Positionspapier zu Big Data in der öffentlichen Verwaltung **(Stand 06.06.2016)**		**White Paper** **Big Data – 1.0.0** **Ergebnis der PG**
Kurzbeschreibung	Der exponentielle Anstieg der weltweit von Mensch und Maschine erzeugten und gespeicherten Daten bringt für die öffentliche Verwaltung neue Herausforderungen aber auch Chancen mit sich. Das Positionspapier soll Grundlageninformationen für strategische Entscheidungen bereitstellen. Besonderes Augenmerk wird auf die strukturellen, rechtlichen, wirtschaftlichen und technischen Aspekte von Big Data im Verwaltungsumfeld gelegt. Das Papier wird durch einen Best Practice Abschnitt ergänzt, um einige konkrete Beispiele für erfolgreich umgesetzte Big Data Projekte aus Verwaltung und Wirtschaft zu zeigen. Das Positionspapier bringt den gemeinsamen Standpunkt der öffentlichen Verwaltung zum Thema Big Data zum Ausdruck. https://www.ref.gv.at/Big-Data.3364.0.html	
Autoren / Koordinierung:	Harald Pirker Daniel Medimorec	Projektteam / Arbeitsgruppe PG-Big Data
Beiträge von:	DI Thomas Burg, Mag. Gerhard Embacher-Köhle, Mag. Martin Hackl, Mag.a Cathrin Kiss, Dr. Alexander Kowarik, DI Michael Mörz, Mag.a Lisbeth Mosnik, BSc., Mag. Wolfgang Schlapschy, Mag. Gregor Schmied	

Version 1.0.0: **06.06.2016**

Inhaltsverzeichnis

Abbildungsverzeichnis

Big Data

Management Summary

Big Data umfasst sämtliche Bemühungen, vorwiegend aber den Einsatz moderner Informations- und Kommunikationstechnologien (IKT), zur Gewinnung von **Erkenntnissen** aus zumeist sehr großen, sich schnell ändernden und unterschiedlich strukturierten Daten. Die betrachteten Daten werden zum Teil durch Menschen (herkömmliche Datenverarbeitung bzw. auch unbewusst über Social Media Plattformen, etc.) jedoch immer öfter durch „intelligente" Gegenstände selbst und automatisiert (mittels Sensoren) gesammelt. Die aus den Analysen der Daten gewonnenen Erkenntnisse sollen dabei helfen, Dinge besser zu verstehen und die richtigen **Entscheidungen** (Betrugsbekämpfung, Verkehrssteuerung, Arbeitsmarktpolitik, Gesundheitsweisen, Krankheitsbekämpfung, etc.) möglichst rasch zu treffen. Wenn dies gelingt, spricht man von einem Mehrwert durch Big Data.

Dementsprechend ist es nicht richtig zu sagen, dass die Digitalisierung und damit verbunden auch Big Data nichts anderes bedeuten, als den Einsatz neuer Werkzeuge, um bestehende Tätigkeiten schneller oder effizienter abwickeln zu können. IT-Expertinnen und Experten – unter Ihnen auch Dr. Viktor Mayer-Schönberger[1] (österreichischer Rechtswissenschaftler und Big Data Experte am Oxford Internet Institute) – kritisieren, dass dieser durchaus verbreiteten Einschätzung ein wesentlicher Aspekt fehlt: die neue **Sichtweise**, in der wir die Welt in der wir leben betrachten und wahrnehmen. Schönberger sieht Big Data als Sprung, welcher weniger technisch als gedanklich ist und damit verbunden ein besseres **Verständnis** über die Wirklichkeit, welches auf Unmengen an zugrunde liegenden Daten und empirischen Analysen beruht. Dadurch können wir **Muster und Zusammenhänge** schneller **erkennen** und besser in unsere Planungen einfließen lassen.

Der öffentlichen Verwaltung kommt im Hinblick auf Big Data eine besondere Verantwortung zu, da sie sich im Antagonismus zwischen Daten schützen und Daten nutzen (zum Zwecke moderner E-Services und möglichst hoher Transparenz) bewegt. Mit dem Einsatz moderner Analyse- und Korrelationsmodelle, wie sie bei Big Data Verwendung finden, erhöht sich grundsätzlich auch die Wahrscheinlichkeit, dass dabei personenbezogene Daten verwendet werden oder durch die Vermengung der unterschiedlichen Daten ein Personenbezug entsteht. Die strenge **Einhaltung der datenschutzrechtlichen Vorgaben** muss deshalb in sämtlichen Phasen eines Big Data Projekts oberste Priorität haben. Dies kann durch das Anwenden geeigneter Mechanismen zur Anonymisierung bzw. Pseudonymisierung von personenbezogenen Daten sowie das Ergreifen ausreichender Datensicherheitsmaßnahmen (Schutz vor Datendiebstahl, Missbrauch oder Manipulation) geschehen. Diese Maßnahmen bilden zentrale Elemente für Big Data und stehen keinesfalls im Widerspruch dazu. So können – wie es auch die Praxis bereits gezeigt hat – auch bei strenger

[1] https://de.wikipedia.org/wiki/Viktor_Mayer-Sch%C3%B6nberger_(Jurist)

Einhaltung der Datenschutzanforderungen **zahlreiche Einsatzmöglichkeiten** für Big Data Ansätze in der öffentlichen Verwaltung identifiziert werden. Dementsprechend sollten Behörden prüfen, wo es Einsatzmöglichkeiten für Big Data Ansätze gibt und diese unter Einhaltung der datenschutzrechtlichen Vorgaben nutzen.

Die für Big Data Analysen notwendigen Wissensgebiete werden in der Praxis meist unter dem Berufsbild „**Data Scientist**" zusammengefasst und umfassen zahlreiche Bereiche wie z. B. analytische Mathematik, Statistik, Wissensmanagement oder Kommunikationsmanagement. Aufgrund der Komplexität von Big Data Vorhaben benötigt es im Regelfall jedoch **interdisziplinäre Teams** zur Lösung der damit verbundenen Aufgabenstellungen. Es wird eine politische Grundsatzentscheidung notwendig sein, ob fehlendes Know How (Stichwort „Data Science") künftig mittels entsprechender Aus- und Fortbildungsprogramme (z. B. über die Verwaltungsakademie des Bundes) angeboten und aufgebaut werden soll, oder man den Weg verstärkter **Kooperationen innerhalb der öffentlichen Verwaltung sowie mit Wissenschaft, Forschung und Wirtschaft** geht und das Know-How im Bedarfsfall zukauft.

Während statistische und mathematische Verfahren relativ stabil bleiben, ändern sich die technischen Rahmenbedingungen permanent. Viele Big Data Technologien (z. B. Hadoop) sind erst kürzlich entstanden und waren somit nicht Teil der Ausbildung älterer IT-Kräfte, was gerade für diesen Bereich **nachhaltige Konzepte** notwendig macht.

Dennoch sind die Herausforderungen, vor denen die öffentliche Verwaltung im Zusammenhang mit Big Data steht weniger technischer, als viel mehr, wie bereits zuvor geschildert, organisatorisch – struktureller Natur. In der Privatwirtschaft ist Big Data bereits seit einiger Zeit ein Geschäftsmodell und der IKT-Markt bietet eine **Reihe von erprobten technischen Lösungen** an.

In der öffentlichen Verwaltung besteht ein breites Spektrum an Handlungsfeldern und Anwendungsgebieten dieser technischen Lösungen. Eine McKinsey Studie zu **Effizienzsteigerungen durch Big Data in der Verwaltung** spricht etwa von einem Sparpotenzial von bis zu 20 Prozent in der öffentlichen Verwaltung. Für ganz Europa wären dies insgesamt bis zu 300 Milliarden Euro.[2] Effizienzsteigerungen sind z. B. im Bereich der **Bearbeitung von BürgerInnenanfragen** zu erzielen. Die eingehenden Anfragen ähneln in der Praxis oft sehr. Wenn Behörden nun eine Frage immer wieder neu bearbeiten und beantworten müssen, würde dies einen sehr hohen Zeitaufwand bedeuten und es würden sich möglicherweise andere Antworten auf ein und dieselbe Frage ergeben. Durch die Verknüpfung unterschiedlicher Kommunikationskanäle und den Zugriff auf verschiedene Datenquellen könnten diese **Anfragen** jedoch **konsistenter, qualitativ hochwertiger und effizienter** bearbeitet werden. Im besten Fall könnten aus dem wachsenden Pool an vorhandenem Wissen Antwortvorschläge sogar automatisch erstellt werden[3]. Dies zeigt wie Big Data künftig einen wesentlichen Beitrag zum Abbau von Routinetätigkeiten leisten und gleichzeitig freie Ressourcen für primäre Tätigkeiten schaffen kann.

[2] Vgl. [KIN11], o. S.
[3] Vgl. [BRZ15], Seite 18

Ein besonderer Innovationsschub kann durch die Verknüpfung von Big Data mit Open (Government) Data zu **Big Open (Government) Data** erreicht werden. Auf der einen Seite steht Big Data, also große, komplexe und raschen Updatezyklen unterworfen Daten, die zur Analyse und dem besseren Verstehen komplexer Probleme herangezogen werden können. Auf der anderen Seite steht Open (Government) Data, welches Daten offen, verfügbar und wiederverwendbar zur Verfügung stellt und eine wichtigste Voraussetzung für datengetriebene Entwicklungen darstellt. Die Verbindung dieser beiden Entwicklungen ermöglicht nun den Zugang und das Arbeiten mit großen (öffentlichen) Datensets und kann als **Innovationsmotor** sowohl für die **Wirtschaft** als auch für die **öffentliche Verwaltung** gesehen werden.

Die sich durch Big Data ergebenden positiven **wirtschaftlichen Aspekte** können auch am Beispiel des **Arbeitsmarktes** erläutert werden. Gerade in diesem Bereich können Big Data-Analysen behilflich sein. Schon jetzt erstellt das Arbeitsmarktservice Prognosen über die Notwendigkeit verschiedener Arbeitskräfte in den bevorstehenden Jahren. Unter Einbeziehung umfangreicherer Datenbasen zu Bildung, Immigration und Bevölkerungsentwicklung, könnten zukünftig noch präzisere arbeitsmarktpolitische Maßnahmen gesetzt und ein Optimum zwischen benötigten und verfügbaren Arbeitskräften erreicht werden.

Um das notwendige Vertrauen der BürgerInnen in Big Data Verfahren und Technologien zu gewährleisten, sollte darauf geachtet werden, dass die Verfahren möglichst transparent gestaltet werden. Auch die **Integrität der Ergebnisse** muss langfristig sichergestellt werden, um das Vertrauen aller zu stärken. Deshalb muss darauf geachtet werden, dass zum einen der richtige Fokus gelegt und die Daten richtig interpretiert werden und zum anderen die hinter den Analysen liegenden Datenmodelle und Algorithmen geprüft sowie die Ergebnisse durch Menschen verifiziert werden.

Klar ist auch, dass Big Data nicht auf alle Fragen Antworten liefern muss. In der Regel sind Big Data Vorhaben mit zum Teil beträchtlichen **Planungs- und Investitionskosten** verbunden. Daher ist es notwendig die Kosten-Nutzenrelation vor der tatsächlichen Umsetzung eines Projekts in jedem Fall vorab zu betrachten.

(1) Begriffsdefinition

„The value of big data lies in our ability to extract insights and make better decisions.“[4]

Big Data umfasst demnach sämtliche Bemühungen, vorwiegend aber den Einsatz moderner Informations- und Kommunikationstechnologien (IKT), zur Gewinnung von Erkenntnissen aus zumeist sehr großen, sich schnell ändernden und unterschiedlich strukturierten Daten. Die betrachteten Daten werden zum Teil durch Menschen (herkömmliche Datenverarbeitung bzw. auch unbewusst über Social Media Plattformen, etc.) jedoch immer öfter durch „intelligente“ Gegenstände selbst und automatisiert gesammelt (mittels Sensoren). Die aus der Analyse der Daten gewonnenen Erkenntnisse sollen dabei im Hinblick auf die gesetzten Ziele einen Mehrwert ergeben.

Ebenso wie das zugrunde liegende Datenmaterial unterscheidet sich der gesetzte zeitliche Fokus von Big Data Projekten. So können sich Big Data Analysen auf die Bewertung des Erfolges von bereits gesetzten Maßnahmen beziehen, oder dabei helfen, aktuelle Herausforderungen und Probleme durch Nutzung bestimmter Echtzeitdaten zu lösen, sowie auch bei der Bewältigung künftiger Aufgaben von großem Nutzen sein, wenn es darum geht, unterschiedliche Maßnahmen in Form von Simulationen durchzuspielen, um die bestmögliche Alternative zu wählen.

Auch wenn Rappa in seiner Definition die Essenz der meisten Big Data Vorhaben recht gut beschreibt, muss seine Einschätzung um einen wesentlichen Aspekt ergänzt werden. Ein umfassender Big Data Ansatz hat sich zwangsläufig auch mit der Frage der Datenarchivierung auseinanderzusetzen und mögliche Wege aufzuzeigen, die Relevanz von organisationseigenen Daten einzustufen sowie die erforderlichen Daten kostengünstig und platzsparend (komprimiert) zu speichern. Beim Einsatz von Komprimierung darf der Zugriff auf die gespeicherten Daten jedenfalls nicht beeinträchtigt werden.

Der Begriff „Big Data“ beschreibt keine neue Technologie, da es auch in der Vergangenheit bereits unzählige digitale Datenbestände gegeben hat. Vielmehr umfasst er die Bemühungen, aus einer rapide zunehmenden Menge an Daten das notwendige Wissen für effizientere Prozesse, bessere Entscheidungen und bürgerfreundlichere Services zu generieren. Verstärkt wird das exponentielle Datenwachstum durch Entwicklungen wie das „Internet of Things“, also die maschinengestützte Erfassung von Daten, welche mit dem rasch wachsenden Markt an sensorgesteuerten Produkten einhergeht. Es ist unbestritten, dass bereits in unmittelbarer Zukunft beinahe jeder Lebensbereich durch die rasanten Entwicklungen in der digitalen Welt verändert werden wird (Quelle: Gartner 2014).

[4] Dr. Michael Rappa Director of the Institute for Advanced Analytics and Distinguished University Professor North Carolina State University, http://analytics.ncsu.edu/?p=4770

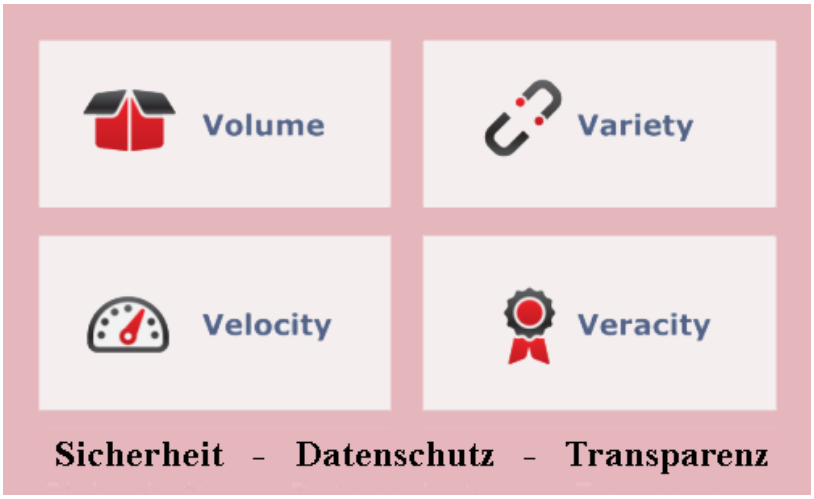
(1.1) Charakteristika

Bei der Planung und Umsetzung von Big Data Projekten gilt es seitens der öffentlichen Verwaltung u.a. folgende Planungsdimensionen mit zu berücksichtigen.

Datenquellen: Die herkömmliche Datenverarbeitung als wesentlichster „Datenproduzent“ wurde schon längst durch Social Media Plattformen, das Internet of Things (IoT) und den Einzug der Sensorik in sämtliche Lebensbereiche ersetzt. Im Bereich der öffentlichen Verwaltung sind es v.a. die Themen Transparenz und Open Government Data, die aus Big Data Sicht neue Potentiale eröffnen.

Dateneigenschaften: Wie bereits in der Begriffsdefinition kurz erwähnt, ist die Komplexität von Big Data Vorhaben nicht nur durch die sukzessive Zunahme potentieller Datenquellen, sondern auch durch die große Vielfalt an unterschiedlichen Datenformaten (**Variety**), den enormen Anstieg der Datenmengen (**Volume**) sowie immer kürzer werdende Produktionszeiten und „Lebensdauern“ der Daten (**Velocity**), in der eine Auswertung einen Mehrwert erzeugt, gekennzeichnet. Von besonderer Bedeutung für Big Data Vorhaben des öffentlichen Sektors ist die Richtigkeit der zu Grunde liegenden Daten (**Veracity**) sowie die Korrektheit der daraus gewonnenen Schlüsse. Traditionell wurden elektronische Datenbestände vor allem in Tabellenform organisiert. Diese „strukturierten“ Daten fügten sich durch die Art ihrer Anordnung also in eine Struktur, welche ihnen bereits eine gewisse Bedeutung gab. Der Großteil, bezogen auf die in Anspruch genommene Speicherkapazität, der mittlerweile zur Verfügung stehenden Daten ist jedoch nicht strukturiert, sondern unstrukturiert. Dies betrifft etwa Textinformationen wie E-Mails, PDF-Dokumente, Daten aus sozialen Netzwerken (Facebook, Twitter, etc.), Audiodateien sowie Bilder und Videos. Sicherheit von IT-Systemen, Datenschutz und Transparenz zählen zwar nicht zu den eigentlichen Charakteristika von Big Data, bilden jedoch den notwendigen Rahmen vor denen Big Data Vorhaben immer betrachtet werden müssen.

