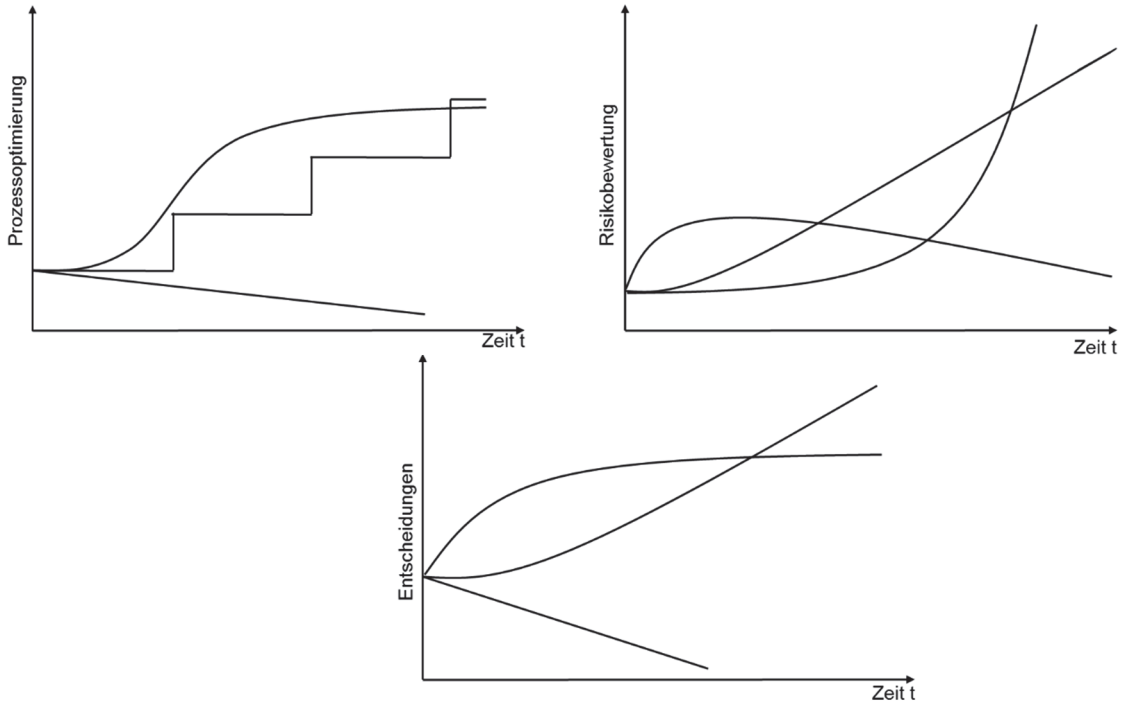


Abbildung 47: Projektionen für die Deskriptoren des Treibers Effizienz

Im Rahmen der *Effizienz* kann die *Prozessoptimierung* unterschiedlichen Projektionen folgen, wie der kontinuierlichen Verbesserung. Hier sei als Beispiel der KVP erwähnt.[155] Eine andere Möglichkeit ist der stufenweise Anstieg, hervorgerufen durch Gesetze oder andere externe Einflussfaktoren wie Neuerungen im Wettbewerb, die dann auf das eigene Unternehmen übertragen werden. Eine

[155] Vgl. VOREST (2013)

abfallende Tendenz der *Prozessoptimierung* ist möglich, wenn die Optimierungsmöglichkeiten erschöpft sind oder aber auf den Prozess weniger Wert gelegt wird.

Die *Risikobewertung* kann ebenfalls durch zwei steigende Projektionen und eine nach einem leichten Anstieg doch fallende Projektion erklärt werden. Dass *Big Data* vor allem im Finanzsektor und konkret bei der Bewertung und Prävention von Kreditkartenbetrugsdelikten eine wichtige Rolle spielt, lässt sich aus den vermehrten Publikationen zu dem Thema entnehmen. So werden in der proaktiven Software Muster von Kreditkartenkäufen für jeden Eigentümer angelegt. Weicht ein Einkauf von dem gewohnten Muster ab, gibt das System Meldung an die Bank oder verständigt sogar selbst den Eigentümer. Darüber werden die Betrugsdelikte nicht unbedingt drastisch weniger, aber es können wesentlich mehr Betrugsdelikte aufgedeckt werden.[156]

Auch im Rahmen der *Entscheidungen* ist die abfallende Projektion einem mangelnden Vertrauen der Menschen in *Big Data* - Analysen geschuldet. Die schwach ansteigende Wurzelfunktion suggeriert das Erreichen eines Plateaus, worüber hinaus eine Entscheidungsfindung nicht mehr auf *Big Data* Tools basiert. Die steigende Projektion geht auf eine anhaltende Nachfrage von Unternehmen nach Datenanalysen zur Entscheidungsfindung zurück. In der kurzen Zeitspanne werden die beiden steigenden Projektionen durch eine Veröffentlichung der BITKOM mit über 30 Beispielen zur Einbindung von *Big Data* in Entscheidungsprozesse gestützt.[157]

In Abbildung 48 sind die Projektionen für die drei Deskriptoren des Treibers *wirtschaftlichen Erfolg* zu sehen.

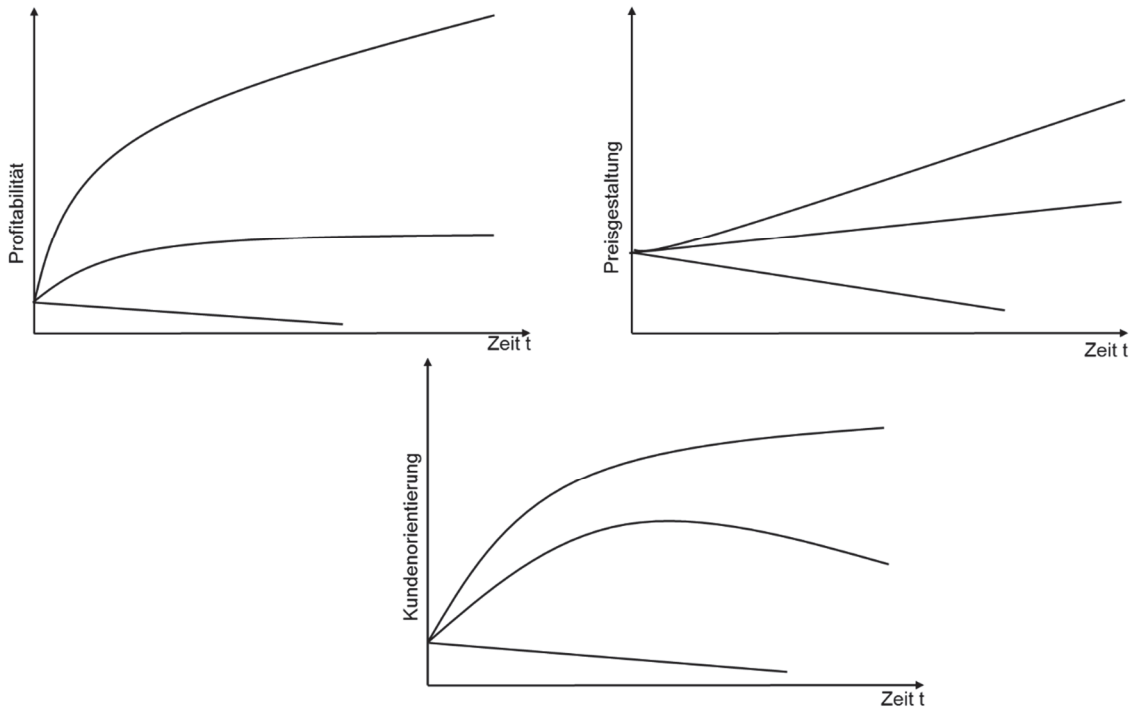


Abbildung 48: Projektionen für die Deskriptoren des Treibers wirtschaftlichen Erfolg

[156] Vgl. Richey (2014)
[157] Vgl. BITKOM (2012), BITKOM (2014)

Die *Profitabilität* benötigt im Allgemeinen im Vergleich zu anderen Deskriptoren eine gewisse Vorlaufzeit, bis sich Ergebnisse nachhaltig durchsetzen. Aus diesem Grund haben die beiden steigenden Projektionen einen schwächeren Anstieg zu verzeichnen. Die *Profitabilität* kann allerdings auch abfallen, wenn die Systeme fehlerhaft sind, die Mitarbeiter nicht mit den Tools umgehen können oder andere Gründe ein negatives Arbeitsklima erzeugen.

Der Deskriptor *Preisgestaltung* stützt sich auf die Relevanz von Echtzeit-Preisanpassungen um optimal Spitzenströme auszunutzen. Auch hier lassen sich wie bei dem Deskriptor zuvor drei Projektionen abbilden, die grob in steigend, gleichbleibend und fallend eingeordnet werden können.

Als letzter Deskriptor muss die *Kundenorientierung* besonders hervorgehoben werden. *Big Data* - Analysen gewähren einem Unternehmen ganz andere Einblicke in die Haltung der Kunden zu dem Unternehmen als dies bisherige Tools möglich machen konnten. Es muss nicht mehr mit kostspieligen Umfragen gearbeitet werden, sondern Bewertungen in den sozialen Netzwerken und anderen Seiten im Internet bieten einem Unternehmen mit Zugang zu Echtzeit-Analyse-Tools eine Fülle von Informationen. Diese können zu angepassten Produkten oder Dienstleistungen führen, Prozesse für den Kunden individueller gestalten oder Wettbewerbsvorteile identifizieren.[158] Aus diesem Grund ist die *Kundenorientierung* als wichtiger Aspekt des *wirtschaftlichen Erfolges* anzusehen und steigt in eine Projektion stark. Die mittlere Projektion tritt ein, wenn sich die Kunden aufgrund der Analysen als zu sehr durchleuchtet fühlen. Dann kann nach einem anfänglichen Anstieg der *Kundenorientierung* diese wieder abfallen, aufgrund von Abwanderungen. Hier ist eine sehr sensible, verantwortungsbewusste und transparente Datenanalyse notwendig. Andernfalls kann die *Kundenorientierung* sogar sinken.

Innerhalb des letzten Einflussbereichs, der *Technik*, werden wie in Tabelle 19 zu sehen insgesamt neun Deskriptoren mit je drei Projektionen betrachtet.

Technik									
Konnektivität			**Know-how**			**IT-Sicherheit**			
Datenmenge	Vielfalt	Daten Akquisition	Analyse der Daten	Visualisierung der Daten	Distribution	Verschlüsselung	Data Masking	Identitäts- und Zugangs-Management	

Tabelle 19: Deskriptoren für den Einflussbereich *Technik* den Treibern zugeordnet

Als Einstieg wird die *Konnektivität* in ihre drei Deskriptoren *Datenmenge*, *Datenvielfalt* und *Datenakquisition* unterteilt. Die Projektionen sind in Abbildung 49 zu sehen.

[158] Vgl. BITKOM (2014)

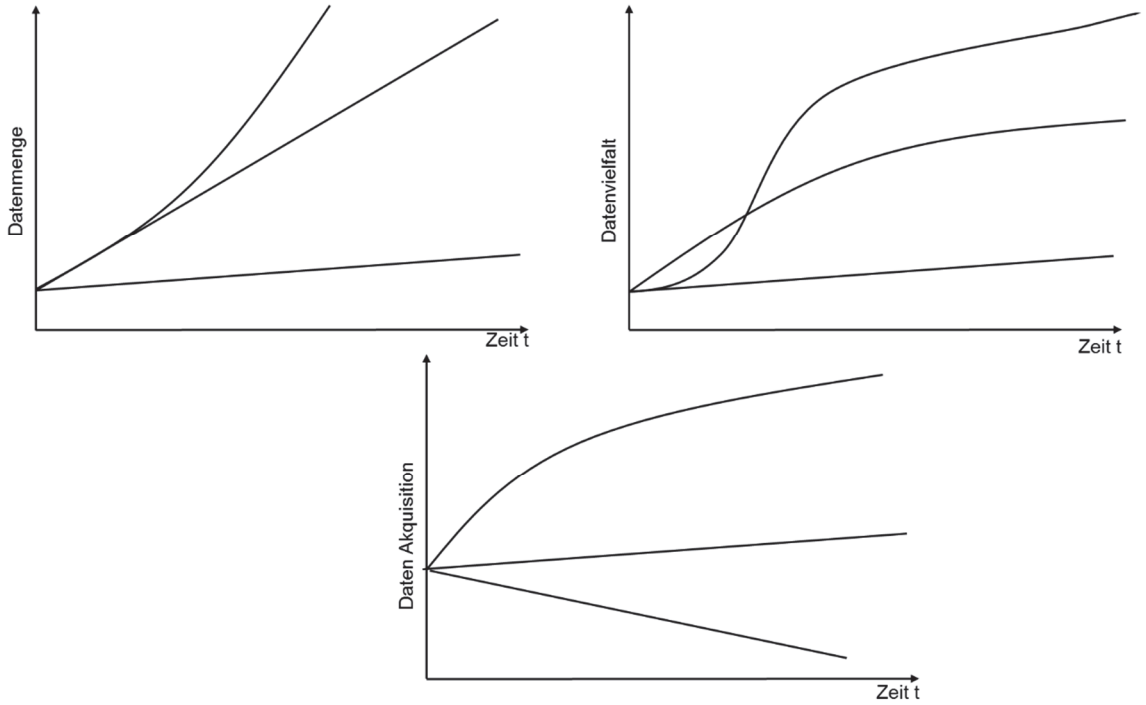


Abbildung 49: Projektionen für die Deskriptoren des Treibers Konnektivität

Die *Datenmenge* wird selbst in der geringsten Projektion steigen, da nicht nur die reine Größenordnung betrachtet wird, sondern auch die wachsende Anzahl der Schnittstellen, die Daten generieren können. Die anderen beiden Projektionen gehen lediglich von einem mittelstarken bis sehr starken Anstieg aus. Dies ist mit aktuellen Erkenntnissen, die einen exponentiellen Anstieg des Datenvolumens vorhersagen, übereinstimmend.[159]

Darüber hinaus gibt es unzählige Dateiformate zum Beispiel im Bereich der Bilderfassung und Bildbearbeitung. So werden die Formate je nach Marke und Modell zum Teil in einem eigenen Format abgespeichert und müssen anschließend erst in eines der gängigen Formate transformiert werden. Dadurch steigt der Bedarf an Programmen in diesem Bereich stark an und die *Datenvielfalt* ebenfalls.

