Jörg Eberspächer
Otto Wohlmuth

Herausgeber

Big Data wird neues Wissen

MÜNCHNER KREIS
Übernationale Vereinigung für Kommunikationsforschung
Supranational Association for Communications Reseach

Jörg Eberspächer
Otto Wohlmuth

Herausgeber

Big Data wird neues Wissen

MÜNCHNER KREIS
Übernationale Vereinigung für Kommunikationsforschung
Supranational Association for Communications Reseach

Impressum

Herausgeber

Prof. Dr.-Ing. Jörg Eberspächer
Technische Universität München
Lehrstuhl für Kommunikationsnetze
Arcisstr. 21
80333 München
joerg.eberspaecher@tum.de

Dr. Otto Wohlmuth
IBM Deutschland R&D GmbH
Manager IBM Systems & Technology Group
Schönaicher Str. 220
71032 Böblingen
wohlmuth@de.ibm.com

Reihenherausgeber:
Münchner Kreis – Übernationale Vereinigung für Kommunikationsforschung e.V.
Tal 16
80331 München
www.muenchner-kreis.de
office@muenchner-kreis.de

Redaktion:
Dipl.-Phys. Volker Gehrling
Münchner Kreis– Übernationale Vereinigung für Kommunikationsforschung e.V.
Tal 16
80331 München
v.gehrling@muenchner-kreis.de

Druck:
Knecht-Druck, München

ISBN 978-3-9813733-7-0

Vorwort

Nach aktuellen Schätzungen verzehnfachen sich die im Internet anfallenden Datenmengen alle fünf Jahre. Das Informationsuniversum („Big Data“) dehnt sich in alle Bereiche aus wie z.B. Handels-, Finanz- und Energiesektor, Telekommunikation, Gesundheitswesen, Automotive, Verkehrsüberwachung und Soziale Netzwerke. Die Datenflut stellt eine große Herausforderung dar, sie bietet aber auch große Chancen, daraus „neues Wissen“ zu machen. Dabei kommt es entscheidend auf die Fähigkeit an, die Datenvielfalt – unter Berücksichtigung der Datenrechte und gesetzlichen Vorgaben – effizient, vielfach sogar in Echtzeit, zu verarbeiten und auszuwerten („Analytics“), um mit dem daraus gewonnenen Wissen dem Anwender Nutzen zu stiften, Wettbewerbsvorteile zu verschaffen und neue Geschäftsfelder zu erschließen. Existierende Technologien und Konzepte stoßen aufgrund ihrer Komplexität und Verarbeitungsgeschwindigkeit an Ihre Grenzen.

Neue adaptive Prozesse, Lösungsansätze und Strategien gewinnen an Bedeutung, die über intelligente Verfahren und lernende Systeme die automatische und schnelle Auswertung von „Big Data“ erlauben. Dieser Einsatz neuer Lösungsansätze wird nicht nur zu einer Optimierung der Technologien, Prozesse und Betriebsmodelle innerhalb großer Unternehmen führen, sondern insbesondere bei KMU’s eine engere Zusammenarbeit zwischen den Unternehmen erfordern, da einzelne Unternehmen aufgrund der Komplexität diese Herausforderungen allein nicht meistern können. Hier müssen neue Wege beschritten werden, ohne die viele Investitionen in die digitale Infrastruktur überhaupt nicht nutzbar sind und der erwünschte technische und ökonomische Fortschritt ausbleibt. Ein wichtiger Punkt ist hierbei auch der Austausch und die Nutzung von Daten und die damit zusammenhängende Frage der Daten-Governance von öffentlichen, privaten sowie vertraulichen Daten.

Der Münchner Kreis hat in seiner Fachkonferenz das Thema mit seinen technischen, ökonomischen und gesellschaftlichen Aspekten zusammen mit Fachleuten aus Industrie und Wissenschaft diskutiert und die Chancen wie auch die Grenzen ausgelotet, die dabei zu erkennen sind. Das Thema wurde von den Experten aus unterschiedlichen Perspektiven beleuchtet, es wurden Lösungskonzepte vorgestellt und diskutiert insbesondere wobei der Nutzen für die Gesellschaft sowie die Veränderungen der Unternehmenslandschaft und die Chancen für KMU’s, kleine und mittelständische Unternehmen, herausgestellt wurden.

Der vorliegende Tagungsband enthält die Vorträge sowie die überarbeiteten Mitschriften der Diskussionen. Allen Referenten und Diskutanten sowie allen, die zum Gelingen der Konferenz und zur Erstellung dieses Buches beigetragen haben, gilt unser herzlicher Dank!

Jörg Eberspächer Otto Wohlmuth

Inhalt

1 Begrüßung und Einführung

Prof. Arnold Picot, Ludwig-Maximilians-Universität München

Heute behandeln wir ein sehr wichtiges, spannendes und unser aller Zukunft prägendes Thema. Zu Beginn möchte ich einige wenige Bemerkungen zu der Thematik machen, um in sehr allgemeiner Weise einzuführen.

Es geht um Daten. Daten sind Zeichen, die auf irgendetwas hinweisen können. Dieses Hinweisen geschieht in der Regel durch eine Zuordnung einer Bedeutung zu diesen Zeichen. Häufig nennt man dieses auch Semantik. Diese Daten, die eine Verbindung herstellen zu irgendetwas, was uns bekannt zu sein scheint, können sich auf alle Arten von Objekten beziehen, nicht lebende oder auch lebende Objekte, also auch auf Menschen, sowie auf alle Arten von Konstrukten, mit denen wir uns in der Welt zu orientieren versuchen, also auf mehr oder weniger anerkannte gedankliche Konstrukte oder Interpretationsmuster, mit denen wir uns umgeben.

Daten können sich auf Zustände zu bestimmten Zeitpunkten beziehen (sogenannte Bestandsdaten). Sie können sich aber auch auf Veränderungen beziehen, sogenannte Bewegungsdaten, die im Zeitablauf stattfinden. Beides gehört meistens zusammen, d.h. dynamische und zeitpunktbezogene Betrachtungen werden verknüpft.

Daten dienen der Dokumentation, d.h. sie sollen uns helfen zu zeigen, was ist oder was war. Insofern dienen Sie der Beschreibung und Erforschung von Gegenwart und Vergangenheit.

Daten werden häufig für bestimmte Zwecke, für bestimmte Problemlösungen eingesetzt. Aus ursprünglich mehr oder weniger zweckneutralen Daten werden dann zweckbezogene Verwendungen, und diese zweckbezogene Verwendung von Daten nennt man auch sehr häufig Information. Wenn man Daten zweckbezogen bündelt und verarbeitet, werden diese Informationen oftmals benutzt, um Entscheidungen zu fundieren und Prognosen zu stellen. Solche Daten gestützten Informationen, Analysen und Entscheidungen ermöglichen den Aufbau von Erfahrung, also eine Erfahrungsbasis, die uns hilft, die Welt möglicherweise besser zu verstehen. Deswegen spricht man dann auch von verbessertem Wissen, das durch diese Vernetzung von Informationen entstehen kann.

Wenn Daten untereinander verwandt sind, also etwas Gemeinsames haben – diese Verwandtschaft kann sich auf formale oder auf inhaltliche Aspekte beziehen –, werden sie häufig in Dateien oder auch in Datenbanken zusammengefasst und dort organisiert und verwaltet. Typischerweise werden Daten mit Hilfe von Programmen und Algorithmen inhaltlich analysiert und verwaltet. Hier kommen dann natürlich die Informatik im engeren Sinne und die Mathematik sehr massiv zum Einsatz.

Die Analyse von Daten geschieht bislang typischerweise offline, also losgelöst vom aktuellen prozesshaften Geschehen, aber zunehmend auch online, d.h. integriert in aktuelle Prozesse und damit in Echtzeit.

All das ist nicht neu. Menschliche Orientierung und menschliches Handeln gründen sich von jeher zu einem erheblichen Teil auf Daten. Die Menschen haben immer versucht, sich irgendwie ein Bild zu machen von dem, was war oder was ist und sich bemüht daraus zusammen mit anderen Faktoren Folgerungen zu ziehen. Allerdings war es in der

Vergangenheit so, dass die Gewinnung interessierender Daten oftmals nur mit sehr großem, ja prohibitiv hohem Aufwand machbar war, wenn sie überhaupt möglich erschien. Das Gleiche gilt für die Analyse gesammelter Daten, die oftmals schwer oder nur mit extrem großem Aufwand möglich war.

Die Digitalisierung nun, die wir alle kennen und mit der wir uns intensiv beschäftigen, hat im Verbund mit den erheblichen technischen Fortschritten in der Informationsverarbeitung und in der Informationsübertragung die Verfügbarkeit von Daten für Wirtschaft und Gesellschaft in einem bislang unvorstellbaren Ausmaß gesteigert. Der Grund dafür, dass nun diese Daten so überaus zahlreich und vielfältig verfügbar und dann auch analysierbar sind, liegt nicht nur allein in der Technik an sich, die uns das eröffnet, sondern er liegt vor allen Dingen in dem enormen Absinken der Kosten, die zur Gewinnung und zur Analyse der Daten aufzuwenden sind. Wir haben es mit einem exponentiellen Rückgang der Kosten zur Gewinnung, Verarbeitung, Speicherung und Übertragung von Daten zu tun und damit wird die Schwelle der Verfügbarkeit dieser Daten immer niedriger gesetzt.

Auf diese Art und Weise stehen die elektronischen Repräsentationen der Welt - d.h. der Zustände, der Prozesse und der Veränderungen dieser Welt, sowohl der physischen Welt wie auch der kulturellen und gedanklichen Konstrukte, mit der wir diese Welt erfahrbar machen - in fast vollständiger bzw. immer vollständigerer Form zur Verfügung.

Die Fruchtbarmachung dieser Potentiale Gesellschaft und Wirtschaft lässt sich m. E. mit dem Schlagwort Big Data belegen, und damit sind wir bei der Überschrift unserer Konferenz: Es geht um die Fruchtbarmachung dieser enorm gesteigerten Potenziale an Daten und ihrer Analysemöglichkeiten.

Big Data stellt natürlich neue Anforderungen etwa an Datenbanksysteme und Algorithmen sowie an die Vernetzung diverser Teilsysteme und Technologien.

Big Data eröffnet neue Optionen der Online- und Offlinegestaltung von Verwaltungs- und Geschäftsprozessen. Man spricht auch von Business Analytics und Business Intelligence. Das alles zunehmend online, also in real time, aber auch natürlich nach wie vor und in vielfältigen Varianten offline.

Big Data gibt Raum für neue Services und neues Unternehmertum, das dazu dient, diese Daten und ihre Analyse und Verwaltung bereitzustellen. Gerade die komplexen und extrem voluminösen Datenmengen können auf diese Art und Weise von Spezialisten in der gewünschten Form verfügbar gemacht werden.

Big Data wirft Fragen auf hinsichtlich der Verfügungsrechte über diese Daten auf, z.B.: Unter welchen Voraussetzungen sind Daten Privatsache? Wann dürfen sie kommerziell gesammelt und verwertet werden? Unter welchen Bedingungen muss Dritten Zugang zu Daten oder auch zu ihren Verarbeitungsstufen gegeben werden? Welche Daten und deren Analyse haben hoheitlichen Charakter und sind insofern ein öffentliches Gut? Welche Daten bedürfen besonderer Sicherheitsvorkehrungen und wie sind diese zu bewerkstelligen?

Big Data ist letztlich eine Art Metapher, die für die fundamentalen Veränderungen unserer gesellschaftlichen und wirtschaftlichen Daseinsbedingungen unter dem Einfluss der Digitalisierung steht. Daraus ergeben sich erhebliche Chancen zur Steigerung des Wissens, zur Verbesserung des Wissenszugangs und zur Erhöhung der Qualität von Geschäfts- und Versorgungsprozessen aller Art. Aber es ergeben sich natürlich auch neue Fragen und Herausforderungen für Forschung, Wirtschaftspraxis und Politik im Allgemeinen.

Meine Damen und Herren, vor diesem Hintergrund möchte der Münchner Kreis am heutigen Tag die Phänomene, die verschiedenen Aspekte und Dimensionen des Themas aus technischer, ökonomischer, gesellschaftlicher Sicht beleuchten und mit Ihnen diskutieren, vor allen Dingen auch die Chancen ausloten, die zu erkennen sind, und die Punkte, die wir dabei wachsam im Auge haben müssen, um die Chancen nachhaltig und sinnvoll zu heben. Dabei geht es um Lösungskonzepte, aber auch um Veränderungen der Unternehmenslandschaft und um Chancen, auch gerade für kleine und mittlere Unternehmen und für die Gesellschaft als Ganzes.

Wir wollen nicht nur, aber auch, technische und theoretische Aspekte beleuchten, sondern insbesondere konkrete Anwendungen und Beispiele zeigen. Zu all dem haben wir, meine ich, ein sehr interessantes Programm zusammengestellt, für dessen Vorbereitung ich dem entsprechenden Programmausschuss unter Leitung von Herrn Wohlmuth von IBM und Herrn Kollegen Eberspächer von der TU München, aber auch anderen, die sehr engagiert im Programmausschuss mitgeholfen haben, ganz herzlich danken.

Meine Damen und Herren, ich hoffe, damit den Boden etwas vorbereitet zu haben für unser Programm und möchte nun in unseren Eröffnungsblock einsteigen. Wir haben es mit drei Vorträgen zu tun. Der eine, der stärker die Geschäftsmodelle, die Anwendungen, die strategischen Perspektiven aus wirtschaftlicher Sicht aufzeigt. Der andere, der die technischen Herausforderungen und Entwicklungen in diesem Feld uns aus der Sicht eines Wissenschaftlers vorstellt. Und der dritte, der uns die rechtliche Verknüpfung mit der Gesellschaft und ihrem Regelwerk nahebringt. Ich glaube, das sind sehr wichtige Keynotes, um damit den Rahmen abzustecken, den wir für die folgenden, stärker anwendungsbezogenen Beispiele, Detailanalysen und Diskussionen benötigen

2 Big Data & Analytics – Herausforderungen und Geschäftsmodelle in einer digitalen Welt

Christian Klezl, IBM, Armonk NY, USA

Ich freue mich sehr als Europäer in New York Gelegenheit zu haben, nach München zu kommen, und auf Deutsch über ein Zukunftsthema zu sprechen. Es ist bei IBM Referenten in den letzten Monaten sehr beliebt geworden, den Zeitrahmen über die letzten 100 Jahre auszurollen. Sie alle wissen wahrscheinlich, dass wir letztes Jahr unser 100jähriges Bestandsjubiläum feiern konnten. Ich will aber gar nicht über IBM sprechen. Ich möchte eingangs zum Thema Big Data eine Analogie aus dem Bereich der Rohstoffe, aus dem Bereich der Erdölindustrie strapazieren.

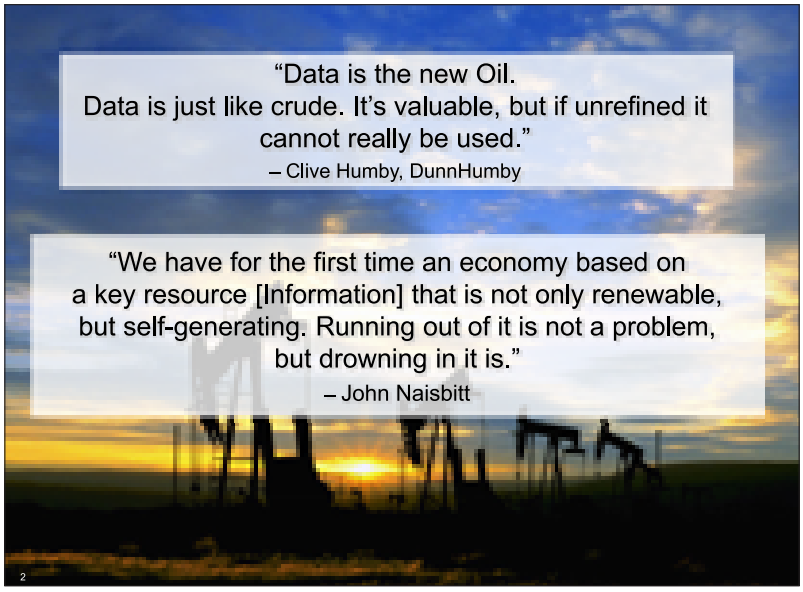

Bild 1

Erdöl in seiner bekannten Form ist über viele Millionen Jahre entstanden und trotzdem hat es die Menschheit erst in den letzten hundert Jahren verstanden, die richtigen Technologien, die richtigen Prozesse, die richtigen Werkzeuge zu entwickeln, um aus einem eigentlich relativ unbrauchbaren Rohmaterial, nämlich Erdöl-Rohöl, einen sehr wesentlichen Grundstoff für Gegenstände des täglichen Lebens zu produzieren. Denken wir an die Kosmetikindustrie, an die Kunststoffindustrie und letztendlich an Rohöl als einen fossilen Brennstoff, der so viele Annehmlichkeiten des täglichen Lebens wie Fahrzeuge bis Mobilität, Heizung usw. antreibt. Ich darf mich den Zitaten hier auf diesen Eingangschart (Bild 1) durchaus anschließen. Ich glaube, die Analogie, die wir hier von Rohöl hin zu einem so wertvollen Grundstoff in unserer heutigen Gesellschaft ziehen können, ist eine sehr große, eine sich fast aufdrängende Parallele zu dem, was wir heute rund um Daten, um das Wachstum und die Explosion der Daten, erleben und wie wir, wie Herr Prof. Picot schon sehr gut darstellt hat, diese Daten letztendlich zu relevanter Information transportieren können.

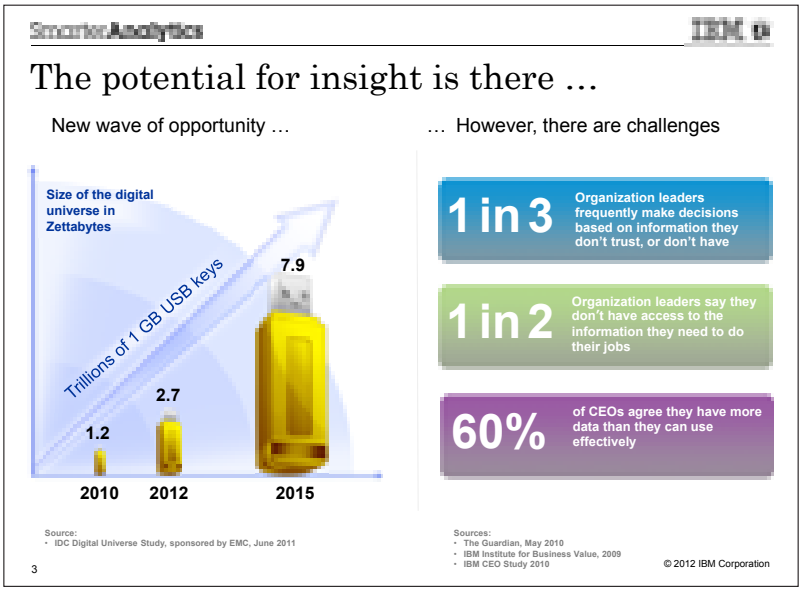

Bild 2

Noch interessanter ist aber vielleicht das zweite Zitat auf diesem Chart (Bild 2), denn auch hier lässt sich eine sehr gute Parallele darstellen. Wenn wir über Daten sprechen, dann sprechen wir im Gegensatz zu Rohöl über einen erneuerbaren Rohstoff. Noch viel spannender als erneuerbar ist, dass es ein sich selbst erneuernder Rohstoff ist. Wie John Naisbitt schon sehr schön dargestellt hat: die Gefahr, die wir beim Rohstoff Daten sehen, ist weniger die Gefahr, dass uns der Rohstoff ausgeht – bekanntlich produzieren wir derzeit alle zwei Jahre das gleiche Datenvolumen, das die Menschheit bisher produziert hat –, sondern die Gefahr ist, dass wir in dem Volumen dieser Daten ertrinken.

Um das Gleiche zu bewerkstelligen, was die Menschheit über die letzten 100 Jahren bei der Entwicklung von Rohöl hin zu einem so wertvollen Grundstoff für unsere Wirtschaft, für unser tägliches Leben, für unsere Gesellschaft geschaffen hat, müssen den gleichen Evolutionsgang in einer viel intensiveren dichteren Zeitspanne verfolgen, bis es letztendlich gelungen ist im Rohölbereich, um aus Daten Information und Mehrwert zu generieren und nicht in dem sich ständig erneuernden Prozess der Datenflut zu ertrinken, mit der wir heute auseinander gesetzt sind.

Ich habe das Potenzial angesprochen, das Wachstum angesprochen. Man kann Daten heute aus verschiedenen Gesichtspunkten betrachten. Es gibt natürlich unzählige Marktforschungsinstitute und Managementconsultingfirmen, die sich mit dem Wachstum dieser Datenmenge auseinandersetzen. IDC ist vielleicht in der stärksten Begrifflichkeit hier prägend geworden mit dieser Verdoppelung des Datenvolumens im Schnitt alle zwei Jahre. Alles, was an Daten in der Menschheit bis 2010 produziert wird, hat sich bis zum Jahr 2012 verdoppelt, wird sich in den nächsten zwei Jahren noch einmal verdoppeln und steigt fast schon exponentiell an über die nächsten Jahre – ein ungeheures Datenwachstum. Es sind Milliarden, eine Billion von USB Sticks im Umlauf, die Daten generieren. Wir generieren täglich über Sensoren, über das Internet der Dinge, wie wir die Begrifflichkeit so schön verwenden, neue Daten, die in irgendeiner Form strukturiert oder unstrukturiert heute vorhanden sind.

Das heißt, das Wachstum dieser Daten, die die Billion an Endgeräten, die heute bereits weltweit wie Fahrzeuge, Haushaltsgeräte, natürlich auch die Informationstechnologie, Büroautomation, Geräte des täglichen Lebens, unsere Mobiltelefone, mit dem Internet kommunizieren – all das integriert, aber auch generiert tagtäglich Daten.

Das Potenzial, das Sie auf der linken Seite dieser Folie dargestellt haben, korrespondiert auf der anderen Seite mit großen Bedenken. Sie haben hier drei verschiedene Umfragequellen, sehr plakativ zusammengefasst. Zum einen die große Sorge von Entscheidungsträgern in der Wirtschaft. Einer von drei Entscheidungsträgern sagt: ich vertraue der Qualität der Daten, mit denen ich tagtäglich zu tun habe, nicht oder nicht ausreichend. Vertrauen in Daten ist aber eine wesentliche Grundvoraussetzung, um qualitative Entscheidungen zu fällen anhand von Daten, die mir zur Verfügung gestellt werden. Noch schlimmer ist es, dass jeder zweite CEO sagt, dass er die Daten hat, die er zu dem Zeitpunkt, wo er eine Entscheidung fällen möchte, nicht hat. Der schlimmste Parameter ist hier vielleicht der ganz rechts unten, wonach 60 % der CEOs sagen, dass sie im Wesentlichen wahrscheinlich mehr Daten zur Verfügung haben, als sie für ihre Entscheidungen überhaupt brauchen. Sie kommen genau in die Theorie von Naisbitt hinein, dass sie in ihren Daten ertrinken und eigentlich den Wald vor lauter Bäumen in unserem täglichen Entscheidungsfindungsprozess nicht mehr sehen.

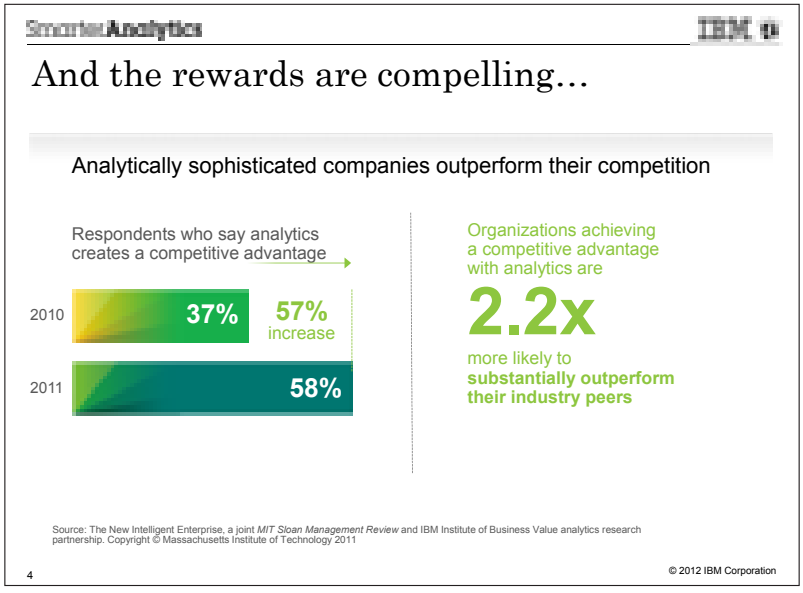

Bild 3

Wir haben also auf der einen Seite ein großes Potenzial, diese Explosion an Rohstoff, an Datenmaterial, und auf der anderen Seite ein sehr hohes Maß an Unzufriedenheit, ein sehr hohes Maß an suboptimalem Zugang, wie wir als Gesellschaft, als Vertreter von Akademien und der Wirtschaft mit diesen Daten im Wesentlichen umgehen. Dabei sollte der Vorteil einer guten produktiven, konstruktiven Nutzung dieser Daten eigentlich auf der Hand liegen. Wir haben hier als IBM und den vielen anderen Instituten, gemeinsam mit dem IMT, schon über mehrere Jahre eine Untersuchung gestartet. Sie sehen hier sehr dramatisch, dass der Glaube und die Zuversicht bei Entscheidungsträgern, dass die richtige Anwendung von Daten im Unternehmen effektive Wettbewerbsvorteile erschließt, sehr stark gestiegen ist (Bild 3). Um 60 % ist von 2010 auf 2011 die Zunahme an Befürworter gestiegen, dass Wettbewerbsvorteile erschlossen werden, wenn ich meine Daten im Unternehmen richtig nutze. Da geht es nicht

um ein Mehr an Daten. Da geht es um ein besseres effizienteres Nutzen der Daten, die ich heute im Unternehmen bereits zur Verfügung habe.

Oder auch 2.2x, sprich eine Verdoppelung des Wettbewerbsvorteils, der durch diese Daten entsteht. In einer anderen Studie geht man davon aus, dass 5 bis 6 % Produktivitätssteigerung durch den effizienten Einsatz von Daten möglich ist. Sie kennen viele Branchen, so wie ich, in denen 5 bis 6 % Produktivitätssteigerung letztendlich über Verlieren oder Gewinnen am Markt oder auch Überleben am Markt, eine sehr starke Aussage in einer sehr margendünnen Industrie treffen. Hier gibt es eigentlich keine Zweifel, dass wir als Gesellschaft ein Mandat haben für die Wettbewerbsfähigkeit unserer Wirtschaft, Gesellschaft entsprechend der Nutzung, um diese Vorteile der Nutzung für uns als Gesellschaft und Individuen wirklich zu erschließen.

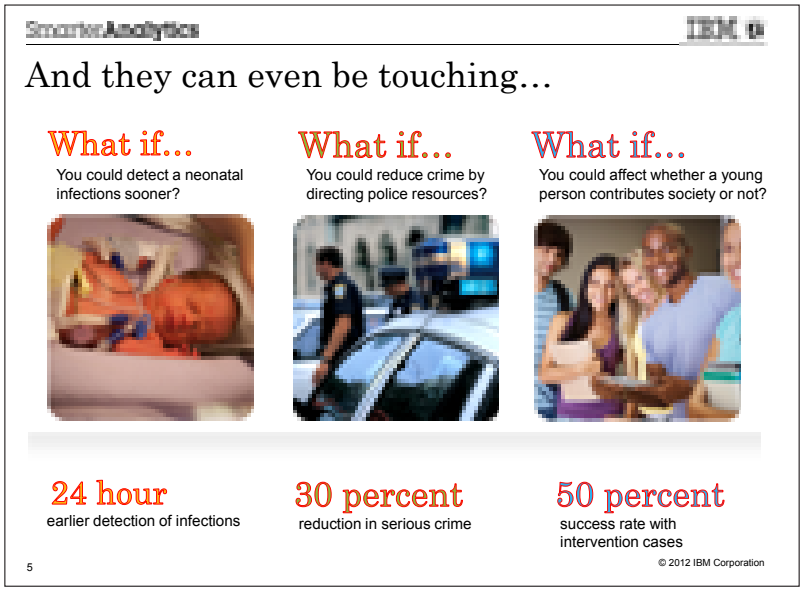

Bild 4

Lassen Sie mich drei Beispiele nennen, die genau in diesen gesellschaftlichen Nutzen hineinspielen (Bild 4). Ich möchte gleich am Anfang postulieren, dass Big Data kein Technologiethema ist. Ich glaube, dass Big Data aus dem Stallgeruch der Festplattenindustrie hinaus muss. Wenn wir als IBM über Big Data sprechen, dann sprechen wir nicht über Wunschträume und Werbeslogans der Harddisk Industrie vor zehn Jahren. Wenn wir über Big Data sprechen, dann sprechen wir über die effiziente Nutzung vorhandener Ressourcen und über den Mehrwert, der über die Verknüpfung und die Analyse dieser Daten entstehen.
Ich zeige Ihnen drei Beispiele aus sehr unterschiedlichen Industrien, aus sehr unterschiedlichen Ländern weltweit, wo solche Big Data Lösungen eingesetzt werden.

Vielleicht eines der emotional bewegendsten Beispiele ist das hier auf der linken Seite, das Thema der Frühgeburten weltweit. Wenn man den Zahlen der WHO glaubt, passieren jedes Jahr 10 % aller Geburten im so genannten Stadium einer Frühgeburt. 15 Millionen Frühgeburten weltweit jedes Jahr. Eine Million davon endet leider tragisch tödlich zumeist innerhalb von 30 Tagen ab dem Zeitpunkt der Geburt. Zumeist ist es so, dass im Fall einer Frühgeburt der Säugling in einen Brutkasten in eine intensivmedizinische Betreuung kommt, wo in jeder Sekunde Hunderte Messdaten abgenommen werden. Wir haben hier ein Projekt

gemeinsam mit Universität Ontario und einem kanadischen Krankenhausträger gemacht und festgestellt, dass 90 MB an Daten pro Patient und pro Tag generiert werden. Tragisch daran ist, dass das meiste davon für den Krankenakt ist. Die wenigsten dieser Daten werden miteinander in Konstellation oder Korrelation gebracht. Die wenigsten dieser Daten werden dazu genutzt, vergleichende Parameter abzutesten und zu versuchen, Krankheitsbilder möglichst frühzeitig zu erkennen.

In dem Projekt, das wir gemeinsam mit diesen beiden Trägern gemacht haben, ging es ganz gezielt darum, diesen Säuglingen nicht zusätzliche Sensoren anzutun, nicht zusätzliche Daten zu messen, nicht zusätzlich Ärzte dazu zu zwingen, IT zu programmieren oder IT-Systeme zu füttern sondern Daten, die es bereits gibt, zu verwenden und in Beziehung zu setzen, um sehr frühzeitig mögliche Probleme für den Säugling im Brutkasten bereits zu erkennen. Bis zu 24 Stunden früher als in der traditionellen Behandlungsweise ließen sich dadurch mit gezielten Therapiemöglichkeiten solche potenziellen Krisensituationen erkennen und damit darauf reagieren. Sie können sich vorstellen, dass 24 Stunden in intensiv medizinischer Terminologie ein unglaublicher Wettbewerbsvorteil ist, um diesen Begriff hier etwas zu missbrauchen, der im Endeffekt über Tod und Leben bei einem Säugling entscheiden kann. Das ist ein ganz maßgebliches gesellschaftliches Mehrwertphänomen, das ich hier herausstreichen möchte.

In einer ganz anderen Branche, auch am nordamerikanischen Kontinent, handelt mein zweites Beispiel. Die Stadt Memphis hat eigentlich keine abweichenden Probleme von den meisten Städten in der industrialisierten Welt heute. Einen Ausgabenstopp im öffentlichen Bereich, gleichzeitig eine steigernde Kriminalität. In der Zahl der Polizeikräfte, im Kapital, in den Assetts, die vorhanden sind, natürlich nur sehr limitierte Ressourcen. Für mich und für viele war neu, dass Kriminalität nicht nach Zufall passiert. Kriminalität passiert nach statistisch erhebbaren Mustern, nach Paradigmen, die man durchaus messen und prognostizieren kann. In einem Projekt mit der Stadtverwaltung von Memphis haben wir versucht, gemeinsam zu realisieren, die bereits existierenden Daten – denken wir an Daten aus der Verkehrsüberwachung, Strafmandate, Meldedaten, Daten aus der leichten Kriminalität – in Bezug zueinander zu bringen, mit Analysewerkzeugen zu veredeln, zu raffinieren und daraus Problemzonen in der Stadt zu definieren. Mit dem gleichen Einsatz der gleichen Ressourcen an diesen Problemzonen ohne deswegen mehr Polizisten auf die Straße zu bringen oder anstellen zu müssen, aber sehr wohl mehr Polizisten weg von der IT hin zum Feldeinsatz auf der Straße zu bringen, konnte die kriminalitätsrate in der Stadt Memphis ganz substanziell gesenkt werden. Wir sprechen dabei nicht von 2 oder 3 % sondern von zweistelligen Reduktionsraten in der Kriminalität, indem man Polizisten dorthin bringt, wo sie nach aller Analyse vermutlich am meisten gebraucht werden.

Ein drittes Thema umfasst einen ganz anderen Bereich in Europa. In England gibt es eine Organisation im Non-Profit-Bereich, die sich sehr intensiv in den industrialisierten Arbeiterstädten mit der Entwicklung von Jugendlichen auseinandersetzt, und die traditionell versucht zu reagieren, wenn Jugendliche, wie die Engländer es so schön blumig formulieren, keinen positiven Beitrag zur Gesellschaft mehr leisten, sprich: wenn sie vom klassischen Bildungsweg abweichen, wenn sie in die Gefahr kommen, von Alkohol, Drogen oder anderem Missbrauch, von einem Hoffnungspotenzial der Gesellschaft zu einem potenziellen Problemfall der Gesellschaft zu werden. Über Datenanalytics, über Big Data und einer Lösung, die wir hier gemeinsam mit diesem Non-Profit-Träger erarbeitet haben, konnten wir ein System bauen, das es schafft, solche Problemfälle pro aktiv zu finden, zu identifizieren und damit dieser Organisation zu ermöglichen, pro aktiv mit diesen Jugendlichen zu arbeiten, bevor sie letztendlich von dieser Schiene fallen und in die Gefahr kommen, dass sie hier von diesem Weg abweichen und damit die Wahrscheinlichkeit, dass man hier noch

positiv konstruktiv eingreifen kann, wesentlich höher ist als wenn bereits der Weg verlassen ist und die ersten Probleme aufgetaucht sind.

Das waren drei ganz unterschiedliche Beispiele. Die Lösung, die dahinter steckt, ist jeweils sehr unterschiedlich konstruiert, sehr unterschiedlich gebaut. Aber ich glaube, dass es drei sehr prägende Beispiele sind, die zeigen, wie sehr es in Big Data nicht um Technologie zum Selbstzweck geht. Technologie ist ein ganz wesentlicher Aspekt, der ermöglichen soll, dass solche Lösungen möglich werden, die letztendlich aber ihren Mehrwert im individuellen und gesellschaftlichen Nutzen von Lösungen wie den drei auf dieser Grafik haben.

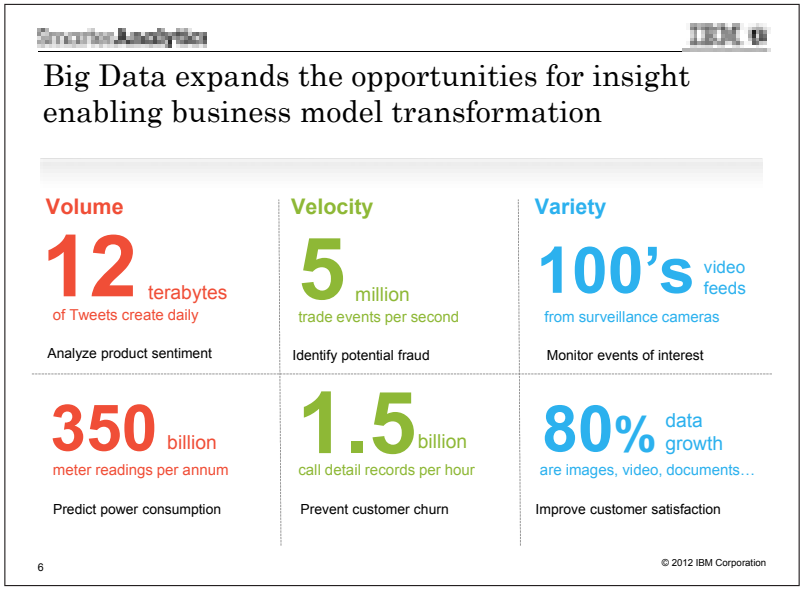

Bild 5

Ich möchte Ihnen gern das Thema Big Data von einer architektonischen Perspektive beschreiben. Ich bin selber kein Techniker sondern komme aus dem Marketing- und Vertriebshintergrund von 15 Jahren IBM in den verschiedensten Bereichen. Wenn wir über Data sprechen – und ich habe vorhin schon etwas provokant formuliert, dass es vielleicht ein Wunschtraumslogan der Festplattenindustrie vor vielen Jahren war -, .sprechen wir nicht nur über Volume (Bild 5). Wir sprechen heute als IBM eigentlich über drei verschiedene Dimensionen von einer Architekturperspektive. Wir sprechen über das Volumen der Daten, über die Geschwindigkeit, die zeitnahe Bereitstellung von Informationsgrundlage und über die Datenvielfalt, mit der wir heute auseinandergesetzt sind. Lassen Sie mich wieder drei Beispiele herausgreifen, die repräsentativ für viele sind, mit denen wir als Menschen es hier tagtäglich zu tun haben.

Zwölf Terabytes an Gezwitscher, an Twitts, täglich im Internet. Wir werden uns alle einig sein, dass davon vieles wertlose insignifikante Information ist. Richtig genutzt, richtig gefiltert, richtig raffiniert, ist diese Information unter Umständen von kritischer Bedeutung für Unternehmen. Denken Sie an ein Konsumgüterunternehmen, das es schafft über den Einsatz von gezielter Analyse herauszufinden, wie über die Neueinführung eines Produktes am Markt getwittert wird. Es ist eigentlich das in die Gegenwart übernommene Abhören des Gesprächs auf der Agora, des Marktplatzes. Was denkt die Öffentlichkeit über mein neues Fahrzeug, mein neues Waschmittel, mein neues Radio, über meinen neuen Song, den ich kreiert habe?

Eine ganz wesentliche Nutzung, die heute bereits tagtäglich im Volumen passiert. Sie müssen 12 Terabytes an Daten, die jeden Tag generiert werden, abfragen. Sie müssen den ganzen Müll an Daten wegfiltern, den Sie nicht wollen und ganz gezielt für Ihr Unternehmen, für Ihr Thema die wertvolle Information herausdestillieren.

350 Milliarden Ablesewerte, die laufend aus den verschiedenen Smart Meters generiert werden, die in den verschiedenen Verbundsystemen und Stromnetzen weltweit installiert sind. Das ist eine ungeheure Menge an Daten, die aber für die Reduktion von Wartungsaufwand, für die Vermeidung von Netzausfällen von Energiebetreibern entscheidungskritisch sein können. Auch hier geht es wieder um das Filtern der richtigen relevanten Informationen aus dem Gesamtvolumen an gewaltigen Datenbeständen.

Ein anderes Beispiel ist die Geschwindigkeit. Hier geht es nicht so sehr um das Volumen, das abgearbeitet werden muss, wobei auch das nicht zu vernachlässigen ist, sondern es geht um die zeitnahe Bereitstellung von Entscheidungs- oder von Lösungsgrundlage. Denken wir an die Telekomindustrie! Das weitgehend bekannte größte Problem der Telekomindustrie ist die Abwanderung von Kunden, das mangelnde Loyalitätsverhalten von Kunden. Die Telekomindustrie hat herausgefunden, dass das größte Problem, das Kunden zur Abwanderung treibt, weniger die Tarifgestaltung sondern in erster Linie die Unzufriedenheit, die Dropraten in Gesprächen, die Leitungsqualität oder das Gespräch selbst ist. Zeitnah eingreifen zu können, zum Zeitpunkt des Netz- oder Gesprächsausfalls bereits zu reagieren, ist hier ein Überlebensfaktor in der Vermeidung von Kundenabwanderung und entwickelt sich zu zeitkritischen Themen.

Beim Thema Datenvielfalt geht es um die gesamten unstrukturierten Daten, Video-Fleets und ähnliche Dinge, die wir tagtäglich produzieren. 80 % des Datenwachstums, das wir heute, kommen letztendlich von unstrukturierten Daten.

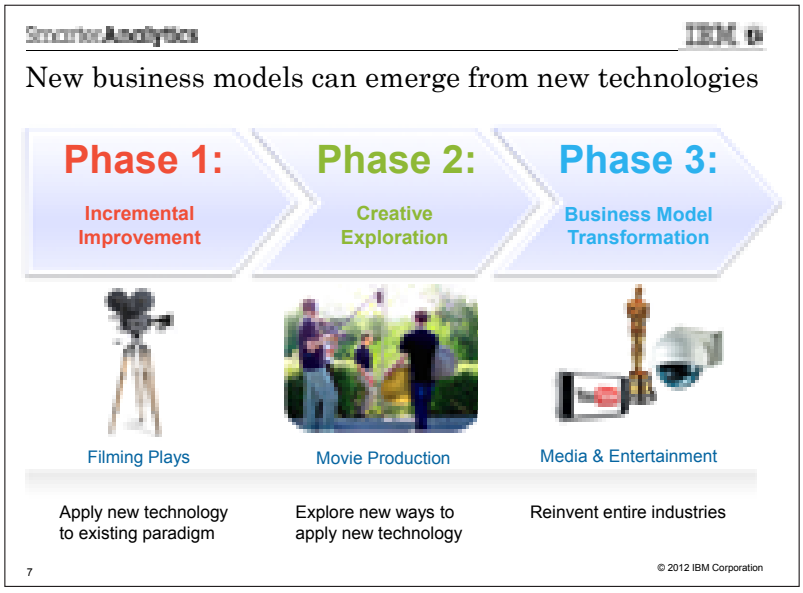

Bild 6

Eine weitere Analogie zurückgreifend auf 100 Jahre ist die Filmindustrie (Bild 6). Ursprünglich ging es vor 100 Jahren in der Filmindustrie um die ersten Schwarz-Weiß-

Stummfilmen, um das Abfilmen von Realität; fahrenden Zügen, bewegende Wolken, ähnliche Dinge. Erst später hat man begonnen, eigene Drehbücher zu schreiben, eine eigene Realität zu generieren und letztendlich erleben wir heute, dass die Filmindustrie eigene Geschäftsmodelle produziert, d.h. die Realität transformiert. Abfilmen von Realität, Kreieren von eigener Realität, Transformation von Realität.

Im Internet ist das Gleiche passiert. Die ersten Websites in der Gestaltung des WorldWideWeb waren wirklich eine Abbildung von Realität. Die Abbildung des Katalogs, der Imagebroschüre, online, digital, waren die ersten Websites, die am Netz verfügbar waren. Später kam das Übergehen in die eigene Schaffung einer Realität, B2B-Prozesse, B2C-Prozesse im Internet und letztendlich heute aber dann Unternehmen, die Born on the Web sind. Die transformieren ganze Industrien dank des Internets.

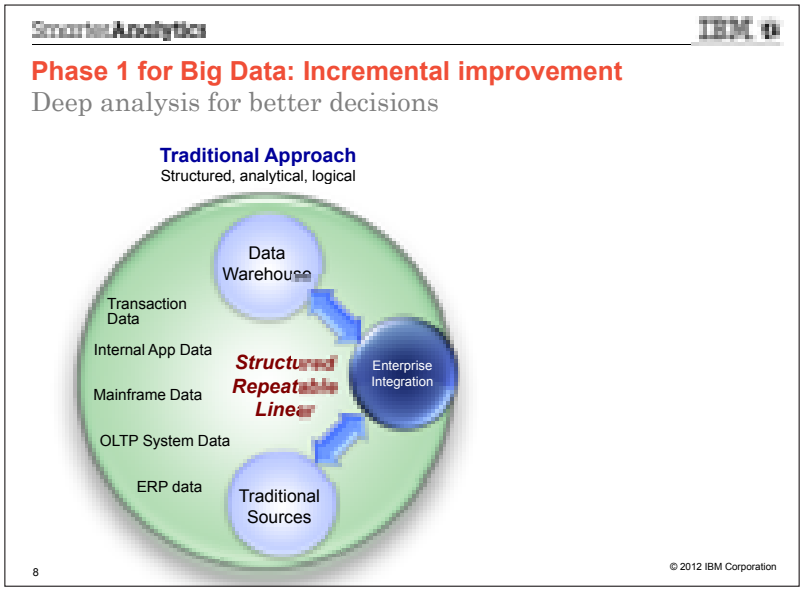

Bild 7

Die gleiche Entwicklung passiert mit Big Data im traditionellen Sinn, wo es darum geht, strukturierte Daten nach dem Modell die Qualität, die ich an Daten hinein füttere bekomme ich auch als Qualität der Analyse heraus, strukturiert in Datenbanken abzugreifen (Bild 7).

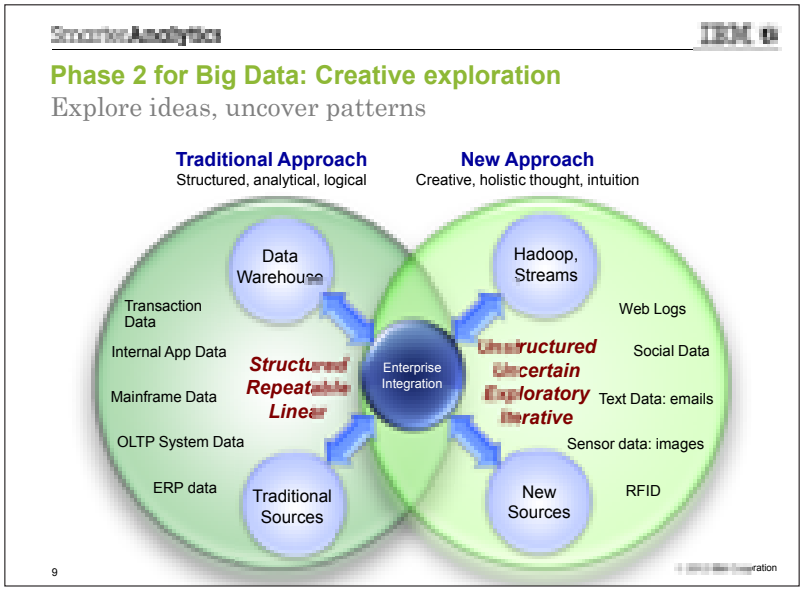

Bild 8

In der zweiten Phase das Kreieren neuer Realitäten, die gesamte Unstrukturierte, Hadoop Streams, alles, was die Twitts, die im Netz existieren, zusammenzuführen (Bild 8). Hier bin ich im Unterschied zur strukturierten Welt nicht mehr in Kontrolle der Qualität der Daten, aber ich bin in der Kontrolle der Methodologie, der Algorithmen, diese Daten zu analysieren und auszuwerten. Hier gehe ich bereits einen neuen Weg in der Zusammenführung dieser zwei Welten.

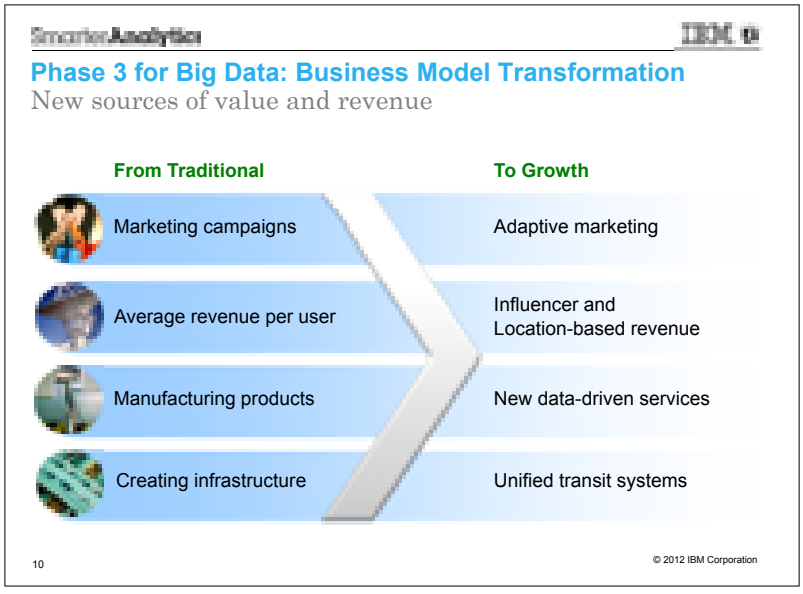

Bild 9

Wirklich spannend wird es, wenn man in den Bereich der Transformation hineingeht (Bild 9). Aus Zeitgründen kann ich diese Beispiele nicht ausführen, aber denken wir an das adaptive Marketing. Wir erleben das alle täglich. Sie tätigen einen Kauf auf Amazon und bekommen plötzlich in den Tagen danach viele Sonderangebote zu verwandten Produkten, zu passenden Produkten. Das Überführen von Marketingkampagnen mit all ihren Streuverlusten, mit ihren Ineffizienzen hin zu sehr adaptiven, sehr zugeschnittenen Marketingprogrammen. Die Bereitstellung von APPs, an der Oberfläche kostenlos, in Wahrheit zahle ich über die Verfügbarkeit meiner individuellen Information. Das sind alles Phänomene, die durchaus auch umstritten sind in manchen gesellschaftlichen Auswirkungen, aber die keine Frage sind im Sinne von einer sehr starken Adaptierung von Systemen auf den Benutzerwunsch hin.
Letztendlich gehen wir in das andere Extrem. In der öffentlichen Verwaltung, wo man von der Schaffung von Infrastruktur hinüberführt in die Integration von Infrastruktur.

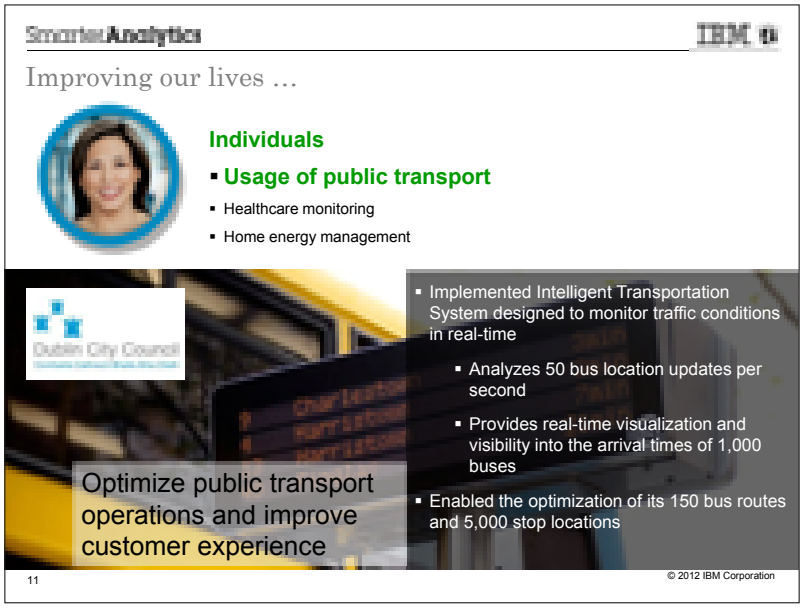

Bild 10

Ich habe Ihnen in den Folien hier drei Beispiele genannt, die wir über die letzten Jahre realisieren konnten. Vielleicht anfangend mit der Stadt Stockholm oder der Stadt Dublin, wo wir gezielt versucht haben, Transportlösungen zu generieren, die übergreifend sind (Bild 10). In Dublin, gemeinsam mit dem State of Dublin und der Stadtverwaltung Dublin die Steuerung der öffentlichen Busbetriebe auf Basis von GPS Daten, auf Basis von Verkehrsanalyse so zu steuern, dass man weniger Ausfälle hat, weniger Abweichung von Routen oder Fahrplänen hat und hier einen effizienteren Einsatz der bestehenden limitierten Ressourcen garantieren kann. In Stockholm, wo wir zusammengeführt Road Pricing und Gebühren für Brücken und Autobahnen mit einer Verkehrsinformation und mit der Steuerung des öffentlichen Verkehrs, das zu 20 % Reduktion im Verkehrsaufkommen in dem Individualverkehrs geführt hat - sehr maßgebliche andere Zugänge zum Thema öffentlicher Infrastruktur.

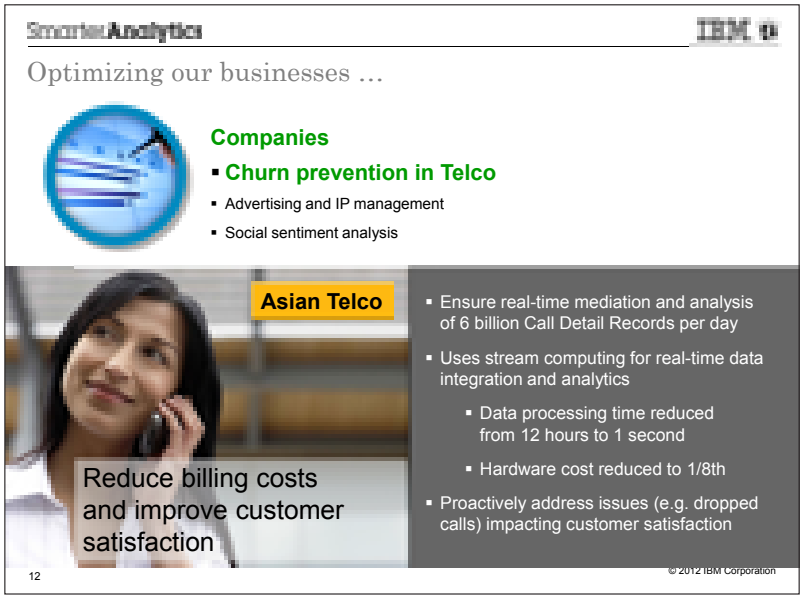

Bild 11

Ein anderes Thema habe ich bereits angesprochen (Bild 11). Ein Telekommodell in Indien, das sich in mit der Datenanalytik genau um diesen Bereich der Kundenabwanderung und der Steigerung der Loyalität kümmert.

Bild 12

Letztendlich gibt es viele Bereiche für gesellschaftliche Auswirkungen. Ich habe aus dem Bereich Energie schon über die Smart Meters gesprochen. Hier das Beispiel von Vestas, einer dänischen Firma, die international Masten aufstellt, die, wenn sie einmal aufgestellt sind,

nicht mehr mobil sind (Bild 12). Es sind keine mobilen sondern feste Assetts und die Nutzung von Analytics, um hier Wetterdaten zu optimieren, um den optimalen Standort zeitnah zu definieren, wo so ein Mast aufgestellt wird, sind wesentliche Faktoren der Wettbewerbsfähigkeit von Vestas, weil es auch aus umweltpolitischer Sicht ganz wesentliche Vorteile hat, die daraus gesellschaftlich erschlossen werden. Keiner braucht einen Windpark, wo die Ausnutzung der Windräder nicht optimal ist.

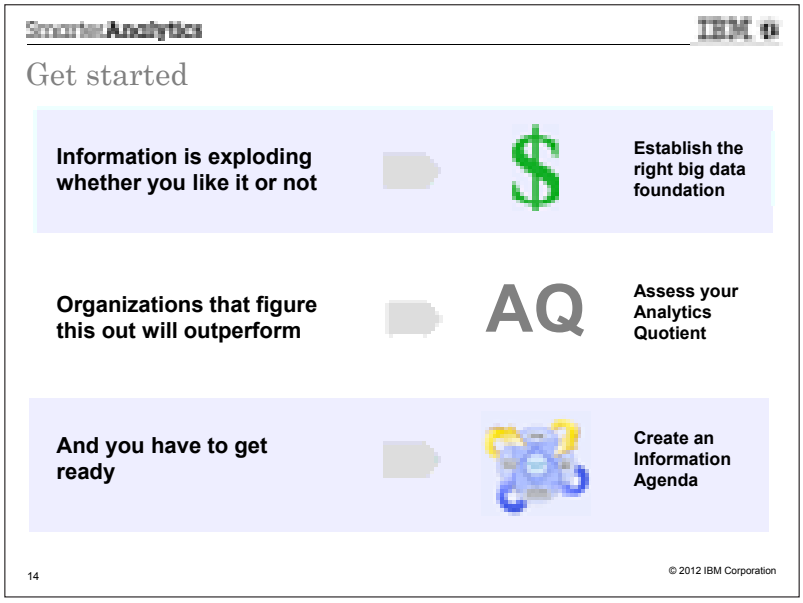

Bild 13

Lassen Sie mich auf diesem Chart (Bild 13) kurz zusammenfassen, was ich als Thema und ein bisschen als Postulat für den Tag heute und für meine Nachredner aufgreifen wollte. Zum einen, wie ich bereits eingangs sagte, ist Big Data Realität, passiert heute und ist kein technologieverliebtes Thema, das zum Selbstzweck existiert. Wir sind erst an der obersten Spitze des Eisbergs und haben hier noch einiges vor uns an der Entwicklung, die uns an Datenvolumen, Datenintensität, Geschwindigkeit und Vielfalt über die nächsten Jahre begleiten wird.

Unternehmen, Organisationen, Verwaltungen, die sich heute pro aktiv mit dem Thema Big Data auseinandersetzen, werden ganz wesentliche Wettbewerbsvorteile für sich erschließen, gesellschaftliche Vorteile für uns alle in der Menschheit erschließen. Das Ganze funktioniert nur, wenn es pro aktiv geschieht. Eine Informationsagenda zu erstellen, die aussagt, was ich in meinem Unternehmen mit dem Rohstoff Information machen will, ist eine wesentliche Grundvoraussetzung.

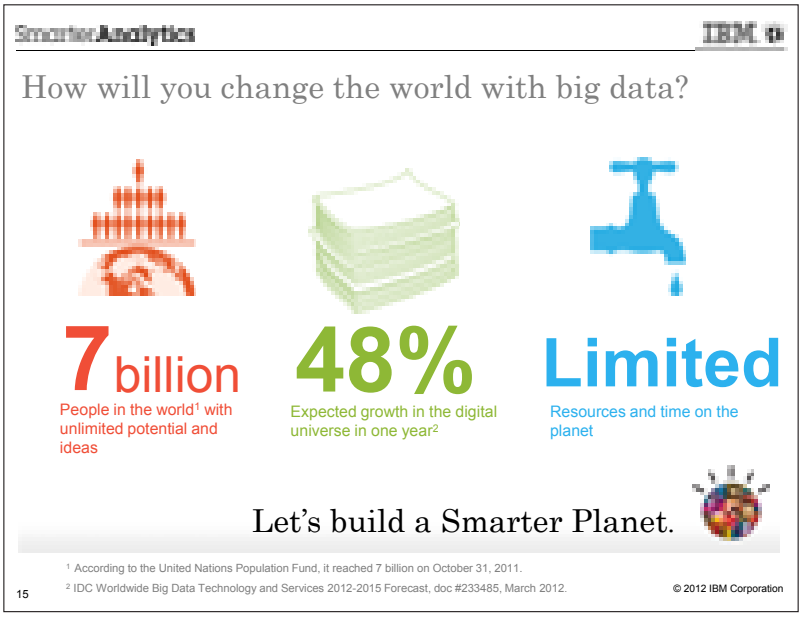

Bild 14

Ich darf mit diesem Chart (Bild 14) enden. Ich habe versucht, an Kundenbeispielen zu zeigen, die letztendlich eines gemeinsam haben. Es geht um die effizientere Nutzung von Ressourcen. Die drei Beispiele hier auf der letzten Folie sind drei ganz wesentliche und kritische Ressourcen, mit denen wir alle tagtäglich auseinandergesetzt sind. Sieben Billiarden Menschen, ein unglaublicher Ressourcenschatz, den es gilt optimal einzusetzen. Am anderen Extrem die limitierten natürlichen Rohstoffe, natürlichen Ressourcen, die uns auf diesem Planeten zur Verfügung stehen und in der Mitte vielleicht dieses Wachstum, dieses sich erneuernden Datenvolumen, die sich alle zwei Jahre verdoppeln und eine unglaubliche Ressource sind und die es gilt, effizient zu nutzen. Sie alle hier in diesem Raum darf ich ein bisschen auffordern, über die Gedanken und Dialog des heutigen Tages zu überlegen: Wo ist mein Platz in dieser Konstellation der Optimierung von Ressourcen durch den Einsatz von Big Data.

3 Forschung, Innovation und Ausbildung in Big Data Analytics – Chancen und Herausforderungen

Prof. Dr. Volker Markl, Technische Universität Berlin

In meinem Vortrag möchte ich über Forschung, Innovation und Lehre im Bereich der Datenanalyse sprechen. Big Data ist derzeit in aller Munde. Es gibt dazu einen sehr prägenden Bericht, eine Studie von McKinsey, die festgestellt hat, dass Big Data die neue Grenze für Innovation, Wettbewerbsfähigkeit und Produktivität insgesamt im nächsten Jahrzehnt darstellen wird.

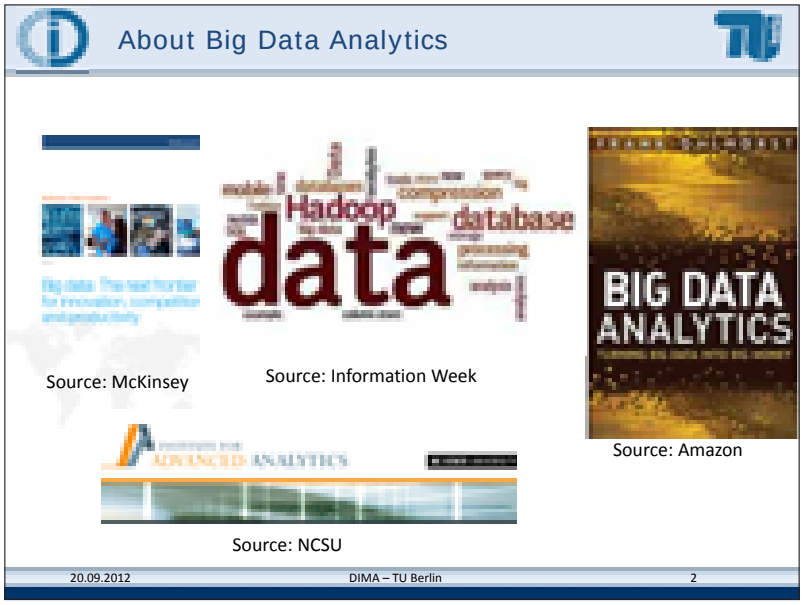

Bild 1

Ich werde versuchen, die sprachliche Verwirrung, die durch den ständigen Gebrauch dieses Hype-Begriffs entstanden ist, aufzulösen und „Big Data“ durch Definitionen und Beispiele zu präzisieren. Jedoch zunächst eine Tag Cloud aus Information Week, welche Technologien wie Datenbanken, Kompression, skalierbare Technologien und große Datenmengen, aber auch spezielle Systeme wie das Hadoop System als wichtige Merkmale von Big Data herausstellt (Bild 1). Dabei sollte angemerkt werden, dass im Bereich der Datenbanksysteme natürlich schon seit Jahrzenten an „Big Data“ geforscht und entwickelt wurde, und kommerzielle Lösungen geschaffen wurden.

Es gibt zu dem Thema Big Data Analytics inzwischen sehr viel Literatur. In den USA und anderswo entstehen derzeit spezielle Studiengänge, die sich damit befassen, Datenanalysten oder sogenannte Datenwissenschaftler, Data Scientists, auszubilden. Hier ist die North Carolina State University ein Beispiel. Aber es gibt auch andere Universitäten, an denen derzeit derartige Studiengänge entstehen.

Ich möchte zunächst darstellen, was überhaupt große Datenmengen sind. Jeder spricht davon. Aber was ist das eigentlich? Ich werde dazu Beispiele geben, „Big Data“ definieren, dann

über grundlegende Technologien zu sprechen, die uns befähigen große Daten zu analysieren, ferner Anwendungen skizzieren, und letztendlich die „Big Data“ Herausforderungen aus der Sicht eines Wissenschaftlers mit starkem Wirtschaftsbezug darstellen, ehe ich mit einen Call to Action meine Vortrag beende.

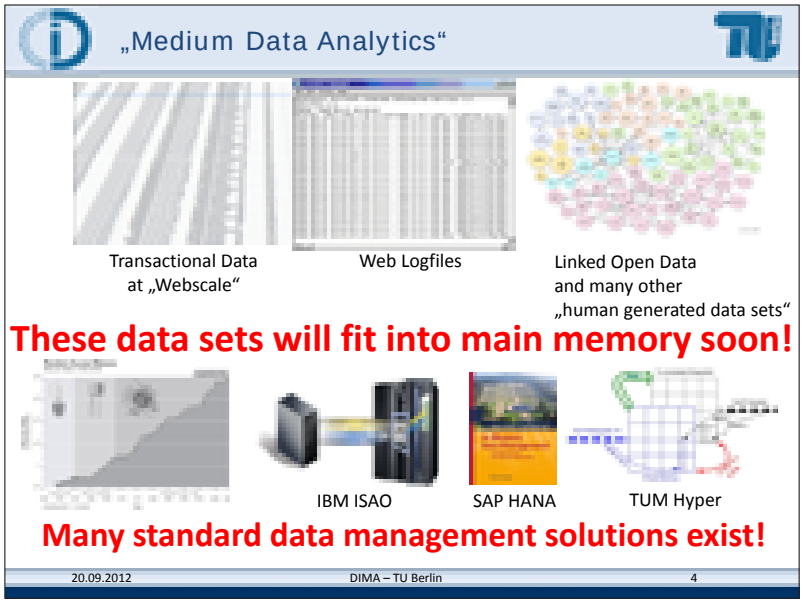

Bild 2

Was sind denn eigentlich große Daten (Bild 2)? Vielleicht sollte man sich erst einmal fragen, was denn mittelgroße Daten sind. Mittelgroße Daten treten in vielen Transaktionssystemen auf, auch auf Web Scale, z.B. bei der Analyse von Web Server Logs oder auch im Rahmen des semantischen Webs oder vielen Anwendungen von Open Data und Linked-Open-Data Dies sind alles Datenmengen, die heutzutage oftmals (fälschlich) als große Daten bezeichnet werden. Denn man kann, dass eigentlich viele dieser Datenmengen bald in den Hauptspeicher von modernen Rechnern passen werden und in vielen Fällen von Standardtechnologien wie klassischen Datenbanksysteme oder Hauptspeicherdatenbanksysteme zur Datenanalyse verwendet werden können. Dabei treten im ersten Moment nicht zwingend die fundamentalen Probleme bezüglich der technologischen Herausforderungen auf, welche ich gleich als Probleme der „großen Datenanalyse“, des „Big Data Analytics“ bezeichnen werde. Der Grund ist, dass wir immer größere Hauptspeicher haben werden. Der Trend ist auch erkannt, verschiedenste größere IT-Anbieter und innovative Start-Ups bieten entsprechende Lösungen an, z.B. hat IBM ein System mit dem Smart Analytics Optimizer, SAP arbeitet am HANA System. Ferner gibt in diesem Bereich interessante Start-ups, in Deutschland beispielsweise die Firma Parstream in Köln, die eine sehr effiziente Datenbank auf Hauptspeicherbasis vertreibt und damit Datenanalysen im mittelgroßen Datenbereich realisiert. Es existieren viele weitere technische Lösungen in diesem Bereich, sowie einige neuartige Systeme.

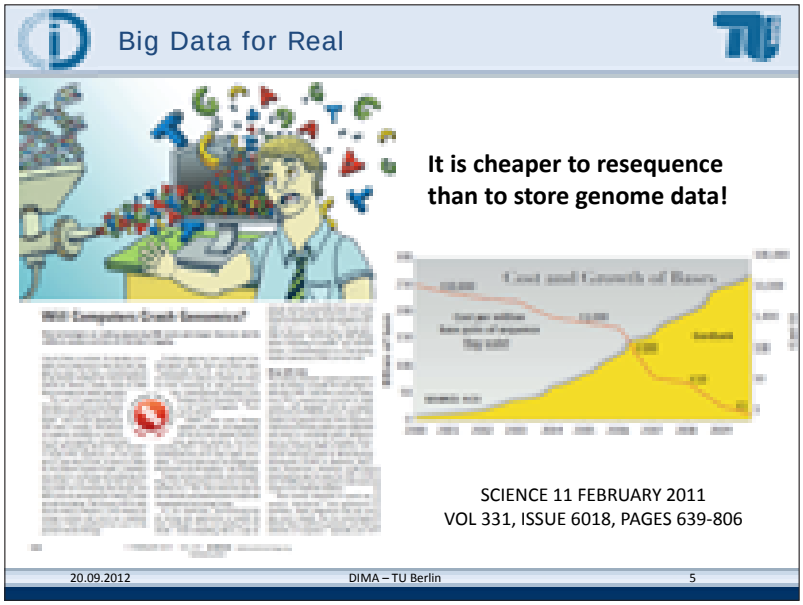

Bild 3

Was sind denn wirklich große Daten? Ich gebe hier ein Beispiel (Bild 3). Ein Artikel aus der Zeitschrift „Science“ hat den Bereich der Gensequenzanalyse beschrieben. Dabei wird festgestellt, dass es inzwischen aufgrund der Preise von Festplatten im Rahmen des Next Generation Sequencing günstiger ist, Daten neu zu sequenzieren, als diese zu speichern. Das ist wirklich ein großes Datenproblem. Ich kann es mir nicht mehr leisten, die Daten zu speichern. Das ist Big Data par excellence.