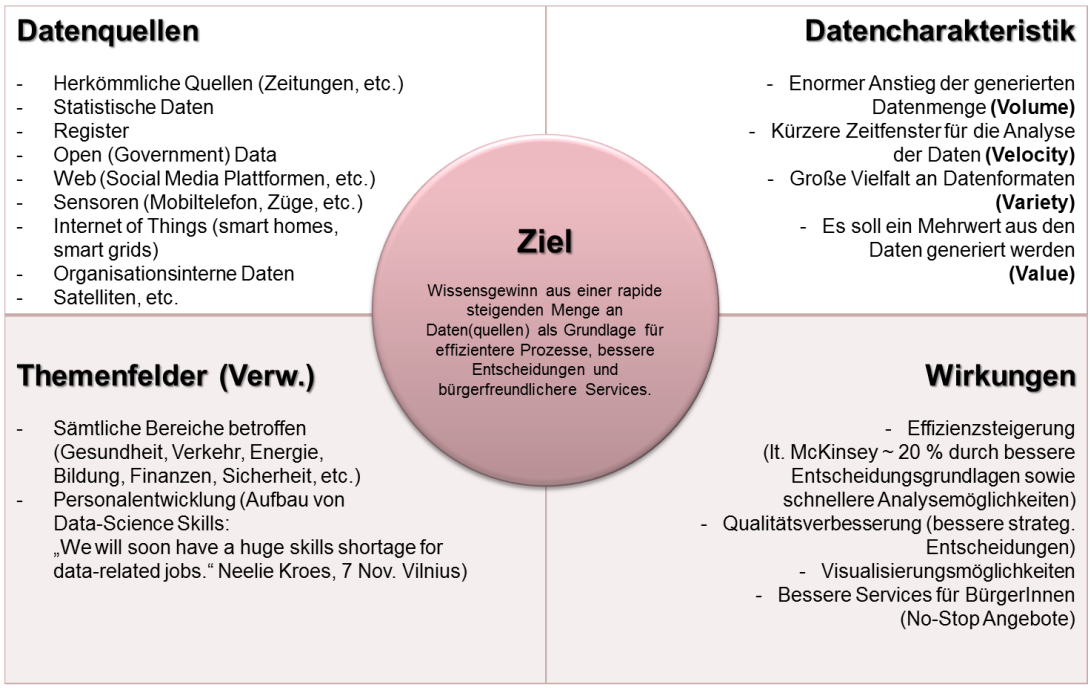
Abbildung 1: Charakteristika von Big Data



Quelle: BRZ, [BRZ15], Seite 7

Themenfelder: Was die möglichen Einsatzfelder betrifft, ist grundsätzlich kein Bereich vom Einsatz moderner Datenanalysetechniken (Big Data) ausgeschlossen. So gibt es neben den „klassischen“ Big Data Bereichen wie Kriminalitäts- und Terrorismusbekämpfung, Gesundheit oder Verkehrsplanung zahlreiche weitere potentielle Einsatzfelder.

Abbildung 2: Charakteristika von Big Data

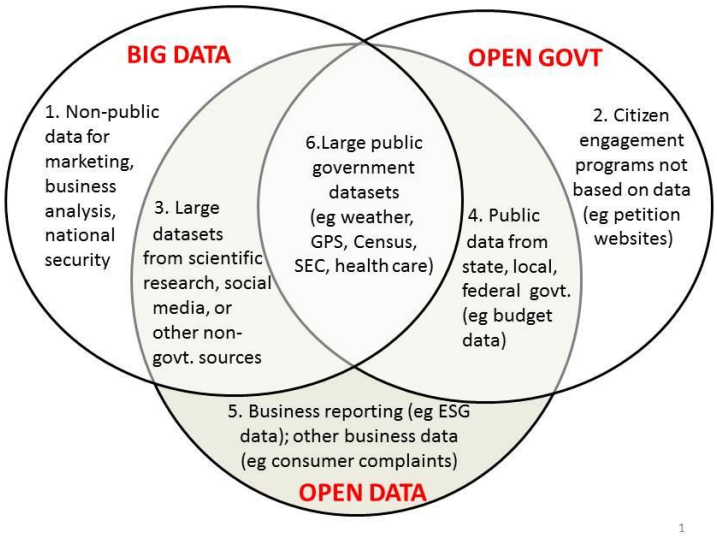


Quelle: eigene Darstellung

(1.2) Abgrenzung zu Open (Government) Data

Nachfolgende Abbildung zeigt im Überblick die wichtigsten Schnittmengen und Unterschiede der Bereich Open Data, Open Government Data und Big Data.

Abbildung 3 Unterschiede und Schnittmengen zwischen Big - Open - Open Gov. Data



Quelle: Gurin, [GUR14], o. S.

(1.3) Anwendungsfelder

Die potentiellen Big Data Anwendungsfelder lassen sich sehr gut anhand ihrer **zeitlichen Perspektive** ableiten. So erscheint eine Strukturierung nach folgenden Clustern sinnvoll:

- rückblickend (**Data Analyses**): beispielsweise zur Bewertung von gesetzten Maßnahmen wie Informationskampagnen. Dieser Bereich weist die deutlichsten Überschneidungen zur herkömmlichen Datenverarbeitung auf und beschäftigt sich vorwiegend mit den Fragen „Was ist passiert?“ bzw. „Warum ist etwas passiert?“. Moderne Analysemethoden bilden gemeinsam mit unterschiedlichsten Visualisierungstechniken die Basis für effiziente Auswertungen, fundierte Entscheidungen und ein professionelles Berichtswesen.

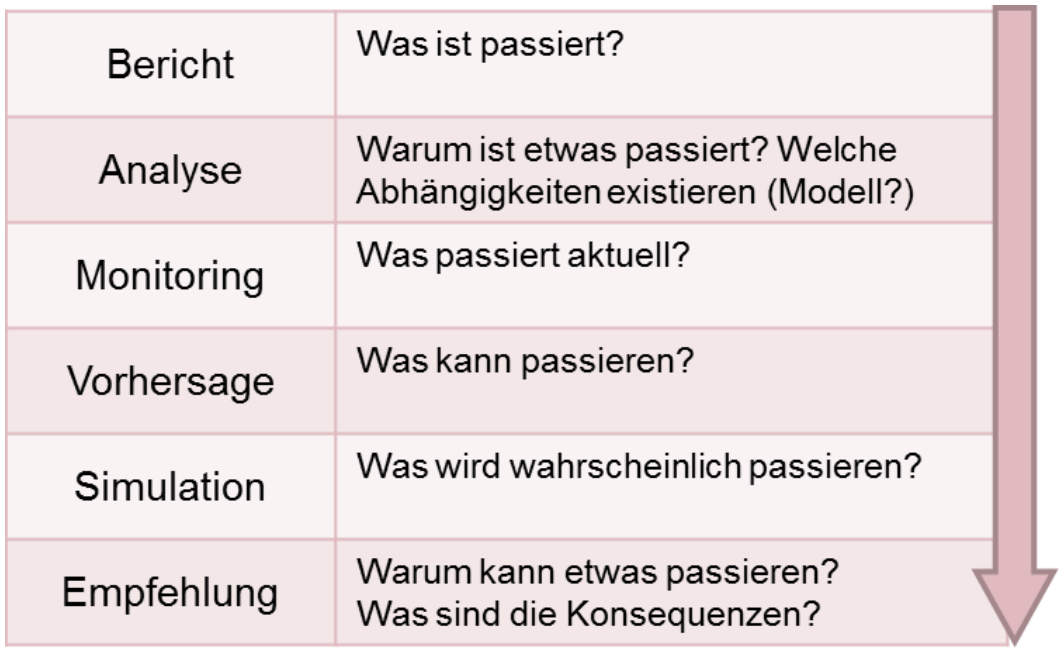
- gegenwartsbezogen (**Service Delivery / Data Compression**): Zahlreiche Monitoring-Systeme nutzen bereits Echtzeitdaten, um beispielsweise im Bereich „intelligenter“ Verkehrssteuerungssysteme oder zur Identifikation von Anomalien (z. B. IT-Sicherheit) bestmögliche und v.a. zeitnahe Lösungen zu liefern sowie rasch auf Gefahren reagieren zu können. Die Kernfragen in diesem Bereich lauten „Was passiert aktuell?“ bzw. „Was kann aufgrund der aktuellen Situation passieren“?

Ein Thema, dessen Bedeutung vor allem im öffentlichen Sektor weiter zunehmen wird, ist die Komprimierung von Daten zur Senkung von Speicherkosten. Zu diesem Zweck können und müssen relevante Daten bereits im Vorfeld auf ihre Eignung zur Lösung einer Problemstellung geprüft und von sogenanntem „Datenmüll“ getrennt werden.

- zukunftsorientiert (**Policy Development**): vorausblickende Auswertungen und Analysen – auch „Predictive Analytics“ genannt – umfassen laut Wikipedia „eine Vielzahl statistischer Techniken wie Modellierung, maschinelles Lernen und Datamining, die gegenwärtige und historische Fakten analysieren, um Voraussagen über die Zukunft oder unbekannte Ereignisse zu ermöglichen“. Dies erlaubt es, z. B. durch Simulation von geplanten Maßnahmen (z. B. Steuerung von Förderungen, Gesetzgebung, Arbeitsmarkt, etc.), bessere Entscheidungen zu treffen und das Risiko von Fehlentscheidungen bzw. -entwicklungen zu minimieren. Im Mittelpunkt stehen die Überlegungen „Was wird/kann wahrscheinlich passieren?“ bzw. „Was sind die Konsequenzen?“.

Ein ganzheitlicher Big Data Ansatz hat sämtliche Phasen der Datenauswertung – von der vergangenheitsorientierten Betrachtung der Daten als Grundlage für Berichte bis zur Simulation von Modellen und daraus abgeleiteten Empfehlungen – zu umfassen, um die damit verbundenen Chancen zu nutzen.

Abbildung 4: Phasen der Datenauswertung



Quelle: Fraunhofer Fokus, [FRAF14], Seite 8

(1.4) Ausgewählte Anwendungsbereiche im Detail

(1.4.1) Predictive Analytics

Predictive Analytics und Advanced Analytics sind Begriffe, welche üblicherweise synonym verwendet werden. Der Bereich der „Predictive Analytics“ bedient sich einer Reihe von Methoden, um Erkenntnisse aus Daten zu gewinnen. Diese Methoden greifen vor allem auf statistische und mathematische Methoden zurück. Beliebt sind auch sogenannte machine-learning Ansätze, welche eher computationaler Natur sind. Ein „Predictive Analytics-Ablauf“ besteht aus mehreren Komponenten welche nachfolgend kurz erläutert werden.

Datenaufbereitung (Data Preparation):

Der erste wichtige Aspekt, welcher auch von einer Predictive Analytics Software abgedeckt werden kann ist die Datenaufbereitung (Data Preparation). Predictive Analytics kann nicht auf Rohdaten angewendet werden. Predictive Analytics Applikationen und Methoden benötigen Daten in bestimmten Strukturen. Eine Ausnahme diesbezüglich würde die Analyse unstrukturierter Daten stellen. Hierbei werden Textdokumente aufbereitet und analysiert und aus den Analysen Erkenntnisse gezogen. Ansonsten werden im Rahmen von Data Preparation ähnliche Tätigkeiten durchgeführt wie in klassischen ETL (**Extract, Transform, Load)** Prozessen, bei denen Daten aus mehreren gegebenenfalls unterschiedlich strukturierten Datenquellen in einer Zieldatenbank vereinigt werden. Ein Teil der für Predictive Analytics notwendigen Datenaufbereitung wird meistens in den entsprechenden Tools durchgeführt, jedoch können beträchtliche Aspekte in den vorgelagerten Datenbanken abgehandelt werden.

Data Understanding:

Bevor statistische Analysen durchgeführt werden ist es wichtig die zugrundeliegenden Daten besser zu verstehen. Hierfür werden meist univariate[5] Analysen, Tabellen und Grafiken angewendet. Bezüglich der Verfügbarkeit solcher Features und der Interaktivität gibt es große Unterschiede zwischen den verschiedenen Predictive Analytics Tools.

Klassische Statistik:

Klassische Statistik bezeichnet jenen Bereich in welchem anhand von Stichproben mithilfe statistischer Methoden Aussagen über zugrundeliegende Populationen gemacht werden. Viele Predictive Analytics Methoden leiten sich aus diesen Ansätzen ab oder bauen darauf auf.

Data Modeling:

Data Modeling stellt die Kernkompetenz des Predictive Analytics dar. In weiterer Folge sollen deshalb die unterschiedlichen Bereiche, welche von den meisten Predictive Analytics Tools abgedeckt sind, kurz beschrieben werden.

Unsupervised Learning:

Unter unsupervised learning versteht man Verfahren zur Identifizierung von Ähnlichkeiten in Datenstrukturen. Diese Art von Analyse wird oft auch als

[5] In der Mathematik bezeichnet univariat eine Gleichung, einen Ausdruck oder eine Funktion, die jeweils nur von einer Variablen abhängen.

Clusteranalyse bezeichnet. Es geht um ein exploratives analysieren der Daten, bei dem Cluster/Gruppen/Segmente identifiziert werden sollen, welche voneinander möglichst verschieden sind. Die Mitglieder eines Clusters sollen einander aber möglichst ähnlich sein. Grob kann zwischen hierarchischen Clusteranalysen und nicht-hierarchischen Ansätzen unterschieden werden. Die einzelnen Predictive Analytics Tools unterscheiden sich in Bezug auf die Verfügbarkeit dieser Art von Algorithmen.

Supervised Learning:

Im Rahmen des supervised learning gibt es immer eine Zielvariable, welche „erklärt“ werden soll. Beispielsweise der Unterschied zwischen Männern und Frauen in Bezug auf deren Einstellung zu einem Produkt oder Unterschiede zwischen Tiergattungen in Bezug auf deren Verhalten. Wenn etwa zukünftiges Verhalten anhand historischer Informationen vorhergesagt werden soll, dann wird dieser Bereich auch als Predictive Analytics bezeichnet. Die entwickelten Modelle können beispielsweise dabei helfen, besonders stornogefährdete KundInnen zu identifizieren, oder potenzielle Steuersünder, Kriminalitätshotspots oder Erdbeben vorherzusagen. Es gibt zahlreiche statistische und machine-learning Methoden, welche hier zur Anwendung kommen. Dazu zählen unter anderem Lineare Regression, Logistische Regression, Entscheidungsbäume (Decision Tree), Neuronale Netze, uvm. Die einzelnen Predictive Analytics Tools unterscheiden sich in Bezug auf die Verfügbarkeit dieser Art von Algorithmen.

Evaluation & Deployment:

Um erstellte Vorhersage und Klassifikations- Algorithmen periodisch anwenden zu können, müssen diese auf die zugrundeliegenden Daten angewendet werden. Dieser Prozess wird Deployment genannt und kann im Predictive Analytics Tool sowie auch in einer Datenbank erfolgen. Wenn etwa ein Algorithmus zur Erkennung von Steuersündern entwickelt wurde, muss dieser täglich, wöchentlich, in jedem Fall aber regelmäßig angewendet werden, um die wahrscheinlichsten Fälle zu identifizieren. Im Zuge dessen wird der verwendete Algorithmus dann auch auf seine Validität überprüft. Dieser Prozess der Evaluation wird oft auch als **„Backtesting“** bezeichnet. Diese Funktionalitäten bedürfen üblicherweise auch einiger Datenaufbereitungsschritte, welche idealerweise im Predictive Analytics Tool abgehandelt werden können.

(1.4.2) Visualisierung

Wie schon im Abschnitt Predictive Analytics erwähnt, besteht ein essentieller Teil von Big Data in der Datenaufbereitung. Seit den 1980er Jahren ist Data Warehousing bekannt und bezieht sich auf die strukturierte Aufbereitung von Daten. Hierfür gibt es zahlreiche ETL (Extraction Transformation Loading) Werkzeuge. Im Rahmen von Big Data wurden diese Prozesse sowohl technisch als auch logisch optimiert. Häufig können diese Prozesse von der eigentlichen Programmierung abstrahiert und über eine visuelle Oberfläche abgehandelt werden. In den letzten Jahren wurden einige kommerzielle und open Source Tools entwickelt, welche speziell für den Aspekt der Visualisierung multivariater Zusammenhänge geeignet sind. Häufig werden die Daten hierbei im Hauptspeicher verarbeitet, was zu schnelleren Reaktionszeiten führt. Die Daten können hierbei in real time bearbeitet und visualisiert werden.

(2) Strukturelle Aspekte

Es wäre falsch zu glauben, dass die Digitalisierung und damit verbunden auch Big Data nichts anderes bedeuten, als den Einsatz neuer Werkzeuge, um bestehende Tätigkeiten schneller oder effizienter abwickeln zu können. Expertinnen und Experten kritisieren deshalb zu Recht, dass dieser durchaus verbreiteten Einschätzung ein wesentlicher Aspekt fehlt: die neue Sichtweise, in der wir die Welt in der wir leben betrachten und wahrnehmen. Experten – unter ihnen auch Dr. Viktor Mayer-Schönberger vom Oxford Internet Institute – sprechen deshalb von Big Data als Sprung, welcher weniger technisch als gedanklich und dementsprechend auch organisatorisch ist. Mit diesem Sprung verbunden ist ein besseres Verständnis über die Wirklichkeit, welches auf Unmengen an zugrunde liegenden Daten und empirischen Analysen beruht und damit die Chance immens verbessert, die richtigen Entscheidungen in sämtlichen Lebensbereichen (Gesundheit, Verkehr, Arbeit, etc.) zu treffen.

(2.1) Ausgangslage

Neben der Euphorie über die vielen Möglichkeiten besteht auch Skepsis gegenüber dem Einsatz von Big Data-Technologien durch die öffentliche Verwaltung. Big Data ist ein relativ junges Phänomen, welches vor allem im organisatorischen Bereich einen bestimmten Grad an Flexibilität voraussetzt. Beispielsweise wird Software heutzutage meist in Form eines „Scrum-Ansatzes“ entwickelt. Das bedeutet, dass Projekte in Einzelschritte zerlegt werden und die „Richtung“ nach Abschluss jedes Meilensteins angepasst werden kann.

Big Data wird aller Voraussicht nach auch im öffentlichen Bereich dazu führen, dass bestimmte Ablaufprozesse restrukturiert werden müssen, neues Fachwissen in den Bereichen Wissensmanagement bzw. „data science“ aufgebaut oder zugekauft werden muss und neue Berufsbilder (z. B. data scientist) und Formen der Zusammenarbeit (interdisziplinäre Teams) etabliert werden müssen. In diesem Kontext werden organisatorische und technologische Kooperationsmodelle in der öffentlichen Verwaltung künftig weiter an Bedeutung gewinnen.

Der Staat hat angesichts der weitreichenden Möglichkeiten der Datennutzung eine besondere Verantwortung. Dabei geht es einerseits um Gefahren und Risiken für das einzelne Individuum sowie gesellschaftliche Minderheiten, etwa durch Überwachung, Datendiebstahl oder Datenmanipulation. Andererseits geht es auch um die Gefahren für die Gesellschaft als Ganzes, um Fragen der Menschenrechte und der Freiheitsgrundsätze.[6] Es gilt bereits vor dem Start eines Big Data Projekts zu klären, welche Verantwortlichkeiten und Prozesse zentral, welche dezentral und welche situationsspezifisch – also in unterschiedlichen Zusammensetzungen – gesteuert werden sollen bzw. in wie weit die Nutzbarmachung von Datenquellen in der jeweiligen Organisation verankert werden kann und soll.

Eine entscheidende Rolle bei der erfolgreichen Etablierung von Big Data kommt der IT zu. Es müssen geeignete Formen (Plattformen, Foren, etc.) definiert bzw. ausgebaut werden, über welche die Kommunikation zwischen Fachbereichen und IT-ExpertInnen in einer Weise vonstattengeht, die der jeweiligen Zielsetzung angepasste Lösungen ermöglicht.