Um Datenanalysen durchführen zu können, müssen Daten erhoben und akquiriert werden. Hierfür müssen alle zur Verfügung stehenden Quellen genutzt werden können und auch zugänglich sein. Dies wird in der Zukunft aufgrund der steigenden Nachfrage möglicherweise Restriktionen unterworfen sein, da die Daten nicht mehr frei zur Verfügung stehen könnten, sondern wie bereits heutzutage im kleinen Maßstab getätigt zum Verkauf angeboten werden. Darüber hinaus kann eine steigende Skepsis oder gar Ablehnung der Bevölkerung in begrenztem Zugang oder im Extremfall sogar zu einer rückläufigen Datenlage führen. Diese Entwicklung ist allerdings nach aktuellem Stand unwahrscheinlich, da eine *Datenakquisition* schon mit Hilfe einer steigenden Anzahl von mobilen Endgeräten gesichert ist.

Für den Treiber *Know-how* können drei Deskriptoren identifiziert werden, die in Abbildung 50 dargestellt sind.

[159] Vgl. SHS (2015)

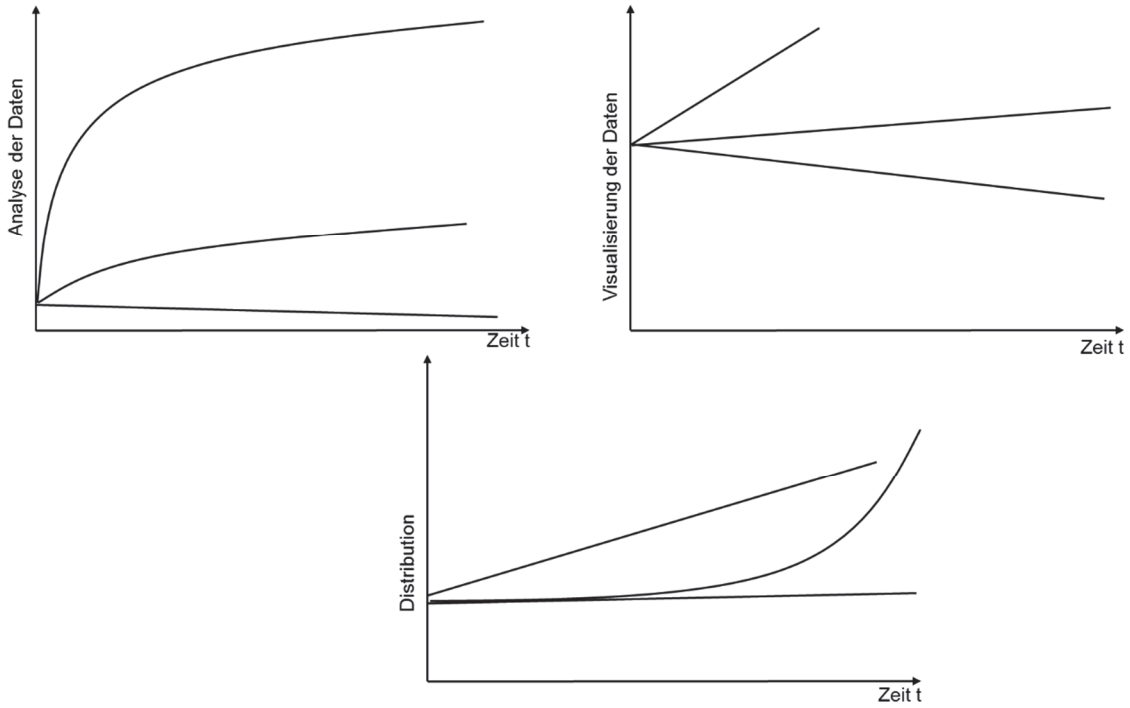


Abbildung 50: Projektionen für die Deskriptoren des Treibers Know-how

Die *Analyse der Daten* wird für die Entscheidungsfindung und die Wissensbasis benötigt. Dabei sieht es in der Zukunft eher nach einem Anstieg, als nach einem Rückgang aus. Dies lässt sich mit der wachsenden Anzahl internationaler Start-ups im Bereich Datenanalysen belegen. Hier seien Quid (San Francisco), Graylog (Hamburg; Houston) und mapegy (Berlin) erwähnt.

Ebenso sind die *Visualisierungen der Daten* auch schon heutzutage ein wichtiges Tool um Sachverhalte klar und deutlich in Kürze zu präsentieren. Diese Entwicklung wird in der Zukunft weiter zunehmen, weshalb eine Projektion sehr stark linear ansteigt.

Die *Distribution* der Daten auf Plattformen mit Hilfe einheitlicher Datentypen bedarf noch einiger Investitionen, die nicht nur finanzielle, sondern vor allem kognitive Natur ist. Dazu zählen vor allem die Datenmanager, die in der Zukunft als wichtiger Berufszweig solche Herausforderungen lösen müssen. Aus diesem Grund starten die Projektionen etwas verhaltener als in anderen Deskriptoren.

Abschließend werden die Deskriptoren für den Treiber *IT-Sicherheit* betrachtet, da dieses Thema wie bereits erwähnt, besonders im deutschen und europäischen Raum einen großen Stellenwert hat.

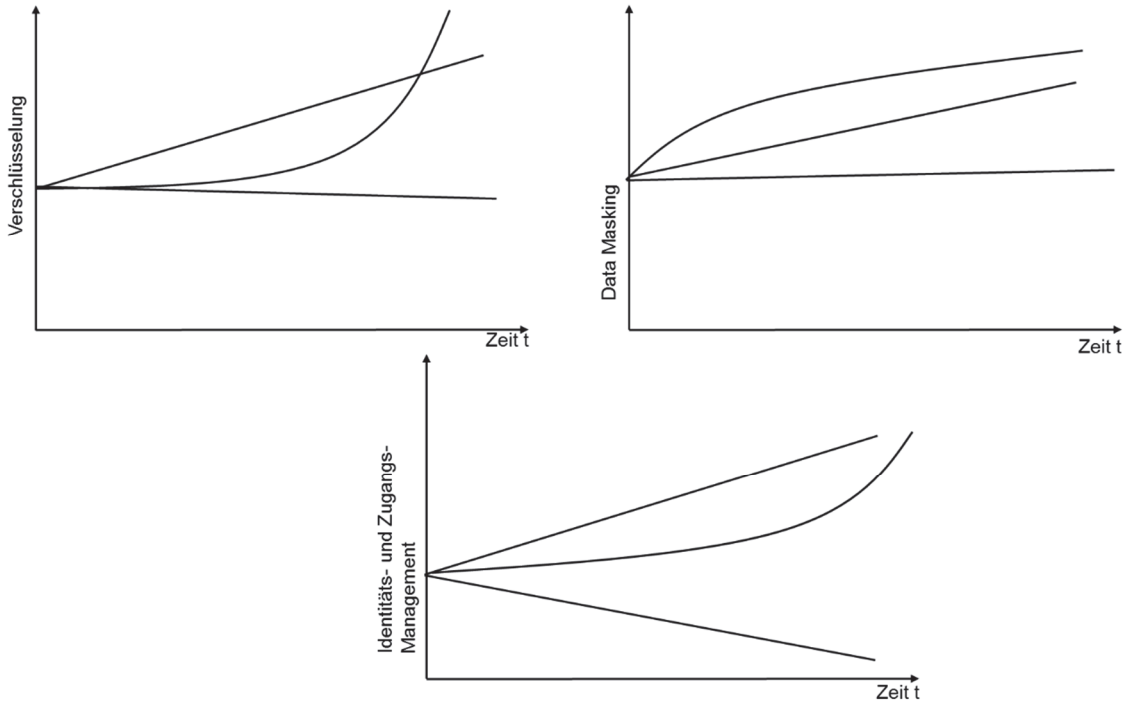


Abbildung 51: Projektionen für die Deskriptoren des Treibers IT-Sicherheit

Die Verschlüsselung von Daten ist in Teilen des Email-Verkehrs und der Regierungsabteilungen bereits Alltag. Die Entwicklung geht zu einer persönlicheren, individuelleren Verschlüsselung, da private Daten vermehrt im Internet oder in Devices mit permanentem Internetzugang gespeichert und verwaltet werden. Somit sehen die Projektionen hier keinerlei gravierende Abnahme vor.

Das *Data Masking* geht in eine ähnliche Richtung, führt die Verschlüsselung allerdings mit Hilfe von *Big-Data-Technologien* durch.[160] Hier sind die Projektionen verhalten zu betrachten, da bisher nur wenige ausgereifte Programme für den kommerziellen Gebrauch existieren. Die Unternehmen lassen ihre Daten intern zwar unter Gesichtspunkten des *Data Masking* verschlüsseln, stellen ihre Software aber nicht zur Verfügung, da sich daraus direkt negative Sicherheitseffekte ergeben können. Somit gilt ähnlich wie beim ersten Deskriptor keinerlei Abnahme, aber eine langsamere Steigung der einzelnen Projektionen.

Der letzte Deskriptor ist ein weiterer Aspekt im inneren Bereich einer Firma. Mit *Identitäts- und Zugangsmanagement* ist die Verwaltung von allen Benutzern innerhalb eines *Big Data* - Systems und deren Zugriffsberechtigungen gemeint. Dabei geht es um die genaue Abgrenzung von read-, write- und execute-Rechten und der Sicherheit der individuellen Ergebnisse jedes einzelnen Benutzers. Sollten in diesem Bereich Fehler auftreten, würden Mitarbeiter die bereitgestellten Plattformen nicht mehr ohne Zurückhaltung benutzen oder gar auf Alternativen umsteigen. Dies würde weder die Vereinfachung noch die Akzeptanz der Technologie innerhalb des Unternehmens positiv beeinflussen. Daher sind die beiden steigenden Projektionen erwünscht und die Zukunft wird dann über die Größe der Steigung bestimmen.

[160] Vgl. BITKOM (2014)

5.4 Alternativenbündelung

Zusammenfassung

Aus den Projektionen werden drei verschiedene Projektionsbündel zusammengefasst. Jedes Projektionsbündel stellt eines der Szenarien dar, die in Kapitel 5.5 näher beschrieben werden.

Nach der detaillierten Beschreibung der Ergebnisse erfolgt in diesem Schritt die Bündelung der Projektionen zu Rohszenarien. Dafür ist in Tabelle 20 eine Zusammenfassung aller bis hierhin erhaltenen Ergebnisse der Szenariotechnik aufgelistet.