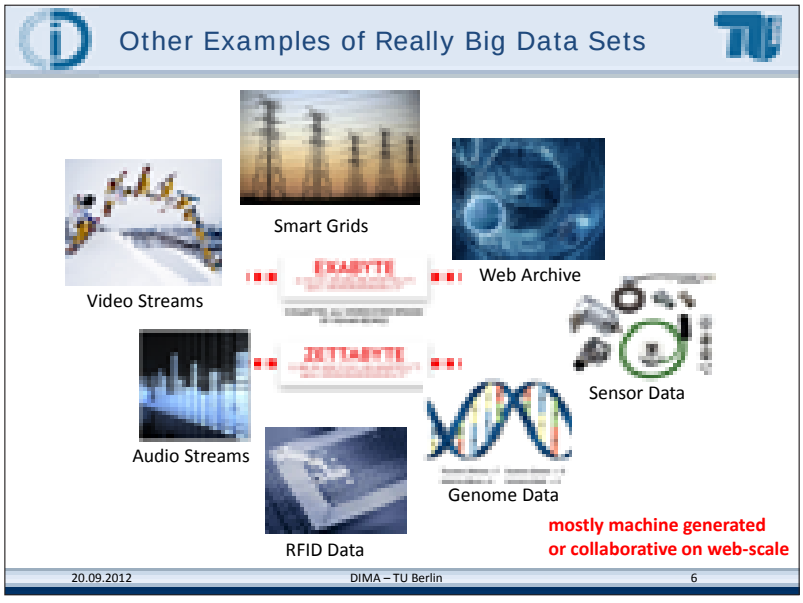

Bild 4

Weitere Beispiele für „Big Data“ sind Daten, die von Sensoren erzeugt werden, oder Daten, die durch Crowds im Internet entstehen (Bild 4). Wir werden riesige Datenmengen erhalten in allen Anwendungen, die mit Smart Grids zu tun haben, bei der Verarbeitung von RFID Daten, Audioströmen oder Videoströmen. Oder auch bei Web Archiven, d.h. wenn ich das Internet historisch auffasse und darüber Analysen machen will. Im Rahmen von Big Data entstehen riesige Datenmengen, die sehr schnell in die Größenordnung von Exabytes und Zettabytes kommen. Ganz wichtig wird in der Zukunft dabei die Analyse von Sensor-, Audio- und Videodatenströmen werden. Ich werde darauf noch zu sprechen kommen.

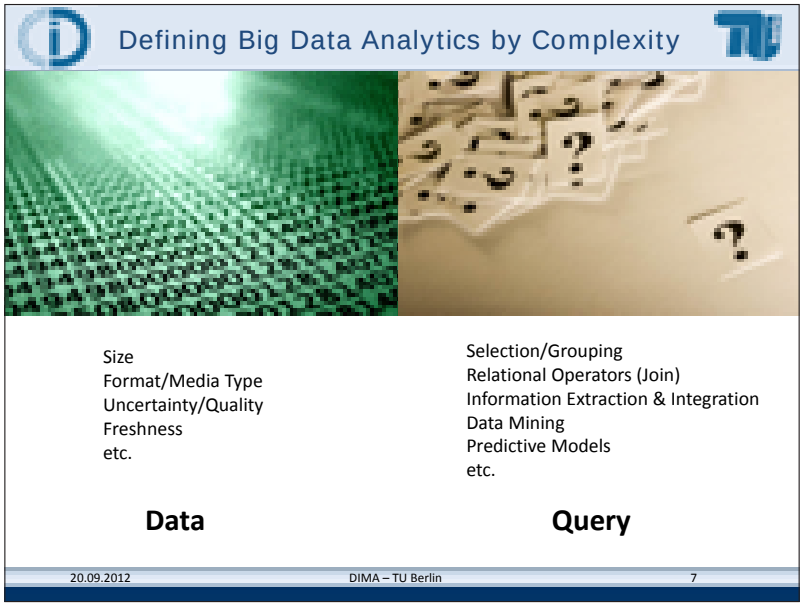

Bild 5

Was bedeutet Big Data Analytics (Bild 5)? „Big“ ist eigentlich ein Ausdruck dafür, dass die Komplexität zugenommen hat. Das Schlagwort „Big Data“ ist plakativ. Die Darstellung der Komplexität kommt durch die Reduktion auf den Begriff „Big Data“ in einigen Bereichen zu kurz. Lassen sie mich daher kurz auf die Komplexität der Datenanalyse und die damit verbundenen neuartigen Herausforderungen von „Big Data Analytics“ eingehen. Dabei ist wesentlich, dass Komplexität zum einen von den Daten her rührt, zum anderen aber auch von den Analysen. Im Hinblick auf Komplexität der Daten, ist es klar, dass die Größe des Datenvolumens ein wesentlicher Aspekt ist. Ein anderer Aspekt, der eine große Rolle spielt, ist die Frage: in welchen Datenformaten oder Medientypen liegen die Daten vor? Analysieren wir klassische relationale Daten, d.h. Daten, die in Tabellen strukturiert sind oder sind die Daten in irgendeiner Form komplexer strukturiert wie z.B. Textdaten, Audio- und Videodaten und andere Datenquellen? Dabei möchte ich noch anmerken, dass heutzutage oftmals sogenannte „NoSQL“ Systeme verwendet werden, um klassische strukturierte Daten, sogenannte relationale bzw. tabellarische Daten, zu analysieren. Im Hinblick auf die Datenverarbeitung sollte jedes Unternehmen beachten, ob dies unter Berücksichtigung von längerfristiger Wartbarkeit und unter Berücksichtigung der wichtigen Datenunhängigkeit von der Anwendung sinnvoll ist. Ich sehe derzeit viele spezialisierte Anwendungssystementwicklungen, die mit open-source NoSQL-Systemen Geschäftsabläufe und Datenhaltung implementieren, obwohl die Form der Realisierung aus Nachhaltigkeitsgesichtspunkten nicht sinnvoll ist.

Ein weiterer wesentlicher Aspekt ist, dass die Daten oft auch eine gewisse Unsicherheit besitzen. Ich bin mir nicht ganz sicher, ob die Information überhaupt so stimmt und ich muss das irgendwie feststellen und validieren. Folglich muss ich mit der Unsicherheit bei der Datenerhebung umgehen, mit der Tatsache, dass Fehler oder Inkonsistenzen in den Daten enthalten sind, Datenqualitätsprobleme verschiedenster Art und Ursache. Das sind komplexe Fragen auf der Datenseite, die man betrachten muss, wenn man „Big Data Analytics"-Systeme baut, sowie, wenn man „Big Data Analytics" betreibt.

Im Rahmen der Datenverarbeitung, bei den Datenanalysen, gibt es andere Formen der Komplexität. Die erste Frage ist, wie komplex die zur Beantwortung einer Fragestellung benötigten Analyseverfahren (d.h., die sogenannten Anfrageverarbeitungsalgorithmen) sind. Das einfachste sind hierbei so genannte Selektionen und Gruppierungen, .d.h. ich nehme eine große Datenmenge, wähle eine Teilmenge davon aus und füge diese in irgendeiner Form zusammen, z.B. Berechnung von Durchschnittswerte. Dazu gibt es inzwischen sehr viele Systeme, die das auch bei riesigen Datenmengen leisten. In der Tat wurde „Big Data" durch die Beantwortung von Fragestellungen mit dieser Komplexität populär. Wichtigster Vertreter ist dabei das Open-Source System Hadoop, das es im Wesentlichen ermöglicht, große Datenmengen im Wesentlichen zu selektieren und zu aggregieren und damit auch schon komplexe Fragen wie Webloganalyse usw. durchzuführen. Das funktioniert recht gut, solange man das System nicht für komplexere Aufgaben missbraucht (was allerdings derzeit aufgrund des Hypes um Hadoop immer häufiger geschieht). Allerdings werden die Anforderungen bzgl. der Anfragen immer größer. Von meinem Vorredner haben wir zu erheblich komplexeren Fragestellungen schon einige Beispiele gehört, wo es darum geht, Datenmengen zu korrelieren, also sogenannte relationale Joins oder Verbünde durchzuführen, oder Informationen aus vielleicht sogar sehr heterogenen Quellen zu integrieren. Dabei kann es sein, dass strukturierte Information erst durch Informationsextraktionen aus Textdaten oder aus anderen Datenquellen wie Audio- oder Videoströmen abgeleitet werden muss. Letztendlich, als höchste Komplexitätsstufe, kann man auch Modelle aus Daten ableiten, um Vorhersagen zu treffen. In dem Feld der „Predicitive Analytics" will man derartige komplexe statistische Vorhersagemodelle aufbauen, um damit Planspiele durchzuführen.

Das sind neue Komplexitäten, die auftreten. Big Data ist ein schönes Schlagwort, in dem der wichtige Aspekt der „Analyse" natürlich ein bisschen zu kurz kommt. Dies ist sicher auch den eingeschränkten Analysefähigkeiten von Hadoop geschuldet.

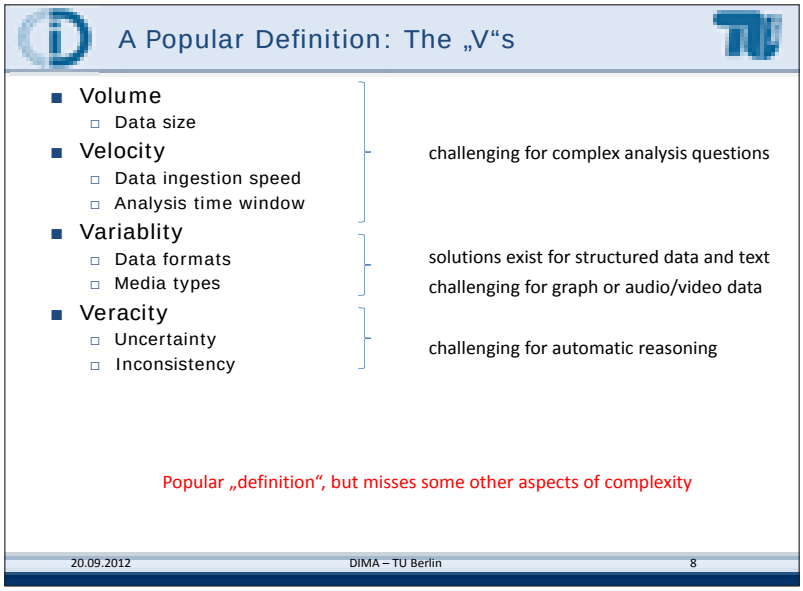

Bild 6

In dem Zusammenhang vielleicht noch einmal eine sehr populäre Definition von Big Data (Bild 6), die das Ganze im Englischen auf drei oder vier „V“s reduziert, d.h. das Datenvolumen (volume), die Geschwindigkeit der Datenverarbeitung (velocity) mit den zwei Aspekten: wie schnell werden die Daten generiert und wie schnell müssen Antworten auf komplexe Analysefragen generiert werden, die Variabilität (variability), die im Wesentlichen unterschiedliche Datenformate bedeutet und die Wahrhaftigkeit der Daten (veracity), d.h. die Fragen der Unsicherheit und vielleicht auch Inkonsistenz der Daten.

In diesem Bereich gibt es natürlich einige Herausforderungen, in Teilen auch harte technologische Forschungsfragen, die man lösen muss, wenn man derartige Systeme bauen oder betreiben will. Es ist z.B. ein Problem, wie ich mit großen Datenvolumina sehr kurze Antwortzeiten, also nahezu Echtzeitanalysen, oftmals auch Online-Analysen genannt, erreichen kann. Das ist einfach für einfache Analysen, wenn es nur darum geht, Daten zu selektieren und zu finden, so wie beispielsweise Google als Beispiel einer großen Datenanalyse aus einem Webindex Webseiten findet. Das ist unkritisch. Diese Technologien beherrschen wir. Schwieriger wird es, wenn die Analysen komplexer werden, wenn wir, wie ich vorhin schon beschrieben hatte, Vorhersagen treffen wollen, Informationen aus verschiedenen Quellen integrieren wollen, Videos verarbeiten wollen, etc. Dabei gibt es durchaus noch offene Probleme, die neue Forschungsansätze und neuartige Technologien erfordern. Es gibt bereits einige für die Anwendung im Markt reife Lösungen für strukturierte Daten und Textdaten. Die Zukunft wird aber vielschichtiger. Audio und Video werden in Zukunft immer wichtiger werden, weil sehr große Datenmengen heutzutage auf Plattformen wie YouTube generiert werden, oder z.B. Telefonmitschnitte in Call Centers gesammelt werden, die analysiert werden sollten, um z.B. die Kundenzufriedenheit mit Produkten zu bestimmen.

Ferner gilt natürlich auch, dass, wenn ich Vorhersagemodelle bauen und automatisch Schlussfolgerungen über die Daten ziehen will, ich mit einer damit verbundenen Unsicherheit umgehen muss, sowie mit möglichen Inkonsistenzen, die in so großen Datenmengen auftreten können.

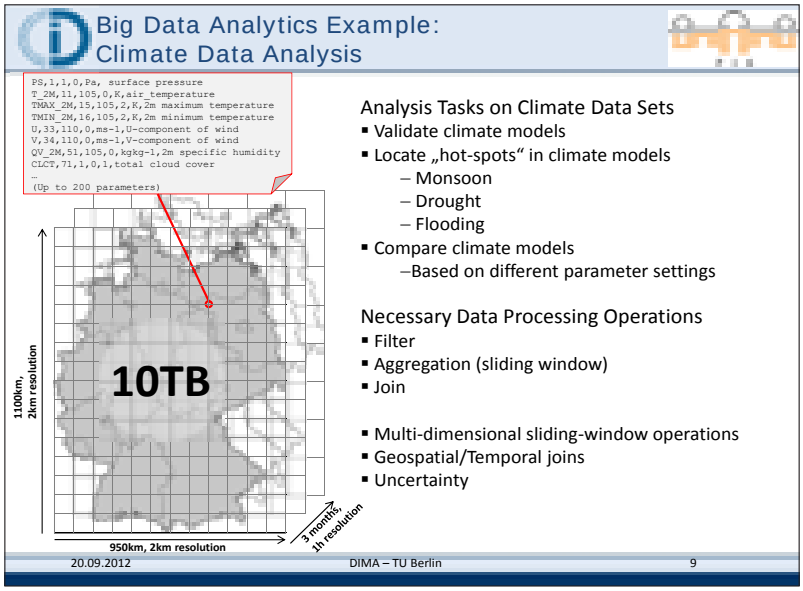

Bild 7

Ich zeige Ihnen jetzt ein Beispiel von „Big Data Analytics“ im wissenschaftlichen Bereich, und zwar die Analyse von Klimadaten (Bild 7). Klimaforschung befasst sich damit, Klimamodelle zu erstellen. Im Wesentlichen kann man sich das als eine erweiterte Wettervorhersage vorstellen, die nicht nur auf eine bestimmte Region abzielt, sondern wirklich global versucht, das Ganze darzustellen, Dinge für die nächsten Jahre abzuleiten. Das basiert auf mathematischen Modellen. Einfach gesagt sind das mehrere Differenzialgleichungen, die auf einem Datenmodell angewendet werden. Dieses Datenmodell ist so beschaffen, dass ich ein Gitter über die Erde lege. Dieses Gitter hat eine bestimmte Auflösung und an den Schnittpunkten von mehreren Gitterzellen berechnet ein Klimamodell bestimmte Parameter für jeden Gitterpunkt als Vorhersage. Dabei muss man bis zu 200 Parameter berücksichtigen, z.B. Oberflächenluftdruck, Lufttemperaturen, Windgeschwindigkeiten, Luftfeuchtigkeit usw. Wenn ich das z. B. für Deutschland mit einem Gitter von zwei Kilometern in Länge und Breite für drei Monate mit einer Stundenauflösung in der Zeit durchführe, bin ich schon bei 10 Terrabyte an Daten. Aber wie schon gesagt, befasst sich Klimaforschung mit der Welt als Gesamtsystem über Jahre oder Jahrzehnte. Dann bis ich sofort bei Petabytes oder sogar noch größeren Datenmengen, die ich hier betrachten muss.

Die interessante Sache ist, dass das Generieren dieser Daten gut beherrscht wird. Man setzt Rechencluster ein, 1000+ Rechner, und führt die Simulation durch. Das Problem ist die Analyse der Daten. Ich habe hier einige Beispielfragestellungen aufgelistet. Besonders interessant finde ich ein Beispiel, das für den Weltklimarat derzeit ein großes Thema ist. Es gibt einige Klimamodelle, die vorhersagen, dass der Monsun in Südostasien in ca. einer Dekade für eine gewisse Zeit, ca. eine Dekade lang, nicht mehr auftreten wird und dann sehr unregelmäßig, sehr sporadisch auftreten wird. Falls dies zuträfe, hätte dies natürlich immense Auswirkungen auf die Landwirtschaft und die Bevölkerung insgesamt in dieser Region. Daher ist es wichtig zu wissen, ob das überhaupt stimmt. Das Problem ist, dass diese Vorhersage basierend auf einem Klimamodell auf Annahmen beruht. Nun müssen Klimaforscher dies validieren und prüfen, was sich ändert, wenn man andere Annahmen macht bzgl. CO_2 Ausstoß oder anderen Aspekten des Klimamodells. Dafür muss ich weitere

Klimamodelle berechnen, was noch einmal einige Petabytes an Daten bedeutet. Dann muss man diese riesigen Datenmengen vergleichen, also in Beziehung setzen, korrelieren. Im Beispiel bedeutet dies, dass ich den Monsun in jedem dieser riesigen Petabyte großen Datenansammlungen finden muss und bewerten muss. Ein Monsun ist ein relativ komplexes Gebilde aus Luftfeuchtigkeit, Druck usw. in dem mehrdimensionalen Datenraum. Das muss man erst einmal finden und dann über verschiedene Modelle vergleichen, technologisch sind hierzu räumliche Joins und Ähnlichkeitssuche auf Petabyte-großen mehrdimensionalen, dichtbesetzten Datenbeständen erforderlich. Dies ist wirklich ein „Big Data Analytics" Problem.

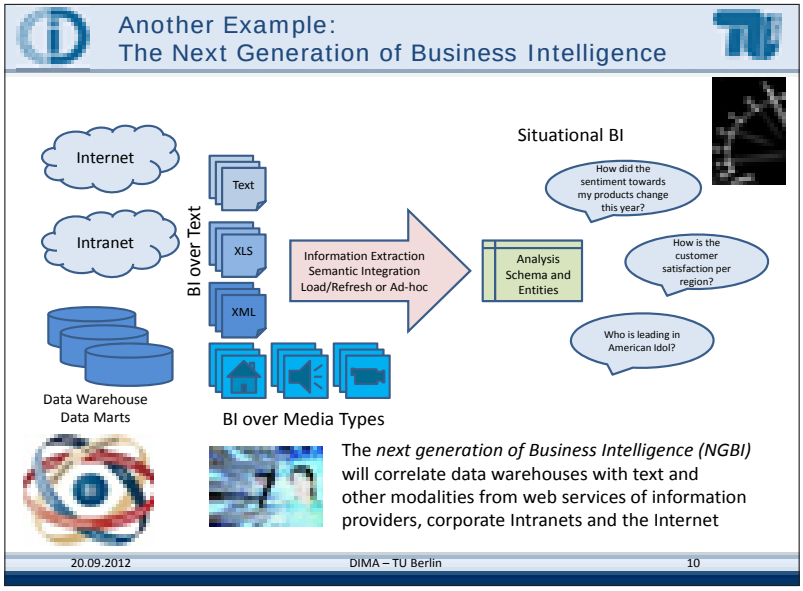

Bild 8

Eine weitere, eher wirtschaftliche Anwendung von Big Data, die schon angeklungen ist, begründet sich durch meine Vision der nächsten Generation von Business Intelligence (Bild 8). Wenn ich mir heute Business Intelligence Anwendungen ansehe, dominiert die klassische Business Intelligence: In einem Data Warehouse verwalte ich meine unternehmensinternen Daten, dazu kaufe ich mir vielleicht ein paar Informationen von externen Datenanbietern und versuche daraus Unternehmensentscheidungen abzuleiten. In Zukunft wird es immer mehr eine Rolle spielen, auf Daten im Internet zu reagieren, aber auch auf Intranets und spezielle Informationen, die in anderen Modalitäten vorliegen, wie z.B. die Call-Centerdaten, die ich eben angesprochen hatte. Und ich möchte aus diesen Daten Informationen ableiten, möglichst zeitnah, um schnell reagieren zu können. Das allerwichtigste Business Intelligence Tool für jeden von uns in diesem Zusammenhang ist heutzutage eine Internet-Suchmaschine wie Google. Wenn wir irgendetwas auf die Schnelle wissen müssen, über die Konkurrenz, über den Ort, wo wir gerade hinreisen, usw., suchen wir Informationen dazu zunächst einmal bei Google oder einer anderen Suchmaschine. Das ist aber ein sehr unstrukturierter Prozess, insbesondere dann, wenn ich komplexere Studien ableiten will, z.B. alle Informationen über den Kunden, den ich gleich besuchen werde.

Ein konkretes Beispiel: ich hatte letztens Besuch von einem Herrn, der lange bei der Weltbank gearbeitet hat und jetzt in Washington D.C. ein Start-up zur Beratung von Investment-

firmen im Bereich nachhaltiger Technologien eröffnen will. Er hätte sehr gern regelmäßig aktuelle Dossiers über nachhaltige Technologien und welche Bewegungen es dort gibt aus dem Web abgeleitet.

Solche Informationen abzuleiten, geschieht heute durch gezielte Websuche, die sehr arbeitsintensiv und aufwändig ist. Eigentlich ist das ein strukturierter Prozess. Ich würde dafür gern Informations-, Integrations- und Extraktionsalgorithmen verwenden und die Informationsbereitstellung über eine große Datenbasis automatisieren, ad-hoc, für alle möglichen Analysefragen.

Das ist eine Variante von Business Intelligence. Im Wesentlichen bedeutet das, dass wir verschiedenste Datenquellen, Texte, aber auch anders strukturierte Daten wie Tabellen, XML-Dokumente und Bilder, Audio- und Videodaten heranziehen, um ähnlich zu einem klassischen Data Warehouse ETL Prozess erst Informationsextraktion zu betreiben und dann diese komplexen Modalitäten semantisch zu integrieren. Danach erzeugt man - im Idealfall ad hoc, in vielen Fällen heutzutage aber zeitlich entkoppelt mit Load and Refresh - ein Analyseschema, wobei das Analyse-Schema, das hier entsteht, im Wesentlichen auf den Extraktionsmethoden basiert, die ich zur Verfügung habe. Danach kann ich unter Anwendung des Analyseschemas Fragen beantworten, wie z.B.: Wie hat sich jetzt die Einschätzung über meine Produkte in diesem Jahr verändert? Das möchte ich wissen basierend auf Blogs oder basierend vielleicht sogar auf YouTube Videos, weil heutzutage immer mehr Videobotschaften (Videoblogs) entstehen.

Oder mich interessiert: Wie messe ich meine Kundenzufriedenheit pro Region für verschiedene Produkte, wie bewerte ich diese? Das möchte ich durch einen Mitschnitt und automatischer Analyse der Beschwerden im Call Center ableiten.

Eine andere mögliche Business Intelligence Frage an das Internet. Ich möchte gern wissen, wer bei „American Idol“, einer etwas seriöseren Variante von “Deutschland sucht den Superstar“, gewinnen wird? Das ist insbesondere für die zukünftige Vermarktung und Werbung mit Stars sehr spannend. Diese Information kann man aus Blogs oder Twitter Feeds ableiten. Wenn ich derartige Fragen ad-hoc beantworten will, schafft dies eine zukünftige Generation von Business Intelligence, die weit über das hinausgeht, was mit heutigen Systemen möglich ist. Es gibt natürlich punktuelle Lösungen, die aber sehr teuer und aufwändig für eine spezielle Fragestellung entwickelt werden. Aber eigentlich möchte ich hier ein System, das mir die Angabe von Anfragen analog zur Suche von Dokumenten im Web oder zur ad-hoc Analyse von Data-Warehouses erlaubt. Dies existiert heutzutage noch nicht, aufgrund verschiedenster technologischer Schwierigkeiten.

Bild 9

Es gibt viele weitere Anwendungen, und ich habe hier noch ein paar Beispiele aufgelistet (Bild 9). Bis jetzt hatte ich über Dinge gesprochen, die für riesige Forschungszentren und Unternehmen relevant sind. Aber daneben ist ein wesentlicher Aspekt, dass Big Data auch nach Hause kommt. Ich habe mir vor einem Jahr ein Smart Home gebaut, und in diesem Smart Home sind sehr viele Sensoren. Es sind Kameras drin. Jeder Taster generiert Informationen. Auf dem Dach ist eine Wetterstation, die generiert Wetterdaten und misst Regen, Windgeschwindigkeiten usw. Damit kann man auch komplexe Steuerungen durchführen, z.B. bei der Heizung, Fenstern, etc. Mein Haus produziert pro Tag viele Gigabyte an Daten, die man alle analysieren kann. Das ist Big Data Analytics zuhause. Das heißt, dass das Big Data umfasst nicht nur große wissenschaftliche Fragestellungen oder Unternehmensentscheidungen, sondern es kommt bis nach Hause. Darin stecken riesige Potenziale für eingebettete Datenbanken und komplexere Analysen.

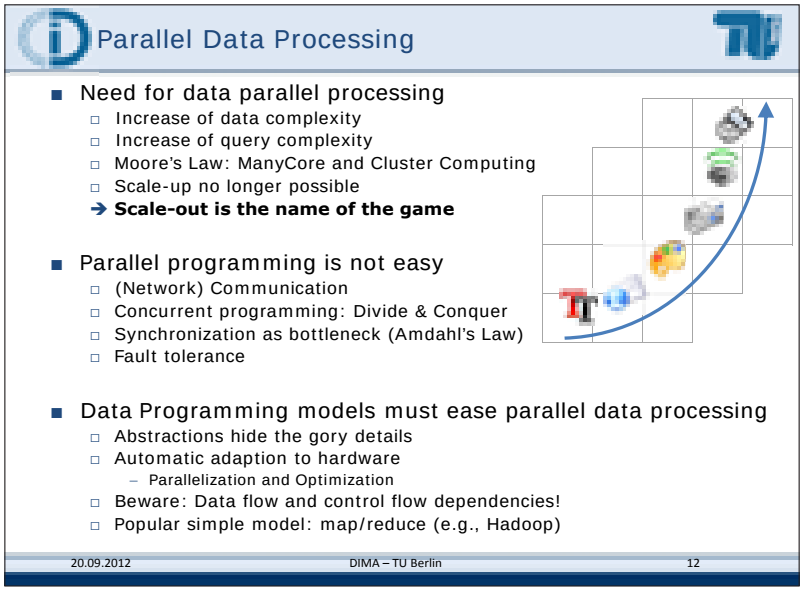

Bild 10

Was bedeutet das jetzt eigentlich von der technologischen Seite (Bild 10) her? Ein ganz wichtiger Aspekt ist dabei die parallele Datenanalyse, die aufgrund der Veränderung der Rahmenbedingungen von Moore's Gesetz die Basis für alle großen Datenverarbeitungsfragen wird. In der Vergangenheit wurden wir durch immer schnellere Rechner, d.h., höhere Taktfrequenzen von Prozessoren verwöhnt. Die nächste Generation von Rechnern lief einfach schneller, und um komplexere Probleme zu lösen, reichte es, bestehende Technologien weiterzuverwenden und einfach einen neuen Rechner zu kaufen, oder sogar ein Anruf bei IBM, um die Taktfrequenz ín meinem Großrechner höher zu setzen. Das klappt heute leider nicht mehr, denn durch physikalische Grenzen ist heutzutage die Möglichkeiten, um Leistungssteigerungen erzielen können, auf Parallelisierung beschränkt. Wenn Sie sich erinnern, so hat vor 10 Jahren jeder Prozessorhersteller mit Gigahertz geworben. Heute wirbt jeder damit, wie viele Rechnerkerne man vielleicht hat. Das bedeutet Parallelverarbeitung. Wir müssen uns damit befassen, wie wir die existierenden Systeme parallelisieren können, um mit diesen großen Datenmengen umgehen zu können. Im Wesentlichen bedeutet das, dass wir uns mit parallelen Algorithmen befassen müssten, was aber überhaupt nicht leicht ist. Hier geht es um das Prinzip von Teile und Herrsche, d.h., ein großes Problem in viele kleine Teilprobleme zu zerlegen, die unabhängig voneinander verarbeitet werden und letztendlich in ein Endergebnis zusammengesetzt werden müssen. Dabei sind Synchronisation und Kommunikation erforderlich, z.B. durch shared memory oder verteilte Dateisysteme, oder Netzwerkkommunikation, die ich auf einmal Rechner im Cluster betreibe. Wichtige Aspekte hier sind nebenläufige Programmierung. Es geht um Synchronisation, die sehr schnell ein Flaschenhals wird, weil einige Probleme sich aufgrund des Gesetzes von Amdahl nicht gut parallelisieren lassen, sobald ich eine serielle Komponente in meinem Programm habe. Sobald ich in größere Rechencluster gehe, wird Fehlertoleranz auch sehr schnell ein Problem. Das ist einfach ein Ergebnis der Zuverlässigkeitstheorie. Wenn ein Rechner mit einer Wahrscheinlichkeit von 99 % funktioniert und ich auf einmal 10.000 habe, ist die Wahrscheinlichkeit, dass alle Rechner funktionieren 99 % hoch 10.000. Das wird sehr schnell sehr klein, d.h. damit muss man sich auch befassen.

Parallelprogrammierung und auch die Programmierparadigmen sind ein Thema. Es gibt ein sehr populäres, sehr einfaches Modell der Parallelverarbeitung durch funktionale Programmierung, genauer gesagt, durch die zwei Funktionen zweiter Ordnung, map und reduce. Dieses Programmiermodell wurde von Google für datenintensive Anwendungen vorgeschlagen und ist die Basis des Open-Source Systems Hadoop. Da gibt es Beispiele, wo man sagt, dass SQL als relationale Sprache nicht immer ausreicht. Es gab dann No SQL Bewegungen. Es gibt Sprachen von XQuery über JAQL über PIG und Hive sowie neuere Sprachen, wie die im Rahmen des Stratosphere-Cloud-Informations-Management-Systems an der TU Berlin entwickelte Programmiermodell PACT und daraus abgeleitete Sprachen wie Meteor. Ich möchte hier nicht in Details gehen, aber es besteht sehr viel Forschungsbedarf an automatischer Optimierung und Parallelisierung von Sprachen zur Datenanalyse, insbesondere, um problemadäquat Informationsextraktionen und –integrationsoperationen mit Verfahren der statistischen Analyse, des Data Mining und anderen Operationen zu kombinieren.

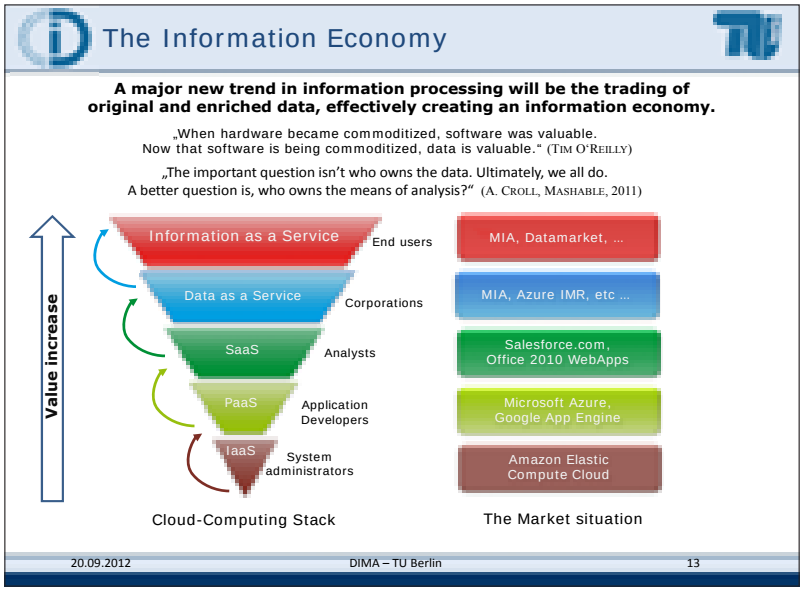

Bild 11

Ich möchte noch auf ein paar weitere Aspekte und die Entstehung der Informationswirtschaft, die auch vorhin schon angeklungen sind, eingehen (Bild 11). Vorher hatten wir auch schon ein paar interessante Zitate gehört, denen ich noch ein paar hinzufügen möchte. Ein großer Trend in der Informationsverarbeitung wird es sein, dass man in Zukunft sowohl angereicherte als auch originäre Daten behandeln wird. Ein sehr schönes Beispiel hierzu ist eine Aussage von Tim O'Reilly, einem der Protagonisten des Web 2.0: „Als Hardware Massenware wurde, war die Software wertvoll. Wir erleben jetzt immer mehr Open Source Bestrebungen. Software wird immer mehr zur Massenware. Jetzt werden die Daten wertvoll.". Daher ist es sehr wichtig, sich darauf zu konzentrieren, was für Daten wir haben und wie wir mit diesen Daten Mehrwert generieren können. Die wichtige Frage dabei ist aber nicht unbedingt, wer die Daten besitzt, was vielleicht ein wichtiger Aspekt ist, sondern wer die Daten analysieren kann. Das wird die große Fragestellung sein. Das heißt, im Cloud Computing bewegen wir uns immer mehr in die Ära von Daten als Dienst oder wirklich wertvolle Informationen als Dienst, im Gegensatz zu Infrastruktur oder Plattform als Dienst.

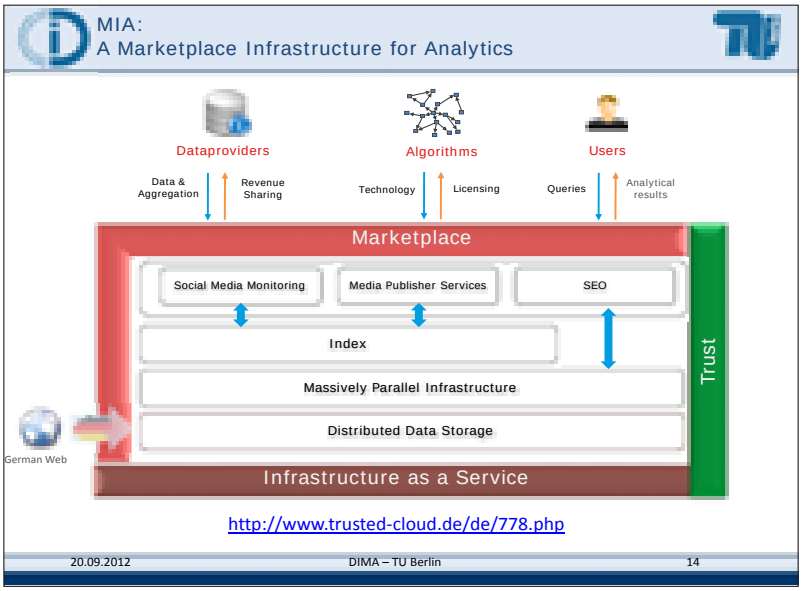

Bild 12

Zwei Beispiele von Forschungsprojekten, die wir am Fachgebiet Datenbanksysteme und Informationsmanagement der TU Berlin zusammen mit Partnern durchführen, möchte ich Ihnen noch kurz aufzeigen. Eines davon ist ein Projekt, das im Rahmen der Trusted-Cloud Initiative des Bundeswirtschaftsministeriums ausgezeichnet wurde, das MIA Projekt (http://www.mia-marktplatz.de/) (Bild 12). In diesem Projekt bauen wir einen Informationsmarktplatz für Daten und Analysen, wo Personen Daten und Analyseverfahren bereitstellen und auswerten können.

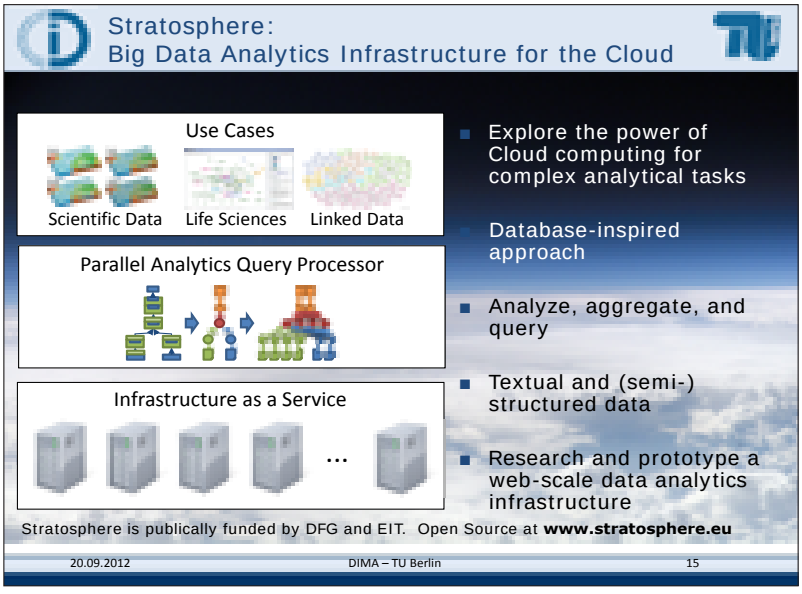

Bild 13

Ein zweites Beispiel (Bild 13) ist das Stratosphere Projekt (www.stratosphere.eu), eine Infrastruktur für Big Data Analytics, ein massiv paralleles Datenverarbeitungssystem, das in einem virtualisierten Cluster oder Clouds von Tausenden von Rechnern komplexe Datenanalysen durchführen kann, mit Informationsextraktion und –integration, wie oben in der Vision der nächsten Generation der Business Intelligence skizziert.

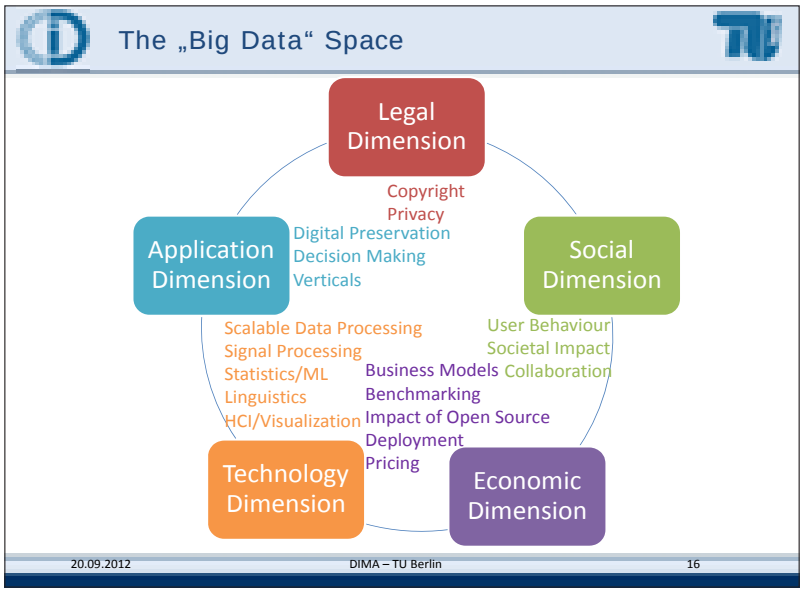

Bild 14

Ganz kurz der Rahmen, in dem sich „Big Data“ bewegt, aus wissenschaftlicher Sicht (Bild 14). Neben großen technischen Herausforderungen und spannenden Forschungsfragen gibt es auch rechtliche Fragestellungen, die hier auftreten, Anwendungsfragestellungen, soziale Fragestellungen, und viele wirtschaftliche Fragestellungen. Mein Vorredner hatte in diesem Zusammenhang schon über Benchmarking, Open Source, Deployment und Pricing gesprochen. Daneben gibt es natürlich spannenden Fragen im Hinblick auf Geschäftsmodelle für Big Data. An der TU Berlin betreue ich einige Gründerteams und Startups, die das Ziel haben, aus Big Data einen wirtschaftlichen Mehrwert und erfolgreiche Unternehmen zu schaffen.

Bild 15

Zu guter Letzt noch ein Call to Action (Bild 15). Ich glaube, dass es sehr wichtig wird, was auch MacKinsey in der Studie vorhergesagt hat, die erforderlichen Talente auszubilden. Das ist ein wesentlicher Aspekt für die Hochschulen. Wir müssen Datenwissenschaftler ausbilden, die tiefe Kenntnisse in bestimmten Datenanalysetechniken aber auch Anwendungskenntnisse haben, die so genannten „T-Shaped Students“, welche mit dieser großen Datenflut in einer immer mehr vernetzten, komplexeren Welt umgehen können. Dafür ist es wichtig, Curricula für Datenanalyse bereitzustellen und auch zu leben.

Daneben ist es im Bereich der Forschung wichtig, diese Datenanalysetechnologien zu untersuchen. Datenanalyse bedeutet im Wesentlichen, Datenmanagement, Data Mining, aber auch Programmiermodelle für parallele Verarbeitung und wie man das abstrahieren, vereinfachen und optimieren kann. Daneben sind als Fundament Statistik und Maschinenlernverfahren wichtig, sowie Methoden der Informationsvisualisierung. Und natürlich brauchen wir Innovation, um unsere Systeme und Anwendungen auch zum Markt tragen zu können, was bedeutet, dass wir einige prominente „Big Data“ Anwendungsfälle demonstrieren müssen, damit immer mehr Menschen und immer mehr Unternehmen gerade hier in Deutschland in die Datenanalyse einsteigen, Start-ups fördern, Technologietransfer insbesondere in kleinere und mittelständische Unternehmen leisten und die rechtlichen Fragestellungen klären.

Damit hoffe ich, dass wir in Deutschland eine Führerschaft in „Big Data“ erreichen können. Momentan haben wir da noch Nachholbedarf, aber Information als Rohstoff für eine Volkswirtschaft ist so bedeutsam, dass wir in diesem Bereich massiv investieren müssen, um unsere Wettbewerbsfähigkeit zu erhalten.

2 Wem gehören die Daten und wer hat außerdem Rechte daran?

Dr. Alexander Duisberg, Bird & Bird LLP, München

1. Einleitung

Die Erschließung des wirtschaftlichen Potenzials, das in der Aufbereitung, Nutzung und Verwertung von Big Data liegt, stellt unmittelbar die Frage nach Eigentum und Verfügungsfreiheit über einzelne Datensätze sowie über strukturierte und unstrukturierte Datensammlungen. Denn soweit Big Data in absehbarer Zeit immer stärker zum Wirtschaftsgut wird, ist die Frage nach dem rechtlichen Schutz und der Verkehrsfähigkeit von – noch zu definierendem oder näher einzugrenzendem – Big Data die wesentliche Voraussetzung für den Aufbau von Wertschöpfungsketten und eines entsprechenden Wirtschaftskreislaufs von (Waren und) Dienstleistungen. Damit rückt die Frage, wem die Daten bzw. Datensammlungen „gehören", also die Frage nach dem Eigentum in den Vordergrund. Bemerkenswerterweise ist diese Frage bisher weitgehend weder gestellt noch erörtert worden, obwohl in hohem Maße die Werthaltigkeit großer Datensammlungen gesehen wird. Im Folgenden sollen einige rechtliche Eckpfeiler gesetzt werden, um Erstellern und Nutzern von Big Data Orientierung und Anregung für die weitere Handhabe und Diskussion zu geben.

Bild 1

2. Was ist Big Data?

Unter „Big Data“ (Bild 1) lässt sich phänomenologisch eine Vielzahl von großvolumigen Sammlungen strukturierter und unstrukturierter Daten in digitalisierter Form fassen, die im öffentlichen und privaten Bereich im Zuge von IKT-basierten Verarbeitungsvorgängen anfallen, wie z.B. Telekommunikationsverkehrsdaten, Weblogs, RFID- und sonstige Sensordaten, Fahrzeug-, Schiffs- und Flugbewegungsdaten, Energiemessdaten, Geo- und Ortungsdaten, Wetter- und Umweltdaten, Finanztransaktionsdaten, statistische Verkaufsdaten, aggregierte Schlüsselwortanalysen etc. pp. Einige der im IKT-Umfeld gängigen Definitionen zum Big Data stellen dabei insbesondere auf die Größe und funktionale Handhabe ab[1], andere betonen die Wachstumsdynamik und Herkunftsquellen[2].

Für die rechtliche Betrachtung ist maßgeblich, dass es sich um eine strukturierte oder unstrukturierte Gesamtheit von digitalisierten Einzeldatensätzen handelt, die in der einen oder anderen Form einer IKT-basierten Verarbeitung ausgesetzt sind. Die eigentliche Größe ist zunächst unproblematisch, vielmehr sind zwei Ansatzpunkte maßgeblich: die rechtliche Einordnung und Verfügungsfreiheit über den einzelnen Dateneintrag bzw. Datensatz, sowie die Unterscheidung zwischen strukturierten und unstrukturierten Datensammlungen. Zugleich wird aus den vorstehend genannten Beispielen deutlich, dass „Big Data“ allenfalls in Ausschnitten und Teilbereichen auch datenschutzrechtliche Relevanz haben kann, mithin der Datenschutz in der Diskussion um Big Data lediglich eine Facette darstellt und in vielen Fällen unerheblich ist.

---

1 “‘Big data‘ refers to datasets whose size is beyond the ability to typical database software tools to capture, store, manage, and analyze.
This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data…“ *(McKinsey & Company,* “Big data: The next frontier for innovation, competition and productivity” (Juni 2011)).

2 “As the amount of data continues to grow exponentially, compounded by the internet, social media, cloud computing and mobile devices, it poses both a challenge and an opportunity for organisations – how to manage, analyse and make use of the ever-increasing amount of data being generated.” *(CEBR (Centre for Economic and Business Research),* “*Data equity: Unlocking the value of big data*“ (April 2011)).

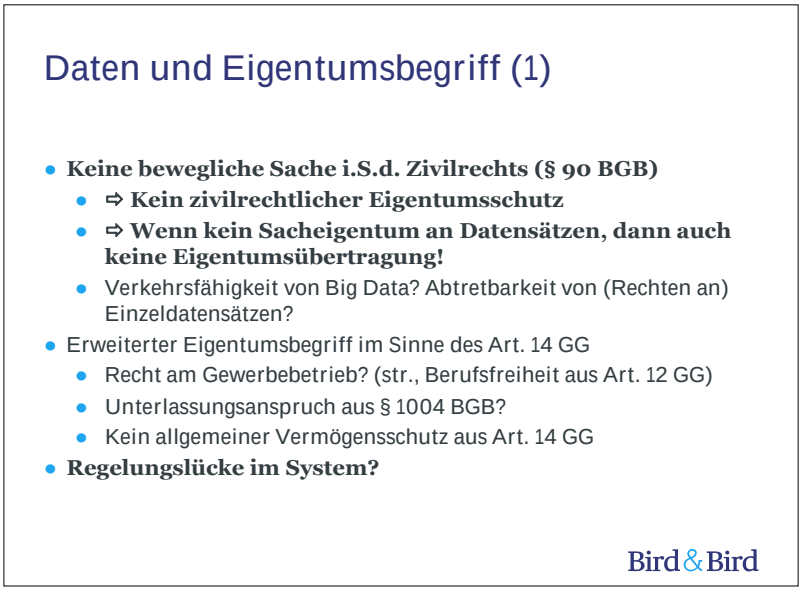

Bild 2

3. Kein zivilrechtliches Eigentum an Daten und Datensätzen

Bemerkenswerterweise ordnet unser Zivilrecht – das im Kern auf dem Bürgerlichen Gesetzbuch von 1900 beruht – Daten keiner rechtlich relevanten Kategorie zu. Der Sachbegriff für körperliche Gegenstände scheidet aus (§ 90 BGB)(Bild 2). Daraus folgt, dass kein zivilrechtlicher Eigentumsschutz für einzelne Daten bzw. Datensätze besteht und diese im Sinne des Sachenrechts auch nicht selbständig übertragbar sind. Dies ist rechtlich ebenso selbstverständlich wie klar, steht jedoch in der Praxis oft in Kontrast zum umgangssprachlichen Verständnis und der Zirkulation von Daten im Wirtschaftsverkehr. Ohne Klarheit über Eigentum oder eigentumsähnliche Ausschließlichkeitsrechte bleibt die Verkehrsfähigkeit von Big Data als Wirtschaftsgut aus rechtlicher Sicht begrenzt.

Ob es sich in unserer zivilrechtlichen Eigentumsordnung um eine planwidrige, ergänzungsbedürftige Regelungslücke handelt, ist nicht mit leichter Hand zu beantworten. Die Verfasser des BGB konnten nicht einmal ahnen, dass 100 Jahre später die Existenz digitalisierter Daten rechtlich diskutiert und bewertet werden würde. Ob diese Lücke allerdings in unserer Zivilrechtsordnung durch eine Anerkennung eigentumsähnlicher Ausschließlichkeitsrechte für jedes Datum und jeden einzelnen Datensatz tatsächlich geschlossen werden kann, ist angesichts der uferlosen Vielzahl und dem exponentiellen Anwachsen von digitalisierten Daten schon in praktischer Hinsicht fraglich.

Gleichwohl bleibt die Frage bestehen, „wem die Daten gehören“ bzw. wie in sinnvoller Weise die wirtschaftliche Verfügung über Datensammlungen gesichert werden kann. Ein maßgeblicher Teil der Antwort darauf ergibt sich aus dem Recht des Datenbankherstellers (siehe dazu unter Ziffer 5). Darüber hinaus – insbesondere soweit die Voraussetzungen für das sui generis Recht des Datenbankherstellers nicht vorliegen – besteht aber im unternehmerischen Bereich in einer Vielzahl von Konstellationen Bedarf, unstrukturierte oder in sonstiger Weise im Unternehmen vorhandene digitalisierte Daten rechtlich zu bewerten und zu schützen. Spätestens im Rahmen von Unternehmenstransaktionen, in denen der digitale Datenbestand

eines Unternehmens zu bewerten ist, stellt sich die Frage, ob es sich dabei um Geschäfts- und Betriebsgeheimnisse (im Sinne des § 17 UWG) bzw. Bestandteile des unternehmerischen Goodwill handelt. Denkbar ist auch, dass bestimmte Datenvorkommen in einem Unternehmen als wesentliches Asset anzusehen sind und als Bestandteil des eingerichteten Gewerbebetriebs im Sinne des Artikel 14 Grundgesetz[3] bzw. gemäß § 823 BGB und analog § 1004 BGB Schutz vor Eingriffen durch Dritte genießen. Ohne die komplexen verfassungsrechtlichen Fragen zu vertiefen, könnte ein grundrechtlicher Eigentumsschutz auch für solche Datenvorkommen bestehen, die ggf. nicht dem Schutz des Datenbankherstellers unterliegen. Gerade bei besonders großen Datenvorkommen mag sich diese Schlussfolgerung aufdrängen – wobei wiederum in der Abgrenzung klar sein müsste, dass der Grundrechtsschutz aus Artikel 14 GG nicht den allgemeinen Vermögensschutz bezwecken kann.[4] Hier steht die Diskussion erst am Anfang und es bleibt abzuwarten, in welchem Umfang die Gerichte und insbesondere das Bundesverfassungsgericht Anlass haben werden, sich mit dieser Frage zu befassen.

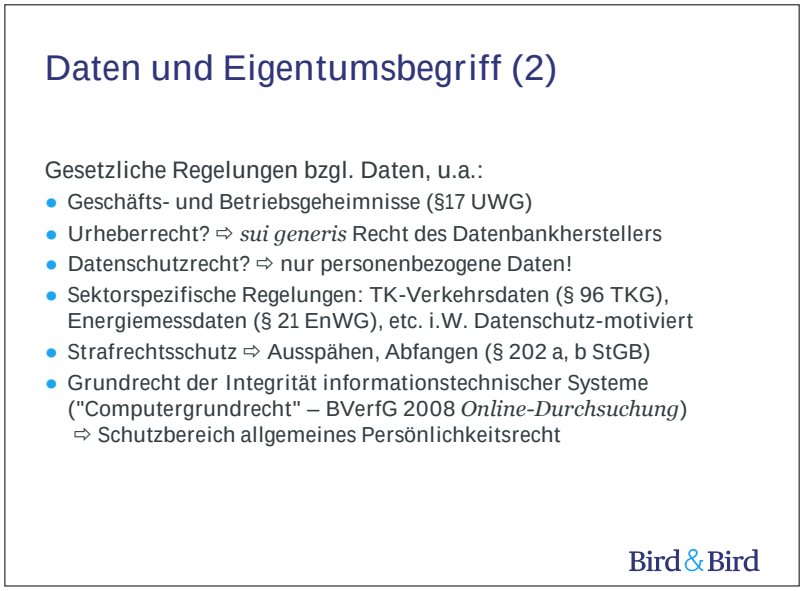

Bild 3

4. Anderweitige gesetzliche Regelungen zum Schutz von Big Data

Der Umstand, dass kein unmittelbarer zivilrechtlicher Eigentumsschutz an Big Data besteht, bedeutet keineswegs, dass einzelne Datensätze und Big Data in anderen gesetzlichen Regelungszusammenhängen unbeachtlich blieben (Bild 3).

[3] Der eingerichtete Gewerbebetrieb fällt nach überwiegender Auffassung unter den grundrechtlichen Eigentumsbegriff des Artikel 14 GGM vgl. Maunz/Papier, Kommentar zu Grundgesetz, Artikel 14 GG Rz. 95ff. m.w.N.; zu den Überschneidungen mit dem Grundrechtsschutz der Berufsfreiheit nach Artikel 12 GG, vgl. Maun/Scholz, a.a.O., Artikel 12 GG Rz. 130ff., 146ff, m.w.N.

[4] Vgl. Maunz/Papier, a.a.O., Artikel 14 GG Rz. 160 m.w.N.

4.1 Betriebs- und Geschäftsgeheimnisse

Auf die Bedeutung des Schutzes im Rahmen betrieblicher Geschäfts- und Betriebsgeheimnisse wurde bereits hingewiesen. § 17 UWG dient als maßgebliche Schutznorm gegen unberechtigte Preisgabe und entsprechende Eingriffe durch Dritte. Sie begründet entsprechende Unterlassungsansprüche im Rahmen der Wettbewerbsordnung und untermauert als Strafnorm den hohen Schutz für Geschäfts- und Betriebsgeheimnisse. Allerdings ist ebenso klar, dass § 17 UWG selbst nicht die Verkehrsfähigkeit von Daten oder Datensammlungen normiert.

4.2 Recht des Datenbankherstellers und allgemeines Urheberrecht

Zentrale Bedeutung gewinnt mit Blick auf die Verkehrsfähigkeit, wie erwähnt, das Recht des Datenbankherstellers (siehe unter Ziffer 5).

Das allgemeine Urheberrecht kann hingegen nur dann behilflich sein, wenn es sich bei den Datensätzen um urheberrechtlich geschützte Werke handelt. Dies ist in ausgewählten Fällen durchaus denkbar,[5] regelmäßig und in einer Vielzahl der Erscheinungsformen von Big Data aber nicht der Fall. Im Ergebnis ist festzuhalten, dass die in der Praxis anzutreffenden Bezugnahmen auf das Urheberrecht und entsprechende Lizenzierungsmodelle fehlgehen können und dann keine Nutzungsberechtigung an Big Data begründen.

4.3 Datenschutz und Datensicherheit

Erhebliche Bedeutung hat natürlich der Datenschutz – allerdings immer nur soweit, als auch personenbezogene Daten betroffen sind (siehe dazu unter Ziffer 6). Besonders zu beachten sind sektorenspezifische Regelungen, wie etwa in der Telekommunikation oder dem Energiesektor (siehe dazu unter Ziffer 9). Der Gesetzgeber setzt zum Teil enge Grenzen für die Speicherung und Nutzbarkeit der großen Datenmengen, die bei den beteiligten Unternehmen anfallen. Hier besteht naturgemäß ein Spannungsverhältnis zwischen der großen Reichweite des Datenschutzes (der beispielsweise auch IP-Adressen erfasst oder erfassen kann)[6] einerseits und den Innovationspotenzialen andererseits, die aus der Nutzung solcher Datenbestände erwachsen.

In den regulierten Industrien werden sich die Innovationspotenziale zum Teil erst durch Anpassungen des regulatorischen Rahmens erzielen lassen, soweit nicht im Wege der Datenanonymisierung abgeholfen werden kann.[7] Im allgemeinen Datenschutzrecht lassen sich einige Konstellationen nach den Grundsätzen der Interessenabwägung nach § 28 Absatz 1 Satz 1 Nr. 2 BDSG lösen, was aber jeweils einer sorgfältigen Prüfung im Einzelfall bedarf.

In diesem Zusammenhang ist auch auf die Regelung zum Umgang mit Datensicherheitspannen hinzuweisen, die – soweit es um personenbezogene Daten geht – weitreichende

---

[5] Ein Bespiel ist insoweit die Veröffentlichung von Satellitenaufnahmen, die die Deutschen Gesellschaft für Luft- und Raumfahrt unter den Bedingungen der creative commons license der Allgemeinheit zur Verfügung gestellt hat (siehe www.dlr.de, www.heise.de vom 1. März 2012).

[6] Vgl. EuGH Urteil vom 24. November 2011, Az. C-7/10, Rz. 51; vgl. zum Diskussionstand *Hoeren/Sieber*, Multimedia-Recht, 30. Ergänzungslieferung 2011, Rz. 82-84; *Spindler/Schuster*, Recht der elektronischen Medien, 2. Auflage 2011 Rz. 8; jeweils m.w.N.; siehe auch Begründungserwägung (24) und Artikel 4 Absatz 1 des Entwurfs der Europäischen Datenschutzverordnung vom 25. Januar 2012 KOM 2012/0011.

[7] Bei Anonymisierung von Daten entfällt die Anwendbarkeit des Datenschutzrechts, § 3 Absatz 6 BDSG; vgl. auch weiter unten bei Ziffer 6.1.

Mitteilungs- und Benachrichtigungspflichten der verantwortlichen Stelle auslösen (siehe § 42 a BDSG, § 109 a TKG, § 15 a TMG, § 83 a SGB X).

4.4 Strafrechtsschutz

Die Ansammlung von Daten bis hin zum Big Data genießt vor Eingriffen Dritter auch im allgemeinen Strafrecht Schutz, insbesondere durch die Tatbestände des Ausspähens und Abfangens von Daten (§§ 202 a, b StGB). Diese Strafnormen schützen vor Hacking und computer-basierten Eingriffen in Datenvorkommen und sichern damit den Bestand von Big Data, insbesondere soweit es in proprietären Unternehmensstrukturen gehalten wird – unabhängig davon, ob es sich dabei der Struktur nach um Datenbanken oder inhaltlich um Geschäfts- oder Betriebsgeheimnisse (§ 17 UWG) handelt.

4.5 Computer-Grundrecht

Das sog. „Computer-Grundrecht" („Grundrecht auf Gewährleistung der Vertraulichkeit und Integrität informationstechnischer Systeme"), das das Bundesverfassungsgericht im Zusammenhang mit Online-Durchsuchungen aus dem allgemeinen Persönlichkeitsrecht entwickelt hat, soll nur kurz erwähnt werden. Als individuelles Grundrecht schützt es den Einzelnen (ggf. auch als Nutzer von Big Data) vor hoheitlichen Eingriffen und kann entsprechend nicht von Unternehmen beansprucht werden.

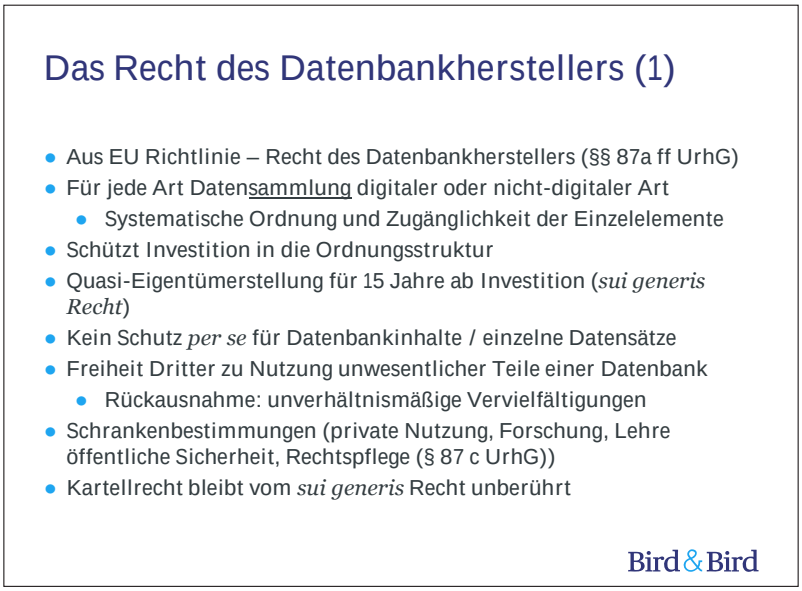

Bild 4

5. Das Recht des Datenbankherstellers (§ 87 a UrhG)

5.1 *sui generis* Recht zum Schutz der Investition

Wie bereits erwähnt nimmt das Recht des Datenbankherstellers (§ 87 a-e UrhG) eine zentrale Rolle hinsichtlich des Schutzes und der Verwertung von Big Data ein (Bild 5). Das Recht des Datenbankherstellers setzt die EU Datenbank-Richtlinie von 1996[8] um. Es gewährt dem Hersteller einer Datenbank auf 15 Jahre begrenzt ausschließliche Nutzungs- und Verwertungsrechte in einer Quasi-Eigentümstellung (sog. *„sui generis"* Schutz). Ziel der gesetzlichen Regelung ist es, die Investition in die Herstellung einer Datenbank zu schützen und eine entsprechende wirtschaftliche Verwertung zu ermöglichen.

Der Inhalt der Datenbank selbst – also die einzelnen Datensätze – erwirbt keinen gesonderten Rechtsschutz aufgrund des § 87 UrhG. Insoweit kommt es darauf an, ob diese Inhalte bzw. Daten eines eigenständigen Rechtsschutzes fähig sind (siehe oben bei Ziffern 3, 4.2). Die Schutzwirkung entsteht mit Herstellung der Datenbank und erfordert keine weitere förmliche Registrierung oder dergleichen. Die wesentliche Änderung einer Datenbank gilt als neue Datenbank und begründet einen neuen, 15 Jahre laufenden *sui generis* Schutz.

Das Gesetz definiert als Datenbank *„… eine Sammlung von Werken, Daten oder anderen unabhängigen Elementen, die systematisch oder methodisch angeordnet und einzeln mit Hilfe elektronischer Mittel oder auf andere Weise zugänglich sind und deren Beschaffung, Überprüfung oder Darstellung eine nach Art oder Umfang wesentliche Investition erfordert."*
Es kommt also im Kern auf eine systematische oder methodische Anordnung von Daten oder „anderen unabhängigen Elementen" an, wobei dafür eine wesentliche Investition erforderlich ist. Eine Vielzahl von proprietären Datenvorkommen ist in der Tat nach Ordnungsprinzipien strukturiert und durch einen entsprechenden Investitionsaufwand getragen. Allerdings hat der Europäische Gerichtshof den sachlichen Anwendungsbereich des *sui generis* Schutzes erheblich eingeschränkt. Investitionen, die in die Erzeugung oder Erhebung von Daten getroffen werden bzw. dort anfallen und dem Aufbau einer Datenbank vorgeschaltet sind, gelten nicht als Investition in die Herstellung der Datenbank selbst.[9]

5.2 Bewertung des Einzelfalls

Damit ergeben sich in einer Vielzahl von Big Data Konstellationen Abgrenzungsfragen, die eine Bewertung im Einzelfall unverzichtbar machen. Sowohl bei der Nutzung von Big Data, das aus öffentlichen Netzen oder anderen frei zugänglichen Quellen bezogen und dann von Unternehmen nach entsprechender Aufbereitung wirtschaftlich verwertet wird, als auch im Rahmen proprietärer Datenvorkommen wird es sehr genau darauf ankommen, die in die Ordnungsstruktur getroffene Investition zu bestimmen und zu dokumentieren, um – über den bloßen Vorgang der Datenerhebung hinaus – den *sui generis* Schutz zu erlangen.

Nur dann kann der Datenbankhersteller seine Datenbank(en) rechtlich abgesichert lizenzieren oder ggf. auch vollständig übertragen, und damit seine Investition amortisieren. Es ist absehbar, dass sich in Zukunft – ob im Rahmen von Datenbanklizenzen, Datenbanktransfers

---

[8] Richtlinie 96/9/EG vom 11. März 1996 über den rechtlichen Schutz von Datenbanken, ABl. L 77, S. 20ff.

[9] Vgl. EuGH vom 9. November 2004 Az. C-203/02, C-338/02, 444/02, *BHB Pferdewetten, / Fixtures-Fussballspielpläne I—II; siehe auch Schricker/Vogel*, Kommentar zum Urheberrecht, 4. Aufl., § 87 a UrhG Rz. 52ff. m.w.N.; *Wandtke/Bullinger/Thum*, Praxiskommentar zum Urheberrecht. 3. Aufl. § 87 a UrhG, Rz. 41, jeweils m.w.N.

oder Unternehmenskäufen – eine Vielzahl von Zweifelsfragen und streitige Auseinandersetzungen dazu ergeben werden.

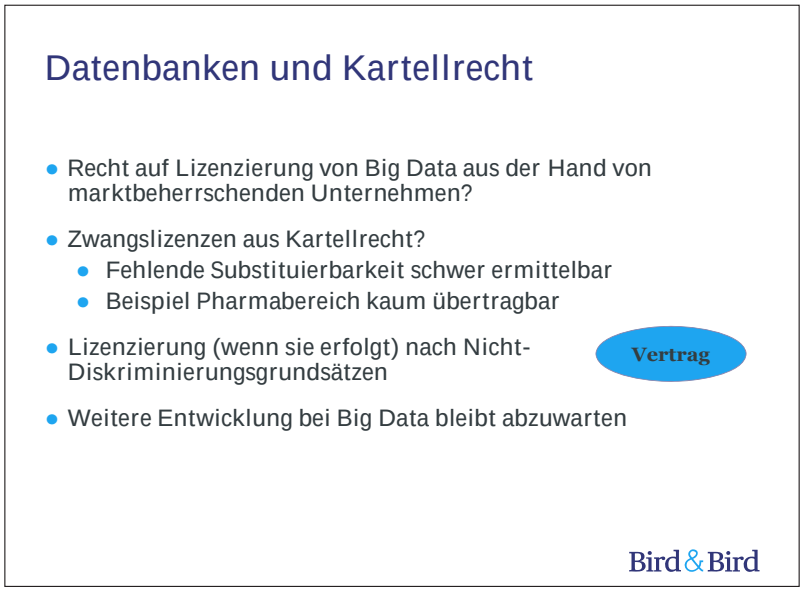

Bild 5

5.3 Wer ist Datenbankhersteller?

Bei komplexen, mehrstufigen Erstellungs- und Verwertungsprozessen rund um Big Data kann sich die Frage stellen, wer die maßgebliche Investition tätigt und wer sich ggf. bestimmter Zulieferleistungen (sowohl hinsichtlich der Datensätze selbst, als auch der zugrunde liegenden Infrastruktur- und Verwertungsleistungen) bedient (Bild 5). Insbesondere wenn es bei Erstellung einer Datenbank in der Zulieferkette zu Transformationen, Migrationen, Rekonfigurationen und Strukturierungen der betroffenen Datensätze kommt, sind vertragliche Regelungen unverzichtbar.

Andernfalls bleibt unklar und einer (ggf. streitigen) Auslegung der tatsächlichen Umstände überlassen, wer nach dem Willen der Parteien die maßgeblichen Investitionen getätigt hat: das Unternehmen, das unter Bezug von Zulieferleistungen eine Datenbank erstellt, oder der Zulieferer, der eine eigene (neue) Datenbank seinem Kunden überlässt.

5.4 Bedeutung für Cloud Computing

Besonders deutlich wird dies in bestimmten Konstellationen des Cloud Computing. Wenn der Cloud-Dienstleister an den Datensätzen des Kunden bestimmte Angleichungen und Konfigurationen vornehmen muss, um sie in sein standardisiertes Verarbeitungsmodell zu überführen, so kann sich bei Vertragsende die Frage stellen, ob der Cloud-Kunde eine neue Datenbank zurückerhält und ihm daran die entsprechenden Rechte des Datenbankherstellers zustehen oder nicht.