[6] Vgl. [BRZ15], Seite 14

(2.2) Chancen/Risiken

„Je mehr Daten sie haben, desto dümmer werden sie“[7]. Die umfangreichsten Daten und besten Technologien nutzen nichts, wenn die falschen Daten betrachtet, der falsche Fokus gelegt oder die falschen Prioritäten gesetzt werden. Dies gilt natürlich nicht nur bei weltpolitischen Problemen, sondern auch im Kleinen.[8] Die diesbezüglichen Herausforderungen von Big Data sind weniger die technischen. Die Nutzung großer Datenmengen (Volume), die Unterschiedlichkeit der Daten (Variety) und der Echtzeitzugriff (Velocity) sind technisch lösbar und in verschiedenen Szenarien längst realisiert. Die Herausforderungen liegen aus struktureller Sicht stärker bei den Rahmenbedingungen, wie etwa im Setzen der richtigen Prioritäten und der Vermeidung von Missbrauch oder Fehlinterpretationen von Daten sowie im Aufbau der erforderlichen Kompetenzen für die korrekte Interpretation der Ergebnisse.[9]

Die Integrität der durch Big Data generierten Antworten sowie die Qualität und Vollständigkeit der Informationen sind primär durch die Algorithmen und deren Umsetzung bedingt. Die Auswirkung für die von einer „Analyse“ betroffene Person oder Sache hängt wiederum primär davon ab, in welcher Umgebung und zu welchem Zweck die Anwendung betrieben wird, bzw. welches Vertrauen die AnwenderInnen in die Antwort setzen. Mit anderen Worten stellt sich die Frage, inwiefern den maschinengenerierten Antworten als Ergebnis einer „Big Data Analyse“ Autorität verliehen wird und ob die Ergebnisse allenfalls noch in Zweifel gezogen werden müssen. Es handelt sich hierbei um Fragen die soziologische, ethische und letztlich wie immer auch rechtliche Dimension von Big Data betreffend.[10]

Auch wenn der richtige Fokus und die richtigen Prioritäten gesetzt werden, so ist nicht garantiert, dass aus Big Data Analysen die richtigen Erkenntnisse abgeleitet werden. Die Technologien beruhen zumeist auf komplexen Algorithmen und Korrelationsanalysen. Dies wirft verschiedenste Fragen auf, die offensichtlichste davon wohl zum Umgang mit Ironie, Parodie oder Satire. Die wenigsten Algorithmen können diese Ausdrucksformen erkennen, obwohl diese gerade in sozialen Medien häufig vorkommen. Doch auch ohne Ironie können Daten falsch interpretiert werden. Dies kann mit schwerwiegenden Folgen verbunden sein, etwa einem falschen Verdacht. Eine weitere zentrale Herausforderung beim Einsatz von Big Data Technologien ist deshalb die Unterscheidung von Korrelation und Kausalität.[11]

„Die Bedeutung der Unterscheidung von Korrelation und Kausalität lässt sich anhand eines Beispiels erläutern. Es wurde in einer Untersuchung in Niedersachsen von 1970 bis 1985 festgestellt, dass sowohl die Zahl der Störche als auch jene der Neugeborenen sank. Ein eindeutige Korrelation, ein kausaler Zusammenhang zwischen Störchen und der Geburt von Kindern lässt sich medizinisch dennoch nicht nachhaltig beweisen.[12]“

[7] William Binney, ehemaliger technischer Direktor der NSA
[8] Vgl. [BRZ15], Seite 15
[9] Vgl. [BRZ15], Seite 14
[10] Vgl. [BEST16], Seite 76
[11] Vgl. [BRZ15], Seite 15
[12] Vgl. [SIE14], o. S.

Wenn die Komplexität von Algorithmen zunimmt, nimmt auch die Komplexität der Erkenntnisse zu. Besondere Skepsis besteht in der Öffentlichkeit daher hinsichtlich des Einsatzes von Predictive Analytics durch die öffentliche Verwaltung. Eine Nutzung von Algorithmen zur Vorhersage des Verhaltens einzelner Individuen ist äußerst heikel. Denn eine Korrelationsanalyse kann immer nur einen Hinweis auf einen Sachverhalt liefern. Big Data ermöglicht lediglich ein mehr oder weniger exaktes Abbild der Wirklichkeit. Die nähere Untersuchung des Sachverhalts obliegt dem Menschen, der die Ergebnisse interpretiert. Es muss also stets eine Verifikation der Erkenntnisse erfolgen. Ebenfalls zu beachten ist, dass Algorithmen, die auf einer bestimmten Datenbasis valide Ergebnisse liefern, bei veränderter Datenlage nicht mehr passend sein könnten. Eine wiederkehrende Verifikation von Algorithmen sowie auch der Datenquellen ist also notwendig. Eine Forderung, die in diesem Zusammenhang immer wieder zu hören ist, zielt darauf ab, dass die den Analysen zugrundeliegenden Algorithmen (sowohl bei Unternehmen als auch bei der öffentlichen Verwaltung) offen gelegt werden sollten, um eine öffentliche Verifikation (sozusagen „Open Source“) zu ermöglichen.[13]

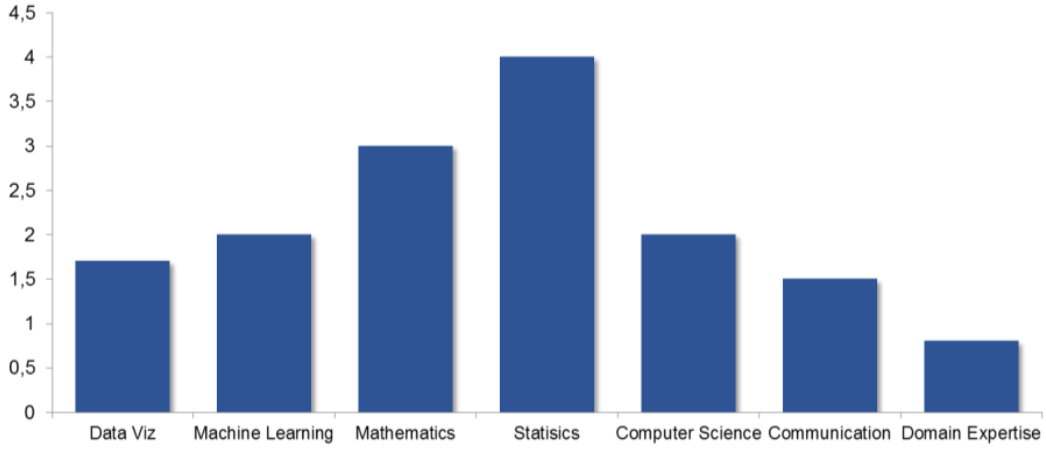
Ein entscheidender Vorteil von Big Data liegt in der flexiblen Vernetzung von Datenquellen und der Anwendung von neuen Technologien, um unbekannte Muster zu identifizieren. Damit verbunden sind auch Auswirkungen auf Standards und bestehende Prozesse. Hadoop Cluster etwa, welche für die parallele Berechnung von Prozessen verwendet werden, machen nur dann Sinn, wenn Berechnungsressourcen flexibel zur Verfügung gestellt werden können. Dazu müsste im öffentlichen Sektor eine strategische Entscheidung zum Einsatz von Cloud-Technologien für unterschiedliche Anwendungsfälle getroffen werden.

[13] Vgl. [BRZ15], Seite 16

(2.3) Organisatorische Aspekte

Entscheidend dafür wie effektiv das Datenpotential genutzt wird sind die Skills der MitarbeiterInnen. Für Big Data Analysen sind verschiedenste Wissensgebiete notwendig, die meist unter dem Berufsbild „Data Scientist“ zusammengefasst werden. Data Scientists wirken als „Übersetzer“ zwischen den Fachanforderungen und den für die Umsetzung erforderlichen analytischen Ansätzen und Technologien. Die unterschiedlichen Anforderungen an solche ExpertInnen sind in nachfolgender Abbildung überblicksmäßig dargestellt.

Abbildung 5: Anforderungen an Data Scientists



Quelle: [SCHU14], Seite 11

Es ist ersichtlich, dass die Anforderungen derart vielschichtig sind, dass diese kaum von einzelnen Personen erfüllt werden können, sondern interdisziplinäre Teams notwendig machen. So sollten FachexpertInnen und Informatik-SpezialistInnen intensiv mit DatenexpertInnen zusammenarbeiten, die das notwendige Know-How in den Bereichen Mathematik und Statistik aufweisen. Für die öffentliche Verwaltung sowie auch deren IT-Dienstleister wird es notwendig, Personal mit den geforderten Skills aufzubauen, anzuwerben oder einzukaufen, um das enorme Potenzial von Big Data auszuschöpfen.[14]

Während statistische und mathematische Verfahren relativ stabil bleiben, ändern sich die technischen Rahmenbedingungen permanent. Viele Big Data Technologien (z. B. Hadoop) sind erst kürzlich entstanden und waren somit nicht Teil der Ausbildung älterer IT-Kräfte. Von großer Bedeutung ist deshalb ein stetiger Austausch in Form von Kooperationen zwischen Wissenschaft, Wirtschaft und Verwaltung.

[14] Vgl. [BRZ15], Seite 41

(2.4) Unterschiedliche Datenquellen

Um einen Eindruck über die Vielfalt der potenziellen Datenquellen zu erhalten sind nachfolgend exemplarisch einige davon aufgezählt. Sofern es sich um personenbezogene Daten handelt, unterliegen diese besonderen datenschutzrechtlichen Einschränkungen (siehe Kapitel „Rechtliche Aspekte“) in ihrer Verwendung.

- Öffentlich zugängliche Daten
  - Zeitungen, Publikationen, Webseiten, Soziale Medien, Blogs, etc.
  - Open Government Data: data.gv.at, publicdata.eu, etc.
  - Statistische Daten: Eurostat, Statistik Austria, etc.
  - Finanzdaten: Weltbank, Europäische Zentralbank, Internationaler Währungsfond, OECD, etc.
  - Gesundheitsdaten: netdoktor.at, PatientsLikeMe.com
- Geschützte Daten
  - Protokolle von Telekommunikationsverbindungen (CDR)
  - Web-Zugriffe (Logdateien)
  - Intranet, Geschützte Social Media-Profile
  - Automatische Erfassungen von RFID-Lesern
  - Kameras, Mikrofone, sonstige Sensoren (Autos, Handys, etc.)
  - Verbrauchsdaten im Energiesektor
  - Wissenschaft: Geologie, Genetik, Klimaforschung und Kernphysik
  - Gesundheitswesen
- Verwaltungsdaten
  - Register: Zentrales Melderegister, Personenstandsregister, Gewerberegister, Firmenbuch, Grundbuch, Waffenregister, etc.
  - Verwaltungsanwendungen
  - Organisatorische Daten (z. B. Kennzahlen)
- Sonstige Datenquellen[15]

[15] Vgl. [BRZ15], Seite 9

(2.5) Empfehlungen für die öffentliche Verwaltung

Neben der Euphorie über die vielen Möglichkeiten besteht auch Skepsis gegenüber dem Einsatz von Big Data-Technologien durch die öffentliche Verwaltung. Big Data ist ein relativ junges Phänomen, welches vor allem **im organisatorischen Bereich** einen bestimmten Grad an **Flexibilität** voraussetzt. Um die mit Big Data verbundenen Möglichkeiten ausschöpfen zu können, ist es notwendig, bestimmte **Ablaufprozesse** (von der Prioritätensetzung bis zur technischen Umsetzung) **und Kooperationsmodelle** zu **restrukturieren**. Gleichzeitig muss das notwendige Fachwissen auch innerhalb der Verwaltung durch Weiterbildungsangebote oder die Aufnahme fachkundiger MitarbeiterInnen aufgebaut oder im Bedarfsfall zugekauft werden.

Sowohl für die Identifikation von potentiellen Anwendungsbereichen als auch die konkrete Umsetzung von Big Data Vorhaben erscheint die **temporäre Installation von interdisziplinären Teams** als besonders geeignet. Dies könnte mit einer stärkeren und permanenten Kooperation mit Wissenschaft, Wirtschaft und Forschung synergetisch verschränkt werden. Eine Plattform „Smart Data Austria" – ähnlich der Cooperation OGD – könnte zum einen für den Big Data Bereich die Funktion einer Informationsdrehscheibe übernehmen und zum anderen versuchen, Big Data als „Ganzes" – vom Bildungsangebot bis zur Umsetzung in kleineren Organisationseinheiten der heimischen Verwaltung – in Österreich voranzutreiben.

Neben den neuen Berufsbildern (z. B. data scientist) und Formen der Zusammenarbeit (interdisziplinäre Teams) auch innerhalb der öffentlichen Verwaltung sollten **organisatorische und technologische Kooperationsmodelle** in der öffentlichen Verwaltung künftig weiter an Bedeutung gewinnen.

(3) Rechtliche Aspekte

Ein wichtiger Aspekt von Big Data liegt in der Verknüpfung von unterschiedlichen Daten. Hier muss im jeweiligen Anwendungsfall entschieden werden, ob die antizipierte Gesamtsicht aus dem Blickwinkel des Datenschutzes erlaubt ist. Im Rahmen von Big Data werden unter Umständen auch zusätzliche und neue Daten generiert. Der Schwerpunkt liegt jedoch in der Erfassung, Verarbeitung und intelligenten Analyse von bestehenden Informationen. Sofern neue Daten generiert werden ist jedenfalls zu prüfen, in welchem rechtlichen Zusammenhang die Datenspeicherung, Analyse und Bereitstellung von Daten erfolgen darf.

(3.1) Ausgangslage

Während die Menschen bei der Nutzung von privaten Internet-Angeboten relativ sorglos mit Ihren privaten Daten umgehen, zeigt sich eine ungerechtfertigt hohe Skepsis was die Verwaltung als Halter von (personenbezogenen) Daten betrifft. Die Verwaltung agiert im Spannungsfeld „Daten schützen - Daten nutzen". Big Data darf deshalb nicht nur vor dem Hintergrund wirtschaftlicher oder technischer Aspekte diskutiert werden, sondern muss rechtliche und ethische Aspekte von Beginn an einbeziehen. Um Auswertungen durchführen zu können, welche unterschiedliche Datenquellen und womöglich personenbezogene Daten beinhalten, muss sichergestellt werden, dass die BürgerInnen durch entsprechende Mechanismen (z. B. Anonymisierung bzw. Pseudonymisierung der Daten) davor geschützt werden, zum oft zitierten „gläsernen Bürger" zu werden.

Mit dem Einsatz von Big-Data-Technologien erhöht sich die Wahrscheinlichkeit, dass personenbezogene Daten verarbeitet werden. Der Einsatz explorativer Verfahren zur Erkennung von Korrelationen (z. B. Alter, Geschlecht, Einkommen zu Kreditwürdigkeit) in einer Menge auf den ersten Blick unabhängiger Daten kann sich daher negativ auf den Einzelnen auswirken. Auch der Rückschluss von Statistiken auf Individuen erlaubt potenziell diskriminierende Eigenschafts-zuschreibungen, die mit der Person nichts zu tun haben.

Für Datensubjekte wird es daher immer schwieriger einzuschätzen, ob Daten wirklich anonym sind und also solche veröffentlicht bzw. verarbeitet werden dürfen, oder ob diese Daten durch „Big Data Methoden", also insbesondere die zielgerichtete Verknüpfung mit anderen großen Datenbeständen, doch wieder auf einzelne Personen rückführbar werden, auch wenn die ursprünglichen Referenzdaten isoliert betrachtet als anonym einzustufen sind.[16]

(3.2) Datenschutz

Jedermann hat Anspruch auf Geheimhaltung der ihn betreffenden personenbezogenen Daten, soweit ein schutzwürdiges Interesse daran besteht. Dieser Grundsatz ist in § 1 Datenschutzgesetz 2000 (DSG 2000) als **verfassungsrechtlich garantiertes Grundrecht** verankert und ist nicht nur gegenüber dem Staat, sondern **gegenüber jedem** durchsetzbar. Die Bestimmungen des DSG 2000 regeln die Verwendung personenbezogener Daten natürlicher und

[16] Vgl. [BEST16], o. S.

juristischer Personen. Werden keine personenbezogenen Daten verwendet, so kommen die Bestimmungen des DSG 2000 nicht zur Anwendung.[17]

(3.2.1) Personenbezogene Daten und Zweckbindung

Ob eine geplante Datenverwendung zulässig ist oder nicht, gehört zu den am häufigsten gestellten datenschutzrechtlichen Fragen. **Gemäß § 6 Abs. 1 Z 2 DSG 2000 dürfen personenbezogene Daten grundsätzlich nur für festgelegte, eindeutige und rechtmäßige Zwecke ermittelt und nicht in einer für diese Zwecke unvereinbaren Weise weiterverwendet werden**. Da viele Big Data-Anwendungen die Sammlung und Auswertung von Daten aus verschiedenen Quellen und die Nutzung für neue Zwecke zum Ziel haben, ist dieser Aspekt besonders zu beachten. [18]

Zweck und Inhalt einer Datenverwendung müssen gemäß § 7 Abs. 1 DSG 2000 grundsätzlich **von der Berechtigung des Auftraggebers umfasst sein** (diese liegt je nach gesetzlicher Zuständigkeit des Auftraggebers vor oder ist von seinen rechtlichen Befugnissen abhängig) **und dürfen die schutzwürdigen Geheimhaltungsinteressen der Betroffenen nicht verletzen**. Während sich die rechtlichen Befugnisse von Unternehmen aus dem Gesellschaftsvertrag oder der Gewerbeberechtigung ableiten lassen, ergeben sich die gesetzlichen Zuständigkeiten der öffentlichen Verwaltung aus den einschlägigen Gesetzen und Verordnungen. Die Weiterverwendung von Daten für wissenschaftliche oder statistische Zwecke ist unter den Voraussetzungen der §§ 46 und 47 DSG 2000 zulässig.[19]

(3.2.2) Schutzwürdiges Geheimhaltungsinteresse

Ein schutzwürdiges Interesse an der Geheimhaltung von personenbezogenen Daten **liegt grundsätzlich immer** vor, außer die personenbezogenen Daten wurden **zulässigerweise veröffentlicht** oder gelten **mangels Rückführbarkeit als anonym**. Darüber hinaus regeln §§ 8 und 9 DSG 2000 weitere Fälle hinsichtlich nicht-sensibler und sensibler Daten in denen schutzwürdige Geheimhaltungsinteressen nicht verletzt werden.[20]

(3.2.3) Gelindestes Mittel und datenschutzrechtliche Grundsätze

Wurde festgestellt, dass die Datenverwendung grundsätzlich zulässig ist, da der Auftraggeber berechtigt ist die zu prüfende Datenverwendungen durchzuführen und schutzwürdige Geheimhaltungsinteressen nicht verletzt werden, so muss nunmehr gemäß § 7 Abs. 3 DSG 2000 geprüft werden, ob der **Eingriff in das Recht auf Datenschutz nur im erforderlichen Ausmaß und mit den gelindesten zur Verfügung stehenden Mitteln** erfolgt und die nachfolgend beschriebenen datenschutzrechtlichen Grundsätze gemäß § 6 DSG 2000 eingehalten werden:

[17] Vgl. [BRZ15], Seite 30
[18] Vgl. [Feiler/Fina13], o. S.
[19] Vgl. [BRZ15], Seite 31
[20] Vgl. [BRZ15], Seite 31