Einfluss-bereich	Gesellschaft					Wirtschaft					Technik								
Treiber	Daten-schutz			Gesetz-gebung		Effizienz			wirt. Erfolg			Konnek-tivität			Know-How			IT-Sicherheit		
Deskrip-toren	# herausgebrachten Gesetze	Liberalisierungsgrad	Grad der Transparenz	# herausgebrachten Gesetze	politische Rahmenbedingungen	Prozessoptimierung	Risikobewertung	Entscheidungen	Profitabilität	Preisgestaltung	Kundenorientierung	Datenmenge	Vielfalt	Daten Acquisition	Analyse der Daten	Visualisierung der Daten	Distribution	Verschlüsselung	Data Masking	Identitäts- und Zugangs-Management
Projek-tionen	langsam steigend	stark steigend	schwach steigend	gleichbleibend	gleichbleibend	s-kurven förmig ansteigend	linear steigend	linear steigend	starke Wurzelfunktion	gleichbleibend	starke Wurzelfunktion	stark linear steigend	steigende S-Kurve	gleichbleibend	schwach steigend	gleichbleibend	gleichbleibend	exp. Steigend	Wurzelfunktion	stark steigend
	exp. Steigend	abfallend	schwach fallend	ansteigend und gleichbleibend	sprunghaft abfallend	stufenweiser Anstieg	abfallend	Wurzelfunktion ansteigend	schwache Wurzelfunktion	linear steigend	ansteigend und abfallend	exp. Steigend	Wurzelfunktion ansteigend	stark ansteigend	stark steigend	stark linear steigend	exp. Steigend	linear steigend	lin. Steigend	schwach steigend
	abfallend	gleichbleibend	stark abfallend	stark abfallend	stark steigend	stark fallend	exp. Steigend	einbrechend	stark abfallend	stark abfallend	stark abfallend	gleichbleibend	gleichbleibend	stark fallend	gleichbleibend	abfallend	linear steigend	abfallend	gleichbleibend	abnehmend

Tabelle 20: Übersicht über Treiber, den dazugehörigen Deskriptoren und deren Projektionen

Die Projektionen lassen sich in drei Alternativenbündel unterteilen. Hierbei wird die Methode des paarweisen Vergleichs herangezogen. Diese Kriterien für Szenarien sind die folgenden:

- Konsistenz und Widerspruchsfreiheit: Sich widersprechende Projektionen können nicht innerhalb eines Szenarios auftreten.
- Stabilität: Durch die Änderung von Projektionen darf nicht das ganze Szenario zum Einsturz gebracht werden.

- Diversität: Die Szenarien müssen grundsätzlich eindeutig voneinander unterscheiden werden können.

Die Projektionen werden anhand eines Paarvergleichs und unter Beachtung der drei Kriterien zu Alternativen gebündelt. Es sollen insgesamt drei Szenarien entstehen. Die drei Bündel sind in Tabelle 21 dargestellt und entsprechen den sogenannten Rohszenarien.

Einfluss-bereich	Gesellschaft					Wirtschaft					Technik								
Treiber	Daten-schutz			Gesetz-gebung		Effizienz			wirt. Erfolg			Konnek-tivität			Know-How			IT-Sicherheit		
Deskrip-toren	# herausgebrachten Gesetze	Liberalisierungsgrad	Grad der Transparenz	# herausgebrachten Gesetze	politische Rahmenbedingungen	Prozessoptimierung	Risikobewertung	Entscheidungen	Profitabilität	Preisgestaltung	Kundenorientierung	Datenmenge	Vielfalt	Daten Acquisition	Analyse der Daten	Visualisierung der Daten	Distribution	Verschlüsselung	Data Masking	Identitäts- und Zugangs-Management
Projek-tionen	langsam steigend	stark steigend	schwach steigend	gleichbleibend	gleichbleibend	s-kurven förmig anst end	linear steigend	linear steigend	starke Wurzelfunktion	gleichbleibend	starke Wurzelfunktion	stark linear steigend	steigende S-Kurve	gleichbleibend	schwach steigend	gleichbleibend	gleichbleibend	exp. Steigend	Wurzelfunktion	stark steigend
	exp. Steigend	abfallend	schwach fallend	ansteigend un eichbleibend	sprunghaft abfallend	stufenweiser Anstieg	abfallend	Wurzelfunktion ansteigend	schwache Wurzelfunktion	linear steigend	ansteigend und abfallend	exp. Steigend	Wurzelfunktion ansteigend	stark ansteigend	stark steigend	stark linear steigend	exp. Steigend	linear steigend	lin. Steigend	schwach steigend
	abfallend	gleichbleibend	stark abfallend	stark abfallend	stark steigend	stark fallend	exp. Steigend	einbrechend	stark abfallend	stark abfallend	stark abfallend	gleichbleibend	gleichbleibend	stark fallend	gleichbleibend	abfallend	linear steigend	abfallend	gleichbleibend	abnehmend

Tabelle 21: Alternativenbündelung der Projektionen

5.5 Beschreibung der Szenarien

Zusammenfassung

Neben einer Visualisierung im Szenariotrichter werden die drei Szenarien „Bytes im Freudentaumel“, „Visionen verwirklichen“ und „Datenkrieg“ verbal ausführlich beschrieben.

Die „Bytes im Freudentaumel“ beschreiben das Szenario mit der höchsten Akzeptanz in der Bevölkerung und der größten Autonomie der Daten. „Die Sicherheit, mit der die Systeme arbeiten ist so sicher, wie kein Staat der Welt.“, so ein Sprecher der Regierung auf Rückfragen zum Thema erlassene Gesetze, „wir sind nun wirklich in einem sich selbst regulierenden Markt und die wenigen Gesetze, die verabschiedet wurden, sind nicht der Rede wert.“ Die enge Verbindung von Technik und Alltag benötigt im Rahmen der IT-Sicherheit keine Neuerungen. Die Sicherheit kommt durch das tiefe Verständnis der Menschen für die Algorithmen.

„Visionen verwirklichen“ beschreibt die realistische Einschätzung der Mensch-Maschine-Interaktion in der Zukunft. Hierbei lautet das Motto: „Die Entwicklung nimmt ihren Lauf und wir gehen im Schleichschritt.“ Forschungszentren machen sich die Systeme zunutze, in dem sie soziale Netzwerke oder Publikationen mit verschiedenen Analysen koppeln und so neue Felder entdecken, die dem menschlichen Auge bis dahin verborgen geblieben sind. Als wichtigster Arbeitsmarkt zeigten sich die Programmierung von Algorithmen und das Verständnis von den Systemen.

Das dritte Szenario beschreibt den „Datenkrieg“, in dem die Datenalgorithmen gegen die Menschen arbeiten. Dies ist auf die versäumten Investitionen in der Ausbildung von Datenmanagern zurückzuführen. Im Verhältnis zu der Menge an Menschen, die die Technik nutzen, wurde die Technik von nur sehr wenigen Menschen verstanden. Die Weiterentwicklungen wurden schneller gefordert, als Verständnis dafür in der breiten Masse entwickelt wurde. Das Resultat sind selbst agierende Systeme, die von der Datenakquise, über die Datenverarbeitung und die Entscheidungsfindung bis hin zu eigenständigen Änderungen auf Basis dieser Analysen in Abläufe und Prozesse eingreift.

5.5.1 Trendszenario: Visionen verwirklichen

Die Balance zwischen Systemen, die in vielen Bereichen eigenständig arbeiten, um Ergebnisse zu produzieren und den Menschen, die diese Ergebnisse und auch die Systeme verstehen, ist eingetreten. Neben einem gesunden Maß an neuen Regelungen, die parallel zur Systementwicklung berechtigt sind, um Missbrauch angemessen zu tadeln, wird weder über- noch unterreguliert. Die Wachstumsraten sind moderat bis gut und in der Wirtschaft ist ein leichter Aufschwung und Enthusiasmus über die moderne Technik und die mit ihr verbundenen Möglichkeiten zu spüren. Das Motto lautet: „Die Entwicklung nimmt ihren Lauf und wir gehen im Schleichschritt.“ Die erzielten Optimierungen innerhalb vieler Firmen sind nicht mehr Eintagsfliegen, sondern lassen sich in vielen Bereichen beobachten. Dadurch wird eine leicht steigende Akzeptanz der Bevölkerung beobachtet, was ein wachsendes Interesse an den Systemen hervorruft. Es wurde bereits an möglichen Schnittstellen

zwischen den reinen Analysesystemen und Robotern geforscht, sodass nicht allein die Wirtschaft von der Vision einer optimal ausgestalteten Supply Chain träumen kann. Auch Forschungszentren machen sich die Systeme zunutze, in dem sie soziale Netzwerke oder Publikationen mit verschiedenen Analysen koppeln und so neue Felder entdecken, die dem menschlichen Auge bis dahin verborgen geblieben sind. Als wichtigster Arbeitsmarkt zeigten sich die Programmierung von Algorithmen und das Verständnis von den Systemen. Mit ausdrücklicher Priorität werden Ausbildungen finanziert und die jungen Arbeitnehmer auf den Umgang mit Systemen geschult. So wird das Bild von ***System und Mensch im Gleichschritt*** geprägt.

5.5.2 Enthusiastisches Szenario: Bytes im Freudentaumel

Als tägliche Begleiter sind die Datenanalysen und Algorithmen nicht mehr aus dem Alltag wegzudenken. Die Algorithmen werden vom Menschen nicht nur angewendet, sondern auch verstanden und können zügig verändert werden. Es wurden Gremien gegründet, die sich ausschließlich mit der Frage der Einbettung der neuen Technologien in den menschlichen Alltag beschäftigt haben. Daraus konnte ein Bildungsstandard erreicht werden, der besonders im Bereich der Datenanalysen einen problemlosen Umgang mit den Systemen Erfolg verspricht. Aus diesem Grund sind Datenschutzrichtlinien oder andere Gesetze nach Meinung der Bevölkerung nicht wichtig. „Die Sicherheit, mit der die Systeme arbeiten ist so sicher, wie kein Staat der Welt.“, so ein Sprecher der Regierung auf Rückfragen zum Thema erlassene Gesetze, „wir sind nun wirklich in einem sich selbst regulierenden Markt und die wenigen Gesetze, die verabschiedet wurden, sind nicht der Rede wert.“ Die enge Verbindung von Technik und Alltag benötigt im Rahmen der IT-Sicherheit keine Neuerungen. Die Sicherheit kommt durch das tiefe Verständnis der Menschen für die Algorithmen. Dadurch entsteht eine vertrauensvolle Symbiose, deren Ziel es ist, eine stabile und stark kundenorientierte wirtschaftliche Struktur zu schaffen. Es gibt keinen Grund seine Daten zum Nutzungsverhalten, zu sozialen Gruppen, Vorleiben und Kontakten nicht preiszugeben. Zum einen ist man einem sozialen Druck ausgesetzt, da mit Hilfe der Daten viele Unternehmen immense Profitabilitätszuwächse verzeichnen können und keiner dem anderen in Algorithmengeschwindigkeit oder Serverleistung hinter her sein möchte. Zum anderen ist die Sicherheit der Maschinen immens, wodurch eine ***Vorsicht absolut nicht notwendig*** ist, sondern eher behindert.

5.5.3 Horrorszenario: Datenkrieg

Die Analysen entwickeln sich selber weiter und sind nicht mehr Entscheidungshilfen, sondern Entscheidungsträger. Die ursprünglich als harmlose, einfache Helfer des Alltags beschriebenen Technologien entwickeln nicht nur ein Eigenleben, sondern beherrschen den Alltag. Daraus entsteht eine wachsende Angst der Bevölkerung der Technik gegenüber, auch weil das Know-how der Menschen im Vergleich zu den Software-Tools stark abnimmt und sich die Analyse-Tools mit Daten selbstversorgen. Dies ist auf die versäumten Investitionen in der Ausbildung von Datenmanagern zurückzuführen. Im Verhältnis zu der Menge an Menschen, die die Technik nutzen, wurde die Technik von nur sehr wenigen Menschen verstanden. Die Weiterentwicklungen wurden schneller gefordert, als Verständnis dafür in der breiten Masse entwickelt wurde. Das Resultat sind selbst agierende Systeme, die von der Datenakquise, über die Datenverarbeitung und die Entscheidungsfindung bis hin zu eigenständigen Änderungen auf Basis dieser Analysen in Abläufe und Prozesse eingreift. Allerdings kann die Datenmenge, -vielfalt und -akquisition nicht wie erwartet stark wachsen, da die Menschen

als wichtigste Datenquelle fehlen. Ganz im Gegenteil versuchen diese mit Hilfe steigender Sicherheitssysteme dem Chaos Herr zu werden. Auch versuchen sie ihre eigenen Daten möglichst sicher zu verschlüsseln, damit eine – wenn auch sehr eingeschränkte – Privatsphäre gesichert werden kann. Die Softwarepakete sind an dem Zeitpunkt nicht alle kompatibel sondern haben unterschiedliche Algorithmen. Dadurch versucht jedes System auf dem Markt mit den geringen Datenmengen die besten Resultate zu erzielen. Da die Menschen und damit auch die Kunden innerhalb der Wirtschaft kaum noch Daten preisgeben und zusätzlich in der Angst gefangen sind von Systemen analysiert zu werden, bricht zuerst die Kundenorientierung ein. Daraus resultiert in einer Kaskade der Einbruch des wirtschaftlichen Handelns. Die Prozessoptimierung und Risikobewertung kann nichtmehr zuverlässig durchgeführt werden, da die Menschen den Urteilen der Systeme nicht trauen, sich ausschließlich überwacht und analysiert fühlen und dadurch den Anweisungen nicht Folge leisten. Schlussendlich sinkt dadurch die Profitabilität rapide. Nichts desto trotz wird versucht, mit Hilfe digitaler Aktualisierungen von Datenschutzgesetzen und anderen staatlichen Anordnungen die Systeme in Schach zu halten. Dies gelingt bereits kaum noch, da ein Überlebenskampf in der digitalen Welt herrscht.