Hier sollten im Regelfall klare vertragliche Regelungen dafür sorgen, dass die Rechte zur Nutzung und Verwertung von Big Data bei demjenigen liegen, der nach dem wirtschaftlichen Verständnis der Parteien tatsächlich die maßgeblichen Rechte daran haben und das Amortisationsrisiko tragen soll. Dies ist auch deswegen besonders wichtig, weil im Rahmen

der §§ 87 a ff. UrhG keine gesetzliche Auslegungshilfe wie etwa die Zweckübertragungslehre des allgemeinen Urheberrechts vorgesehen ist.[10]

5.5 Insolvenz auf Dienstleisterseite

Für den Fall der Insolvenz – gerade auch im Falle einer Insolvenz des Cloud-Provider – wird man damit rechnen müssen, dass vertragliche Abreden über die Eigenschaft des Datenbankherstellers als schuldrechtliche Regelungen zwischen den Parteien dem Wahlrecht des Insolvenzverwalters gemäß § 103 InsO unterliegen. Danach könnte der Insolvenzverwalter, der an die Stelle des insolventen Dienstleisters tritt, frei entscheiden, ob es im Interesse des Gläubigerschutzes vorteilhafter ist, die Datenbankherstellereigenschaft beim Kunden bzw. Auftraggeber zu belassen und einen entsprechenden Dienstleistungsvertrag fortzuführen – oder auf die tatsächliche Feststellung der Datenbankherstellereigenschaft abstellen, die möglicherweise dazu führt, dass der insolvente Dienstleister eigene Rechte an der Datenbank hat und diese im Rahmen der Insolvenz frei verwerten kann. Hier bleibt abzuwarten, wie die insolvenzrechtliche Rechtsprechung in Zukunft entscheiden wird.

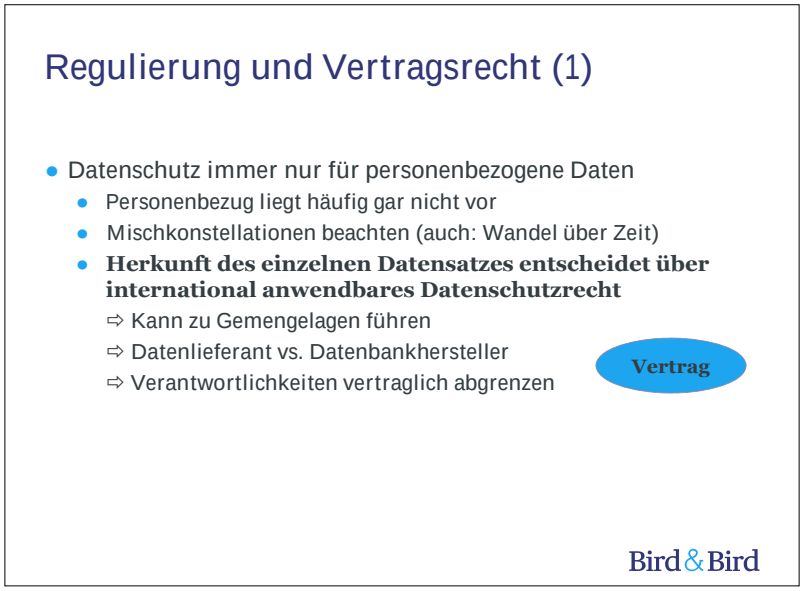

Bild 6

[10] Gemäß § 31 Absatz 5 UrhG bestimmt sich der Umfang der eingeräumten Nutzungsrechte nach dem von beiden Parteien zugrunde gelegten Vertragszweck. Dieser Grundsatz ist aber nicht ohne Weiteres auf §§ 87 a ff. UrhG und die Frage anzuwenden, wer unter zwei Parteien als der Investor und damit Hersteller einer Datenbank anzusehen ist. Siehe allgemein *Schricker*, Kommentar zum Urheberrecht, 4. Aufl., § 31 UrhG Rz. 64, 74 ff.

6. Datenschutz

6.1 Personenbezug und Anonymisierung

Wie bereit erwähnt (siehe oben Ziffer 4.3), kommt dem Datenschutz im Rahmen von Big Data hohe Bedeutung zu (Bild 6). Allerdings ist er immer nur dann einschlägig, wenn Big Data tatsächlich auch die Verarbeitung von personenbezogenen Daten erfasst. Gemäß § 3 Absatz 1 BDSG sind personenbezogene Daten *„Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person“*.[11] Die Reichweite ist groß, was unter die Personenbestimmbarkeit fällt. Die EU Datenschutzrichtlinie 95/46/EG führt dazu weiter aus: *„Als bestimmbar wird eine Person angesehen, die direkt oder indirekt identifiziert werden kann, insbesondere durch Zuordnung zu einer Kennnummer oder zu einem oder mehreren spezifischen Elementen, die Ausdruck ihrer physischen, physiologischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität sind.“* [12]

Zwar kann in einer Vielzahl von Big Data Konstellationen die Bestimmbarkeit einer natürlichen Person gegeben sein – noch dazu, wenn sich über Zeit technische Möglichkeiten auftun, aus zunächst unscheinbaren Angaben Rückschlüsse auf die Identität von Einzelpersonen zu ziehen.[13] Es ist aber auch klar, dass der Personenbezug dann nicht gegeben ist, wenn zuvor personenbezogene Daten vor einer Verarbeitung vollständig anonymisiert werden bzw. der Rückschluss auf das Individuum nicht oder nur mit unverhältnismäßig hohem Aufwand möglich ist.[14] Reine Unternehmensinformationen sind dagegen – entgegen verbreiteter laienhafter Einschätzung – in keiner Weise datenschutzrechtlich relevant.

Daraus ergibt sich zum einen, dass in einer Vielzahl von Big Data Konstellationen jeder Personenbezug hinsichtlich der verarbeiteten Datensätze entfällt und das Datenschutzrecht damit nicht anwendbar ist. Zum anderen kann es aber ebenso gut sein, dass Big Data in Mischkonstellationen sowohl personenbezogene als auch nicht personenbezogene Daten enthält bzw. sich aufgrund dynamischer Technologieentwicklung über Zeit der Personenbezug später doch ohne unverhältnismäßigen Aufwand herstellen lässt. Insoweit ist es im Einzelfall unverzichtbar, die in Rede stehenden Datenkategorien von Big Data darauf zu überprüfen, ob sie Personenbezug haben oder über Zeit haben könnten.

6.2 Internationale Gemengelagen

Es tritt datenschutzrechtlich eine weitere Herausforderung hinzu: Die Frage, welches nationales Datenschutzrecht auf einzelne Datensätze Anwendung findet, richtet sich im Grundsatz

---

[11] Die Bestimmung setzt Artikel 2 a) der Europäischen Datenschutzrichtlinie 95/46/EG vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr Amtsblatt Nr. L 281 vom 23/11/1995, S. 31ff. um und hat in sämtlichen EU Staaten entsprechenden Eingang in die nationalen Datenschutzgesetze gefunden. Die Auslegung der nationalen Datenschutzbehörden, was als personbezogene Daten anzusehen ist, variiert. Die gemeinsame Arbeitsgruppe der nationalen Datenschutzbehörden, die sog. „Artikel 29 Gruppe“, hat es sich zur Aufgabe gesetzt, hier für eine Angleichung der Auslegung zu sorgen. Mit der anstehenden EU Datenschutzverordnung ist eine weitergehende Vereinheitlichung zu erhoffen (siehe Artikel 4 Absatz 1 der Entwurfsfassung vom 25. Januar 2012).

[12] Artikel 2 a) der EU Datenschutzrichtlinie 95/46/EG, a.a.O.

[13] Zur Frage der IP-Adressen siehe bereits oben unter Ziffer 4.3.

[14] Siehe § 3 Absatz 6 BDSG: *„Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.“*

nach dem Sitz der verantwortlichen Stelle.[15] Sofern große Datenvorkommen aus einer Zusammenstellung von Datenerhebungen bestehen, die in unterschiedlichen Ländern zuvor von unterschiedlichen verantwortlichen Stellen durchgeführt wurden, kann es also innerhalb derselben Datenbank zu datenschutzrechtlichen Gemengelagen kommen. Aufgrund der unterschiedlichen Auslegungen und Anforderungen, u.a. an die technischen Sicherungsmaßnahmen, kann dies im Extremfall zu einer Divergenz der Compliance-Anforderungen führen.

In der Praxis zeigt sich, dass die daraus resultierenden Abgrenzungsfragen und Risikoerwägungen in Zuliefer- und Kundenverhältnissen zu erheblichen Problemen und schwierigen Diskussionen um den angemessenen Ausgleich bzw. passende Haftungsfreistellungen führen. Auch hier sind sorgfältige vertragliche Regelungen von entscheidender Bedeutung, um streitigen Auseinandersetzungen bestmöglich vorzubeugen.

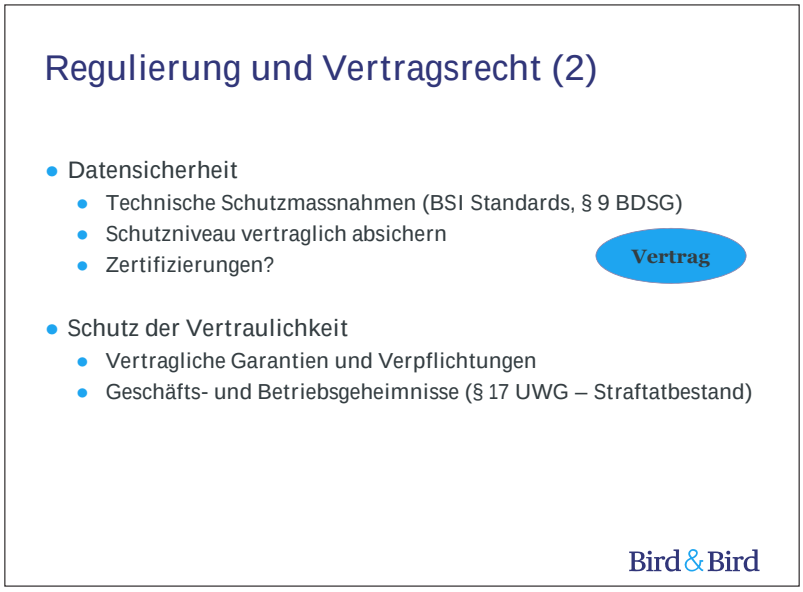

Bild 7

7. Datensicherheit und Schutz der Vertraulichkeit

Wie bereits erwähnt, sind die Anforderungen an die Datensicherheit im Umgang mit Big Data von elementarer Bedeutung. Regulatorisch folgt dies – soweit das Datenschutzrecht einschlägig ist – aus den Anforderungen an technische und organisatorische Maßnahmen gemäß § 9 BDSG (Bild 7). Sektorenspezifische Anforderungen – gleichermaßen für personenbezogene

[15] Gemäß der Kollisionsnorm des § 1 Absatz 5 BDSG gilt das BDSG nicht nur für Datenerhebungen, die von ausländischen verantwortlichen Stellen innerhalb Deutschlands durchgeführt wurden. Entsprechend gespiegelte Kollisionsregeln finden sich gemäß Artikel 4 der EU Datenschutzrichtlinie 95/46/EG in allen anderen EU Mitgliedstaaten.

und nicht personenbezogene Daten – können hinzutreten.[16] Auch die Anforderungen des Bundesamtes für Sicherheit in der Informationstechnik (BSI) an den IT Grundschutz sowie entsprechende technische Standards, einschließlich internationaler ISO Zertifizierungen Standards können einschlägig sein bzw. aufgrund entsprechender vertraglicher Vereinbarung Verpflichtungen begründen, das bestmögliche Schutzniveau im Umgang mit Big Data zu wahren.

Auf die Besonderheiten im Umgang mit Datensicherheitspannen in Auslagerungsfällen wurde bereits hingewiesen (siehe oben Ziffer 4.3). Hier sollte der Inhaber der Datenbank sicherstellen, dass von ihm beauftragte Dienstleister sämtlichen Unterrichtungspflichten, denen der Inhaber der Datenbank als verantwortliche Stelle im Sinne des Datenschutzrechts unterliegt, unverzüglich im Innenverhältnis nachkommen und die verantwortliche Stelle von etwaige Sanktionen freistellen, die aus einer unterlassenen oder verspäteten Befolgung von Compliance-Anforderungen folgen.

Im Übrigen sind selbstverständlich im Rahmen der Überlassung von Big Data die üblichen vertraglichen Schutzvorkehrungen – insbesondere mit Blick auf Geschäfts- und Betriebsgeheimnisse – durch pönalisierte Vertraulichkeitsvereinbarungen etc. zu treffen.

Bild 8

[16] Nur beispielhaft: Siehe § 109 TKG, oder für die Finanzdienstleister die MA Risk der BAFin, z.Zt in Überarbeitung, aus der eine Vielzahl von Anforderungen zur Mindestabsicherung technischer Ausfallsrisiken etc. hervorgehen (siehe Entwurf vom 26. April 2012 – http://www.bafin.de / SharedDocs/Veroeffentlichungen/DE/Konsultation/2012/kon_0112_ueberarbeitung_marisk_ba.html.).

8. Musterinformation des VDA über Datenspeicher im Fahrzeug

Ein bemerkenswertes Beispiel (Bild 8) zum rechtlichen Umgang mit relativ hohen Datenaufkommen stellt die Muster-Information des VDA und der Datenschutzbehörden über Datenspeicher im Fahrzeug vom April 2012 dar.[17] Diese Musterinformation soll jeweils in die Betriebsanleitungen neuer Fahrzeugen aufgenommen werden.

Ziel ist es, den Fahrzeughalter oder auch Fahrzeugnutzer darüber zu informieren, dass die Hersteller bzw. Vertragswerkstätten im Bedarfsfall Zugriff auf verschiedene technische, elektronisch aufgezeichnete Informationen über die Fahrzeugnutzung nehmen, die Auskunft über die technischen Zustände des Fahrzeugs geben (z.B. Füllstände von Systemkomponenten, Radumdrehungszahl, Geschwindigkeit, Querbeschleunigungen, Defektmeldungen an Licht und Bremsen, Auslösung von Airbags, Einsetzen der Stabilitätsregelungssysteme, Außentemperaturen). Während diese technische Informationen der Erkennung und Behebung von Fehlern und der Optimierung von Fahrzeugfunktionen dienen, können sie bei Werkstattaufenthalten und im Rahmen sonstiger Serviceleistungen durchaus auch Rückschlüsse auf die Fahrerpersonen ermöglichen (z.B. im Rahmen der Rekonstruktion von Unfallhergängen).

Es ist aus rechtlicher Sicht eigentlich offen, was mit dieser Musterinformation erreicht wird. Die Nutzung der technischen Daten ist im Regelfall – sofern ein Personenbezug auf den Halter oder Fahrzeugführer möglich ist – im Rahmen der Durchführung eines Wartungs- oder Reparaturvertrages gesetzlich gerechtfertigt[18]. Sobald diese Informationen mit anderen Informationen wie Unfallprotokollen, Unfallschäden, Zeugenaussagen verbunden werden können, stellt sich die Frage einer gesetzlichen Rechtfertigung neu. Ob die Nutzung der technischen Daten bzw. die Herstellung einer solchen Verknüpfung überhaupt zulässig ist, kann nicht allgemein beantwortet werden. Die Rechtfertigung aus § 28 Absatz 1 Satz 1 Nr. 1 BDSG dürfte regelmäßig nicht greifen und auch eine Rechtfertigung im Rahmen einer Interessenabwägung nach § 28 Absatz 1 Satz 1 Nr. 2 BDSG dürfte bei nachteiligen Folgen für den Betroffenen (z.B. Nachweis eines verkehrsregelwidrigen Verhaltens) entfallen. Entsprechend kann die Nutzbarkeit dieser Daten für solche, außerhalb einer Serviceleistung liegenden Zwecke – trotz der Muster-Information – im Ergebnis an datenschutzrechtlichen Erwägungen scheitern.

---

[17] Muster-Information über Datenspeicher im Fahrzeug, hrsg. vom VDA und den Datenschutzbehörden der Länder, vom 6. Februar 2012 (http://www.lda.bayern.de/lda/datenschutzaufsicht/lda_aktuell.htm).

[18] Gemäß § 28 Absatz 1 Satz 1 Nr. 1 BDSG sind Datenverarbeitungen zulässig, wenn sie *„für die Durchführung eines rechtsgeschäftlichen Schuldverhältnisses mit dem Betroffenen erforderlich“* sind.

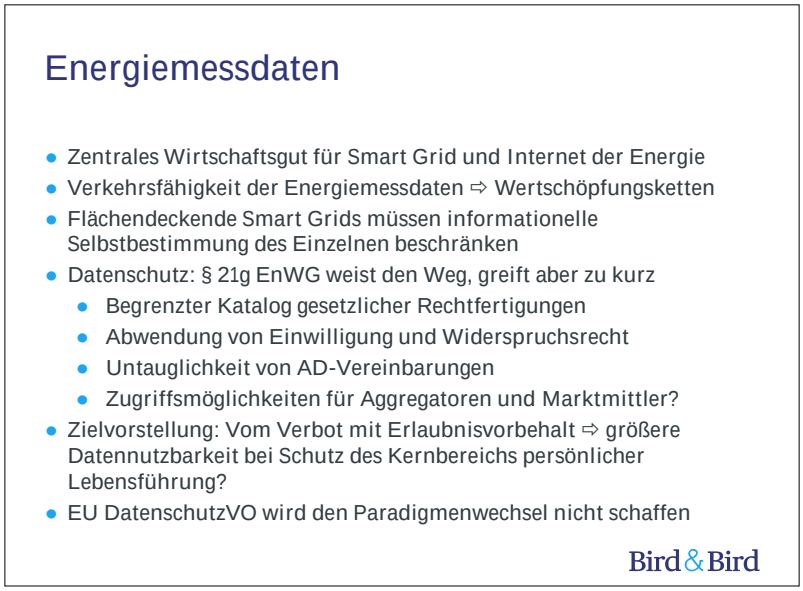

Bild 9

9. Energiemessdaten als Voraussetzung von Smart Grids

Die sich im Zuge der Energiewende abzeichnende Entwicklung von Smart Grids wirft eine Vielzahl von datenschutzrechtlichen Fragestellungen auf, die hier nicht im Einzelnen erörtert werden können.[19] In aller Kürze soll aber festgehalten werden, dass die hohen Volumina von Energiemessdaten, die mittels der Smart Meter anfallen werden, eine erhebliche datenschutzrechtliche Trageweite haben (Bild 9).

Mit § 21 g EnWG und den noch anstehenden Ausführungsverordnungen wird der Ansatz der datenschutzrechtlichen Spezialgesetzgebung verfolgt. Nach der derzeitigen gesetzlichen Regelung zeichnet sich ab, dass die Entwicklung des Internet der Energie und damit einhergehende mehrstufige, multi-direktionale Datenströme und Verarbeitungsvorgänge von Energiemessdaten datenschutzrechtlich mit dem herkömmlichen Instrument der Auftragsdatenverarbeitung adäquat abgebildet und abgesichert werden können. Zudem greift § 21 g EnWG in derzeitiger Fassung zu kurz, was die Entfaltung von Intermediären und die Entwicklungschancen peripherer Dienstleistungen im Internet der Energie betrifft.[20]

Der exponentielle Anstieg der Datenvolumina durch Smart Grids und im Rahmen des Internet der Energie bietet Anlass, den herkömmlichen datenschutzrechtlichen Ansatz des Verbots mit Erlaubnisvorbehalt in Frage zu stellen und neu zu bewerten. Die Umkehrung hin zu einer größeren oder sogar allgemeinen Freigabe der Datennutzung unter dem Vorbehalt und der Wahrung eines Schutzes des Kernbereichs der persönlichen Lebensführung wäre an dieser Stelle wegweisend. Die Europäische Datenschutzverordnung wird diesen – nach zu-

[19] Dazu näher Duisberg, „Datenschutz im Internet der Energie“, in: Peters/Kersten/Wolfenstetter (Hrsg.), „Innovativer Datenschutz“, 2012.

[20] Dazu Duisberg, a.a.O., m.w.N. zu Stellungnahmen in der Literatur.

nehmender Auffassung im juristischen Schrifttum – fälligen Paradigmenwechsel[21] nach dem Diskussionsstand zum derzeitigen Entwurf (noch) nicht schaffen.

Bild 10

10. Open Data und Big Data aus dem öffentlichen Sektor

Der Nutzung von Big Data aus öffentlich zugänglichen Quellen wird eine entscheidende Rolle in der Freisetzung von Innovation rund um Big Data zugemessen (Bild 10).

10.1 Zugriff auf Daten und Informationen im Internet

Dies betrifft zum einen die im Internet frei verfügbaren Datenbanken und sonstigen Informationsquellen. Gerade mit Blick auf die oben stehenden Ausführungen zu Datenbanken und dem Urheberrecht (siehe Ziffern 4.2 und 5) ist allerdings der Fehlvorstellung vorzubeugen, dass alle im Internet einsehbaren Daten und Informationen ohne Weiteres frei verwendbar und daraus einschränkungslos Auswertungen und eigene Datenbanken erstellt werden können.

Zum einen sind ggf. vorhandene Nutzungsbedingungen zu beachten; zum anderen kann es bei der Übernahme und Verwertung von Fremdinhalten oder ganzer Datenbanken zu Rechtsverletzungen kommen, selbst oder gerade wenn von dem Betreiber der Datenquelle keine Nutzungsbedingungen bereitgehalten werden. Auch in datenschutzrechtlicher Hinsicht ist nicht jedes personenbezogene Datum, das etwa auf sozialen Netzwerken einsehbar ist, für die Erstellung und Übernahme in eigene Datenbanken und zur Durchführung entsprechender Datenanalysen etc. ohne Weiteres frei verwendbar.[22] Auch hier ist eine Prüfung der Zugriffs-

[21] *Schneider*, „Hemmnis für einen modernen Datenschutz“, in: Anwaltsblatt, Heft 4, 2011, S. 233ff.; *Heckmann*, Smart Life - Smart Privacy Management, K&R 2011, S. 1ff.

[22] Ob allerdings das geplante und angesichts des öffentlichen Protests dann eingestellte Forschungsprojekt des Hasso-Plattner Instituts, bestimmte Analyseverfahren basierend auf Einträgen in Sozialen Netzwerken durchzuführen, tatsächlich wie behauptet mit dem Datenschutzrecht unvereinbar gewesen wäre, steht dahin und ist hier nicht abschließend zu beurteilen (http://www.sueddeutsche.de/digital/kritik-an-neuem-scoring-verfahren-facebook-projekt-der-schufa-abgeblasen-1.1377327).

und Nutzungsberechtigungen vor einer unternehmerischen Tätigkeit geboten, die auf solche Datensammlungen aufbaut.

10.2 Weiterverwendung von Daten aus dem öffentlichen Sektor

Besondere Bedeutung kommt darüber hinaus der Freigabe von Big Data durch öffentliche Stellen zu. Hier hat schon vor langer Zeit die EU Richtlinie über die Weiterverwendung von Informationen des öffentlichen Sektors („PSI-Richtlinie“)[23] die Grundlagen gelegt, die im Informationsweiterverwendungsgesetz („IWG“) in deutsches Recht umgesetzt wurden. Die Leitprinzipien sind dabei die Grundsätze der Transparenz, Nichtdiskriminierung, Kostendeckung (ohne Gewinnerzielung der öffentlichen Hand). Die Preisgabe öffentlicher Informationen und Daten steht unter diversen Vorbehalten (siehe § 1 IWG), zu denen u.a. der (in § 1 Absatz 1 IWG allgemein gesetzte) Vorbehalt der Abwägung mit Sicherheitsinteressen des Staates steht.

10.3 Weiterentwicklung der PSI-Richtlinie

Die Erfolge in der Preisgabe von Informationen und Daten des öffentlichen Sektors – wie etwa Statistiken, Wirtschafts- und Umweltdaten wie auch Archivinhalte und Sammlungen von Büchern und Kunstwerken – sind nach Einschätzung der Kommission bislang beschränkt und das damit verbundene Innovationspotenzial insbesondere für kleinere und mittlerer Unternehmen (KMU) noch weitgehend unerschlossen.[24] Die Kommission hat daher 2011 angestoßen, die PSI-Richtlinie weiterzuentwickeln und ihr stärkere Geltung zu verschaffen, u.a. mit Blick auf die EU Richtlinie über den Zugang der Öffentlichkeit zu Umweltinformationen („Aarhus-Richtlinie“), die Richtlinie zur Geodateninfrastruktur („INSPIRE Richtlinie“) und die Richtlinie für die Einführung intelligenter Verkehrssysteme im Straßenverkehr.[25]

Ziel ist es, über den Weiterverwendungsanspruch hinaus die Freigabe öffentlicher Datenbestände voranzubringen und insbesondere den Mitgliedstaaten eine Verpflichtung zur technischen Ermöglichung und Gestattung der Weiterverwendung von allgemein zugänglichen Dokumenten für gewerbliche und nichtgewerbliche Zwecke aufzuerlegen. In dem Zusammenhang sollen die Mitgliedstaaten angehalten werden, die Verwendung offener behördlicher Lizenzen[26] zu fördern sowie Dokumente in maschinenlesbarem Format zusammen mit den zugehörigen Metadaten verfügbar zu machen. Der Kostendeckungsgrundsatz bleibt im Wesentlichen unverändert, wobei für Dokumente im Grundsatz lediglich die durch Vervielfältigung und Weiterverbreitung verursachten Zusatzkosten erhoben werden dürfen.[27]

---

[23] EU Richtlinie 2003/98/EG über die Weiterverwendung von Informationen des öffentlichen Sektors, ABl. L 345/ S. 90ff.

[24] Zum Änderungsvorschlag für die PSI-Richtlinie siehe KOM (2011) 877 vom 12. Dezember 2011.

[25] Richtlinie 2003/4/EG vom 23. Januar 2003 über den Zugang der Öffentlichkeit zu Umweltinformationen, ABl. L 41, S. 26ff.; Richtlinie 2007/2/EG vom 14. März 2007 zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft, ABl. L 108, S. 1ff.; Richtlinie 2010/40/EU vom 7. Juli 2010 zum Rahmen für die Einführung intelligenter Verkehrssysteme im Straßenverkehr und für deren Schnittstellen zu anderen Verkehrsträgern, ABl. L 207, S. 1ff.

[26] Wie z.B. für urheberrechtlich geschützte Dokumente durch die creative commons Lizenzen (siehe dazu näher www.creativecommons.org).

[27] In begründeten Ausnahmefällen dürfen für die Weiterverwendung zusätzliche Gebühren erhoben werden, auch wenn diese über den genannten Zusatzkosten liegen, wenn diese nach objektiven, transparenten und nachprüfbaren Kriterien und mit Zustimmung der in Artikel 4 Absatz 4 der PSI-Richtlinie genannten unabhängigen Behörde festgelegt sind; siehe Änderungsvorschlag für die PSI-Richtlinie, a.a.O., S. 20.

Soweit es sich um strukturierte Daten und Datenbanken handelt, kann die Überlassung durch Vertrag und ggf. auch durch spezifisch geregelte Nutzungsbedingungen erfolgen, wobei diese dann – soweit die Inhalte keinen Urheberrechtsschutz genießen – als Datenbanklizenzen auszugestalten sind.[28]

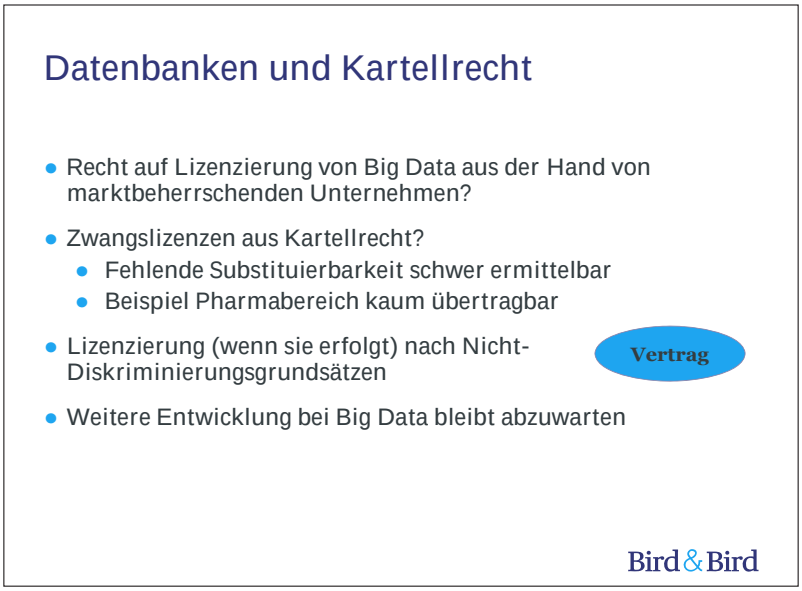

Bild 11

11. Datenbanken und Kartellrecht

Die Regelungen zum sui generis Schutz für Datenbanken stellen ausdrücklich klar, dass das damit verbundene Ausschließlichkeitsrecht die Bestimmungen des Kartellrechts unberührt lässt.[29] (Bild 11) Die Kommission hat vorhergesehen, dass das Ausschließlichkeitsrecht aus §§ 87 a ff. UrhG den Missbrauch einer beherrschenden Stellung erleichtern könnte, *„insbesondere in Bezug auf die Schaffung und Verbreitung neuer Produkte und Dienste, die einen Mehrwert geistiger, dokumentarischer, technischer, wirtschaftlicher oder kommerzieller Art aufweisen.“*[30]

Neben dem Diskriminierungsverbot bei der tatsächlichen Erteilung von Datenbanklizenzen (bei denen man in Anlehnung an Patentlizenzen an die FRAND-Grundsätze denken würde)[31] könnte sich für Big Data, das sich in der Hand marktbeherrschender Unternehmen befindet, womöglich sogar die Frage stellen, ob Dritte einen Anspruch auf Lizenzerteilung gegen den Willen des Datenbankinhabers geltend machen könnten.[32] Zwar dürfte es – außerhalb des öffentlichen Sektors – im Einzelnen nicht leicht möglich sein, den relevanten Markt

[28] Siehe Änderungsvorschlag für die PSI-Richtlinie, a.a.O., S. 19.
[29] Siehe Artikel 13 der EU Datenbankrichtlinie 96/6/EG vom 11. März 1996, ABl. L 77.
[30] Siehe Erwägungsgrund 47 der EU Datenbankrichtlinie, a.a.O.
[31] „FRAND“ steht im Patentrecht für Lizenzerteilungen, die „fair, reasonable and non-discriminatory“ erfolgen.
[32] Gemäß Artikel 31 des TRIPS Abkommens sind Zwangslizenzen bisher auf Patente beschränkt.

einzugrenzen und die fehlende Substituierbarkeit darzutun. Es bleibt insoweit aber – mit verhaltener Skepsis – abzuwarten, ob Dritte in Zukunft faktische Monopolstellungen von Inhabern extrem großer Datenvorkommen, wie sie etwa bei Suchmaschinenbetreibern und den Betreibern von sozialen Netzwerken anfallen, mit Erfolg angreifen könnten.

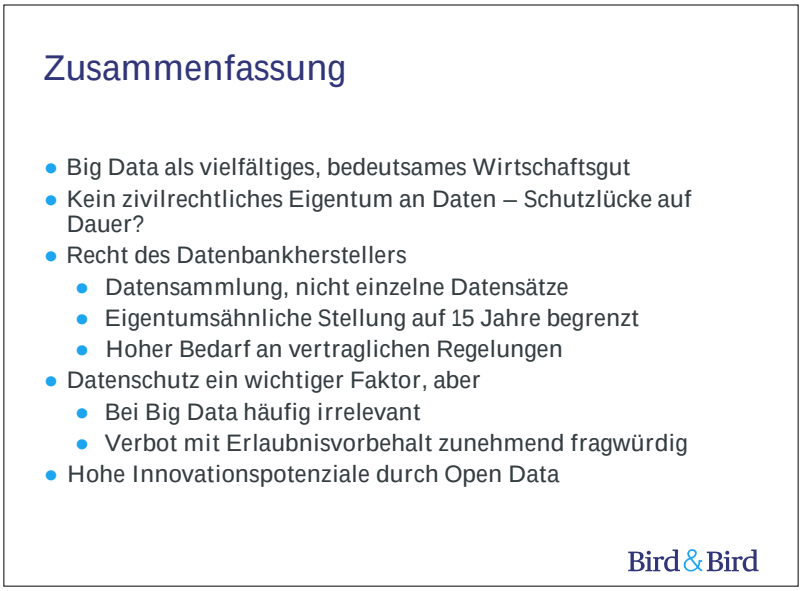

Bild 12

12. Zusammenfassung

Big Data wird zu einem zunehmend bedeutsamen Treiber von Innovation. Um als Wirtschaftsgut Platz zu greifen, muss die Verkehrsfähigkeit von Big Data – also insbesondere dessen Verwertung und Übertragbarkeit – gesichert sein (Bild 12). Indem einzelne Datensätze grundsätzlich nicht individuell eigentumsfähig sind, kommt es mithin maßgeblich auf das Recht des Datenbankherstellers an. Hier ergibt sich aus der einschränkenden Rechtsprechung des EuGH jedoch keineswegs, dass jede große Datensammlung per se auch unter den Schutz des sui generis Rechts aus § 87 a UrhG fällt.

Soweit das Recht des Datenbankherstellers besteht, folgt aus dem sui generis Schutz – sowohl auf der Erstellungs- und Zulieferseite als auch auf der Verwertungsseite – ein hoher Bedarf an vertraglichen Regelungen. Hier steht die Rechtspraxis in vieler Hinsicht erst am Anfang.

Der Datenschutz ist ein wichtiger Faktor in der rechtlichen Analyse von Big Data. Allerdings bleibt er in einer Vielzahl von Konstellationen vollständig irrelevant. Soweit es zu Mischkonstellationen und international integrierten Datenbanken kommt, können sich datenschutzrechtlich hochkomplexe Gemengelagen ergeben. Die Austarierung der Compliance-Risiken führt zu differenzierten vertraglichen Gestaltungen.

Big Data führt verstärkt vor Augen, dass der herkömmliche Ansatz des Verbots mit Erlaubnisvorbehalt zunehmend in Frage steht. Der Schritt zu einer weitergehenden Freigabe der Datennutzung unter Sicherung und Wahrung eines klaren Schutzbereichs für Daten, die den

Kernbereich der persönlichen Lebensführung betreffen, ist für die Zukunft eine denkbare und erstrebenswerte Vision.

Dem öffentlichen Sektor kommt in der Freigabe von Daten und Dokumenten aus Behörden und öffentlichen Einrichtungen eine maßgebliche Rolle in der Ermöglichung und Beschleunigung von Innovation zu. Hier ist mit der Überarbeitung der PSI-Richtlinie ein wichtiger Zukunftsschub zugunsten von Big Data und Innovation gerade im KMU-Umfeld zu erwarten.

5 Innovative Anwendungsfälle für Datenanalyse

Dr. Volker Rieger, Detecon International GmbH, Bonn

Vor über zehn Jahren habe ich mich in der Automobilindustrie mit der Verbindung von Informations- und Kommunikationstechnologie mit dem Fahrzeug beschäftigt. Damals steckte das Thema Telematik – oder „Connected Car“, wie es heute teilweise genannt wird -, noch in den Kinderschuhen. Das Spannende für mich war damals weniger die technischen Herausforderungen, sondern die Branchenkonvergenz zwischen Automobilindustrie einerseits und den ICT- und Consumer Electronics-Branchen. Ich glaube, die vielen Beispiele, die wir heute Morgen schon gesehen haben und die, die ich Ihnen noch vorstellen möchte, zeigen, dass genau diese Konvergenz eines der zentralen Themen von Big Data und entsprechenden Anwendungen ist.

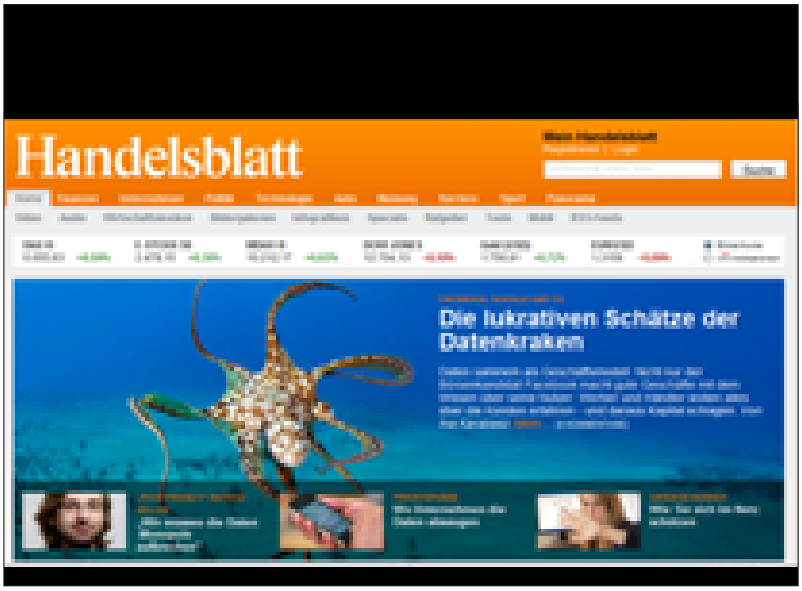

Bild 1

Während wir den Vortrag mit dem Programmausschuss abgestimmt haben, bin ich eines Tages mit diesen Schlagzeilen des Handelsblatts Online konfrontiert worden (Bild 1). Erst einmal sehe ich es grundsätzlich als eine positive Entwicklung, dass das Thema Datenanalyse von Meinungsbildnern dargestellt und damit wahrgenommen wurde. Sie sehen aber natürlich schon an den Begriffen wie „Datenkraken“, „Monopole aufbrechen“, „Daten absaugen“, „sich im Netz schützen“, dass die öffentliche Diskussion leider noch häufig mit negativen Aspekten belegt ist. Ich für meine Person und - ich vermute - die meisten von Ihnen hier im Raum auch, glauben an die Chancen von Big Data. Die Vorredner haben die Potenziale dargestellt, und ich möchte dies aufgreifen und durch weitere Beispiele ergänzen. Ich hoffe, dass wir gemeinsam damit auch die öffentliche Diskussion beeinflussen können.

Ich habe einen etwas anderen Blick auf das Thema als meine Vorredner. Datenanalyse sehe ich nicht so sehr unter den Aspekten „Wettbewerbsvorteile erzielen“ und „interne Unternehmensentscheidungen durch gute Informationen aufbereiten“. Dies sind natürlich wichtige Einsatzfelder und auch bedeutende Treiber des Themas Big Data. Aber ich möchte Ihren

Blick auch noch auf einen anderen Aspekt lenken. Daten und aus ihnen gewonnene Informationen sind in vielen Industrien inzwischen elementarer Teil der Wertschöpfung geworden. Damit stiften die Daten einen direkten Nutzen für die Kunden. Dies ist aus meiner Sicht ein ganz entscheidender Mechanismus, um die Akzeptanz im Markt zu erhöhen.

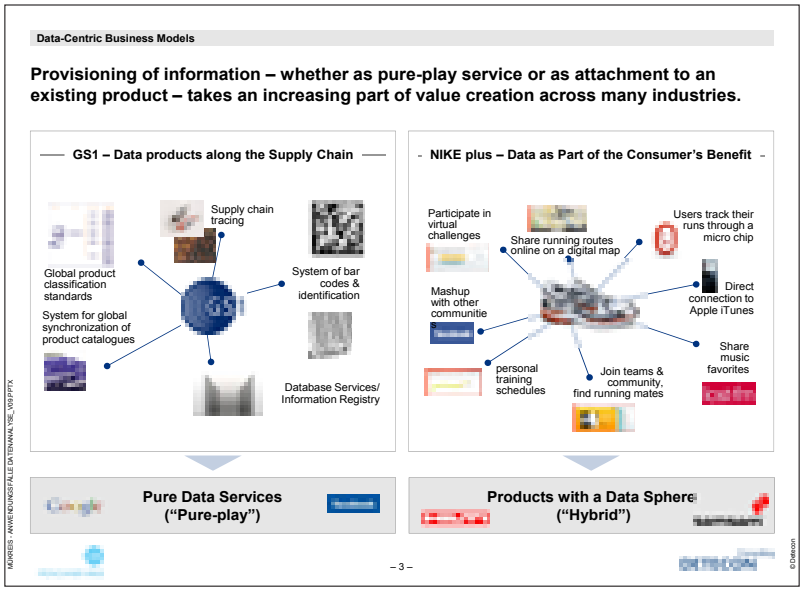

Bild 2

In unseren Projekten und Diskussionen unterscheiden wir zwei grundlegende Ansätze der Wertschöpfung mit Daten, die Sie auf Bild 2 dargestellt sehen. Zum einen gibt es Unternehmen, die im Wesentlichen nur mit Daten handeln, für die Daten der primäre Teil der Wertschöpfung sind. Hierfür gibt es ganz klassische Beispiele. Das GS1-Netzwerk ist die Organisation hinter den Daten, die Sie auf den Strichcodes an der Supermarktkasse finden. In diese Kategorie gehören auch die großen neuen Player, die in aller Munde sind wie Google oder Facebook - letzteres jetzt ganz aktuell noch einmal durch den Börsengang.
Ein viel spannenderer Bereich für mich ist das, was auf der rechten Hälfte von Bild 2 dargestellt ist. Viele Unternehmen gehen dazu über, etwas, das wir „Datensphäre“ nennen, rund um ihr Produkt aufzubauen. Wir haben das Beispiel Fahrzeugdaten in der Automobilindustrie bereits von meinem Vorredner dargestellt bekommen.

Ein anderes Beispiel, welches ich seit längerem nutze, ist die Firma Nike, die rund um ein Produkt, den Turnschuh, den man erst einmal überhaupt nicht mit Informations- und Kommunikationstechnologie in Verbindung bringen würde, eine Datenwelt aufgebaut hat. Viele von Ihnen werden das sicher kennen: den Sensor im Schuh, der über iPhones, iPads und inzwischen auch andere Geräte die Bewegungsdaten ausliest, und die dann in einem Internetportal bereitgestellt werden. Dadurch entsteht für den Läufer, den Sportbegeisterten, eine völlig andere wertliche Anmutung dieses Turnschuhs: Strecken können abgespeichert werden; Vergleichswettkämpfe gegen andere Sportler können durchgeführt werden; oder ich kann mir meine Lieblingsmusik zu den Trainingsstrecken bereitstellen. Das ist ein schönes Beispiel - andere sind inzwischen nachgezogen, auch die deutschen Sportartikelhersteller - dafür, wie über Daten Mehrwert für den Verbraucher, für den Nutzer erzeugt werden kann. Auf diesen Aspekt möchte ich auch meine Beispiele im weiteren Teil des Vortrags fokussieren.

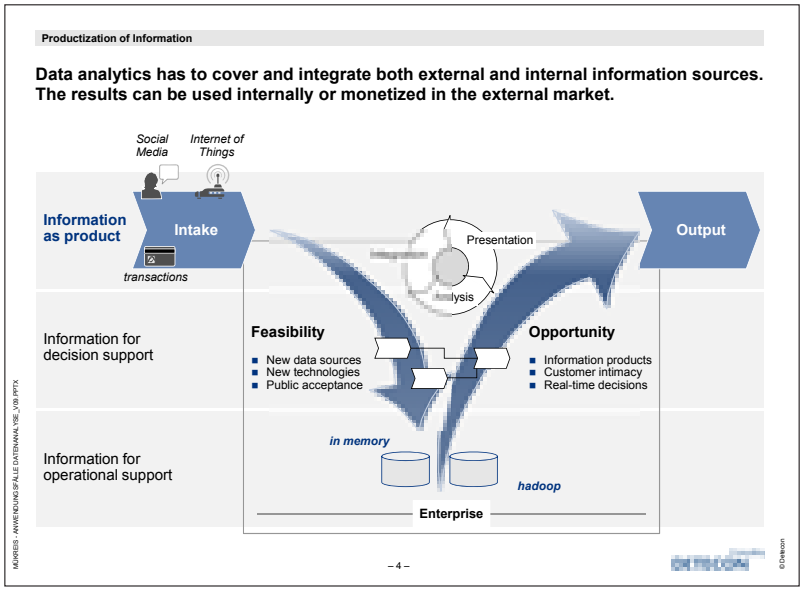

Bild 3

Wenn man solche Art von Wertschöpfung über Daten erzielen will, ist der Blick darauf, wo die Daten entstehen und wo sie herkommen, wichtig (Bild 3). Man findet im Kontext Datenanalyse heute häufig überwiegend die unternehmensinterne Sicht, vielleicht auch historisch aus der Rubrik „Business Intelligence“ geprägt. Diese Sicht betrachtet üblicherweise Unternehmensdaten, die in den Geschäftsprozessen anfallen und in geeigneten Speichersystemen und Datenbanken abgelegt sind. Für die Wertschöpfungs-Perspektive, die mir wichtig ist, gehört erstens jetzt neu und zusätzlich dazu, dass es auch Daten gibt - und in Zukunft immer mehr geben wird -, die außerhalb des Unternehmens entstehen oder die außerhalb des Unternehmens nutzbar sind, auf die man zugreifen kann. Das sind zum Beispiel Social-Media-Daten oder Daten aus dem Internet der Dinge, die in vielfältigen Sensoren anfallen. Als zweites ist es auch ganz wichtig, dass Unternehmen ihre Daten nicht nur für interne Entscheidungen nutzen können, sondern dass Sie sie monetarisieren können, dass die Daten Teil der Produktwelt werden und damit einen konkreten Mehrwert für die Kunden dieser Unternehmen darstellen können. Die wesentlichen Schritte der Verbindung finden Sie in dem Kreis in der Mitte von Bild 3 dargestellt. Dies umfasst zuerst die Integration der Daten, dann die Analyse und abschließend die Aufbereitung, so dass dann wirklich ein nutzbarer Wert für die Kunden, für die Verbraucher entsteht.

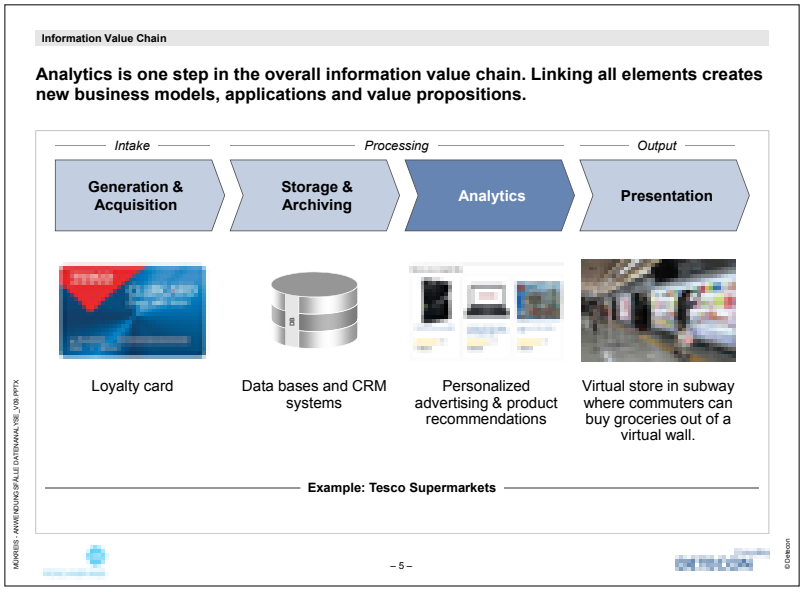

Bild 4

Wenn wir über Wertschöpfung mit Daten sprechen, dann hilft auch ein Blick auf eine Wertschöpfungskette. Auf Bild 4 finden Sie die, die wir für datenbasierte Wertschöpfung in unseren Projekten nutzen, dargestellt. Parallel ist sie am Beispiel von Tesco, der britischen Supermarktkette, illustriert. Daten müssen erst einmal generiert oder akquiriert werden. Dann müssen Sie in einer sinnvollen Form gespeichert werden. Als nächstes müssen Analysen, Auswertungen, durchgeführt werden. Abschließend – und das ist ein sehr wichtiger Schritt - müssen sie dann für die Nutzer – und das sind nicht immer nur die Manager, die Entscheider in den Unternehmen, sondern auch für die Kunden des Unternehmens - dargestellt werden.

Das Beispiel von Tesco kennen Sie vielleicht schon, es wird vielfach in der Literatur dargestellt. Tesco war federführend bei dem Einsatz von Kundenkarten. Wie hinlänglich bekannt ist, entstehen dabei zahlreiche Daten, die gespeichert und ausgewertet werden. Die koreanische Tochter Tesco Homeplus nutzt diese nun in einem revolutionären Ansatz, bei dem die Kunden nicht in den Laden kommen müssen, sondern der Laden zu den Kunden kommt. Tesco Homeplus hat virtuelle Regale in U-Bahnstationen angebracht. Dort sind die am häufigsten gekauften Produkte bildlich dargestellt. Diese fotografiert man mit seinem Smartphone ab und bekommt sie dann kurze Zeit später direkt nach Hause geliefert.

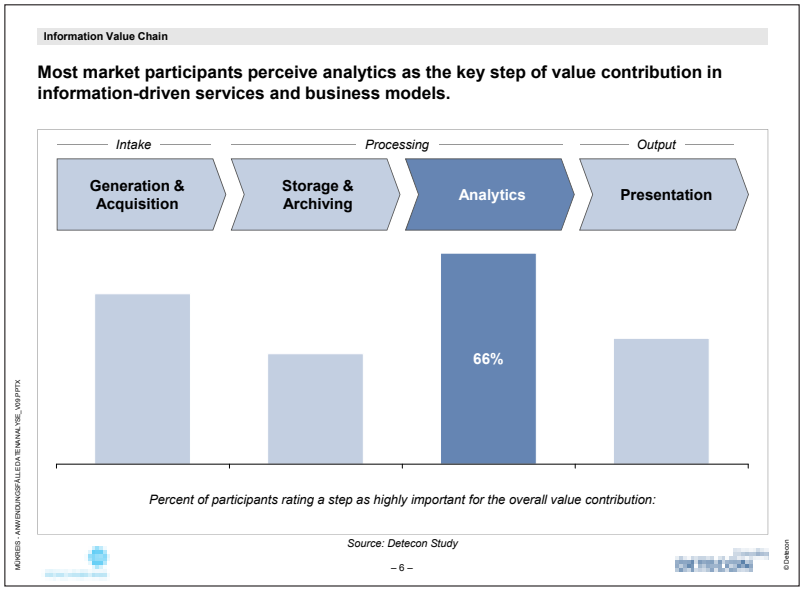

Bild 5

Vor einiger Zeit haben wir eine Studie mit sowohl Unternehmen, die schon heute Datenwertschöpfung betreiben, also Daten im Markt anbieten, als auch solchen, die angebotene Daten als Anwender nutzen, durchgeführt. Unter anderem haben wir die Frage gestellt, wo die größten Wertbeiträge entlang der Wertschöpfungskette gesehen werden (Bild 5). Dabei ist deutlich geworden, dass der wesentliche Bereich der Analysebereich ist – und das ist somit zu Recht das Thema unserer Konferenz. Aber – und auch diese Botschaft möchte ich Ihnen heute mitgeben - die Erzeugung und die Akquisition der Daten sind ebenfalls wichtige Bereiche.

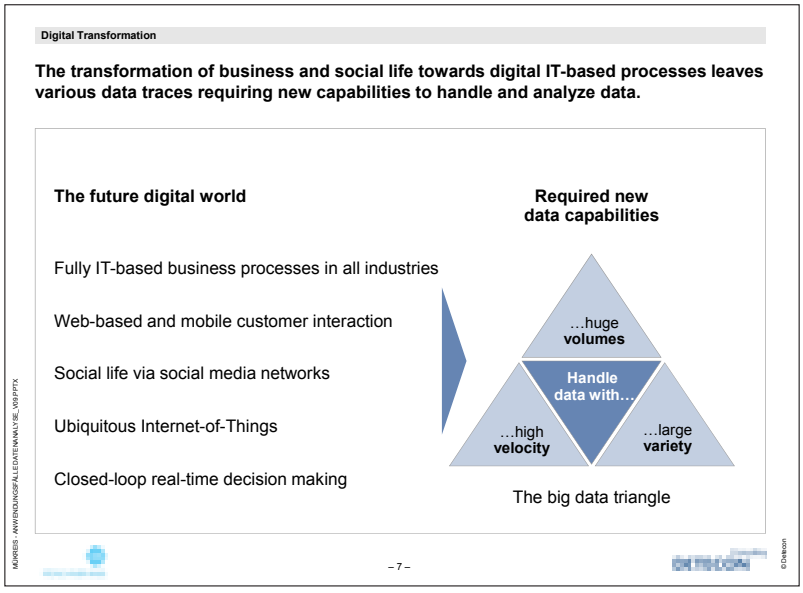

Bild 6

Wo kommen nun die Daten her (Bild 6)? Zum einen entstehen natürlich dadurch viele Daten, weil die Geschäftsprozesse in vielen Industrien immer weiter digitalisiert werden. In manchen Branchen ist das bereits heute annähernd zu 100 % der Fall. Aber es gibt auch andere Bereiche - als Extremfall mag die öffentliche Verwaltung gelten - wo noch sehr viel mit Papier gearbeitet wird. Der Trend zur Digitalisierung schreitet aber immer weiter voran. Eine wichtige Quelle ist dadurch entstanden, dass in vielen Branchen die Interaktionen mit dem Kunden elektronisiert wurden, dass wir webbasierte Schnittstellen für den Vertrieb - Webshops - aber auch für die sonstige Kundeninteraktion CRM, Customer Self Service, usw. haben.

Die dritte Quelle, Social Media, ist schon von meinen Vorrednern angesprochen worden. Dort entstehen einfach „spontan“ Daten - unabhängig von Unternehmens- oder wirtschaftlichen Prozessen. Eine weitere Quelle ist das sogenannte „Internet der Dinge“, also Maschinendaten, die von vielfältigen Sensoren und Microcontrollern in Geräten wie Autos aber auch Waschmaschinen erzeugt werden. Es gibt Abschätzungen, wonach es über 50 Milliarden Microcontroller gibt – Tendenz steigend -, die zukünftig miteinander vernetzt sind und Daten austauschen. Damit haben wir wichtige Treiber aus der Angebotssicht dargestellt. Ein weiterer Treiber, der schon bei meinen Vorrednern anklang, ist der steigende Bedarf nach kurzfristigen Entscheidungen in Wirtschaft und Politik, für die die Entscheider „in Echtzeit“ Daten nutzen wollen. Diese fünf Punkte zusammen ergeben dann Big Data, mit den drei Vs, die meine Vorrender ebenfalls schon vorgestellt haben.

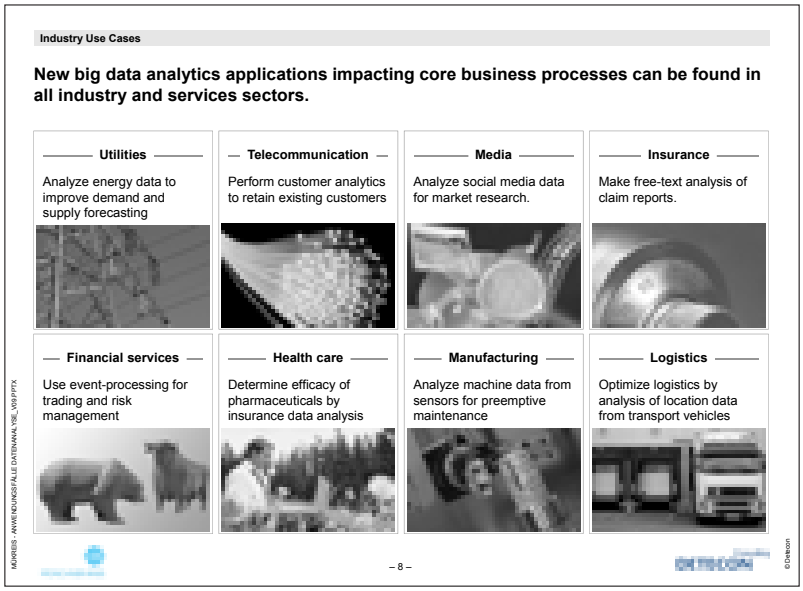

Bild 7

Damit komme ich zu der Übersicht möglicher Anwendungsfälle von Big Data (Bild 7). Einige sind bei meinen Vorrednern erwähnt worden, auch weil sie momentan die zentralen Beispiele in der öffentlichen Diskussion sind. Hierzu zählt die Energieversorgung, bei der vielfältige Daten entlang der Energiewertschöpfungskette - in der Erzeugung, in der Verteilung aber auch im Verbrauch, wo Smart Meter momentan ein zentrales Schlüsselkonzept sind, um die Energiewende zu managen und zu steuern – anfallen und genutzt werden können. In diesem Bereich gibt es sehr viele spannende Anwendungsbeispiele. Gerade hier in

München gibt es mit der Entelios AG ein sehr interessantes Start-up Unternehmen, das im Thema Demand Response Management unterwegs ist. Hierbei schaltet man Großverbraucher wie Kühlaggregate oder Pumpen kurzfristig ab oder zu, wenn die Energieerzeugung durch Solar oder Wind volatil ist. Dadurch lassen sich kurzfristige Angebots- oder Bedarfsspitzen ausregeln. Demand Response Management benötigt umfangreiche Messungen und Daten und insbesondere auch Analysen bis hin zu Prognosen, wann solche Energieverbrauchsschwankungen eintreten werden.

Die Telekommunikation ist auch schon erwähnt worden. Gerade das Thema Customer Analytics spielt in dieser Branche eine wesentliche Rolle, um dem Kundenschwund zu begegnen. Im Medienbereich gibt es vielfache Anwendungen, die bis in das Thema Social Media hineingehen. Versicherungsunternehmen setzen sich ebenfalls mit dem Thema Big Data auseinander. In der Versicherungsbranche hat man es häufig mit unstrukturierten Daten zu tun. Zum Beispiel können Schadenfälle, die Kunden ihrer Versicherung beschreiben, analysiert werden. Hierbei findet die Analyse überwiegend auf Basis unstrukturierter Daten – Textdokumente – statt. Die Daten werden für die Schadenabwicklung ausgewertet, aber auch um Muster zu erkennen, die für die Prävention genutzt werden können.

Das Thema Big Data im Finanzbereich ist in den letzten Jahren auch ausführlichst und in vielen Facetten in der Öffentlichkeit diskutiert worden. Dabei sind natürlich auch die Risiken im Bereich automatisierter Finanzanalyse deutlich geworden. Analysen sind letztlich immer ein Blick in die Vergangenheit verbunden mit statistischen Annahmen und Risikomodellen. Wenn diese dann nicht greifen oder auf falschen Annahmen beruhen, kann es mächtige Schieflagen geben, die wir alle in jüngster Zeit im Finanzsektor deutlich wahrgenommen haben.

Der Gesundheitsbereich ist ein wichtiges Anwendungsfeld und wurde auch bereits erwähnt. In Deutschland besteht großes Potenzial für Datenanalysen im gesamten Bereich Maschinenbau. Viele Unternehmen sind bereits dazu übergegangen, ihren gesamten Maschinenpark zu vernetzen. Damit kann zum einen die Produktion auf Basis dieser Maschinen optimiert werden. Vielfach wird die Vernetzung aber auch von den Herstellern eingesetzt, um Wartungsdienstleistungen anzubieten. Dies geht soweit, dass inzwischen bereits Managed Service-Geschäftsmodelle angeboten werden. Hierbei wird nicht mehr die Maschine verkauft, sondern der Betrieb der Maschine als Dienstleistung angeboten. Beispielsweise wird nicht mehr das Messgerät verkauft, sondern es werden Tausend Messungen in der Stunde als Dienstleistung angeboten. Gerade für solche Anbieter ist es enorm wichtig, permanent Dateninformation über die Betriebsparameter der Maschine zu haben und diese dann auch zu analysieren, um vorausschauend Wartungsarbeiten durchführen zu können. Ich denke, wir werden in den nachfolgenden Vorträgen hierzu einiges spannendes hören.

Das Thema Logistik deckt auch den Automobilbereich ab. In der Logistik sind telematische Lösungen, bei denen komplexe Lieferketten mit vielfältigen Daten entlang des Logistikprozesses - sei es auf der Ebene eines einzelnen Pakets oder eines Container oder eines LKWs - gesteuert werden, inzwischen gang und gäbe. Sie alle kennen die Geräte, mit denen der Paketzusteller bei Ihnen zuhause vor der Tür steht und in die das letzte Datum am Abschluss einer langen Lieferkette eingegeben wird.

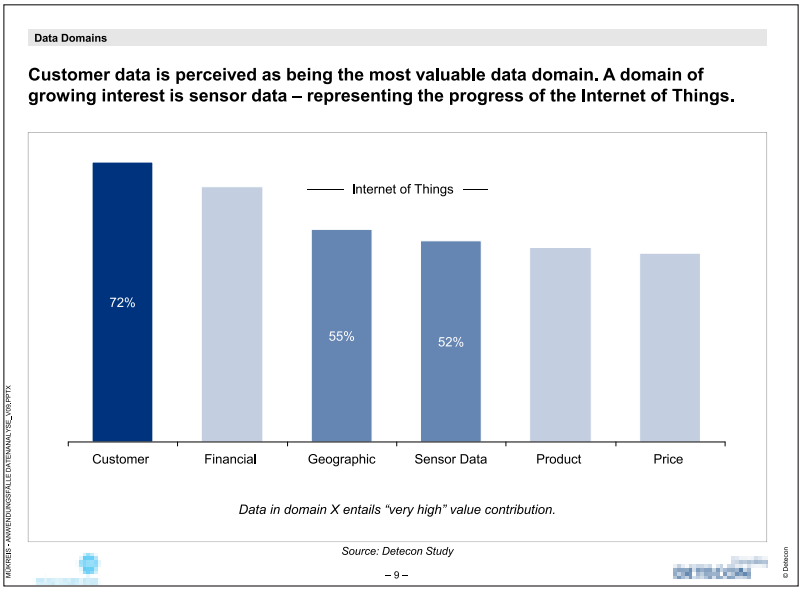

Bild 8

Ich möchte Ihren Blick noch einmal darauf richten, welche Arten von Daten es gibt (Bild 8). Bedeutsam sind natürlich die Kundendaten, die auch in unserer Studie, aus der ich hier noch einmal Ergebnisse präsentiere, als wichtigster Typ genannt wurden. In diesem Bereich müssen wir besonders sensibel vorgehen und die datenschutzrechtlichen Aspekte berücksichtigen. Ich möchte Ihren Blick jedoch vor allem auch auf die weiteren Arten von Daten lenken, in denen auch umfangreiches Potenzial liegt. Hierzu zählen Finanzdaten, geografische Daten, Maschinendaten, Produktdaten, Preisinformationen, die alle genutzt werden können, um für Kunden und Unternehmen Mehrwert zu schaffen.

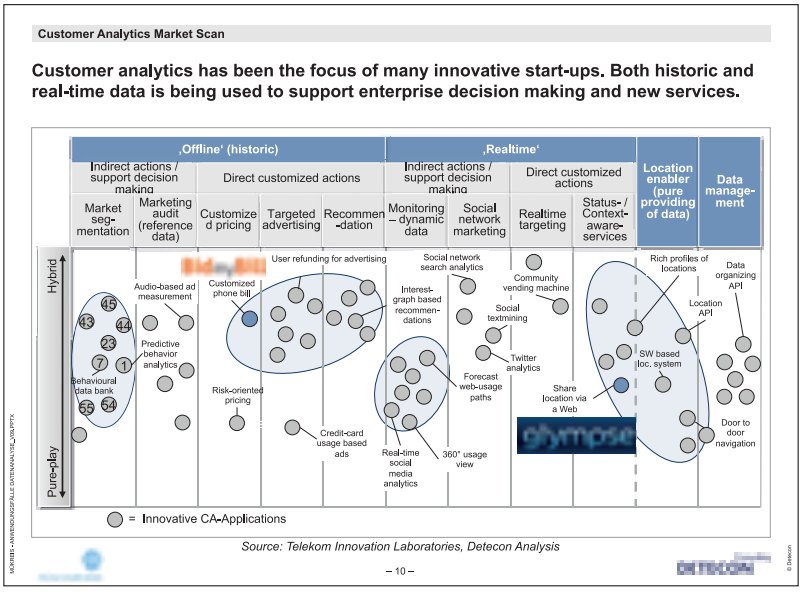

Bild 9

Für den Bereich Kundendaten haben wir gemeinsam mit der Deutschen Telekom - mit den Telekom Innovation Laboratories - im letzten Jahr eine Studie gemacht und uns dabei eine Übersicht verschafft, welche Arten von Anwendungen heute schon im Markt zu finden sind (Bild 9). Dabei haben wir unseren Blick nicht nur auf Business Intelligence-Lösungen fokussiert sondern auch solche Lösungen betrachtet, die das Wissen um den Kunden im Sinne des Verbrauchers wieder an ihn oder sie zurück spielt. Vielfach finden sich hier Start-ups, die Geschäftsmodelle entwickelt haben, um aus Kundendaten oder kundennahen Daten einen Mehrwert zu schaffen. Die Anwendungsfälle lassen sich grob in solche mit historischen Daten, die offline ausgewertet werden, und solche unter Nutzung von Echtzeitdaten aufteilen. In beiden Bereichen entstehen auch Daten, die überwiegend für die unternehmensinterne Entscheidungsfindung genutzt werden. Aber – und das möchte ich betonen - es gibt schon bereits heute sehr viele Beispiele im Markt, die ich leider nicht alle einzeln vorstellen kann, wo die Daten genutzt werden, um den Kunden einen Mehrwert zu liefern. Gleiches gilt auch für den Echtzeitbereich.

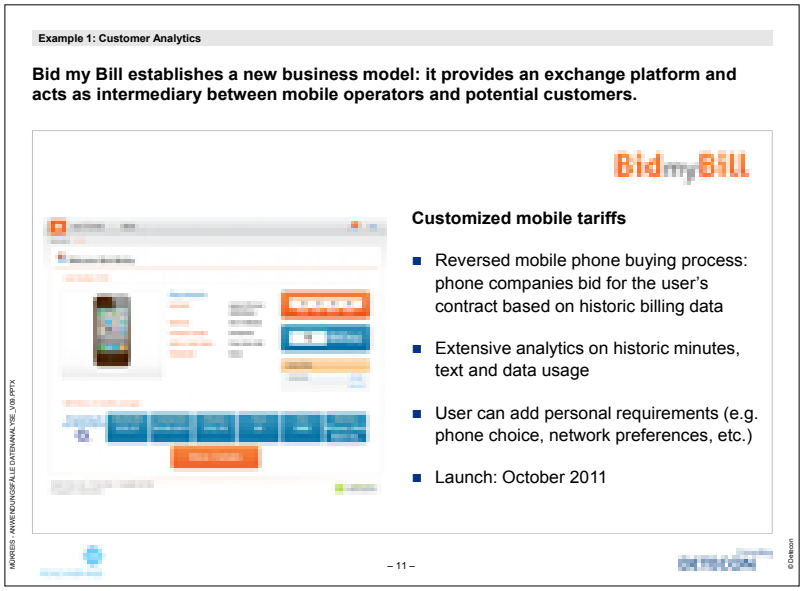

Bild 10

Einige ausgewählte dieser Beispiele möchte ich nun kurz in Snapshots vorstellen. Das englische Unternehmen BidmyBill (Bild 10) bietet Verbrauchern einen Dienst, bei dem diese ihre Verbrauchsdaten aus dem Mobilfunk in ein Portal hochladen können. Dieses führt dann Auswertungen durch und ermöglicht es, einen optimal geeigneten Telefontarif angeboten zu bekommen. Bestands- und Verkehrsdaten in der Telekommunikation sind traditionell ein sehr sensibles Thema, zu dem es durch das Fernmeldegeheimnis und auch im Telekommunikationsgesetz sehr hohe Datenschutzanforderungen gibt. Für mich sieht man aber an diesem Beispiel, dass, wenn man ihnen einen Mehrwert liefert, Kunden sogar bereit sind, ihre Daten hochzuladen und auswerten zu lassen, weil sie dann eine günstigere Telefonrechnung haben können.

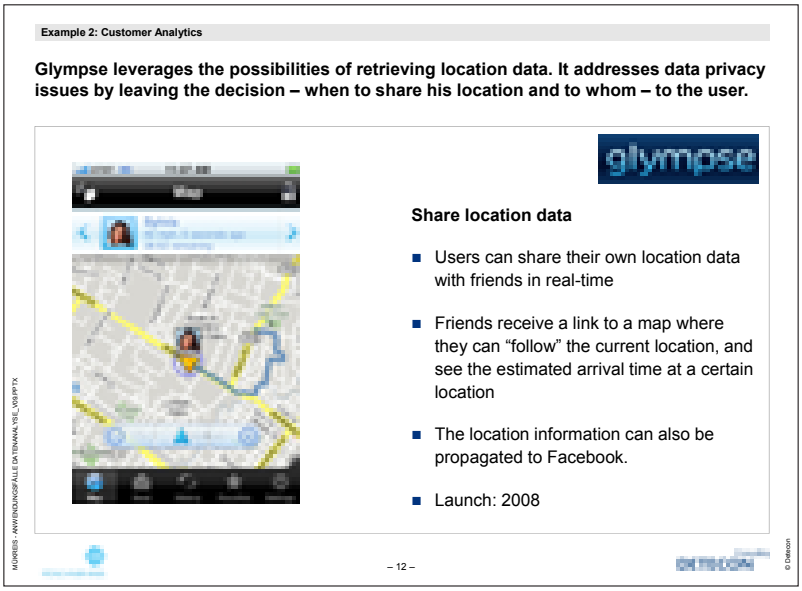

Bild 11

Ein anderes Beispiel, auch ein Start-up, diesmal aus den USA, ist die Firma Glympse (Bild 11). Glympse bietet einen Dienst aus dem Bereich Location Based Services an. Durch das Senden eines „Glympses" können sie Ihren Standort in Echtzeit von anderen verfolgen lassen. Sicher weckt das zum Teil Big Brother-Assoziationen. Glympse geht jedoch davon aus, dass es zahlreiche Anwendungsfälle gibt, in denen Nutzer dieser Art der Datenanalyse als Mehrwert empfinden. Beispiele sind, dass Sie Ihre Freunde über Ihre ungefähre Ankunftszeit informieren wollen oder dass Eltern Ihren Kindern Freiräume einräumen möchten, aber gleichzeitig ein „Auge auf sie halten" möchten. Glympse ist seit dem Jahr 2008 im Markt aktiv.