- Gemäß Z 1 dürfen Daten nur **nach Treu und Glauben** und auf rechtmäßige Weise verwendet werden. Dies liegt insbesondere dann vor, wenn die Betroffenen über die Umstände des Datengebrauchs und das Bestehen und die Durchsetzbarkeit ihrer Rechte informiert wurden.
- Gemäß Z 2 dürfen Daten nur für **festgelegte, eindeutige und rechtmäßige Zwecke ermittelt** und nicht in einer für diese Zwecke unvereinbaren Weise weiterverwendet werden. Die Weiterverwendung für wissenschaftliche oder statistische Zwecke ist unter gewissen Voraussetzungen zulässig (§§ 46 und 47 DSG 2000).
- Gemäß Z 3 dürfen **Daten nur soweit sie für den Zweck der Datenverarbeitung wesentlich sind**, verwendet werden und über diesen Zweck nicht hinausgehen.
- Gemäß Z 4 dürfen Daten nur so verwendet werden, dass sie im Hinblick auf den Verwendungszweck im Ergebnis **sachlich richtig** und wenn nötig auf den neuesten Stand gebracht sind.
- Gemäß Z 5 dürfen Daten nur solange in personenbezogener Form aufbewahrt werden, als dies für die Erreichung der Zwecke, für die sie ermittelt wurden, erforderlich ist. Eine längere **Aufbewahrungsdauer** kann sich aus besonderen gesetzlichen, insbesondere archivrechtlichen Vorschriften ergeben.[21]

(3.2.4) Bereichsabgrenzung

Nicht im Datenschutzgesetz selbst geregelt und doch datenschutzrechtlich relevant für öffentliche Auftraggeber ist die Bereichsabgrenzung, die in § 9 E-Government-Gesetz (E-GovG) festgelegt ist und im bereichsspezifischen Personenkennzeichen (bPK) seine Ausgestaltung findet. **Das E-Government-Gesetz legt fest, dass die Identifikationsfunktion des bereichsspezifischen Personenkennzeichens (bPK) auf jenen staatlichen Tätigkeitsbereich beschränkt ist, dem die Datenanwendung zuzurechnen ist, in der das bPK verwendet werden soll**. Sie bewirkt eine Trennung zwischen den staatlichen Aufgabenbereichen, wodurch bereichsübergreifende Auswertungen ohne weitere Voraussetzungen (Kommunikation im Wege verschlüsselter bPK in Fällen, in denen die bereichsübergreifende Kommunikation rechtlich vorgesehen ist) ausgeschlossen sind: Gemäß § 9 Abs. 2 erster Satz E-GovG ist die Abgrenzung der staatlichen Tätigkeitsbereiche für Zwecke der Bildung von bPK so vorzunehmen, dass zusammengehörige Lebenssachverhalte in ein- und demselben Bereich zusammengefasst werden und miteinander unvereinbare Datenverwendungen (§ 6 Abs. 1 Z 2 DSG 2000) innerhalb desselben Bereichs nicht vorgesehen sind. Details zu der Abgrenzung der staatlichen Tätigkeitsbereiche sind in der **Bereichsabgrenzungsverordnung**[22] definiert.[23]

[21] Vgl. [BRZ15], Seite 33 f.
[22] https://www.digitales.oesterreich.gv.at/e-government-bereichsabgrenzungsverordnung
[23] Vgl. [BRZ15], Seite 34

(3.2.5) Anonymisierung und Pseudonymisierung

Eine für viele Anwendungsszenarien relevante Möglichkeit der Datenverwendung ist jene der Anonymisierung oder Pseudonymisierung. Das DSG 2000 unterscheidet zwischen „personenbezogenen“ und „indirekt personenbezogenen“ Daten. „Indirekt personenbezogen" sind Daten dann, wenn der Personenbezug der Daten derart ist, dass der Auftraggeber die Identität der Betroffenen/des Betroffenen mit rechtlich zulässigen Mitteln nicht bestimmen kann. **Nicht personenbezogene (anonymisierte) Daten dürfen unter keinen Umständen – auch dann nicht, wenn verschiedene Datensets kombiniert werden – auf eine Person rückführbar sein und unterliegen nicht dem Anwendungsbereich des DSG 2000**. Dies wird angesichts der vielfältigen Big Data Technologien und Einsatzszenarien eine immer größere Herausforderung.

Der deutsche Bundesverband für Informationswirtschaft, Telekommunikation und neue Medien (BITKOM) hat notwendige Anforderungen an eine **Anonymisierung** (auf das deutsche Datenschutzgesetz bezogen, doch weitgehend auf die Situation in Österreich anwendbar) wie folgt zusammengefasst[24]: „Ursprünglich personenbezogene Daten können dadurch anonymisiert werden, dass Identifikationsmerkmale gelöscht oder bestimmte Merkmale aggregiert werden. Aggregation von Merkmalen heißt, dass exakte Angaben durch allgemeinere ersetzt und die Daten dann zusammengefasst werden: Beispielsweise eine Gruppenbildung anhand des Geburtsjahres anstelle des genauen Geburtsdatums oder anhand einer weiträumigen Gebietsangabe anstelle der Adressangabe. Eine Anonymisierung kann auch dadurch vorgenommen werden, dass aus einem Bestand personenbezogener Daten einzelne Angaben ohne Personenbezug herausgefiltert werden.“ Hinsichtlich der **Pseudonymisierung** fasst BITKOM die Ansätze wie folgt zusammen: „**Bei der Anonymisierung werden Identifikationsmerkmale gelöscht, bei der Pseudonymisierung nur ersetzt**. […] Bei der Pseudonymisierung gibt es grundsätzlich zwei Verfahrensarten: Erstens die Erzeugung von Zufallswerten und deren Zuordnung zum Betroffenen mittels einer Referenzliste. Für die Inhaberin bzw. den Inhaber dieser Liste bleiben die Daten dann personenbezogen. Zweitens die Erstellung von Pseudonymen durch Hash-Verfahren mit geheimen Parametern. Wenn die Rückrechnung der ursprünglichen Daten mit sehr hohem Aufwand verbunden ist, spricht man auch von einer Einweg-Pseudonymisierung.“

Auch auf die erwähnte Gefahr der Kombination von Datensets geht BITKOM ein: „Bei der **Weitergabe der Daten an interessierte Dritte** wird teilweise die Auffassung vertreten, dass hier eine **Anonymisierung gar nicht möglich** sei, da die Drittunternehmen eventuell über eigene Datenbestände verfügten, die es dem Drittunternehmen erlaubten, die betroffene Person zu identifizieren. Diesbezüglich ist anzumerken, dass die Verhältnismäßigkeit des Aufwands zur De-Anonymisierung nur anhand der Umstände des Einzelfalls und nicht pauschal beurteilt werden kann. Wenn Fallkonstellationen denkbar sind, in denen die Zusammenführung von Datenbeständen zur Wiederherstellung eines Personenbezugs ausreicht, bedeutet dies nicht, dass die Anonymisierung generell unmöglich wäre.“ [25]

[24] Vgl. [BIT13], Seite 26
[25] Vgl. [BRZ15], Seite 35

(3.2.6) Automatisierung von Entscheidungen

Eines der potenziellen Anwendungsfelder von Big Data Anwendungen ist die Automatisierung von Prozessen und damit potenziell auch von Entscheidungen. Damit einhergehende Gefahren (beispielsweise durch die Verwechslung von Korrelation und Kausalität) tritt das Datenschutzgesetz in § 49 Abs. 1 DSG 2000 entgegen: **„Niemand darf einer für ihn rechtliche Folgen nach sich ziehenden oder einer ihn erheblich beeinträchtigenden Entscheidung unterworfen werden, die ausschließlich auf Grund einer automationsunterstützten Verarbeitung von Daten zum Zweck der Bewertung einzelner Aspekte seiner Person ergeht, wie beispielsweise seiner beruflichen Leistungsfähigkeit, seiner Kreditwürdigkeit, seiner Zuverlässigkeit oder seines Verhaltens.“**[26] Ausnahmen sind vorgesehen, wenn eine ausschließlich automatisiert erzeugte Entscheidung gesetzlich ausdrücklich vorgesehen ist, die Entscheidung im Rahmen des Abschlusses oder der Erfüllung eines Vertrages ergeht und dem Ersuchen des Betroffenen auf Abschluss oder Erfüllung des Vertrages stattgegeben wurde oder die Wahrung der berechtigten Interessen des Betroffenen durch geeignete Maßnahmen – beispielsweise die Möglichkeit, seinen Standpunkt geltend zu machen – garantiert wird.[27]

(3.2.7) Datensicherheit

Das Datenschutzgesetz regelt in § 14 DSG 2000 auch Datensicherheitsmaßnahmen und gibt diesbezügliche vor: „Dabei ist je nach der Art der verwendeten Daten und nach Umfang und Zweck der Verwendung sowie unter Bedachtnahme auf den Stand der technischen Möglichkeiten und auf die wirtschaftliche Vertretbarkeit sicherzustellen, dass die **Daten vor zufälliger oder unrechtmäßiger Zerstörung und vor Verlust geschützt** sind, dass ihre **Verwendung ordnungsgemäß** erfolgt **und** dass die **Daten Unbefugten nicht zugänglich** sind.“ Die Methode, in Bezug auf Big Data für Datensicherheit zu sorgen, unterscheidet sich datenschutzrechtlich grundsätzlich nicht von „herkömmlichen“ Anwendungen. Wie bei allen Anwendungen muss auch in Big Data Projekten eine **Schutzbedarfsfeststellung** durchgeführt und der Schutzbedarf in der Lösung berücksichtigt werden. **Vertraulichkeit, Verfügbarkeit und Integrität müssen entsprechend § 14** DSG 2000 sichergestellt werden.[28]

Ein weiterer zu beachtender Aspekt betrifft Anwendungen, in denen Daten unterschiedlicher Quellen zusammengeführt werden: „In einer Big Data-**Anwendung werden typischerweise personenbezogene Daten zusammengeführt, die vorher** in separaten IT-Systemen gespeichert wurden. Jedes dieser separaten Systeme verfügte idealerweise über Sicherheits-vorkehrungen, die den mit den jeweiligen Datenkategorien verbundenen Risiken und damit den Anforderungen des § 14 DSG 2000 entsprachen. Soll durch die Implementierung der Big Data-Anwendung das Sicherheitsniveau nicht gesenkt werden, so muss die Big Data-Anwendung jedenfalls jene Sicherheit bieten, die das sicherste der zuvor verwendeten IT-Systeme bot. **Das vorherige Maximum wird bei Big Data so zum neuen Minimum.“**[29] Weiters ist eine Risikoerhöhung

[26] Vgl. [BRZ15], Seite 35 f.
[27] Vgl. [BRZ15], Seite 35 f.
[28] Vgl. [BRZ15], Seite 36
[29] Vgl. [Feiler/Fina13], o. S.

aufgrund der konzentrierten Sammlung von Daten zu beachten. Besonders zu berücksichtigen ist darüber hinaus auch die Frage, wer bei organisationsübergreifenden Verwendung von Daten für den Gesamtdatenbestand verantwortlich ist.[30]

(3.2.8) Transparenz (Nachvollziehbarkeit der Datenverarbeitung)

Das Konzept der informationellen Selbstbestimmung erfordert die Kontrollierbarkeit der Datenverarbeitung. Die Verarbeitung personenbezogener Daten ist heute in der Regel jedoch nicht unmittelbar durch die Betroffenen kontrollierbar. Sie findet auf Servern „im Verborgenen" statt. Eine rechtswidrige Verwendung personenbezogener Daten ist daher oftmals den Betroffenen gar nicht bekannt.[31]

Angesichts der Gefahren und Risiken für die Privatsphäre sowie der Skepsis gegenüber Big Data-Anawendungen in der Bevölkerung ist daher folgender Aspekt bedeutend: das ebenfalls im Datenschutzgesetz festgelegte **Recht auf Auskunft sowie auf Richtigstellung und Löschung von Daten**. In § 26 DSG 2000 ist hinsichtlich des Auskunftsrechts festgelegt, dass ein Auftraggeber jeder Person oder Personengemeinschaft, die dies schriftlich verlangt und ihre Identität in geeigneter Form nachweist, Auskunft über die zu dieser Person oder Personengemeinschaft verarbeiteten Daten zu geben hat. Außerdem hat gemäß § 27 DSG 2000 jeder Auftraggeber unrichtige oder entgegen den Bestimmungen des DSG 2000 verarbeitete Daten richtigzustellen oder zu löschen, und zwar aus eigenem Antrieb, sobald ihm die Unrichtigkeit von Daten oder die Unzulässigkeit ihrer Verarbeitung bekannt geworden ist, oder auf begründeten Antrag des Betroffenen. Die sich aus diesen Vorgaben ergebenden **Pflichten** sind natürlich auch in Big Data Projekten zu beachten.[32]

In der Praxis fällt oft auf, dass sowohl Auskunftsrecht als auch das Recht auf **Richtigstellung und Löschung** bei den BürgerInnen oft nicht ausreichend bekannt sind. Sie könnten aber **Basis für ein größeres Vertrauen seitens der Bevölkerung gegenüber datenorientierten Anwendungen der öffentlichen Verwaltung** sein und damit auch die Bereitschaft zur Nutzung dieser Anwendungen erhöhen. Denkbar wäre etwa ein möglichst flächendeckend eingesetztes Shared IT Service, das verfahrensübergreifend einfache Einsicht in gespeicherte Daten und Zugriffe sowie die Möglichkeit bietet, die Löschung der gespeicherten Daten zu beantragen.[33]

[30] Vgl. [BRZ15], Seite 36
[31] Vgl. [BEST16], Seite 123
[32] Vgl. [BRZ15], Seite 37
[33] Vgl. [BRZ15], Seite 37

(3.2.9) Big Data vor dem Hintergrund der EU-Datenschutz-Grundverordnung

Die EU-Datenschutz-Grundverordnung (EU-DSGVO) wurde am 14. April 2016 vom Europäischen Parlament angenommen nachdem der Rat am 08. April 2016 seinen Standpunkt in erster Lesung festgelegt hatte. Sie ist am 04. Mai 2016 im Amtsblatt der Europäisches Union veröffentlicht worden und tritt damit am 24. Mai 2016 in Kraft. Anwendbar ist sie ab dem 25. Mai 2018. Die EU-DSGVO wird die Datenschutz-Richtlinie 95/46/EG (DSRL) ersetzen. **Ziel der EU-Datenschutz-Grundverordnung ist es, die datenschutzrechtlichen Vorschriften zu aktualisieren und zu modernisieren**.

Das Datenschutzpaket umfasst zwei Bereiche:

- eine Datenschutz-Grundverordnung
- eine Richtlinie über den Schutz personenbezogener Daten bei der Verarbeitung zum Zwecke der Strafverfolgung

Die **Datenschutz-Grundverordnung regelt die Rechte natürlicher Personen sowie die Pflichten derjenigen, die die Daten verarbeiten bzw. für die Verarbeitung der Daten verantwortlich sind**. Ferner ist darin festgelegt, wie die Einhaltung der Vorschriften gewährleistet werden soll und welche Sanktionen bei Verstößen gegen die Vorschriften zu verhängen sind.

Mit Anwendungsbeginn der neuen europäischen Datenschutzregeln wären die oben getroffenen datenschutzrechtlichen Erwägungen zu der Verwendung von Big Data neu zu bewerten und gegebenenfalls an die geänderte Rechtslage anzupassen.

(3.3) Empfehlungen für die öffentliche Verwaltung

Mit dem vermehrten Einsatz moderner Technik, insbesondere aber bei der Umsetzung von Big Data Vorhaben, erhöht sich die Wahrscheinlichkeit, dass dabei personenbezogene Daten verwendet werden oder entstehen.

Die **strenge Einhaltung der datenschutzrechtlichen Vorgaben** darf deshalb in keiner Phase eines Big Data Projekts außer Acht gelassen werden und sollte immer oberste Priorität haben. Dazu zählen insbesondere das Anwenden geeigneter **Mechanismen zur Anonymisierung bzw. Pseudonymisierung** von Daten sowie das Ergreifen ausreichender **Datensicherheitsmaßnahmen** zum Schutz vor Datendiebstahl, Missbrauch oder Manipulation.

Um das notwendige Vertrauen der BürgerInnen in Big Data Verfahren und Technologien zu gewährleisten, ist darauf zu achten, diese möglichst transparent zu gestalten.

Ein Risiko birgt auch das Speichern der Daten mit sich, wenn diese nicht lokal sondern in der **Cloud**[34] gespeichert werden. Diese virtuellen Speicher liegen oft im Ausland, was unter Umständen Auswirkungen auf die Datensicherheit und den Datenschutz haben kann.

Die öffentliche Verwaltung hat eine besondere Verantwortung und muss mit dieser Verantwortung gewissenhaft umgehen. Das bedeutet darauf zu achten, dass der

[34] Vgl. [Cloud 12], o. S.

richtige Fokus gelegt und die Daten richtig interpretiert werden sowie auch, dass die hinter den Analysen liegenden Modelle und Datenquellen immer wieder geprüft und die Ergebnisse durch Menschen verifiziert werden.

(4) Wirtschaftliche Aspekte

Alleine im Jahr 2011 wurden mehr Daten erzeugt als in der Geschichte der Menschheit[35]. Nur ein kleiner Teil dieser Daten (15%) liegt strukturiert und verwertbar vor[36]. Derzeit besteht Zugriff auf nur einen sehr kleinen Teil der potentiell nützlichen Daten. Ein weiterer nennenswerter ökonomischer Effekt ist der zu erwartende Innovationschub, der sich durch die Verknüpfung von Big Data und Cloud Services ergeben wird. Entsprechend der Vorhersage einer International Data Corporation (IDC) Studie sollen bis 2020 die Ausgaben in Unternehmen für cloudbasierte Big Data Analysetechnologien viereinhalb Mal schneller wachsen, als Ausgaben für In-house Lösungen[37].

Für Gesellschaft und Wirtschaft ergeben sich Vorteile, wenn diese Daten nutzbar und zugänglich gemacht werden könnten: Neue Dienste können generiert, verbesserte Produkte angeboten, Arbeitsabläufe effizienter gestaltet werden. Internationale Studien zeigen, dass Unternehmen durch datengetriebene Innovationen einen Wettbewerbsvorsprung erzielten und ihre Produktivität erheblich steigern konnten. Es ergeben sich neue Perspektiven und Nutzen für IKT-Unternehmen, aber auch Datenproduzenten und NutzerInnen aller anderen privaten und öffentlichen Sektoren. IDC[38] prognostiziert 430 Milliarden Produktivitätsanstieg weltweit bis 2020 für Organisationen, die auf Datenanalyse setzen und handlungsfähige Informationen liefern, im Vergleich zu weniger analytisch orientierten Organisationen.

(4.1) Ausgangslage

Datengetriebene Innovationen bieten große Chancen insbesondere für Klein- und Mittelunternehmen. Ein besonderes Charakteristikum von Daten ist, dass sie nicht ortsgebunden sind und weltweit allen Unternehmen und Organisationen zur Verfügung gestellt werden können.

Im Vergleich zu den USA hat die digitale Wirtschaft in Europa die Datenrevolution schleppend aufgenommen, weil in Europa Zutrittsschranken für KMUs entstehen und Innovation behindert werden. Es mangelt an Datenfachleuten, die technische Fortschritte in konkrete Geschäftsmöglichkeiten umsetzen können, an entsprechender Grundinfrastruktur, große Datensätze sind unzureichend zugänglich und rechtliche Rahmenbedingungen (z.B. Datenschutz, -sicherheit, -besitz) zu komplex.