5.6 Störereignisse und deren Wirkung auf die Szenarien

Zusammenfassung

Zwei mögliche Störereignisse (sog. Wild Cards) könnten die Entwicklung der Projektionen nachhaltig verändern: „globaler Stromausfall“ und „künstliche Intelligenz“. Ein “globaler Stromausfall“ würde sämtliche Einflussbereiche negativ beeinflussen und allen voran die *Wirtschaft* und *Gesellschaft* in eine Ur-Stimmung zurückwerfen. „Künstliche Intelligenz“ wiederum unterstreicht die Vorteile der Algorithmen und nimmt hier positiven Einfluss auf die Szenarien.

Im Folgenden werden Störereignissen beschrieben, welche die beschriebenen Szenarien nachhaltig beeinflussen können. Störereignisse sind solche Ereignisse, welche eine geringe Eintrittswahrscheinlichkeit haben bei gleichzeitig hoher Auswirkung.

1. Globaler Stromausfall

Als gesellschaftliches Störereignis wird nachfolgend der **globale Zusammenbruch des Stromnetzes betrachtet**. Durch solch einen Zustand und den damit einhergehenden Zusammenbrüchen von Kommunikationskanälen (Internet, analoges Telefon) in globalem Ausmaß, treten weitreichende Folgen für die Gesellschaft und die *Wirtschaft* eines Landes ein. Wie in „Blackout“ von Marc Elsberg 2012 eindrücklich beschrieben, würden nach nicht einmal fünf Tagen ohne Stromversorgung sämtliche zivile Versorgungen im Bereich der Medizin und Ernährung zusammenbrechen.[161] Die Ursachen, Folgen und nötige Präventivmaßnahmen konkret für Deutschland wurden 2010 in einem Arbeitsbericht des Büros für Technikfolgen-Abschätzung beim deutschen Bundestag ausgearbeitet.[162]

[161] Vgl. Elsberg (2012)
[162] Vgl. TAB (2010)

Außerdem resultieren die zunehmende just-in-time-Produktion und die datenbasierte Produktionsüberwachung in einer starken Abhängigkeit ganzer Handelsstränge von Datenanalysen.

Tabelle 22 beschreibt in der mittleren Spalte die Projektionen im Ist-Zustand während des Szenarios „Datenkrieg“ und gibt mit Hilfe der Pfeile in der letzten Spalte die Veränderung durch das Störereignis „globaler Stromausfall“ an.

Datenkrieg		
globaler Stromausfall	Anzahl der verabschiedeten Gesetze steigt an	
	Bereitschaft der Menschen Daten preiszugeben ist gering	
	Transparenzbereitschaft der Unternehmen stagniert	
	Effizienz geht zurück	
	Unternehmen beklagen wirtschaftliche Einbußen und Misserfolge	
	Datenmenge und –vielfalt stagniert	
	Möglichkeiten neue Daten zu gewinnen sehr eingeschränkt	
	Datenverarbeitung und -aufbereitung gleichbleibend	
	Sicherheitsmaßnahmen im IT-Bereich stark wachsend	
Σ	Akzeptanz der datenbasierten Entwicklung in der Bevölkerung ist gering	

Tabelle 22: „Datenkrieg“ mit Störereignis „globaler Stromausfall“

Es ist eindeutig zu erkennen, dass nicht nur innerhalb einiger wichtiger Deskriptoren die bereits negativen Ausprägungen verstärkt werden (z. B. *Effizienz* und *wirtschaftlicher Erfolg*), sondern dass auch in Summe die Ablehnung gegenüber datenbasierten Entwicklungen, die zwangsweise an die Strom- und Versorgungsnetze gekoppelt sind, weiter zunimmt. Auf die Gesetze und die Datenmenge hat das Störereignis keinerlei gravierenden akuten Auswirkungen, da das Augenmerk im Falle eines globalen Ausfalls nicht auf der Erlassung neuer Gesetze liegt. Zusätzlich können Datenmengen nur händisch erfasst werden, was die Stagnation innerhalb des Szenarios bestätigt aber nicht aktiv unterstützt. Weniger verstärkt werden dagegen die Transparenzeigenschaft der Unternehmen und die

Datenverarbeitung, da aufgrund der fehlenden Infrastruktur besonders letzteres nur manuell möglich ist, was nicht nur zu einer Stagnation des Prozesses führt, sondern zu einer Rückentwicklung.

In Tabelle 23 und Tabelle 24 werden die Auswirkungen des globalen Stromausfalls auf die Szenarien „Bytes im Freudentaumel“ und „Visionen verwirklichen“ dargestellt.

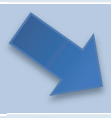
Bytes im Freudentaumel		
globaler Stromausfall	Anzahl der verabschiedeten Gesetze sinkt	
	Bereitschaft der Menschen Daten preiszugeben ist sehr hoch	
	Transparenzbereitschaft der Unternehmen sinkt	
	Effizienz steigt mit unterschiedlichen Geschwindigkeiten an	
	Unternehmen erfreuen sich an wachsender Profitabilität durch Kundenfokus	
	Datenmengen stehen in vielerlei Formaten zur freien Verfügung	
	Möglichkeiten neue Daten zu gewinnen unbegrenzt	
	Analyse, Visualisierung und Distribution der Daten steigt stark an	
	Sicherheitsmaßnahmen im IT-Bereich stark abnehmend	
Σ	Akzeptanz der datenbasierten Entwicklung in der Bevölkerung ist immens	

Tabelle 23: „Bytes im Freudentaumel“ mit Störereignis „globaler Stromausfall“

In diesem Szenario zeigt sich deutlich die Entwicklung in eine weniger überschwängliche Akzeptanz. Durch die nahezu völlige Abhängigkeit von den datenbasierten Analysen würde ein globaler Stromausfall nicht nur die *Wirtschaft* extrem schädigen, sondern auch das persönliche Miteinander. Kommunikation auf analogem Wege sind möglicherweise verlernt worden oder die Mittel stehen nicht mehr zur Verfügung. Dadurch sinkt die Akzeptanz der Bevölkerung in die neuen Technologien. Außerdem werden die gesetzlichen Standards für Präventivmaßnahmen erhöht, was eine gezieltere Kontrolle der Technologien sicherstellt.

Visionen verwirklichen		
globaler Stromausfall	Anzahl der verabschiedeten Gesetze bleibt stabil über die Jahre	
	Bereitschaft der Menschen Daten preiszugeben ist gleichbleibend	
	Transparenzbereitschaft der Unternehmen steigt schwach an	
	Effizienz steigt mit moderaten Geschwindigkeiten an	
	Unternehmen verzeichnen Erfolge durch wachsende Kundenorientierung	
	Datenmengen stehen in vielerlei Formaten zur Verfügung	
	Möglichkeiten der Daten Akquisition sind überschaubar	
	Analyse wird vom Geheimtipp zum Modus Operandi; Visualisierung und Distribution der Daten steigt schwach an	
	Sicherheitsmaßnahmen im IT-Bereich orientieren sich am aktuellen Stand	
Σ	Akzeptanz der datenbasierten Entwicklung in der Bevölkerung ist geprägt von Vorsicht und Interesse im Gleichgewicht	

Tabelle 24: „Visionen verwirklichen“ mit Störereignis „globaler Stromausfall“

Im Trendszenario halten sich die vertretbare Zurückhaltung und das technologische Interesse an der datenbasierten Entwicklung die Waage. Dadurch kommt es im Falle eines Ausfalls zu nicht so starken Schwankungen innerhalb der Bevölkerung. Die *Wirtschaft* wird nichts desto trotz auch hier die meisten Verluste erleiden, da die Infrastruktur zu großen Teilen nicht mehr vorhanden ist. Die Bevölkerung ist aber nicht in dem Maßen auf sich allein gestellt, wie in den anderen Szenarien. Dadurch können viele Projektionen als gleichbleibend oder nur marginal verändert angesehen werden. So werden im Rahmen der Gesetzgebung keine ad hoc – Gesetze verabschiedet werden und auch die Datenpreisgabe wird im bisherigen Umfang weitergehen.

2. Künstliche Intelligenz

Als dritte WildS. Card wird der wachsende Einfluss von Systemen künstlicher Intelligenz in der Welt gesehen. Die erste bedeutende Entwicklung wird im Jahre 1997 öffentlich, als Garri Kasparow, damals amtierende Schachweltmeister von dem IBM Computer „Deep Blue“ in sechs Partien besiegt wird. Von da an ging die Entwicklung weiter, bis zum aktuellsten Beispiel dem IBM Watson. Die Programmierung erlaubt es diesem System selbst komplizierteste Ratespiele, wie das amerikanische Quiz-Show Spiel

„Jeopardy“ zu verstehen und gegen Menschen zu gewinnen.[163] Was bisher hauptsächlich in sportlichen Bereichen oder der Produktion von Maschinen erfolgreich angewendet wurde, kann mit Hilfe immer kürzeren Entwicklungszyklen auch im Automobilbereich beim selbstständigen Fahren beobachtet werden. Der Vorstellung sind keine Grenzen gesetzt, wie dem Buch „The Singularity Is Near“ von Raymond Kurzweil zu entnehmen ist.[164] Roboterähnliche Maschinen in Pflegeberufen oder bei der Kinderbetreuung sind Mensch-Maschine-Interaktionen auf höchstem technologischem Niveau, die in der Zukunft wahr werden könnten.

Datenkrieg		
künstliche Intelligenz	Anzahl der verabschiedeten Gesetze steigt an	
	Bereitschaft der Menschen Daten preiszugeben ist gering	
	Transparenzbereitschaft der Unternehmen stagniert	
	Effizienz geht zurück	
	Unternehmen beklagen wirtschaftliche Einbußen und Misserfolge	
	Datenmenge und –vielfalt stagniert	
	Möglichkeiten neue Daten zu gewinnen sehr eingeschränkt	
	Datenverarbeitung und -aufbereitung gleichbleibend	
	Sicherheitsmaßnahmen im IT-Bereich stark wachsend	
Σ	Akzeptanz der Datenanalysen in der Bevölkerung ist gering	

Tabelle 25: „Datenkrieg“ mit Störereignis „künstliche Intelligenz“

Die Systeme der künstlichen Intelligenz werden vermehrt als Datensammler eingesetzt. So gehen sie dem Menschen im Alltag zur Hand, dennoch beobachten sie die menschlichen Abläufe und passen sich immer weiter selbstständig an. Daraus resultiert ein wachsender Argwohn den Menschen den Systemen gegenüber. Nichts desto trotz benötigen die Menschen die maschinellen Helfer, da die wachsende Anzahl an Aufgaben am Tag vor allem in der Parallelität sonst nicht zu meistern ist. Auch

[163] Vgl. n-tv; dpa (2011)
[164] Vgl. Kurzweil (2013)

wenn die Sicherheitsmaßnahmen und Überwachungen der Systeme engmaschig ausgelegt sind, bleibt die Akzeptanz der Bevölkerung gering.

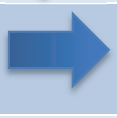
Bytes im Freudentaumel		
künstliche Intelligenz	Anzahl der verabschiedeten Gesetze sinkt	
	Bereitschaft der Menschen Daten preiszugeben ist sehr hoch	
	Transparenzbereitschaft der Unternehmen sinkt	
	Effizienz steigt mit unterschiedlichen Geschwindigkeiten an	
	Unternehmen erfreuen sich an wachsender Profitabilität durch Kundenfokus	
	Datenmengen stehen in vielerlei Formaten zur freien Verfügung	
	Möglichkeiten neue Daten zu gewinnen unbegrenzt	
	Analyse, Visualisierung und Distribution der Daten steigt stark an	
	Sicherheitsmaßnahmen im IT-Bereich stark abnehmend	
Σ	Akzeptanz der datenbasierten Entwicklung in der Bevölkerung ist immens	

Tabelle 26: „Bytes im Freudentaumel“ mit Störereignis „künstliche Intelligenz“

Aufgrund der Vereinfachung und Entlastung im Alltag, erfreuen sich die Systeme großer Beliebtheit in der Bevölkerung. Die *Wirtschaft* profitiert zudem von der unerschöpflichen Datenmenge, welche die Systeme in der Interaktion mit den Menschen sammeln. Lediglich die rechtlichen Aspekte müssen etwas verschärft werden, damit einem Missbrauch der Systeme entgegengewirkt werden kann.