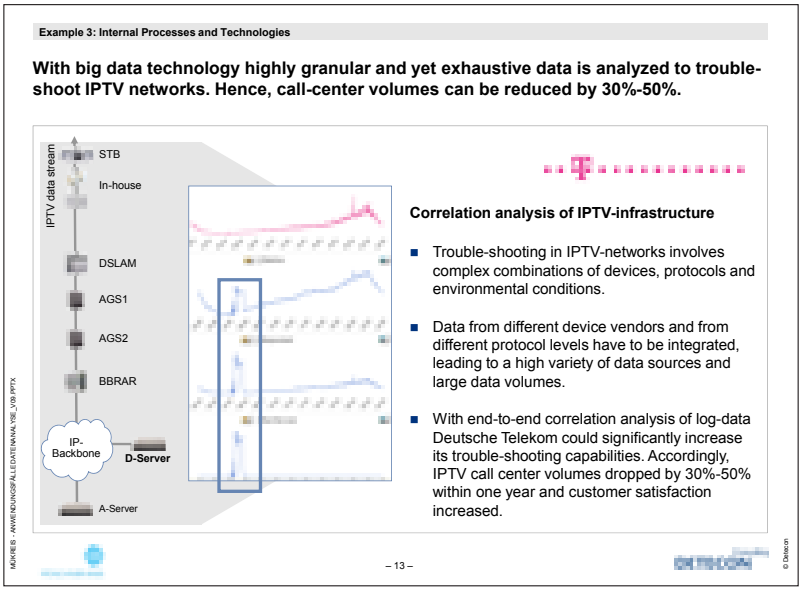

Bild 12

Ein weiteres Beispiel – diesmal eher aus dem internen Bereich – kommt aus der Deutschen Telekom im Zusammenhang mit dem Produkt Entertain (Bild 12). Sie kennen vermutlich das IPTV-Angebot der Deutschen Telekom, ein sehr komplexes technisches Produkt. Es wurde eine Datenanalyse durchgeführt mit dem Ziel, die Qualität des Dienstes zu verbessern. Auf Bild 12 sehen sie die verschiedenen technischen Elemente, die ich hier nicht alle im Einzelnen erläutern möchte, die notwendig sind, um das Fernsehbild über die Telefonleitung auf den Bildschirm zu bringen. Entlang dieser Produktionskette wurden umfangreiche Analysen durchgeführt und dadurch Fehler aufgespürt und Qualitätssicherung betrieben. Dabei konnte die Qualität der Dienste signifikant gesteigert werden. Sichtbar wurde dies daran, dass die Anzahl der Anrufe in der Hotline um 30 bis 50 % zurückgegangen sind – auch hier also eine Analyse, die konkreten Mehrwert für die Verbraucher erzeugt hat, wenn auch auf indirekte Weise.

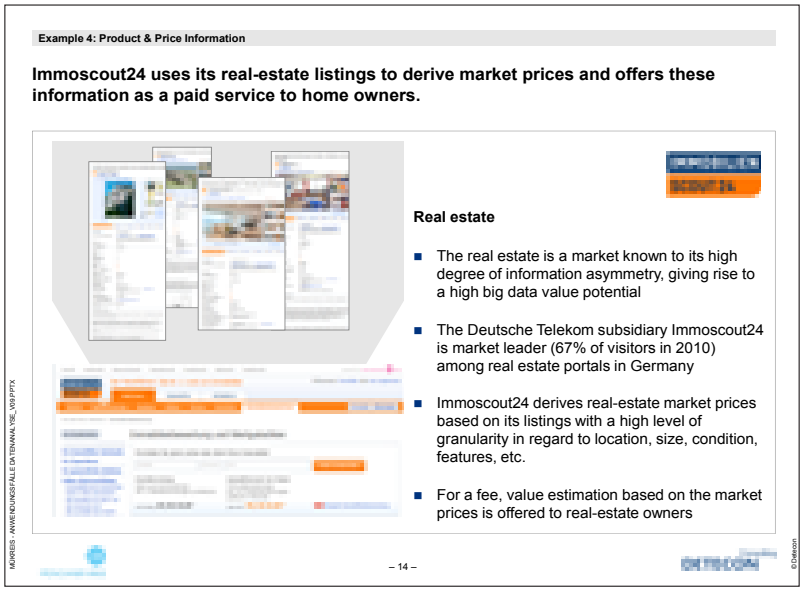

Bild 13

Mein vorletztes Beispiel, ist ein Angebot von Immobilienscout, auch einem Unternehmen im Konzern Deutsche Telkom (Bild 13). Immobilienscout24 ist primär ein Maklerportal, auf dem Wohnungen und Häuser zum Mieten und zum Kauf angeboten werden. Dort fallen viele Daten an, die gar nicht zum primären Geschäftsmodell gehören. Immobilienscout ist dazu übergegangen, diese Daten auszuwerten und gegen Gebühr - also das Thema Monetarisierung von Daten -, Marktinformationen über Hauspreise in gewissen Regionen, Stadtteilen, Stadtbezirken, Trends anzubieten.

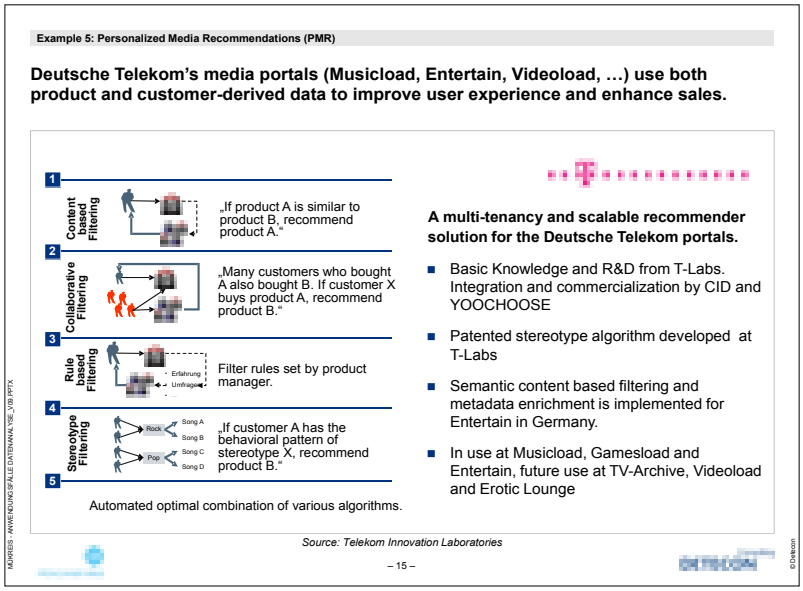

Bild 14

Mit meinem letzten Beispiel möchte noch das Thema Empfehlungen – neudeutsch Recommendations – ansprechen (Bild 14). Recommendations sind im Kern dieses Feature von Amazon, das Sie alle gut und vermutlich schon seit geraumer Zeit kennen: „Kunden die dieses Produkt gekauft haben, haben auch jenes gekauft". Bei den Telekom Innovation Laboratories ist eine neue innovative Recommendation Engine für den Multimediabereich entwickelt worden, die produktspezifische Informationen verknüpfen kann und auch ein Stereotyping der Nutzer durchführen kann. Damit kann identifiziert werden, ob jemand eher ein Rockmusikfan oder ein Popmusikfan ist, und auf dieser Basis können Empfehlungen gegeben werden. Auch hierbei gibt es einen Mehrwert für das Unternehmen – gesteigerte Absätze – aber auch für die Nutzer, die bessere Vorschläge erhalten und denen ein besseres Musikangebot bereitgestellt werden kann.

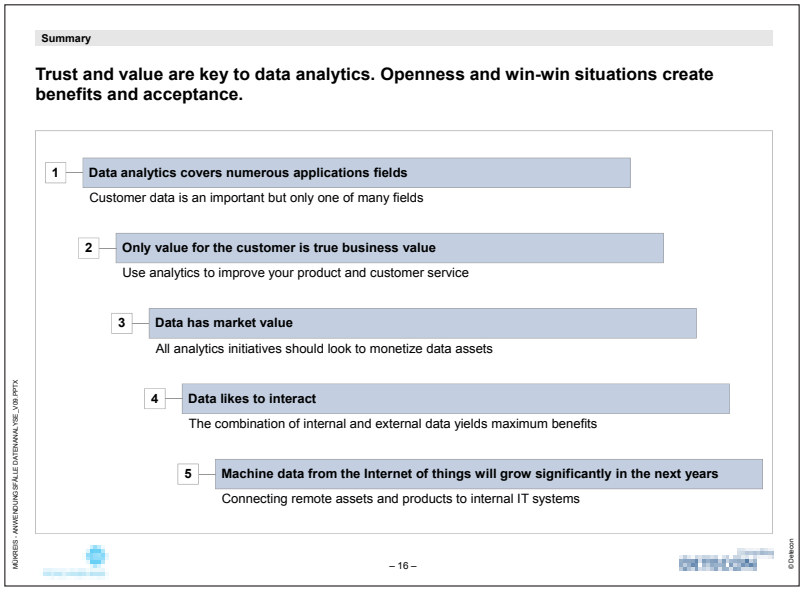

Bild 15

Damit bin ich am Ende meines Vortrags (Bild 15). Zusammenfassend kann man sagen, dass es vielfältige Anwendungsfelder gibt, die ich jeweils nur anreißen konnte. Mein Plädoyer ist, dass wir darauf achten sollten, mit den Daten wirklich Mehrwert für die Kunden zu erzeugen. Daten lassen sich häufig direkt monetarisieren, was explizit den Mehrwert demonstriert. Wichtig ist die Verknüpfung von Daten aus unterschiedlichen Domänen, speziell von unternehmensinternen Daten mit externen Daten. Meine persönliche Überzeugung ist, dass das Thema Maschinendaten, Internet der Dinge, in den nächsten Jahren ein extrem wichtiger Treiber sein wird.

6 Big Data Perspektiven am Beispiel der Analyse von Sensordaten aus Hochgeschwindigkeitszügen

Prof. Dr. Volker Tresp, Siemens AG, München

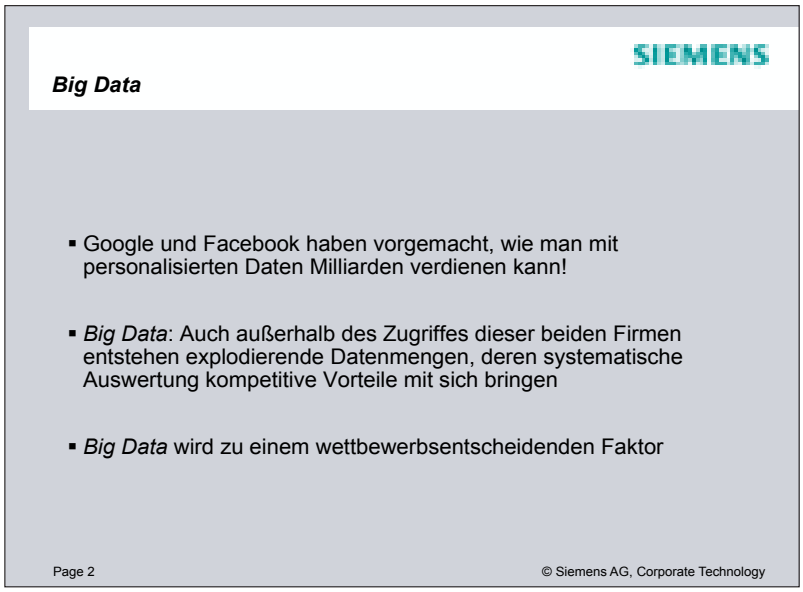

Bild 1

Warum findet Big Data so viel Interesse (Bild 1)? Ein Grund ist sicherlich, dass Google und Facebook uns vorgemacht haben, wie man mit personalisierten Daten Unsummen verdienen kann. Das eigentlich Überraschende hierbei ist, dass die enormen Geschäftspotentiale von Suchmaschinen so lange übersehen wurden. Schon vor Google gab es zum Beispiel die Suchmaschine Altavista, die heute nur noch wenigen ein Begriff sein dürfte. Auch Google selber hatte anfänglich nicht das eigentliche später höchstprofitable Geschäftsmodell gesehen.

Das allgemeine Interesse an Big Data beruht natürlich darauf, dass auch außerhalb des Zugriffes dieser beiden Firmen explodierende Datenmengen entstehen, deren systematische Auswertungen kompetitive Vorteile mit sich bringen können. Jeder hat den Verdacht, dass Big Data vielleicht sein eigenes Geschäft beeinflussen wird, und man will dabei natürlich zu den Googles gehören und nicht zu den Altavistas. Big Data hat das Potential, in einigen Branchen zu einem wettbewerbsentscheidenden Faktor zu werden. Hinzu kommen neue Geschäftsmöglichkeiten, die es vor Big Data so noch gar nicht gegeben hat.

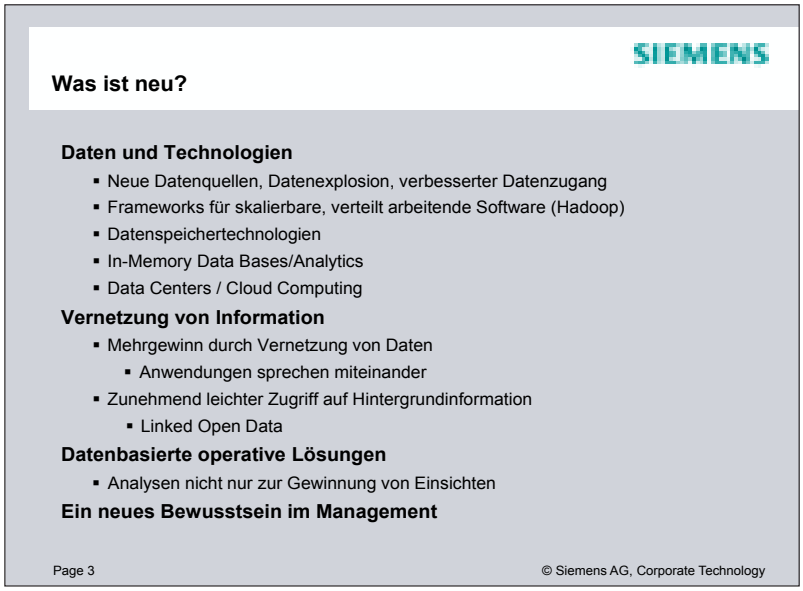

Bild 2

Was ist neu (Bild 2)? Einmal natürlich eine tatsächliche Explosion verfügbarer Daten in existierenden und neu entstehenden Anwendungen, wie zum Beispiel durch Gene Sequenzierung in der personalisierten Medizin. Oft handelt geht es bei Big Data um personengebundene Informationen, wie benutzer- oder patientenspezifische Daten, und das Ziel sind extrem personenspezifische Dienste. Die Schlüsselobjekte müssen aber nicht Personen sein. Im Internet der Dinge werden Informationen zu Ort, Zustand, und Kontext von möglicherweise allen verkauften Produkten einer Produktlinie über das Internet zugänglich und auswertbar. Oder man betrachte das Internet der Sensoren, wo Umweltdaten aufgenommen werden, zum Beispiel zur Früherkennung von Problemen und zur Analyse der Umweltbelastung. Ähnlich vielversprechende Perspektiven ergeben sich im entstehenden Internet der Services.

Ein weiterer Aspekt sind neue Technologien, die Big Data unterstützen, und die Anbieter entsprechender Plattformen sind im Big Data Umfeld augenblicklich sehr sichtbar. Es entstehen Rahmenwerke für skalierbare, verteilt arbeitende Software, wie Hadoop, ebenso neue Datenspeichertechnologien und In-memory Datenbanken und In-memory Analytics Lösungen. Enabling Technologien sind zusätzlich Datenzentren und Cloud Computing.

Ein erheblicher Mehrwert entsteht in der Verknüpfung von Daten. Zu einem erhobenen Datensatz können heute bereits komfortabel Hintergrundinformationen aus unterschiedlichsten Datenquellen hinzugefügt werden, wie aus Geo-Datenbanken, Wetterinformationsdiensten, Eventkalendern, oder Datenbanken mit allgemeinem Hintergrundwissen wie Wikipedia und dessen formalisierten Varianten, DBpedia, Yago, Freebase und dem Google Graph.

Die Verlinkung von Informationsquellen kann die Auswertung und die operative Nutzung der anwendungsspezifischen Daten erheblich verbessern. Aus der Semantic Web Community heraus ist hier die Linked Open Data (LOD) Initiative entstanden, als Verknüpfung diverser Datenquellen von wachsender Quantität aber auch Qualität.

Wie vielleicht schon deutlich geworden ist, geht es bei Big Data nicht nur um den Erkenntnisgewinn, sondern primär um die operative Nutzbarmachung der Daten. Zum Beispiel ist Google nicht nur daran interessiert, welche Benutzerklassen es gibt, viel wichtiger ist es, dem Google-Kunden personenspezifisch zur richtigen Zeit die richtige Werbung zu zeigen, also Hintergrundinformationen operativ zu nutzen.

Neu ist auch, dass das Management offen wird für das Konzept der Datenintelligenz, das heißt dem Konzept, dass Lösungen aus den Daten heraus entstehen.

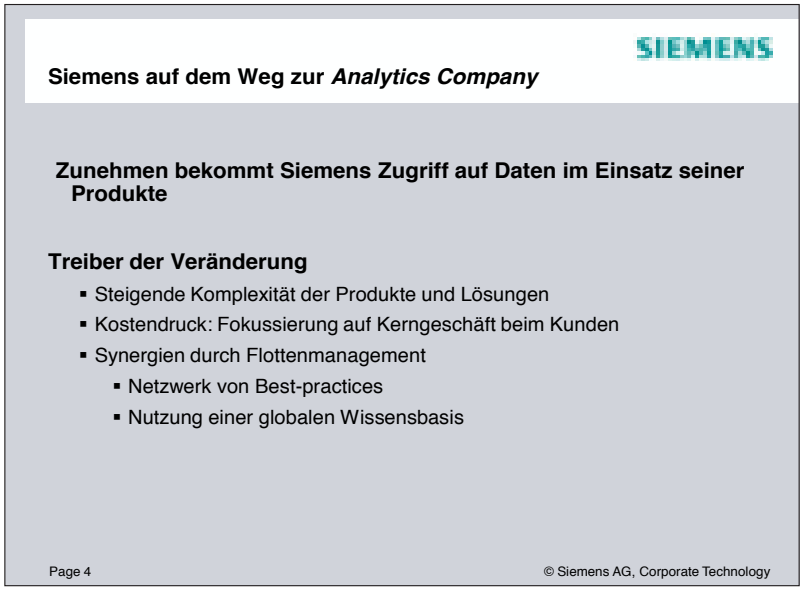

Bild 3

Siemens ist auf dem Weg zu einer Big Data Company (Bild 3). Ein Hauptgrund ist, dass Siemens zunehmend Zugriff auf operative Daten aus dem Einsatz seiner Produkte bekommt. Das war vor einige Jahren bei weitem noch nicht so: Nur in einigen wichtigen Großanwendungen war Siemens auch in die Problemlösungen integrativ mit eingebunden.
Was sind die Treiber der Veränderungen? Ein Grund ist die steigende Komplexität der Produkte und Lösungen, die natürlich der Hersteller schon von vornherein sehr gut versteht und beherrscht, mit der sich aber der Kunde vielleicht schwer tut. Ein zweiter Grund ist der wachsende Kostendruck beim Kunden, der sich natürlich lieber auf sein Kerngeschäft fokussiert würde als sich mit Wartung und Instandhaltung zu beschäftigen. Ein dritter Grund sind Synergien im Flottenmanagement, die Hersteller ausnutzen können. Wenn Siemens z.B. MR- oder CT-Geräte im Healthcare Bereich weltweit betreut, entsteht ein einzigartiges Netzwerk von Best Practices und eine globale Wissensbasis über Problemfälle und Lösungen.

Bild 4

Wartung und Instandhaltung sind der Entry Point, der Türöffner, der weitere geschäftliche Möglichkeiten eröffnet (Bild 4). Über Wartung und Instandhaltung bekommt Siemens ein tiefes Kundenverständnis, wodurch ein kundenspezifisches Angebot ermöglicht wird: Wenn der Siemens Service genau weiß, wie jemand in einer Klinik ein MR-Gerät benutzt, dann kann Siemens diesem Kunden natürlich genau sagen, dass ein MR Gerät der neuen Generation für dessen Probleme maßgeschneidert ist und eine Neuanschaffung sich in kurzer Zeit rechnen würde. Eine technisch tiefe Beratung des Kunden ist möglich. Und Siemens gewinnt natürlich auch ein Verständnis über den Einsatz der eigenen Produkte im Allgemeinen. Dies kann dann die Grundlage zur Entwicklung gänzlich neuer datenbasierter Lösungen sein.

Bild 5

Am Beispiel von Hochgeschwindigkeitszügen möchte ich die angesprochenen Punkte illustrieren (Bild 5). Das Beispiel ist gewählt, weil wir aus der Forschung heraus mit dieser Division im Moment Themen im Big Data Umfeld vorantreiben. Erst ein bisschen Hintergrund. Die Marke ICE ist laut Deutscher Bahn eine der erfolgreichsten Deutschlands. Der Wiedererkennungswert ist enorm. Die neueren Hochgeschwindigkeitszüge in der Linie der Siemens Velaro Reihe sind reine Siemensprodukte mit einer installierten Basis in Spanien, China und Russland. Neben hoher Sicherheit, Verlässlichkeit und Leistungsfähigkeit besitzt der Velaro eine beeindruckende Energieeffizienz und verbraucht umgerechnet nur etwas 0,33 Liter Benzin pro Sitzplatz auf 100 Kilometer. Ab Herbst 2012 bis 2014 sollen auch in Deutschland 16 Modelle der neuesten ICE Generation in Betrieb genommen werden.

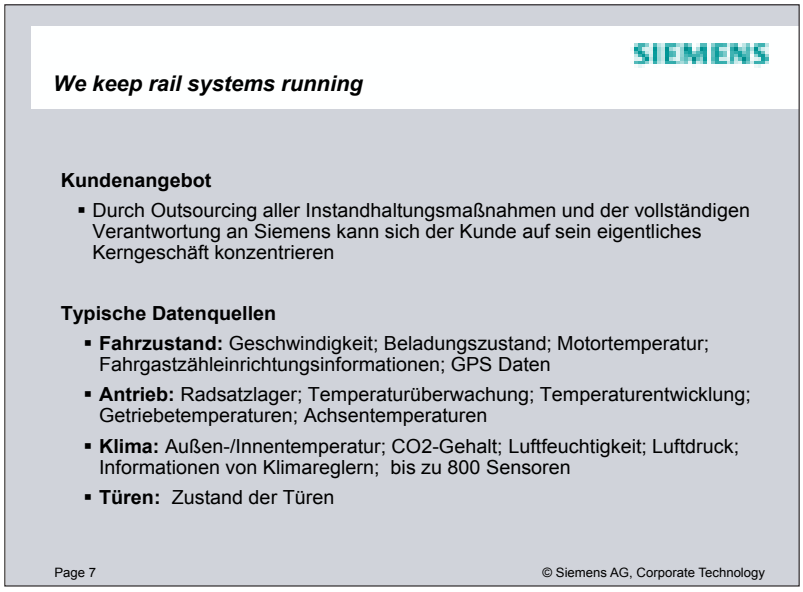

Bild 6

Bei vielen Kunden übernimmt Siemens vollständig die Instandhaltungsmaßnahmen im Betrieb (Bild 6). Hierzu werden eine Vielzahl von Daten gesammelt und ausgewertet, wie Informationen zum Fahrzustand (Geschwindigkeit, Beladungszustand, Motortemperatur, Antrieb, Informationen zu Fahrgastzahlen, GPS Daten), zum Antrieb (Radsatzlager; Temperaturüberwachung; Temperaturentwicklung; Getriebetemperaturen; Achsentemperaturen), zu Klima (Außen-/Innentemperatur; CO_2-Gehalt; Luftfeuchtigkeit; Luftdruck; Informationen von Klimareglern; insgesamt bis zu 800 Sensoren) und zum Zustand der Türen.

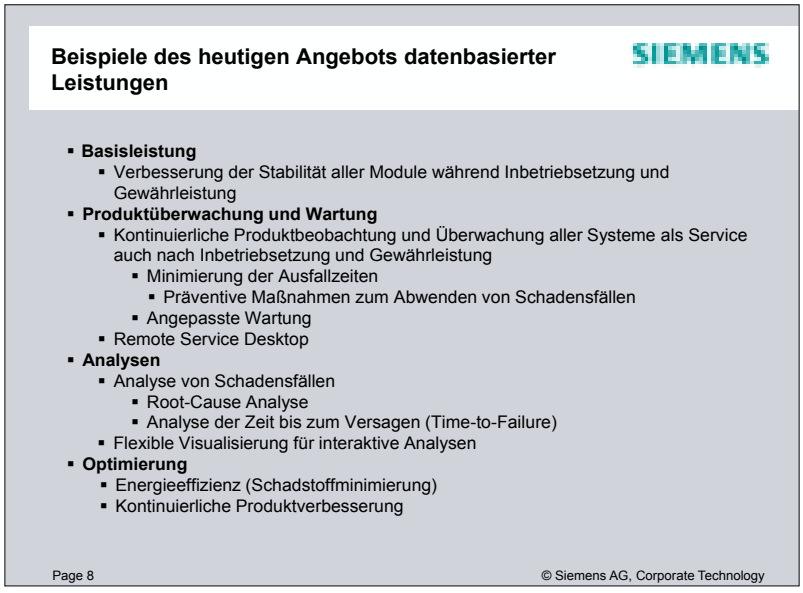

Bild 7

Hier sind einige der Angebote aus dem gegenwärtigen Leistungsspektrum (Bild 7).
Die Basisleistungen, die bei jedem Kunden erbracht werden, umfassen Arbeiten zur Verbesserung der Stabilität aller Module während Inbetriebsetzung und Gewährleistung.
Ein weiterführendes Angebot umfasst die kontinuierliche Produktbeobachtung und Überwachung aller Systeme als Service auch nach Inbetriebsetzung und Gewährleistung. Ziele sind hier eine Minimierung der Ausfallzeiten durch präventive Maßnahmen zum Abwenden von Schadensfällen und eine an den Zustand des Systems angepasste Wartung. Ein Remote Service Desktop ermöglicht einen Gesamtüberblick über die gesamte installierte Basis oder Flotte.

Ein nächstes Leistungsmerkmal betrifft Analysen, wie die Analyse von Schadensfällen durch eine Root-Cause Analyse oder eine Analyse der Zeit bis zum Versagen eines Bauteils (Time-to-Failure). Unterstützt werden diese Leistungen durch eine flexible interaktive Visualisierung.

Ein weiteres Angebot betrifft die Optimierung, z. B. zur Verbesserung der Energieeffizienz, und zur Verringerung von Schadstoffen, als Teil eines allgemeinen Ziel einer kontinuierlichen Produktverbesserung.

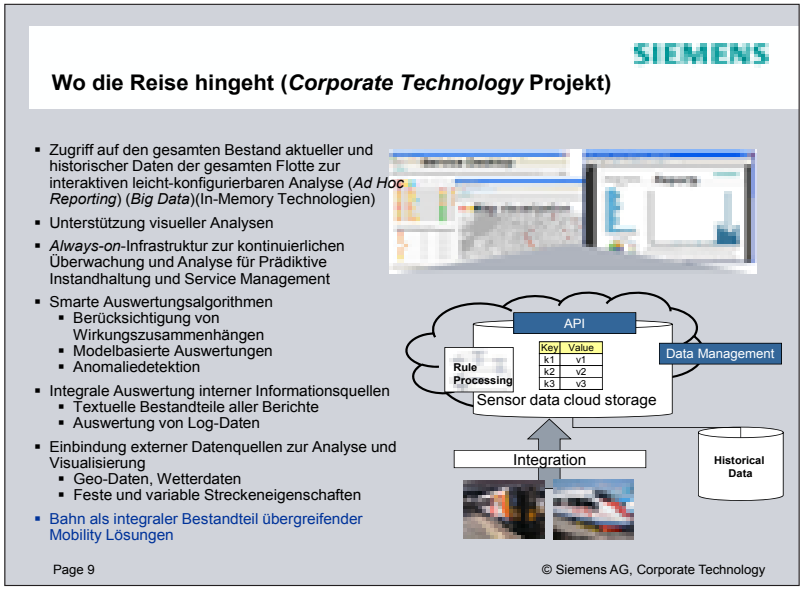

Bild 8

Wo geht die Reise hin (Bild 8)? Hier sind einige Ideen im Rahmen von unserem Projekt aus der Corporate Technology mit der Division.

- Zugriff auf den gesamten Bestand aktueller und historischer Daten der gesamten Flotte zur interaktiven leicht konfigurierbaren Analyse, Stichwort ad hoc Reporting.
- Always-on-Infrastruktur zur kontinuierlichen Überwachung und Analyse. Diese soll es ermöglichen, dass auf Daten von überall auf der Welt jederzeit zugegriffen werden kann und Analysen angestoßen werden können.
- Smarte Auswertungsalgorithmen unter Berücksichtigung von Wirkungszusammenhängen, eine modellbasierte Auswertungen, und Algorithmen zur Detektion von Anomalien.

- Integration weiterer interner Informationsquellen, wie durch eine Auswertung von Textfeldern in Berichten und Log Dateien.
- Einbindung externer Datenquellen zur Analyse und Visualisierung. Hierzu gehören Geo-Daten und Wetterdaten. So können Temperaturanstiege eventuell durch Terraineigenschaften oder Wettereinfluss bereits erklärt werden. Externe Datenquellen sind ebenfalls feste und variable Streckeneigenschaften: Wenn zum Beispiel durch Glatteis ein Schlupf entsteht, erklärt dies möglicherweise ebenfalls ein auffälliges Temperaturprofil.

Natürlich sollte man den Bahnverkehr als integrativen Bestandteil übergreifender Mobility Lösungen sehen, unter Einbindung des kommunalen Verkehrs.

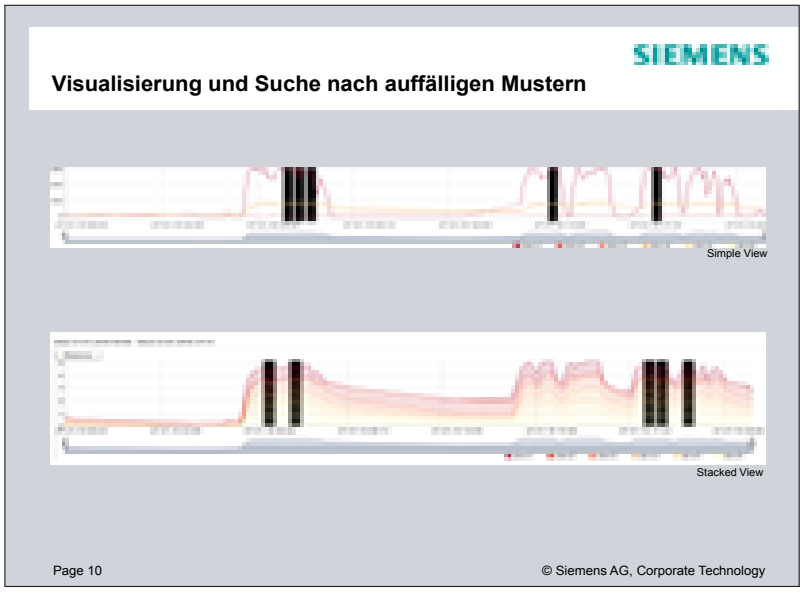

Bild 9

Bild 9 soll einen Eindruck über die aufgenommen Daten ermöglichen. Wir sehen zeitliche Geschwindigkeitsprofile, an denen man die Bewegung eines Zuges gut verfolgen kann, mit Standzeiten und Ruhezeiten. Ebenfalls sieht man Temperaturprofile. Man kann sich nun als Benutzer mit Visual Analytics Anwendung einen bestimmten Bereich herauswählen, der auffällig erscheint, und man bekommt dann sofort automatisch angezeigt, in welchen anderen Zeiträumen ähnliche Auffälligkeiten vorhanden sind.

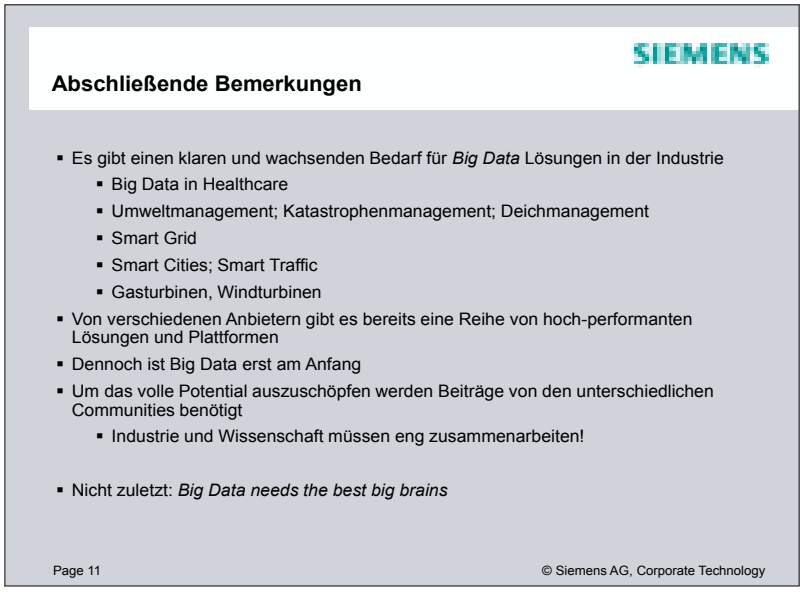

Bild 10

Es gibt aus unserer Sicht einen klaren und wachsenden Bedarf für Big Data Lösungen in der Industrie (Bild 10).

- Gegenwärtig nehmen wir Patientendaten in Kooperationen mit Kliniken systematisch auf, um durch die Auswertung dieser Daten langfristig Prozesse zu verbessern, vergleichende Studien zu ermöglichen und Angebote zur Entscheidungsunterstützung abzuleiten. Abzusehen ist die kommende Datenexplosion in der Translational Biotechnology mit Genom-Daten und Daten aus High-Throughput Experimenten. Die schiere Menge der anfallenden Daten überfordert jeden Arzt und eine Nutzbarmachung verlangt nach einer automatisierten ganzheitlichen Auswertung.
- Ein weiteres Thema ist die zu erwartende Datenflut im Umweltmanagement und im Katastrophenmanagement zum Beispiel in der Überwachung von Deichen. Interessante Möglichkeiten erlauben eine sensorische Auswertung von Smart Phone Daten, z.B. zur Früherkennung von Erdbeben, wo auch wenige Sekunden Vorsprung bereits interessant sein können.
- Smart Metering in Smart Grid ist ein Big Data Thema.
- Smart Cities mit der wachsenden Zahl von Sensoren, die dort zur Verfügung stehen, auch im Zusammenhang mit Smart Traffic und Elektromobilität.
- Dann aber auch klassische Siemens Themen wie Gasturbinen und Windturbinen, die zunehmend mit Sensorik ausgestattet werden

Von verschiedenen Anbietern gibt es bereits eine Reihe von hoch-performanten Lösungen und Plattformen für Big Data. Aber Big Data ist natürlich erst am Anfang. Diese Plattformen sind ein wesentlicher Bestandteil der Lösungen, aber sind noch nicht die Lösung selber. Um das volle Potential ausschöpfen zu können, werden Beiträge von den unterschiedlichen Communities benötigt, sowohl auf der industriellen Seite als auch auf der akademischen Seite, d.h. Industrie und Wissenschaften müssen eng zusammenarbeiten.

Big Data braucht die besten Talente. Google, Microsoft und Facebook sind da durchaus wegweisend und ziehen Top-Talente an über F&E Zentren mit langfristigen Perspektiven, kreativem Freiraum und spannenden Themen. Bei Big Data geht es um Kreativität und ungewöhnliche Ideen. Auch der Erfolg von Google beruht nicht nur auf einer neuen Geschäftsidee, sondern nicht zuletzt auch auf innovativen technischen Ideen, wie dem PageRank Algorithmus.

7 Qualitätsmanagement im Automobilbau ohne Datenanalyse – undenkbar

S. Meinzer, J. Prenninger, A. Deicke, BMW AG, München

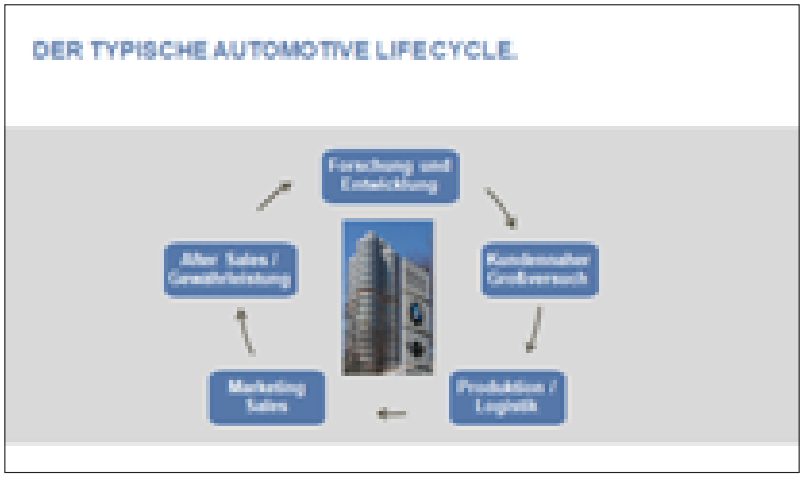

Bild 1

Worum geht es (Bild 1)? Es geht um die Daten aus dem / im Fahrzeug, zur Steigerung der Sicherheit, Kontrolle von Verschleiß, aber natürlich auch Problemidentifikation. Wir reden hier im Wesentlichen von technischen Daten, personenbezogene Daten gibt es in diesem Falle nicht. Das ist auch mit den zuständigen Behörden diskutiert. Ein kleiner Hinweis: wir bekommen die Kundendaten auch gar nicht. Es gibt leider aus BMW Sicht gesehen in der BMW-Händlerwelt 80 DMS, Dealer-Management-Systeme. In diesen Daten, DMS System, ist der Kundendatensatz drin. Auf den haben wir überhaupt Zugriff. Die Werkstattwelt ist hier ganz anders. Davon rede ich. In der Werkstatt gibt es weltweit ein einheitliches System, weil die Spezifika der Werkstattabläufe ganz nahe an der Automobilindustrie hängen. Das dürfte bei allen Fahrzeugherstellern so sein. Von diesen Daten rede ich. Was gibt es bei uns zu sagen? Ich rede nicht über Forschung und Entwicklung. Da war ich vor zehn Jahren. Dass da riesige Datenmengen entstehen, ist jedem klar. Das ist nicht das Neue. Wovon wir jetzt reden, ist, dass wir am Beginn der Kundennutzungsphase (erste Serienfahrzeuge aber vor Kundenauslieferung) einen Großversuch als Neuigkeit eingebaut haben. Auch in der Produktion gibt es inzwischen erhebliche Unterstützungsmöglichkeiten durch Big-Data IT. Bei Marketing Sales kommt jeder gleich auf die Idee, Facebook usw. Dazu sage ich noch etwas und natürlich speziell zu unserer Welt aus Sicht After Sales.

Bild 2

Was hat BMW getan? AVAQS (Bild 2). Jetzt fragen Sie mich bitte nicht, was diese Abkürzung tatsächlich bedeutet. Da hat jemand in Analogie gedacht - man beobachte, was da draußen so alles los ist und wenn dann irgendeiner eine Rakete startet oder in unserem ein Problem hochkommt - dann möchten wir das möglichst früh erkennen.

Was steht dahinter? Letztendlich dem Ingenieur – das ist das Wesentliche hier – einen pragmatischen Zugriff zu geben auf diesen riesigen Datenmengen, die wir gewöhnt sind. Die haben wir schon immer. Im Wesentlichen kommen diese ganz stark aus den Gewährleistungsabrechnungsabläufen oder auch aus Flottenabrechnungen usw. Da gibt es eine ganze Menge von Daten, die geschichtlich schon immer in der Automobilindustrie da waren. Nur sind die üblicherweise sehr schwierig erschließbar oder gar korrelierbar. Sie kennen alle hier ITPM. Der Ingenieur, der ganz schnell ein IT-System haben will, um diese Daten interpretieren zu können, ärgert sich, wenn er ITPM hört, weil das für Ihn bedeutet, dass er in 3 Jahren oder in 5 Jahren für soundso viel Euro einmal ein intelligentes System bekommt, das ihm irgendwelche Daten, zum Beispiel aus dem ICE, in der richtigen Art und Weise auswertet.

Das Neue bei AVAQS ist, dass es eben nicht so ist. Hier gibt es intelligente Plattformen der IT Industrie, die es uns erlauben, verschiedenste Datensätze miteinander zu korrelieren. Natürlich brauchen wir dazu einen wissenden Menschen. Sie kennen den schönen Satz über der Elektrik Ingenieure: „Wer misst, misst Mist“. Also, wer Unsinn in die Datenauswertung eingibt, der bekommt natürlich auch hinten nur Unsinn heraus. Deswegen brauchen wir beide. Wir brauchen den Ingenieur, der weiß, wovon er redet und wir brauchen den Datenanalytiker, der weiß, was rausgenommen werden muss, Null Kilometer Daten zum Beispiel in diesem Falle.

Was sind das für Datenmengen? Wir haben heute schon viele Terabyte/Petabyte. Ich habe jetzt gerade Zetabyte und so etwas gelernt, so weit sind wir noch nicht. Petabytes. Das sind schon ganz heftige Datenmengen. Seit ungefähr zehn Jahren werden im Prinzip alle Fahrzeuge bei uns abgespeichert. Nicht die aktuellen Daten, natürlich bereinigt und komprimiert. Das ist einfach speichertechnisch nicht machbar.

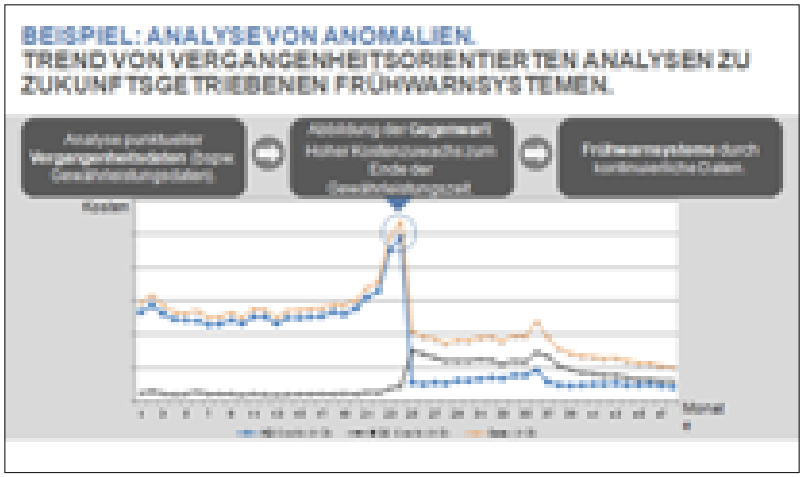

Bild 3

Wie schaut es traditionell aus (Bild 3). Traditionell heißt das ganz einfach, dass Sie alle wunderschön fahren und hoffentlich kein Problem haben. Und wenn Sie ein Problem haben, dann kommen Sie in die Werkstatt, und das tun Sie natürlich immer spätestens vor Ablauf der Gewährleistung. Danach wird es etwas weniger. Auf diesem Weg können wir auf die Daten immer nur dann zugreifen, wenn Sie in der Werkstatt sind. Beispiel ICE – eine kontinuierliche Datenübertragung wäre ein Traum für den Ingenieur und vor allem in der Phase der Erprobung – ich rede jetzt nicht von der reinen Engineeringerprobung im Sinne Prototypen, die in Lappland oder wo immer fahren, sondern seriennah, kundennah. Hier kommen wir eben leider erst, wenn Sie in die Werkstatt kommen, an diese im Fahrzeug abgespeicherten Informationen.

Was steht da drin? Es steht nicht der Ort drin, wo das Fahrzeug war, keine Bewegungsdaten. Aber da steht z. B. Ölwechsel, Verschleißdaten der Bremsen, Busdaten, Batteriezustände – ein ganz kritisches Thema, auf das ich gleich noch komme.

Wie sieht das in Zukunft aus? Ich spreche immer in Bezug auf kundenahen Großversuch mit internen Fahrzeugen und was man davon auch im Kundenbetrieb nutzen kann bzw. könnte. Da kann man sich vieles denken, was aber heute nicht getan wird.

Bild 4

Jedes Fahrzeug erzeugt Daten (Bild 4). In den Fahrzeugbordnetzen sind verschiedenste Busse enthalten, langsame, schnelle, deterministische und kontinuierliche Daten. Die Kunst des Themas ist, die Daten im Fahrzeug so vorzustrukturieren, dass ich einerseits keine Daten, keinen Informationsinhalt verliere und andererseits natürlich möglichst konkret Probleme identifizieren kann. War jetzt die Batterie leer oder nicht, wenn das Fahrzeug nicht anspringt? Das ist natürlich eine wesentliche Information. Warum war sie leer? Weil der Lichtschalter an war? Das ist geht übrigens im BMW gar nicht mehr. Wenn er neu genug ist, können Sie das Licht anlassen, denn er schaltet es von selber ab.

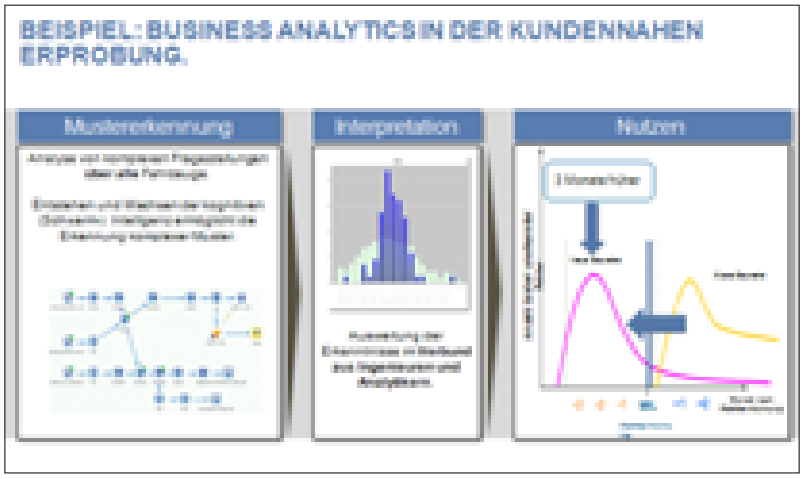

Bild 5

Das ist eine große Herausforderung an die Ingenieure, hier Informationen im Fahrzeug so bereitzustellen, dass wir damit später auch etwas anfangen können (Bild 5). Diese auf GSM oder UMTS zu übertragen, ist natürlich kein Thema. Das wissen Sie alle hier. Die dann zu korrelieren mit den verschiedensten Datenbanken, die bei uns vorhanden sind - im Wesentlichen fahrzeugbezogene Themen wie Konstruktionsdaten, Baudaten. Jedes Fahrzeug ist verschieden. Da sind Teilerückverfolgungsdaten drin usw. Dann kann man korrelieren.

Das Fahrzeug mit dem entsprechenden Baustand hat dieses oder jenes Problem. Das wird intern bei uns natürlich stark genutzt für das Engineering. Das sind im kundennahen Großversuch, der KMG Phase, die ersten wirklichen Serienfahrzeuge, die vom Band kommen. Ich fahre solche Fahrzeuge immer gern, weil ich dadurch das neueste habe. Diese Fahrzeuge sieht aber nie ein Kunde, weil sie verschrottet werden. Es handelt sich um Größenordnungen zwischen 50 und 100 manchmal sogar 500 Fahrzeuge. Zusätzlich werden dann auch die ersten Fahrzeuge für die Dealer-Drive-Events, die Vorstellungsfahrzeuge für die Händler – auch einige 100 Fahrzeuge –, mit dem gleichen System intern überwacht.

Wenn man dann an echte Kunden denkt, kann man theoretisch die Datenübertragung auch verwenden. 40 % unserer Fahrzeuge haben die technischen Voraussetzungen dafür im Fahrzeug. Das wird aber ab Band ausgeschaltet. Was ist ab Band eingeschaltet? Das sind die Themen, die Sie zu Ihrem Nutzen brauchen. Wenn Sie sich z.B. die BMW Remote APP aus dem Apple Store runterladen, können Sie schauen, wo denn Ihr Fahrzeug ist. Auf manchen amerikanischen Großparkplätzen mag das ganz interessant sein, es wiederzufinden.

Ein ganz wichtiges Thema für uns ist hier der Datenschutz; der Kunde bestätigt mit dem Runterladen dieser Applikation und dem entsprechenden Eintrag in unsere Datenbank sein Einverständnis. Damit ist eine ganz wesentliche Datenschutzanforderungen erfüllt. Wir sehen diese Daten wieder nur anonym – auf das Fahrzeug bezogen, auf Namen und Adressen haben wir keinen Zugriff.

Ein ganz wichtiger Anwendungsfall tritt ein, wenn Sie liegenbleiben, was Ihnen hoffentlich nie passiert. Was ist die häufigste Liegenbleibeursache? Die häufigsten Ursachen sind ein leerer Tank, bzw. eine leere Batterie. Erstaunlicherweise sind wir alle nicht in der Lage, die vielen Lämpchen im Auge zu behalten. Es ist für den mobilen BMW Service natürlich von entscheidender Bedeutung, ob er mit dem Abschleppwagen hinfahren muss oder ob es vielleicht ausreicht, einen 5-Liter Kanister Benzin zu dem Kunden zu bringen. Deswegen ist einer der wesentlichen Daten, die hier übertragen werden, der Tankinhalt. Zum Thema Batterie haben wir in den neuen Fahrzeugen auch einen Batteriemonitor drin, der die Batterie überwacht, die sogenannten State of Charge, SoC, und State of Health, SoH Daten, der dann hoffentlich früh genug meldet, ob diese Batterie den nächsten Winter überlebt oder nicht. Das ist noch in der Erprobung und nicht so ganz banal, denn ob dieses Fahrzeug in Florida, Griechenland oder in Skandinavien fährt, hat bei einer Meldung der Batterie am 1.November noch eine erhebliche unterschiedliche Bedeutung, sprich: wir brauchen intelligente Datenverarbeitung dahinter, nicht im Fahrzeug sondern dahinter. Dann bekommen Sie die Information „Achtung, diese Batterie wird voraussichtlich den nächsten Winter nicht überleben“. Die Entscheidung was zu tun ist, liegt natürlich beim Kunden, dazu geben wir nur eine Empfehlung.

Wie läuft das? Den Datenmenschen hier ist das sicherlich viel klarer als mir. Ich sehe das aus der Anwendungssicht. Ich bin Ingenieur, Elektrotechniker. Über die verschiedensten Methoden werden die Peaks während des KMGs raus gefiltert. Da sitzen wirklich typischerweise am Abend Ingenieure da und checken während eines Dealer Drive Events alle Fahrzeugdaten, meistens so 300 bis 400, die über drei bis vier Wochen fahren und allen Händlern vorgestellt werden. Innerhalb dieses Zeitraums werden jeden Tag z.B. beim Anlassen oder wenn während der Fahrt irgendwas passiert, die Daten übertragen und am Abend laufen Checkprogramme um Auffälligkeiten zu identifizieren.

Das sind Fahrzeuge nach Serienanlauf, nach SoP, Start of Production, die allerdings nicht verkauft werden. Wir sind inzwischen in der Lage, hier auch sehr schnell zu reagieren. Ein Teil der Fehler ist aus dem E/E Bereich, Electric-Electronic, oft Software.

Warum geht es letztendlich? Sie wollen als Kunde keine Kinderkrankheiten haben. Und wir wollen das auch nicht, weil es uns Geld kostet, wenn diese Fahrzeuge mit diesen Fehlern zu früh im Feld auftauchen. Also haben wir gemeinsam nichts anderes als das Interesse, den Fehler so schnell wie möglich vor dem ML – Model Launch bereits zu identifizieren und natürlich dann auch zu beseitigen.

Das geht im Softwarebereich sehr schön. Aber auch im Prozessbereich, in den Werken kann man eine ganze Menge machen. Denken Sie an Klappern, wo dann Filz verklebt wird oder andere kurzfristige Maßnahmen durchgeführt werden oder Werkzeuge geändert werden, um ganz schnell aus den Problemen herauszukommen. Das sind Datensätze um die 5000 pro Tag während dieser Phase. Ein Datensatz ist relativ klein, es sind Datenmengen im Kilobytebereich/Fahrzeug und Event. Die Menge entsteht über die Anzahl der Fahrzeuge bzw. Anzahl der Events.

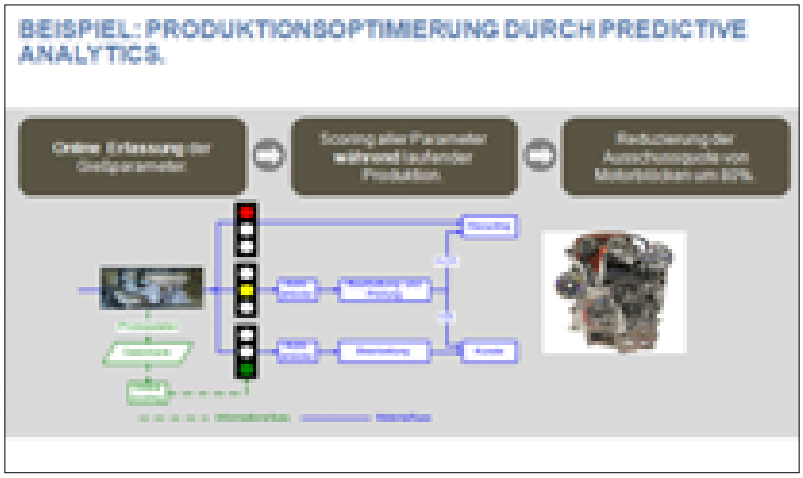

Bild 6

Die nächste Phase ist die Produktion (Bild 6). Der Motor ist ein schönes Beispiel. Keiner von uns möchte einen Motordefekt haben. Ein Motoreffekt kann z.B. entstehen durch Gussfehler. Der schöne große Motorblock dort vorn aus Aluminium, ein schönes großes Gussteil von 250 Kilo beim V8. Wenn da irgendetwas bei der Produktion des Gusses nicht stimmt, dann bekommt der Motor irgendwann einmal ein Problem. Er verzieht sich oder bekommt Risse. Heute wird jeder Motor geröntgt. Röntgen kostet Zeit und Aufwand, 7 Minuten. Die Anlagen in der Multiplikation sind schlimm genug. Invest ist in der Automobilindustrie aber nicht unbedingt das ganz große Thema, aber Prozesszeit – 7 Minuten pro Motor. Und das bei 7000 Motoren pro Tag.

Wenn hier oben steht 80 % reduzieren, dann heißt das nicht, dass 80 % der standardmäßigen Produktion Ausschuss gewesen wäre. Wir reden hier von Prozent, von kleinen Zahlen. Diese gilt es auch noch rauszuholen, indem der Prozess im Gießprozess von vornherein so überwacht wird, dass gar kein Schrott entstehen kann. Das ist ein großes Thema, über das eine Promotion geschrieben wurde, um hier noch einmal ganz erhebliche Effizienzen herauszuholen und dem Kunden einen Motor zu bieten, der auch weit über den Gewährleistungs/Kulanzzeitraum hinaus problemlos läuft und um intern natürlich Kosten zu drücken.

Gewährleistungskosten sind in der Automobilindustrie, wir haben gerade die Bilanz 2011 veröffentlicht, ein sehr großes Thema. Jeder Euro, der hier investiert wird, um GW-Kosten zu vermeiden, ist natürlich gutes Geld.

Bild 7

Mit meinem nächsten Beispiel kommen wir wirklich zu den Kunden (Bild 7). Jetzt wollen ihm wir ein Auto verkaufen. Wenn Sie heute in den BMW Internet-Konfigurator gehen, wird der Ihnen keinen Spaß mehr machen. Er ist alt und entspricht nicht – ich sage es mit Absicht - den modernen Spielanforderungen. Jeder von uns möchte ein bisschen Spaß haben im Internet. Deswegen gibt es jetzt einen neuen Konfigurator. Es klingt ein bisschen banal diese Bildchen anzuzeigen, ist es aber nicht.

Wir reden von 15 Basismodellen. Ein Modell ist, was Sie hier umgangssprachlich als Karosserievariante beschreiben. Ein 3er, 7er, X5 sieht schon einmal anders aus und das ist logischerweise ein entscheidender Unterschied. Denken Sie weiter an Farbvarianten, Ausstattungsvarianten, Sonderausstattungen. Wir haben ca. 2000 Sonderausstattungen. Eine Sonderausstattung ist wiederum abhängig vom Zulassungsland. In einemLand heißt Fernbedienung 435 MHz, im anderen 315Mhz usw. Das ergibt in der Kombinatorik und dann auch noch abgelegt in die Ländervarianten – es macht keinen Sinn, einem chinesischem Kunden einen Diesel anbieten zu wollen, den es da einfach nicht gibt – ein extrem komplexes Werk. Ich glaube, dass es sehr gut gelungen ist. Sie können es sich heute schon im Pilotbetrieb unter bmw.at anschauen und genießen, dass Sie dann wirklich relativ schnell Ihr Fahrzeug konfiguriert haben.

Warum ist das für uns so wichtig? Unser Business Modell ist das individuelle Fahrzeug, Ihr Fahrzeug. Natürlich gibt es auch Flottenfahrzeuge, wo einer 10 Stück in Grün bestellt. Das gibt es bei BMW natürlich auch. Unsere ganz große Stärke liegt aber darin, Ihnen Ihr ganz spezielles Fahrzeug anzubieten. Das sehen Sie im Internet. Sie können es sich vorher anschauen, drehen und wenden, ändern und träumen. Sie können natürlich die Daten dann auch Ihrem BMW Händler der Wahl Daten schicken. Weit mehr als der vorhin angesprochene – langweilige - Katalog.

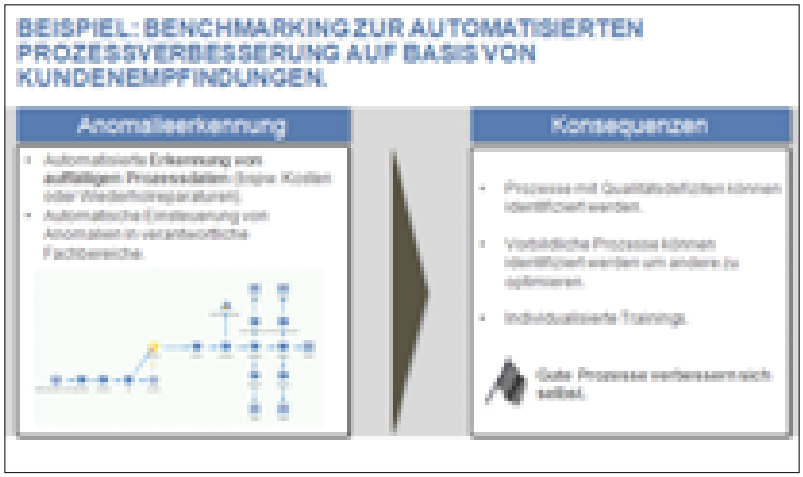

Bild 8

Mein nächstes Thema ist die IT (Bild 8). Wir reden von Kundendaten, aber wir haben keine Adresse. Es gibt in der Automobilindustrie standardmäßig so genannte Clubabfragen. Da wird von einer unabhängigen Organisation die Kundenzufriedenheit abgefragt. Diese Daten bekommen wir nur anonymisiert. Aber das ist für uns eine **der** Datenquellen überhaupt, um Kundenzufriedenheit zu analysieren. Wenn Sie - ich hoffe, dass Ihr Händler es tut - bei einem BMW Händler reparieren oder die normale Wartung hatten, dann sollte er Sie hinterher anrufen. Das geht nicht zu 100 %, aber ungefähr 40 bis 50% aller Kunden werden hinterher angerufen.

Warum tun wir das? Kundenzufriedenheit ist ein enorm wichtiges Thema. Natürlich wollen wir wissen, ob der Auftrag zur Zufriedenheit ausgeführt wurde, aber vor allen geht es um die Wiederholreparaturen. In der Automobilindustrie ist das größte Thema der Kundenunzufriedenheit, dass Sie das Fahrzeug hinbringen und sagen, was gemacht werden muss. Sie bekommen das Auto zurück, und es war nichts. Dieser Kunde ist garantiert verärgert. Und wenn Sie das zweimal mit ihm machen, ist es ganz aus. Wir sagen so ganz einfach: „dass das erste Fahrzeug vom Sales verkauft wird, das zweite Fahrzeug verkauft der Servicemitarbeiter“. Der gesamte Druck der Automobilindustrie gilt der Vermeidung von Wiederholreparaturen.

Jetzt kommt leider ein kleines Problem, was hier unten steht, Gefühle, Meinungen. Was der Kunde sagt, was eine Wiederholreparatur ist, hat unter Umständen überhaupt nichts damit zu tun, was technisch dahinter war. Der schlimmste Fall in unserem Verständnis ist, dass der Kunde wiederholt hin musste. Warum? Der Ölwechsel war fällig. Danach mussten die Winterreifen gewechselt werden. Der dritte Termin war die Rückgabe des geleasten Wagens. Innerhalb von sechs Wochen macht das drei Termine, der Kunde ist verärgert, aber die Lösungsansätze liegen in ganz anderen Bereichen als der reinen Fahrzeugtechnik. Hier die Kundenerwartungen herauszufinden und das auch noch mit Ländererwartungen zu korrelieren, ist uns wichtig. Dieses Jahr wird China vermutlich der größte Markt der Welt für BMW sein, aber die Kunden dort haben kein über Jahrzehnte erlerntes Gefühl für Autos. Wenn chinesische Kunden Fehler beschreiben, dann tun sie das in anderer Art und Weise, die nichts mit dem zu tun hat, wie wir die alle mit Autos aufgewachsen sind schon als Jugendliche. Daran muss man denken. Hier bieten die IT-Technologien uns jetzt Möglichkeiten, endlich auch voranzukommen. Was wir uns natürlich vorstellen könnten, wäre eine Video-

auswertung, wie vorhin angesprochen. Das haben wir aber leider noch nicht. Da funktionieren die Datenmengen leider nicht.

After Sales: Gewährleistung habe ich schon erwähnt. Hier ist unser Repeat Repair Thema. Dr. Watson wurde schon angesprochen. Wir haben mit IBM gemeinsam einen Vorläufer, eine Visibility-Studie laufen, was man darüber tun wird. Die Idee ist einfach, dass heute eine Reparaturanleitung, ein Reparaturprozess einmal von einem Ingenieur beschrieben wird. Ob 50.000 Mechaniker in der Welt das mit der Zeit nicht besser wissen? Ich behaupte jedenfalls, dass das so sein muss. Je älter das Fahrzeug wird, je größer die Stückzahlen sind, desto sicherer wird dieses Statement richtig sein. Wir haben heute aber keinerlei Möglichkeit, diese Daten zurückzuführen. Wie sagt ein Mechaniker Ihnen, der unter anderen Bedingungen, weil die Lohnstrukturen anders sind oder wie auch immer, dass eine Reparatur vielleicht wesentlich kostengünstiger dargestellt werden könnte. Wenn Sie ein 10 Jahre altes Fahrzeug haben, dann fragen Sie das Internet. Das weiß üblicherweise, was mit Ihrem Fahrzeug los ist. Das ist mir gerade selber passiert mit einem 15 Jahre alten Fahrzeug.

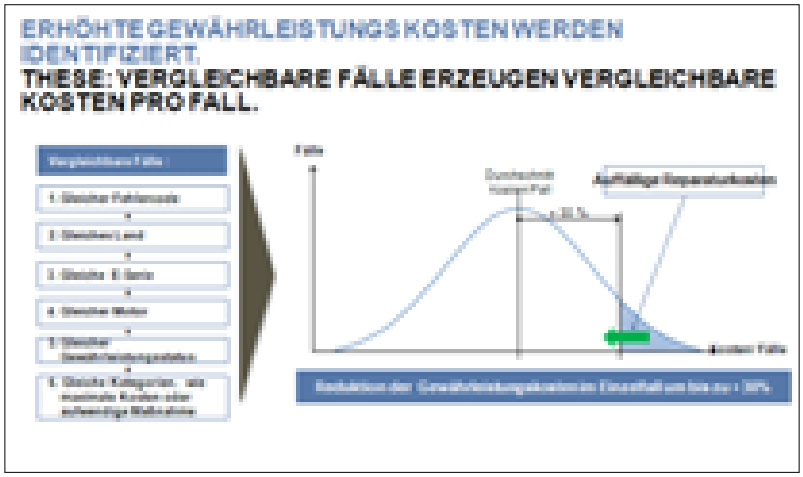

Bild 9

Worum geht es hier? Wieder die Gewährleistungskosten (Bild 9). Es gibt ganze Abteilungen, die mit entsprechender Analysesoftware diese Fälle hier anschauen. Irgendetwas ist schief gegangenen? Die gleiche Reparatur sollte eigentlich, da unsere technischen Vorgaben, Medien, Abläufe, Prozesse, Tools weltweit 100 % einheitlich sind, in etwa immer gleich lang dauern. Wir haben keine in irgendeiner Form länderangepassten Reparatursysteme. Man sollte also meinen, dass es in irgendeiner Form eine Gaußsche Glockenkurve wäre, in der die Reparaturzeiten liegen. Zeit ist Geld. Es gibt aber immer Auffälligkeiten, wo etwas nicht funktioniert hat. Der Prozess ist nicht sauber beschrieben oder ein Händler oder viele haben gemerkt, dass man das besser machen kann. Letztendlich die Gewährleistungskosten zu reduzieren, ist natürlich für uns das A und O.

Vielleicht als kleine Zahl, die Sie leicht selber nachprüfen können. Wenn Sie die Bilanzen der letzten sieben, acht Jahre anschauen, dann steht da in etwa immer die gleiche Gewährleistungskostenzahl drin. Wenn Sie aber auch wissen, dass sich bei BMW die Stückzahlen verdoppelt haben und andererseits der Gewährleistungszeitraum in vielen Ländern verdoppelt wurde, wissen Sie, was das bedeutet.

Bild 10

Ich will nicht sagen, dass das nur bei BMW so ist, die gesamte deutsche Automobilindustrie ist besser geworden ist (Bild 10). Aber hier sind durch moderne Methoden wirklich Welten passiert, um letztendlich dem Kunden zu helfen. Um den geht es. Der kommt einfach nicht wieder, wenn wir ihn nicht entsprechend bedienen. Wir müssen sein Kundenfeedback noch vielmehr bearbeiten, und dazu brauchen wir Ihre Hilfe. Dazu gehört z.B. auch bei Modelleinführungen, Facebook zu befragen wie der neue 3er im Moment z. B. gefällt. Das wird natürlich gemacht und auch entsprechend darauf reagiert, welche Kritik auch immer da war. Manchmal geht es erst in drei Jahren, wenn dann eine Modellüberarbeitung an das Band kommt. Unser großer Kampf Gewährleistungskosten und in der Fertigung beherrschte Prozesse. Irgendwann heute Morgen wurde gesagt: ungefähr 5 % Effizienzsteigerung pro Jahr durch Big Data und ich bin sicher, dass das stimmt.

8 Semantische Technologien und Maschinelles Lernen zur Verwaltung und Nutzung von Big Data

Prof. Dr. Rudi Studer, Karlsruher Institut für Technology (KIT),
FZI Forschungszentrum Informatik, Karlsruhe

Von meinem Hintergrund her ist es nicht überraschend, dass bei mir Semantik eine Rolle spielt, weil das ein Thema ist, das wir seit Ende der 90er Jahre sehr massiv im Karlsruher Umfeld in verschiedenen Institutionen betreiben. Ich möchte auch aufzeigen, wie diese Methoden mit Lernverfahren in letzter Zeit an manchen Stellen sehr gute Synergien liefern.

Nach einer kurzen Einleitung will ich in drei Themengebiete einsteigen. Zum einen möchte ich darüber sprechen, was wir im Bereich Datenintegration auf einer eher semantischen Ebene machen können. Wie können wir die Vielzahl von Daten, die überall vorhanden sind, zusammenbringen? Wie können wir bei solchen großen heterogenen Datenmengen interessante Dinge in diesen Daten finden? Diese Fragestellung betrachten wir unter dem Stichwort Semantic Search. Dann möchte ich aufzeigen, dass wir bei unstrukturierten textuellen Daten einen Mehrwert durch Kombination mit den Lernverfahren erzielen können. Das ist mein Programm für den heutigen Vortrag.