Das Potential für datengetriebene Innovationen ist sehr hoch; eine OECD Studie stellt fest, dass höhere Investitionen in Big Data und Förderung von Data Sharing und die Wiederverwendung von Daten zu einer besseren Nutzung der Datenanalyse in wirtschaftlicher und sozialer Hinsicht führen werden. Laut Prognosen von IDC werden bis 2020 Ausgaben im Bereich visuelle Aufbereitung der Daten und Datenaufbereitung durch NutzerInnen selber 2,5 mal schneller

[35] http://www.grimme-institut.de/imblickpunkt/pdf/IB-Big-Data.pdf, Seite 2
[36] Vgl. [DatBank14], Seite 495
[37] Vgl. [IDC15], Seite 2
[38] Vgl. [IDC15], Seite 2

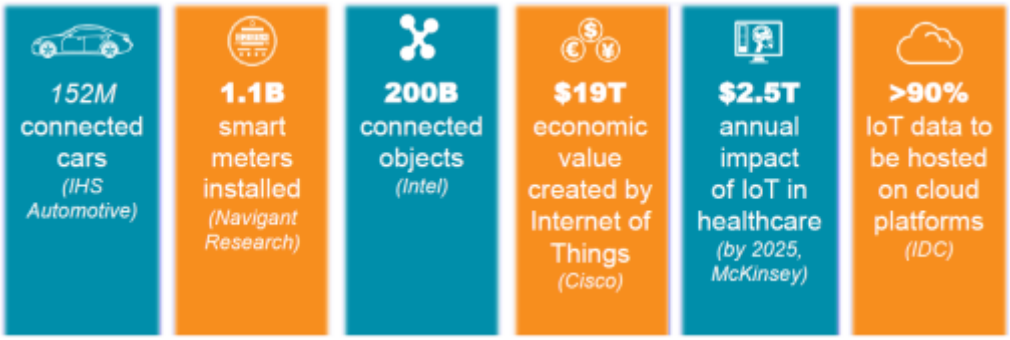
wachsen als für traditionelle IT-Tools mit ähnlichen Funktionen. Bis 2020 wird die jährliche durchschnittliche Wachstumsrate für Big Data-Services 23% betragen[39].

(4.2) Marktsituation

Die International Data Corporation (IDC) erwartet, dass sich der weltweite Big Data Markt von 9,8 Milliarden USD im Jahr 2012 auf 32,4 Milliarden USD im Jahr 2017 steigern wird. Das entspricht einer jährlichen Wachstumsrate von 27 Prozent. Der österreichische Markt wurde 2013 mit etwa 23 Mio. Euro beziffert und könnte laut Prognosen des Austrian Institute of Technology (AIT) bis 2017 auf 73 Millionen Euro jährlich anwachsen. Hinsichtlich der Erwartungen an Big Data hoffen über 50 Prozent der Unternehmen durch Big Data die eigenen Daten besser aufbereiten und analysieren zu können sowie Trends und Verhaltensmuster rascher zu erkennen. Darüber hinaus erwartet man sich Prozessoptimierungen, Kostenreduktionen und verbesserte Profitabilitätsanalysen.[40]

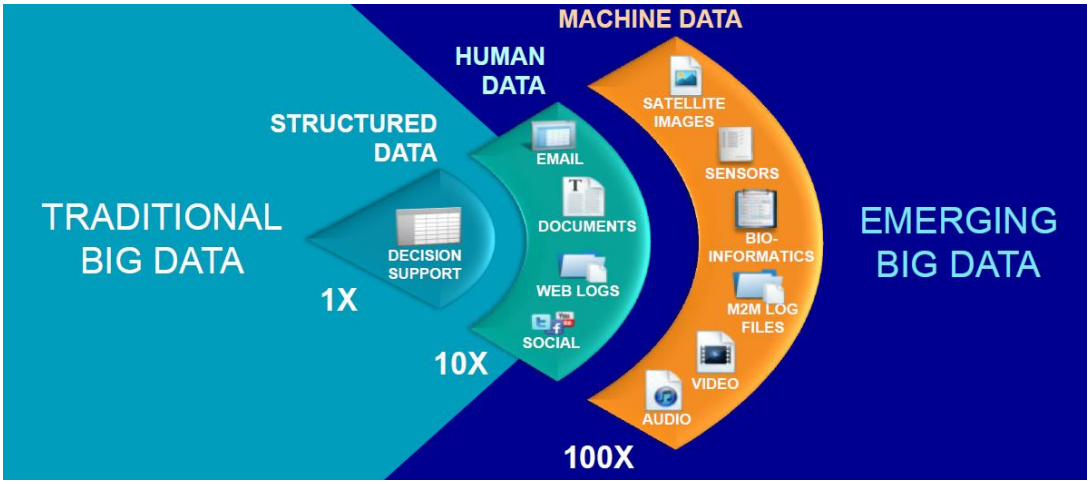
Die beiden nachfolgenden Grafiken verdeutlichen die enormen Veränderungen, die unsere Gesellschaft durch Big Data bis 2020 erfahren wird.

Abbildung 6: Das Internet der Dinge wird bis 2020 allgegenwärtig.



Quelle: Präsentation Hitachi Forum, Wien 3.11.2015

Abbildung 7: Human vs. Machine Data



Quelle: Präsentation Hitachi Forum, Wien 3.11.2015

[39] Vgl. [IDC15], Seite 2
[40] Vgl. [BMVIT2014], Seite 73 ff.

Es gibt einen massiven Trend zur Vernetzung verschiedener Systeme und Geräte, die – mit Sensoren ausgestattet - untereinander kommunizieren, riesige Datenmengen generieren, austauschen und Aktionen in der realen Welt auslösen. Das Internet der Dinge stellt die Verknüpfung der virtuellen Welt mit der realen Welt dar, führt zu einer rasant wachsenden Menge an Daten und ist somit ein wesentlicher Treiber von Big Data.

Die Entwicklung hin zum Internet der Dinge erfordert ein neues Denken. Cisco spricht dabei vom Internet of Everything und hat errechnet, dass 2014 gerade einmal 10 Milliarden Dinge mit dem Internet verbunden sind. In den nächsten Jahren warten jedoch 1,5 Billionen Objekte darauf, ebenfalls mit Sensoren ausgestattet und vernetzt zu werden. Das daraus resultierende Wachstumspotential beziffert Cisco auf **14,4 Billionen US-Dollar**. Aus der Digitalisierung beziehungsweise aus dem Prozess der Erzeugung und Analyse von Big Data leiten sich neue Geschäftsmodelle ab. Denn mit den neuen Daten und den großen Datenmengen erwachsen selbst neue Aufgaben, die zum Teil erst noch erschlossen werden müssen und volkswirtschaftlich ein enormes Potential darstellen. Die Erhebung, die Verarbeitung und Weiterverwertung von Daten und nicht zu vergessen der Schutz von Daten sind unternehmerische Felder, auf denen noch viel brach liegt[41].

Im Bereich des öffentlichen Sektors geht es um die Frage, welcher intelligent vernetzter Objekte der öffentliche Sektor bedarf. Zugleich muss Investitionssicherheit, Kompatibilität und Zukunftsfähigkeit sichergestellt werden. Weitere Fragestellungen sind, in wieweit intelligent vernetzte Objekte auch Steuerungs- und Kontrollaufgaben eigenständig übernehmen sollen. Es wird zentrale Bereiche geben, etwa die Gesetzgebung oder Rechtsprechung, wo automatisierte Entscheidungen prinzipiell abgelehnt werden.[42]

Das Thema Big Data ist Realität und vor allem in Wirtschaft und Wissenschaft gelebte Praxis. Auch in der öffentlichen Verwaltung gibt es bereits seit einiger Zeit Bemühungen, die Möglichkeiten in Zusammenhang mit Big Data Mechanismen stärker für sich zu nutzen, um Entscheidungsgrundlagen rascher bereitstellen zu können, Services z. B. durch die Einbindung von Echtzeit-Daten zu verbessern und Wirkungen (z. B. arbeitsmarktpolitische Maßnahmen) mittels Simulationen besser vorhersagen zu können.

Die massiven digitalen Veränderungen erfordern eine neue Art des Denkens, der fortlaufenden Innovation, Neuerfindung und Anpassung – einen fortdauernden Prozess der sämtliche Phasen der Datenauswertung umfasst.

(4.3) Anwendungsfeder

Die Potenziale für erfolgreiche Big Data Anwendungen sind insbesondere auch aus wirtschaftlicher Sicht interessant. Nachfolgend werden einige Beispiele genannt.

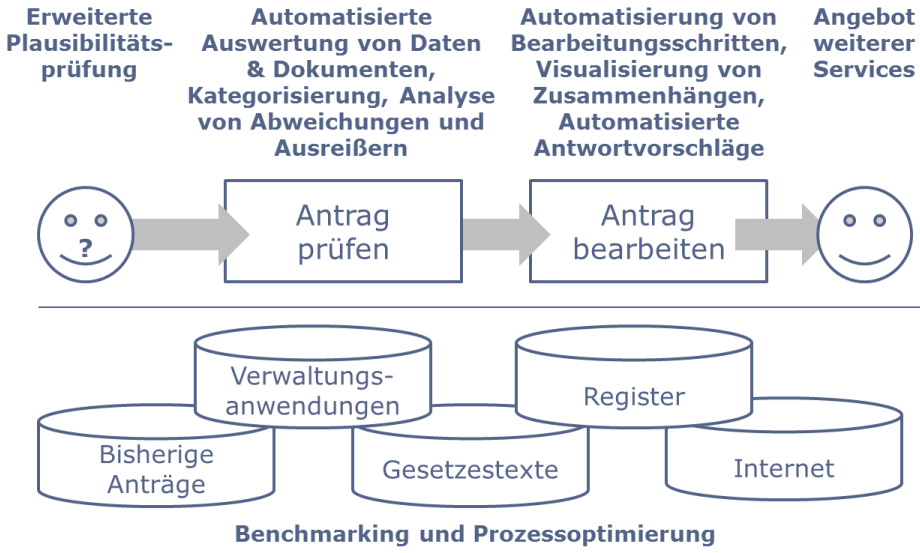
(4.3.1) Effizienzsteigerung und Verwaltungsreform

Big Data kann mehr Qualität und Effizienz in den Verwaltungsprozessen bewirken. Eine Studie des Business Application Research Centers BARC zeigt, dass

[41] https://bigdatablog.de/2014/12/10/big-data-das-internet-der-dinge-zwingen-unternehmen-zum-umdenken/
[42] https://www.zu.de/info-de/institute/togi/assets/pdf/ZU-150914-SmartGovernment-V1.pdf

45 Prozent der befragten Unternehmen durch Big Data-Analysen bereits Prozesskosten senken konnten. McKinsey spricht von einem Sparpotenzial von bis zu 20 Prozent in der öffentlichen Verwaltung. Für ganz Europa wären dies insgesamt bis zu 300 Milliarden Euro.[43] Beispiele dafür gibt es bisher sowohl national als auch international kaum, doch selbst wenn nur zwei bis drei Prozent an Kosteneinsparungen erzielt werden könnten, wäre das Potenzial beeindruckend. Aufgrund der besonderen Relevanz und der vielversprechenden Aussagen wurde in den Untersuchungen zu dem vorliegenden Bericht auf dieses Handlungsfeld ein Schwerpunkt gelegt. Es wurden einige Anwendungsbeispiele identifiziert, anhand derer in Verwaltungsprozessen Effizienzsteigerungen erzielt werden können. In nachfolgender Grafik sind diese mittels eines stark vereinfachten schematischen Prozesses einer Antragsstellung zusammengefasst. Die Beispiele für geeignete Anwendungsbereiche und Prozesse sind vielfältig; etwa Patentanträge, die Vergabe von Förderungen und Beihilfen, Prozesse der Krankenkassen, etc.[44]

Abbildung 8: Anforderungen an Data Scientists



Quelle: [BRZ15], Seite 28

(4.3.2) Services für BürgerInnen und Unternehmen

Durch die intelligente Nutzung von Daten ist es möglich, BürgerInnen künftig eine höhere Servicequalität oder auch neue Services anzubieten. Wie Amazon, Google & Co kann auch die öffentliche Verwaltung Prozesse und Services bei höherer Qualität effizienter gestalten, teilweise sogar automatisieren. Ein Beispiel für eine Big Data-Anwendung ist das Anfragenmanagement, etwa um BürgerInnenanliegen schneller, in höherer Qualität und mit weniger Aufwand zu beantworten. Bei BürgerInnenanfragen hat man es mit einer Unmenge an Daten zu tun. Dabei sind sich Anfragen oft sehr ähnlich. Wenn die Verwaltung nun eine Frage immer wieder neu beantwortet, ergeben sich daraus hoher Zeitaufwand und möglicherweise immer wieder andere Antworten auf dieselbe Frage. Durch die Integration

[43] Vgl. [KIN11], o. S.
[44] Vgl. [BRZ15], Seite 27

unterschiedlicher Kommunikationskanäle und den Zugriff auf verschiedenste Datenquellen können Anfragen konsistenter, qualitativ hochwertiger und effizienter bearbeitet werden. Dabei können im besten Fall aus dem gewaltigen Pool an vorhandenem Wissen (z.B. Verwaltungsanwendungen wie dem ELAK, aus dem Internet oder aus bisherigen Antworten etc.) sogar Antwortvorschläge automatisch erstellt werden.[45]

Weiters könnten One Stop-Shops **durch intelligente Vernetzung** von Daten und Automatisierung von Prozessen zu „**No Stop-Shops**" weiter entwickelt werden. Bei einem Umzug könnte mit der Änderung im Zentralen Melderegister (ZMR) ein Informationsaustausch zur Adressänderung automatisiert erfolgen, sei es in Bezug auf die öffentliche Verwaltung (z.B. Grundbuch) oder privat (Banken, Versicherungen etc.). Beim Hausbau könnte mit der Baubewilligung auch die Beantragung von Förderungen angestoßen werden und mehr. Und auch für Unternehmen wären etwa bei der Unternehmensgründung durch Vernetzung von Daten und Automatisierung weitere Optimierungen möglich.[46]

(4.3.3) Modernisierung der Gesetzgebung

Bereits heute muss für jedes Gesetz und jede Verordnung eine sogenannte „**wirkungsorientierte Folgenabschätzung**" durchgeführt werden. Verschiedene Werkzeuge unterstützen etwa dabei abzuschätzen, um wie viel sich die Anzahl der armutsgefährdeten Menschen verringert, wenn die Lohnsteuer gesenkt wird. In Zukunft lassen sich auf einer breiten Datenbasis und durch den Einsatz von Big Data Technologien solche und schwierigere Fragen zeitnaher und präziser beantworteten. Der Gesetzgeber wird also auch **komplexe Szenarien** in Echtzeit **ex ante** analysieren sowie die potenziellen Auswirkungen bewerten können. Weiters wird die Wirksamkeit der gesetzlichen Maßnahmen **ex post** besser evaluierbar sein.

Darüber hinaus ist die **öffentliche Meinung** stärker in die Arbeit der Verwaltung einbeziehbar. Es können anonymisiert Stimmungen und Aussagen zur sozialen Lage analysiert werden, etwa aus öffentlich zugänglichen Daten wie Zeitungen, Blogs und sozialen Medien. Darauf basierend könnten „Alarmfunktionen" eingerichtet werden, etwa wenn zu einem bestimmten organisatorischen (z.B. Wartezeiten bei einer bestimmten Behörde) oder technischen System (z.B. Pendlerrechner) besonders häufig kritische Diskussionen entdeckt werden. So wäre es möglich, präventive Maßnahmen zu treffen, um potenzielle Probleme bereits frühzeitig abzufangen. Die **Erwartungen und Wünsche** von BürgerInnen könnten besser berücksichtigt werden. Außerdem könnten **Trend-Analysen** zu gesellschaftlich relevanten Themen durchgeführt und neu auftauchenden Problemen proaktiv begegnet werden. Nicht zuletzt sind neue Formen der Demokratie im Bereich der **E-Partizipation** äußerst datenintensiv. Als ein sehr weitreichendes Beispiel sei „**Liquid Democracy**" genannt, wo die BürgerInnenInnen ihre Stimmen zu einzelnen Sachfragen an politische Vertreter „delegieren" (und auch wieder entziehen) können.[47]

[45] Vgl. [BRZ15], Seite 20
[46] Vgl. [BRZ15], Seite 20
[47] Vgl. [BRZ15], Seite 21 f.

(4.3.4) Staatliche Infrastruktur

Einige Paradebeispiele der Anwendungsmöglichkeiten für Big Data-Technologien können unter dem Überbegriff „staatliche Infrastruktur“ zusammengefasst werden. Die Handlungsfelder reichen von der Gesundheit über Verkehr und Energie bis hin zur Bildung. Nachfolgend nur ein paar der denkbaren bzw. zum Teil bereits realisierten Big Data Szenarien:

- **Gesundheit** (Früherkennung von Epidemien, Diagnostik, Kosteneinsparungen durch die Optimierung von Therapien und Medikation)
- **Verkehr** (Verkehrsleitsysteme basierend auf Mobilfunk-, Wetter-, Sensordaten, Planung des Platzbedarfs und Erstellung der Fahrpläne im Bereich der öffentlichen Verkehrsmittel)
- **Energie** (Green-IT, Optimierung im Energieverbrauch durch Smart Metering, Platzierung von Windrädern, etc.)
- **Bildung** (Optimierung von Lehrmethoden, Mobile Learning, uvm.)

(4.3.5) Sicherheit und Kriminalitätsbekämpfung

Das klassische Einsatzgebiet von Big Data Technologien in der öffentlichen Verwaltung liegt wahrscheinlich in der Prävention und der Bekämpfung von Kriminalität. Dabei ist der Fächer der Einsatzgebiete sehr breit und reicht von der Terrorismusbekämpfung und Cyber-Security über Predictive Policing bis hin zur Betrugsbekämpfung. Die Einsatzmöglichkeiten erstrecken sich über verschiedene Ressorts und Anwendungsgebiete:

- **Inneres:** Kriminalitätsbekämpfung, Terrorismusbekämpfung, Predictive Policing, Bekämpfung von Cyber-Crime, etc.
- **Justiz:** Ermittlungstätigkeiten (z.B. Wirtschafts- und Korruptionsstaatsanwaltschaft), Unterstützung der Gerichte, etc.
- **Finanz:** Zoll, Steuerhinterziehung, Geldwäsche, Finanzmarktaufsicht, Insiderhandel, Glücksspiel, Wettbetrug, etc.
- **Weitere:** Förderungen und Beihilfen, Katastrophenschutz, etc.

(4.4) Chancen und Risiken

Innerhalb des öffentlichen Sektors besteht eine breite Aufgabenvielfalt. Der Einsatz von datengetriebenen Innovationen kann in zahlreichen Bereichen zu Produktivitäts-, Prozess- und Effizienzsteigerungen führen. Zum Beispiel liegt ein Nutzungsaspekt von Big in der Einbindung verfügbarer Daten in den statistischen Produktionsprozess: Durch die immer geringere Bereitschaft an konservativen Erhebungen teilzunehmen („survey fatigue“) müssen neue Datenquellen erschlossen werden, um die notwenigen statistischen Ergebnisse, die als Entscheidungsgrundlage dienen, bereitstellen zu können.

Datengetriebene Innovationen unterstützen die öffentliche Hand, soziale und globale Herausforderungen zu adressieren: Diese betreffen u.a. Themen wie zum Beispiel Klimawandel, Naturkatastrophen, Gesundheit, demografischer Wandel und Wasser.