Visionen verwirklichen		
künstliche Intelligenz	Anzahl der verabschiedeten Gesetze bleibt stabil über die Jahre	
	Bereitschaft der Menschen Daten preiszugeben ist gleichbleibend	
	Transparenzbereitschaft der Unternehmen steigt schwach an	
	Effizienz steigt mit moderaten Geschwindigkeiten an	
	Unternehmen verzeichnen Erfolge durch wachsende Kundenorientierung	
	Datenmengen stehen in vielerlei Formaten zur Verfügung	
	Möglichkeiten der Daten Akquisition sind überschaubar	
	Analyse wird vom Geheimtipp zum Modus Operandi; Visualisierung und Distribution der Daten steigt schwach an	
	Sicherheitsmaßnahmen im IT-Bereich orientieren sich am aktuellen Stand	
Σ	Akzeptanz der datenbasierten Entwicklung in der Bevölkerung ist geprägt von Vorsicht und Interesse im Gleichgewicht	

Tabelle 27: „Visionen verwirklichen“ mit Störereignis „künstliche Intelligenz“

Die Gesetzeslage verschiebt sich in die Präventivgesetzgebung, um Missbrauch der künstlichen Intelligenzen zu verhindern. Die Menschen arbeiten an vielen Schnittstellen mit den Systemen und sind sich den Fähigkeiten und der Notwendigkeit bewusst. Aufgrund der gleichmäßigen und auf Verständnis basierenden Entwicklung solcher Intelligenzen, im Gleichschritt mit der Menschheit, empfinden die Menschen die Systeme nicht als Bedrohung. So gibt es in der *Wirtschaft* keinen unerwarteten Effizienzschub, sondern ein gleichmäßiges Ansteigen von Fähigkeiten und Nutzungsmöglichkeiten.

Die Veränderungen innerhalb der Szenarien sind vor allem im Bereich der *Wirtschaft* als einheitlich zu betrachten. So verschieben sich die einzelnen Projektionen innerhalb dieses Treibers nicht gegeneinander. Anders im Bereich von *Know-how* oder *Konnektivität*. Hier sind auch gegenläufige Entwicklungen zu beobachten, wie z.B. in Tabelle 22 zu sehen.

6 Ableitung robuster Technologien und Implikationen für das strategische Management

Zusammenfassung

Innerhalb des Szenario-Transfers sollen robuste Technologien abgleitet werden, die über alle Szenarien Bestand haben. Hierfür wird die Unternehmensperspektive eingenommen, zukünftige Marktsegmente gebildet und mögliche Produktfelder abgeleitet. Zu den robusten Technologien zählen laut vorliegender Analyse die Basis-Software-Pakete in Form von Programmiersprachen (Java, C-Familie, Python) und Infrastrukturen (*Hadoop*, *MapReduce*) sowie kundenspezifische Lösungen, zumeist im Bereich der Analyse von großen Datenmengen.

Zusätzlich wird eine Roadmap für die nächsten 5 Jahre formuliert, in der die Entwicklung folgender drei Bereiche „Datenbank-Systeme“, „Verarbeitung großer Datenmengen“ und „Echtzeitanalysen“ aufgezeigt wird.

Außerdem werden Implikationen für das strategische Management abgleitet. Im Rahmen des Szenarios „Visionen entwickeln“ vor allem Aufgaben im Bereich der Mitarbeiterschulung und der Thematisierung des Themas „Big Data“. Auch in den Kundenbeziehungen sollte das Thema priorisiert werden, um die Akzeptanz neuer Technologien im *Big Data*-Kontext zu erhöhen.

Ziel dieses Schrittes ist es, emergente Technologien zu identifizieren, welche über alle drei Szenarien Bestand haben werden. Diese Technologien werden als robust bezeichnet. Quellen hierfür sind die Erkenntnisse aus der bibliometrischen Analyse und der vorangegangenen Schritte der Szenariotechnik.

Methodisch wird ein strategisches Roadmapping durchgeführt. Hierzu werden innerhalb der drei Marktszenarien Marktleistungen und Produktfelder für ein Logistikunternehmen bestimmt, denen emergente Technologien zugeordnet werden. Die emergenten Technologien werden aus der bibliometrischen Analyse abgeleitet und innerhalb von Technologieklassen bestimmt.

6.1 Emergente Technologien ableiten

Im Bereich der Autorenanalysen wird deutlich, dass kundenspezifische Lösungen auf unterschiedlichen Software-Systemen sehr gefragt sind (vgl. Kapitel 4.2.1, S. 33, Kapitel 4.2.2, S. 43). Inzwischen weit verbreitete Systeme wie *Hadoop*, aufbauend auf *MapReduce*, seien an dieser Stelle als herausragende Wegbereiter der nachfolgenden Software erwähnt (vgl. Kapitel 4.2.1, S. 35). Es gibt also Basis-Software-Pakete und individuelle, auf Kundenwünsche zugeschnittene Anwendungen, die sich der Struktur der Basispakete bedienen.

Darüber hinaus kann jegliche Software entsprechend ihrem primären Ziel klassifiziert werden. Dies kann den Fachrichtungen in Kapitel 4.2.2 entnommen werden. Für die Klassifizierung können die Begriffe „Hardware & Architecture“, „Theory & Methods“, „Software Engineering“ und „Information Systems“ bzw. „Interdisciplinary Applications“ verwendet werden. In Abbildung 52 ist eine Zuordnung von Anwendungen zu den entsprechenden Klassen zu sehen. Hierbei erfolgte

die Zuordnung der Teilbereiche in der gleichen Reihenfolge wie die Klassen. Somit wird inhaltlich „Hardware & Architecture“ zu „Hardware“, „Theory & Methods“ zu „Big Data Distributions“, „Software Engineering“ zu „Data Management Components“, „Information Systems“ zu „Analytics and Visualizations“ und „Interdisciplinary Applications“ zu „Services“ zugeordnet. Damit ergeben sich die emergenten Technologien zu den fünf, in Abbildung 52, dargestellten Säulen.

Hardware	Big Data Distributions	Data Management Components	Analystics and Visualizations	Services
•Storage •Servers •Networking •Vendors: Dell, HP, IBM, Cisco	•Community Hadoop distribution •Enterprise Hadoop distributions •Non-Hadoop Big Data frameworks •Vendoers: Cloudera, IBM, MapR, LexisNexis, Microsoft	•NoSQL databases •Data integration •Data quality and governance •Vendors: DataStax, IBM, Informatica, Syncsort	•Analytic development platforms •Advanced analytics applications •Data visualization tools •BI applications •Vendors: Karmaspere, Tresata, Datameer, SAS	•Consulting •Training •Software maintenance •Hardware maintenance •Hosting/cloud •Vendors: Think Big Analytics, Amazon, Web Services Accenture

Abbildung 52: Technologien der Anwendungsfelder von Big Data[165]

Es ergibt sich für ein Logistikunternehmen in jedem Szenario eine spezifische Kombination an Marktleistungen und dazugehörigen Technologien. Danach werden aus den bereits gewonnenen Erkenntnissen zukünftig notwendige Kompetenzen abgeleitet, mit denen die Unternehmen ihre bestehenden Geschäftsmodelle je nach Szenario neu ausrichten können. Abschließend werden für die szenariospezifischen Marktleistungen Technologien aus der vorangegangen Technologievorausschau herangezogen, die für deren Umsetzung hilfreich bzw. notwendig sind. Daraus ergeben sich für die Szenarien die Übersichten in Tabelle 28.

[165] Eigene Darstellung in Anlehnung an Kalakota (2011)

Szenarien	„Bytes im Freudentaumel“	„Visionen verwirklichen“	„Datenkrieg“
Markt	**Kanal:** Webshop **Kunden:** überregionale und lokale Kunden **Wettbewerber:** internationale Investorengruppen	**Kanal:** Webshop, persönlicher Kontakt vor Ort **Kunden:** überregionale und lokale Kunden **Wettbewerber:** lokale und internationale Versandhäuser	**Kanal:** persönlicher Kontakt vor Ort **Kunden:** lokale Kunden **Wettbewerber:** lokale Versandshops
Marktleistungen	Vorausschauende Lieferung von jeglichen Waren auf allen Frachtwegen Global umspannendes Logistiknetz Online Präsenz Digitaler Support und Vertragsabwicklung	Lieferung von jeglichen Waren auf allen Frachtwegen Kooperationsnetzwerk Online Shop Individuelle, persönliche Vertragsabwicklung	Lieferung von Waren auf von Menschen kontrollierbaren Wegen Kleines, überschaubares Netz Flyer und Printmedien Persönliche Vertragsabwicklung
Technologien	**Datenbanksysteme** **Verarbeitung großer Datenmengen** **Echtzeitanalysen** Visualisierung & Analysen *Big Data* Verteilsysteme	**Datenbanksysteme** **Verarbeitung großer Datenmengen** **Echtzeitanalysen** Services	**Datenbanksysteme** **Verarbeitung großer Datenmengen** **Echtzeitanalysen** Datenmanagement Komponenten

Tabelle 28: Strategische Roadmap für Big Data

6.2 Robuste Technologien identifizieren und den Einfluss auf jedes Szenario darstellen

Die emergenten Technologien werden nun priorisiert und als robuste Technologie bezeichnet, falls sie in jedem Marktszenario vorhanden sind. Ausgehend von den verschiedenen Schwerpunkten innerhalb der einzelnen Szenarien, lassen sich nicht alle Technologien übergreifend als gleich wichtig bezeichnen. Als robuste Technologien aus der Bibliometrie lassen sich allerdings die Programmiersprachen Java, C (C++, C#), Python und die Internetprogrammiersprachen JavaScript und PHP identifizieren. Darauf aufbauend werden die Datenbanksysteme, Echtzeitanalysen und die Verarbeitung großer Datenmengen ermöglicht. Diese robusten Technologien sind in Tabelle 28 fett gedruckt.

Innerhalb der nächsten Jahre wird die Größe und Zuverlässigkeit der Datenbanken entscheidend sein. Darüber hinaus tritt eine zunehmende Automatisierung ein, die zu selbst denkenden Datenbanken führt. Innerhalb des Szenarios „Visionen entwickeln“ ist die Zukunft mehr auf das Verständnis der Datenbanken ausgerichtet und auf das Miteinander weniger auf ein unkontrolliertes Handeln der Systeme. Daran schließt die Verarbeitung großer Datenmengen an, die im Falle von „Bytes im Freudentaumel“ mit Hilfe von immer neuen Programmen bewältigt wird. Innerhalb der anderen beiden Szenarien geschieht das weniger auf die Nachfrage der Menschen hin, sondern im Fall von

„Datenkrieg“ durch inkrementelle Veränderungen, die die Programme selbst vornehmen und im Fall von „Visionen entwickeln“ nur nach sorgfältiger Prüfung der Notwendigkeit.

In allen drei Szenarien sind die Echtzeitanalysen sehr wichtig. Dennoch gibt es auch hier Abstufungen in den Anforderungen an die Programme. Während im „Datenkrieg“ wieder die Programme selbst die Geschwindigkeit bestimmen und Verbesserungen einführen, ohne dass der Mensch aktiv davon in Kenntnis gesetzt wird, steht das Verständnis im Szenario „Visionen verwirklichen“ nach wie vor an erster Stelle. „Bytes im Freudentaumel“ lässt den Systemen mehr Spielraum, ist aber hauptsächlich an der Verkürzung der Zeiträume interessiert, wodurch der technologische Druck hier am höchsten ist.

6.3 Implikationen für das strategische Management ableiten und Maßnahmen bestimmen

Ausgehend von den Erkenntnissen aus dem vorherigen Szenario-Transfer, werden im letzten Schritt der Szenariotechnik Handlungswege aufgezeigt und strategische Maßnahmen für ein Logistikunternehmen, ein Spezialmaschinenbauer (Voith Paper) und einen Gesamtlösungsanbieter (Siemens) abgeleitet. Hierbei geht es nun vor allem um das Szenario „Visionen verwirklichen“, um ein möglichst reales Bild zu zeichnen.

Die Entwicklung hin zu einer wachsenden Anzahl von softwareunterstützten Entscheidungen ist eindeutig innerhalb der deutschen und industriell fortgeschrittenen Welt erkennbar. Die Entwicklung kann in fünf Phasen (Roadmap, Pilot, Engineering, Applications, Automation)[166] eingeteilt werden. Dabei geht man von zwei Extremen aus. Zum einen von der klassischen Geschäftsführung, deren Datenkonsum auf Recherchen, Controlling und Feedbackbögen beschränkt ist. Zum anderen von der durch Datenanalysen getriebenen Firma, deren Entscheidungen ausschließlich auf Analysen gestützt ist. Zwischen diesen beiden Extremen lassen sich aus den Ergebnissen der Szenarien die oben genannten drei Unternehmenstypen einordnen und Chancen und Risiken ableiten.

Für das Logistikunternehmen besteht ein Risiko darin, dass eine neu verwendete Technologie nahezu fehlerfrei funktionieren muss, da davon der gesamte Betrieb abhängig ist. Auf der anderen Seite sind die internationalen Kontakte und Kooperationspartner als Chance zu verstehen, die das Risiko aufgrund von Erfahrungsaustausch abmildern können.

Im Spezialmaschinenbau, z. B. bei Voith Paper, liegt das Risiko eher auf der organisatorischen denn auf der technologischen Seite. Die traditionsbewusste und ingenieurswissenschaftliche Kultur kann die Entrepreneurship-Kultur und iterative Prozesse behindern. Als Chance wiederum sind für den Markt ganz neue Ansätze zu nennen, die – wenn sie einmal im Unternehmen angenommen werden – zu Differenzierungsmerkmalen gegenüber dem Wettbewerb führen können. Dazu gehören nicht nur Lösungen für den Kunden, sondern auch für interne Prozesse.