Ich will nichts mehr über große Datenmengen im Allgemeinen sagen, denn für uns ist eher relevant, dass die Menge der strukturierten, semantisch beschriebenen Daten im Webkontext aufgrund vielerlei Entwicklungen signifikant zunimmt. Mit am prominentesten ist die Entwicklung, die unter dem Stichwort Linked Open Data (LOD)-Initiative bekannt ist, die seit einigen wenigen Jahren am Laufen ist und eigentlich zu einer zweiten Charakterisierung des Web führt. Klassischerweise reden wir vom Web als einer immensen Sammlung von Dokumenten, die untereinander verlinkt sind. Die Entwicklung mit Linked Open Data führt aber zu einem Web of Data-Begriff, bei dem eine Vielzahl von strukturierten Daten miteinander verknüpft sind.

Was können wir durch diese Datenintegration an Mehrwert erreichen? Wie können wir gerade durch semantische Suche auch Anfragen an großen Datenbeständen bearbeiten, die mit dem Stichwort Long-Tail Queries charakterisiert werden? Dazu werde ich noch etwas sagen. Wie können wir sowohl durch semantische Modellierung im Hintergrund wie auch durch passende Lernverfahren neue Beziehungen in diesen Daten auffinden?

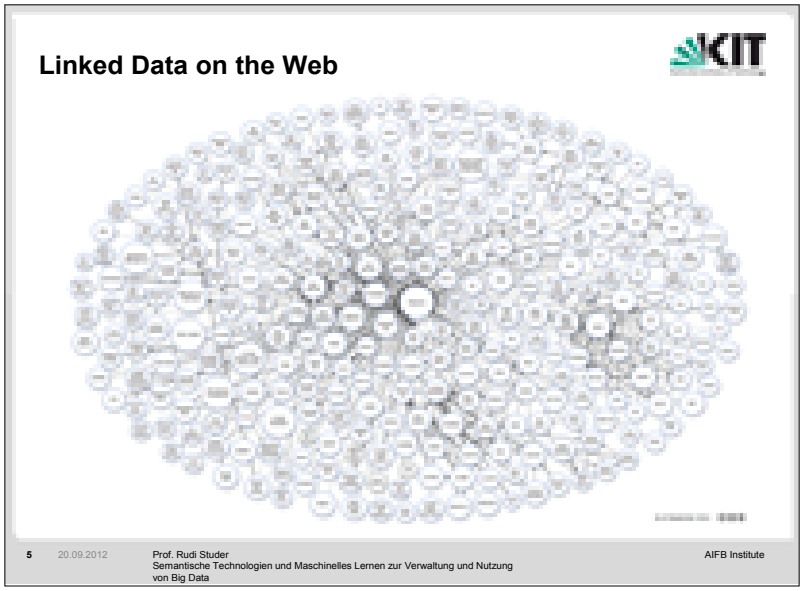

Bild 1

Wenn wir uns die Entwicklung von Linked Data im Web anschauen (Bild 1), so war das eine Initiative, die eigentlich einmal damit startete, dass man die Fakten, die man in Wikipedia in diesen kleine Boxen auf der rechten Seite findet, in einer strukturierten Art und Weise bereitstellen wollte, um sie besser im Datenbanksinne zugreifbar, verarbeitbar zu machen. Das führte zu dieser DBpedia-Entwicklung, die insbesondere aus Berlin durch einige Kollegen vorangetrieben wurde. Es entstand eine rasante Entwicklung, in der eine Vielzahl von Datenquellen sukzessive bereitgestellt wurden, die auf dem RDF-Datenformat basieren und dann mit den bereits existierenden Datenbeständen verknüpft wurden. Das ist ein sehr explosionsartiger Prozess, sowohl was die Anzahl der Daten als auch was die Anzahl der Datenverknüpfungen angeht.

Wenn man das aktuelle Bild vom Herbst vergangenen Jahres anschaut, dann sehen Sie diese Linked Data-Wolke mit Wikipedia im Zentrum. Wenn Sie in die Details gehen würden, würden Sie sehr viele Gouvernment-Daten finden, die in den USA oder UK durch die aktuellen Regierungen gepusht, veröffentlicht werden. Das Ziel ist es, viele Daten, die man in den Ministerien sowieso parat hat, öffentlich zugängig zu machen und zu einer weiteren Verarbeitung, sei es im kommerziellen oder nicht kommerziellen Umfeld, zur Verfügung zu stellen. Wenn Sie dort oben hinein zoomen würden, dann würden Sie Dinge aus dem kulturellen Bereich, Musik, Film in dieser Ecke finden. BBC stellt eine ganze Menge an Daten bereit und ist sehr prominent vertreten.

Unten würden Sie viele Daten im Bereich Life Sciences finden, ein Bereich, der in den letzten Jahren sehr stark in die Daten-getriebenen Analysen hineingegangen ist. Für die Uni-Welt würden Sie im rechten Teil der Linked Data-Wolke viele Quellen über Publikationen finden.

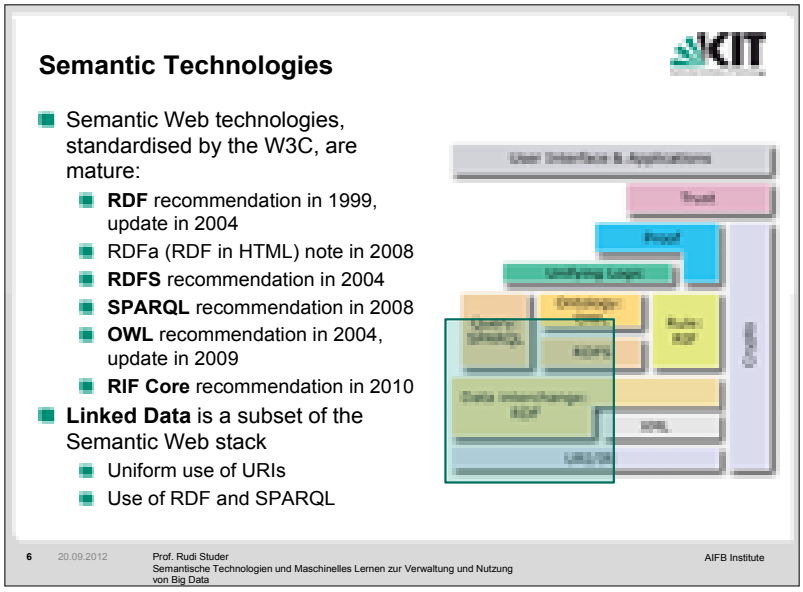

Bild 2

Was sind die Standards, die u.a. bei Linked Open Data eine Rolle spielen (Bild 2)? Manche Leute reden hier vom Semantic Web-Stack, der von Tim Berners-Lee als Begriff eingeführt wurde. Wenn wir uns das klassische Web anschauen, haben wir die URIs für die Identifikation der Dokumente und Entitäten auf dem Web, haben XML, um Inhalte zu beschreiben und miteinander zu verknüpfen. Was Sie hier farbig sehen, sind die semantischen Beschreibungsebenen, die in den letzten sieben, acht Jahren insbesondere auch vom W3C, der Standardisierung zugeführt wurden. Da haben wir als Basis-Datenmodell das Ressource Description Framework, ein grafbasiertes Datenmodell, das heutzutage die Datengrundlage für Linked Data ist und das seit 2004 ein W3C Standard ist.

Wenn man eingeschränkte konzeptuelle Strukturen bereitstellen möchte, gibt es - aufbauend auf RDF - RDF Schema, was uns im Wesentlichen erlaubt, taxonomische Strukturen und ein paar Querbeziehungen zu definieren. Das ist der erste Einstieg in die Ontologiewelt, die dann komplett durch die Ontologiesprache OWL abgedeckt wird. OWL basiert auf Beschreibungslogiken in verschiedenen Nuancen und stellt uns reichhaltige Modellierungskonstrukte bereit, um Zusammenhänge in Anwendungsdomänen zu definieren.

Ganz wichtig für unsere spätere Diskussion ist die Anfragesprache SPARQL, die es uns erlaubt, RDF-basierte Inhalte anzufragen, analog zu dem, was Sie aus der Datenbankwelt kennen. Wenn wir von relationalen Datenbanken sprechen, kennt jeder SQL als Anfragesprache, um diese Datenbankinhalte abzufragen. SPARQL ist die Abfragesprache, die es uns erlaubt, auf diese grafbasierten Daten im ähnlichen Sinne zuzugreifen. Als eine der jüngsten Standardisierungen haben wir das RIF, Rule Interchange Format, das dazu dient, in eingeschränktem Maße regelhafte Beschreibungen in einer standardisierten Art und Weise bereitzustellen.

Weshalb erwähne ich das? Für viele Leute ist die ganze Semantikwelt immer noch mit der Idee verbunden, dass sie in der Forschung recht und schön sei. Kann man sie aber in der Praxis einsetzen? Ein wichtiger Schritt sind die verschiedenen Sprachen, die in den letzten

Jahren durch W3C standardisiert worden sind und die Basis für die Entwicklung zugehöriger Tools bilden.

Was hat das mit Linked Data zu tun, wo Sie vorher diese schöne Linked Data-Wolke gesehen haben? Linked Data basiert vom Datenformat her zu 100% auf dem Ressource Description Framework. Diese vernetzten grafbasierten Daten können durch die SPARQL-Anfragesprache angefragt werden, so dass man da auch ein definiertes Sprachmittel hat, um Inhalte aus diesen RDF-basierten Quellen herauszuziehen. So touchiert man manchmal etwas die Ontologiewelt, die an dieser Stelle aber nur sehr rudimentär ausgeprägt ist. Deshalb ist der Ist-Zustand, den wir erreicht haben, durch wohldefinierte Standards und Datenquellen charakterisiert, die, was Anzahl der Quellen und Umfang der Inhalte angeht, in den letzten Jahren sehr stark explodiert sind.

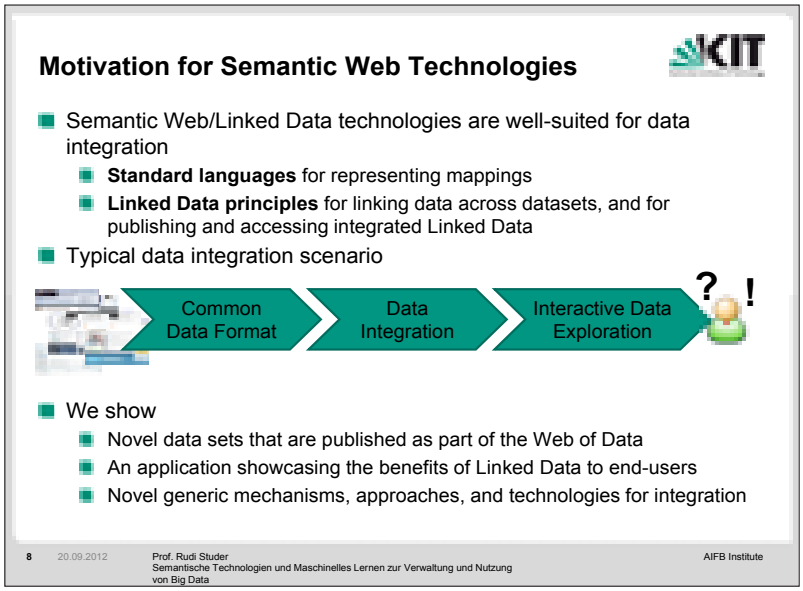

Bild 3

Damit komme ich zu meinem ersten Thema, wie wir jetzt Datenintegration auf einer eher semantischen Ebene machen können (Bild 3). Wichtig ist zum einen, dass wir Standardsprachen haben, aber nicht nur, um die einzelnen Quellen zu beschreiben, sondern auch um Beziehungen zwischen diesen einzelnen Quellen herzustellen. Wir hatten gesehen, dass diese Datenquellen miteinander verknüpft sind. Die Frage ist, wie wir diese Verknüpfungen beschreiben können. Dafür können wir diese standardisierten Sprachen benutzen.

Was für eine Vorgehensweise haben wir? Wir können Datenquellen in Standarddatenformate transformieren, sie über die Bereitstellung entsprechender Abbildungen zwischen diesen Datenfeldern integrieren und interaktiv auf diese Daten zugreifen. Was ist unter dem Anwendungsgesichtspunkt daran interessant? Zum einen, dass wir inzwischen eine Vielzahl von Quellen vorfinden. Dann können Sie sehen, wie man diese Daten mit gängigen Applikationen verknüpfen kann und was für neuartige Ansätze wir benötigen, um mit diesen Daten umgehen zu können.

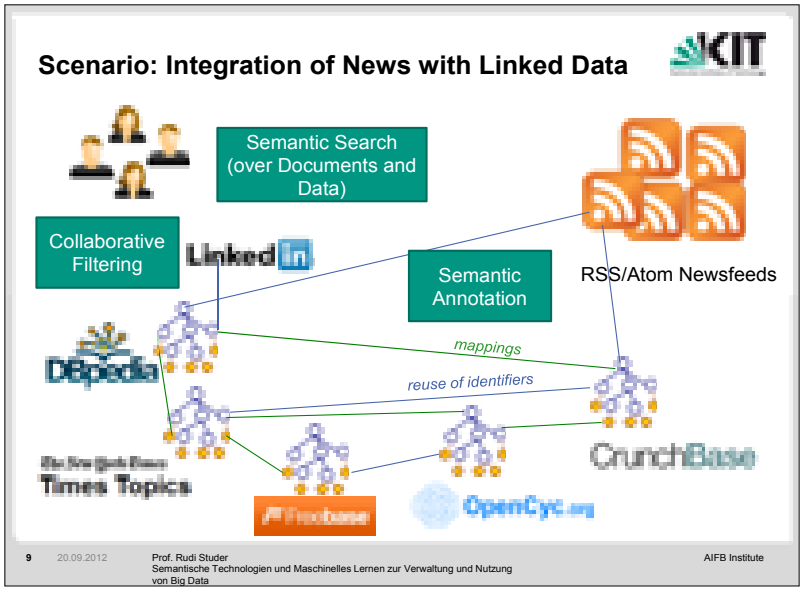

Bild 4

Ich will Ihnen ein kleines Szenario aufzeigen (Bild 4), damit Sie sehen, was wir mit dem Mehrwert meinen, den wir über dieses Web of Data, das wir als Datenquelle parat haben, erzielen können. Hier sehen Sie eine einfache News-Anwendung, in der Sie es gewohnt sind, eine Vielzahl von News zu bekommen, sei es im Finanzbereich oder anderen Themenbereichen. Die Frage ist immer, was ich an Inhalten aus diesen textlichen Nachrichten, die irgendwo auf der Welt mit den Kunden, mit den Lieferanten, ausgetauscht werden, an Analysen herausziehen kann. Heutzutage kann man gut Entitäten in solchen Quellen erkennen. Wenn wir als Beispiel die IBM nehmen, ist es sehr leicht aus solchen textlichen Quellen zu erkennen, dass da IBM als Unternehmen angesprochen wird. Ich könnte z.B. eine Verknüpfung herstellen von dieser textlichen Nachrichtenquelle zu der Datenquelle CrunchBase. CrunchBase ist z.B. eine der Quellen, in denen man Informationen über Technologiefirmen vorfindet, in dieser strukturierten RDF-basierten Art und Weise, wo ich dann gängige Informationen über IBM finden könnte. Vielleicht könnte ich auch herausfinden, dass der Sitz von IBM in Armornk, USA ist und mir dann DBpedia anschauen, was ich über Armornk finde. So kann ich heutzutage über alle Städte naheliegende Informationen finden. Dieses Szenario setzt aber voraus, dass ich in der Lage bin, die Strukturen, die ich in dieser CrunchBase-Datenquelle habe, passend mit den Strukturen, die ich in DBpedia vorfinde, zu verknüpfen. Eine weitere technische und inhaltliche Herausforderung besteht auch darin, wenn unterschiedlichen Quellen, z.B. CrunchBase und NYTimes, unterschiedliche URIs für IBM verwenden, zu erkennen, dass in beiden Quellen dasselbe Unternehmen angesprochen wird. NYTimes ist auch eine der Firmen, die sehr stark in die Bereitstellung von Linked Data eingestiegen ist und wo Sie sehr viele Informationen zum ganzen Business- und Politikbereich finden. Es ist naheliegend, dass Sie Informationen über IBM finden, wenn Sie auf Datenquellen von NYTimes gehen.

Eine der Herausforderungen ist, zu erkennen, ob die Entität, die ich bei CrunchBase vorfinde, die gleiche ist, die ich bei NYTimes vorfinde und wiederum die gleiche, die ich vielleicht bei Freebase vorfinde. Das sind alles Quellen, die über Firmen, Personen, Städte und dgl. Aussagen machen können. Erkennen von identischen Entitäten oder von ähnlichen Entitäten ist eine der Herausforderungen.

Wir haben hier ein Szenario diskutiert, in dem wir sehen, dass wir diese Nachrichtenquelle mit sehr vielen Informationen anreichern können, die heutzutage auf dem Web da sind. Man muss nur beachten, dass man diese Quellen in der korrekten Art und Weise miteinander verknüpfen kann. Das ist der Mehrwert, den wir durch das Web of Data haben, wenn wir das mit den sowieso entstehenden Nachrichten geeignet verknüpfen können.

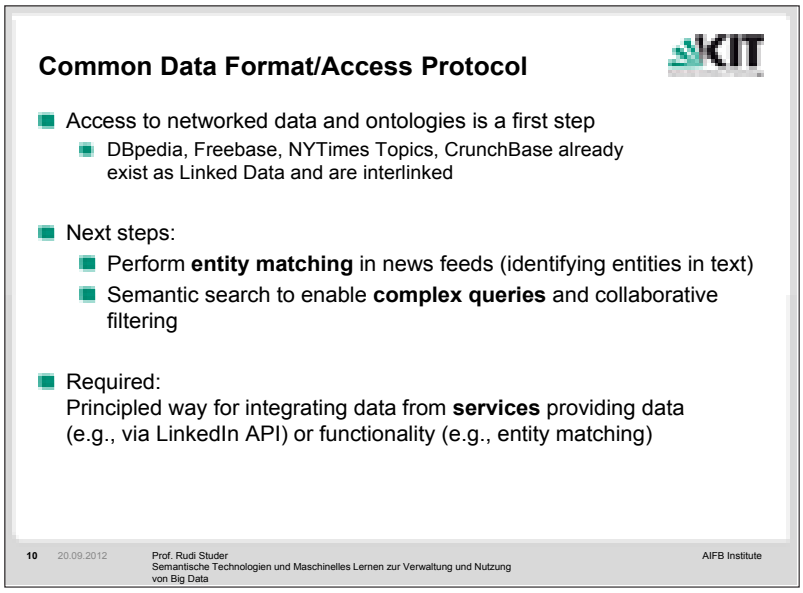

Bild 5

Vor welcher Herausforderung stehen wir also (Bild 5)? Wie wir auf diese Quellen zugreifen können, wenn wir SPARQL parat haben. Wie wir aus Texten Entitäten herausfinden. Wir werden noch sehen, dass das aber nicht immer geht. Wie wir komplexe Anfragen on Top von diesen Quellen stellen können, werden wir noch bei Semantic Search diskutieren.

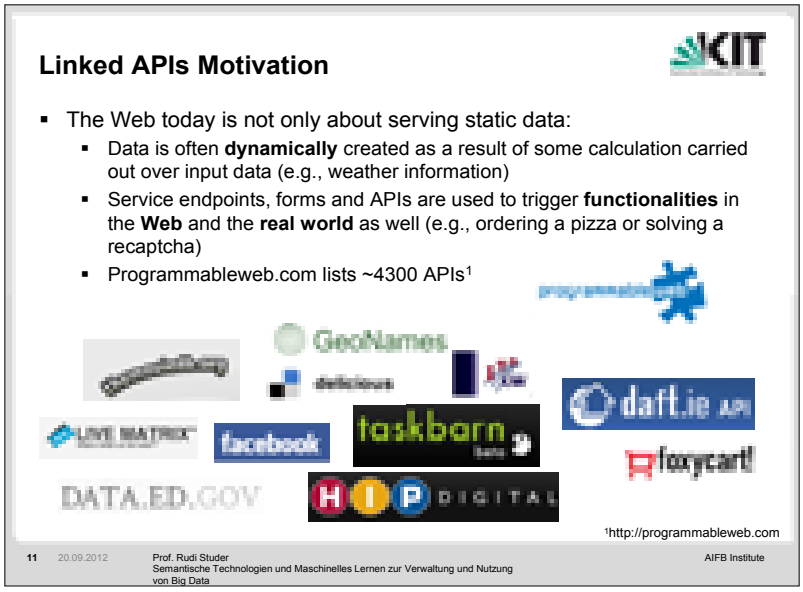

Bild 6

Wenn Sie aber schauen, was sich wirklich auf dem Web abspielt, was an Dynamik drin ist, dann gibt es noch andere Aspekte zu berücksichtigen (Bild 6). Sie alle kennen inzwischen diese ganzen APIs, die von vielerlei Anbietern angeboten werden, sei es Wetterinformation, sei es Geoinformation, was sicher Herr Abecker später noch im Detail diskutieren wird. Wir sind eigentlich nicht mehr nur auf dieser statischen Webwelt unterwegs, sondern auch in der Webwelt, in der wir Services haben, die uns Funktionalitäten bereitstellen. Dann ist die Frage, ob wir eigentlich in der Lage sind diese beiden Welten, diese statische Linked Data-Welt und diese Servicewelt irgendwie vernünftig miteinander zu verknüpfen. Ich habe hier ein paar Beispiele von Anbietern aufgeführt, die Ihnen solche Services bereitstellen.

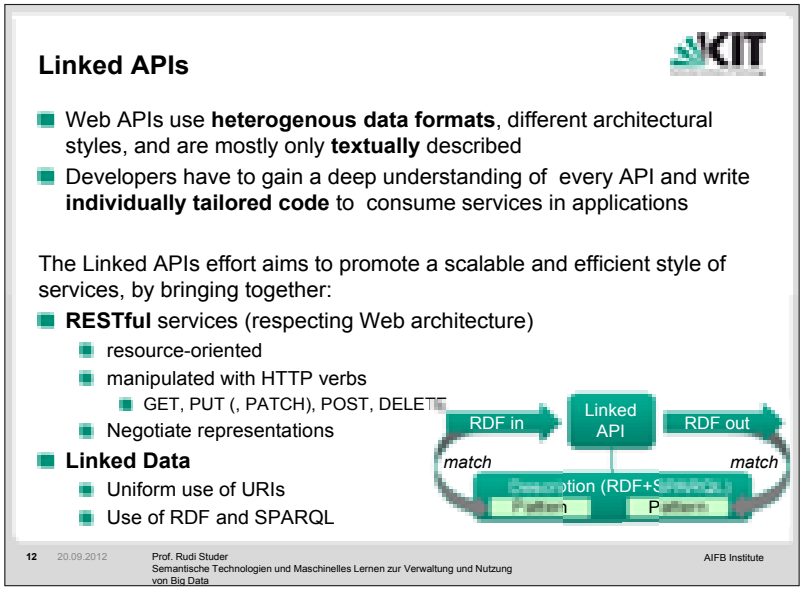

Bild 7

Welche Entwicklung spielt sich im Forschungsbereich in den letzten zwei bis drei Jahre ab (Bild 7)? Klassischerweise sind die APIs dadurch charakterisiert, dass außer einer syntaktischen Charakterisierung der Ein-/Ausgabeschnittstellen alles, was Funktionalitäten angeht, in textlichen Beschreibungen beinhaltet ist. Wenn Sie verschiedene APIs miteinander verknüpfen wollen, müssen Sie ziemlich tief in die Programmierebene einsteigen, um den passenden „Glue Code“ programmieren zu können.

Was ist die Idee, die mit den Linked APIs verbunden ist? Wir haben zum einen die Linked Data-Welt, die darauf basiert, dass man RDF-Daten hat, die mit SPARQL-Anfragen abgefragt werden können und die URIs haben, um die Entitäten zu identifizieren. Die verknüpfe ich jetzt mit der Servicewelt, aber eher mit einer leichtgewichtigen Servicewelt, nämlich mit RESTFul Services. Diese repräsentieren einen Trend, der die letzten Jahre relativ stark an Bedeutung gewonnen hat. Dieser Ansatz erlaubt es uns auch, gezielt gewisse Funktionalitäten durch entsprechende Patterns anzusprechen und auch Angaben zu machen, in welchen Formaten der Service die Eingaben erwartet und in welchen Formaten er die Ausgaben zurückliefert. Was wir mit diesen Linked APIs anstreben und auch prototypisch realisiert haben, ist die Idee, dass wir die Linked Data Formate RDF und SPARQL nehmen und sie mit den RESTful Services verknüpfen. Die Kombination besteht darin, dass alles, was die Linked APIs als Eingabe verarbeiten, durch passende Transformation in RDF daherkommt. Wir können dazu über SPARQL charakterisieren, welche Teilmenge von Daten als Eingabe produziert werden sollen. Das Gleiche gilt in Bezug auf die Daten, die der Service als Ausgabe liefert: wenn wir eine Wetterinformation über den Flughafen München anfragen, dann liefert der Service die Wetterinformation im RDF Format zurück, so dass wir damit in der Lage sind, die dynamische Servicewelt mit der statischen Linked Data-Welt auf einer einheitlichen Basis zu verknüpfen. Das ist der Mehrwert, den wir erzielen können.

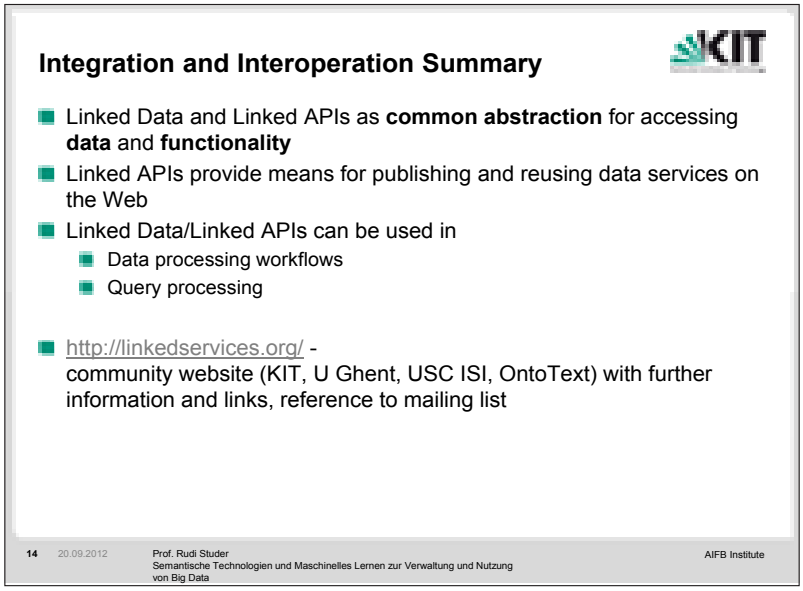

Bild 8

Das war eine Entwicklung, die in den letzten zwei Jahren Fuß gefasst hat (Bild 8).

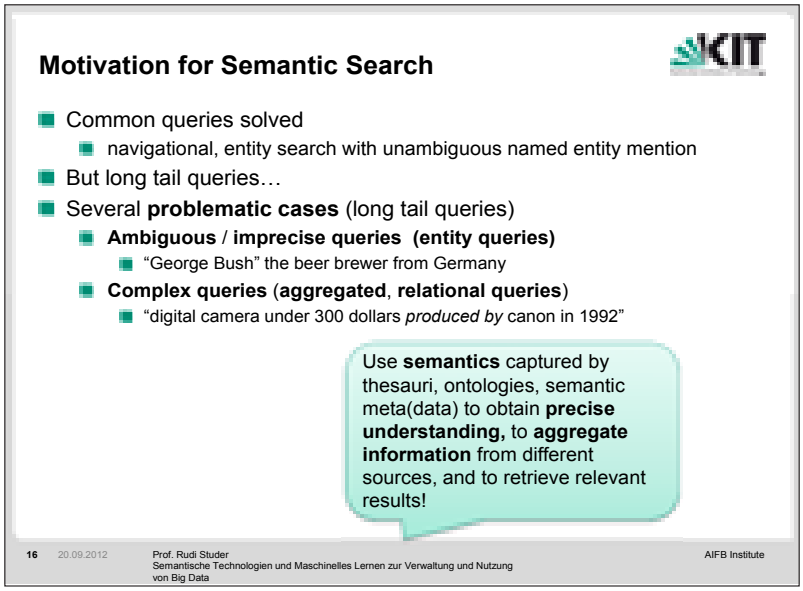

Bild 9

Lassen Sie mich mein zweites Thema aufgreifen, Semantic Search (Bild 9). Weshalb reden wir über Semantic Search? Wir alle kennen z.B. Google oder Bing. Wenn Sie nach Bayern München fragen, um ein Münchner Beispiel aufzugreifen, wird Ihnen Google genau die Information liefern, die Sie haben wollen. Wenn Sie nach George Bush fragen und den ehemaligen Präsidenten der USA meinen, würden Sie auch genau das finden, was Sie haben wollen. Aber vielleicht gibt es auch Interesse an dem Bierbrauer in Deutschland, der auch den Namen George Bush hat. Da wird Ihnen Google nicht auf der ersten, zweiten oder dritten Seite die passende Information liefern. Dies sind die sogenannten long tail queries, die selten auftreten, aber trotzdem durch die Vielzahl der einzelnen Fragen relevant sind, auch für viele Businessbereiche relevant sind, wenn Sie an Amazon und dgl. denken. Die können Sie eigentlich durch die klassischen Technologien nicht gut abdecken. Auch Fragen zu bestimmten Produkten mit bestimmten Eigenschaften, sind nicht so leicht mit den klassischen Anfragetechniken abzudecken. Da ist die Frage, was wir eigentlich über Semantik an Mehrwert für die Suche bereitstellen können.

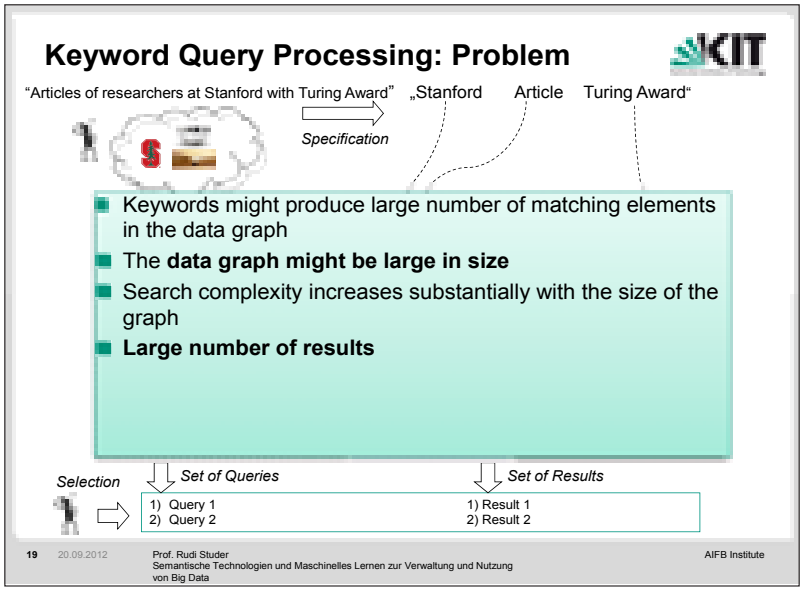

Bild 10

Ich möchte an einem kleinen Beispiel diskutieren, was die Techniken grundsätzlich leisten können, ohne hier jetzt ins Detail gehen zu können (Bild 10). Ich bin zum Beispiel an Artikeln oder Publikationen von Forschern interessiert, die in Stanford arbeiten und schon einmal den Turing Award bekommen haben. Die Insider unter Ihnen wissen, dass wir über John McCarthy reden, das Paradebeispiel an der Stelle. Wenn ich das in Google eintippe, kann Stanford relevant sein, Article steht für Publikation, Turing Award wollte ich eigentlich auch haben. Wenn ich diese drei Keywords angebe, dann sollte das Google eigentlich herausfinden, dass ich damit diese Anfrage gemeint habe. Dann schaue ich im Web of Data nach Informationen. Wenn ich bei Freebase bin, finde ich manches über Institutionen, auch etwas über Stanford zum Beispiel. Bei DBLP weiß jeder Forscher, dass er alles zu Publikationen findet, was er an der Stelle benötigt. Bei DBpedia finde ich sicher einen Eintrag über Turing Awards. Eigentlich habe ich die Informationen parat, aber in drei zunächst einmal losgelöst nebeneinander stehenden Datenquellen. Jetzt kann ich schauen, ob ich Personen finde, die irgendetwas mit Turing Award zu tun haben. Da finde ich diesen John McCarthy hier bei DBpedia. Dann finde ich irgendwas zu Publikationen. Da taucht auch irgendjemand auf, der den Namen John McCarthy hat. Was weiß ich jetzt über die Verknüpfung von dieser Person 1 mit Person 3? Ist das die gleiche Entität? Meine ich nur, dass die ähnlich sind, aber vielleicht doch nichts miteinander zu tun haben, weil es zufälligerweise zwei solche John McCarthys gab, die diese Eigenschaften haben. Da habe ich hier gewisse Unsicherheiten.

Hier ist einmal mit Wahrscheinlichkeiten bewertet, wie sicher ich eigentlich bin, dass der John McCarthy bei DBLP derjenige ist, der bei DBpedia auftaucht. So setzt sich das an der Stelle fort. Im Idealfall hätte ich jetzt eine Verknüpfung, bei der ich sehe, dass die Person, die ich bei Freebase finde, mit Stanford assoziiert ist. Das ist dieser John McCarthy, der diese Publikationen hat und der auch den Touring Award bekommen hat. Das wäre an dieser Stelle der Idealfall.

Dummerweise, wenn wir uns diese drei Keywords oben anschauen - Stanford, Article, Turing Award - kann jeder von Ihnen, wenn er diese Keywords liest, sich sagen, dass das

eine Anfrage nach Zeitungsartikeln über die Turing Award-Verleihung ist, die in Stanford stattgefunden hat. Das ist auch eine natürliche Leseart für diese drei Keywords.
Damit sehen Sie, wenn ich diese Anfrage interpretiere und nicht nur eine Leseart habe sondern eine zweite, dritte, vierte usw., da muss ich irgendwie erkennen können, was die am nahe liegendste Lesart ist, die ich an der Stelle bereitstellen muss.

Was sind die Herausforderungen? Wir haben sehr große Datenbestände an der Stelle. Ich muss dann irgendwie in der Lage sein, mit diesen großen Datenbeständen umzugehen.

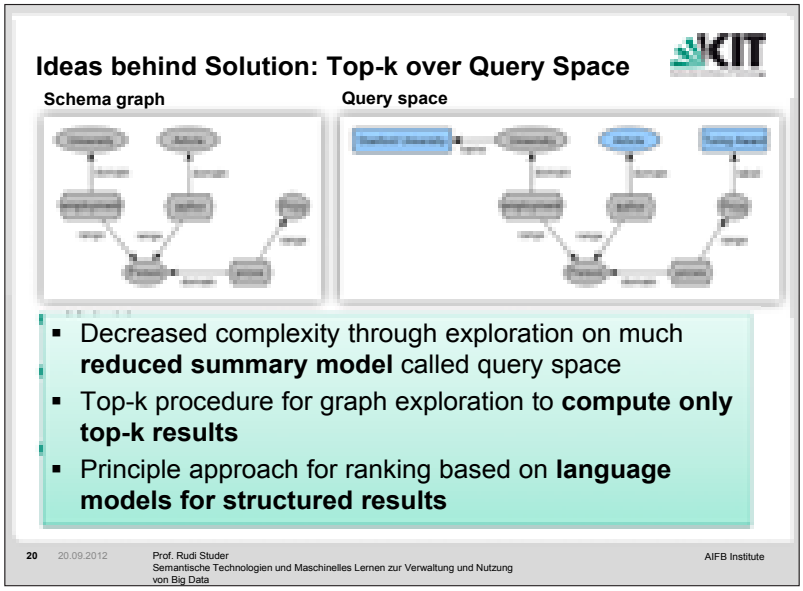

Bild 11

Ich möchte einmal andeuten, wie wir das machen können (Bild 11). Um von diesen riesengroßen Datenbeständen wegzukommen, kann ich die Daten ein bisschen reduzieren, indem ich immer noch die wesentlichen semantischen Strukturen erhalte, aber z.B. alles nur noch auf einer Schemaebene mache. Ich habe Entitätstypen wie Universitäten oder Autoren, aber ich habe die Instanzen aus dem Schemagraph entfernt und erweitere den dann aufgrund der Anfrage zu dem, was wir an dieser Stelle Query Space nennen. Hier müssen wir dann in der Lage sein, durch passende Matchingtechniken zu erkennen, dass das Keyword Stanford, das wir eingegeben haben, irgendwie zu „University“ von dem Schemagraph passt und Turing Award auf die „Prize“-Entität Bezug nimmt.

Was ist der Vorteil? Der Suchraum ist kleiner, so dass ich auf einer kleineren Datenmenge arbeite. Außerdem gibt es– wir kennen das auch aus dem Datenbankbereich - diese Top-k Algorithmen, die uns nicht alle Resultate liefern, sondern die besten k Resultate liefern. Da ist die Frage, wie ich das Ranking machen kann, was uns wirklich sicher die besten k Antworten liefert. Das sind aber noch einmal andere Themenbereiche, die da eine Rolle spielen: Wie sehen solche Top-k Algorithmen aus, die dann auf diesen Graphen effizient operieren können?

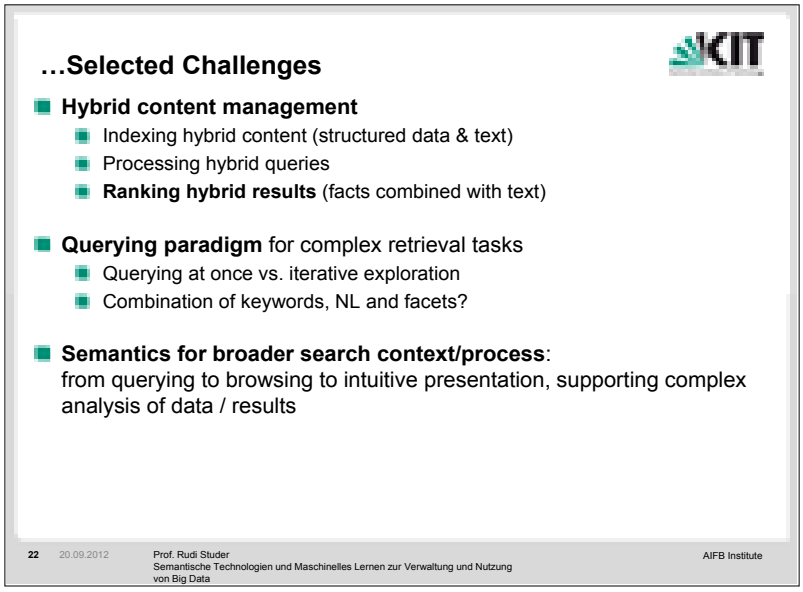

Bild 12

Soviel zu diesem semantischen Suche-Bereich (Bild 12). Aktuelle Herausforderungen sind, wenn ich nicht nur die strukturierten Daten habe, sondern auch Textdaten. Wie kann ich eigentlich die Dinge miteinander verknüpfen?

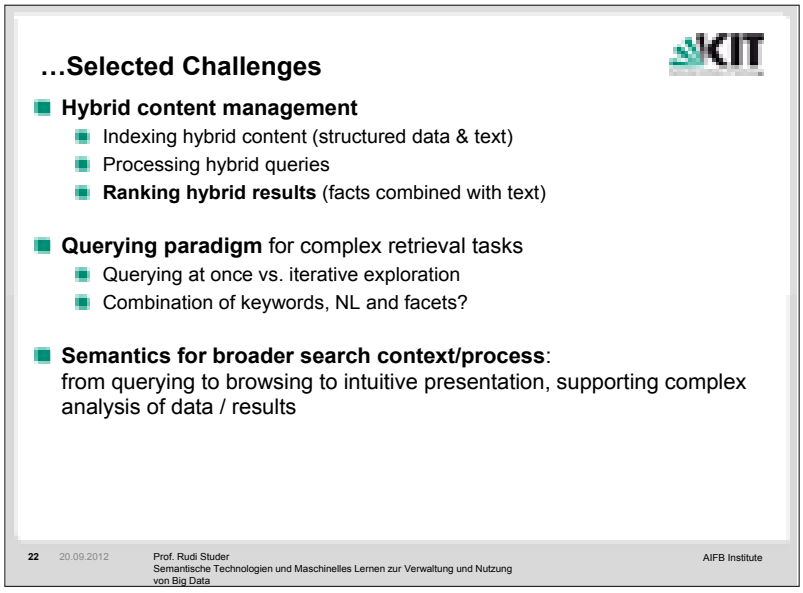

Bild 13

Lassen Sie mich abschließend einen Blick in den Learning Bereich werfen, wo ich eigentlich zwei Themenfelder ansprechen möchte, zum einen den Themenbereich Textmining (Bild 13). Da kann ich aus Textquellen solche Entitäten wie IBM oder wie Stanford University heraus-

finden. Wie kann ich die dann mit den Daten verknüpfen, die ich schon habe? Oder wie kann ich z.B. durch passende Clustering-Algorithmen ähnliche Entitäten oder ähnliche Fakten oder Ereignisse aus solchen Quellen herausfinden?

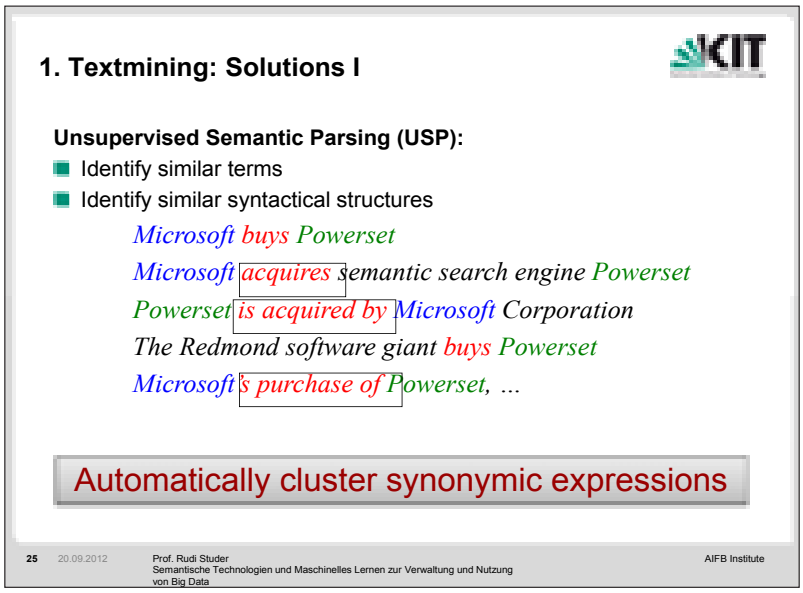

Bild 14

Wenn wir noch einmal an textliche Quellen denken - gerade vorhin hatte ich über solche News-Nachrichten gesprochen (Bild 14). In Finanznachrichten könnten wir auf dieses Ereignis zurückkommen: „Microsoft buys Powerset“. Das könnte auch in allen möglichen syntaktischen Ausprägungen in verschiedenen Texten stehen. Da entsteht die Frage, ob ich irgendwie erkennen kann, dass es sich eigentlich immer um denselben Fakt handelt, nämlich Microsoft hat Powersoft gekauft. Ich möchte hier nicht fünf Fakten haben, sondern erkennen können, dass es immer dieselbe Aussage, dasselbe Faktum ist, was sich in der realen Welt zugetragen hat.

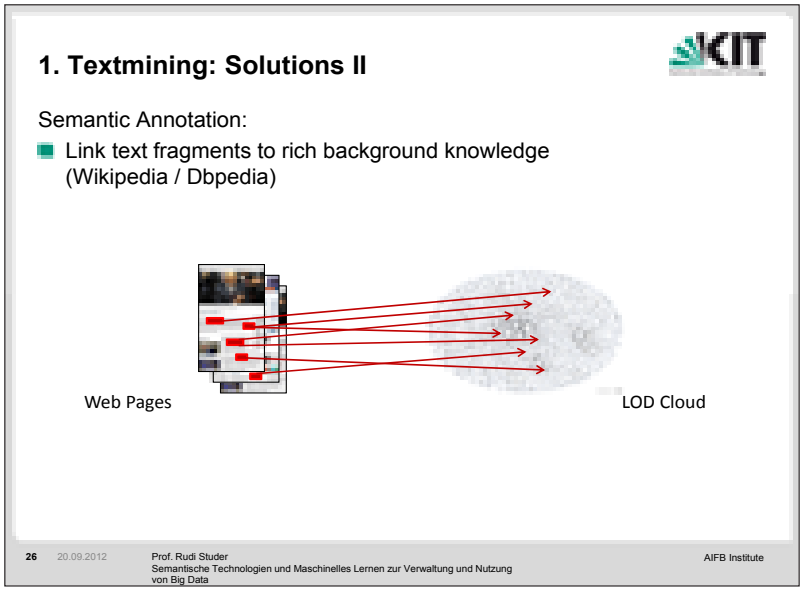

Bild 15

Das zweite ist (Bild 15), wie ich solche Textfragmente mit meiner strukturierten Datenwelt verknüpfen kann, mit diesem Linked Data, um einfach einen Mehrwert um diese Nachricht herum aufbauen zu können.

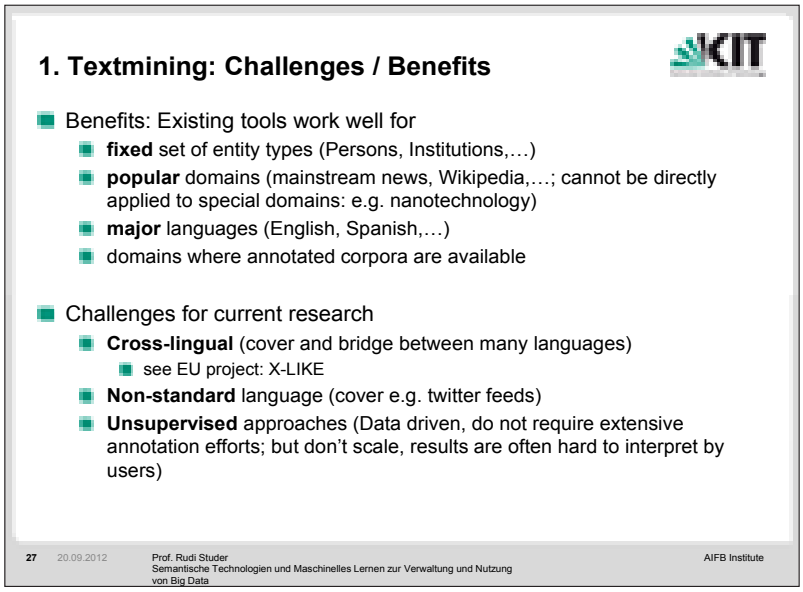

Bild 16

Was wir heutzutage relativ gut können, ist, ist Entitätserkennung für wenige Entitätstypen zu machen (Bild 16). Personen können wir ziemlich gut erkennen, Institutionen auch. Aber wenn Sie in andere Bereiche gehen, nimmt die Qualität ziemlich schnell ab. Wir können das auch für gängige Datenquellen machen, wenn wir an Wikipedia u. ä. denken. Wir haben

gerade ein Projekt im Bereich Nanotechnologie zu dem Thema gemacht. Da sieht es schon sehr viel schlechter aus, was die Entitätserkennung angeht.

Jeder von Ihnen denkt heutzutage in Englisch. Das gibt es sehr vieles, was algorithmisch geht und gut ausgearbeitet ist. Auch die passenden Lexika sind da. Aber wenn Sie in andere Sprachen gehen, z.B. Indien mit 25 Sprachen, und die Grundlage für Lexika usw. anschauen, sieht es ganz düster aus. Da muss man noch sehr viel Basisarbeit leisten, um überhaupt solche Dinge hinzubekommen.

Wenn wir in Europa unterwegs sind, ist alles wichtig, was Mehrsprachigkeit angeht. Sie bekommen eine englische Nachricht, die nächste in Spanisch, die dritte in Französisch und wollen die miteinander verknüpfen. Das ist etwas, was wir in einem EU Projekt untersuchen, wie wir Informationen aus verschiedenen Sprachen miteinander verknüpfen können.
Heute Morgen tauchte schon mehrfach Twitter auf. Wenn man sich die Sprachausdrücke anschaut, die da auftauchen, sind die von einer ganz anderen Bauart als wenn Sie einen sauber ausformulierten Pressetext von IBM lesen. Der hat eine ganz andere Struktur, was die Sprachqualität angeht. Man muss auch in der Lage sein, damit umgehen zu können.

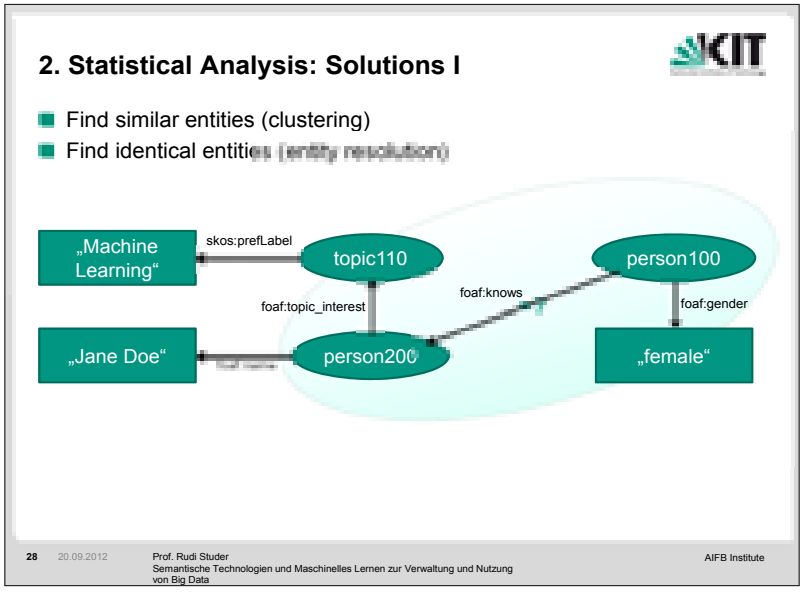

Bild 17

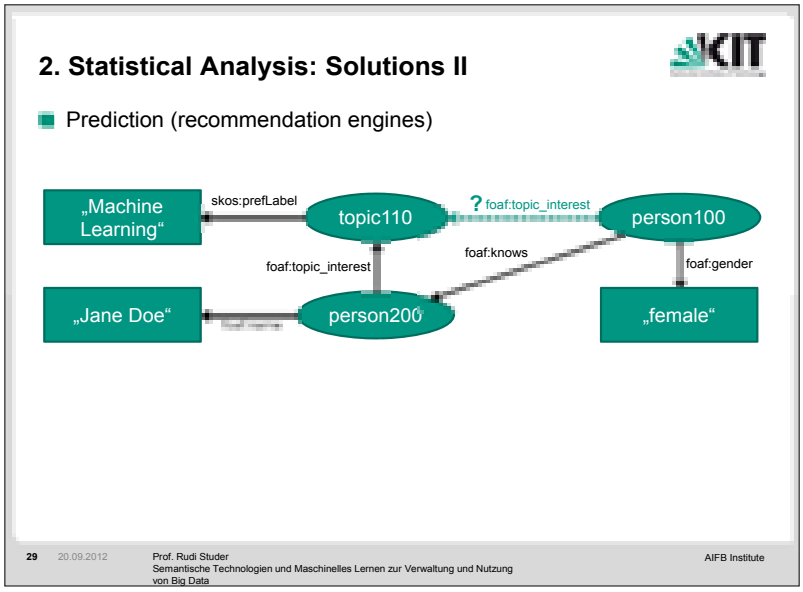

Bild 18

Lassen Sie mich noch den letzten Punkt ansprechen (Bild 17, 18). Es geht darum wie wir herausfinden können, ob gewisse Entitäten vielleicht etwas miteinander zu tun haben. Das ist auch eine der Herausforderungen, die in dem Kontext existieren. Ein kleines Beispiel hierzu: Ich habe zwei Personen. Von der einen weiß ich, dass sie weiblich ist. Von der zweiten weiß ich, dass sie Jane heißt und sich zufälligerweise für Machine Learning interessiert. Da ist die Frage, ob wir irgendetwas an Zusammenhängen zwischen diesen beiden Entitäten, dieser Person 100 und dieser Person 200 herstellen können. Sind es vielleicht die gleichen Personen? Beide weiblich war die Information, die man schon aus dem Namen herauslesen konnte. Oder haben die vielleicht die gleichen Interessen? Wenn sich die Jane für Machine Learning interessiert, ist dann vielleicht die Person 100 auch an Machine Learning interessiert, aufgrund von Ähnlichkeiten, die Sie in Facebook, bei Twitter, oder sonst wo über diese Person auffinden? Das sind Fragen, die wir an der Stelle angehen.

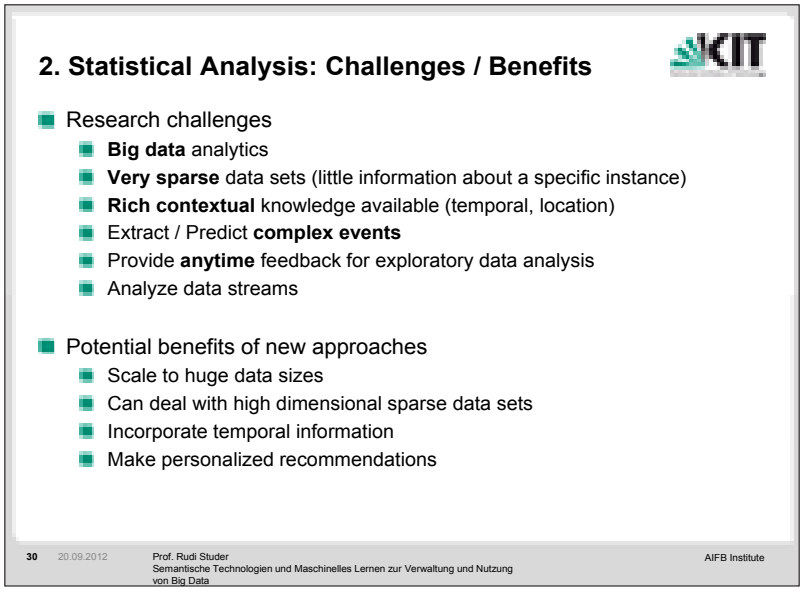

Bild 19

Was sind die Herausforderungen, wenn man an diese Learning-Algorithmen denkt (Bild 19)? Wir haben oftmals sehr dünn besetzte Datenquellen parat. Über IBM werden wir viel finden. Über Madonna werden wir auch viel finden. Aber es gibt viele Entitäten, für die Sie nur ein paar Einträge finden. Da ist die Frage, was Sie in diesem Kontext wirklich herausfinden können.

Um noch einen zweiten Punkt herauszugreifen: Anytime Algorithmen . Wenn ich riesige Datenmengen analysiere, dann will ich nicht drei Minuten warten bis die Resultate komplett kommen, sondern will vielleicht nach 5 Sekunden die ersten Resultate haben und dann immer bessere Resultate bekommen. Das sind zwei Punkte, die an der Stelle eine Rolle spielen.

Damit möchte ich zum Schluss kommen. Ich denke, dass ich aufzeigen konnte, dass wir einen Mehrwert durch das Web of Data haben, das uns eine Vielzahl von Daten im Web bereitstellt. Ich hatte aufgezeigt, dass man die Dinge nicht nur statisch sondern auch dynamisch betrachteten sollte, wenn man an die Anbieter von APIs denkt. Dass sie durch semantische Suchansätze in der Lage sind, auch Anfragen sinnvoll zu beantworten, die in diesem Long Tail sind, also selten an der Stelle gestellt werden aber trotzdem durch die Verknüpfung von Datenquellen gut zu verarbeiten sind und dass wir über Lernverfahren in der Lage sind, Ähnlichkeiten, Zusammenhänge zu erkennen, zum Beispiel auch manchmal Vorhersagen zu machen, wie man Entitäten miteinander verlinken könnte; oder auch Event Processing, was ich heute nicht angesprochen habe und das, wenn wir an die vorgestellten BMW- und Siemens Anwendungen denken, auch eine wichtige Rolle spielt: Pro aktiv zu erkennen, ob in den nächsten Wochen sukzessive Probleme entstehen und darauf aufbauend vorbeugende Maßnahmen ergreifen zu können.

Soviel mein knapper Einblick in drei methodische Themenbereiche, die für viele zukünftige Anwendungsszenarien zunehmend von Bedeutung sind. Ich bin gern bereit, später noch Fragen zu beantworten.

9 Alexandria – die kollaborative Wissensmaschine

Florian Kuhlmann, Neofonie GmbH, Berlin

Alexandria, wir nennen so die kollaborative Wissensmaschine, ist das größte Forschungsprojekt, was die Neofonie GmbH je durchgeführt hat. Das Projekt fand im Rahmen des Theseus Programms statt. Anhand dieses Beispiels können wir uns anschauen, wie Technologien aus der reinen Forschung auch in KMUs genutzt werden können, am Beispiel meines Arbeitgebers Neofonie. Ich will auch darauf eingehen, was das Ganze mit Big Data zu tun hat.

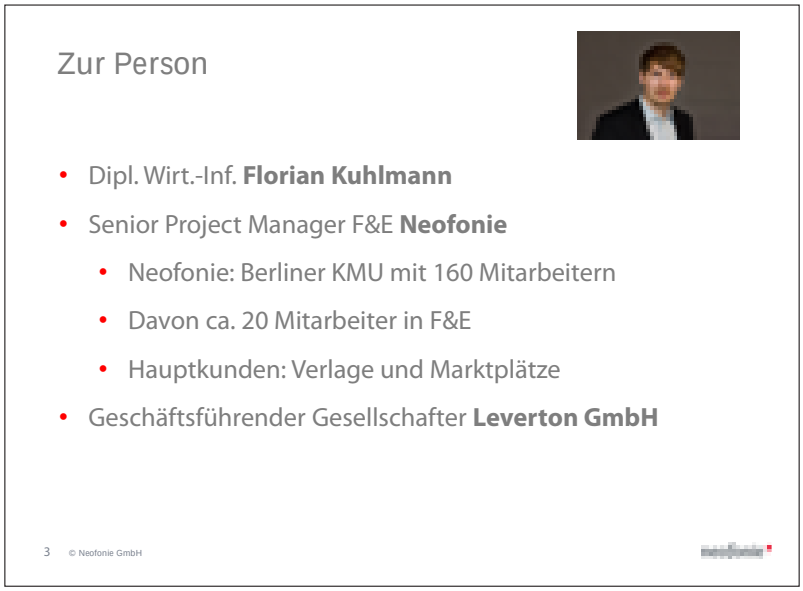

Bild 1

Zunächst kurz zu meiner Person (Bild 1). Ich arbeite seit sechs Jahren für Neofonie als Projektleiter. Neofonie ist ein KMU aus Berlin. Wir beschäftigen derzeit 160 Mitarbeiter, kommen aus dem Bereich Suchtechnologien. Traditionell haben wir einen großen Schwerpunkt auf Forschung und Entwicklung mit derzeit 20 Mitarbeitern, was für ein Unternehmen unserer Größe relativ viel ist. Hauptkunden der Neofonie kommen aus dem Verlagsbereich sowie Marktplätzen wie z.B. eBay, Kalaydo. Verlage sind immer noch der Hauptbestandteil unserer Kunden. Nebenbei habe ich noch ein Start-up in einem ganz anderen Bereich gegründet. Hier geht es um die Analyse von Rechtsdokumenten, was aber heute nicht das Thema ist.

Bild 2

Das Thema ist Alexandria, einer der sechs Theseus Use Cases (Bild 2). Theseus war das größte IKT Forschungsprojekt in Deutschland und wird dieses Jahr abgeschlossen. Der Schwerpunkt bei Alexandria lag auf der Kombination von semantischen Technologien mit Social Media. Das Projekt selbst ist Ende 2011 abgeschlossen worden, und unter alexandria.neofonie.de kann man heute frei zugänglich einen kleinen Showcase finden, wo man einen Teil der Technologien, die wir entwickelt haben, ausprobieren kann.

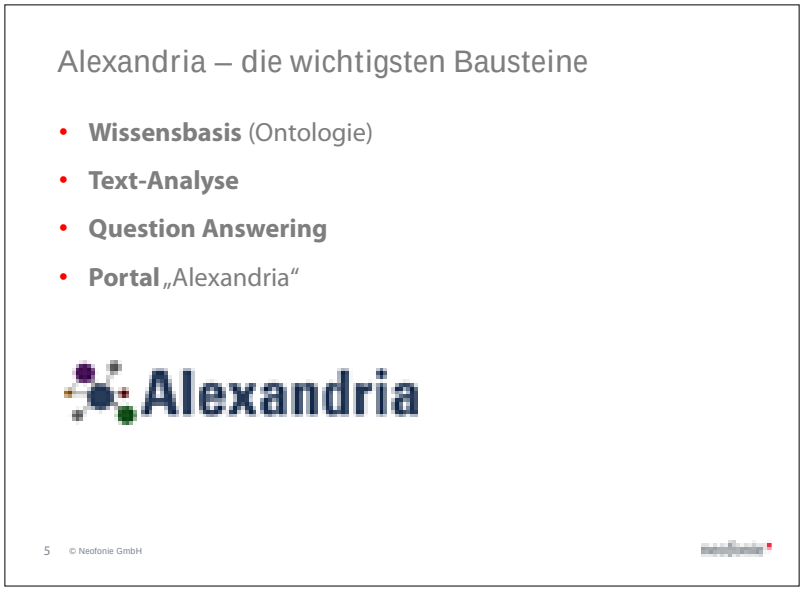

Bild 3

Die wichtigsten Bausteine, die bei Alexandria entstanden sind, sind zum einen eine Wissensbasis mit strukturierten Daten, eine so genannte Ontologie, die ich gleich im Detail vorstellen möchte (Bild 3). Des Weiteren Werkzeuge zur Textanalyse, auch relevant für heute. Was ich heute aus Zeitgründen nicht vorstelle, ist das Question Answering, eine Art semantische Suche, die es erlaubt, mit natürlich sprachlichen Anfragen Analysen auf eine Wissensbasis durchzuführen, Aber das können Sie selbst ausprobieren, wenn Sie möchten.

Bild 4

Die Wissensbasis ist das Herz von Alexandria (Bild 4), und besteht aus einem Schema, wo wir definieren, welche Konzepte in der Wissensbasis auftauchen dürfen, Personen, Orte, Werke, Organisationen usw. und welche Arten von Beziehungen es geben kann. Dies ist per Hand modelliert. Was nicht manuell geschieht, ist die Befüllung mit Daten. Dazu komme ich später. Wenn die Daten einmal drin sind, sieht das Ganze so aus. Wir haben mittlerweile ein Netzwerk mit ungefähr sieben Millionen verschiedenen Entitäten. Das sind 2,2 Millionen Personen, eine Million Orte, ungefähr 700.000 Unternehmen usw. Und wir haben ungefähr 100 Millionen Fakten und Beziehungen zwischen diesen verschiedenen Entitäten.

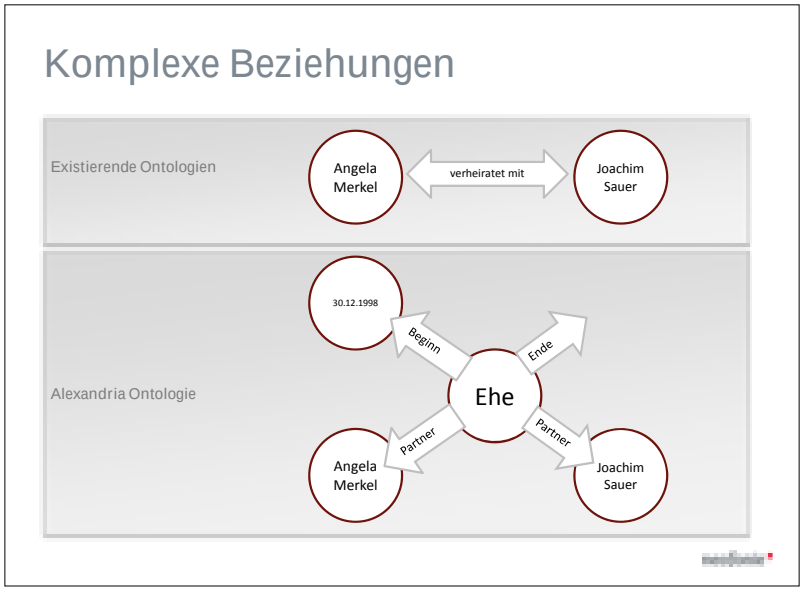

Bild 5

Eine Besonderheit, die wir in Alexandria abbilden, ist, dass wir komplexe Beziehungen in unserer Ontologie erlauben (Bild 5), d.h. wir haben nicht nur binäre Beziehungen, also z.B. „Angela Merkel verheiratet mit Joachim Sauer“, wie es herkömmliche Ontologien bieten. Wir haben hier mit einem kleinen Kunstgriff, in dem wir Beziehungen zu Konzepten erhoben haben, die Möglichkeit, Wissen von nahezu beliebiger Komplexität abzubilden, d.h. wir sagen nicht, dass es eine Beziehung „verheiratet mit“ gibt, sondern es gibt ein Konzept Ehe. An das Konzept Ehe hängen wir beliebig viele Attribute, und der Ehepartner ist jetzt eines dieser Attribute. Beginn der Ehe und Ort der Eheschließung sind weitere Attribute. Dies erlaubt Analysen, die mit dem einfachen binären Modell, was wir hier oben sehen, nicht möglich sind.

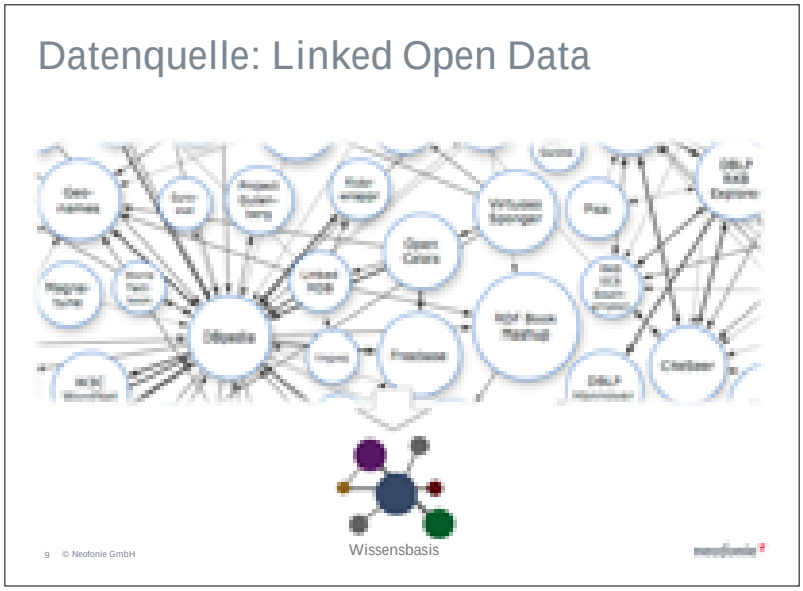

Bild 6

Was sind die Datenquellen (Bild 6)? Das haben wir heute schon mehrfach gesehen, Linked Open Data. Wir fragen ständig mehrere Quellen aus der Linked Open Data-Cloud an. Hauptquellen sind derzeit Freebase, DBpedia und Geonames, und wir kopieren die Daten und bilden diese auf unsere eigene Wissensbasis ab.

Bild 7

Die zweite Datenquelle ist in Zukunft hoffentlich die Community (Bild 7). Das Portal ist öffentlich zugänglich. Man kann die Daten, die in dem Portal vorhanden sind, nahezu

beliebig editieren. Es gibt einen Qualitätsmanagementprozess usw. Das kann man sich auf dem Portal anschauen.

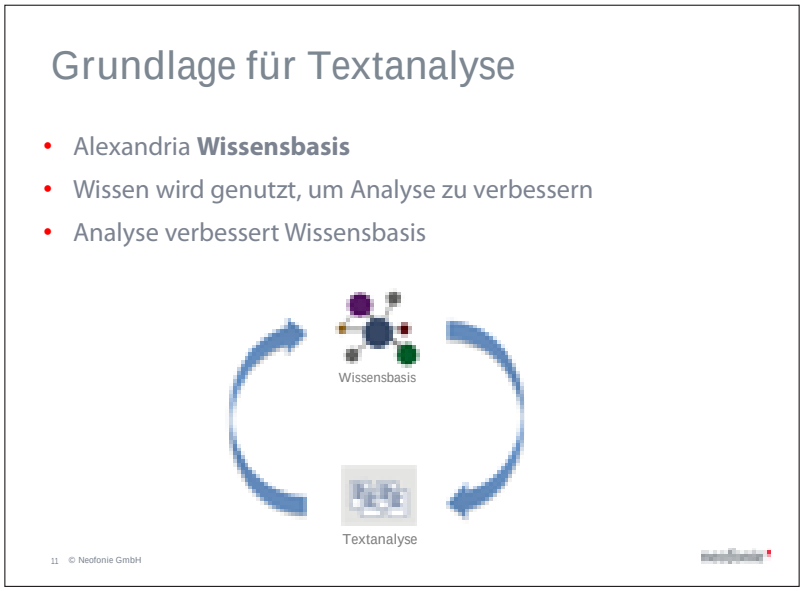

Bild 8

Was ist der Nutzen der Wissensbasis (Bild 8)? Man kann das vorhandene Wissen analysieren. Es ist spannend, Beziehungen zwischen Personen, Orten usw. herauszufinden Für uns als KMU ist aber der Hauptnutzen der Wissensbasis, dass diese die Grundlage für Textanalyse bildet. Das heißt, wir nutzen Informationen aus der Wissensbasis, um Textanalysen zu verbessern. In Zukunft soll es so sein, dass das Ganze zu einem Kreislauf wird. Wir wollen die Informationen, die wir durch die Textanalyse gewinnen, zurück in die Wissensbasis führen, so dass die Wissensbasis weiter anwächst und sich dadurch die Qualität der Textanalyse verbessert usw.