Die potentiellen Anwendungsgebiete für Datenquellen sind also vielfältig. im Rahmen des White Papers der BRZ „Big Data in der öffentlichen Verwaltung“ wurden diese in sechs Handlungsfelder unterteilt:

Services für BürgerInnen und Unternehmen

Modernisierung der Gesetzgebung
Wirtschaft und Arbeit
Staatliche Infrastruktur
Sicherheit und Kriminalitätsbekämpfung
Effizienzsteigerung und Verwaltungsreform

Gerade in der Zusammenarbeit zwischen Wirtschaft und Verwaltung sind Anwendungsmöglichkeiten denkbar. So könnten verschiedenste Daten herangezogen werden, um ein frühzeitiges Erkennen von Indikatoren für Probleme größerer Unternehmen oder ganzer Branchen zu ermöglichen. Es wäre denkbar, Geschäftsbeziehungen, Beteiligungen und Wirtschaftsdaten zu kombinieren, um Risiken abzuschätzen. Daraus ergeben sich auch neue (präventive) Möglichkeiten der **Finanzmarktaufsicht**. Ein weiterer Schwerpunkt könnte die Identifikation von Förderschwerpunkten aus Wirtschaftsdaten, Patenten und Infrastrukturdaten darstellen.[48]

Neben den positiven Effekten müssen auch die potenziellen Risiken mitgedacht werden. So wie der technische Fortschritt zu Zeiten der Industriellen Revolution zu einem Wegfall vieler Arbeitsplätze in bestimmten Bereichen wie der handwerklichen Fertigung geführt hat, wird es auch im Zusammenhang mit Big Data mittelfristig zu Wegfall bestimmter Routinetätigkeiten kommen. Aber ähnlich wie damals werden auch die heutigen Entwicklungen zum Entstehen neuer Bildungszweige und Berufsbilder führen, z.B. Data Scientists. Diese Chancen gilt es frühzeitig zu nutzen und Österreich als Wissensstandort und –exporteur im Bereich Big Data zu etablieren.

In ihrer Summe kann die Nutzung von Big Data in der Verwaltung zur weiteren Steigerung der Lebensqualität, zur Verbesserung der Standortqualität und zur Stärkung der BürgerInnenorientierung beitragen. Am Beispiel des Anwendungsgebietes Arbeitsmarkt soll dies weiter verdeutlicht werden. Die bereits mehrfach geäußerte Notwendigkeit des Aufbaus von Know How und entsprechender Fachkräften ist ein wesentlicher Aspekt für die Steigerung der Attraktivität und der Sichtbarkeit des Standorts Österreich, sowie für die Generierung von Wertschöpfung ist die Bereitstellung und Weiterentwicklung von hoher Kompetenz am Arbeitsmarkt. Diese hohe Kompetenz ist in einem innovationsstarken Bereich wie Big Data erforderlich, um neue Ideen und Unternehmen zu entwickeln, bestehende zu stärken und internationale Unternehmen anzuziehen.

Für die effiziente Verwendung von Daten ist Kompetenz in der Interpretation und Aufbereitung der Daten notwendig, welche derzeit teilweise am österreichischen Arbeitsmarkt nicht verfügbar ist. Hierfür wird als wichtiges Ziel die Weiterentwicklung und Festigung der vorhandenen Kompetenzen definiert, welches unter anderem durch eine Stärkung von wissenschaftlich fundierten Ausbildungen und dem Angebot an zusätzlichen hochqualitativen Weiterbildungsmaßnahmen bestehen kann.[49]

OECD Schätzungen zeigen, dass in den meisten Ländern Datenfachkräfte 2013 unter 1% der Gesamtbeschäftigten ausmachten. Laut einer IDC-Studie wird der

[48] Vgl. [BRZ15], Seite 22
[49] Vgl. [BMVIT14], Seite 100 und Seite 141

Fachkräftemangel anhalten und sich auf den Bereich Data Scientist, Data Architekt und Datamanager beziehen.

Daten und Intelligente Datenanalyse werden die Automatisierung von immer mehr Tätigkeiten ermöglichen. Big Dat hat somit Auswirkungen auf den Arbeitsmarkt, weil intellektuell anspruchsvolle Aufgaben wie z.B. Diagnose auf Basis von medizinischen Bildern automatisiert abgewickelt werden können. Immer mehr Arbeitsplätze mit mittlerem Einkommen werden von dieser Entwicklung negativ betroffen sein.

Es werden nicht alle Arten von Jobs von der Digitalisierung betroffen sein und neue Arten von Skills und Arbeitsplätzen werden sich entwickeln (wie z.B. Dataspezialisten). Die datengetriebene Wirtschaft wird vermehrt Arbeitsplätze in folgenden Bereichen benötigen:

- Lösen unstrukturierter Probleme, inkl. Probleme ohne regelorientierte Ansätze.
- Verarbeitung neuer Informationen, insb. Decison making.
- nicht routinemäßige und manuell geprägte Tätigkeiten, die sowohl „Sehen“ und Feinabstimmung der Muskeln erfordern.

Während die Lösung von unstrukturierten Problemen und das Arbeiten mit neuer Information hochwertigen Arbeitsplätzen vorbehalten sein wird, werden nicht routinemäßig und manuell geprägte Tätigkeit immer wichtiger für Jobs im Niedriglohnbereich[50].

(4.5) Kosten/Nutzen-Überlegungen

Es sollte für jeden Bereich der Verwaltung gesondert hinterfragt werden, inwiefern der Einsatz von Intelligenter Datenanalyse in den Verfahrensabläufe sinnvoll sind und wie diese Lösungen konkret aussehen. Zeitnah könnten jene Bereiche bestimmt werden, in denen die Unterstützung durch Intelligente Datenanalyse ermöglicht, dass intelligenter, effizienter und effektiver verwaltet, gehandelt, gestaltet und zusammengearbeitet werden kann. Dies ist zum Beispiel der Fall, bei verbesserter Situationswahrnehmung, sensorgestützter Entscheidungsanalyse, Prozessoptimierung, Ressourcenverbrauchsoptimierung.

Kostenstrukturen können sich durch den Null-Grenzkosten Effekt verändern und variable Kosten können so minimiert werden. Lassen sich Aufgaben durch den Einsatz von Intelligenter Datenanalyse schneller und kostengünstiger erledigen, kann dies auf lange Sicht das Budget der Verwaltung entlasten[51.] Zu berücksichtigen ist, dass der Einsatz von Intelligenter Datenanalyse die manuellen Aufwände nur dann senkt, wenn sich regelmäßig wiederkehrende Aufgaben automatisieren lassen und dadurch Zeit für individuelle Fragestellungen frei wird.

Es werden neuartige kooperative Ansätze durch Big Dat ermöglicht, die die dynamische Selbstorganisation stärken. Dies kann zur Auflösung von klassischen Zuständigkeits- und Fachbereichen führen.

[50] Vgl. [OECD15], Seite 7
[51] Vgl. [SmartGov15], Seite 9

Es gibt zahlreiche Studien, die eine 5-10-prozentige schnellere Produktivitätssteigerung bei Unternehmen feststellen, die Big Data Technologien einsetzen im Vergleich zu jenen Unternehmen, die Big Dat Technologien nicht einsetzen. Jedoch können diese Schätzungen nicht generalisiert werden. Je nach Sektor variieren die eingeschätzten Effekte von datengetriebenen Innovationen. Die Effekte sind abhängig von komplementären Faktoren, wie zum Beispiel der Verfügbarkeit von qualifizierten Fachkräften und der Verfügbarkeit und Qualität der verwendeten Daten. Diese Studien leiden an einer gewissen Selektionsverzerrung. Dies macht es schwierig, die Effekte von datengetriebenen Innovationen von anderen Faktoren auf Organisationsebene loszulösen. Es werden daher Studien benötigt, die die Auswirkungen von datengetriebenen Innovationen auf Produktivitäts- und Effizienzsteigerung bewerten[52].

(4.6) Innovation

Die Entwicklung hin zum Internet der Dinge– einhergehend mit den dramatisch sinkenden Kosten für die Datensammlung, -speicherung und -verarbeitung und zunehmende Rechenleistung – bedeutet, dass Intelligente Datenanalyse vermehrt ein Motor für Innovation ist und potentiell eine wichtige neue Quelle für Wachstum darstellt.

Intelligentes Datenmanagement schlägt die Brücke von reinen Daten zu Information und Wissen. Im Vordergrund steht die Verknüpfung und Nutzbarmachung der vorhandenen und neu hinzukommenden Daten zur Realisierung innovativer Dienste und Anwendungen.

Auch der bereits zuvor mehrfach im Positionspapier genannte Bedarf der Einrichtung entsprechender Bildungsangebote, um sich den neuen wirtschaftlichen Anforderungen stellen zu können, ist zu beachten.

(4.6.1) Forschungsthemen

Aktuelle Forschungsthemen liegen in folgenden Bereichen:

- **Datenanalyse und Integration** wird die Verarbeitung und Analyse von Daten in beliebiger Form (z.B. Bilder, Videos, Tondokumente, menschliche Sprache) behandelt. Herausforderungen sind auch Aggregation bzw. Fusion von multimodalen bzw. heterogenen Daten.
- **Semantische Verarbeitung** erweitert Daten um Struktur und ermöglicht das Verstehen und den Umgang mit strukturierten Daten auf vielfältige Weise.
- durch geeignete Wissens-Extraktion und -Abstraktion wird die **Automatisierung von Wissensprozessen** ermöglicht, bzw. deren effizientere, kostengünstigere Ausgestaltung.
- **Kognitive Systeme** modellieren menschliche geistige Leistungen und erforschen darauf aufbauend kognitive technische Systeme. 2020 werden 50% aller Business Analytics Software auf cognitive computing basieren.

[52] Vgl. [OECD15], Seite 29

- Algorithmen für **Prädiktion** aus Daten (Maschinelles Lernen, Reasoning, Entscheidungsuntersützung).
- fortgeschrittene Schnittstellentechnologien bis zu **Brain Computer Interfaces**.
- "**Privacy by design**" und Möglichkeiten und Grenzen im Hinblick auf die Risiken von Big Data.

(4.6.2) Wertschöpfung - Schaffung eines Daten-Service-Ökosystems und Datenmärkte

Ein wichtiger Baustein für die Ermöglichung von Daten getriebenen Innovationen ist die Schaffung eines funktionierenden Daten- und Service-Ökosystems. Ein Daten-Service-Ökosystem schafft Win-Win-Situationen für alle Stakeholder (Öffentlicher Sektor, GU, KMU, Start Ups, ForscherInnen, EntrepreneureInnen, BürgerInnen,…), macht Dienste und Daten zugänglich und interoperabel und besteht idealerweise aus folgenden Elementen: Dateninkubator (erleichtert KMUs Zugang zu Daten), Grundinfrastruktur für eine datengesteuerte Wirtschaft (Cloud Computing, HPC, 5G,…), Leuchtturminitiativen für bestimmte Anwendungsfelder wie z.B. Smart City, Industrie 4.0) und Data Curation (Datenpflege und -wiederherstellung).

Hinsichtlich des offenen Zugangs existieren in Österreich bereits übergreifende Initiativen, die zur Auffindbarkeit von Open Data beitragen und öffentliche sowie private Daten zur Wiederverwendung zur Verfügung stellen (data.gv.at, opendataportal.at).

Ein Ökosystem, das Dienste und Daten zugänglich und interoperabel macht, soll ermöglichen, dass auch proprietäre Daten in kontrollierter Art und Weise geteilt werden. Regionen und Gemeinden könnten verbesserte und effizientere Dienste anbieten.

Datenmärkte sollen den Handel mit Daten ermöglichen. Daten könnten angereichert, verknüpft, getauscht, versteigert, verkauft oder gekauft werden. Ein Daten-Service Ökosystem soll auch Dienste (Services) anbieten, u.a. zur Datenbeschaffung, zur Datenbewahrung und Qualitätsverbesserung und -erhaltung aber auch Veredelung bieten. Ebenso sind Dienste für Zitierung, Saldierung und ggf. Lizenzierung notwendig.

(4.6.3) Investitionen in die Zukunft

Der Zugang zu Technologien der Intelligenten Datenanalyse ist kritisch für die Realisierung des Potentials von datengetriebenen Innovationen.

Die Entwicklung der internationalen Patentanmeldungen beweisen, dass Technologien im Bereich datengetriebenen Innovationen rapide im Wachsen begriffen sind[53].

[53] Vgl. [OECDSTI15], Seite 4: "Since 2007, the number of patent filings related to the IoT,big data analytics, and quantum computing and telecommunication have grown at two digit rates; in 2012, thelatest year for which data is available, at more than 40% year-on-year."

Jedoch sind diese auf eine begrenzt Anzahl von Ländern konzentriert, nämlich USA gefolgt von Kanada, Frankreich, Deutschland, Korea, Japan, UK und China. Länder mit erweiterten Kapazitäten zu Technologien der Intelligenten Datenanalyse werden davon profitieren.

Eigentumsrechte von autonomen Maschinen und Systemen werden durch IP-Rechte definiert werden[54].

(4.7) Empfehlungen für die öffentliche Verwaltung

Das enorme Potential von Big Data erfordert, dass die bedeutenden wirtschaftlichen und sozialen Herausforderungen unserer Gesellschaft in einem regierungsweiten und partizipativen Ansatz adressiert werden. Die Vorteile von Big Data sollen maximiert und die mit Big Data in Verbindung stehenden Risiken und Barrieren gemildert werden - insbesondere die Auswirkungen auf Arbeitsplätze, geistige Schutzrechte, Wettbewerb und Steuerwesen.

Ein Spannungsverhältnis besteht insbesondere zwischen Privacy-Aspekten und Innovation. Daher sollen Lösungen gefunden werden, die Big Data Anwendungen ermöglichen und gleichzeitig Innovationen generieren und Datenschutzaspekte berücksichtigen. Es geht um die Frage, wie große Datenmengen mit Personenbezug für verwaltungsinterne Innovationen genutzt werden können. Die Verknüpfung personenbezogener Daten mit Daten aus anderen Datenquellen (unterschiedliche Datenarten, unterschiedliche IT-Verfahren) sollen das Verwaltungshandeln und -prozesse unterstützen ohne, das Grundrecht auf Datenschutz zu beeinträchtigten. Dazu sind die rechtlichen Voraussetzungen zu schaffen und Kooperationen zwischen den Verwaltungsorganisationseinheiten zu schließen. Diese organisationsübergreifende Zusammenarbeit kann für die Verwaltung aber auch für die BürgerInnen neue Innovationsfelder erschließen.

Nach einer Kosten-/Nutzen- und Risiken Analyse sollen Big Data-Lösungen in der Verwaltung eingesetzt werden. Die in diesem Kapitel dargestellten Anwendungsfelder verdeutlichen die vielfältigen möglichen Einsatzgebiete. Als eine Einsatzmöglichkeit der öffentlichen Verwaltung, bei der der Einsatz von Big Data bereits positive Ergebnisse aufzeigt, ist der Arbeitsmarkt genannt worden.

Entsprechend der dargestellten Chancen und Risiken in Kapitel (4.4) empfiehlt eine OECD-Studie[55] das frühzeitige Adressieren von Arbeitsmarktaspekten und Ungleichheit, die aufgrund der technologischen Veränderungen auch Änderungen in den Qualifikationen bedingen. ArbeitnehmerInnen innerhalb wie außerhalb der Verwaltung sollen dabei unterstützt werden, sich auf die datengetriebene Wirtschaft einzustellen. Ungleichheit kann zum zentralen Thema werden, insbesondere, wenn es um den Zugang zu dringend benötigten hochqualifizierten Arbeitskräften geht und die Vorteile nur auf einen bestimmten Kreis limitiert sind.

Aus wirtschaftlicher Sicht soll die Verwaltung in Ihrer strategischen Ausrichtung vor allem folgende Bereiche verfolgen:

- Bildung, Weiterbildung, Training im Bereich Data Science, Open Source, sowie Privacy und Security.

[54] Vgl. [OECDDDI15], Seite 15
[55] Vgl. [OECDSTI15], Seite 7

- Adressieren von Risiken, die Ungleichheit im Einkommen am Arbeitsmarkt verschärfen durch Sozial- und Steuerpolitik.

Investitionen sind neben Aus- und Fortbildungsmaßnahmen auch im Bereich Innovation und damit im Forschungsbereich notwendig. Die daraus resultierende Wertschöpfung hängt ganz wesentlich von einem funktionierenden Daten- und Service-Ökosystems und der Beteiligung aller relevanten Stakeholder (Öffentlicher Sektor, KMU, Start Ups, ForscherInnen, Wirtschaftstreibende, BürgerInnen,…) ab. Solch ein Daten-Service-Ökosystem schafft für alle beteiligten eine erfolgversprechende Konstellationen und macht Dienste und Daten zugänglich und interoperabel.

Ein wesentlicher Erfolgsfaktor ist hierbei natürlich auch eine stärkere Kooperation innerhalb der Verwaltung, wie sie zuvor auch schon bei den strukturellen Aspekten gefordert wurde. Nur so kann die österreichische Verwaltung die sich durch Big Data ergebenden wirtschaftlichen Chancen nutzen und Österreich als Wissensstandort und Wissensexporteur in diesem Bereich etablieren.

(5) Technische Aspekte

Oft werden technische Aspekte als Herausforderungen beim Einsatz von Big Data in der öffentlichen Verwaltung gesehen. Die Herausforderungen sind jedoch zumeist weniger technisch, als viel mehr, wie bereits in den Abschnitten zuvor geschildert, organisatorisch – struktureller Natur. In der Privatwirtschaft ist Big Data bereits seit einiger Zeit ein Geschäftsmodell und der IKT-Markt bietet eine Reihe von erprobten technischen Lösungen an.

(5.1) Ausgangslage

Die Explosion der Menge an Daten und Informationen in unserer Gesellschaft stellt IT-Anwendungen vor die Herausforderung, immer mehr Daten in stetig kürzerer Zeit speichern, analysieren und verarbeiten zu müssen. Die seit Jahrzehnten erfolgreich eingesetzten relationalen Datenbanksysteme (RDBMS) und Business Inteligence (BI) Tools können diese Anforderungen oft nur noch ungenügend, bzw. nur unter Einsatz hoher Hardwareinvestitionen erfüllen.

Obwohl diese bewährten Systeme auch weiterhin einen Platz in der modernen IT-Welt haben werden, gibt es Anwendungsfälle, die den Einsatz alternativer Ansätze und Technologien erfordern.

Dieser Abschnitt soll nicht nur Technologien und Konzepte vorstellen, welche einerseits die Verbreitung von Big Data vorantreiben, sondern andererseits auch jene, welche besagte Flut an Daten und Informationen erfolgreich bewältigbar machen.

Grundsätzlich sind diese Technologien alle noch sehr jung und entwickeln sich auch dementsprechend schnell weiter. Es ist hierbei vor Allem auch wichtig, dass in der universitären Ausbildung auf diese Aspekte Wert gelegt wird und auch mit Unternehmen, welche diese Technologien einsetzen (Google, Runtastic,…etc.) kooperiert wird.

(5.2) Chancen/Risiken

Großes Potential liegt in der Möglichkeit zur Umsetzung von Anforderungen die bis vor kurzem noch undenkbar waren und möglichen Kostenersparnissen im Vergleich zu traditionellen Technologien.