An dem Beispiel Siemens als Gesamtlösungsanbieter wird deutlich, dass mehr als ein Standbein zu haben von Vorteil ist. Dadurch ist das Unternehmen nicht so abhängig von einer Technologie wie kleinere Betriebe und zusätzlich besteht die Chance der innerbetrieblichen Synergienutzung. Ein Risiko ist die zu prozessgetriebene Umsetzung, woraus lange Implementierungszeiten bedeutet, was auch für das Change Management zutrifft.

[166] Vgl. Think Big (2015)

Aktuell befindet sich vor allem die deutsche Wirtschaft in den unteren Stufen. So startete Voith im Jahre 2015 die Initiative Papermaking 4.0 für die Division Voith Paper und befindet sich zwischen einem *Big Data* Pilot und dem *Big Data* Engineering. In einem Ende 2015 durchgeführten Interview mit Henning Reich von Siemens[167] bescheinigte dieser innerhalb des Konzerns unterschiedliche Entwicklungsstufen. Grundsätzlich sei das Bewusstsein für jeden im Betrieb wichtig, so Reich. Dies gelte für alle Business Units und führt dazu, dass auch in fast allen Business Units bereits das Level der Roadmap erreicht sei. Darüber hinaus gebe es mit den „Sinalytics" den Ansatz, eine interne Wissensbasis für Lösungen im Bereich *Big Data* zu schaffen. Davon profitiere schon die Medizintechnik, so Reich weiter, die mithilfe von *Computer Aided Diagnostics* (CAD) Ärzten als Entscheidungshilfen zur Verfügung stehe. Auch im Bereich der Gasturbinen werde eine weitere Technik, das *Condition Monitoring* (CM) zur Zustandsüberwachung angewendet. Beide Bereiche bedienen sich zusätzlich des *Cloud Computings*, wodurch ein Remote-Zugang möglich sei, schloss Reich. Aktuell im Übergang zwischen der Pilot- und der Engineering-Phase befinde sich der Mobility-Sektor, da dort auch aufgrund des öffentlichen Interesses intensiv an Verbesserungen gearbeitet werde. Grundsätzlich, so Reich, werde man versuchen in den Bereichen *Machine Learning*, *Cloud Computing* und Zustandsüberwachungen Datenmengen zu generieren, um dann mit geeigneten Analysen dem Kunden Optimierungspotenziale aufzuzeigen.

Für alle Firmen, besonders für den langfristigen Erfolg dieser Entwicklung und reibungslose Übergänge in die nächsten Phasen, ist es unumgänglich so früh wie möglich die Förderung der fachspezifischen Berufe einzuleiten. Dazu gehören neben der Implementierung von ganz neuen Ausbildungsmöglichkeiten auch der Weiterbildungsmöglichkeiten der aktuell eingestellten Kräfte. Zusätzlich muss ein Unternehmen den Kundenstamm nachhaltig pflegen, um möglichst viele Kundendaten zu gewinnen. Besonders in der Anfangsphase von Analyseprogrammen müssen eventuell Kundendaten manuell eingefügt werden. Dieser Aufwand muss frühzeitig eingeplant oder gar proaktiv vermieden werden. Wie Jörg Bald, Geschäftsführer bei der Waldemar Winckel GmbH & Co. KG 2014 in einem Interview erklärte: „Informationen sind das Öl der Zukunft"[168] Zudem muss über die Sicherheit aufgeklärt werden, da viele Menschen innerhalb der Bevölkerung eine Abneigung gegen zu große Transparenz hegen.

[167] Interview mit Henning Reich am 04.12.2015 im Rahmen des CE ProMM in Ravensburg
[168] Vgl. Bald (2014)

7 Fazit und Ausblick

Industrie 4.0 ist als Megatrend für die zukünftige Ausrichtung von Wirtschaft und Gesellschaft nicht mehr wegzudenken. Ausgehend von der globalen bibliometrischen Analyse und der Szenariotechnik haben sich als wichtigste Erkenntnisse vor allem der Umgang mit der *IT-Sicherheit* und dem *Datenschutz* herauskristallisiert. Auch wurde die thematische Vielfalt des Begriffes *Big Data* anhand der deskriptiven Statistiken deutlich (vgl. Kapitel 4.2.3, S. 51) und der Wissensbasen deutlich (vgl. Kapitel 4.3.1). Neben theoretischer Forschung im Bereich Software-Architekturen und Programmiersprachen wurden auch die Sozialwissenschaften beleuchtet. Die Schlüsseltechnologien für eine erfolgreiche Umsetzung von *Big Data* in beispielsweise einem Logistikunternehmen befinden sich im theoretischen Cluster des Forschungsportfolios und lauten Datenbanksysteme, Verarbeitung großer Datenmengen und Echtzeitanalysen. In diesen Bereichen muss das Unternehmen Kompetenzen aufbauen, um *Big Data* in Zukunft gewinnbringend nutzen zu können.

Auffällig ist die Diskrepanz zwischen den Ergebnissen der Methode im Bereich der Datensicherheit. Dieser Umstand erschwert es vor allem den deutschen Unternehmen im Hinblick auf die vierte industrielle Revolution den Anschluss an den globalen Wettbewerb zu meistern. Kooperationen müssen vorangetrieben werden, sodass die *Big Data* - Architekturen auf internationales Niveau, wie bei den erwähnten Firmen aus dem Silicon Valley, angehoben werden können. Darüber hinaus wird in der Zukunft nicht nur das Interesse am eigentlichen Programmieren solcher Systeme, sondern auch das Verständnis der Ergebnisse ansteigen. Bereits heute müssen Ausbildungsberufe und Studiengänge auf diese wachsende Nachfrage ausgerichtet werden.

Daraus entstehen innerhalb der Unternehmen neue Geschäftsfelder, die den Fortbestand des Unternehmens sichern. Neben dem Change-Management innerhalb der Unternehmen muss auch die Aus- und Weiterbildung der Angestellten einen höheren Stellenwert in der Zukunft einnehmen. Der Wandel soll nicht zu einer Entfremdung von bestehenden Werten führen, sondern die Offenheit für neue Modelle zulassen und so die Aufnahme von *Big Data* - Technologien erleichtern. Dazu gehört als ganz konkrete Aufgabe für die nächsten 24 Monate, den vorhandenen Mitarbeiterstamm mit der Thematik vertraut zu machen, die Angst zu nehmen und gleichzeitig eine Begeisterungskultur für Neues zu schaffen.

8 Anhang

Projektionskatalog:

Einflussbereich Gesellschaft	
Treiber: Datenschutz	
1 Anzahl herausgebrachter Gesetze	
Neuerungen im Datenschutzgesetz, Neuerungen zu Sonderfällen	A exponentiell steigend Die Nachfrage nach Sicherheit im Rechtlichen erhöht sich.
	B langsam steigend Die rechtlichen Instanzen reagieren auf akute Gesuche nach Gesetzen aus Gesellschaft und Politik.
	C abfallend Gesetze werden weder ernst genommen noch benötigt.
2 Liberalisierungsgrad	
vom Kunden/Nutzer freiwillig bereitgestellte Daten	A stark steigend Die Kunden vertrauen der wachsenden Individualisierung und den Analysen.
	B gleichbleibend Vorsicht und Notwendigkeit der Datenpreisgabe halten sich die Waage.
	C abfallend Das Vertrauen der Kunden ist nicht mehr vorhanden oder schwindet rapide.
3 Grad der Transparenz	
Grad der Transparenz aus Sicht des breitstellendes Unternehmen	A schwach steigend Die Unternehmen geben freiwillig oder aufgrund äußerer Umstände Einsicht in Prozessabläufe.
	B schwach fallend Die Unternehmen versuchen ihre Betriebsabläufe weitestgehend unter Verschluss zu halten.
	C stark abfallend Mit allen Mitteln wehren sich die Unternehmen gegen eine Offenlegung von Abläufen und versuchen diese vermehrt zu schützen.
Treiber: Gesetzgebung	
1 Anzahl herausgebrachter Gesetze	
allgemeine Gesetze, alles was nicht in den Bereich Datenschutz fällt	A ansteigend und gleichbleibend Aufgrund neuer Einflüsse werden Gesetze erlassen; anschließend stagniert die Anzahl.
	B gleichbleibend Die rechtlichen Instanzen verabschieden in bekannten Größenordnungen.

	C stark abfallend Gesetze werden weder ernst genommen noch benötigt.
2 politische Rahmenbedingungen	
interne und externe Zusammenschlüsse von Staaten	A stark steigend Der Druck auf einen Staat nimmt von Innen und Außen zu.
	B gleichbleibend Periodische Schwankungen der Beziehungen mit eingerechnet, gibt es keinerlei Druckausübung.
	C sprunghaft abfallend Die Staaten haben innerhalb und außerhalb keinerlei Berührungspunkte.

Tabelle 29: Projektionskatalog für den Einflussbereich Gesellschaft

Einflussbereich Wirtschaft	
Treiber: Effizienz	
1 Prozessoptimierung	
Verminderung von Kosten durch optimierte interne und externe Prozesse	A stufenweiser Anstieg Es werden in definierten Abständen die Prozessverbesserungen eingeführt.
	B S-förmiger Anstieg Nach einer Anlaufphase und einem Aufschwung geraten die Optimierungen allmählich ins Stocken.
	C stark fallend Aufgrund mangelnder Pflege oder Weiterentwicklung ist keine Optimierung möglich.
2 Risikobewertung	
Verbesserung der Bewertung von Risiken mit Hilfe von Datenanalysen	A exponentiell steigend Aufgrund wachsender Konkurrenz wird die Bewertung von wirtschaftlichen Risiken immer wichtiger.
	B linear steigend Die Risikobewertung komplettiert das Gesamtpaket der Analysen, ist aber nicht allein ausschlaggebend.
	C abfallend Risiken sind entweder unkalkulierbar geworden oder so unwichtig, dass die Investitionen woanders sinnvoller eingebracht werden können.
3 Entscheidungen	
Big Data Tools als Entscheidungsgrundlage	A linear steigend

	Aufgrund wachsender Konkurrenz wird die Unterstützung der Entscheidungsfindung mit *Big Data* Tools immer wichtiger.
	B Wurzelfunktion Nach anfänglicher Euphorie stagniert die Entwicklung.
	C einbrechend Die Analysen waren zu unzuverlässig, als dass weiter damit gearbeitet werden konnte.
Treiber: wirtschaftlicher Erfolg	
1 Profitabilität	
Profit pro verkaufter Einheit	A starke Wurzelfunktion Die neuen Technologien schöpfen ungeahnte Potentiale aus den bisherigen Abläufen.
	B schwache Wurzelfunktion Einige Schwachstellen der Technologien führen zu Unregelmäßigkeiten und Einbußen beim Profit.
	C stark abfallend Die Analysen waren zu unzuverlässig, als dass weiter damit gearbeitet werden konnte.
2 Preisgestaltung	
Preisanpassung in Echtzeit für maximalen Absatz	A linear steigend Echtzeitanalysen ermöglichen tagesgenaue Produktplatzierung und optimale Ausschöpfung der Potentiale.
	B gleichbleibend Es können keine großen Zeitsprünge erwartet werden.
	C stark abfallend Mit unflexiblen Analysen kann nicht weiter gearbeitet werden.
3 Kundenorientierung	
individuellere und angepasste Implementierungen am Markt	A starke Wurzelfunktion Der Aufschwung durch neueste Analysen aus den sozialen Netzwerken und dem Internet ist überall zu spüren.
	B ansteigend und abfallend Nach anfänglicher Euphorie stagniert die Analysekompetenz.
	C stark abfallend Die Analysen waren zu ungenau, als dass weiter damit gearbeitet werden konnte.