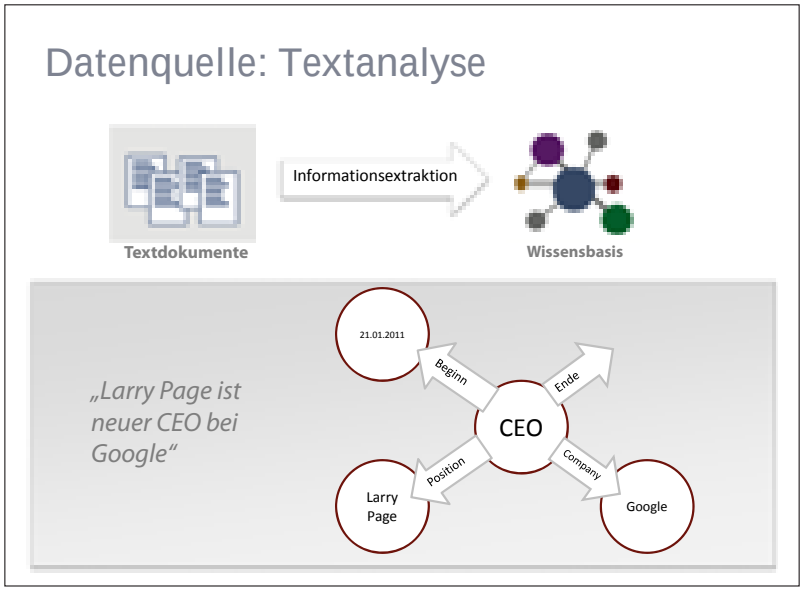

Bild 9

Was ist Textanalyse eigentlich (Bild 9)? Für uns ist das die Überführung von unstrukturierten Daten in strukturierte. Wenn wir in einem Textdokument einen Satz finden wie „Lary Page ist neuer CEO bei Google“, dann wollen wir das Ganze in unser Modell überführen, d.h. das Konzept CEO wollen wir hier mit Attributen füllen. Wir wollen Lary Page als Attribut Position und das Datum des Auffindens als Startdatum angeben. Überführung in strukturierte Daten ist das große Ziel der Informationsextraktion, wenn die Daten einmal strukturiert vorliegen, kann ich die Daten dazu beliebig auswerten. Bei Textdokumenten ist das nur begrenzt möglich.

Bild 10

Wie hilft die Wissensbasis bei der Textanalyse (Bild 10)? Ich hatte versprochen, dies zu zeigen. Ein großes Problem ist, dass erkannte Namen, die erst einmal nur durch Zeichenketten repräsentiert werden, nicht eindeutig sind. Wir haben das schon im Vortrag vorher gehört. Bei Alexandria haben wir z.B. den Namen Peter Müller 18mal vergeben, 18 verschiedene Personen mit jeweils unterschiedlichen Eigenschaften. Ein prominenteres Beispiel ist Sarah Connor, wo es noch komplizierter ist. Wir haben Sarah Connor einmal als Sängerin und einmal als fiktive Figur im Terminator Universum, d.h. hier gibt es Filme und da gibt es eine Figur Sarah Connor, die durch verschiedene Schauspieler repräsentiert wird.

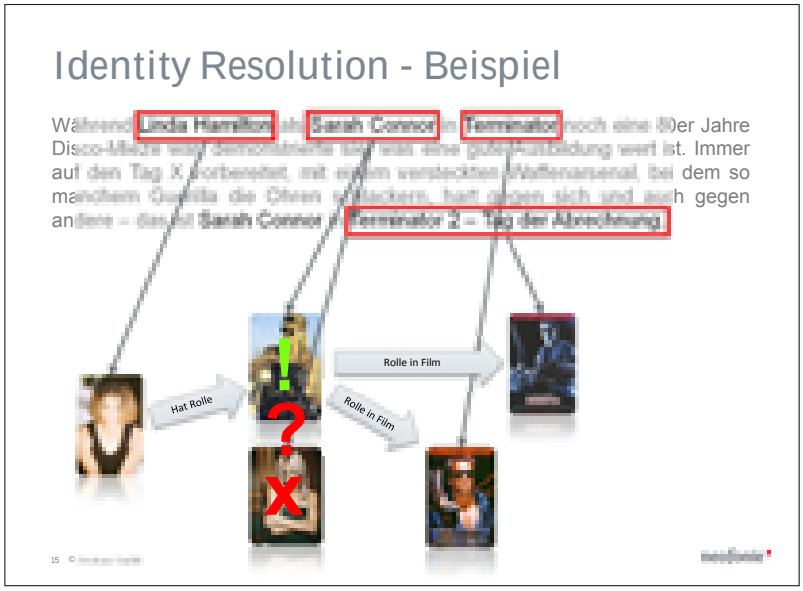

Bild 11

Hier sehen wir einen Beispielstext, in dem es um eine der Sarah Connors geht (Bild 11). Wir müssen den Text erst einmal durchlesen: „Während Linda Hamilton als Sarah Connor in Terminator noch eine 80er Jahre Discomieze war, demonstrierte sie, was eine gute Ausbildung wert ist.“ Ich lese jetzt nicht weiter. Wir als Mensch mit Verstand für den Kontext können beim ersten Satz durch eine Interpretation des Kontextes ziemlich schnell feststellen, dass es offensichtlich hier um die Sarah Connor aus Terminator geht und nicht um die Sängerin Sarah Connor. Für die Maschine ist dies etwas schwieriger. Sie muss sich den Kontext selbst erschließen. Wir haben ein Verfahren entwickelt, was sich zunächst einmal alle gefundenen Entitäten anschaut und gegen eine Wissensbasis abgleicht. Alles, was hier im Text fett gedruckt ist, sind gefundene Entitäten. Bei Linda Hamilton haben wir Glück. Da gibt es nur einen Eintrag in der Wissensbasis. Hier sind wir uns relativ sicher, dass es die Linda Hamilton ist, die gemeint ist. Bei Sarah Connor sieht es ein bisschen schwieriger aus. Wie schon erwähnt, haben wir zwei verschiedene Sarah Connors. Das ist das Entscheidungsproblem, vor dem wir stehen. Terminator 1 und Terminator 2 als Filme sind auch wieder eindeutig. Hier haben wir ein paar Anhaltspunkte. Wenn wir jetzt die Wissensbasis analysieren und die Vernetzungsinformationen abbilden, sehen wir, dass nur die obere Sarah Connor überhaupt vernetzt ist mit den anderen gefundenen Entitäten. Das ist für uns die Entscheidungsgrundlage zu sagen, dass es sehr wahrscheinlich die obere Sarah Connor ist, nämlich die Kunstfigur aus dem Terminatorfilm.

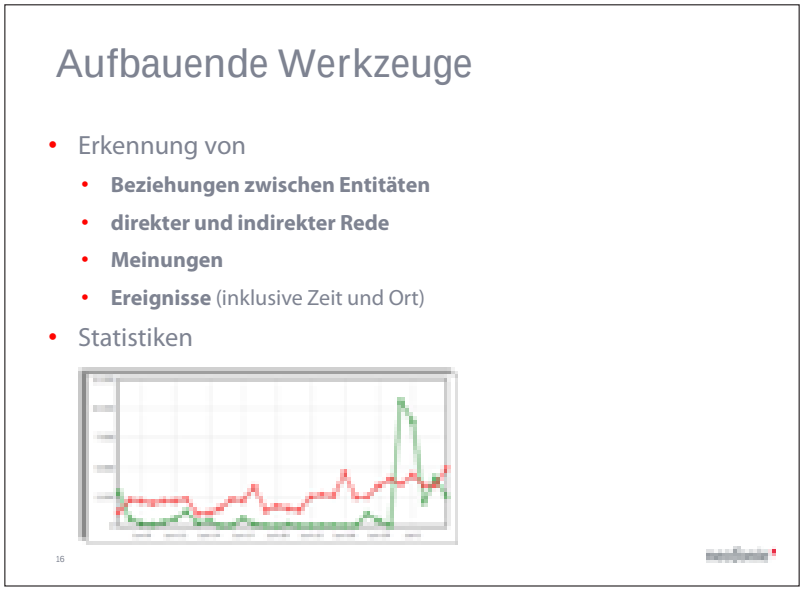

Bild 12

Aufbauend auf dieser Identifizierungskomponente haben wir in Alexandria diverse Werkzeuge entwickelt die einen echten Mehrwert bringen (Bild 12). Wir können Beziehungen zwischen Entitäten ermitteln. Wir können direkte und indirekte Rede extrahieren und auch den Entitäten zuordnen: Wer hat was gesagt. Bei der tiefen Extraktion von Meinungen stehen wir noch am Anfang, aber es klappt schon ganz gut. Wer hat welche Meinung mit welcher Polarität, negativ oder positiv, über wen? Wir können neue Ereignisse herausfinden inklusive Zeit und Ort, wo und von wann bis wann die Ereignisse stattgefunden haben. Wir können Statistiken bilden, was auch interessant ist. Wir können im Zeitverlauf schauen, wie sich die Popularität von bestimmten Personen, Orten, Ereignissen entwickelt.

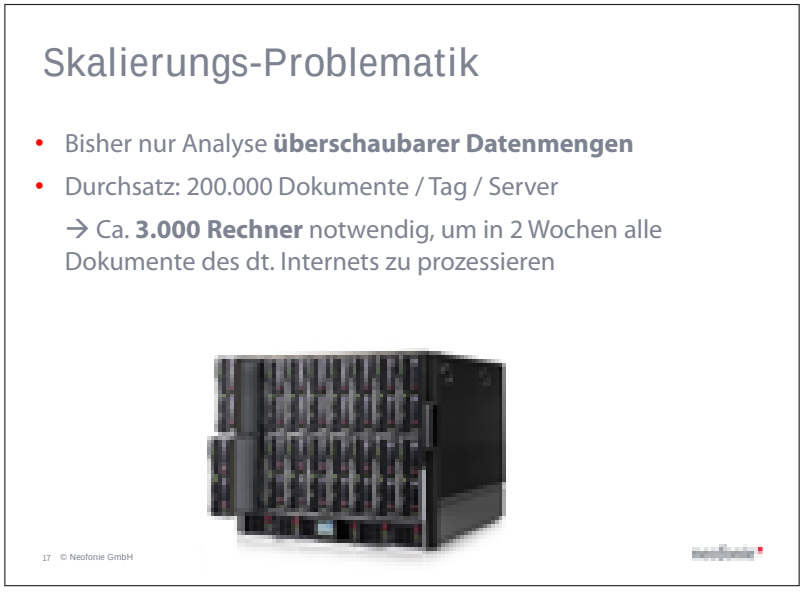

Bild 13

Ein Problem ist, dass die bisher beschriebenen Technologien nur für den Einsatz auf kleinen, überschaubaren Datenmengen gut funktionieren (Bild 13). Das klappt ganz gut mit vielleicht ein paar Hundert oder Tausend Dokumenten pro Tag. Wenn man sich die Frage stellt, wie können wir z.B. einen Fall bearbeiten, wo jemand eine Analyse über das ganze Internet haben will, ist die Antwort recht einfach. Wir haben das einmal hochgerechnet.

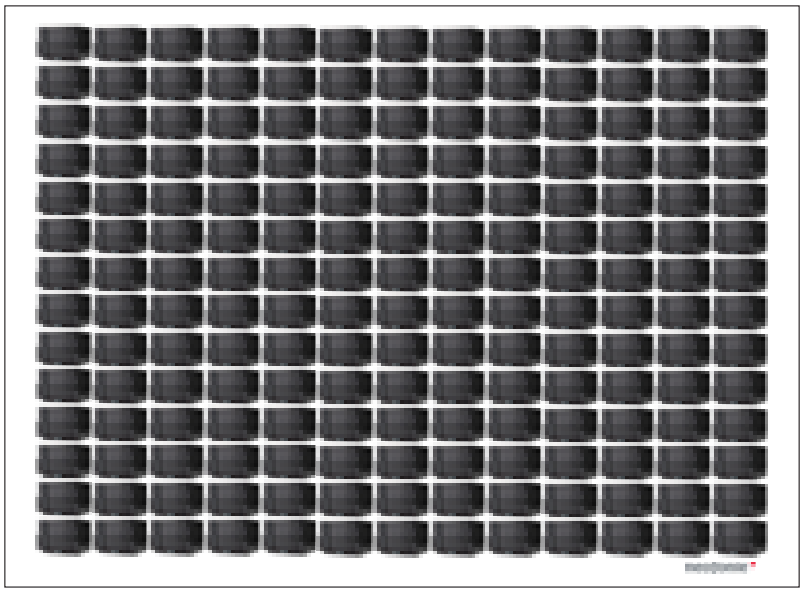

Bild 14

Um das deutsche Internet in zwei Wochen mit den bestehenden Technologien zu analysieren, brauchen wir ungefähr 3000 Rechner (Bild 14). Das ist für ein KMU unserer Größe nicht erschwinglich, auch nicht mit Cloud Technologien.

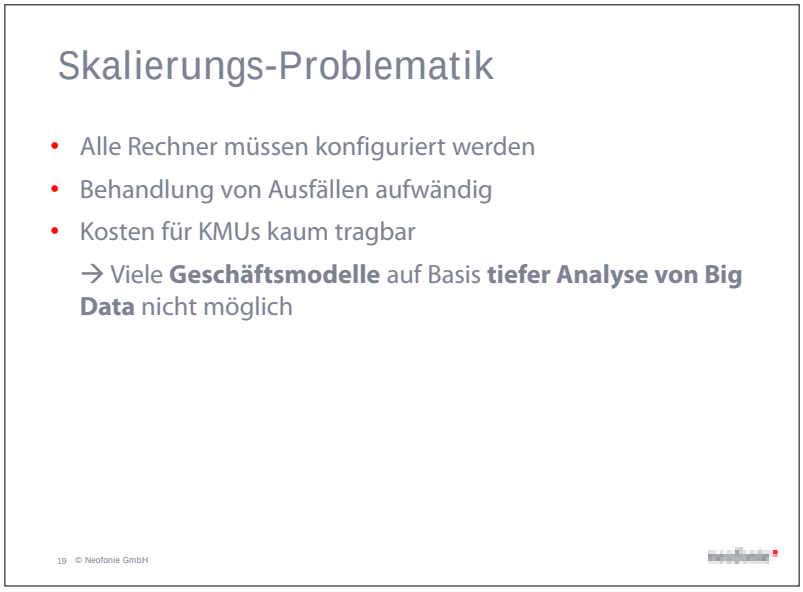

Bild 15

Auch diese für nur zwei Wochen anzumieten, übersteigt unsere Dimensionen bzw. die unserer Kunden (Bild 15). Geschäftsmodelle auf Big Data sind so nicht möglich. So wie die Technologien implementiert sind, ist es noch zu kostenintensiv, als dass wir als KMU uns das leisten könnten.

Bild 16

Aus diesem Grunde haben wir das aktuelle Projekt MIA zusammen mit der TU Berlin und anderen Partnern ins Leben gerufen (Bild 16). Davon haben wir heute auch schon kurz gehört. Hier geht es darum, mit Hilfe von Cloud Technologien, einen Marktplatz für Information

und Analysen zu schaffen. Die Forschungsschwerpunkte sind semantische Analysen auf Big Data mit neuen Technologien, die in dem Big Data Universum entstanden sind, voranzutreiben und die Skalierbarkeit erheblich zu verbessern. Darauf aufbauend soll eine Abfragesprache angeboten werden, die in „Near Real Time“ Analysen auf solchen Voranalysen ermöglicht, z.B. „zeig mir alle Produkte, die in Zusammenhang mit Firma XY erwähnt werden“, das Ganze mit fairen Abrechnungsmodellen. Dazu werden wir eine Plattform auch für andere KMUs später bereitstellen, die Analysen durchführen können und wo vielleicht auch neue Geschäftsmodelle erst entstehen, die vorher gar nicht möglich gewesen sind.

Das Projekt läuft seit Anfang des Jahres bis Ende 2014. Die Leitung liegt bei der TU Berlin, aber Neofonie hat einen wesentlichen Anteil bei der Implementierung. Weitere Partner sind u.a. Fraunhofer und noch drei weitere KMUs aus Deutschland.

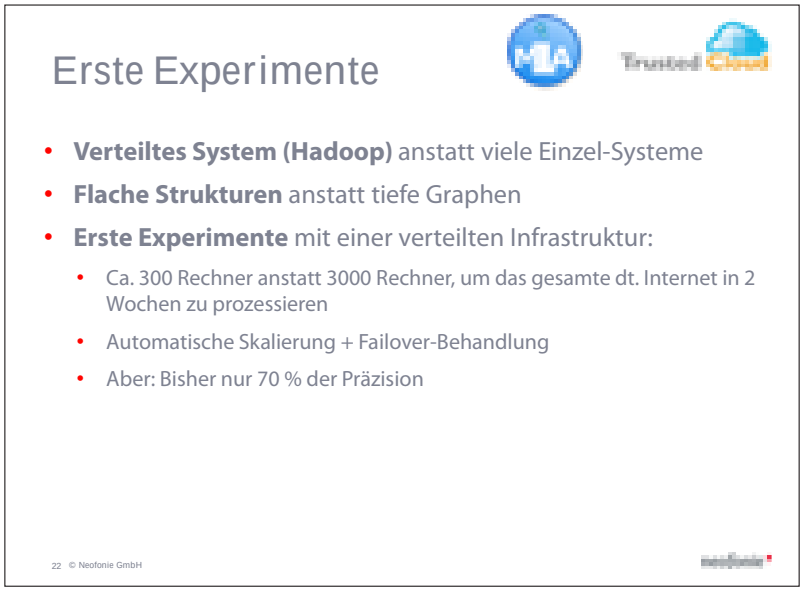

Bild 17

Erste Experimente, die wir hier mit Big Data Technologien wie z.B. Hadoop durchgeführt haben, sind sehr positiv (Bild 17). Wir konnten hier den Ressourcenbedarf auf nur 10 % minimieren, d.h. anstatt 3000 Rechnern brauchen wir nur noch 300 Rechner, um das deutschsprachige Internet in zwei Wochen zu analysieren. Das ist immer noch zu viel für uns. Aber es ist ein Anfang, und wir haben gerade erst mit dem Forschungsprojekt angefangen, d.h. hier ist sicherlich noch Luft nach oben bei der Skalierbarkeit.

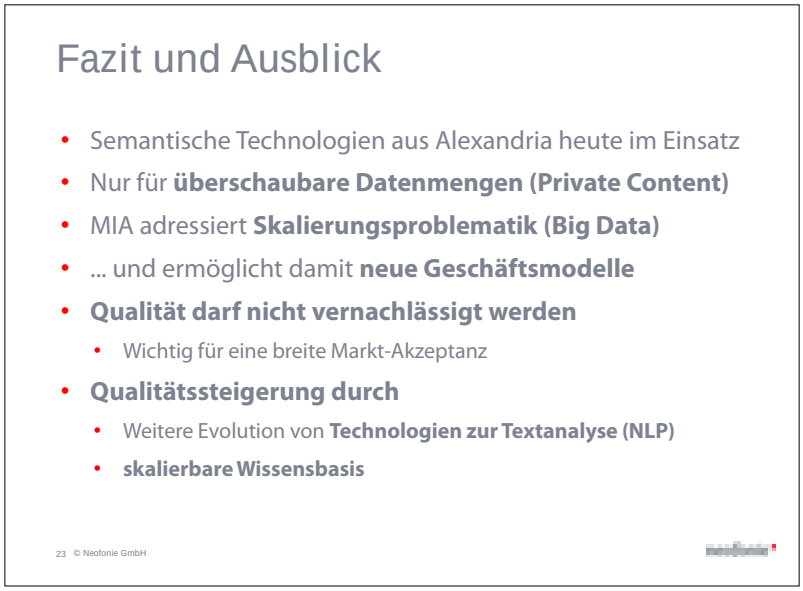

Bild 18

Damit möchte ich zum Fazit und zum Ausblick kommen (Bild 18). Die semantischen Technologien, die wir in Alexandria entwickelt haben, sind heute bereits im Einsatz. Stern.de ist z.B. einer unserer ersten Kunden. Weitere Kunden werden sicherlich dieses Jahr folgen. Aber eine Einschränkung liegt hier bei der Datenmenge. Wir schaffen es, vielleicht 400.000 Dokumente am Tag zu analysieren, aber wenn jemand eine Analyse über das ganze Internet haben möchte, müssen wir noch ein bisschen warten, bis das Projekt MIA diese Problematik adressiert hat, was hoffentlich dann im Jahr 2014 der Fall sein wird. Wir erhoffen uns, neue Geschäftsmodelle darauf aufbauend zu ermöglichen.

Mir ist noch wichtig zu sagen, dass bei manchen Basis-Technologien die Weiterentwicklung nicht vernachlässigt werden darf. Es gibt viele neuen Hype-Themen. Es gab Cloud Computing. Jetzt gibt es Big Data usw. Die Grundlagen der semantischen Technologien sind gut erforscht, aber da ist noch wesentlich Luft nach oben. Die Fortschritte hier sind evolutionär, nicht revolutionär. Man muss hier ausdauernd sein, um z.B. die Präzision der Verfahren weiter zu steigern. Aber dies ist teilweise Grundlagenforschung, was ein KMU mit unserer Größe nicht übernehmen kann. Das muss in der Grundlagenforschung geschehen. Die Skalierbarkeit ist ein wichtiger Faktor, aber die Qualität der Dienste ist mindestens genauso wichtig für eine breite Marktakzeptanz der Technologien. Für viele Anwendungsfälle reicht eine Präzision von unter 90% nicht aus; Verfahren zum Parsen von Textdokumenten erreichen derzeit aber kaum höhere Werte. Wenn es wirklich um sensible Analysen geht, kann das hier ein K.O. Kriterium sein.

Ein anderes Thema, was uns in Zukunft noch einmal begegnen wird, ist eine Wissensbasis wie wir sie in Alexandria aufgebaut haben, und die wesentlich zur Qualität der Textanalyse beiträgt, auch skalierbar zu machen. Das passiert heute de facto nicht. Der Trend geht hier zu In-Memory-Technologien, da nur diese auch schnell genug sind, komplexe Anfragen schnell zu beantworten. Aber wenn wir durch maximalen Hauptspeicherausbau an die In-Memory Grenze stoßen, ist das Ende der Skalierbarkeit erreicht. Dann geht es nicht mehr weiter.

10 Einfaches Finden und Analyse von Geo- und Umweltdaten

Dr. Andreas Abecker, disy Informationssysteme GmbH, Karlsruhe

Überblick

Ich möchte meinen Vortrag gliedern, wie folgt:

1. Zunächst skizziere ich das Tätigkeitsfeld der disy Informationssysteme GmbH, nämlich Datenmanagement und Berichtssysteme für Geo- und Umweltdaten.
2. Dann erkläre ich kurz, worum es bei Geodaten geht und womit sich die Informatik bei deren Verarbeitung besonders beschäftigt.
3. Es folgt eine Darstellung der Anwendungsgebiete für Geodaten im Allgemeinen und in der Umweltinformatik im Besonderen.
4. Ein erster Exkurs erläutert die Idee der offenen (freien) Geodaten bzw. freien Verwaltungsinformationen, wirft die Frage auf, ob freie Geodaten tatsächlich wie erhofft das Fundament für eine zukünftige florierende Geodatenwirtschaft bilden können und zeigt auf, wie damit einhergehende Initiativen die Menge und Komplexität verfügbarer Geodaten erhöhen können.
5. Nachdem bis dahin Geo- und Umweltdaten im Allgemeinen das Thema waren, wird zwar bezweifelt, dass es sich hier tatsächlich um „Big (!) Data“ im gängigen Sinne handelt, gleichzeitig werden aber auch verschiedene Trends aufgezeigt, die in der näheren Zukunft tatsächlich „große“ Geodatenströme erzeugen könnten, nämlich kostengünstige Fernerkundungstechnologien, kabellose Sensornetze, das Social Web und die Smart City Idee.
6. Ein zweiter Exkurs erläutert die Idee des „Participatory Sensing“ zur Datenerzeugung bzw. -sammlung, die auch ein Zukunftstrend sein könnte, der in naher Zukunft die Verfügbarkeit georeferenzierter Daten erhöht.
7. Abschließend möchte ich zusammenfassen und die Themen meines Vortrags auch in den Kontext der anderen Redner in dieser Vortragsreihe stellen.

1 - Hintergrund: die disy Informationssysteme GmbH

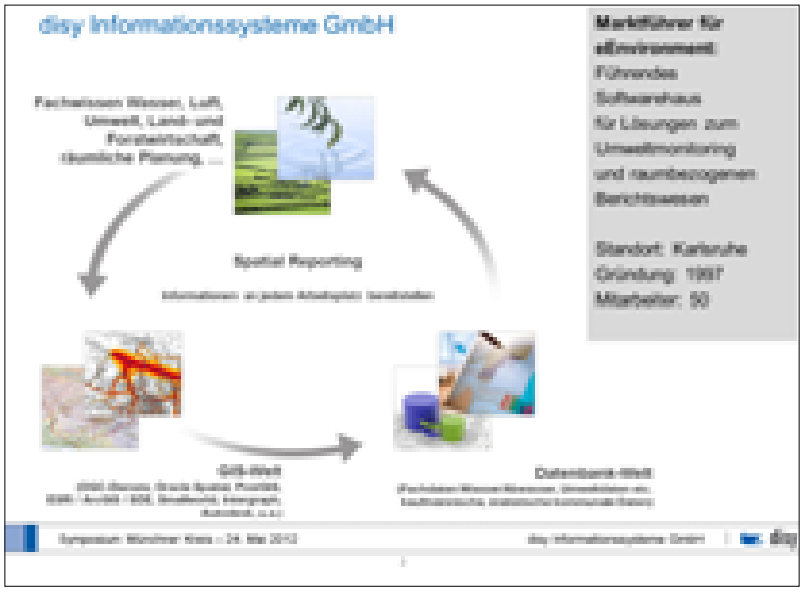

Bild 1

Die disy Informationssysteme GmbH (Bild 1) in Karlsruhe ist ein mittelständischer Software-Lieferant und –Dienstleister mit zurzeit etwa 50 Vollzeitmitarbeitern/innen. disy erstellt auf der Basis eigener Produkte Datenintegrationslösungen und Datenauswertungslösungen im Bereich der Umweltdaten und davon berührter bzw. angrenzender Gebiete, wie Land- und Forstwirtschaft, kommunales Datenmanagement, Lösungen für Stadtwerke usw. Deshalb kommen zurzeit fast all unsere Kunden aus dem Bereich der öffentlichen Verwaltung; es sind dies verschiedene Bundes- und Landesämter, Landkreise, einige große Kommunen. Dort drehen sich die Projekte in aller Regel um Berichtspflichten zu Umweltdaten, wie Lärm- und Luftverschmutzung, Wassergüte und ähnliches. Insofern ist zunächst einmal ein guter Teil des Alleinstellungsmerkmals von disy die langjährige Erfahrung in diesem vertikalen Marktsegment, also das Hintergrundwissen zur Fachlichkeit der Umweltthemen. Technisch gesehen, basiert die Herangehensweise von disy auf einem „klassischen" Data Warehousing Ansatz, bei dem Methoden zum Reporting und der Business Intelligence auf integrierte Daten aus verschiedenen Quellen angewandt werden. Weiterhin haben aber im Umweltbereich fast alle Daten (und auch deren Auswertungen) in natürlicher Weise einen relevanten Orts- bzw. Raumbezug. Man betrachtet z.B. die Gewässerqualität entlang des Flusslaufs und in der Nähe gewisser Einleiterstellen, die Lärmentwicklung zu beiden Seiten der Eisenbahnstrecke, oder die Luftverschmutzung relativ zu den Schornsteinen eines Emittenten (unter Beachtung unterschiedlicher Windrichtungen); man vergleicht zwei Städte hinsichtlich ihrer Feinstaubbelastung, zwei Landkreise hinsichtlich ihres Wertstoffaufkommens im Hausmüll usw.

Aus dieser Bedeutung von Orts- und Raumaspekten hat sich ergeben, dass disy technologisch seine Verfahren zum Data Warehousing, Reporting und Business Intelligence im Bereich der Geoinformation ansiedeln musste. Es geht also immer um Sachdaten mit Orts- oder Raumbezug; das überträgt sich auf die Auswertungen, so dass wir den disy Ansatz als „Spatial Reporting" (räumliches Berichtswesen) bezeichnen. Aber lassen Sie mich zunächst kurz das „Material" beschreiben, mit dem wir arbeiten, nämlich Geodaten.

2 - Ausgangsbasis: Geodaten und Geoinformationssysteme (GIS)

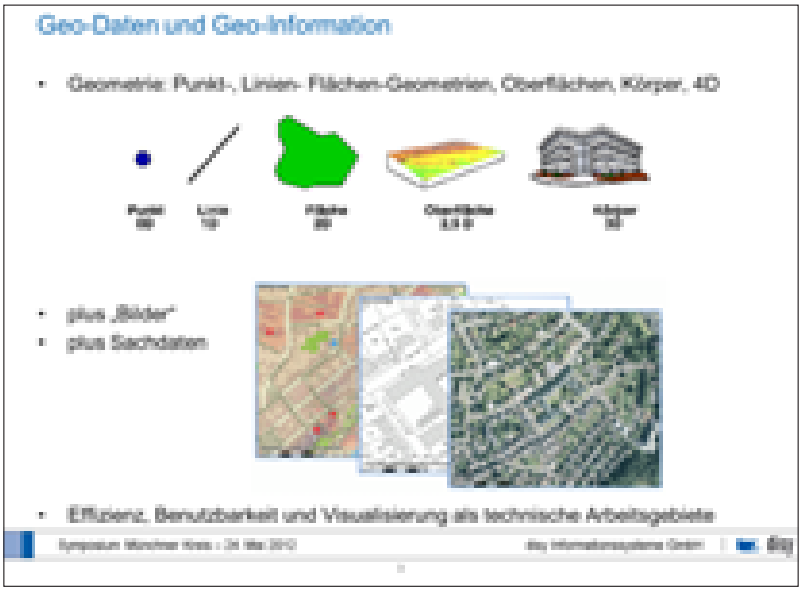

Bild 2

Um die technischen Fragestellungen etwas besser zu verstehen, sei zunächst einmal der Begriff der Geodaten bzw. Geoinformation kurz erläutert (Bild 2). Grundsätzlich geht es bei Daten in Geographischen bzw. Geoinformationssystemen (GIS) zumeist um die Erfassung und Verarbeitung von geometrischen Objekten, die sich über Punktkoordinaten beschreiben lassen (sog. Vektordaten). Die einfachsten möglichen Geometrien sind Punkte und Linien. Linien können schon beispielsweise eine Straße repräsentieren, Mountainbike-Routen, usw. Ein geschlossener Linienzug ergibt ein Polygon, also eine Fläche, wie z.B. ein Bundesland, ein Landkreis, ein Flurstück oder ein Naturschutzgebiet.

Diese Vektordaten zu Punkt-, Linien- und Flächengeometrien werden häufig zusammengeführt mit sogenannten Rasterdaten, das sind pixelbasierte „Bildern" im weitesten Sinne, typischerweise Landkarten, Geländefotos, Pläne o.ä.. In einem gewissen Ausmaß lassen sich natürlich Raster- und Vektordaten ineinander überführen (Erzeugung einer anschaulichen (Raster-)Kartendarstellung aus Vektordaten oder Erzeugung von Vektordaten aus Satelliten- oder Luftbildern durch Bildanalyse). Oder es können Darstellungen von Vektordaten in ein Rasterbild hinein projiziert werden, z.B. Gemarkungsgrenzen.

Zusätzlich zu Raster- und Vektordaten zu geographischen Objekten gibt es in aller Regel eine große Menge von Sachdaten mit Orts- oder Raumbezug, die für eine gewisse Fachanwendung relevant sind. Das könnten z.B. Messwerte zur Luftqualität am Ort einer Messstation sein, Daten zur Verkehrsdichte und deren zeitlichen Verlauf auf einem Autobahnabschnitt oder sozio-ökonomische Daten zur Bevölkerung eines bestimmten Stadtquartiers. Sachdaten können natürlich beliebig komplex strukturiert und vernetzt sein, man könnte sich beispielsweise vorstellen, dass zu einem Acker außer seiner Lage auch Informationen zum Besitzer, zur Bodenstruktur, Bepflanzung, Pestizid-Einsatz, Erträgen der letzten Jahre usw. abgespeichert sind. Vektor-, Raster- und Sachdatenvisualisierungen können natürlich auch sinnvoll in Informationsgraphiken oder Fachkarten integriert werden. So könnte man in ein Luftbild

einer Siedlung die Straßen und Hausnummern projizieren und bei den einzelnen Dachflächen durch künstliche Einfärbung darstellen, welchen Wärmeverlust durch das Dach die Untersuchungen mit Wärmebildkameras ergeben haben.

Soweit haben wir bisher nur ein- und zweidimensionale Geometrien betrachtet. Der nächste Schritt sind sog. zweieinhalb-dimensionale Geometrien (2,5D). Bei diesen kann jedem Punkt im zweidimensionalen Modell ein weiterer einfacher Attributwert zugeordnet werden, z.B. für eine Höhen- oder eine Zeitangabe. Damit lässt sich z.B. für eine Mountainbike-Route auch das Geländeprofil mit darstellen oder in einer Planungskarte für Windkraftanlagen die Windgeschwindigkeit in 40, 60 und 80 Meter Höhe über Grund beschreiben.

Echte 3D-Datenbanken verarbeiten dann als Objekte erster Ordnung (mit spezifischen Operatoren und eigener Semantik) wirkliche dreidimensionale Körper. Typisches Beispiel sind „Bauklötzchenbilder“, z.B. für 3D-Stadtmodelle, mit denen man die Stadtplanung anschaulicher darstellen kann, die aber z.B. auch als Basis für Lärmausbreitungsberechnungen dienen können. Ein anderes Beispiel wären 3D-Modelle eines Grundwasserkörpers unter einer Tiefbaustelle. Die Fortsetzung sind dann 4D-Datenbanken, mit denen sich zurzeit primär noch die Forschung befasst. Hier geht es um Körper, die sich in der Zeit bewegen, z.B. eine Schadstoffwolke bei einem Chemieunfall oder eine Verunreinigung in einem Wasserkörper, die sich ausbreitet.

Soviel zur Datengrundlage. Aber brauche ich dafür wirklich spezifische Software-Werkzeuge, wieso kann ich Geodaten nicht mit konventionellen Werkzeugen (relationale Datenbanken u.ä.) verarbeiten? Das geht *im Prinzip* natürlich schon. Die Forschungs- und Entwicklungsthemen der Geoinformatiker und GIS-Entwickler betreffen hauptsächlich die Effizienz, Benutzerfreundlichkeit und problemangemessene Verständlichkeit von GIS-spezifischen Operationen. So könnte man beispielsweise natürlich ein 4D-Phänomen auch „naiv“ in einer relationalen Datenbank repräsentieren; aber die Anfragebearbeitung für 4D-typische Fragestellungen würde voraussichtlich extrem ineffizient. Eine Darstellung in konventionellen Informationssystemen würde auch von Haus aus keine *natürlichen* Anfrageformen oder Verarbeitungsbefehle bereitstellen, um problemadäquat Aufgabenstellungen formulieren zu könne. Im 2D-Bereich wären Beispiele für raumbezogenen Anfragen: Gib mir alle Städte in 10 km Abstand Fahrstrecke von Autobahnabfahrt XY. Finde alle Hausbesitzer von Häusern in 30 km Umkreis einer Chemieanlage. Bestimme alle Fast-Food Restaurants innerhalb von 60 Minuten Autofahrt auf Bundesstraßen!

Hier geht es also um inhärent raumbezogene Anfrage- und Verarbeitungsmöglichkeiten. Thema der GIS-Forschung und –Entwicklung ist es, Systeme so aufzubauen, dass entsprechende, benutzerfreundliche Funktionalitäten effizient implementiert werden können. Häufig befasst man sich auch mit Visualisierungs- und Darstellungstechniken, um komplexe Sachverhalte und Phänomene nachvollziehbar darzustellen – traditionell durch zweidimensionale Visualisierungen, also Kartendarstellungen, Grafiken, Bilder, neuerdings auch durch in 3D-Darstellungen (z.B. in Virtual-Reality Umgebungen) oder eingeblendet in Darstellungen der realen Welt („Mixed Reality“, „Augmented Reality“). Als Beispiel sei das Projekt MAGUN an der HTW Berlin genannt, bei dem man sich „durch sein Smartphone“ anschauen kann, wie eine bestimmte Gegend bei verschiedenen Hochwasserständen aussehen würde.

3 - GIS-Anwendungen

Bild 3

Gehen wir zu wichtigen Anwendungsbranchen und -feldern für GIS-Technologien im Allgemeinen und Umweltinformationssystemen im Besonderen über (Bild 3):

Logistik und Verkehr sind natürlich die naheliegendsten Anwendungsfelder für Geoinformation, also insbesondere Routenplanung, auch in Kombination mit Echtzeitinformationen (Verkehrsdichte, Baustellen, …), um schnell, günstig oder umweltschonend ein gewisses Ziel zu erreichen, eventuell auch verkehrsträgerübergreifend. Die öffentliche Verwaltung nutzt Geodaten für vielfältigste planerische Tätigkeiten, im Rahmen der Raum- und Umweltplanung oder der Stadtplanung, für Planfeststellungsverfahren und Bauplanung, Flurneuordnung, zahlreiche Berichts- und Überwachungszwecke und vieles mehr.

Ein kommerziell sehr bedeutendes Anwendungsfeld ist das sog. Geo-Marketing. Es gibt Firmen, die hausgenau aufgelöste sozio-ökonomische Datenbeständen zur Demographie einzelner Siedlungen besitzen, auf deren Basis man punktgenau Werbemaßnahmen planen kann. In diesen Bereich zählen auch Standortplanungen für Geschäftsstandorte oder Gebietsplanungen für Vertriebsgebiete.

Auch im Bereich der Land- und Forstwirtschaft gibt es vielfältige visionäre Anwendungen, z.B. im sog. „Precision Farming“. Es gibt heutzutage im Experimentalstadium schon Ernte- oder auch Düngemaschinen mit hochentwickelter eingebauter Sensorik, welche aus dem Grün einer Pflanze eine Hypothese darüber aufstellen kann, wie gut oder schlecht es dieser speziellen Pflanze gerade geht. Mit diesem Wissen kann man diese einzelne Pflanze gezielt düngen kann dadurch den Düngemitteleinsatz (genauso wie bspw. den Pestizideinsatz) punktweise optimieren – was ökonomisch und ökologisch sinnvoll ist. Denkt man das Ganze etwas weiter und kombiniert z.B. eine solche „in-situ“ Beobachtung (vor Ort) mit Daten der Erdbeobachtung von einem Satelliten aus, könnte sich z.B. herausstellen, dass an einem bestimmten Ort die gesamte Gegend gerade eine schlechte Vegetation zeigt (boden- oder klimabedingt), woraus sich eventuell schließen lässt, dass eine bestimmte Düngestrategie gar

nicht erfolgversprechend sein kann, sondern man grundsätzlichere Maßnahmen ergreifen muss; oder es lässt sich eine herannahende Schlechtwetterfront oder ein naher Schädlingsbefall sichten, was wiederum zu anderen Behandlungsstrategien führt. Hier kann man schon zukünftige Szenarien erahnen, die sich in den Bereich der „ Big Data“ bewegen, wenn man nämlich Beobachtungen vor Ort (wie dem Düngegerät) mit Hintergrundwissen und Daten der Erdbeobachtung kombinieren möchte, und das am liebsten in Echtzeit.

Weitere naheliegende Anwendungsgebiete für Geodaten finden sich im Katastrophen- und Zivilschutz (sowohl in der Frühwarnung als auch bei der Notfallplanung und im Notfallmanagement), beim Militär, im Immobilienbereich und im Tourismus (auf persönlichen Präferenzen basierte Reiseplanung oder Auswahl von Routen, Hotels, etc., ggf. mit auf das Smartphone übertragbarem Reiseplan und lokationsbasierten Empfehlungen für Restaurants, Freizeitangebote, Hotels, etc.).

Wir haben soweit primär Anwendungen von *Geodaten* betrachtet. Vertieft man etwas mehr in den Bereich der Umweltinformatik, also das angestammte Arbeitsfeld von disy, ergeben sich noch einige andere Anwendungsgebiete. Typischerweise geht es dabei um Anwender aus der öffentlichen Verwaltung, die nationale oder europäische Berichtspflichten (z.B. im Rahmen der europäischen Umgebungslärmrichtlinie, der europäischen Wasserrahmenrichtlinie, der Hochwasserrisikomanagementdirektive usw.) erfüllen müssen. Langfristig kann man aber auch davon ausgehen, dass Umweltmonitoring und umweltbezogenes Reporting auch für Firmen zunehmende Relevanz gewinnen wird – hier geht es z.B. um die ökologische Bewertung von Produktlebenszyklen oder um den CO_2-Fußabdruck von Produkten und Produktionsprozessen (als Schlagwort sei hier das Thema „Corporate Environmental Management Information System, CEMIS“ genannt).

Wissenschaft und Politik übernehmen auch zunehmend Konzepte der integrierten ökologisch-ökonomischen Betrachtung komplexer Sachverhalte, z.B. beim Integrierten Küstenzonenmanagement (engl. „Integrated Coastal Zone Management, ICZM“) oder beim Integrierten Wasserressourcenmanagement (IWRM). Betrachtet man zum Beispiel letzteres, so ist es heute von Interesse, auf der Basis von Pegelstandsmessungen und von Wettervorhersagen in einem Flusseinzugsgebiet die Hochwasserwellen flussabwärts vorherzusagen; in wenigen Jahren wird man vielleicht zusätzlich zu Sensornetzen an Flüssen und wasserwirtschaftlichen Anlagen (Staubecken, Schleusen, …) flächendeckend sog. „Smart Meters“ bei den Wasserendverbrauchern und an strategisch wichtigen Stellen im Verteilnetz haben, welche in (Nah-) Echtzeit feingranulare Verbrauchsdaten liefern, die es – zusammen mit intelligenten Prognosemodellen – erlauben, die aktuellen und die in naher Zukunft erwarteten Wasser-Angebote und -Verbräuche so gegeneinander abzugleichen, dass sich das Gesamtsystem hinsichtlich der Versorgung und des Energieverbrauchs optimieren lässt.

Dabei kennt man heutzutage den Begriff des „Smart Metering“ noch eher aus dem Bereich der Stromversorgung, mit der Zielsetzung des sog. „Smart Grid“ – zu dessen intelligenter Steuerung es natürlich auch in hohem Maße raumbezogener Daten und raumbezogener Datenanalysen (z.B. gebietsbezogene Bedarfsprognosen und gebietsbezogene Erzeugungsprognosen aus erneuerbaren Quellen, Schwachstellenanalyse im Niederspannungsnetz, etc.) erfordert, um Energieangebot und -nachfrage im Stromnetz vernünftig zusammenzubringen.

Im Zuge der „Energiewende“ findet man in diesem Kontext in den letzten Jahren auch eine explodierende Anzahl wissenschaftlicher Publikationen zur Bestimmung des Energieerzeugungspotenzials an einem bestimmten Punkt oder in einer Region, für verschiedene Typen regenerativer Energien. Hierfür sind natürlich Geodaten von größter Bedeutung. Das betrifft offensichtlich Windgeschwindigkeiten über Grund für Windenergieanlagen; aber auch die

„Rauigkeit“ eines Geländes hat einen Einfluss auf die Effizienz einer Windenergieanlage. Geodaten kann man auch nutzen für Sichtbarkeitsanalysen von Windrädern, also die Frage, wie die Landschaft aus verschiedenen Standpunkten optisch durch neue Bauwerke verändert wird.

In ähnlicher Weise können Geodaten noch für viele andere planerischen Aufgaben bei regenerativen Energien genutzt werden. So sind für Solaranlagen auf Hausdächern in Wohngebieten deren Sonnenstunden, die Sonnenexposition und Dachneigung relevant, aber auch Verschattungsmöglichkeiten durch umliegende Gebäude – dabei geht es um hochgradig spezifisches räumliches Schlussfolgern; für die Erzeugung von Biogas ist es nicht nur wichtig für die betriebswirtschaftliche und ökologische Sinnhaftigkeit, welche Mengen verwertbarer Biomasse in der Umgebung anfallen und wie diese sich zur Anlage transportieren lassen, sondern auch, wo Verwerter für das entstehende Biogas existieren, zum Beispiel in Form nahegelegener Blockheizkraftwerke. Denkt man noch weiter und plant noch intelligentere Unterstützungssoftware, müssen natürlich auch raumplanerische Gegebenheiten (Vorzugsflächen, Ausschlussflächen) und Aspekte der Flächenkonkurrenz beachtet werden (ist eine Landnutzung für Energiepflanzenanbau einem Nahrungsmittelanbau vorzuziehen, wenn ja, mit welcher Pflanze, usw.). Die Nutzer solcher Geodaten zur Unterstützung planerischer Aufgaben obliegen naturgemäß zunächst den öffentlichen Verwaltungen, aber ebenso könnten hier z.B. auch potenzielle Investoren für Energieanlagen aktiv werden.

4 - Open Data und die „Emerging Data Economy“

Als Exkurs, der eng mit dem Thema Geodaten zusammenhängt, der aber auch als ein Startpunkt für die Verfügbarkeit von „Big Geodata“ vorstellbar ist, lassen Sie mich kurz auf aktuelle Trends und Entwicklungen im Bereich offener Geodaten bzw. offener Verwaltungsinformation („Public Sector Information, PSI“ / „Open Government Data, OGD“) eingehen. Seit vielen Jahren wird von einschlägigen Fachverbänden und Initiativen gefordert, Daten der öffentlichen Verwaltung, die steuerfinanziert gesammelt bzw. erzeugt wurden, im großen Stile kostenfrei offenzulegen, um damit eine Initialzündung für die sog. „Data Economy“ zu geben, bei der ein ganzes Ökosystem von Datenverwertern und Mehrwertdienstleistern entstehen könnte, das auf der Basis der offenen Verwaltungsdaten neue Anwendungen (und insbesondere Smartphone-Apps zum lokationsbasierten Datenkonsum) für Bürger oder die Wirtschaft schaffen würde.[1] Dies würde technisch noch erleichtert, wenn man den Prinzipien der *„Linked Open Data (LOD)“* folgen würde, die Prof. Studer in seinem Vortrag schon vorgestellt hat. In der Forschung und in innovationsnahen Kreisen genießen die Themen Open Data und LOD seit einigen Jahren eine enorme Bedeutung, die auch immer noch wächst.

Meines Erachtens ist es aber noch eine recht ungeklärte Frage, ob dies tatsächlich zeitnah geschehen wird und was man dafür tun muss. Zumindest gibt es zunehmend Initiativen und Rahmenbedingungen, die dies begünstigen. Im Zusammenhang mit Geo- und Umweltdaten sind es mindestens die folgenden europäischen Richtlinien, Initiativen und Großprojekte, die hier förderlich wirken:

- **SEIS: Shared Environmental Information System**[2] fördert ein gemeinsames europäisches Umweltinformationsnetzwerk mit einfachem und freiem Zugang zu umweltrelevanten Informationen für verschiedene Stellen, inklusive der interessierten Öffentlichkeit.

[1] Siehe z.B. http://www.cloud-finder.ch/uploads/media/2012-06-OGD_Studie_Schweiz.pdf für eine recht aktuelle Studie.
[2] Siehe z.B. http://www.umweltbundesamt.at/umweltsituation/umweltinfo/ui_initiativen/seis/

- **INSPIRE: Infrastructure for Spatial Information in the European Community**[3] ist eine EU-Direktive, welche die Mitgliedsstaaten verpflichtet, ihre Geodaten schrittweise in interoperablen web-basierten Geodateninfrastrukturen verfügbar zu machen (INSPIRE verlangt jedoch nicht die kostenfreie Bereitstellung für den Bürger). Sie wird in Deutschland beispielsweise durch die Geodatenzugangsgesetze des Bundes und der Länder umgesetzt.
- **GMES (Global Monitoring for Environment and Security) und GEOSS (Global Earth Observation System of Systems)** sind von der EU und anderen Staaten getragene, langjährige Initiativen, um die Erdbeobachtung von Satelliten aus zu befördern und mit Methoden der Fernerkundung Fragestellungen zum Klimawandel, zu Katastrophen- und Krisenmanagement, zur Überwachung der Meeresumwelt u.v.m. besser beantworten zu können.

All diese Initiativen tragen dazu bei, dass langfristig sehr viel breitere Kreise aus Verwaltung, Wissenschaft und Wirtschaft einfacher und kostengünstiger auf mehr, vielfältigere und qualitativ hochwertigere Daten über die Erde und die Umwelt zugreifen können. Somit steigen die Chancen, dass mittelfristig auch im Bereich der Geo- und Umweltdaten tatsächlich „Big (!) Data" zur Verfügung stehen und nicht nur eher „Medium-Sized Data". Aber dazu weiter unten mehr. Zunächst wollen wir noch bei der Frage bleiben, ob sich aus offenen Geodaten der öffentlichen Verwaltung zwangsläufig eine Daten-Ökonomie entwickeln wird. Meine klare Antwort hierzu: *Ich weiß es nicht!* Zurzeit sehe ich noch enttäuschend wenig konkrete Anzeichen und nur wenige überzeugende Geschäftsmodelle. Allerdings scheinen mir einige Aussagen zu diesem Themenkomplex belastbar zu sein:

- Hemmnisse für die Entstehung einer Datenökonomie sind heute m.E. nicht primär *technischer* Natur, sondern eher administrativer, juristischer oder sozio-ökonomischer Art (also z.B. Lizenzfragen, Haftungsfragen oder auch die Frage, ob man technisch, fachlich und personell die Verwaltung in den Stand versetzt, politisch gewünschten Veröffentlichungspflichten sinnvoll nachzukommen).
- Eine kritische Frage beim nächsten Schritt scheint mir die nach Geschäftsmodellen und *Anreizsystemen* zu sein.
    - Warum sollte jemand, der die Daten heute schon verkaufen kann, sie morgen verschenken? Was Daten der öffentlichen Verwaltung angeht, kann dies durch politische Entscheidung erfolgen. Für (bereits existierende oder noch zu schaffende) private Datenbestände ist die Frage schon schwieriger zu beantworten. Vorstellbar wären immerhin kollaborativ erzeugte Datenpools (vgl. weiter unten, z.B. „Social Web" oder „Participatory Sensing"; ein sehr interessantes Beispiel hierfür ist schon heute die „Volunteered Geographic Information (VGI)", insbesondere Open Street Map) zur allgemeinen Nutzung oder von der Wissenschaft bereitgestellte Datenbestände. Aktuell ist hier aber die Lage noch eher die, dass traditionelle Geschäftsmodelle (Routing-Anwendungen) gegenüber freien Angeboten verschwinden. Dass in diesem Kontext auch neue Arbeitsplätze oder volkswirtschaftlicher Nutzen in nennenswerter Menge entstehen, steht noch aus. Hier wäre allerdings auch die volkswirtschaftliche Forschung gefragt, vielleicht ist ja der gesamtwirtschaftliche Nutzen freier Routing-Anwendungen größer als die Geschäftsverluste vormals proprietärer Anbieter?
    - Welche Bezahlmodelle sind für Daten-Apps oder Mehrwertdienste realistisch und welche Preise können hier erzielt werden? Offensichtlich ist man heutzutage noch

[3] Sie z.B. http://de.wikipedia.org/wiki/Infrastructure_for_Spatial_Information_in_the_European_Community und http://www.geoportal.de/

eher selten bereit, für Daten im Netz zu zahlen. Man braucht also entweder sehr geringe Nutzungskosten und sehr große Nutzerzahlen, werbefinanzierte Modelle oder Sponsormodelle (z.B. dass eine Krankenkasse die Pollenwarnungen für ihre Mitglieder finanziert).

Insgesamt ist das Gebiet der offenen Daten sicherlich ein hochspannendes Feld (und einige Staaten – USA, Großbritannien, … – haben uns hier schon einiges voraus), auf dem auch die ein oder andere deutsche Kommune in jüngster Zeit sehr aktiv wird (z.B. Berlin, München, Stuttgart u.a.). Auch die EU und verschiedene Bundesministerien fördern hier immer wieder Forschungsprojekte, Wettbewerbe und Pilotanwendungen. Trotzdem denke ich, dass wir hier noch einige gute Ideen brauchen, um schnell einen neuen Wirtschaftsboom zu sehen. Das betrifft insbesondere die nichttechnischen Fragestellungen und Anwendungsideen.

Zu Ende des Exkurses zu „Open Data“ und zur Frage, ob „Open Governmental Data“ oder „Linked Governmental Data“ tatsächlich die Initialzündung für eine „Data Economy“ geben können, sei der Vollständigkeit halber auch erwähnt, dass sehr häufig auch die Wissenschaft der Nutznießer von offenen (und auch von großen) Datenmengen sein kann und dass auch dies der Gesellschaft zugutekommen kann. In den letzten Jahren gibt es bspw. zunehmend Untersuchungen im Bereich der „Geomedizin“ oder der „Medizinischen Geographie“, wo es darum geht, aus Umgebungsfaktoren statistisch belegbar herauszufinden, welchen Einfluss diese auf den medizinischen Status von Populationen haben. Soviel zum Exkurs in Sachen „Open Data“. Jetzt aber noch einmal eingehender zum Begriff der „Big Data“.

5 - Big Geodata?!

Soweit haben wir uns noch allgemein mit der Situation der „Geodaten an sich“ befasst. Sind das aber auch Big Data im Sinne dieses Symposiums? Betrachtet man die typischen Definitionsansätze für Big Data, werden häufig die Kriterien Variety, Velocity und Volume genannt. Dazu kann man sagen, dass man im Geodatenbereich heutzutage eine hohe Variabilität (Variety) von Datenformaten, -quellen usw. haben kann, aber für einzelne, spezifische Anwendungen eher selten tatsächlich haben wird. Was die Geschwindigkeit von Datenerzeugung und -austausch (Velocity) angeht, gibt es durchaus einige Anwendungen, wo bereits in Echtzeit oder Nah-Echtzeit Datenströme entstehen (z.B. in der Verkehrsstromüberwachung), aber auch hier gehen wir in der nahen und mittleren Zukunft von einer deutlichen Zunahme aus. Betrachtet man schließlich das prominenteste Attribut, die Datenmenge (Volume), gibt es zwar natürlich einzelne Anwendungsgebiete mit sehr großem Datenaufkommen (z.B. Klimamodelle), aber konkrete einzelne Anwendungen funktionieren doch in der Regel eher mit mittleren Datenmengen. Das bedeutet insgesamt, dass für alle drei definitorischen Kriterien die Tendenz dahin geht, dass das Attribut „big“ zwar heute oder in naher Zukunft schon zutreffend sein könnte, dass es aber für die heute üblichen Anwendungen gar nicht notwendig ist, über das Maß von „medium-sized“ hinauszugehen. Trotzdem möchte ich folgende Prognose wagen:

Die Menge an Geodaten wird wachsen. Das Wachstum selber wird sich beschleunigen. Es gibt weitere Aspekte, die dafür sprechen, dass die Menge verfügbarer Geo- und Umweltdaten, auch Echtzeitdaten zu diesen Themen, schnell (und schneller werdend) wachsen wird. Dazu einige Bemerkungen:

1) Sog. „Unmanned Aerial Vehicles“ (UAV, unbemannte Luftfahrzeuge: Drohnen, ferngesteuerte Heli- und Quadrokopter jeder Größe, etc.) werden kleiner, preisgünstiger und

technisch leistungsfähiger – dasselbe gilt für die von solchen Geräten zur Fernerkundung nutzbare Sensorik (Kameras) und die Software zur Steuerung und Datenauswertung (Bild 4). Damit wird es immer einfacher und kostengünstiger, auch für kleinere und kommerzielle Anwendungsfälle Fernerkundungsmethoden zur Datenerzeugung zu nutzen, die bisher nur dem Militär und der öffentlichen Verwaltung zugänglich waren.

2) Auch kostengünstige und teilweise in Echtzeit kabellose mit Basisstationen kommunizierende Sensoren / Sensornetze für die verschiedensten Phänomene werden immer kostengünstiger und gleichzeitig leistungsfähiger (z.B. zur Messung von Luft- oder Wasserverschmutzungen, zur Beobachtung rutschungsgefährdeter Steilhänge u.v.m.). Im Extremfall geht das bis zum Forschungsansatz des sog. „Smart Dust", wo man mit Nanotechnologie winzige Sensorknoten baut, die man dann zu Zehntausenden verbreiten kann, um flächendeckend gewisse Phänomene zu beobachten.

Bild 4

3) Das Thema „Social Web", also die Menge der von Internetnutzern kollaborativ erzeugten Daten und Informationen, bspw. in Form von Blogs oder Tweets, ist ja schon des Öfteren in diesem Symposium aufgetaucht. Auch solche benutzergenerierte Informationen haben natürlich sehr häufig einen expliziten oder impliziten (und dann evtl. durch automatische Georeferenzierung maschinell herleitbaren) Orts- oder Raumbezug. Auch solche Informationen können u.U. fallbezogen in großer Zahl entstehen; im Beispiel von Naturkatastrophen und Rettungsaktionen gibt es viele nützliche Ideen, wie sich Beobachtungen von außen (z.B. durch Fernerkundung) und vor Ort (durch Bürger eines Katastrophengebiets) sinnvoll kombinieren lassen, um Rettungskräften ein vollständigeres und detaillierteres Lagebild zu geben.

Verschiedene der oben genannten Aspekte, insbesondere die der kabellosen Sensornetze, spielen bei der Vision der „Smart City" zusammen, die früher oder später, in kleinen, aber merklichen Schritten, sicher in einigen der großen Metropolen oder der führenden Innovationshauptstädte umgesetzt wird, um die drängenden Probleme der Ballungszentren beim Verkehr, Energieeinsparung, Umweltschutz usw. zu lösen. Je weiter solche Ansätze um sich

greifen, (deren Nützlichkeit häufig vom Nah-Echtzeitaspekt lebt), desto mehr Daten werden über Kurz oder Lang für „Big Data Analytics“ zur Verfügung stehen. Und praktisch alle in solchen Szenarien anfallenden Daten haben in natürlicher Weise einen Orts- oder Raumbezug, der auch häufig relevant für die sinnvolle Nutzung der Daten ist.

Beispiel für Sensordatenfusion in Echtzeit: Eye on Earth.

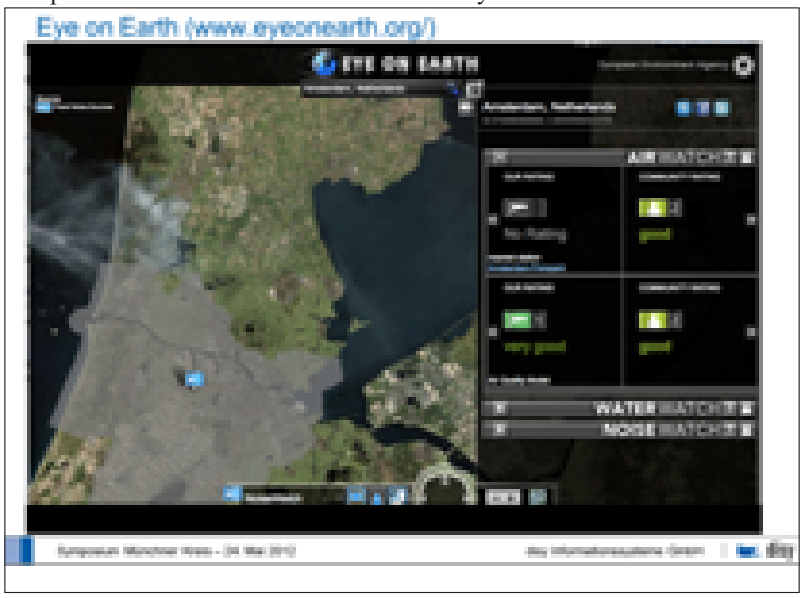

Bild 5

Interessante Perspektiven ergeben sich immer, wenn man versucht, verschiedene solche Datenquellen zusammenzuführen. Als sehr interessantes Beispiel für einen solchen Ansatz lassen Sie mich auf das europäische Leitprojekt („Integrating Project, IP“) „Eye on Earth“ hinweisen (Bild 5). Dort kombiniert man einerseits Umweltmessdaten aus offiziellen staatlichen Messnetzen (Luftqualität aufgrund der Beobachtung von Stickoxiden, Feinstaubbelastung und Ozonwert, Wasserverschmutzung und Lärmbelastung an 22.000 Messpunkten europaweit) und andererseits von Bürgern per SMS beigesteuerte Benutzerbewertungen zu diesen drei Themen. Auch wenn man am momentanen Stand der Umsetzung sicherlich noch vieles kritisieren kann, verwendet das Projekt meiner Meinung nach einen sehr interessanten Grundansatz und realisiert beispielhaft auch eine leistungsfähige cloud-basierte Software-Architektur. Sowohl die Idee (Zusammenführung staatlicher Daten und benutzergenerierter Information in Echtzeit) als auch der Umsetzungsansatz können sicher Anregungen für weitere ähnliche Anwendungen geben.

6 - Exkurs: Participatory Sensing zur Geodatenerzeugung

Bild 6

Ich hatte weiter oben unter anderem zwei Trends genannt, die m.E. ein wachsendes Volumen an Geodaten begründen, nämlich die zunehmend leistungsfähigeren Sensornetze und das Social Web. Gewissermaßen in der „Schnittmenge“ beider Trends liegt die Idee des „Participatory Sensing“ (auch „Urban Sensing“, „Citizen Sensing“, „People-as-Sensors“ u.v.m.),[4] welches aktuell von disy zusammen mit dem Forschungszentrum Informatik Karlsruhe im vom Bundesministerium für Bildung und Forschung (BMBF) geförderten KMU-innovativ Projekt „PartSense“ [5] untersucht und vorangetrieben wird (Bild 6). Die einfache Grundidee ist die, dass die Abertausenden von Menschen, die täglich überall mit einem leistungsfähigen, Internet-fähigen und GPS-bewehrten Computer inklusive einer Menge hochkomplexer Sensorik unterwegs sind – nämlich einem Smartphone – diese Ausrüstung auch nutzen, um im Rahmen definierter Mess- oder Datenerfassungskampagnen bestimmte Informationen von Interesse vor Ort aufzunehmen und einer zentralen Auswertestelle zu melden. Man wendet also die Web2.0-Idee des „Crowdsourcing“ auf Aufgaben zur Datensammlung an. Das beginnt bei einfachen Apps zur *Mängelmeldung*, mit denen Bürger Mängel an ihrer
städtischen Infrastruktur (defekte Straßenbeleuchtung, schlecht geschaltete Verkehrsampeln, überfüllte Müllcontainer, Schlaglöcher etc.) vor Ort mit einem Foto dokumentieren und mit Zeitstempel und Ortsangabe an die Stadtverwaltung schicken können. Die Illustration oben zeigt ein Beispiel für den häufig untersuchten Fall der *Lärmerfassung*: Bürger nehmen mit ihrem Smartphone-Mikrophon unterwegs Lärmpegel auf und ein zentraler Server kann daraus eine empirisch basierte Lärmkartierung erstellen. Ähnliche Pilotprojekte gibt es z.B. zur *Luftqualität* (mit mobilen Sensoren, die über Bluetooth für die Datenweitergabe mit einem Smartphone kommunizieren), im *Naturschutz* (Sichtungen seltener Pflanzen oder Tiere) u.v.a.m. Erfasst man nicht nur an einem Punkt verortete Daten, sondern z.B. eine gefahrene *Mountainbike-Strecke* (inkl. Höhenprofil), kann man sogar mithilfe des Erschütterungssensors im Smartphone Hypothesen liefern, wie die Oberflächenbeschaffenheit dieser Strecke ist.

[4] Siehe z.B. http://www.wilsoncenter.org/sites/default/files/participatory_sensing.pdf
[5] Participatory Sensing für Firmen und die öffentliche Verwaltung. Gefördert vom BMBF unter dem Förderkennzeichen 01IS11029A und 01IS11029B.

Insgesamt gehen wir davon aus, dass die Zukunft noch vielfältigste Anwendungsideen für Participatory Sensing sehen wird. Diese erhöhen natürlich das Aufkommen georeferenzierter Datenströme stark.

7 – Zusammenfassung und Abschluss

Lassen Sie mich die wesentlichen Themen des Vortrages, auch im Kontext der anderen Vorträge dieser Sitzung (von Herrn Prof. Dr. Studer, Herrn Dr. Steinacker und Herrn Kuhlmann), noch einmal abschließend bewerten (Bild 7). Ich habe einerseits dargestellt, dass Geo- und Umweltdaten bereits heute an vielen Stellen eine wichtige Rolle zur zukunftsfähigen Gestaltung unserer Wirtschaft und Gesellschaft spielen. Mit Bezug auf die typischen Kriterien der „Big Data“ (Volume, Velocity, Variety) reden wir da jedoch bei den heutigen Anwendungen noch eher über mittelgroße denn über „sehr große“ Dimensionen. Ich habe allerdings auch ausführlich beschrieben, dass m.E. Entwicklungen in den Bereichen drahtlose (Nah-Echtzeit-)Sensornetzwerke, kostengünstige Fernerkundung durch Satelliten und UAV, benutzererzeugte Geodaten (Volunteered Geographic Information, Social Web, Participatory Sensing), Smart City und Open Governmental Data mittelfristig zu einer Explosion der für Bürger, Wirtschaft und Wissenschaft verfügbaren Geo- und Umweltdaten führen kann, welche dann die Themen der Big Data fraglos berührt.

Bild 7

Wenn dies passiert, sind zweifellos vielfältigste Fragen der Sensordatenfusion, der automatischen Georeferenzierung, der semantischen Datenintegration usw. zu lösen. Für intelligente Auswertungen solcher Daten braucht es sicherlich ein tiefergehendes Verständnis und effiziente Implementierungen der Semantik orts- und raumbezogener Aussagen (und Anfragen) wie auch effiziente Methoden zur Erkennung komplexer Ereignismuster in Datenströmen. Um schließlich wirklich intelligente computerbasierte Entscheidungsunterstützung auf der Basis von Geo- und Umweltdaten zu schaffen, sind darüber hinaus komplexe Ansätze aus Expertensystemen, Operations Research und Wissensmanagement zu verknüpfen.

Die technischen Ansätze hierzu in der Geoinformatik hinken an manchen Stellen der Kerninformatik etwas hinterher. Allerdings gibt es sehr viele analoge oder eng verwandte Entwicklungen in Kerninformatik und Geoinformatik, namentlich im Bereich Semantisches Web und Internet der Dienste, welche eine weiterführende Konvergenz begünstigen. So stellen die „Semantiker" das Hintergrundwissen eines Anwendungsgebiets häufig in sog. Ontologien dar. Dagegen gibt es in der Umweltinformatik seit vielen Jahren eine große Tradition zu Umweltthesauri, welche zumindest als leichtgewichtige Wissensorganisationssysteme und für die Lieferung von Hintergrundwissen verstanden und genutzt werden können. Ein weiterer zentraler Aspekt des Semantischen Web ist die Idee ausdrucksstarker Metadaten für elektronische Daten, um deren Maschinenverarbeitbarkeit zu erleichtern. Hierzu gibt es in Geodateninfrastrukturen viele Vorarbeiten und allgemein akzeptierte Standards. Grundsätzlich ist die Standardisierung von Protokollen, Datenmodellen und Repräsentationssprachen auch einer der wichtigsten pragmatischen Erfolgsfaktoren des Semantischen Web. Gleiches gilt in der Geoinformatik, nur dass man sich dort weniger mit den Standards des World Wide Web Consortium (W3C) befasst als mit denen des Open Geospatial Consortium (OGC). Schließlich basieren auch verteilte Geodateninfrastrukturen in aller Regel auf dienstebasierten Web-Architekturen. Das Thema der Linked Environmental Data diffundiert langsam aus dem Semantischen Web in den Umweltinformatik-Bereich und findet dort zunehmend Interesse.

Bild 8

Viele dieser Themen haben wir im THESEUS KMU-Projekt HIPPOLYTOS ansatzweise untersucht, um die Nutzung von Umwelt- und Geodaten mit Techniken des Semantischen Web einfacher und effektiver zu machen (Bild 8). Dennoch gibt es meines Erachtens in der Zukunft noch genügend Ansatzpunkte für die nutzbringende Verquickung semantischer Technologien mit Themen der Geo- und Umweltinformatik. Der Übergang zu Big Data kann danach den nächsten Quantensprung darstellen, hisichtlich der Chancen ebenso wie hinsichtlich der Herausforderungen!