So problematisch es wäre an Big-Data-Herausforderungen mit traditionellen relationalen Datenbanksystemen heranzugehen, so falsch wäre es ebenso, Big Data-Technologien als Lösung sämtlicher Probleme zu sehen. Aus diesem Grunde versucht dieser Abschnitt auch ein Gefühl zu vermitteln, welches Werkzeug für welche Herausforderung eingesetzt werden können.

(5.3) Driving Technologies

Die explosionsartige Entwicklung von Daten wird durch einige Technologien mit hervorgerufen, welche somit als Triebfeder für Big Data zu sehen sind.

(5.3.1) Social Media

Social Media (auch Soziale Medien) sind digitale Medien und Technologien (vgl. Social Software[56]), die es NutzerInnen ermöglichen, sich untereinander auszutauschen und mediale Inhalte einzeln oder in Gemeinschaft zu erstellen.

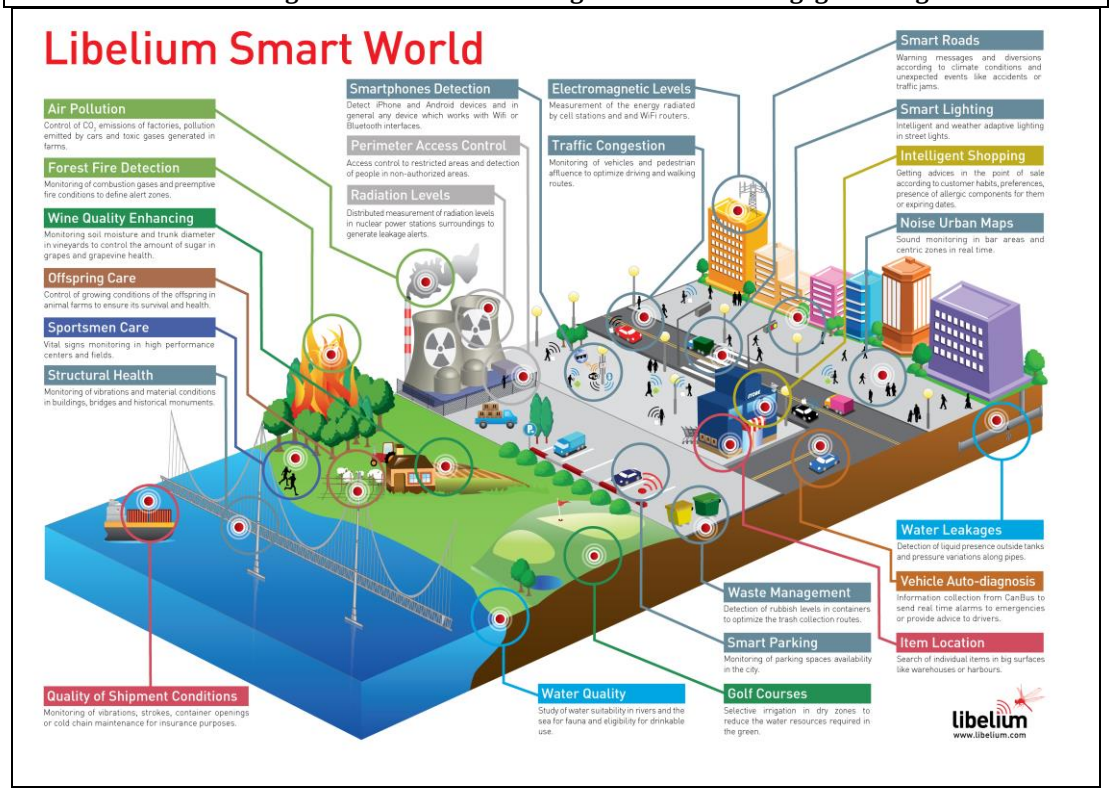
Dabei entstehen Unmengen von Daten, die einerseits gespeichert und abgefragt werden müssen, andererseits aber auch durchaus offline analysiert werden können, um wertvolle Informationen zum Vorschein zu bringen. Viele Hersteller von BI-Tools und Predictive Analytics-Software bieten Schnittstellen an, um auf solche Daten zuzugreifen. [57]

(5.3.2) Internet of Things (IoT)

Internet of Things, oder das Internet der Dinge (siehe auch Kapitel 4.3.1), ist ein Begriff, welcher die vernetzte Welt, in welcher wir leben, beschreibt. Geräte wie Mobiltelefone, Fahrzeuge, Wohnraum, etc. werden immer häufiger mit Sensoren ausgestattet, welche Informationen erfassen, speichern und unter einander austauschen – unten stehende Abbildung versucht dies zu illustrieren.

Möglich gemacht wird das IoT durch Technologien wie RFID (Radio-frequency identification) und IPv6 (Internet Protocol Version 6) die es erlauben beliebige Objekte zu vernetzen. Dabei werden enorme Mengen an Daten generiert, deren effiziente Speicherung, Abfrage und Weiterverarbeitung den Einsatz von Big Data Technologien bedingt.

Abbildung 9: Das Internet der Dinge wird bis 2020 allgegenwärtig.



[56] https://de.wikipedia.org/wiki/Social_Software
[57] https://de.wikipedia.org/wiki/Social_Media

Quelle: Forbes[58] (http://www.forbes.com)

(5.4) Enabling Technologies

Verschiedenste Technologien können zur Bewältigung der Big Data Herausforderungen herangezogen werden. Diese umfassen ein breites Spektrum von Anwendungsfällen, von der simplen Speicherung großer Datenmengen über die Analyse bis hin zur Vorhersage künftiger Entwicklungen.

(5.4.1) NoSQL

Der Begriff NoSQL steht für „Not Only SQL“. SQL bezeichnet dabei die Structured Query Language – eine deklarative Programmiersprache welche zur Abfrage von Daten in relationalen Datenbankmanagementsystemen verwendet wird. Unter dem Begriff NoSQL werden verschiedenste Datenbanksysteme und –technologien zusammengefasst, welche grundsätzlich auf die Verarbeitung von nicht relational erfassbaren Daten spezialisiert sind.

Darüber hinaus sind diese Datenbanksysteme auf verteilte und horizontale Skalierbarkeit ausgelegt, um performant große Datenmengen (im Bereich TerraByte und darüber) zu verarbeiten.

Eine weitere, wesentliche Eigenschaft von NoSQL-Datenbanken ist auch die Flexibilität, was zum Großteil damit zusammenhängt, dass Sie die BenutzerInnen oder EntwicklerInnen nicht dazu zwingen ein gewisses Schemaformat einzuhalten.

Die Webseite nosql-database.org listet derzeit mehr als 225 NoSQL-Datenbanken, welche sich in den technologisch zugrundeliegenden Konzepten teilweise sehr deutlich unterscheiden. Davon lässt sich ein Großteil einer bzw. mehrere der nachfolgend beschriebenen Kategorien zuordnen, wobei diese Liste exemplarisch die wichtigsten Kategorien aufzeigt und keinen Anspruch auf Vollständigkeit erhebt.

(5.4.2) Spaltenorientierte Datenbanken

In relationalen Datenbanken werden die Daten zeilenorientiert gehalten. Dies entspricht dem zeilenweise zusammenhängen von Daten und Einfügen von neue Daten als zusätzliche Zeilen.

ID	Name	Gehalt
1	Meier	2500
2	Müller	1700
3	Huber	1900

Hingegen werden Daten bei spaltenorientierten Datenbanken folgendermaßen abgelegt:

ID	Name	Gehalt
1, 2, 3	Meier, Müller, Huber	2500, 1700, 1900

Einer der größten Vorteile der spaltenorientierten Speicherung ist die Möglichkeiten zur Komprimierung, was bei den heutigen Datenmengen ein sehr

[58] http://www.forbes.com/sites/jacobmorgan/2014/05/13/simple-explanation-internet-things-that-anyone-can-understand/#390072cf6828

wichtiges Kriterium darstellt. Außerdem bietet sie gewisse Vorteile bei der Analyse der Daten, dem Caching und Operationen auf Attributebene wie Filterung, Sortierung und Aggregation.

Hinsichtlich der Performance von Abfragen sind spaltenorientierte Datenbanken vorzuziehen, wenn bei der Abfrage eher einzelne Spalten oder viele Zeilen benötigt werden.

Zeilenorientierte Datenbanken hingegen spielen ihre Stärken beim Einfügen von (größeren Mengen von) Datensätzen, sowie beim Verknüpfen mehrere Tabellen (Joins, karthesisches Produkt) sowie beim Auslesen ganzer Tabellen aus.

Grundsätzlich eigenen sich spaltenorientierte Datenbanken sehr gut, um relativ einfache und auf Listen basierende Datenmodelle abzubilden. Damit scheinen sie zum Beispiel ideal geeignet, um Sensor- oder Telemetrikdaten zu speichern.

Bekannte Vertreter spaltenorientierter Datenbanken sind zum Beispiel[59]:

- SAP HANA
- Oracle: Oracle 12c Enterprise Edition mit In-Memory
- Apache Cassandra
- Sybase IQ

(5.4.2.1) Document Store

Bei dokumentenorientierten Datenbanken werden Daten in Form von Dokumenten gespeichert.

Ein Dokument ist dabei eine hierarchische Ansammlung von sogenannten Key/Value-Paaren, die aber zusammen ein gewisses (möglicherweise auch durchaus komplexes) Datenmodell beschreiben. Dokumente werden dabei in der Regel in ein binäres Format überführt, indiziert und persistiert. Relationen zwischen einzelnen Dokumenten können über Dokumentreferenzen hergestellt werden.[60]

Als Dokumentformat kommt hier sehr häufig JSON (Java Script Object Notation) zum Einsatz.

Folgendes exemplarische JSON-Dokument zeigt zum Beispiel wie Marker in Google-Maps als JSON-Dokument abgebildet werden können[61]:

```
{"markers": [
.....    {
.....              "point":new GLatLng(40.266044,-74.718479),
.....              "homeTeam":"Lawrence Library",
.....              "awayTeam":"LUGip",
.....              "markerImage":"images/red.png",
.....              "information": "Linux users group meets second ...",
.....              "fixture":"Wednesday 7pm",
.....              "capacity":"",
.....              "previousScore":""
.....    },
.....    {
.....              "point":new GLatLng(40.294535,-74.682012),
.....              "homeTeam":"Applebees",
```

[59] Vgl. [BRZ16], o.S.
[60] Vgl. [JAVA16], Seite 22ff
[61] http://www.sitepoint.com/google-maps-json-file

```
 .....              "awayTeam":"After LUPip Mtg Spot",
 .....              "markerImage":"images/newcastle.png",
 .....              "information": "Some of us go there …
 .....              "fixture":"Wednesday whenever",
 .....              "capacity":"2 to 4 pints",
 .....              "tv":""
 .....    },
] }
```

Im Gegensatz zu relationalen Datenbanken besteht bei dokumentorientierten Datenbanken kein Schemazwang. Daher können Dokumente also beliebig um neue Elemente erweitert werden und eine Datenbank kann auch problemlos verschiedene Arten von Dokumenten (z.B. Twitter-Feeds, Facebook-Posts, GoogleMaps-Marker, …) Seite an Seite speichern. Lediglich die angebundene Anwendung muss damit natürlich umgehen können.

Bekannte Vertreter dokumentenorientierter Datenbanken sind zum Beispiel die OpenSource-Produkte MongoDB oder Couchbase. Auch das IBM-Produkt Notes/Domino, dessen Anfänge schon in die 1990er Jahre zurückreichen, kann durchaus zu dieser Kategorie gezählt werden. Ein ebenfalls recht interessanter Vertreter dieser Kategorie ist das kommerzielle Produkt MarkLogic, welches zum Beispiel mit Features wie ACID-Konformität und XA-Transaktionen aufwartet, welche eigentlich in der Welt der relationalen Datenbanken zu finden sind.

(5.4.2.2) Graphen Datenbanken

Graphendatenbanken bilden in erster Linie nicht Daten ab, sondern vor allem ihre Beziehungen zueinander. Eine solche Datenbank besteht also immer aus Knoten die durch Kanten miteinander verbunden sind, wobei sowohl Knoten als auch Kanten Properties haben können.

Graphen Datenbanken finden zum Beispiel häufig Verwendung in sozialen Netzwerken oder Logistikanwendungen und spielen ihre Stärke vor allem dann aus, wenn bei den abzubildenden UseCases hauptsächlich die Beziehungen zwischen Objekten im Vordergrund stehen. Ein sehr bekannter OpenSource-Vertreter dieser Kategorie ist zum Beispiel Neo4j.

(5.4.2.3) Key value/Tuple Store

In solchen Systemen werden Daten als einfache Schlüssel/Wert-Paare gespeichert und lassen sich auch nur über den Schlüssel wieder finden. Vereinfacht dargestellt ist ein Key/Value-Store eine zweispaltige Datenbank, wo in einer Spalte der Schlüssel liegt und in der anderen der eigentliche Wert, auf den über den Schlüssel zugegriffen wird.

Der populärste Vertreter von Key/Value bzw. Tupel-Stores ist Redis. Weitere Vertreter sind z.B. Memcached oder Oracle NoSQL. Auch das Hadoop zu Grunde liegende FileSystem HDFS (Hadopp File System) zur Speicherung der von Hadoop analysierten Daten kann zu dieser Kategorie gezählt werden.

Vorteile von Key/Value bzw. Tupel-Stores sind Skalierbarkeit und effiziente Datenhaltung (für dafür geeignete Daten wie Web 2.0 Daten, die nicht miteinander verbunden sind). Die Verarbeitung und Haltung von komplexen relationalen Daten ist nicht die Stärke von Key Value / Tupel Stores.[62]

[62] Vgl. [BRZ16], o. S.

(5.4.3) In-Memory Technologien

In-Memory-Technologie verschiebt Daten- und Informationsquellen von Datenbanken in den Arbeitsspeicher, damit die Ergebnisse von Analysen und Transaktionen sofort verfügbar sind. Die Elemente der In-Memory-Technologien sind nicht neu, aber verbesserte Hard-und Software-Technologien haben es möglich gemacht, Realtime-Enterprise Anwendungen mit In-Memory-Technologien zu realisieren.

In-Memory-Datenbanken werden sehr oft als sogenannte Appliances, also als Paket aus Hard- und Software angeboten, können aber auch durchaus auf zertifizierter Hardware von Drittherstellern installiert werden. Die verwendete Hardware zeichnet sich dabei meist durch äußerst großzügig dimensionierte Speicher und Rechenkapazitäten aus, so bietet zum Beispiel eine Oracle Big Data Appliance Maschine (X5-2) bis zu 13 TB Hauptspeicher pro Rack die gleichzeitig von 648 Cores bearbeitet werden können. [63]

Ebenso wird bei solchen Systemen aber auch versucht die im Hauptspeicher zu haltenden Daten zu vermindern. Dies kann zum Beispiel durch den Einsatz von spaltenorientierter Speicherung erreicht werden, indem man die dann mögliche Datenkompression nutzt und auf dann nicht nötige Indizes verzichtet.

Bekannte Vertreter sind zum Beispiel:[64]

- SAP Hana
- Oracle: Oracle 12c Enterprise Edition mit In-Memory

(5.4.4) Predictive Analytics

Die größte Herausforderung im Big Data-Umfeld ist es sicherlich, die massenhaft vorhandenen Daten nicht nur zu speichern, um sie später wieder abzufragen, sondern aus ihnen mittels Analysen, Data Mining, Aggregation und Korrelation neue Information und Wissen zu schaffen.

Nach der Definition der Analysten von Forrester[65] kann zu Predictive Analytics jede Lösung gezählt werden,

> mit deren Hilfe sich aussagekräftige Muster und Abhängigkeiten in Datenbeständen identifizieren lassen,
> und auf diese Weise mögliche zukünftige Ereignisse vorhersagen sowie potenzielle Handlungsmöglichkeiten bewerten lassen.

Quasi eine Vorstufe für Predictive Analytics stellt das Data Mining dar, dessen Aufgabe es ist Muster in Datenbeständen zu erkennen. Eine Unterform davon ist das Text Mining - also das Erkennen von Zusammenhängen und Mustern in Texten. Dazu sind Mittel wie Klassifizierung (Clustering) und Modellierung von Entscheidungsbäumen, neuronale Netze sowie Assoziationsanalysen nötig.

Predictive Analytics gibt sich aber nicht nur mit dem Erkennen von Mustern zufrieden, sondern versucht daraus durch statistische Berechnungen, Simulationen und weiteren fortschrittlichen Methoden Vorhersagen zu ermitteln.

---

[63] Vgl. [DOAG15], o. S.
[64] Vgl. [BRZ16], o. S.
[65] http://www.forrester.com/rb/research

Wichtig für den Einsatz von Predictive Analytics sind laut Forrester bestehende Data Warehouse oder ETL-Prozesse (Extract, Transform, Load), da einerseits dort bestehende Daten mit in Analysen einbezogen werden sollen und andererseits Synergieeffekte dadurch entstehen, dass Funktionen zu Predictive Analytics in bestehende BI-Landschaften integriert werden. Viele große BI-Hersteller bieten inzwischen die Integration von Big Data-Komponenten an. Insbesondere auch die Nutzung von Visualisierungstools vorhandener BI-Systeme kann einen großen Mehrwert bringen.

Entsprechende Predicitve Analytics Software-Suiten welche für sämtliche Arbeitsschritte, von der Aufbereitung der Daten über die Analyse bis hin zur Visualisierung, Unterstützung bieten werden zum Beispiel von den folgenden Herstellern meist im Rahmen ihrer BI-Produkte angeboten[66]:

- Oracle
- IBM
- Terradata
- SAS
- SAP

Jenseits des für Predictive Analytes nötigen fortschrittlichen statistischen und algorithmischen Rüstzeugs werden auch Technologien benötigt, um mit den Mengen an Daten umzugehen. Hierfür gibt es verschiedenste Ansätze, die im Folgenden kurz beschrieben werden.[67]

(5.4.4.1) Business Intelligence (BI)-auf Basis In-Memory-DB

Die bereits in Abschnitt (5.4.3) beschriebenen In-Memory-Technologien wie SAP Hana können ebenso zur Unterstützung im BI-Prozess herangezogen werden.

(5.4.4.2) BI auf Basis MMP-DB

Gleichfalls können traditionelle BI-Prozesse auf Basis von sogenannten MMP „massively parallel processing“ Datenbanken wie zum Beispiel Greenplum von Pivotal laufen.[68]

(5.4.4.3) Apache Hadoop

Hadoop ist ein in Java geschriebenes Framework und ermöglicht es, auf Clustern von Computern / Servern rechenintensive Operationen mit großen Datenmengen unter Verwendung des Map/Reduce-Algorithmuses durchzuführen. Hadoop selbst besteht dabei hauptsächlich aus zwei Komponenten:

- HDFS: Dabei handelt es sich um das verteilte Filesystem von Hadoop auf welchem die zu berechnenden Daten gespeichert werden. Es handelt sich hierbei um eine Key value/Tuple Store.

[66] Vgl. [BRZ14], Seite 22ff
[67] http://www.computerwoche.de/a/mit-predictive-analytics-in-die-zukunft-blicken,2370894 und [BRZ16], o.S.
[68] https://en.wikipedia.org/wiki/Massively_parallel_(computing)

- Map/Reduce: Dabei handelt es sich um einen von Google eingeführtes Programmiermodell für nebenläufige Berechnungen über große Datenmengen auf verteilten Systemen.