Tabelle 30: Projektionskatalog für den Einflussbereich Wirtschaft

Einflussbereich Technik	
Treiber: Konnektivität	
1 Datenmenge	
Anzahl von datengenerierenden Schnittstellen	A stark linear steigend Die Anzahl verbundener Endgeräte steigt rasant.
	B exponentiell steigend Der Anstieg erfolgt zeitverzögert aufgrund von diversen Kundennutzungsverhalten.
	C gleichbleibend Die Bereitschaft der wachsenden Nachfrage an Daten durch mobile Endgeräte gerecht zu werden ist gering.
2 Vielfalt	
unterschiedlichste Datentypen	A schwache Wurzelfunktion Die aktuelle Entwicklung macht einige neue Datentypen notwendig. Danach ist keine Steigerung mehr zu verzeichnen.
	B steigende S-Kurve Nach einer Anlauf- und Wachstumsphase, stagniert die Anzahl der Dateiformate.
	C gleichbleibend Trotz neuer Technologien sind keine weiteren Formate notwendig.
3 Daten Akquisition	
aus allen zur Verfügung stehenden Quellen	A stark ansteigend De Anzahl verbundener Geräte steigt rasant.
	B gleichbleibend Trotz neuer Entwicklungen sind keine maßgeblichen Hortungen von Daten ersichtlich.
	C stark fallend Die Daten sind fehlerhaft oder schädlich und zerstören im Extremfall die akquirierenden Systeme.
Treiber: Know-how	
1 Analyse der Daten	
für Entscheidungsfindung, Vergangenheitserklärung und Prognosen	A stark steigend Start-ups und private Programmieren ebnen den Weg für immer neue Analysen.
	B schwach steigend Nur eine geringe Anzahl an Firmen widmet sich der Entwicklung neuester Tools.
	C gleichbleibend

	Die Notwendigkeit von Analysesystemen ist gering und die Entwicklung begrenzt.
2 Visualisierung der Daten	
eingängige Ergebnispräsentation	A stark linear steigend Nachfrage nach eingängigen Visualisierungen ist sehr groß.
	B gleichbleibend Die bisherigen Methoden reichen aus.
	C abfallend Die präferierte Variante sind klassische Zahlen und Fakten. Die Visualisierung gerät ins Hintertreffen.
3 Distribution	
einheitliche Plattformen zur Weitergabe der generierten Daten	A linear steigend Die Vereinheitlichung und der schnelle Zugang zu Daten gewinnen immer mehr an Priorität.
	B exponentiell steigend Nach anfänglichen technischen Problemen, können schließlich einheitliche Plattformen bereitgestellt werden.
	C gleichbleibend Die Nachfrage ist nicht vorhanden oder die technischen Möglichkeiten sind eingeschränkt.
Treiber: IT-Sicherheit	
1 Verschlüsselung	
Chiffrierung ohne Verfremdung	A linear steigend Wachsendes Sicherheitsbedürfnis begründet den Anstieg.
	B exponentiell steigend Die Notwendigkeit wurde zeitversetzt erkannt.
	C abfallend Das Vertrauen in die Sicherheit der nicht verschlüsselten Systeme wächst, das Gefühl des ungehinderten Austausches setzt sich durch.
2 Data Masking	
Anonymisierung bzw. Verfremdung von Daten mit Hilfe von *Big-Data-Technologien*	A linear steigend Wachsendes Sicherheitsbedürfnis begründet den Anstieg.
	B Wurzelfunktion Nach anfänglicher Begeisterung bleibt das Niveau gleich.
	C gleichbleibend Der State oft he Art ist ausreichend um ein sicherer Gefühl innerhalb der Datenwelt zu erfahren.

3 Identitäts- und Zugangsmanagement

Verwaltung von Benutzern innerhalb eines *Big Data* Systems und deren Zugriffsberechtigungen	A stark steigend Das Verlangen nach Sicherheit ermöglicht den Sicherheitsfirmen immer weiter Programme und Methoden zu entwickeln.
	B schwach steigend Die Technik ist bereits auf einem guten Stand, was zu einer nur mäßigen Nachfrage führt.
	C abnehmend Das Vertrauen in die Systeme wächst, das Gefühl des ungehinderten Zugangs setzt sich durch.

Tabelle 31: Projektionskatalog für den Einflussbereich Technik

Literaturverzeichnis

ACATECH (2011)
ACATECH (2011): *Cyber-Physical Systems. Innovationsmotoren für Mobilität, Gesundheit, Energie und Produktion* (acatech POSITION), Heidelberg u.a.: Springer Verlag 2011

Apache (2015)
Apache Software Foundation (2015): http://www.apache.org/, zuletzt geprüft am 20.06.2015

Bald (2014)
Bald, Jörg (2014): RFID-Systemintegration: Jörg Bald, Geschäftsführer Waldemar Winckel, im Gespräch mit „RFID im Blick“
http://www.rfid-im-blick.de/de/201411122395/rfid-systemintegration-joerg-bald-geschaeftsfuehrer-waldemar-winckel-im-gespraech-mit-rfid-im-blick.html, zuletzt geprüft am 03.08.2016

Baron (2013)
Baron, Pavlo (2013): Big Data für IT-Entscheider. Riesige Datenmengen und moderne Technologien gewinnbringend nutzen: Carl Hanser Verlag GmbH Co KG

Barton und Court (2012)
Barton, Dominic; Court, David (2012): Making advanced analytics work for you. In: *Harvard business review* (Vol. 90, No. 10), S. 78–83
https://hbr.org/2012/10/making-advanced-analytics-work-for-you, zuletzt geprüft am 20.06.2015

BCG (2015)
BCG (2015): Porfoliomatrix. Die BCG-Porfoliomatrix unter Mitarbeit von Bruce Henderson. Hg. v. Boston Consulting Group
http://www.bcg.at/bcg_osterreich/geschichte/klassiker/portfoliomatrix.aspx, zuletzt geprüft am 09.07.2015

BITKOM (2012)
BITKOM (2012): Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. BITKOM-Arbeitskreis Big Data. Berlin

BITKOM (2014)
BITKOM (2014): Big-Data-Technologien - Wissen für Entscheider. BITKOM Leitfaden. Berlin

BMBF (2014)
BMBF - Bundesministerium für Bildung und Forschung, (2013): Zukunftsbild "Industrie 4.0", Bonn

Boyd und Crawford (2012)
Boyd, Danah; Crawford, Kate (2012): Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. In: *Information, Communication & Society* 15 (5), S. 662–679

Brecht (2015)
Brecht, Leo (2015): Prozessmanagement III: Business Engineering und Analytics. Vorlesungsskript. Universität Ulm, Ulm. Institut für Technologie- und Prozessmangement

Brynjolfsson und McAfee (2014)
Brynjolfsson, Erik; McAfee, Andrew (2014): The second machine age. Work, progress, and prosperity in a time of brilliant technologies. First Edition. New York: W. W. Norton & Company

Burt et al. (2006)
Burt, G.; Wright, G.; Bradfield, R.G.; van der Heijden, K. (2006): The role of scenario planning in exploring the environment in view of the limitations of PESTEL and its derivatives. Int. Stud. Manage. Organ. 36 (3), S. 50–76

Bushman et al. (2013)
Bushman, Frederic D.; Barton, Spencer; Bailey, Aubrey; Greig, Caitlin; Malani, Nirav; Bandyopadhyay, Sourav et al. (2013): Bringing it all together. In: *AIDS* 27 (5), S. 835–838

Cavoukian, Jonas (2012)
Cavoukian, Ann; Jonas, Jeff (2012): Privacy by Design in the Age of Big Data. In: *Privacy by Design* https://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf, zuletzt geprüft am 23.06.2015

CGC (2013)
CGC International Conference on Cloud and Green Computing (2013): Karlsruhe, Germany

Chang et al. (2014)
Chang, Yu-Wei; Hsu, Ping-Yu; Wu, Zeng-Yuan (2014): Exploring managers' intention to use business intelligence. The role of motivations. In: *Behaviour & Information Technology* 34 (3), S. 273–285

Chelmis et al. (2013)
Chelmis, Charalampos; Wu, Hao; Sorathia, Vikram; Parasanna, Viktor K. (2013): Semantic Social Network for the Enterprise. Hg. v. Computing and Informatics–Special Issue on Computational Intelligence for Business Collaboration

Chen (1976)
Chen, Peter Pin-Shan (1976): The Entity-Relationship Model - Toward a Unified View of Data. In: *ACM Transactions on Database Systems* Vol. 1 (No. 1), S. 9–36

Computerwoche (2012)
Computerwoche (2012): Big Data - Nutzen für verschiedene Geschäftsbereiche Deutschland 2012 | Umfrage. Hg. v. Statista - Das Statistik Portal
http://de.statista.com/statistik/daten/studie/257964/umfrage/umfrage-zum-nutzen-von-big-data-fuer-verschiedene-geschaeftsbereiche/, zuletzt geprüft am 10.06.2015

Dean und Ghemawat (2008)
Dean, Jeffrey; Ghemawat, Sanjay (2008): MapReduce: simplified data processing on large clusters. In: *Commun. ACM* 51 (1), S. 107

Deutscher Bundestag (2013)
Deutscher Bundestag (2013): Anzahl der vom Bundestag verabschiedeten Gesetze von 1949 bis 2013 (7. Wahlperiode bis 17.)Wahlperiode). Hg. v. Statista - Das Statistik Portal http://de.statista.com/statistik/daten/studie/155525/umfrage/anzahl-der-gesetzesbeschluesse-auf-initiative-von-bundestag-bundesrat-und-bundesregierung/, zuletzt geprüft am 20.06.2015

DUDEN (2013)
DUDEN (2013): Data-Mining. Hg. v. Bibliographisches Institut GmbH http://www.duden.de/rechtschreibung/Data_Mining, zuletzt geprüft am 20.06.2015

Elsberg (2012)
Elsberg, Marc (2012): Blackout. Morgen ist es zu spät. Blanvalet Verlag; 1. Auflage, München

EMC (2015)
EMC Corporation (2015): Daten - Volumen der weltweit generierten Daten bis 2020 (in Exabyte). Hg. v. Statista - Das Statistik Portal http://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/, zuletzt geprüft am 10.06.2015

EY (2013)
EY (2013): Welche drei Städte weltweit bieten die besten Voraussetzungen, die "Googles" der Zukunft hervorzubringen? Hg. v. Statista - Das Statistik Portal http://de.statista.com/statistik/daten/studie/249423/umfrage/umfrage-zu-zukuenftig-wichtigen-standorten-fuer-technologieunternehmen/, zuletzt geprüft am 17.06.2015

Fabian et al. (2015)
Fabian, Benjamin; Ermakova, Tatiana; Junghanns, Philipp (2015): Collaborative and secure sharing of healthcare data in multi-clouds. In: *Information Systems* 48, S. 132–150

Fawcett und Waller (2014)
Fawcett, Stanley E.; Waller, Matthew A. (2014): Supply Chain Game Changers—Mega, Nano, and Virtual Trends—And Forces That Impede Supply Chain Design. In: *J Bus Logist*

Gartner (2013)
Gartner (2013): IT Glossary - Big Data Definition http://www.gartner.com/it-glossary/big-data, zuletzt geprüft am 18.06.2015

Gartner Inc. (2014)
Gartner Inc. (2014): Hype Cycle for Emerging Technologies 2014. Stanford, USA http://www.gartner.com/newsroom/id/2819918, zuletzt geprüft am 24.06.2015

Google (2015a)
Google Trends (2015a): Websuche - Big Data. Google Inc.
https://www.google.de/trends/explore#q=Big%20Data, zuletzt geprüft am 18.06.2015

Google (2015b)
Google Trends (2015b): Websuche - Industrie 4.0. Hg. v. Google Inc.
https://www.google.de/trends/explore#q=Industrie%2040, zuletzt geprüft am 16.06.2015

Google Scholar (2015a)
Google Scholar (2015a): H-Index, Jinjun Chen
https://scholar.google.com.au/citations?user=mQ9vJ24AAAAJ&hl=en, zuletzt geprüft am 21.06.2015

Google Scholar (2015b)
Google Scholar (2015b): H-Index, Yi-Ke Guo
https://scholar.google.com/citations?user=wZXp3kMAAAAJ&hl=de, zuletzt geprüft am 21.06.2015

Google Scholar (2015c)
Google Scholar (2015c): H-Index, Jian Li
https://scholar.google.com.au/citations?user=MbgxsAwAAAAJ&hl=en, zuletzt geprüft am 21.06.2015

Google Scholar (2015d)
Google Scholar (2015d): H-Index, Volker Markl
https://scholar.google.com/citations?user=el2k_oIAAAAJ&hl=de, zuletzt geprüft am 21.06.2015

Google Research (2015)
Google Research (2015): Jeff Dean
http://research.google.com/people/jeff/, zuletzt geprüft am 21.06.2015

Greive (2015)
Greive, Martin (2015): Grexit oder nicht? Die ultimative Pro- und Contra-Liste. Hg. v. Die Welt
http://www.welt.de/wirtschaft/article142740290/Grexit-oder-nicht-Die-ultimative-Pro-und-Contra-Liste.html, zuletzt geprüft am 21.06.2015

Havemann (2009)
Havemann, Frank (2009): Einführung in die Bibliometrie. 1. Aufl. Berlin: Ges. für Wiss.-Forschung c/o Inst. für Bibliotheks- und Informationswissenschaften

Heilmann und Liegl (2013)
Heilmann, Dirk; Liegl, Thomas (2013): Big Data und Datenschutz. Der Umgang der Deutschen mit persönlichen Daten und die Konsequenzen für den Einsatz von Big Data-Analysen. In: *Handelsblatt Research Institute*

Hide (2015)
Hide, Winston (2015). Department of Biostatistics. Harvard University
http://www.hsph.harvard.edu/winston-hide/, zuletzt geprüft am 23.06.2015

Howe et al. (2008)
Howe, Doug; Costanzo, Maria; Fey, Petra; Gojobori, Takashi; Hannick, Linda; Hide, Winston et al. (2008): Big data: The future of biocuration. In: *Nature* 455 (7209), S. 47–50