11 Intelligente Geschäftsanbahnung für Produkte und Personen mit Linked Open Data im WWW

Dr. Achim Steinacker, intelligent views GmbH, Darmstadt

Intelligente Geschäftsanbahnung für Produkte und Personen mit Linked Open Data im WWW. Vielleicht direkt einen Satz dazu. Zum einen, wie ich seit heute Morgen weiß, hat das eigentlich mit Big Data gar nichts zu tun. Ich dachte, dass wir damit schon immer eine ganze Menge mit Daten machen. Aber für die Dimension heute Morgen ist es wohl doch ein bisschen klein. Zum anderen geht es jetzt eigentlich mehr darum, wie ich denn diese ganzen Standards rund um Linked Open Data, die auch Herr Studer vorgestellt hat, tatsächlich innerhalb eines Unternehmens einsetzen. Das dritte ist, dass wir den Theseus Kontext doch noch nicht so ganz verlassen haben. Was ich über das Projekt erzählen werde und woher es kommt, speist sich aus sehr vielen Ideen, die man in Theseus bearbeitet hat.

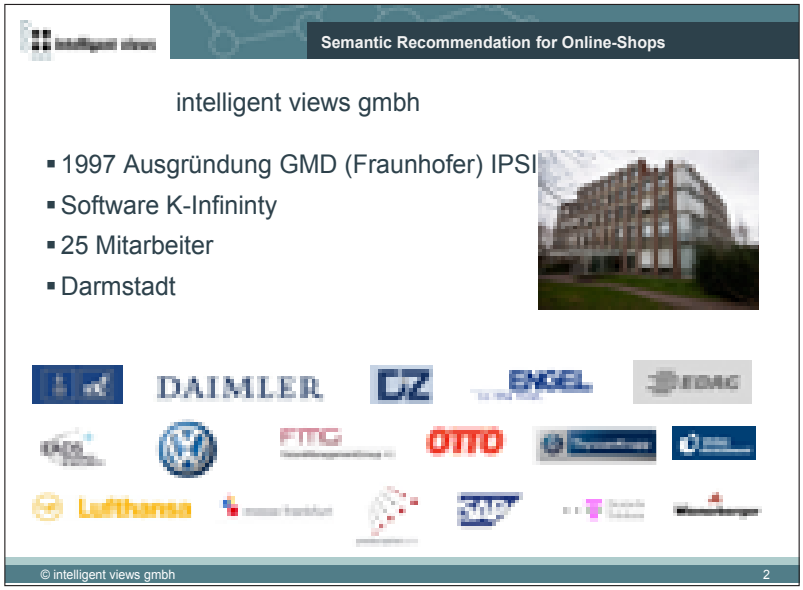

Bild 1

Ein Satz zu unserer Firma intelligent views (Bild 1). Wir sind eine Ausgründung aus dem Forschungsinstitut „Integrierte Publikations- und Informationssysteme (IPSI)“ der damaligen Gesellschaft für Mathematik und Datenverarbeitung (GMD). Wir sind mit einer eigenen Basistechnologie, unserer Software K-Infinity, seit 1997 am Markt und bearbeiten eigentlich alle Themen rund um semantische Technologien. Wir haben 25 Mitarbeiter, sitzen in Darmstadt, und verdienen unser Geld mit Industrieprojekten, quer durch alle Branchen und Unternehmensgrößen. Aufgrund unserer Herkunft sind wir noch relativ stark in Forschungsprojekten involviert, z.B. Theseus.

Bild 2

Forschungsprojekte sind für uns immer eine Herausforderung, aber für die akademischen Partner noch sehr viel mehr als für uns, weil wir einen ganz guten Überblick darüber haben, was tatsächlich in der Industrie relevant und umsetzbar ist (Bild 2). Die Hochschulen mögen es dann doch nicht so, wenn wir sagen, dass 2/3 von dem, was sie erzählen, in der Industrie keinen interessiert und auch keinerlei Relevanz hat.

Das andere, was denen nicht so wirklich gefällt, hat man auf den Folien von Herrn Studer gesehen. Die ganzen Standards rund um Semantik sind alle noch relativ jung, vielleicht teilweise drei, vier Jahre alt. Jetzt gibt es uns seit 1997, d.h. wir haben uns bei vielen Sachen immer eigene Sachen gebaut und halten uns nicht an Standards. Wenn wir mit unseren eigenen Ideen ankommen, wird das nicht unbedingt so gern gesehen, weil man sich an das anpassen muss, was wir uns schon vor ein paar Jahren überlegt haben.

Ausgehend von dem Theseus Projekt, wo Semantik eher in Reinkultur mit Fokus auf den Diensten, aber auch in großen Umgebungen behandelt wurde, haben wir uns gefragt, ob das richtig relevant ist und ich damit zu einem Kunden gehen und zum Kauf empfehlen kann. Vielleicht müssen wir da doch einen oder zwei Schritte weiter zurückgehen und das ganze simpler machen und die Beschreibungsmodelle, die wir für unsere Entitäten in der realen Welt haben, vereinfachen und die Verfahren und Algorithmen, die darauf aufbauen, vereinfachen. Damit hat uns der Link Open Data Ansatz und die Standards, die da verwendet werden, interessiert. Man hat bei Herrn Studer gesehen, dass das nur einen kleinen Teil dieses Semantic Web Stacks abdeckt, der aber unserer Ansicht nach völlig ausreicht.

Bild 3

Daher haben wir zusammen mit Prof. Hepp von der Bundeswehruniversität in München das BMBF-Projekt Intelligent Match, aufgesetzt, das bis zum 31.12.2012 läuft (Bild 3). Prof. Hepp ist ein starker Treiber der Themen rund um die kommerzielle Beschreibung von Produkten und Dienstleistungen auf Basis von Linked Open Data. Er ist Autor von Goodrelation, einer reduzierten kleinen Ontologie, mit der man Merkmale und kommerzielle Eigenschaften von Produkten beschreiben kann. Sehr viel davon wurde von schema.org aufgegriffen, einer Vereinigung der großen Suchmaschinenanbieter, die dadurch in der Lage sind, diese Beschreibungen, die in Webseiten integriert sind, zu verstehen und auch auswerten zu können. Das führt dazu, wenn man in den großen Suchmaschinen einen Suchbegriff eingibt, man nicht nur den Link zu Amazon sieht, sondern gleich ein Bild, einen Preis und vielleicht auch Bewertungen, weil ich einfach in meinen Webseiten diese Information schematisch hinein kodiere.

Die Idee war, einen kommerziellen Partner mit in das Projekt aufzunehmen, der den Nutzen und die Notwendigkeit sieht, mit solchen strukturierten Informationen umzugehen. Der eine ist die Messe Frankfurt, seit zwei Jahren erzählt mir der IT Chef der Messe Frankfurt nun, dass strukturierte Daten das Öl des neuen Jahrtausends sind.

Der andere Partner ist die Otto Gruppe, deren Motivation darin besteht, sich gegenüber den Mitbewerbern abzugrenzen. Da sie das nie über den Preis kann und auch nicht will, sucht sie nach Möglichkeiten, um qualitativ bessere Angebote zu machen. Um das zu können, muss ich mir die Mühe machen, meine Dinge auch qualitativ besser und genauer zu beschreiben. Diese Mühe machen wir uns. Man sieht es auf einer anderen Folie beim konkreten Anwendungsfall.

Zuvor habe ich aber noch ein andres Beispiel. Heute Morgen kam der Spruch, dass die Kosten zur Gewinnung qualitativer Daten ständig sinken. Wenn dazu jemand bei einem unserer Kunden fragt, so hat der dazu bestimmt eine völlig andere Meinung. Entsprechend des Fachgebiets oder der Domäne sinken die Kosten für die Gewinnung der Daten nicht im

geringsten. Das war vor 20 Jahren ein richtiges Geschäft und ist es heute noch. Ich muss manuell und redaktionell arbeiten. Die Qualität, die mir eine automatische Analyse bringt, ist für solche Ansprüche, wie es die Messe Frankfurt hat, einfach nicht ausreichend, und ich muss manuell nacharbeiten

Bild 4

Bild 5


Semantic Recommendation for Online-Shops
Eine schwierige Aufgabe …
Vielzahl von Standards zur Beschreibung von Gütern und ihre Komplexität die erschwert eine automatisierte Produktrecherche und -beschaffung spezieller Güter über das WWW.
Wikipedia: „… eine Klammer zum Verschließen eines Briefumschlags …“
© intelligent views gmbh

Bild 6

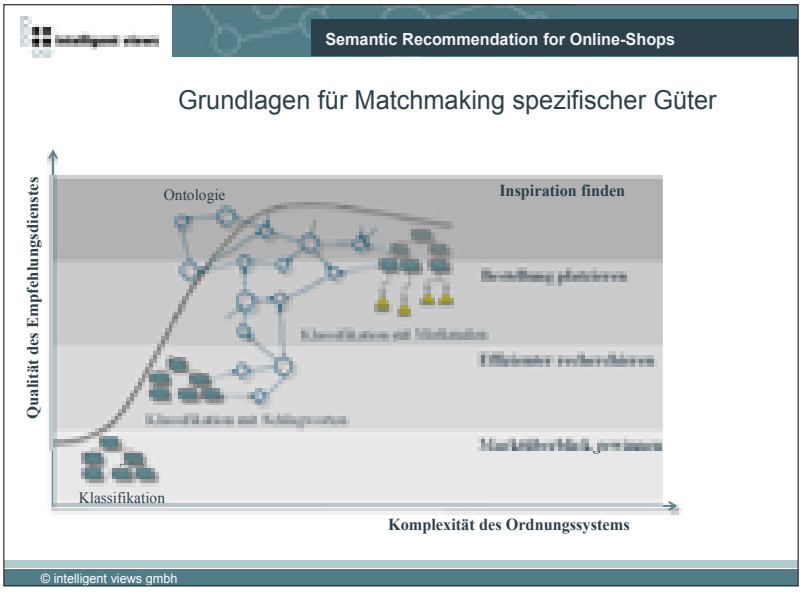
Semantic Recommendation for Online-Shops
Grundlagen für Matchmaking spezifischer Güter
Qualität des Empfehlungsdienstes
Ontologie
Inspiration finden
Klassifikation
Komplexität des Ordnungssystems
© intelligent views gmbh

Bild 7

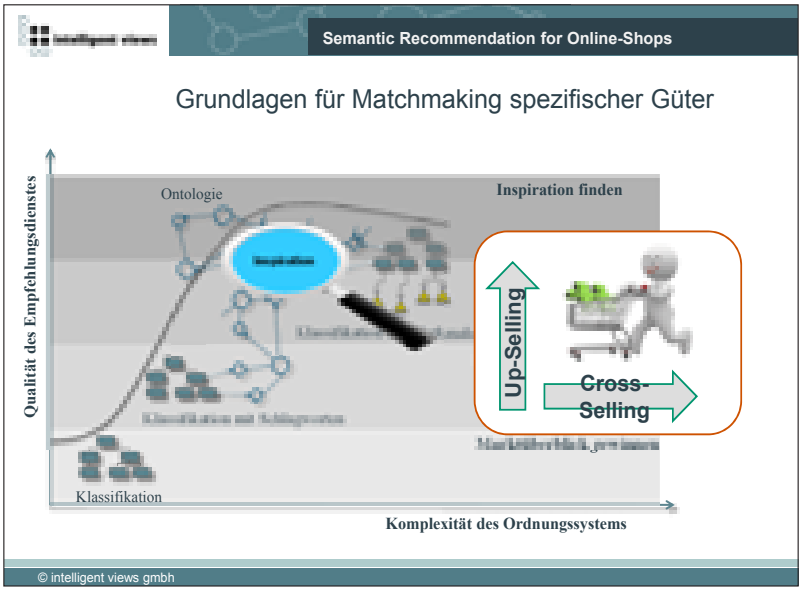

Bild 8

Die Idee war, sich einen Beschaffungsprozess anzusehen, wie dieser in der Regel abläuft, dass ich mir einen Marktüberblick verschaffe, eine echte Recherche mache, zu einem Angebotsvergleich komme und irgendwann einmal so etwas in den Betrieb überführe (Bilder 4 bis 8). Der Ansatz ist in dem Moment, wo die Inspiration stattfindet – Inspiration heißt, dass ich merke, dass ich etwas brauche, von dem ich noch nicht wusste, dass ich nicht darauf verzichten kann. Wie komme ich an die Stelle dran? Was habe ich für Möglichkeiten, wenn ich die Produktbeschreibung habe, so eine Inspiration zu schaffen, solche Empfehlungsdienste, wie ich das von Amazon her kenne, mit den semantischen Beschreibungen, die ich über meine Produkte habe, zu realisieren.

Für die Verkaufsseite ist es das Problem, die Produkte, die Beschreibungen, die vorhanden sind, so zu platzieren, dass wirklich Zwischenhändler oder Marktplätze etwas damit anfangen können. Auf der Beschaffungsseite habe ich immer größere Probleme aus meinem stetig wachsenden Angebot über die unterschiedlichsten Quellen auch das zu finden, was ich eigentlich brauche.

Das beginnt schon damit, dass die Idee, die Eigenschaften von Produkten oder Dienstleistungen anhand von standardisierten Klassifikationssystemen zu beschreiben, nicht neu ist. Sie ist leider schon so alt, dass es eine Unmenge an Standards gibt. Ein Ausgleich zwischen solchen verschiedenen Standards ist schlicht nicht möglich und auch nicht machbar. Ich als Betreiber bekomme am Ende von meinen Lieferanten eine Produktbeschreibung natürlich in einem anderen Standard zur Verfügung gestellt. Wenn es wenigstens eine standardisierte Beschreibung wäre, wäre das schön, denn 95% aller Informationen, die ich von meinen Lieferanten bekomme, sind einfach Freitexte. Ich kann selbst versuchen, meine strukturierten Eigenschaften herauszuholen. Selbst wenn so etwas verwendet wird, sind die auch noch inkompatibel oder einfach nicht aufeinander abgebildet.

Mit solchen Klassifikationssystemen kann ich maximal einen Marktüberblick gewinnen. Wenn ich versuche zu verallgemeinern, indem ich Schlagworte und Synonyme mit dazu nehme, kann ich die Qualität der Recherche erhöhen. Richtig interessant wird es, sobald ich meine Klassifikationssysteme auf solche merkmalsbasierten Klassifikationen umstelle, weil ich dann wirklich strukturiert fragen kann und auch wirklich sagen kann: Ja, ich habe ein Foto und wunderbar, dass Du mir Blitze vorschlägst, aber hast Du auch welche, von denen ich genau weiß, dass sie passen? Weil ich strukturiert abfragen kann, dass die physikalischen Eigenschaften dieser beiden Dinge aufeinander passen und ich sie dafür verwenden kann. Das letzte ist, wenn ich so etwas richtig ausbaue, indem ich formale oder Domain Ontologien mit dazu nehme, bin ich wirklich in der Lage und habe genug Dinge, um wirklich auch Empfehlungen aussprechen zu können für bestimmte Benutzer, über die ich Informationen habe oder die mir zumindest ein bisschen erklärt haben, wo sie sich bewegen, z.B. über ihre Kaufhistorie.

In Intelligent Match ging es darum, was für einen Aufwand ich habe, um meine Produkte, meine Daten tatsächlich auf so einen Level zu bringen und welche Arten von Empfehlungen ich zu welchem Zweck damit wirklich realisieren kann. Der erste Punkt ist die Inspiration, dass ich ein Up-Selling oder Cross-Selling damit machen kann.

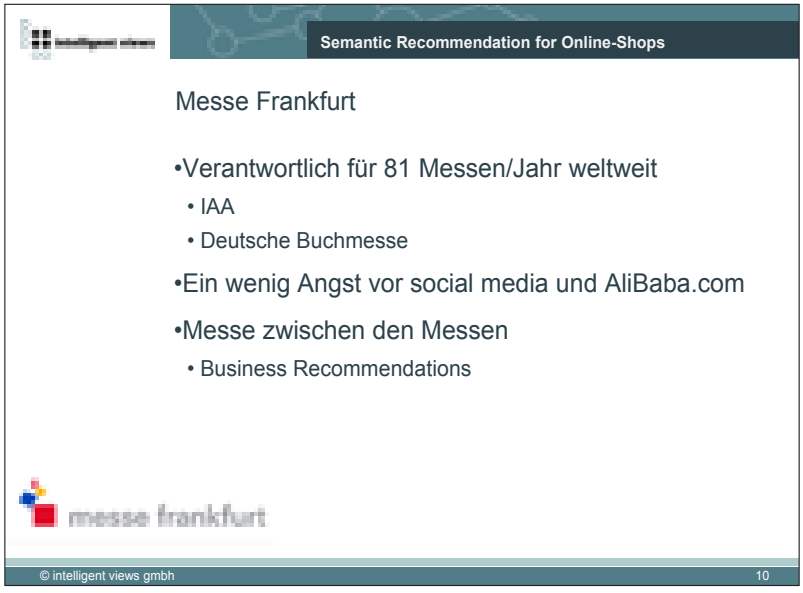

Bild 9

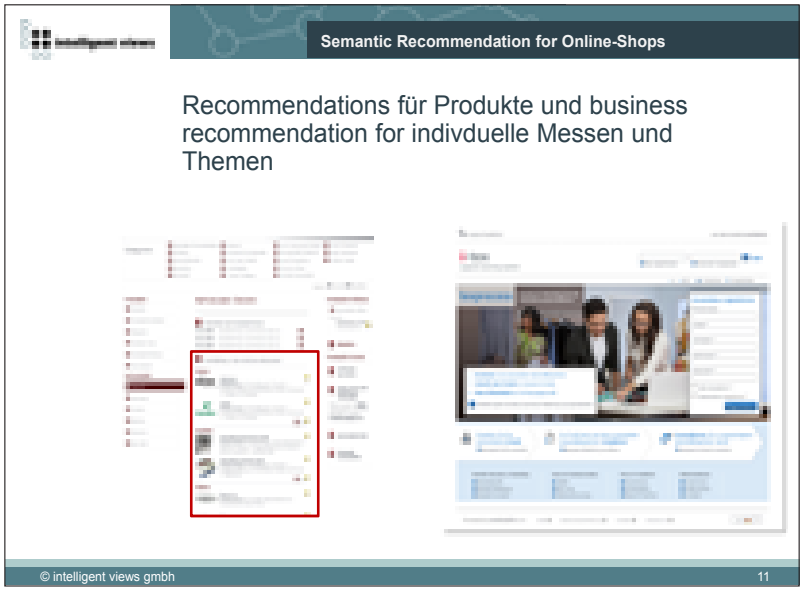

Bild 10

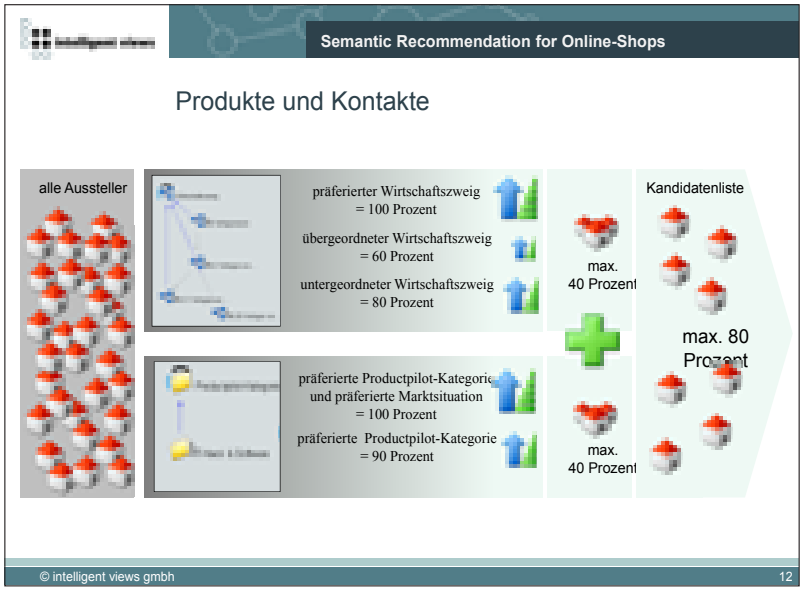

Bild 11

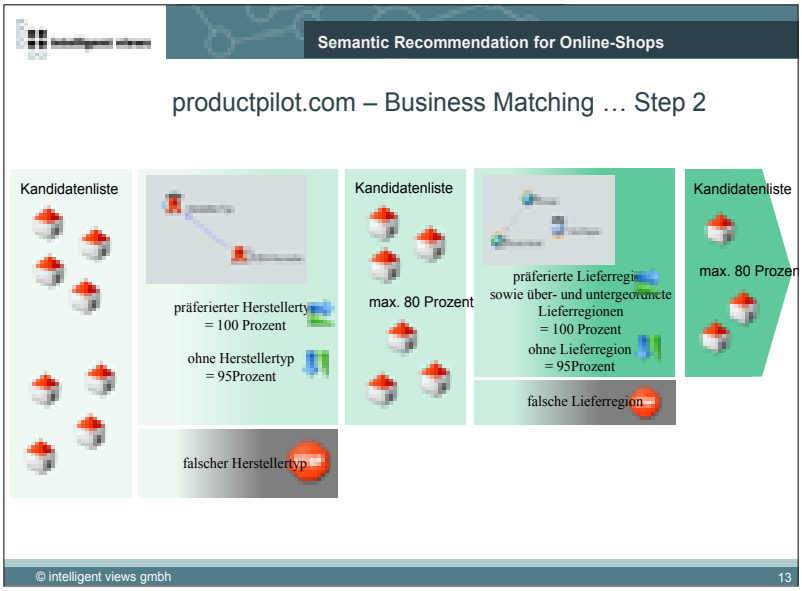

Bild 12

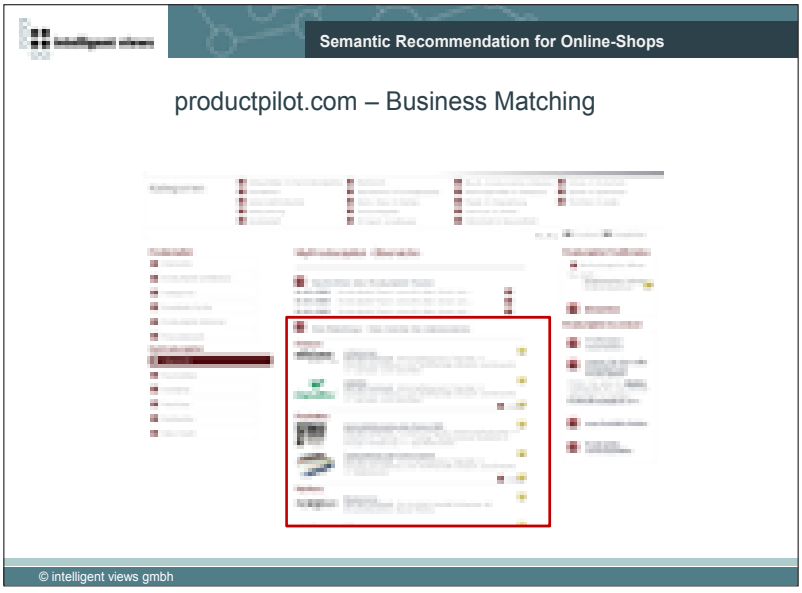

Bild 13

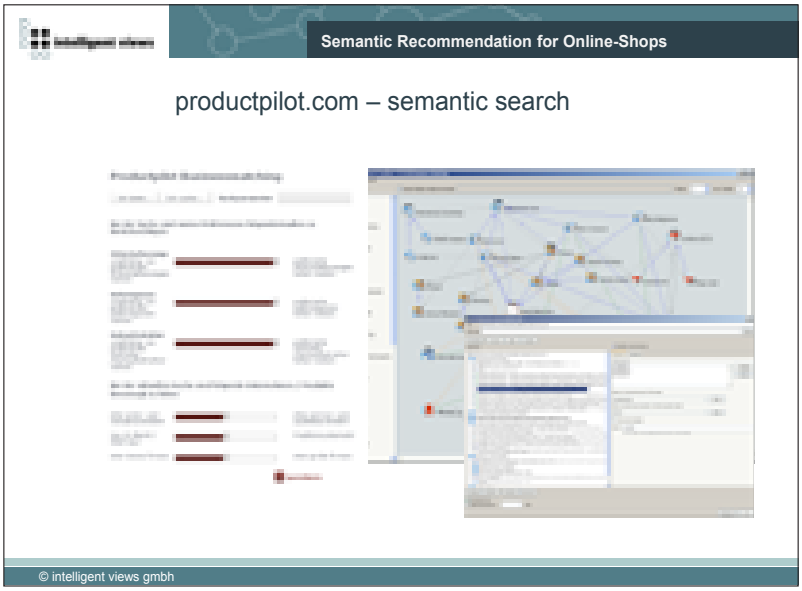

Bild 14

Der erste Kunde war die Messe Frankfurt, verantwortlich für 81 Messen im Jahr weltweit (Bilder 9 bis 14). Was die Messe in Frankfurt ein bisschen umtreibt, ist die Angst vor Social Media und vor allem das große Zittern vor AliBaba.com. Es gehört im Moment zu großen Teilen Yahoo, die versuchen, ihre Anteile wieder loszuwerden. AliBaba ist eigentlich eine Beschaffungsplattform für den kompletten asiatischen Markt, weniger Beschaffung als dass Sie da direkt Produkte kaufen können, sondern dass sich da der komplette asiatische Markt, die Produzenten dort mit Produktbeispielen und Informationen präsentieren. Der Ansatz ist, dass man nicht mehr zwei Jahre lang auf alle Messen im asiatischen Raum reisen muss um einen möglichen Partner in China zu finden, sondern sich durch eine Recherche über das Portal eine ausgesuchte Liste erstellt, die dann direkt angesprochen werden können. Die Messe Frankfurt hat daraufhin beschlossen, selbst so etwas zu machen. Sie setzen Portale auf, um zum einen unseren Ausstellern noch besser die Möglichkeit zur Präsentation zu geben, aber auch um so etwas wie Messe zwischen Messen darzustellen. Fachspezifisch bezogene Kommunikationsplattformen, die jeweils immer nur für eine bestimmte Messe aktiv sind und bei denen man weiß, dass die Leute dort unterwegs sind, die man normalerweise auf der Messe trifft. Darauf aufbauend ist ein Vorschlagswesen zu realisieren, so dass ich als möglicher Anbieter oder Aussteller nicht lange recherchieren muss, an wen ich mich eigentlich wenden könnte, sondern ich auf diese Plattform gehe, ein paar Informationen eingebe und mir dann eine Liste von Personen vorgeschlagen wird, die ich auf der nächsten Messe am besten einmal treffen sollte, welche Stände ich ablaufen sollte oder wen ich am besten direkt kontaktiere.

Das Modell besteht aus den Produktbeschreibungen der Hersteller, die mit Hilfe von Produktklassifikations Standards wie eCl@ss oder UNSPSC erstellt sind, ergänzt um persönliche Informationen und Informationen über Firmen, die über eine NACE Wirtschaftszweigklassifikation zusammenhängen. Auf diesem semantischen Modell wurden Recommendation Verfahren entwickelt, mit dem Hintergrund einer einfachen Generierung themenspezifischer Portale, die auf Knopfdruck eine Plattform erstellt, die dann automatisch auch noch die passenden Recommendations generiert.

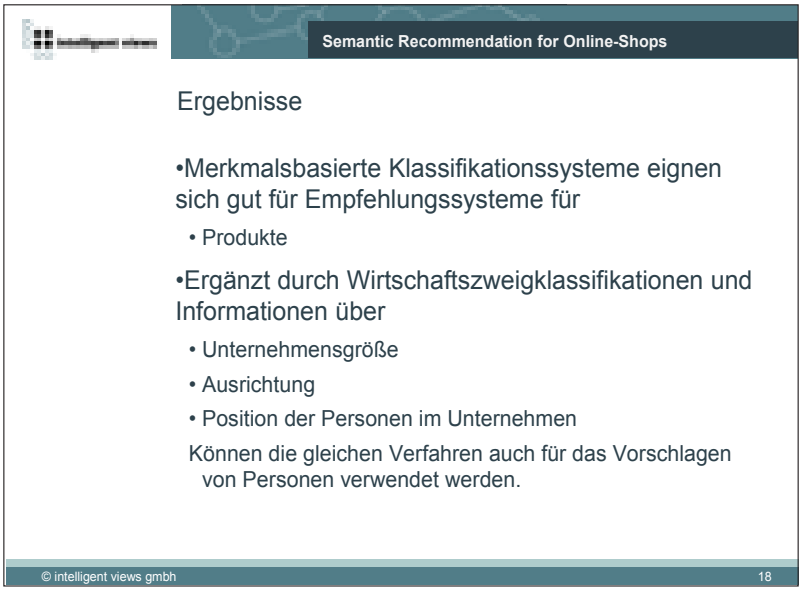

Bild 15

Die Ergebnisse zeigen (Bild 15), dass sich reine merkmalsbasierte Klassifikationssysteme schon sehr gut als Grundlage für Empfehlungssysteme für Produkte eignen. Die Ergebnisse solcher Vorschläge sind den Vorschlägen, die bspw. Amazon liefert, mindestens gleichwertig. Ergänzt durch Wirtschaftszweigklassifikation und weitere Informationen über Unternehmen, Ausrichtung und Position können die Verfahren damit auch für das Vorschlagen von Personen verwendet werden.

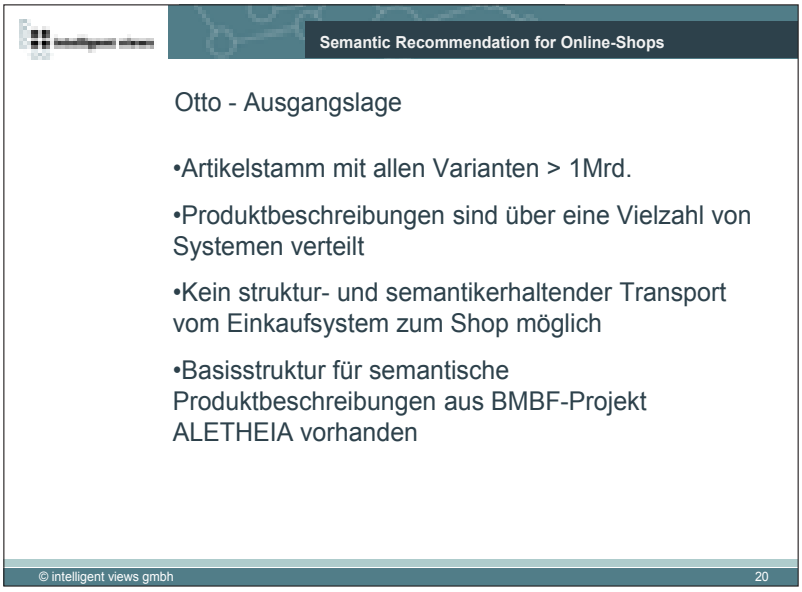

Bild 16

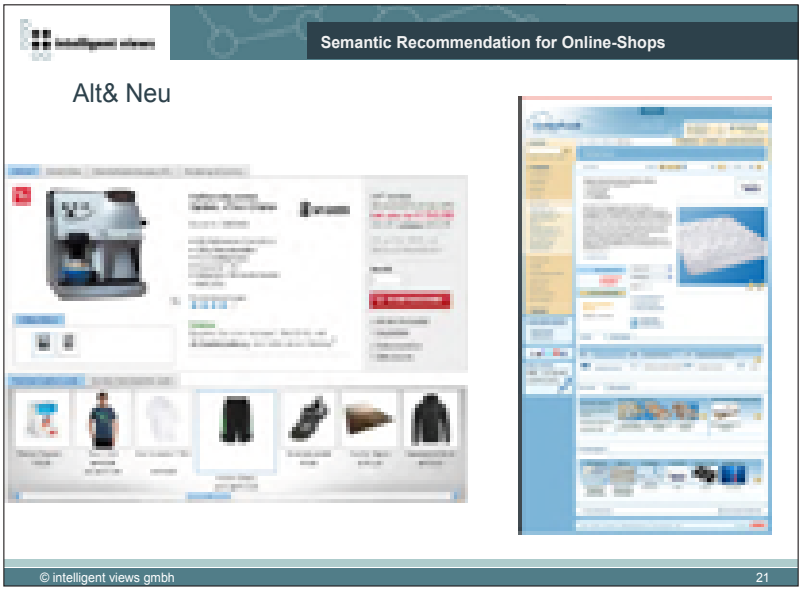

Bild 17

Das andere Beispiel ist die Otto Gruppe (Bilder 16 und 17), Universal-Versender seit 1949, zweitgrößter Online-Händler weltweit nach Amazon. Auch hier steht man vor dem Problem, dass nahezu alle strukturierten Daten, die in den Einkaufsystemen vorliegen, auf dem Weg zum Online-Shop schrittweise verlorengehen, bis zum Schluss nur noch unstrukturierte Texte vorliegen mit denen keine hochwertigen Empfehlungen mehr möglich sind. Otto hatte allerdings eine Basisstruktur für semantische Produktbeschreibung. Es kommt aus dem BMBF Projekt ALETHEIA.

Bild 18

Bild 19

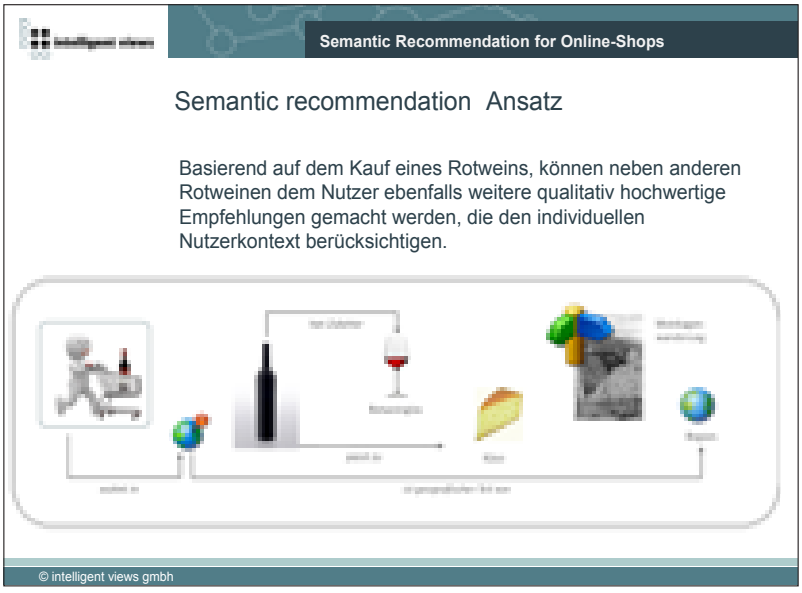

Bild 20

Was sich auch bei Otto gezeigt hat, ist, dass die Entwicklung generischer Recommendation Algorithmen sehr stark vom Kontext der Produkte abhängt (Bilder 18 bis 20). Ein ganz simples Beispiel: Sie haben einen Onlineshop für Bohrmaschinen. Wenn Sie wissen, dass ein Mann letzte Woche eine Bohrmaschine gekauft hat, schlagen Sie dem keine mehr vor, sondern der bekommt jetzt Bohrer vorgeschlagen.

Wenn Sie allerdings einen Onlineshop für einen Jagdausstatter haben und Sie wissen, dass der letzte Woche ein Messer gekauft hat, dann schlagen Sie dem sehr wohl wieder ein Messer vor, denn es könnte ein Messersammler sein, und die gibt es viel häufiger als Bohrmaschinensammler.

Bild 21

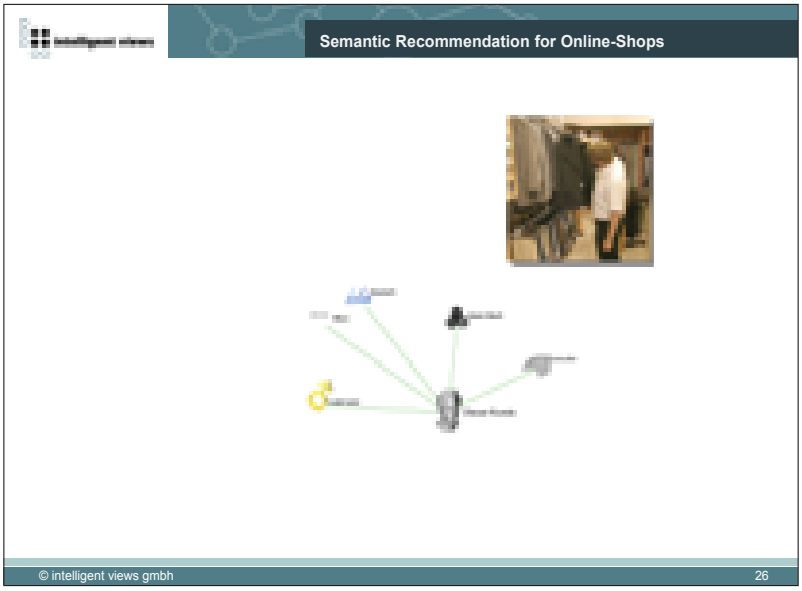

Bild 22

Bild 23

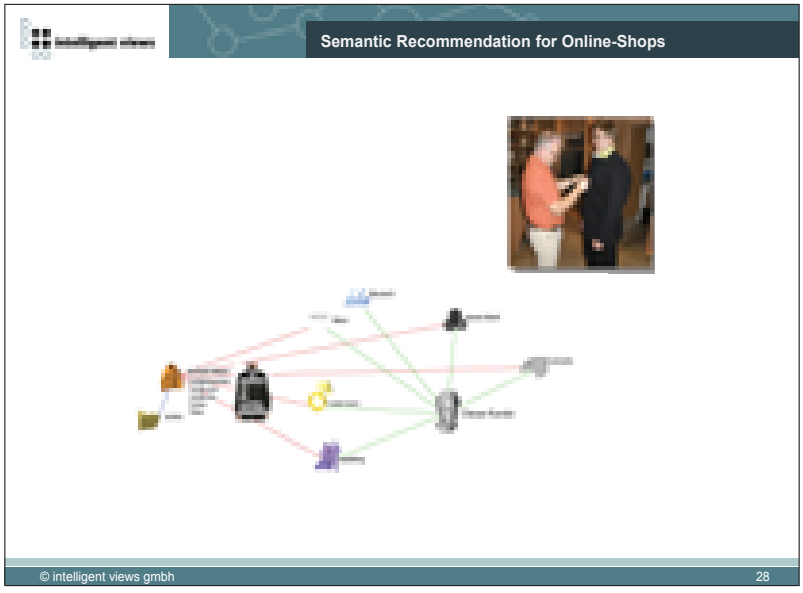

Bild 24

Bild 25

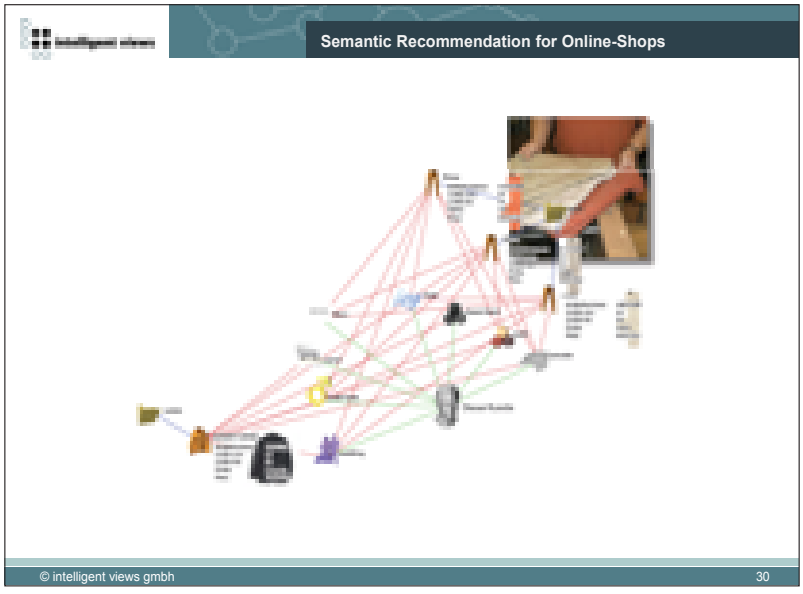

Bild 26

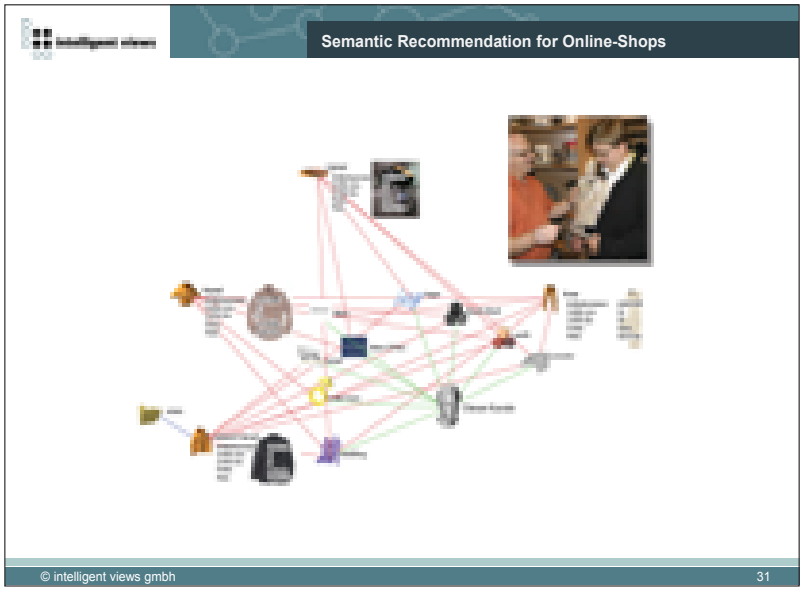

Bild 27

Bild 28

Bild 29

Bei Otto ist man für den Bereich Mode dazu übergegangen bei Verkäufern in Fachgeschäften über die Schulter zu gucken und zu versuchen nachzuvollziehen wie eigentlich ein echter Verkäufer mit so einer Situation umgeht, dass der Kunde herein kommt und einen Kauf tätigen möchte (Bilder 21 bis 29). Was schaut er sich alles an? Was für Kleidung trägt er aktuell?

Der Verkäufer hat zu Beginn eine beobachtende Rolle und versucht nur Informationen zu sammeln. Nach einer gewissen Zeit spricht er den Kunden an und fragt nach dem Anlass. Im Beispiel geht es um eine Hochzeit und der Kunde hätte gern etwas Klassisches. Dann kommt der erste Vorschlag zum Jackett und es wird noch nach Alternativen gesucht, die aber schon möglichst nah an der ersten Wahl des Kunden liegen. So versucht der Verkäufer sich dem Ganzen anzunähern. Einem guten Verkäufer geht es natürlich nicht nur um das Jackett, sondern er schlägt noch passende Hosen vor, die sowohl zum Anlass passen als auch zu dem bereits Angeschauten. Und zu einer Hose gehören noch ein Gürtel, Hemd usw., denn man heiratet schließlich nur einmal oder wird nur selten eingeladen.

Diesen Prozess hat man ein paar Mal durchprobiert, sich dann mit den Verkäufern unterhalten und daraus eine Mode-Ontologie für Otto erstellt, d.h. nicht nur direkte Eigenschaften sondern auch abgeleitete Eigenschaften, Merkmale und insbesondere Zielgruppen und Anwendungen.

Diese ganzen strukturierten Eigenschaften sind gut und schön, aber der eigentliche Witz kommt über Zielgruppe und vor allem Anwendungen, warum man das kauft.

Bild 30

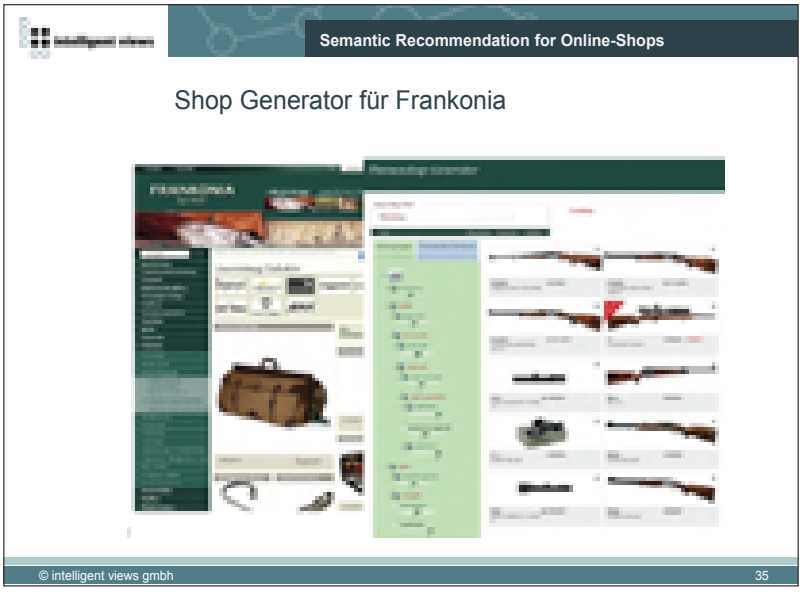

Bild 31

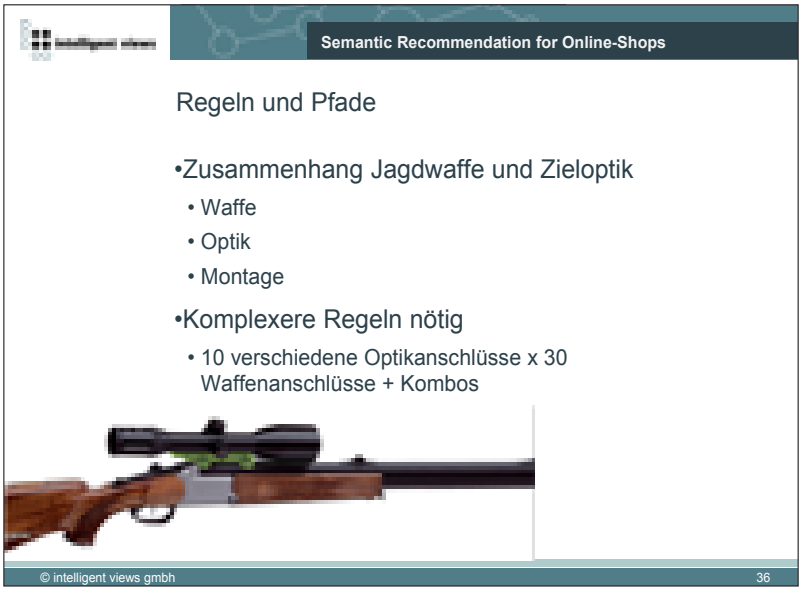

Bild 32

Bild 33

Bild 34

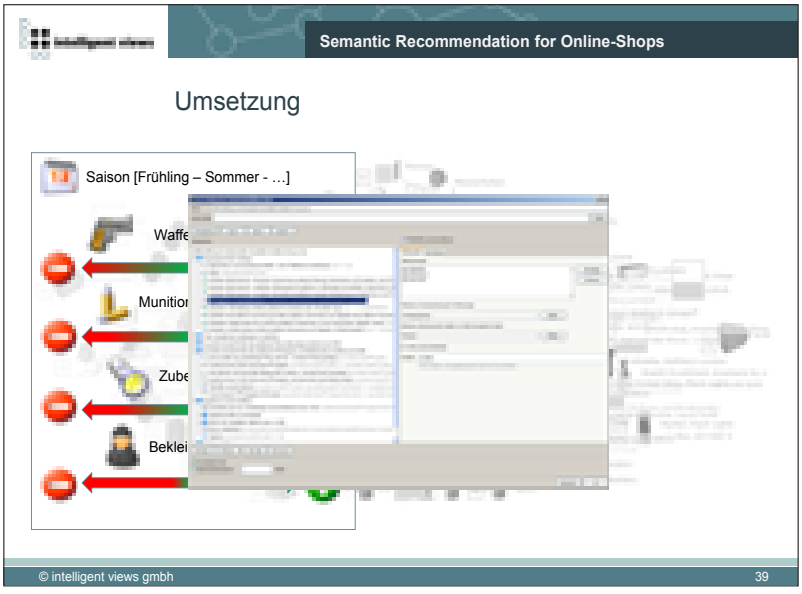

Bild 35

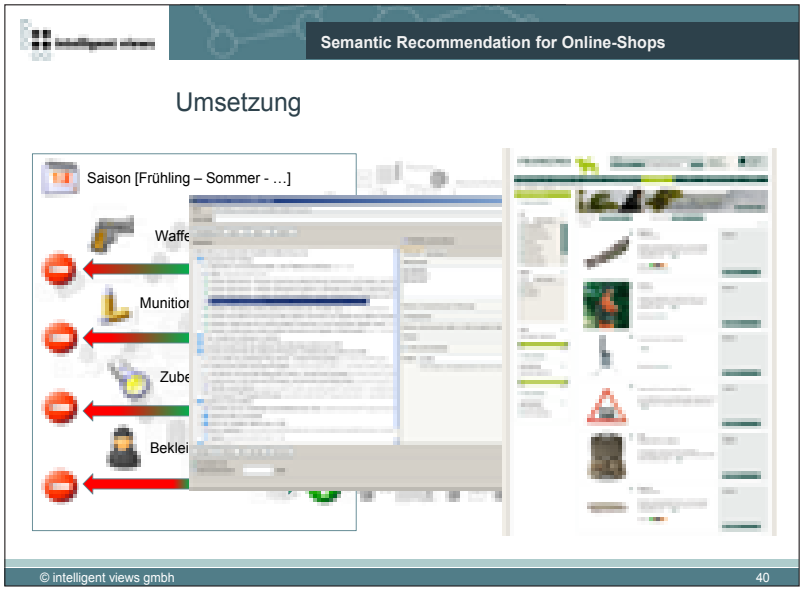

Bild 36

Konkret im Einsatz ist das System beim Online-Shop einer Tochter der Otto Group, dem Jagdausstatter Frankonia (Bilder 30 bis 36). Er wurde deswegen ausgesucht, weil die Jägercommunity ein extrem anspruchsvoller Kundenkreis ist. Schlechte Vorschläge, die beispielsweise keine aktuellen Schonzeiten für Wild berücksichtigen, werden von diesen Kunden abgelehnt und führen im schlimmsten Fall dazu, dass die Kompetenz von Frankonia in den Augen ihrer Kunden leidet. Weil Frankonia hauptsächlich hochpreisige Waren führt, wird erwartet, dass auch bei Vorschlägen im Online-Shop alle für die Produkte und den jeweiligen Einsatzzweck relevanten Merkmale und Eigenschaften bei der Empfehlung berücksichtigt werden. Dies führte bei Frankonia zu einer vollständigen und formalen Beschreibung der Warenwelt. Damit ist man mittlerweile in der Lage, einen Shop Generator zu bauen, der ausgehend von mehreren Themen, selbstständig Warensortimente für Themenshops erzeugen kann.

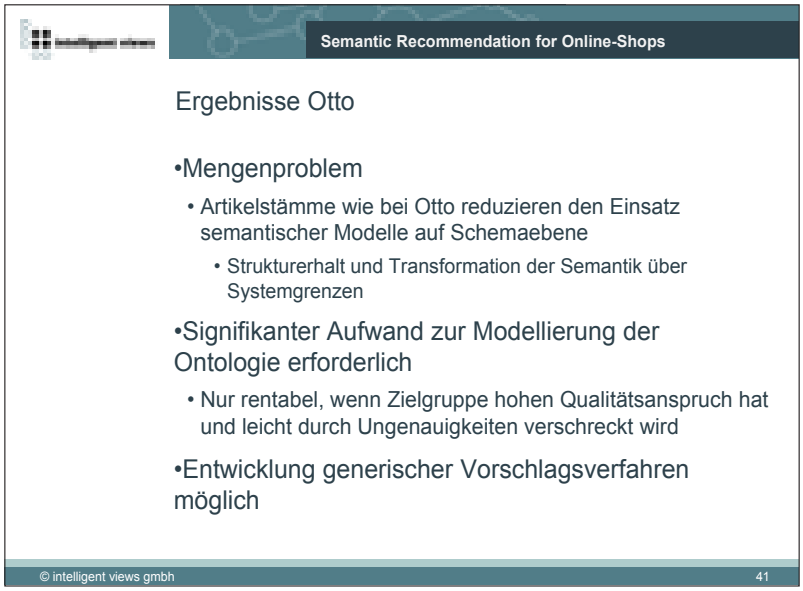

Bild 37

Ich komme noch zu den Ergebnissen zu Otto und damit zum Zusammenhang zu Big Data (Bild 37). Handelshäuser wie Otto haben ein echtes Mengenproblem. Millionen Artikel in unterschiedlichen Varianten reduzieren den Einsatz von semantischen Modellen auf die Schemaebene und erzwingen eine Vorberechnung der Vorschläge. Es gibt einen signifikanten Aufwand zur Modellierung der Ontologie, der dazu noch stark von den Qualitätsansprüchen abhängt. Dies macht solche Ansätze auch zu einer vertrieblichen Herausforderung – da man gegen statistische Ansätze argumentieren muss. Diese versprechen eine Black-Box, die keinerlei manuellen Eingriff erfordert, sondern nach einer gewissen Trainings- und Beobachtungszeit autonom in der Lage ist, Produktvorschläge zu liefern. Wenn deren Qualität akzeptabel ist, lohnt der Aufwand für die Erstellung eines semantischen Modells nicht. Wenn aber die Zielgruppe hohe Anforderungen an die Qualität der Vorschläge stellt, und schlechte oder falsche Vorschläge die potentiellen Kunden vertreiben, ist die Verwendung eines semantischen Modells ohne Alternative.

12 DISKUSSION
Veränderung der Industrielandschaften

Leitung: Dr. Wolf v. Reden, Fraunhofer Institut für Nachrichtentechnik HHI, Berlin

Dr. v. Reden:
Wir haben in dieser Session versucht, Ihnen Glanz und Elend der Semantik zu zeigen. Wir haben dabei den Weg zu den Big und Media Data zurückgefunden. Wir haben sogar rechtliche Aspekte noch einmal erwähnt. Jetzt sind Sie dran. Jetzt wollen wir die Diskussion eröffnen. Wenn Sie Fragen, Vorschläge oder Thesen haben , die zu dieser Session passen, bitte jetzt!

Dr. Helbig:
Ich hätte eine kurze Frage an Herrn Steinacker in dem Fall Otto. Werden die Merkmale aus den Produkten in irgendeiner Form generiert oder muss jemand hingehen, der Produktmanager, und jedes seiner Produkte in diese Ontologie bewertend eingeben?

Dr. Steinacker:
Soweit möglich werden die natürlich automatisch generiert. Eine Struktur gab es schon immer ein bisschen, auch bevor man das in semantischen Modellen gemacht hat. Da werden natürlich Abbildungen gemacht. Es gibt auch auf Schemaebene Abbildungen, dass der eine ganz billig Rot sagt und der andere den CMYK Wert drin hat. Dass es da eine Clearingstelle gibt, wo ich dann anfragen kann, was meine Produktbeschreibung in dem System wäre. Dass so eine automatische Transformation stattfindet. Es ist also weniger die Produktbeschreibung. Der größte Aufwand ist immer die Erstellung der Ontologie des semantischen Modells an sich, des Domainwissens, und je besser man vorher strukturiert hat desto weniger Aufwand habe ich. Wenn ich nur einen Katalogtext bekomme, wo das Ding im Fließtext drin steht, kann ich da maximal Vorschläge machen. Aber die muss ich redaktionell prüfen.

Dr. Helbig:
Setzt Otto das jetzt in größerem Stil ein oder hat das explorativen Charakter gehabt?

Dr. Steinacker:
Das ist tatsächlich bei frankonia.de, bei schlafwelt.de in Betrieb. Der eine oder andere ist in Vorbereitung, das wirklich Otto weit auf otto.de auszurollen.

Prof. Eberspächer:
Ich habe eine Frage an Herrn Kuhlmann, die an diese Frage anschließt. Wie viel Pflege ist bei dem System eigentlich nötig? Wenn sich jetzt die Beziehungen, die da abgebildet sind, ändern? Wie lang dauert das, bis das aktualisiert ist?

Herr Kuhlmann:
Wir haben verschiedene Systeme, um die Daten zu aktualisieren. Es gibt zum einen die Möglichkeit, die Informationen aus den integrierten Linked Open Data-Quellen zu extrahieren. Wenn dort neue Informationen reinkommen, tauchen diese Minuten später automatisch bei uns auf. Es gibt auch die Möglichkeit, händisch über entsprechende Tools einzugreifen. Auf der Plattform alexandria.neofonie.de kann man die Daten ändern, kann neue Daten einfügen, kann Datensätze löschen. Die dritte Möglichkeit ist, dass wir Tausend deutschsprachige Nachrichtenquellen durch unsere Textanalyse analysieren, was ich vorher kurz präsentiert habe, und versuchen, Informationen rauszuziehen, auch Beziehungsinforma-

tionen usw. Aber da ist die Qualitätsbarriere noch zu hoch, als dass wir die Daten so ungesehen in die Wissensbasis aufnehmen wollen würden.

Prof. Eberspächer:
Aber jetzt Ihr Beispiel erweitert: Wenn jemand sagt: mich interessiert vor allem, welche Autos die Personen fahren oder wo sie Urlaub machen. Also ganz klare Aussagen. Wie kriegt man das rein? Muss man dann zuerst das Schema ändern? Machen das dann Sie?

Herr Kuhlmann:
Ja, ganz genau. Alles, was sich konzeptionell ändert, an neue Arten von Beziehungen, neue Arten von Entitäten, ich will heute Produkte abbilden, o.ä., da muss manuell gearbeitet und das Schema erweitert werden. Wenn dies geschehen ist, und wenn Datenquellen für diese Arten von Daten vorliegen, dann kann ich mit Hilfe von entsprechenden Werkzeugen Mappings herstellen und die Daten dann wieder automatisch befüllen.

Prof. Eberspächer:
Aber im Prinzip kann man sich schon vorstellen, dass das auch mal ein von jedermann nutzbares Werkzeug wird, wo ich sage, welche Beziehungen ich eigentlich gern dargestellt haben möchte, und dann klick ich mir das zusammen?

Herr Kuhlmann:
Nee, leider nicht.

Prof. Eberspächer:
Schade. Ja, nicht heute, aber später mal?

Herr Kuhlmann:
Man könnte es machen, und technisch ist es kein Problem. Das Problem liegt eher darin, dass man so etwas wie eine Ontologie Modellierung nicht Laien überlassen sollte, die entweder von Modellierung keine Ahnung haben oder von der Domäne keine Ahnung haben. Selten gibt es Personen, die auf beiden Gebieten Experten sind. Insofern ist der Ontologie Modellierungsprozess immer etwas, wo mehrere Personen an einem Tisch sitzen, der langwierig ist, wo viel diskutiert wird, wo man sich die Köpfe einschlägt, wo Philosophie hinter steckt. Das will man nicht offen für jedermann zugänglich machen. Was passiert ist – jetzt nicht nur negativ gesehen, aber wenn man sich das Kategoriensystem bei Wikipedia einmal anschaut, was so über die Jahre gewachsen ist, da ist keine Einheitlichkeit drin. Das reicht nicht, um Analysen darüber zu fahren. Es ist natürlich gewachsen. Das ist auf jeden Fall ein Problem.

Herr Totzke, Siemens Enterprise Communication:
Ich habe eine Frage an Herrn Prof. Studer und zwar zum W3C Semantic Stack, den Sie vorgestellt haben. Inwieweit gibt es da im Markt eine Durchdringung, und gibt es auch eine Koordination der Standardisierung? Ich denke beispielsweise daran, was MPEG z.B. hinsichtlich Multimedia-Tagging heute macht, so dass diese Systeme später auch besser beispielsweise über diesen Stack integrierbar sind.

Prof. Studer:
Was diese Standardisierung angeht, das hatten Sie aus dem Vortrag von Herrn Abecker entnehmen können, da gibt es konkurrierende Standardisierungsgremien. W3C ist u.a. auch bekannt geworden durch die Basisstandardisierung, was die Web-Infrastruktur angeht und auf dem baut dieser Semantik Web-Stack auf, den ich Ihnen kurz vorgestellt hatte.

Wenn Sie in die Praxis schauen, dann sehen Sie, dass gerade RDF dabei ist, im breiteren kommerziellen Umfeld Fuß zu fassen. Wenn Sie sich heutzutage beispielsweise Google anschauen, die Snippits, die Sie als Ergebnisse bekommen, dann basieren sie teilweise inzwischen auf diesen strukturierten RDF-Daten. Oder wenn Sie an große Retailer in den USA denken, tauchen RDF-strukturierte Daten auch als Basis auf, um Produktbeschreibungen zu machen mit der Erfahrung, dass sich auch Ihre Geschäftsergebnis dadurch verbessern. Da ist gerade aus meiner Sicht der Weg geschafft, dass diese Standards sich wirklich im realen kommerziellen Umfeld ausbreiten. Wenn Sie an die Web-Ontologiesprache OWL denken, dann finden Sie diese in spezifischen Branchen, was ich auch vorhin erwähnt hatte, zum Beispiel im Live Science Bereich, wo man es sowieso gewohnt ist, in solchen taxonomischen Strukturen zu denken. Wenn Sie an Medizin, Biologie und dergleichen denken, finden Sie heutzutage sehr große Ontologien, die auf diesen Standards basieren. In anderen Domänen überhaupt nicht. Man kann aber nicht sagen, dass die irgendwie flächendeckend sind, es hängt von den spezifischen Branchen ab, in denen man unterwegs ist.

Dr. Wohlmuth:
Ich habe eine Frage zu der Aussage, dass bekannte Suchmaschinenhersteller die genannten semantischen Methoden auch tatsächlich einsetzen. Wie groß sehen sie denn die Chancen, dass Internetnutzer die Möglichkeiten und die Features, die Metasprachen und Semantik bieten, in naher Zukunft einfach über die bekannten Portale nutzen können. Und dies, ohne auf spezielle Portale wechseln zu müssen, die dann nur sehr kleine Datenmengen enthalten und manuell aktualisiert werden müssen?

Prof. Studer:
Es gibt eine Initiative von ein paar großen Playern, was Schema.org angeht, wo man sich auf eine Basisstruktur geeinigt hat, was gewisse semantische Beschreibungen angeht, kein formal definierter Standard sondern wie am Markt üblich ein Standard, der durch ein paar Marktführer definiert wurde. Das ist ein Bereich, wo man sieht, dass auch über Unternehmensgrenzen hinweg eine erste Konsensbildung stattfindet, was noch relativ einfach gehaltene semantische Strukturen angeht, die dann verschiedene Player im Markt benutzen und dadurch Dinge kompatibel machen.

Dr. Steinacker:
Ich denke, ein anderes Beispiel ging erst letzte Woche durch die News Seiten, dass Google mit seinem Knowledgegraph im Prinzip etwas Ähnliches vorhat, was jetzt Wikipedia macht, nämlich strukturiert und dass da auch signifikant Geld rein fließt in eine manuelle und editierte Version, in einen manuellen Aufbau. Das zeigt, dass das ganz klar seinen Weg findet und zeigt natürlich auch, dass man in der Lage wäre, so etwas alles automatisch aus Inhalten zu generieren, wenn man denn genug Inhalte hätte. Wenn das jemand könnte, dann jemand wie Google und Bing, weil die auf den ganzen Inhalten sitzen. Dass die jetzt Geld in die Hand nehmen und diesen Ansatz von Wikipedia sponsern, zeigt natürlich auch, dass da nach wie vor manueller Aufwand erforderlich ist. Das zeigt aber natürlich auch, dass die sich etwas davon versprechen und dass sie es tun werden.

Prof. Studer:
Darf ich noch eine Ergänzung anführen? Zu dem gerade vor kurzem von Wikipedia, und zwar der deutschen Organisation, aufgesetztem Wiki Data-Projekt, das seit Anfang April am Laufen ist und gerade zum Ziel hat, die Faktenhaltung, die Sie in Wikipedia derzeit rudimentär vorfinden, auf einer semantischen Basis zu machen. Das ist gerade ein großes Projekt, was in Berlin durchgeführt wird und das Ziel hat, in ungefähr einem Jahr die

Faktenhaltung von Wikipedia, und zwar dann quer Beet, nicht nur für die deutsche Wikipedia, auf dieser semantischen Basis zu machen. Die Idee ist, dass ein Faktum einmal erfasst wird und dann in alle möglichen Wikipedias eingespielt werden kann und damit der Pflegeaufwand und Wartungsaufwand deutlich reduziert wird. Ich gehe davon aus, dass das Projekt erfolgreich sein wird. Das wäre ein ganz wesentlicher Schritt, um diese semantische Fundierung der Datenbereitstellung voranzutreiben.

Dr. v. Reden:
Ich hätte noch eine Frage. Es scheint, aus den jetzt gehörten Äußerungen, relativ einfach zu sein, aus strukturierten Daten semantische Erkenntnisse und damit auch einen Mehrwert zu generieren. Wie ist es bei unstrukturierten Daten? Wir haben einige bunte Beispiele. Neue Produkte werden in irgendwelchen Webforen von Kunden bewertet, oft mit einer Sprache, die grammatikalisch nicht sehr korrekt ist, bzw. zuweilen - auf gut Deutsch - ziemlich rotzig daherkommt. Wie kann eine Syntaxmaschine das sinnvoll interpretieren, zumal die meisten Syntaxmaschinen sowieso amerikanischen Ursprungs sind? Wie können die aus einer nur bedingt deutschen Sprache interpretierbare Entitäten erstellen und damit doch noch verwertbare Aussagen erzeugen?

Dr. Steinacker:
Ich sage einen Satz. Sie sagen bestimmt das Gegenteil. Meiner Ansicht nach geht es schlicht nicht.

Dr. v. Reden:
Trotzdem kenne ich solche Auswertungsgrafiken. Es gibt Firmen, die solche Tools anbieten und das mit überraschenden Ergebnissen. Man kann dabei die benutzten und ausgewerteten Stellen des analysierten Textes bis ins Einzelne verfolgen. Vielleicht wirkt das Verfahren erst durch das Gesetz der großen Zahl, das ab ca, 25 ausgewerteten Bewertungen gelten würde. Dabei werden dann ironische oder schlichtweg daneben liegende Bemerkungen durch eine Gaußverteilung ausgemittelt. Ich habe Syntaxmaschinen gesehen, die bei einer Smartphone Bewertung aus etwa 100 Foreneinträgen zu einem mir scheint vernünftigen Gesamtergebnis kommen. Wie funktioniert das?

Herr Kuhlmann:
Es ist immer die Frage des Anspruchs dahinter, aber auch eine Frage der Domäne, die ich betrachte. Die Präzision, die da erreicht wird, schwankt sehr stark. Je nachdem wie tief ich fragen möchte, wie sehr Subjektivität dabei eine Rolle spielt. Das sind alles einzelne Dinge, die ich berücksichtigen muss. Welche Quellen gucke ich mir an? Gucke ich mir nur Nachrichtenquellen an oder das gesamte Web? Guck ich mir Blogbeiträge an oder, noch schlimmer, gucke ich mir Twitterbeiträge an? Die Qualität schwankt. Man kann es nicht sagen. Man muss sich den einzelnen Use Case anschauen, ob mir die Qualität da ausreicht. Welche Informationen will ich haben und welchen Anspruch habe ich an die Qualität? Das muss ich abwägen, und für manche Use Cases funktioniert es heute schon gut. Im Medizinbereich z.B. oder bei uns die Nachrichtenanalyse funktioniert gut, wobei gut hier auslegbar ist. Zu 100% funktioniert es nicht und wird auch lange noch nicht so sein. Es wird noch Jahre, Jahrzehnte, vielleicht Jahrhunderte dauern, bis man soweit kommt. Aber die Verfahren werden besser und man überschreitet langsam bei der Präzision bestimmter Teilproblematiken die 90% Grenze.

Prof. Studer:
Ich würde auch noch gern eine Antwort ergänzen. Ich hatte auf einer Folie gesagt, dass man in manchen Bereichen derzeit schon relativ gut zuhause ist. Da würde ich meinem Kollegen

zustimmen. Was so etwas unkonventionellere Textklassen angeht, ist man sicher noch in den Anfängen der Textanalyse. Wir haben aber in der Forschung noch ein paar Jahre vor uns, wo wir neue Methoden und Verfahren entwickeln können. Von daher wird man inkrementell Fortschritte machen, was weitere Domänenabdeckung oder andere sprachliche Textklassen angeht. Aber Wunder sollte man nicht erwarten. Das geht irgendwie inkrementell voran und hängt wirklich von der Domäne ab, ob die Qualität dessen, was man an Analyse liefern kann, schon hinreichend ist, um das in der Praxis einsetzen zu können.

Dr. v. Reden:
Okay, das Leben ist eben inkrementell. Mit dieser sehr tröstlichen Aussage möchte ich Sie gerne in die Pause entlassen.

13 Von der Quizshow ins Geschäftsleben: IBM Watson Analytics im Gesundheitswesen

Thomas Hampp, IBM Deutschland Research & Development GmbH, Böblingen

Ein paar warnende Worte für Mediziner zu Beginn. Dieser Vortrag bespricht Anwendungen von IBM Watson im Gesundheitswesen. Der Vortragende ist aber kein Mediziner, sondern auf der KI-, Software- und Ingenieursseite zuhause. Eine weitere Entschuldigung auch an die Ingenieure: Der Schwerpunkt dieses Vortrags ist es nicht zu erklären, wie Watson technisch funktioniert. das wird nur sehr kurz angerissen. Das Thema in diesem Vortragsblock ist es darzustellen, wie Analysetechnologien in verschiedenen Domänen zur praktischen Anwendung kommen. Deswegen werde ich auf die Problemstellungen und vor allem deren Lösungen in der medizinischen Domäne eingehen, ohne tiefe medizinische oder technische Details zu geben.