Hutter (2013)
Hutter, Thomas (2013): Facebook: 240 Milliarden Fotos – über 7 Petabytes Foto-Daten pro Monat http://www.thomashutter.com/index.php/2013/01/facebook-240-milliarden-fotos-uber-7-petabytes-foto-daten-pro-monat/, zuletzt geprüft am 20.06.2015

IEEE (2013)
IEEE International Conference on Big Data (2013): Silicon Valley, CA, USA

Ituski et al. (2013)
Ituski, Hiroshi; Matsubara, Hitoshi; Arita, Kazuki; Omi, Kazunari (2013): Effective Clusterization of Political Tweets Using Kurtosis and Community Duration. In: 2013 International Conference on Social Computing (SocialCom). Alexandria, VA, USA, S. 928–931

Kalakota (2011)
Kalakota, Ravi (2011): What is a "Hadoop"? Explaining Big Data to the C-Suite. Hg. v. practicalanalytics.wordpress.com
https://practicalanalytics.wordpress.com/2011/11/06/explaining-hadoop-to-management-whats-the-big-data-deal/, zuletzt geprüft am 25.06.2015

Keese (2014)
Keese, Christoph (2014): Silicon Valley. Was aus dem mächtigsten Tal der Welt auf uns zukommt. Orig.-Ausg. München: Knaus

Kessler (1963)
Kessler M. M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1), S. 10-25

KIT (2014)
KIT - Karlsruher Institut für Technologie (2014): H-Index. Unter Mitarbeit von Dr. Claudia Kramer und Gabriele Kloska. Hg. v. Karlsruher Institut für Technologie
http://www.bibliothek.kit.edu/cms/h-index.php, zuletzt aktualisiert am 03.02.2014

Kurzweil (2013)
Kurzweil, Ray (2013): The Singularity Is Near, zuletzt geprüft am 23.06.2015

LaValle et al. (2011)
LaValle, Steve; Lesser, Eric; Shockley, Rebecca; Hopkins, Michael S.; Kruschwitz, Nina (2011): Big Data, Analytics and the Path From Insights to Value 21 (52 (2)), S. 20–32

LeFevre et al. (2005)
LeFevre, K.; DeWitt, D. J.; Ramakrishnan, R. (2005): Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, S. 49-60. ACM

Lohr (2012)
Lohr, Steve (2012): The Age of Big Data. Hg. v. The New York Times http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=2&scp=1&sq=Big%20Data&st=cse, zuletzt geprüft am 22.06.2015

Manyika et al. (2011)
Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles et al. (2011): Big data: The next frontier for innovation, competition, and productivity

Marchand-Maillet (2015)
Marchand-Maillet, Stéphane (2015): Centre Universitaire d'Informatique. Université de Genève. Genf http://viper.unige.ch/doku.php, zuletzt geprüft am 23.06.2015

McAfee und Brynjolfsson (2012)
McAfee, Andrew; Brynjolfsson, Erik (2012): Big Data: The Management Revolution. In: *Harvard business review* (90), S. 60-6, 68, 128

McKinsey (2012)
McKinsey Global Institute (2012): Creating competitive advantages from big data in retail

Mell und Grance (2011)
Mell, Peter; Grance, Tim (2011): The NIST Definition of Cloud Computing. Hg. v. NIST. National Institute of Standards and Technology http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf, zuletzt geprüft am 04.05.2015

Mitchell (2006)
Mitchell, Tom M. (2006): The Discipline of Machine Learning. Hg. v. School of Computer Science. Carnegie Mellon University. Pittsburgh http://homepages.math.uic.edu/~xutu1116/Machinelearning/MachineLearning.pdf, zuletzt geprüft am 20.06.2016

MLIA (2015)
MLIA Group (2015): Machine Learning and Instrument Autonomy (MLIA) Group. Hg. v. Jet Propulsion Laboratory. California Institute of Technology. Pasadena California http://ml.jpl.nasa.gov/public/mls/, zuletzt geprüft am 23.06.2015

Mohammed et al. (2012)
Mohammed, Yassene; Mostovenko, Ekaterina; Henneman, Alex A.; Marissen, Rob J.; Deelder, André M.; Palmblad, Magnus (2012): Cloud parallel processing of tandem mass spectrometry based proteomics data. In: *J. Proteome Res.* 11 (10), S. 5101–5108

n-tv, dpa (2011)
n-tv; dpa (2011): Chronik: Meilensteine der IBM-Geschichte http://www.n-tv.de/ticker/Computer/Chronik-Meilensteine-der-IBM-Geschichte-article3543221.html, zuletzt geprüft am 23.06.2015

O'Leary (2013)
O'Leary, Daniel E. (2013): Artificial Intelligence and Big Data. In: *IEEE Intell. Syst.* 28 (2), S. 96–99

Reuters (2012)
Thomson Reuters (2012): Subject Area Terms. Hg. v. Web Of Science Help https://images.webofknowledge.com/WOKRS56B4/help/WOS/hp_subject_area_terms_easca.html, zuletzt aktualisiert am 29.03.2012, zuletzt geprüft am 20.06.2015

Richey (2014)
Richey, Erin (2014): Big data brings big changes to fraud detection. Hg. v. DELL https://powermore.dell.com/business/big-data-brings-big-changes-fraud-detection/, zuletzt geprüft am 21.06.2015

Rouse (2014)
Rouse, Margaret (2014): MapReduce. Hg. v. Searchenterprisesoftware http://www.searchenterprisesoftware.de/definition/MapReduce, zuletzt aktualisiert am März 2014

Rudell et al. (2014)
Rudell, Jacques; Bhagavatula, Venumadhav; Wesson, William (2014): Future integrated sensor radios for long-haul communication. In: *IEEE Commun. Mag.* 52 (4), S. 101–109

Rudin und Wagstaff (2014)
Rudin, Cynthia; Wagstaff, Kiri L. (2014): Machine learning for science and society. In: *Mach Learn* 95 (1), S. 1–9

Sakr et al. (2013)
Sakr, Sherif; Liu, Anna; Fayoumi, Ayman G. (2013): The Family of MapReduce and Large-Scale Data Processing Systems. In: *ACM Comput. Surv.* 46 (1), S. 1–44

Schadt et al. (2010)
Schadt, Eric E.; Linderman, Michael D.; Sorenson, Jon; Lee, Lawrence; Nolan, Garry P. (2010): Computational solutions to large-scale data management and analysis. In: *Nat. Rev. Genet.* 11 (9), S. 647–657

Schrage (2012)
Schrage, Michael (2012): Was Manager an Big Data nicht verstehen. Hg. v. Harvard Business Manager http://www.harvardbusinessmanager.de/blogs/a-862658.html, zuletzt geprüft am 31.07.2016

SHS (2015)
SHS VIVEON AG (2015): Big Data - Neue Herausforderung für Unternehmen. Unter Mitarbeit von Stefan Gilmozzi, Dr. Jörg Seelmann-Eggebert. Hg. v. SHS VIVEON AG
http://www.shs-viveon.com/de/ihr-fachbereich/bi-dwh/big-data.html, zuletzt geprüft am 21.06.2015

Simmhan (2013)
Simmhan, Yogesh; Aman, Saima; Kumbhare, Alok; Liu, Rongyang; Stevens, Sam; Zhou, Qunzhi; Prasanna, Viktor (2013): Cloud-Based Software Platform for Big Data Analytics in Smart Grids. In: *Comput. Sci. Eng.* 15 (4), S. 38–47

Small (1973)
Small H. G. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265-269

Stelzer et al. (2015)
Stelzer, Birgit; Meyer-Brötz, Fabian; Schiebel, Edgar; Brecht, L.eo (2015): Combining the scenario technique with bibliometrics for technology foresight: The case of personalized medicine. Technological Forecasting and Social Change (98), S. 137-156

TAB (2010)
TAB, Büro für Technikfolgen-Abschätzung beim deutschen Bundestag (2010): Gefährdung und Verletzbarkeit moderner Gesellschaften – am Beispiel eines großräumigen Ausfalls der Stromversorgung. Endbericht zum TA-Projekt (Nr. 141)

Tankard (2012)
Tankard, Colin (2012): Big data security. In: *Network Security.* (2012(12), S. 1353-4858

Think Big (2015)
Think Big (2015): Bringing Your Big Data Strategy to Life. Hg. v. THINK BIG, A TERADATA COMPANY
http://thinkbig.teradata.com/hadoop_nosql_services, zuletzt geprüft am 25.06.2015

Tirunillai und Tellis (2014)
Tirunillai, Seshadri; Tellis, Gerard J. (2014): Mining Marketing Meaning from Online Chatter. Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. In: *Journal of Marketing Research* 51 (4), S. 463–479

TU Berlin (2013)
TU Berlin - Technische Universität Berlin (2013): (Hg.): Prof. Dr. rer. nat. Volker Markl. Institut für Softwaretechnik und Theoretische Informatik Fachgebiet Datenbanksysteme und Informationsmanagement
http://www.dima.tu-berlin.de/menue/staff/volker_markl/parameter/de/, zuletzt geprüft am 20.06.2015

Van Der Aalst et al. (2012)
Van Der Aalst, Wil M. P. (2013): Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining. In: Van Der Aalst, Wil, John Mylopoulos, Michael Rosemann, Michael J. Shaw, Clemens Szyperski, Minseok Song et al. (Hg.): Asia Pacific Business Process Management, Bd. 159. Cham: Springer International Publishing (Lecture Notes in Business Information Processing), S. 1–22

Van Raan (2003)
Van Raan, Anthony FJ (2003): The use of bibliometric analy-sis in research performance assessment and monitoring of interdisciplinary scientific de-velopments. Hg. v. Citeseer. Leiden University http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.629&rep=rep1&type=pdf, zuletzt geprüft am 21.06.2015

Veltman et al. (2013)
Veltman, Joris A.; Cuppen, Edwin; Vrijenhoek, Terry (2013): Challenges for implementing next-generation sequencing-based genome diagnostics: it's also the people, not just the machines. In: *Personalized Medicine* 10 (5), S. 473–484

Vorest (2013)
VOREST AG (2013): Bausteine ihres Erfolgs. KVP (kontinuierlicher Verbesserungsprozess) - Fakten, Wissen und Infos
http://kvp.me/kvp_kaizen/, zuletzt geprüft am 21.06.2015

Voß (2015)
Voß, Jan (2015): Vorratsdatenspeicherung: Fakten-Check und Chronologie. Hg. v. politik-digital.de http://politik-digital.de/vorratsdatenspeicherung-fakten-check-und-chronologie, zuletzt geprüft am 21.06.2015

Wainer et al. (1974)
Wainer, Howard; Gruvaeus, Gunnar; Blair, Michael (1974): TREBIG: A 360/75 FORTRAN program for three-mode factor analysts designed for big data sets. In: *Behav. Res. Meth. & Instru.* 6 (1), S. 53–54

Wall Street Journal (2015)
Wall Street Journal (2015): Ranking der wertvollsten digitalen Start-ups | 2015. Hg. v. Statista - Das Statistik Portal
http://de.statista.com/statistik/daten/studie/157889/umfrage/ranking-der-wertvollsten-digitalen-start-ups/, zuletzt geprüft am 17.06.2015

Waller und Fawcett (2013)
Waller, Matthew A.; Fawcett, Stanley E. (2013): Data Science, Predictive Analytics, and Big Data. A Revolution That Will Transform Supply Chain Design and Management. In: *J Bus Logist* 34 (2), S. 77–84

Walsh (2005)
Walsh, P.R. (2005): Dealing with the uncertainties of environmental change by adding scenario planning to the strategy reformulation equation. Manag. Decis. 43 (1), S. 113–122

Wikibon (2013)
Wikibon (2013): Big Data - Umsatz führender Anbieter weltweit 2013 | Statistik. Hg. v. Statista - Das Statistik Portal
http://de.statista.com/statistik/daten/studie/302303/umfrage/umsatz-der-fuehrenden-anbieter-von-big-data-loesungen-weltweit/, zuletzt geprüft am 10.06.2015

Yesudas et al. (2014)
Yesudas, M.; Menon, G.; Ramamurthy, V. (2014): Intelligent operational dashboards for smarter commerce using big data. In: *IBM J. Res. & Dev.* 58 (5/6), S. 13:1-13:10

Zhang et al. (2013)
Zhang, Xuyun; Liu, Chang; Nepal, Surya; Yang, Chi; Dou, Wanchun; Chen, Jinjun (2013): SaC-FRAPP: a scalable and cost-effective framework for privacy preservation over big data on cloud. In: *Concurrency Computat.: Pract. Exper.* 25 (18), S. 2561–2576

Zhao et al. (2014)
Zhao, Jiaqi; Wang, Lizhe; Tao, Jie; Chen, Jinjun; Sun, Weiye; Ranjan, Rajiv et al. (2014): A security framework in G-Hadoop for big data computing across distributed Cloud data centres. In: *Journal of Computer and System Sciences* 80 (5), S. 994–1007