Wir kommen nun zur medizinischen Problemstellung. Die kann man mit einer einfachen Frage beginnen. Sind Sie eigentlich zufrieden mit Ihrem Arzt? Nicht nur damit, dass er sich Zeit nimmt, aber auch mit seiner fachlichen Qualität? Ist er immer informiert und findet die richtigen Diagnosen? Wenn Ihr Arzt so gut ist und Sie wirklich zufrieden sind, dann ist sicher seine Praxis voll. Aber wann bildet sich ein solch populärer Arzt eigentlich fort? Das selbe kann man zu Ärzten in Kliniken fragen: Glauben Sie, dass Sie in der Klinik immer ideal versorgt werden, dass alle Entscheidungen, die für Sie wichtig sind, so getroffen werden, dass sie auf der Basis der neuesten Fachkenntnisse erfolgen? Und dass sie immer evidenz-basiert und wohlbegründet erfolgen?

Die meisten Leute würden wahrscheinlich ‚nein' sagen. Das ist auch kein Wunder und kein Problem der Ärzte. Wir haben in dieser Konferenz viel von den sich vergrößernden Informationsfluten gehört. Im medizinischen Bereich wird die Zahl zitiert, dass sich das Fachwissen alle fünf Jahre verdoppelt. Es gibt sogar Zahlen dass sich das gesamte Wissen der Menschheit alle zwei Jahre verdoppelt. Es ist klar, dass selbst ein engagierter Haus- oder Klinikarzt, seine Fortbildungszeit nicht entsprechend anpassen kann. Wenn man vor diesem Hintergrund den Umfragen glauben will, dass der Großteil der Ärzte weniger als fünf Stunden im Monat mit dem Lesen von Fachartikeln verbringt, dann kann sich ein Patient durchaus unwohl fühlen bei der Diagnose oder Behandlung.

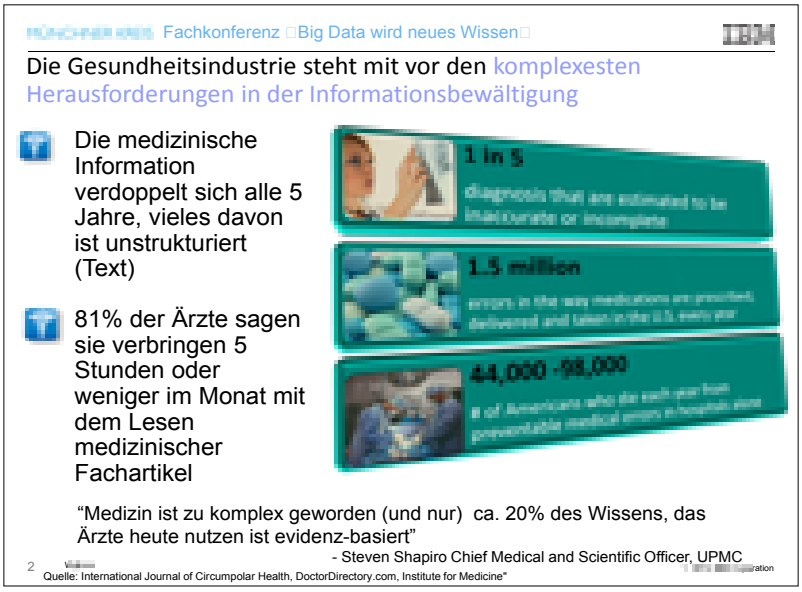

Bild 1

Bild 1 zeigt einige Zahlen aus den USA, die ich als Nichtmediziner nur unkommentiert wiedergeben kann. Eine von fünf Diagnosen ist entweder unvollständig oder sogar falsch. 1,5 Millionen Fehler gibt es beim Verabreichen von Medikamenten. Erschreckend ist die riesige Zahl ist von ca. 50.000 bis 100.000 Amerikanern, die jedes Jahr an vermeidbaren medizinischen Fehlern sterben. Da scheint es durchaus ein reales Problem zu geben. Einer der Gründe mag sein, dass die Medizin einfach zu komplex geworden ist. Viele Entscheidungen werden nicht evidenz- oder faktenbasiert getroffen. Entscheidungen auf Intuition und Erfahrung zu treffen kann durchaus etwas Gutes sein. Aber wenn Faktenwissen gar nicht erst zur Verfügung steht, dann kann das Kopfwissen nicht gegen die ganzheitliche Wahrnehmung und das Bauchgefühl abgewogen werden. Und das ist dann sicher nicht mehr gut.

Was kann eine Lösung für dieses Problem sein? Wenn die Mediziner nicht mehr mit dem Lesen nachkommen, dann wäre es schön, wenn man ein System hätte, das Informationen stellvertretend nachlesen kann. Und lesen tut man eben geschriebene Texte.

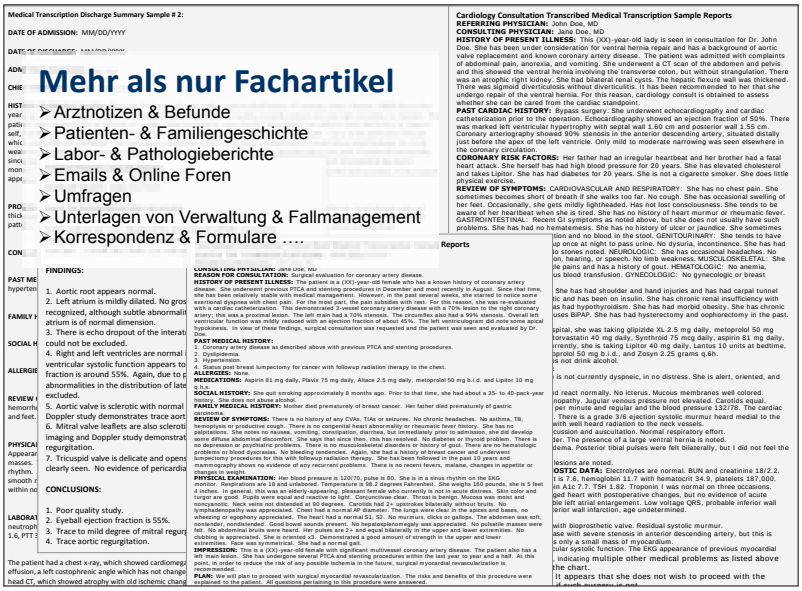

Bild 2

Der größte Teil des Wissens, das wir also so beklagenswert fehlend oder kaum mehr leistbar identifiziert haben, liegt in der medizinischen Welt in Texten vor (Bild 2). Es sind nicht nur die Fachartikel, die ein Arzt zur Fortbildung braucht. Im Alltag von niedergelassenen und Klinikärzten, gibt es auch noch Ärztenotizen, Befunde, Familiengeschichten, Berichte aus der Pathologie, elektronische Textdaten, Unterlagen aus der Verwaltung und noch vieles Andere mehr. Bild 2 zeigt einige Beispiele von solchen Texten.

Zugegebenermaßen ist nicht alle medizinische Information Text. Es gibt natürlich Information wie Laborberichte. Das sind Zahlen. Blutwerte, Zuckerwerte. Ein Blutbild ist ein einziger großer strukturierter Datensatz und kein Aufsatz. Aber es gibt eben auch sehr, sehr viel Text. Das heisst ein System, das bei der Lösung von unseres Problem helfen soll, wäre gut bedient, wenn es auch sehr gut mit Text umgehen könnte.

In diesem Fokus auf Text unterscheidet sich der hier vorgestellte analytische Ansatz von anderen Ansätzen, die auf dieser Tagung besprochen wurden. Nicht alle neuen Ansätze im Bereich Analytics und Big Data haben notwendigerweise viel mit Text zu tun. Sie erinnern sich sicher an den Vortrag mit der Vorstellung des IBM Systems, das Frühgeborene überwacht: Dieses System analysiert Datensätze aus einer Vielzahl von Sensoren zur Früherkennung von gesundheitlichen Problemen. Diese Anwendung verwendet Big Data, maschinelles Lernen, tiefe Analyseverfahren, aber sie fokussiert nicht auf Text und verwendet keine Watson Technologien. Nicht alles, was wir als IBM machen, und nicht alles was Big und Analytics ist, ist auch Watson. Ich spreche hier speziell über Watson-artige Analytiken, und das sind Problemstellungen, bei denen Text ein entscheidender Teil der Lösung ist.

Soweit haben wir also ein ernstes Problem identifiziert und wir haben einen Hinweis, wo eine Lösung liegen könnte: Wissen aus Texten könnten ein Schlüssel sein. Jetzt wechsle ich für kurze Zeit das Thema und stelle Ihnen eine amerikanische Quizshow vor, in der IBM einen Computer gegen menschliche Gegenspieler antreten ließ.

Bild 3

Das klingt nach reiner Unterhaltung aber wir werden sehen, dass aus diesem Spaß ganz wertvoller Ernst und Nutzen entstehen kann (Bild 3). Vielleicht wird auch klar, dass IBM dieses Quizshowprojekt nicht nur zum Zweck des Marketings gemacht hat. Ich will sicher nicht sagen, dass IBM den Marketing- und Imagegewinn bereut. Aber wir werden sehen, dass dieses Projekt von Anfang an darauf ausgerichtet war etwas zu erschaffen, das jenseits des Spielgewinns auch Probleme in der realen Welt lösen kann.

IBM hat also mit seinem Computer an dieser Quizshow teilgenommen und glücklicherweise auch gewonnen. Das war übrigens nicht sicher und auch für IBM spannender, als es vielen in der Firma lieb war. Denn ‚Jeopardy' ist als Spiel wirklich die ultimative Herausforderung für Forscher, die sprachverstehende Frage-Antwort-Systeme entwickeln und das aus ganz verschiedenen Gründen. Um zu verstehen warum, muss man sich in Deutschland erst einmal klarmachen, dass ‚Jeopardy' nicht ‚Wer wird Millionär' ist. Jeopardy hat keine Multiple Choice Fragen. Jeopardy hat nicht den freundlichen Herrn Jauch, der die Kandidaten Minuten lang überlegen lässt und zwischendrin Scherze macht. Jeopardy wird gegen Profispieler gespielt. Da gibt es „Ligen". Es gibt Leute, die diese Art von Spiel betreiben seit sie 16 Jahre sind. Diese Spieler leben davon, solche Fragen zu beantworten. Bei diesem Spiel geht es um Reaktionszeiten und um sehr breites Wissen. Multiple Choice Quizz könnte man durch Auswürfeln und Raten gewinnen. Ein Computersystem, das „Wer wird Millionär" gewinnt, kann eine Studentengruppe übers Wochenende schreiben.

Bild 4

Jeopardy dagegen ist sehr viel schwieriger. Die größten Herausforderungen sind auf dem Bild 4 dargestellt. So muss das System z.B. eine genaue Textanalyse durchführen, die auch mit subtilen Bedeutungsunterschieden, Ironie, Rätseln, Wortspielen umgehen kann. Ein anders Problem ist die Breite des abgefragten Wissens. In Jeopardy gibt es Kategorien wie Sport, Politik, Kultur, Kochen, die alten Römer, Mode des 20. Jahrhunderts usw. Die Kategorien sind klassischer sogenannter „Long Tail“: Viele Kategorien sind ganz selten. Am Anfang hat das Forscherteam bei IBM evaluiert, ob man das notwendige Wissen manuell semantisch so modellieren kann, wie es z.B. im vorhergehenden Vortrag für die Anwendung bei der Frankonia-Jagd dargestellt wurde. Dieser Ansatz hat sich aber als chancenlos herausgestellt bei der Vielzahl und Offenheit der möglichen Themen bei Jeopardy. Das war genau einer der Gründe, wieso Jeopardy als Forschungsprojekt so interessant war. Ansätze, die eine vollständige oder auch nur weitreichende Modellierung aller beteiligten Kategorien voraussetzen waren nicht anwendbar, weil der Aufwand für Erstellung und Pflege der Modelle zu groß war. Das IBM Team musste also einen Ansatz finden, der in der Lage war, Antwortqualität basierend auf einer flachen Ontologie oder einer oberflächlichen Semantik zu erbringen. Natürlich werden in Watson auch die semantischen Verfahren verwendet, dass die in den Vorgängervorträgen erwähnt wurden. Watson verwendet ontologische Technologien wie SPARQL, RDF etc. Watson greift auch auf die semantisch modellierten DBpedia Fakten zu. Aber darauf alleine konnte das Forscherteam kein System aufbauen, das Gewinnchancen in Jeopardy hat. Watson verwendet einen eklektischen Ansatz der verschiedene Verfahren kombiniert und semantische Technologien sind nur ein Teil davon. Ein solches Verstehen von Texten aus einem breiten Themenspektrum ohne aufwendige manuelle Modellierung ist für viele Anwendungsgebiete auch jenseits der Quizshow entscheidend. Nicht zuletzt in der Medizin wie wir noch sehen werden.

Eine weitere Herausforderung des Jeopardy-Spiels ist, dass es bei Jeopardy nicht möglich ist zu gewinnen, wenn der Spieler nicht weiß, ob er eine Antwort nicht weiß. Er muss wissen, ob die Antwort, die ihm in den Sinn kommt, wahrscheinlich falsch ist und ein gutes Gefühl dafür haben, dass es z.B. 35 % sicher ist, dass ein Antwortkandidat richtig ist. Warum sind

diese Verlässlichkeitsabschätzungen in Jeopardy so wichtig? In Jeopardy gibt es die Regel, dass ein Spieler, der sich um eine Frage bewirbt und diese dann falsch beantwortet, den Geldwert der Frage abgezogen bekommt. Das heißt, ein Spieler sollte sich besser nicht um eine Frage bewerben, wenn er seinen Antwortkandidaten als unsicher einschätzt. Außer natürlich wenn der Spieler momentan weit im Rückstand liegt und etwas riskieren muss. Aus diesen Überlegungen wird klar, dass für eine erfolgreiche Spielstrategie eine verlässliche Konfidenzeinschätzung der Antwort entscheidend ist. Nur wenn ein Spieler diese Fähigkeit hat, kann er die ideale Strategie spielen. Wir werden später sehen, dass diese Verlässlichkeitsabschätzung auch für die Medizin und die Diagnose sehr wichtig sind.

Als letzte Herausforderung, die Jeopardy so besonders macht, kommt noch die hohe Geschwindigkeit des Spiels dazu. Antworten müssen sehr schnell gegeben werden. Für ein Computersystem schließt das Verfahren aus, wie sie z.B. in akademischen Systemen verwendet werden, die zwar 98 % Ergebnisqualität haben können, aber dafür über Nacht an der Antwort rechnen müssen. Bei einem TV Spiel würde kein Zuschauer so lange warten wollen. Nachdem sich ein Spieler um eine Frage beworben hat, muss dieser eine Antwort mit einer verlässlichen Konfidenzabschätzung im Durchschnitt in unter 3 Sekunden haben. Nur dann hat er eine faire Chance zu gewinnen. Geschwindigkeit ist aber auch in realen Anwendungen wichtig, z.B. eben in Dialogsystemen.

Zusammenfassend kann man sagen, dass Jeopardy Anforderungen stellt, die gar nicht spezifisch für eine Quizshow sind: Ein System muss schnell sein, es muss breites Wissen haben, mit Konfidenzen umgehen können. und wir brauchen ein robustes Verstehen von Texten.

Könnte man diese Herausforderungen auch meistern mit den Suchtechnologien, die wir alle täglich benutzen? Viele Leute sind überzeugt, dass man ‚Wer wird Millionär' gewinnen kann, wenn man die Antwortvorschläge in einer Internetsuchmaschine nachschlagen dürfte. Wenn das so ist, wo ist dann das technische Problem?

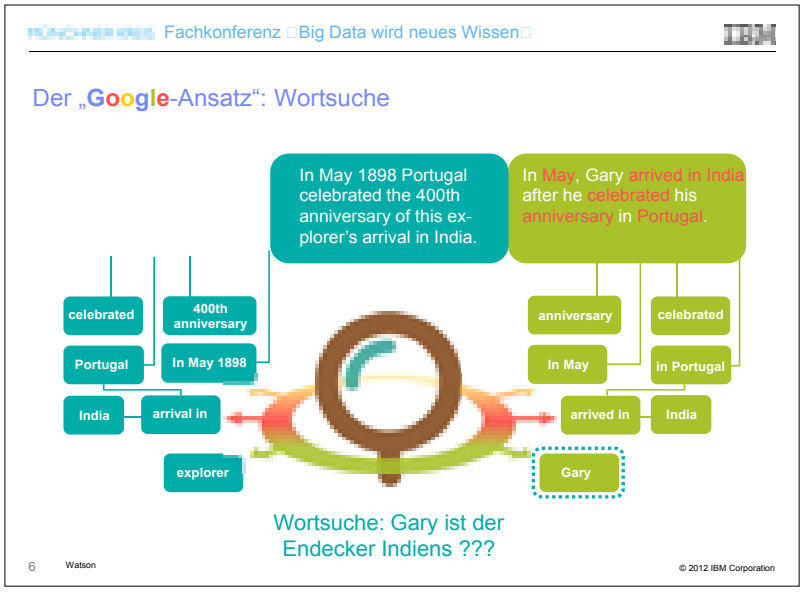

Bild 5

Bild 5 zeigt links oben eine typische Jeopardy „Frage“ - diese haben traditionell kein Fragezeichen am Ende–, bei der man herauszufinden soll, wer der erwähnte Entdecker ist. Würde man mit einem klassischen Stichwortansatz suchen, dann findet eine Suchmaschine typischerweise Textpassagen, wie die rechts dargestellte, die ganz viele gemeinsame Worte mit der Frage hat. Daraus müsste man dann schließen, dass „Gary“ Indien entdeckt hat. Das ist natürlich falsch. Es ist bei komplexen Fragen also nicht mit einfacher Stichwortsuche getan.

Was man stattdessen anwenden muss, sind die Verfahren, die heute Nachmittag schon erwähnt wurden: Erst einmal muss eine sprachliche Analyse durchgeführt werden. Syntaxanalyse, Referenzauflösung und semantische Zusammenhänge ermittelt werden. Das Ergebnis der Analyse kann dann gegen eine Wissensbasis verglichen werden. Die Wissensbasis ist aufgebaut worden aus Texten, die sich das System im Vorfeld einverleibt hat.

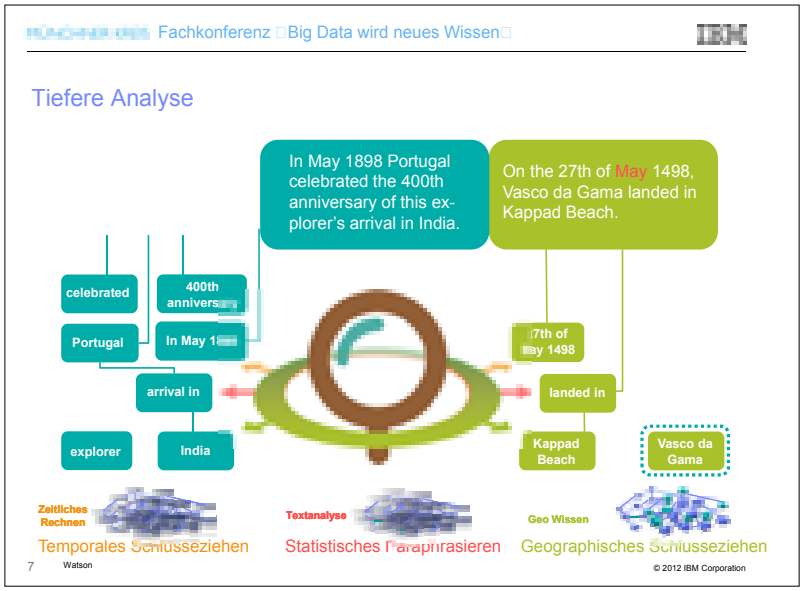

Bild 6

Bild 6 zeigt rechts oben einen Satz mit einem Antwortkandidaten aus dieser Wissensbasis. Nach dem klassischen Keyword Ansatz hat dieser Satz nur genau ein Wort mit der Frage gemeinsam, d.h. traditionelle Suchtechnologien würde es als mögliche Antwort nicht sehr hoch einordnen.

Aber auf tieferen Ebenen, wenn ein System Datumsangaben und Datumsarithmetik versteht, kann es ermitteln, dass hier verwandte Zeiten erwähnt sind, weil das Jahr 1898 minus der 400 Jahre Jubiläum aus der Frage zu der Jahreszahl 1498 aus der Antwort passt. Um solche Bezüge herzustellen, braucht das System ein Datumsverständnis. Es muss zumindest so viel semantisch über Datumsangaben wissen, dass es solche Arithmetik und Bezüge herstellen kann. Solche Verfahren werden auch in ontologischen Ansätzen verwendet werden und sie sind auch für Watson notwendig. Aber Datumsverständnis ist nicht sehr themenspezifisch und es ist nicht alles, was für das Herstellen eines Bezuges zwischen Frage und Antwort notwendig ist. Andere Verfahren helfen bei anderen Problemen: „Arrival“ und „landed“ hat auf der Oberfläche von Stichworten nichts miteinander zu tun. Aber ein System kann herausfinden, dass diese Worte Synonyme sein können. Dieses Wortwissen kann man in Lexika

hinterlegen oder, wenn man ein wartungsfreies System bauen will, kann man versuchen, über statistische Analysen automatisch zu ermitteln, dass diese beiden Worte wahrscheinlich das gleiche bedeuten. Das kann ein System z.B. daran erkennen, dass sie häufig in ähnlichen Wortzusammenhängen vorkommen. Der reduzierte Wartungsaufwand gegenüber einem manuell gepflegten Lexikon ist besonders dann wichtig, wenn ein System jenseits der Allgemeinsprache auf neue Domänen und kompliziertere Bereiche, wie z.B. der Medizin angewendet werden soll.

Ein dritter Bezug wird hier mittels Geowissen hergestellt. Hier ist eine Geodatenbank hinterlegt und mit deren Hilfe kann das System etablieren, dass es eine „Located-in“ Relation zwischen „Kappad Beach“ und „India“ gibt. Das Beispiel lässt erkennen, dass eine Vielzahl an verschiedenen tieferen Analyseverfahren jenseits des Key Word Matching angewendet werden muss, um die richtigen Bezüge herzustellen. Kein einzelnes Verfahren würde ausreichen.

In dem Beispiel sind ganz unten die jeweiligen Wissensbasen dargestellt. Diese sollen Berge von Wissen symbolisieren. Aber wo kommt dieses Wissen ursprünglich her? Wie kommt ein System wie Watson an solches Wissen? Es wäre nach den Regeln des Spieles erlaubt gewesen, das Wissen von Hand einzugeben. 10.000 von Helfern bei IBM hätten die Fakten in Tabellen eintippen können. Es war mit dem Veranstalter von Jeopardy nur vereinbart, dass das IBM System die Fragen live beantwortet und dabei nicht mit dem Internet verbunden sein darf. Was aber im Vorfeld passiert, was auf den Festplatten des Systems liegt, welche Wissensbasis in welchem Format, war IBM überlassen. Das IBM Team hat aber keine manuellen Verfahren zum Wissenserwerb verfolgt, weil automatische Verfahren besser zu den Forschungsperspektiven und strategischen Ideen passen. Die Idee war einen Ansatz zu verfolgen, bei dem sich das System das Wissen selbst „anliest“. Genau das was Ärzte aus Zeitgründen immer seltener leisten können.

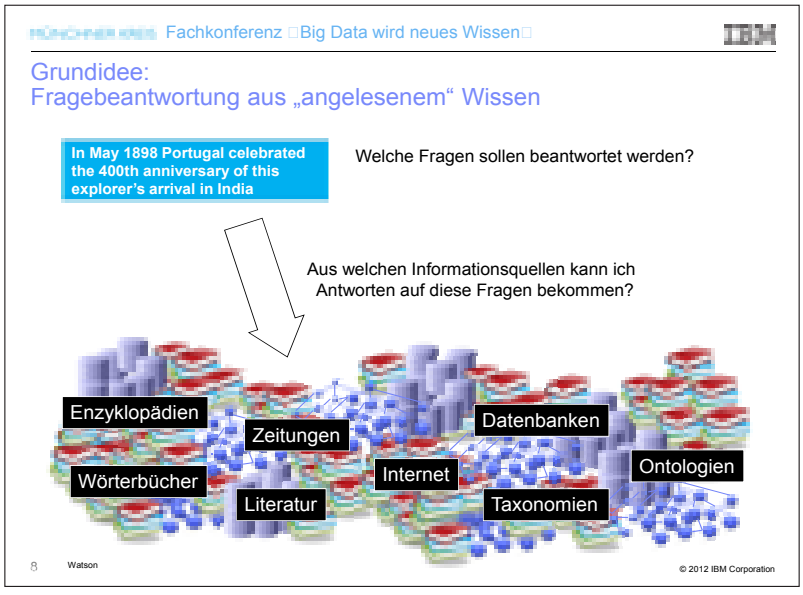

Bild 7

Anlesen klingt für Menschen sehr leicht, aber was heißt denn Anlesen für ein Computersystem? Es heißt, das als Input für die Wissensbasis Texte zur Verfügung stehen, welche die Fakten enthalten, die für die Fragenbeantwortung notwendig sind. Bild 7 zeigt die wichtigsten Quellen, die Watson in Jeopardy verwendet hat, im Überblick. Neben Texten gibt es da auch eine Handvoll von strukturierten DBPedia Einträgen und die ebenfalls strukturierte Internet Movie Data Base. Aber der Kern sind 10 Jahrgänge New York Times und die Wikipedia als einfacher Text. Diese Quellen wurden aufgenommen in der Annahme, dass sie voraussichtlich das Wissen enthalte, das man für Jeopardy Antworten braucht. In Jeopardy ist noch nie nach astrophysikalischen Details gefragt worden und auch biochemische oder physikalische Forschungsergebnisse sind kein Thema. Antworten für Fragen aus diesen Bereichen wären in Fachjournalen zu finden. Diese waren für Jeopardy aber nicht relevant. Das System wird also so aufgesetzt, dass es nur die Quellen liest, die voraussichtlich das Wissen enthalten, das für die Fragedomäne wichtig ist. Im Fall der Quizshow waren das Themen aus dem Bereich des Allgemeinwissens wie sie in Zeitungen und der Wikipedia eben vorkommen.

Anders als beim Menschen heißt Lesen für Watson maschinelles prozessieren. Dabei wird nicht nur eine einzelne, homogene Wissensbasis aufgebaut. Wie erwähnt ist das Watson-System eklektizistisch im Ansatz und kombiniert viele Verfahren. Da wird sowohl ein spezieller Passageretrieval Index aufgebaut als auch noch ein konventioneller Lucene Suchindex. Es wird auch noch ein Triple Store mit extrahierten Fakten gefüllt. Eine ganze Batterie von Algorithmen wird auf die Texte losgelassen und jeder einzelne kann seine eigene Wissensbasis aufbauen. Initial ist dabei eine große Menge von Texten zu verarbeiten, später immer nur die neu dazu kommenden. Wenn allerdings die Leseverfahren zum Wissenserwerb verbessert werden müssen alle Text neu analysiert werden.

Dieses „verstehende“ Lesen beherrscht das System nicht ohne Vorbereitung. Watson ist keine universelle künstliche Intelligenz, der nur einige Basisaxiome mitgegeben werden müssen und die darauf aufbauend sofort Wissen aus beliebigen Textsorten und Bereichen erwerben kann. Das ist analog zum menschlichen Lesen bei dem ein Schüler ca. 16 bis 18 Jahre zur Schule gehen muss, bevor er eine Zeitung wirklich verstehend lesen kann. Entsprechend muss auch Watson auf einen Bereich wie z.B. Allgemeinwissen aus Zeitungen konfiguriert werden. Danach kann es alle Zeitungsartikel oder Wikipedia Artikel lesen. Aber medizinische Texte wird es nur sehr schlecht verarbeiten können. Genauso wenig wie ein Schüler, der Zeitungstexte gut versteht, einen Medizinfachartikel lesen kann. Zur Anpassung an einen neuen Bereich muss investiert werden in Dinge wie das Erlernen eines Fachvokabulars. In diesen Anpassungsprozessen werden z.B. Synonyme erlernt. Es gibt also einen Initial- und Pflegeaufwand für jede neue Domäne. Aber wenn eine Domäne erst einmal konfiguriert ist können neue Texte automatisch verarbeitet werden.

Der Bereich des Wissenserwerbs aus größeren Textmengen ist der Berührungspunkt von Watson und Big Data. Allerdings ergeben zehn Jahre New York Times und die gesamte Wikipedia gar nicht so viele Daten. Watson ist kaum „Medium Data“. Technologisch läuft der Wissenserwerb aber auf der Big Data Plattform Hadoop. Der Grund für die Verwendung dieser Skalierungsplatform ist aber stärker die Skalierung zur Bewältigung der Tiefe der Analyse und weniger der Breite über große Datenmengen.

Die bisherigen Ausführungen haben hoffentlich klarer werden lassen warum sich IBM in dem Bereich Quizshow engagiert hat. Mit einem System für Quizshows kann man viele Verfahren erproben und entwickeln, die sich auf andere Bereiche übertragen lassen.

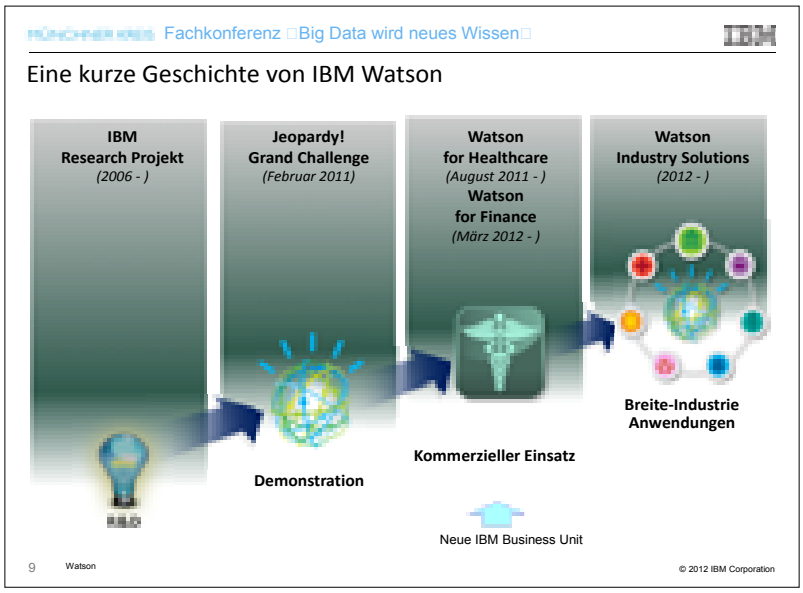

Bild 8

Bild 8 zeigt den Verlauf des Projektes Watson von der Vergangenheit bis in die Zukunft. In der Öffentlichkeit ist das Projekt vor allem mit dem Jeopardy Gewinn 2011 bekannt geworden. Begonnen hat das Watson Projekt aber schon 2006 und schon davor gab es eine Gruppe in IBM Research, die an Question-Answering-Systemen geforscht hat. D.h. der hier gezeigte Startpunkt 2006 war nicht bei null sondern bei einem Forscherteam, das auf akademischem Weltniveau Question-Answering gemacht hatte. Von da ausgehend waren noch fünf Jahre Arbeit von ca. 25 Personen notwendig, um mit Watson Jeopardy gewinnen zu können.

IBM hat damit bewiesen, dass solche Probleme lösbar sind und hat damit auch dem ganzen Bereich der natürlichen Sprachverarbeitung und der Expertensysteme zu frischem Wind verholfen. Aber für eine kommerzielle Firma kommt die wahre Herausforderung erst jetzt, weil es eben nicht nur darum geht eine Quizshow gewinnen zu wollen sondern auch praktische Anwendungsprobleme in der realen Welt lösen können.

Wir haben schon mehrfach von medizinischen Anwendungen gehört. Aber es gibt inzwischen auch „Watson for Finance“. Es gibt mit der Citygroup einen Partner, mit dem wir gemeinsam versuchen, die Watson Technologie auch für Probleme im Finanzbereich anzuwenden. Weitere Anwendungsbereiche werden folgen.

Im medizinischen Bereich ist der Partner der IBM für Watson Well Point. Mit dieser Krankenversicherung läuft im Moment eine Pilotphase. In Produktion ist das Projekt aber noch nicht und deswegen gibt es noch keine abschließenden Zahlen zur Qualität und Nutzen des Systems.

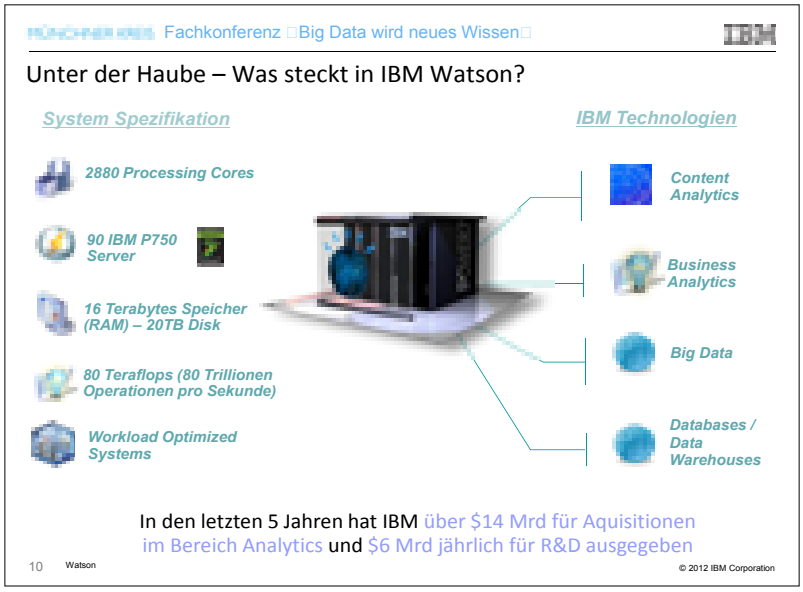

Bild 9

Bild 9 zeigt das Watson System für die Quizshow. Es enthält 90 IBM Power 750 Blades. Ohne ins Detail zu gehen soll folgender Punkt betont werden: Watson ist einerseits ein ziemlich großes und leistungsfähiges System aber es ist kein Supercomputer, der in der Liga der schnellsten Rechner der Welt vorne mitspielen könnte. Es ist sehr gute, weit ausgebaute Industriehardware von der Stange. Das ist ein Unterschied zu dem IBM Schach Challenge. Damals hat ein IBM Computer zum ersten Mal einen menschlichen Großmeister im Schach besiegt. Damals musste IBM die Rechnerhardware erst entwickeln mit der das möglich war. Für Watson war der Schwerpunkt die Software.

Man sieht das Watson viel Hauptspeicher hat, wogegen von Festplattenkapazität auf Big Data Niveau gar nichts zu sehen ist. Das liegt daran, dass wir hier das Watson Laufzeitsystem sehen und nicht das System, das sich im Vorfeld das Wissen anliest. Das Laufzeitsystem, das in der Quizshow Fragen beantwortet, bekommt nur die fertige, kondensierte Wissensbasis aufgespielt und arbeitet nahezu komplett im Hauptspeicher. Das ist notwendig, weil nur drei Sekunden Zeit zur Beantwortung der Frage bleiben und Festplattenzugriffe einfach zu langsam wären. Das ganze große System ist übrigens nur für einen Single User Betrieb ausgelegt.

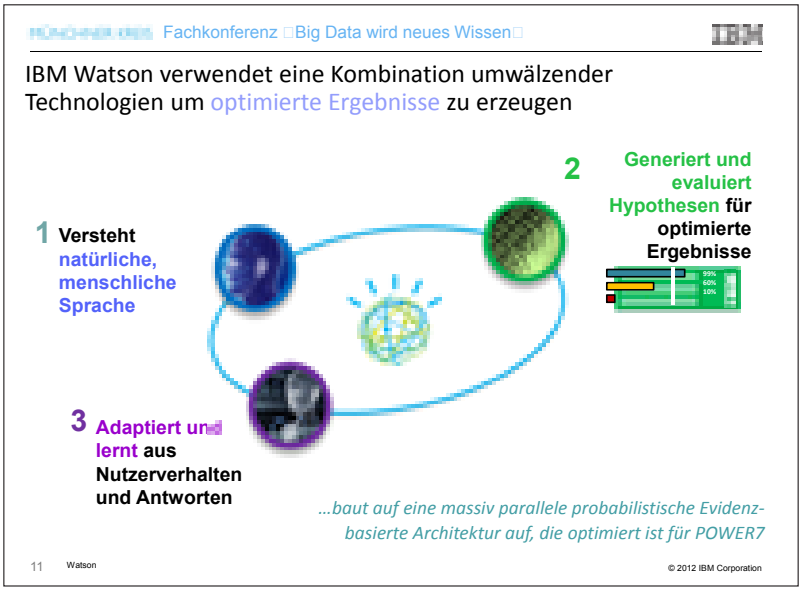

Bild 10

Bild 10 versucht herauszustellen was Watson von anderen Analysesystemen unterscheidet. z.B. von dem Monitoring System für Frühgeborene von dem mein Vorgänger berichtet hat. Drei Kernpunkte machen Watson aus. Es kann natürliche menschliche Sprache verarbeiten. Ob die gesprochen oder geschrieben vorliegt, ist dabei nicht zentral. Das ist letztendlich eine Frage der Vor- und Nachverarbeitung mittels „Speech to Text“ oder „Text to Speech“ Verfahren.

Ein zweiter zentraler Punkt, der weniger offensichtlich ist, liegt im Evaluieren von Hypothesen. Diese Hypothesen müssen daraufhin bewertet werden mit welcher Wahrscheinlichkeit sie korrekt sind. Wir haben gesehen dass es für das Gewinnen in Jeopardy wichtig ist, dass diese Wahrscheinlichkeitsaussagen zuverlässig sind. Bei Watson spricht man daher von sogenannten verlässlichen Konfidenzen. Das ist an sich schon schwierig, aber für ein System wie Watson besonders herausfordernd, weil einer der zentralen Punkte in der Architektur von Watson ist, dass es ganz heterogene Analysetechnologien zusammenfasst. Jede dieser Technologien hat aber ihre eigenen sogenannten Konfidenz-Scores auf jeweils einer eigenen Skala. Die einzelnen Teile liefern eben nicht wahrscheinlichkeitstheoretisch sauber kombinierbare Vorhersagen. Das Verfahren mit dem Watson die Summe all dieser heterogenen Abschätzungen nimmt und daraus eine verlässliche Konfidenz ermittelt, ist wahrscheinlich wichtiger jeder einzelne Algorithmus im System.

Der dritte zentrale Punkt ist das maschinelle Lernen. Dabei muss man sagen, dass Lernen in Watson nicht mit einem zentralen, generischen General Purpose Learner implementiert ist. Da gibt es keinen Algorithmus, der beliebige Dinge lernen kann, so wie vielleicht ein Kind Fußball genauso lernen kann wie Mathematik und Französisch. In Watson geht es um Lernen an vielen verschiedenen Stellen im System. Da gibt es Lernverfahren, die lernen z.B. Synonyme über Textzusammenhänge. Es gibt auch Lernverfahren, die im Verlauf der Quizshow lernen, welche Antworttypen für eine bestimmte Fragekategorien erwartet werden. Wenn z.B. alle bisherigen korrekten Antworten in einer Kategorie Städte waren dann lernt das

System nicht mit Bundesländern zu antworten selbst, wenn diese eigentlich ein höheres Konfidenzranking hätten.

Der gesamte Ablauf in Watson findet massiv parallel statt. Nur so ist die notwendige Antwortgeschwindigkeit zu erreichen. Deswegen ist Watson auf einer Integrationsarchitektur namens UIMA aufgebaut, die genau diese Parallelität unterstützt. Und es läuft auf einer Hardware, die solche parallelen Verarbeitungen sehr gut ausführen kann.

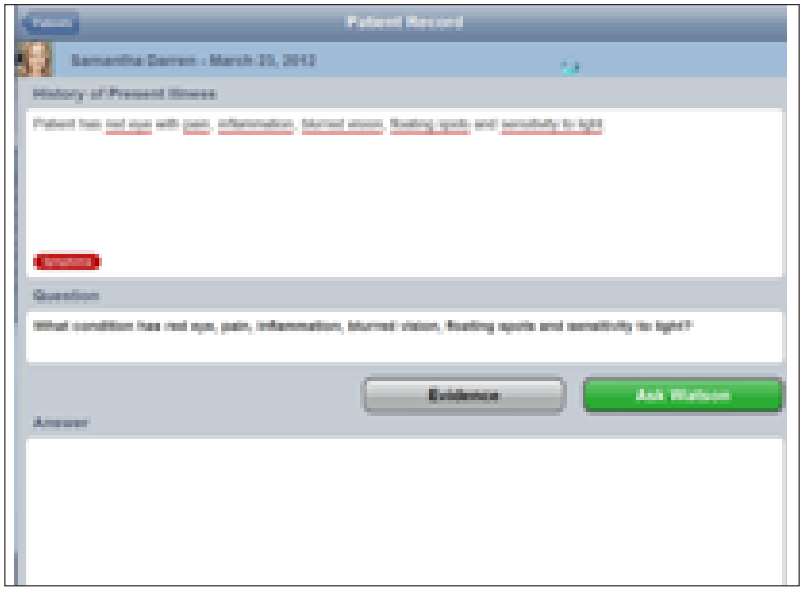

Bild 11

Bild 11 zeigt ein Beispiel, das veranschaulichen soll, wie sich Watson Technologien im medizinischen Bereich anwenden lassen. Dieses Beispiel ist nicht in genau dieser Form bei Well Point im Einsatz, sondern dient zur Illustration. Wir haben hier eine hypothetische Patientin mit Augenproblemen. Oben sieht man eine Beschreibung ihrer Symptome. Diese Symptombeschreibung ist ganz ähnlich wie eine Jeopardy Frage: „Welche Krankheit hat Symptome von rote Augen, Schmerzen, Entzündung, verwaschenes Sehen, Flecken im Auge und Lichtempfindlichkeit?“

Das System kann versuchen in der Wissensbasis von medizinischen Fachartikeln, Fallbeschreibungen, Studien und Lehrbüchern die wahrscheinlichsten Krankheiten zu suchen.

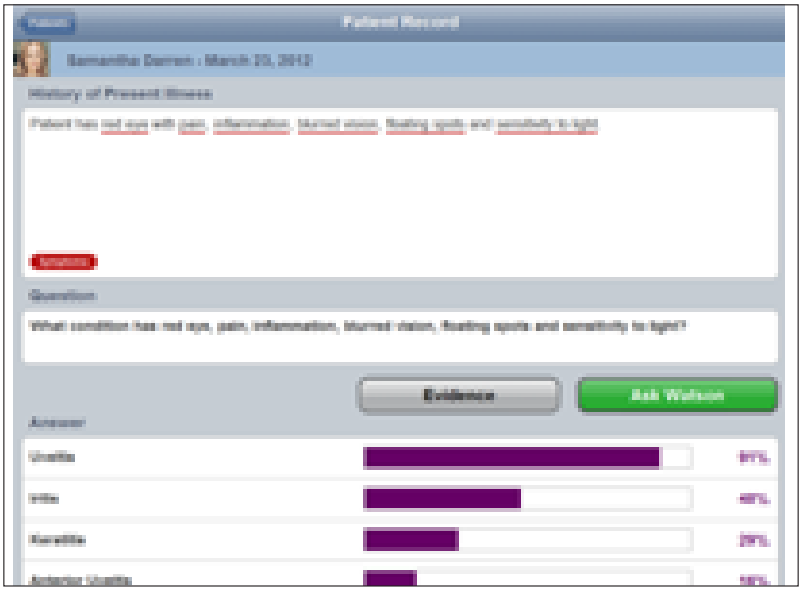

Bild 12

Bild 12 zeigt beispielhafte Antworten. Nicht jeder wird wissen was Uveitis ist. Es handelt sich dabei um eine Entzündung einer Haut im Auge. Auch ein menschlicher Mediziner wäre sicher auf ähnliche Diagnosen gekommen. Der Typ Frage, der hier dargestellt ist, kommt aus Fragekatalogen für Medizinstudenten im Grundstudium. Das ist noch nicht die Sorte Fragen, bei denen Watson typische menschliche Diagnoseleistungen übertreffen wird, es zeigt aber die Richtung.

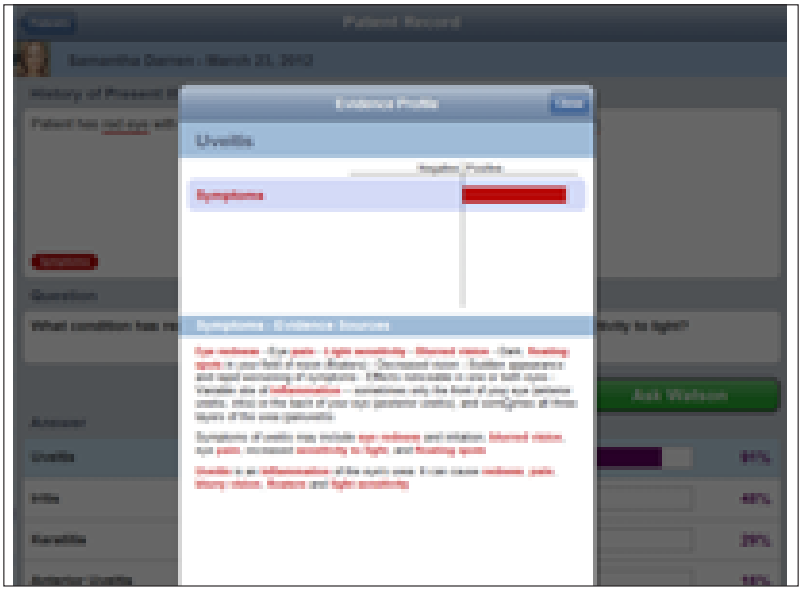

Bild 13

Bild 13 zeigt einen Aspekt von Watson der im medizinischen Bereich sehr wichtig ist, der aber in Jeopardy gar nicht erlaubt war. In einem Diagnosesystem muss ein Arzt eine Watson Diagnose hinterfragen können: Warum wurde diese Krankheit diagnostiziert? Warum mit 1 % Wahrscheinlichkeit? Das System soll darauf mit den Belegstellen antworten, um die Diagnose zu rechtfertigen. In Jeopardy gibt es solche Nachfragen nicht. Eine Antwort in der Quizshow ist richtig oder falsch. Aber in den Anwendungen als arztunterstützendes System das Zweitmeinungen liefert ist es wichtig Begründungen zu liefern. Ganz speziell in den Fällen, wo Watson eine unerwartete Diagnose liefert. Die Begründung kann Vertrauen in die abweichende Meinung schaffen und ein Element der Wissenserweiterung für den Arzt darstellen. In einem System, das jenseits des Medizinstudiums eingesetzt wird, muss die Begründung sogar noch ausführlicher sein als hier dargestellt. Es müssten Fachartikel, Referenzquellen, Studien und mehr Details dargestellt werden.

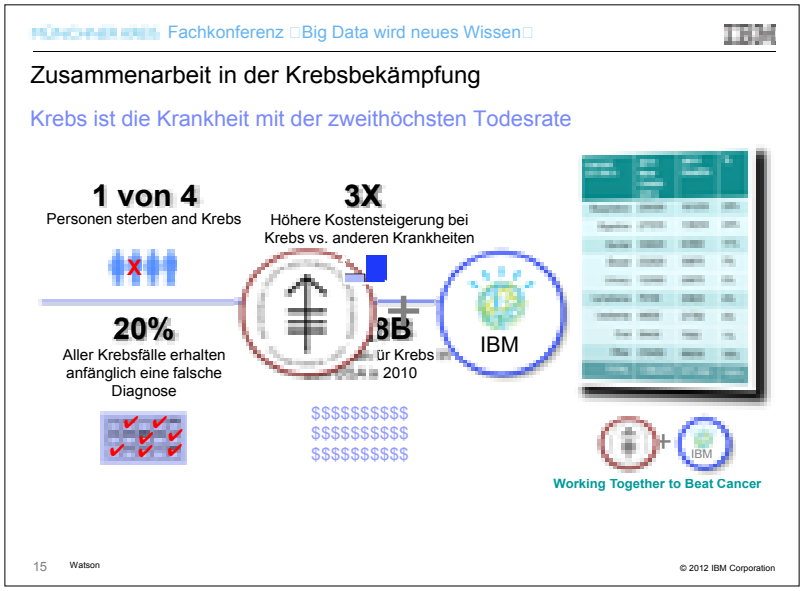

Bild 14

Das Beispiel sollte die grundsätzliche Idee der Diagnoseassistenz anschaulich gemacht haben. Aber an welchen Anwendungen arbeitet IBM nun konkret? Ein Thema ist die Krebsdiagnose. Bild 14 zeigt die Begründung, warum es besonders wichtig ist, etwas gegen die Krankheit Krebs zu unternehmen. Das ist der Fokus eines Projekts das IBM Research mit dem Memorial Sloan Kettering Cancer Center -einer großen, auf Krebs spezialisierte Klinik in den USA- durchführt.

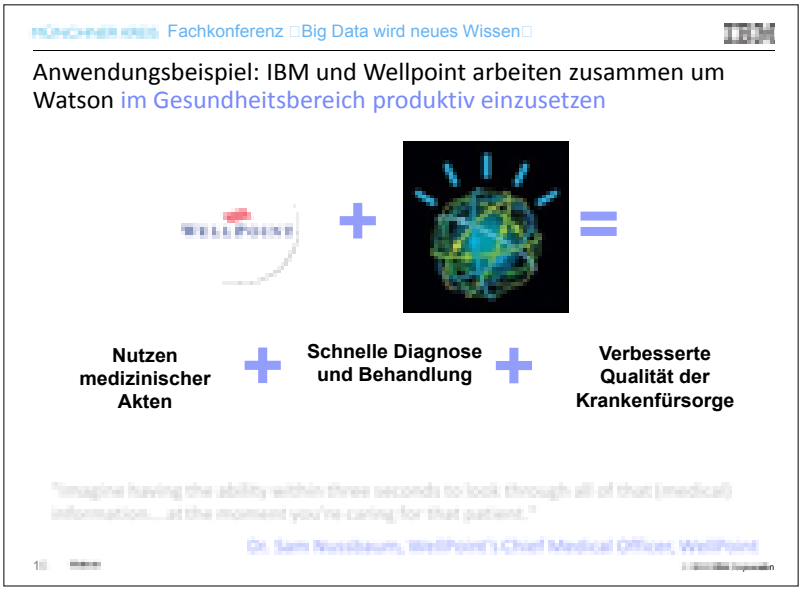

Bild 15

Bild 15 zeigt, dass eines der Szenarien das Wellpoint mit IBM umsetzen will das oben illustrierte Diagnoseszenario ist. Es geht aber um eine komplexere Variante als die eben vorgestellte.

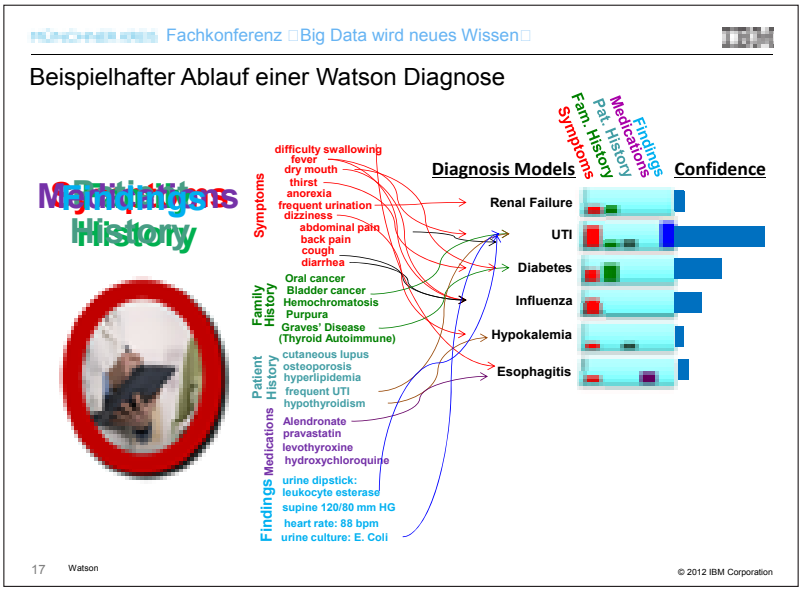

Bild 16

Bild 16 illustriert diese komplexere Variante der Diagnoseunterstützung. Auch hier gibt es verschiedenste Symptome, und das System hat ein Diagnosemodell gelernt. Hier sind die Symptome Nierenversagen, Harnwegsinfektion, Diabetes, Grippe und Hypokalemie. Aber es geht hier noch weiter: Das System erfragt auch die Abwesenheit von bestimmten Symptomen. Auch das ist etwas, das in Jeopardy nicht vorkommt, aber für ein Diagnosesystem wichtig ist. Es gibt aber noch mehr Input: Die Familiengeschichte, die ich erfragen muss, oder die vielleicht schon elektronisch vorliegt. Dann kommt noch die Patientengeschichte hinzu und wieder ändert sich die Gesamtgewichtung der Diagnosekandidaten. Es kommt noch Wissen hinzu, welche Medikamente der Patient bereits eingenommen hat. Auch das kann die Diagnose verändern. Als letztes werden noch konkrete Laborbefunde berücksichtigt. Letztendlich wird in der Gesamtsicht eine Harnwegsinfektion als wahrscheinlichste Krankheit diagnostiziert. Wie im vorherigen Beispiel werden also Prädiktoren aus verschiedenen Bereichen integriert und zu einer Gesamtevidenz kombinieret.

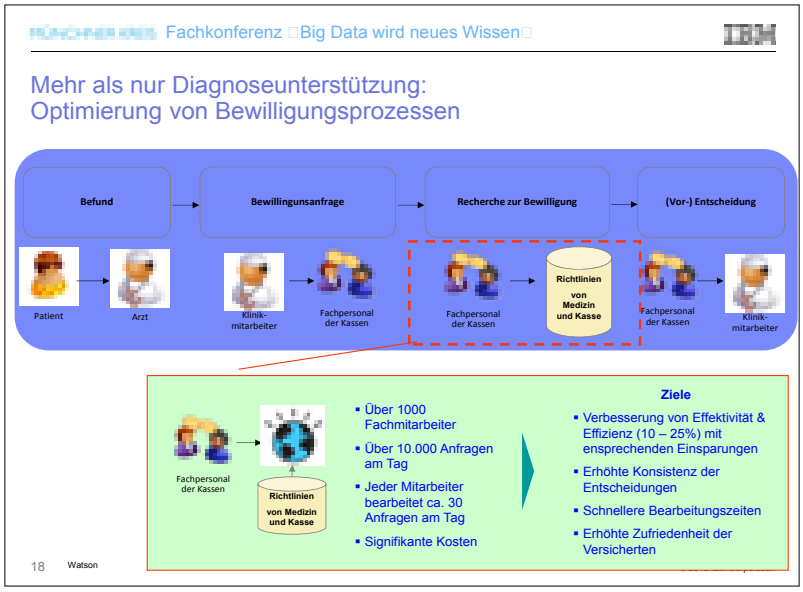

Bild 17

Diagnoseunterstützung ist aber nicht das einzige Anwendungsszenario im Gesundheitsbereich. Bei Wellpoint wird als Erstes ein ganz anderes Szenario umgesetzt, das in mancher Beziehung einfacher ist und das einen klaren finanziellen „Return on Investment“ hat. Bild 17 gibt einen Überblick dazu. In den USA gibt es im Gesundheitssystem etwas, das sich Preauthorization nennt – hier mit Bewilligungsprozess übersetzt. In Deutschland hat das keine direkte Entsprechung. Aber in amerikanischen Kliniken bekommt ein Patient nicht nur auf Entscheidung eines Arztes ein Medikament oder ein teures Verfahren wie MRI oder Röntgen oder eine andere Maßnahme wie z.B. eine Prothese. Das muss vorher von der Krankenkasse durch Spezialisten begutachtet und genehmigt werden. Ein Arzt, der eine Diagnose und einen Behandlungsvorschlag hat, muss das speziell geschulten Mitarbeitern der Krankenkasse vorlegen. Diese Personen führen dann eine Recherche durch um zu ermitteln, ob die Maßnahme in diesem ganz speziellen Fall medizinisch notwendig ist. Nur was nach Sachlage der aktuellen Vorschriften medizinisch notwendig ist, wird bewilligt. Eine Maßnahme, die medizinisch notwendig ist für einen 40jährigen Mann, der Raucher ist, eine Vorgeschichte von Übergewicht hat und schon dreimal operiert wurde, ist nicht notwendigerweise medizinisch notwendig bei einer 20jährigen Frau ohne Vorgeschichte. Die

Vorschriften und die medizinische Sachlage auf Grund derer entschieden wird ändert sich ständig, die Entscheidung muss aber immer wohlfundiert und nachvollziehbar sein. D.h. diese Recherchen sind relativ kompliziert, aufwendig und damit teuer.

Für diese Verfahren hat Wellpoint über 1000 Fachmitarbeiter mit über 10.000 Anfragen am Tag. Jeder Mitarbeiter kann aber nur 30 Anfragen am Tag bearbeiten. Das macht klar was für einen kostenintensiven Prozess dieses Verfahren darstellt. Es macht auch klar, dass es hier große Einsparpotentiale gibt. Wenn außerdem noch eine erhöhte Konsistenz der Ergebnisse und kürzere Wartezeiten erreicht werden, dann ergeben sich auch zufriedenere Patienten und bessere Qualität.

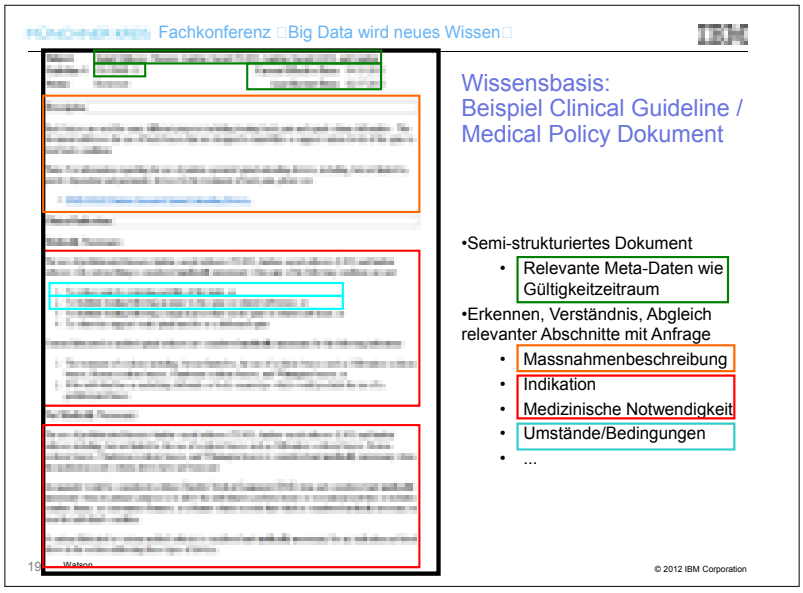

Bild 18

Dieses Einsatzszenario mag zwar nichts mit Diagnose zu tun haben. Es hat aber durchaus mit Texten zu tun. Bild 18 zeigt eine Vorschrift. Das ist ein langes Dokument. Man nennt es semi-strukturiert, weil es eine Vielzahl separater Felder hat, in denen wichtige strukturierte Informationen stehen: Wie lange ist das Dokument gültig? Wann ist es zum letzten Mal validiert worden? Welche Beschreibungen, welche Indikationen sind relevant? Die aufgeführten Umstände und Bedingungen müssen mit der aktuellen Patientensituation und der aktuellen Diagnose abgeglichen werden.

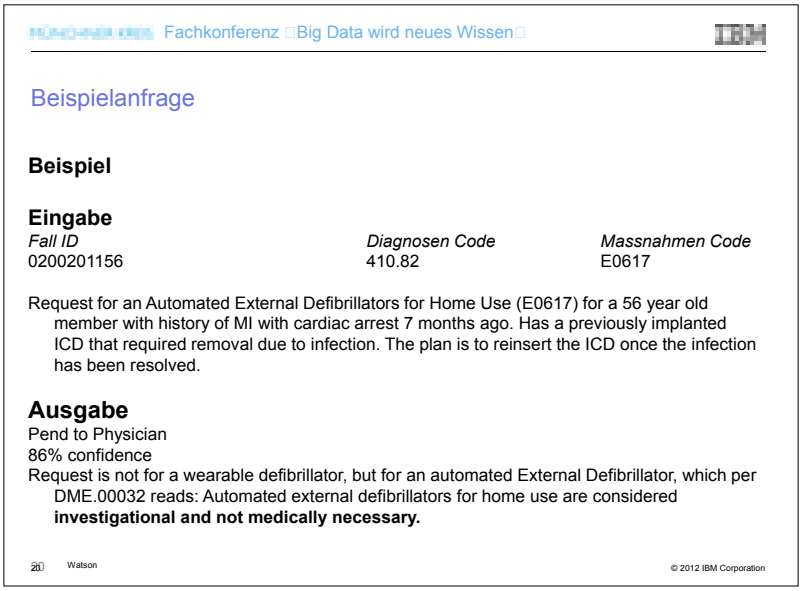

Bild 19

Bild 19 zeigt ein Beispiel von Eingabe und Ausgabe für das System. Hier wird ein automatisierter externer Defibrillator für den Hausgebrauch beantragt. Die Antwort von Watson ist, dass unter den angegebenen Umständen die Maßnahme medizinisch nicht notwendig ist. Es wird auch eine Konfidenz und Belegstelle zur Begründung gezeigt zusammen mit Verweisen auf weitere Details. Watson wird in diesem Szenario allerdings nie letztendlich negativ entscheiden dürfen. Ein Nein von Watson bedeutet hier, dass ein Arzt hinzugezogen werden muss. Watson darf also nur automatisch bewilligen, nicht automatisch ablehnen. Es war die Entscheidung von Wellpoint, das Verfahren so aufzusetzen und das ist wahrscheinlich sehr weise. Auch in diesem ganz anderen Szenario spielen aber ganz ähnliche Komponenten eine Rolle wie in der Diagnose. Es geht um Konfidenzen. Es geht um Textverständnis. Es geht um Lernen.

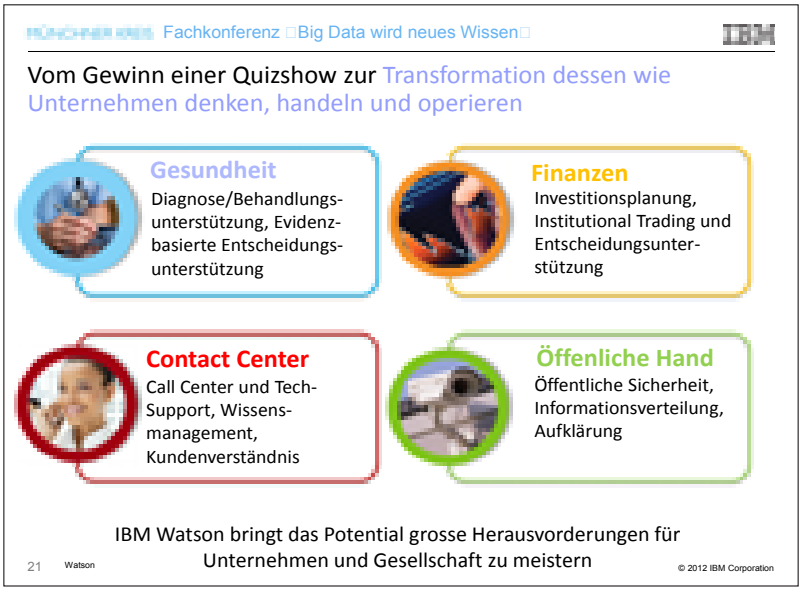

Bild 20

Aber Watson soll natürlich nicht nur im medizinischen Bereich eingesetzt werden. Investitionsberatung im Finanzbereich ist wahrscheinlich ähnlich komplex wie Diagnose im medizinischen Bereich. Contact Center, Wartung von komplizierten technischen Geräten, Dokumentanalyse bei der öffentlichen Hand sind weitere Gebiete in denen wir Anwendungen von Watson in der Zukunft sehen (Bild 20).

14 Schlusswort

Prof. Dr. Jörg Eberspächer, Technische Universität München

Meine Damen und Herren. Ich habe den Watson ganz einfach gebeten, er soll mir eine sehr kurze Zusammenfassung dieses Tages formulieren. Es kam folgendes heraus: tolle Konferenz mit spannendem Thema, sehr gute Referenten und sehr aufmerksames Publikum. Ich muss sagen, dass das sehr bestellt klingt... Ist aber sehr zutreffend!

Meine Damen und Herren, ich fasse nicht zusammen, aber ich meine, wir haben ein breites Spektrum gesehen und gehört, in der Tat von sehr guten Referenten aus den unterschiedlichsten Bereichen, und dafür möchte ich ganz herzlich danken. Ein Satz ist mir besonders in Erinnerung geblieben, der vielleicht doch zu erwähnen ist: „Man braucht einen wissenden Menschen". Ich denke, es ist ganz tröstlich dass wir sowohl für die Generierung dieser Algorithmen als auch für die Interpretation der Ergebnisse doch noch wissende Menschen brauchen!

Ich möchte damit schließen und vor allem dem Programmausschuss danken unter der Leitung von Herrn Wohlmuth und vielen anderen Experten, nicht nur von IBM, sondern von den anderen beteiligten Unternehmen und Institutionen. Dann natürlich unserem Backoffice draußen für die perfekte Organisation. Und nicht zuletzt Dank an Sie alle für Ihr Kommen und auch die Mitwirkung in der Diskussion. Ich möchte noch darauf hinweisen, dass die Slides in den nächsten Tagen von der Homepage des MÜNCHNER KREISES heruntergeladen werden können. Es wird auch wieder einen Tagungsband in Form eines Buches geben, was aber ein paar Monate dauert.

In Ihren Unterlagen haben Sie eine Vorankündigung. Am 10. Oktober veranstalten wir im M.O.C. München im Rahmen der Communications World die Konferenz „Personal Communications – Wie soziale Netzwerke und neue Technologien die interpersonelle Kommunikation revolutionieren", mit dem Schwerpunkt Geschäftskommunikation. Und dann am 22. November „Smart Business Networks", wo wir diese relativ neue Form der Wertschöpfungsnetzwerke behandeln, die im Internet immer stärkere Bedeutung haben. Ich hoffe, dass wir uns da wiedersehen!

Damit schließe ich die Konferenz. Auf Wiedersehen!

Liste der Referenten und Moderatoren

Dr. Andreas Abecker
disy Informationssysteme GmbH
Leiter Innovationsmanagement
Erbprinzenstr. 4
76133 Karlsruhe
andreas.abecker@disy.net

Axel Deicke
Ex Vice President Service Engineering
BMW Group
Hamsterweg 9
85598 Baldham
axel.deicke@gmail.com

Dr. Alexander Duisberg
Bird & Bird LLP
IT Commercial
Pacellistr. 14
80333 München
alexander.duisberg@twobirds.com

Prof. Dr.-Ing. Jörg Eberspächer
Technische Universität München
Lehrstuhl für Kommunikationsnetze
Arcisstr. 21
80333 München
joerg.eberspaecher@tum.de

Thomas Hampp
IBM Deutschland R&D GmbH
Senior Technical Staff Member
Schönaicher Str. 220
71032 Böblingen
thomas.hampp@de.ibm.com

Christian Klezl
IBM
Vice President Corporate Strategy
Bld CHQ1, Office 2B-17N-1
One New Orchard Road
Armonk, NY 10504-1722, USA
klezl@us.ibm.com

Florian Kuhlmann
Neofonie GmbH
Senior Project Manager R&D
Robert-Koch-Platz 4
10115 Berlin
florian.kuhlmann@neofonie.de

Prof. Dr. Volker Markl
Technische Universität Berlin
Fakultät IV
Juister Weg 6
14199 Berlin
sekr@dima.tu-berlin.de

Prof. Dr. Dres. h.c. Arnold Picot
Ludwig-Maximilians-Universität
Institut für Information, Organisation
und Management
Ludwigstr. 28
80539 München
picot@lmu.de

Dr. Wolf v. Reden
Fraunhofer Institut für Nachrichtentechnik
HHI
Einsteinufer 37
10587 Berlin
wolf.von.reden@hhi.fraunhofer.de

Dr. Volker Rieger
Detecon International GmbH
Managing Partner
Oberkasseler Str. 2
53227 Bonn
volker.rieger@detecon.com

Dr. Achim Steinacker
intelligent views GmbH
Presales Manager
Julius-Reiber-Str. 17
64293 Darmstadt
a.steinacker@i-views.de

Prof. Dr. Rudi Studer
KIT Campus Süd
Institut AIFB, Geb. 11.40
Kaiserstr. 12
76131 Karlsruhe
studer@kit.edu

Prof. Dr. Volker Tresp
Siemens AG
CT IC 4
Otto-Hahn-Ring 6
81739 München
volker.tresp@siemens.com

Dr. Otto Wohlmuth
IBM Deutschland R&D GmbH
Manager IBM Systems & Technology Group
Schönaicher Str. 220
71032 Böblingen
wohlmuth@de.ibm.com

ISBN 978-3-9813733-7-0