Stark vereinfacht beschrieben, werden dabei jeweils eine Map- und eine Reducefunktion spezifiziert. Zuerst werden die Daten auf die Knoten des Rechensystems verteilt, auf welchen dann die Map-Funktion ausgeführt wird. In der sogenannten Shufflephase werden nach gewissen Schlüsseln zusammengehörige Daten dann auf einem Knoten gesammelt, auf welchem dann jeweils die Reduce-Funktion aufgerufen wird, die das Endergebnis ausgibt. Der entsprechende Wikipediaeintrag zum Thema Map/Reduce[69] erläutert das Prinzip sehr anschaulich anhand des klassischen Big Data Beispiels „WordCount".

Rund um Hadoop hat sich ein großes Ökosystem an Tools entwickelt, welches bei der Entwicklung von Map/Reduce-Jobs und dem Anzapfen von Datenquellen unterstützen (einige Beispiele sind Flume, Sqoop, Hive, Jaql und Pig). Viele Hersteller wie zum Beispiel Microsoft, IBM, Terradata oder SAS integrieren und erweitern Hadoop und Teile seines Ökosystems im Rahmen ihrer Produkte. Andere Hersteller wie Hortonworks oder Cloudera bieten spezielle Hadoopdistributionen an.[70]

(5.4.4.4) Apache Spark

Bei Apache Spark handelt es sich wie bei Hadoop ebenfalls um ein OpenSource-Framework, welches die verteilte Analyse von Big Data-Inhalten erlaubt. Es ist etwas jünger als Hadoop und wird von einigen Seiten als dessen Nachfolger proklamiert, da es performanter und leichter zu benutzen sein soll.

Darüber hinaus kann es im Gegensatz zu Hadoop nicht nur im Batchmodus laufen, sondern auch real-time streaming data verarbeiten. Außerdem ist Spark auch nicht wie Hadoop auf ein eigenes FileSystem zur Speicherung der Daten beschränkt, sondern kann auf verschiedenen anderen NoSQL-Datenbanken wie eben HDFS, Apache Cassandra oder auch Amazon S3 operieren. Ebenfalls interessant sind die native Unterstützung von In-memory-Technologien und die Möglichkeit Daten in mehreren iterativen Schritten zu verarbeiten. Diese Eigenschaften machen Spark zusätzlich auch für eine Onlineverarbeitung interessant, was es deutlich von Hadoop abhebt.

Die Funktionalität von Apache Spark lässt sich über Module erweitern. So existieren zum Beispiel Erweiterungen, um relationale oder graphenorientierte Datenbanken in den Analyseprozess einzubinden, oder Module welche Algorithmen im Bereich Machine Learning bereitstellen. [71]

(5.5) Abgrenzung zu traditionellen RDBMS und BI-Systemen

Die in Abschnitt (5.4.1) beschriebenen NoSQL-Technologien unterscheiden sich zum Teil stark von traditionellen relationalen Datenbanksystemen.

[69] https://de.wikipedia.org/wiki/MapReduce#Beispielhafte_Berechnung
[70] **Fehler! Hyperlink-Referenz ungültig.**https://de.wikipedia.org
[71] http://www.infoq.com/articles/apache-spark-introduction

Augenscheinlichstes Unterscheidungsmerkmal ist natürlich der Verzicht auf das zeilenbasierte, relationale Datenmodell.

Ebenso prüfen moderne RDBMS welche Daten sich in einer Datenbanktabelle speichern lassen und welche nicht. Dies wird durch gewisse Constraints erreicht, welche der Ersteller des Datenmodells vorgibt und die festlegen welche Daten sich im erstellten Modell speichern lassen. NoSQL-Technologien sind hier großzügiger und erlauben es den BenutzerInnen verschiedenste Daten nebeneinander zu speichern. Was auf den ersten Blick enorm vielversprechend aussieht bedeutet aber, dass angebundene Anwendungen sicherstellen müssen, dass sie auch mit den verschiedenen Datenmodellen umgehen können. Es wird also Logik und Verantwortung für die Kontrolle der Datenintegrität von der Datenbankschicht in die Applikationsschicht verschoben, was sich aber in bestimmten Fällen lohnen kann.

Einer der größten Unterschiede ist, dass NoSQL-Datenbanken in der Regel (bis auf einige Ausnahmen wie MarkLogic oder Neo4j) auf das sogenannte ACID-Prinzip verzichten. ACID steht als Akronym für:

Atomicity: dies sagt aus, dass eine verändernde Anfrage an eine Datenbank entweder in ihrer Gänze erfolgreich beendet wird oder als Gesamtes schief geht. Ein Zwischenstatus ist nicht erlaubt.

Consistency: damit soll ausgesagt werden, dass sich eine Datenbank während einer Transaktion immer in einem konsistenten und validen Stand befindet und es nicht möglich ist, dass die verwendeten Daten außerhalb der Transaktion verändert werden.

Isolation: gibt an, dass BenutzerInnen, welche sich in einer gewissen Transaktion mit gewissen Daten einer Datenbank befindet, nicht von anderen zugreifenden BenutzerInnen beeinflusst werd.

Durability: gibt an, dass alle Daten die in einer erfolgreich durchgeführten Transaktion geschrieben wurden sich auf nichtflüchtigem Speicher befinden.

NoSQL-Systeme hingegen stellen weniger hohe Anforderungen an die Konsistenz der Daten und arbeiten meist nach dem sogenannten BASE-Prinzip (siehe auch CAP-Theorem[72]):

Basically Available: sagt aus, dass die Datenbank garantiert für alle BenutzerInnen sofort und zu jeder Zeit verfügbar ist.

Soft State: drückt aus, dass Daten im System während einer Transaktion verändert werden können, da zum Beispiel ein anderer Knoten neuere Daten liefert.

EventuallyConsistent: gibt an, dass sich Daten zwar in einem Soft State befinden können, schließlich irgendwann aber durch das System (Replikation) in einen konsistenten Zustand überführt werden.

Dies erlaubt NoSQL-Technologien die einfache und günstige horizontale Skalierung. sodass enorme Datenmengen bei äußerst hoher Verfügbarkeit verwaltet werden können.

In-Memory-Datenbanken großer Hersteller wie Oracle oder SAP verzichten zwar ebenfalls zumindest teilweise auf die zeilenorientierte Speicherung, geben aber das ACID-Prinzip dabei nicht auf und können daher natürlich zur Speicherung

[72] https://de.wikipedia.org/wiki/CAP-Theorem

kritischer Geschäftsdaten verwendet werden. Darüber hinaus sind die Daten in solchen Datenbanken auch weiterhin über SQL zugreifbar. SAP zum Beispiel bietet inzwischen die SAP Business Suite auf Basis von SAP Hana an.[73]

Klassische BI-Prozesse arbeiten meist auf Daten die aus Online-Systemen in die BI-Umgebung gespielt werden. Big Data Technologien können hier als Ergänzung in verschiedenen Phasen des BI-Prozesses eingesetzt werden.

(5.6) Integration in BI-Prozess

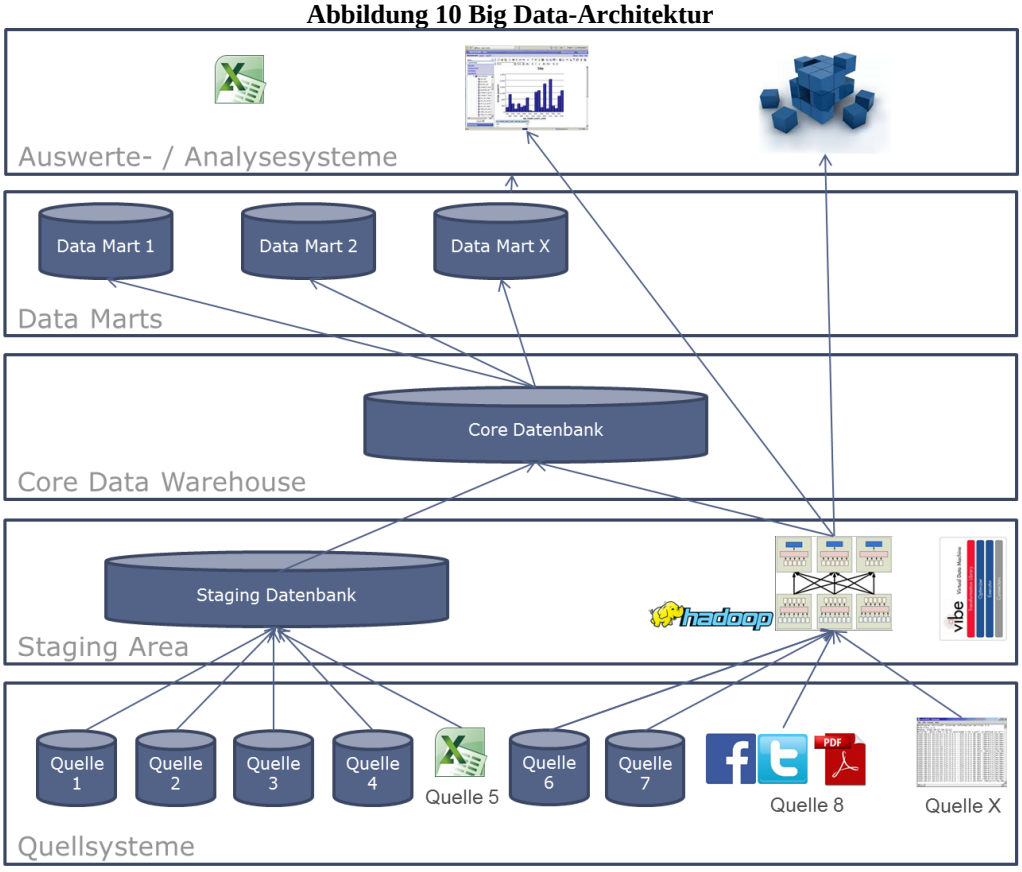
Im Bereich der öffentlichen Verwaltung sowie in größeren Unternehmen existieren meist bereits sehr umfangreiche Data-Warehouses die mittels eines ausgereiften BI-Prozess nach dem ETL-Prinzip (Extract/Transform/Load) gefüllt und abgefragt werden.

Vereinfacht gesagt werden dabei Daten aus verschiedenen Systemen und Anwendungen entnommen (Extract), in einer eigenen Umgebung, der sogenannten Staging-Area entsprechend behandelt (Transform) und dann ins Datawarehouse eingespielt (Load). Auf diesem Datenbestand werden dann entsprechende Abfragen und Analysen (vlg. Predictive Analytics) abgesetzt bzw. Teile der so gewonnenen Daten wieder in eigene Datenbanken exportiert (man spricht dann von sogenannten Data-Marts).

Zur Integration von Big Data Technologien wie Hadoop oder Spark und zur Zusammenfassung von semantischer Suche, Text Mining, Predictive Analytics und Data Warehouses zu einer stimmigen Big Data Lösung wurde im Zuge einer strategischen Initiative zum Thema Big Data im BRZ die folgende Big Data-Architektur vorgeschlagen[74]:

[73] Vgl. http://www.golem.de/news/business-suite-on-hana-sap-stellt-r4-vor-das-nicht-so-heisst-1301-96843.html
[74] Vgl. [BRZ14], o.S.

Abbildung 10 Big Data-Architektur



Quelle: BRZ 2014

Das klassische Datawarehouse wird hierbei um Daten angereichert, welche aus Big Data Quellen gewonnen wurden. Solche Quellen könnten zum Beispiel die schon vorgestellten NoSQL-Datenbanken darstellen. Mittels Hadoop oder Spark können diese Daten dann entsprechend aufbereitet und analysiert werden, sodass sie danach ins Core Datawarehouse übernommen werden. Dort werden sie dann mit den in den eingesetzten BI-Suites bereits vorhandenen mächtigen Mitteln zum Data- und Textmining bzw. Predictive Analytics analysiert. Grundsätzlich sind aber natürlich auch durchaus weitreichendere Integrationsmöglichkeiten denkbar.

(5.7) Technologiewahl und Umsetzung

ACID-Prinzipien machen relationale Datenbanken zu einem idealen Speicherort für wichtige Geschäftsdaten.. Dort hat jeder einzelne Datensatz einen hohen Stellenwert und dessen konsistente und dauerhafte Speicherung ist enorm wichtig.

Der Verzicht auf diese sehr harten Prinzipien und einige weitere Besonderheiten erlauben es NoSQL-Technologien enorme Datenmengen bei äußerst hoher Verfügbarkeit und geringem TCO (total cost of ownership) zu verwalten.

Der Preis dafür ist aber sehr oft ein Verzicht auf gewisse Garantien und Sicherheitseigenschaften welche bei der Verarbeitung von kritischen Geschäftsdaten unabdingbar sind.

Werden jedoch zum Beispiel Messdaten von Sensoren verarbeitet und analysiert, wo einzelne Messwerte in der Menge der Daten nicht entscheidend sind, können NoSQL-Systeme ihre Stärken ausspielen. Hier empfehlen sich spaltenorientierte Systeme wie Apache Cassandra als Mittel der Wahl für eher einfache, listenorientierte Datenstrukturen und dokumentenorientierte Systeme wie zum Beispiel mongoDB für komplexere Strukturen.[75]

Sollen große Mengen wichtiger Geschäftsdaten sicher und konsistent in dokumentenorientierter Form gespeichert werden, könnte MarkLogic einen Blick wert sein.

Graphenorientierte Datenbanken wie Neo4j können dann Einsatz finden, wenn bei den zu modellierenden Daten die Beziehungen zwischen Objekten im Mittelpunkt stehen.

In-Memory-Datenbanken wie SAP Hana oder Oracle 12c In-Memory würden sich aufgrund deren hybriden Ansatzes sowohl für die Verarbeitung von kritischen Geschäftsdaten, als auch für weniger wichtige Massendaten eigenen. Hier darf aber der Kostenfaktor nicht außer Acht gelassen werden.

Predictive Analytics und Data Mining fand bisher meist eher offline im Zuge von BI-Prozessen statt. Apache Spark könnte hier einen Wendepunkt darstellen und komplexe Abfragen auch direkt auf Basis eines Online-System ermöglichen.

Soll eine vorhandene BI-Lösung um Big Data Aspekte erweitert werden bieten die meisten Hersteller entsprechende Lösungen an, um eine bestmögliche Integration zu erreichen. Dies kann die Lernkurve entsprechend senken, da weiterhin mit vertrauten Tools gearbeitet werden kann. Eventuell können aber auch OpenSource-Produkte eine kostengünstige Alternative darstellen. Der Einsatz von Big Data-Technologien empfiehlt sich hier besonders im Staging-Bereich der BI-Prozesse.

(5.8) Empfehlungen für die öffentliche Verwaltung

Beim Aufsetzen von Big Data Projekten sollte auf die Abgrenzung zwischen Small Data und Big Data geachtet werden. Grundsätzlich sollte die Frage betrachtet werden, ob die anstehenden Anforderungen wirklich den Einsatz entsprechender Big Data Technologien rechtfertigen. Die in Kapitel (1) definierten vier Charakteristika von Big Data helfen dabei eine Anwendung als Big Data Anwendung zu definieren, wobei die Grenzen zwischen sogenannten Small und Big Data fließend sind.[76]

Bei den eingesetzten technischen Lösungen muss die Kompatibilität der verwendeten Werkzeuge beachtet werden. Bei Bedarf empfiehlt es sich daher zu prüfen, ob bestehende Infrastruktur um kompatible Big Data-Aspekte ergänzt werden kann, um bereits vorhandenes Know-How zu nutzen. Die Basiskomponente Big Data etwa beinhaltet SAP HANA, welches SAP essentielle Berechnungsschritte in die Datenbank auslagert, was zu einer Optimierung der Performance der Applikation führt.

Open Source ist die treibende Kraft für Interoperabilität. Open Source Software nimmt offene Standards auf und verbessert daher die Interoperabilität. Die Förderung einer gesunden Open Source Wirtschaft in Österreich könnte durch

[75] Vgl. [Java16], Seite 22ff
[76] Vgl. [BRZ15], Seite 8

strengere Regelungen in Bezug auf Projektimplementierungen für den öffentlichen Sektor unterstützt werden. Wann immer möglich, soll Open Source Sofware in der öffentlichen Verwaltung Closed Source Software vorgezogen werden. Sollten Closed Source Lösungen implementiert werden, sind zumindest Open Standards zu implementieren.

Es gibt auch zahlreiche Open Source Lösungen wie Hadoop, Spark, Hive, R, etc., welche lizenzfrei sind. Diese Open Source Applikationen werden häufig auch von Unternehmen vertrieben, wobei in diesem Fall Lizenzkosten anfallen. Der Vorteil liegt darin, dass Wartung und Funktionalität gewährleistet wird. Vor Allem im Bereich Big Data zahlt sich jedoch der Vergleich von Oen Source zu kommerziellen Lösungen aus.

Eine sorgsame Technologieauswahl ist nicht nur im Big Data Bereich von Bedeutung. Eine allgemein gültige Empfehlung für die Wahl einer bestimmten Big Data Technologie kann aufgrund des umfassenden Angebots nicht gegeben werden – diese Wahl hängt immer vom speziellen Einsatzzweck ab und bedingt mitunter vielleicht auch einer Kombination mehrerer Technologien. In Abschnitt (5.7) wurde aber versucht einen kurze Guideline zu geben was wofür einsetzbar ist.

Wie auch in allen vorhergehenden Kapiteln wird auch im Zusammenhang mit den technischen Aspekten eine klare Empfehlung in Richtung Kooperation ausgesprochen. Die Vielzahl der zur Verfügung stehenden Technologien sowie die vielfältigen Fertigkeiten die nötig sind, um aus den Massen von Daten wertvolle Informationen zu gewinnen, machen Big Data zu einem komplexen Thema. Ebenso wichtig wie die verwendeten Technologien ist die Beherrschung der zu Grunde liegenden statistischen Algorithmen und Modelle.

Bei Big Data Projekten sollte deshalb bereits in der Konzeptionsphase auf ExpertInnenen (interdisziplinäre Teams) zurückzugegriffen werden. Kooperationen innerhalb der Verwaltung sowie zwischen Verwaltung, Wirtschaft und Wissenschaft sollen im Big Data Bereich die Nachhaltigkeit (Ausbildungsmaßnahmen, Projekte, etc.) sichern.

(6) Ausblick / Empfehlung

Big Data beschreibt eine Palette an Technologien zur Analyse großer Datenbestände, die nicht immer neu sind, jedoch vor dem Hintergrund stetig wachsender Datenmengen – getrieben von zunehmender Durchdringung elektronischer Systeme und dessen Vernetzung durch das Internet – in teilweise stark geänderten Rahmenbedingungen zum Einsatz kommen. Ebenso wie Host Applikationen im Laufe der Zeit durch Data Warehouses abgelöst worden sind, werden Big Data Technologien in weiterer Folge die bestehenden Systeme nach und nach ablösen.

Um sich den Herausforderungen moderner Verwaltungen erfolgreich stellen und damit **Zukunftssicherheit** gewährleisten zu können, sollten diese **Technologien auch im öffentlichen Bereich** zum Einsatz kommen. Ein besonderer Vorteil in der Nutzung von Big Data liegt in der Möglichkeit, heute dort Zusammenhänge in Daten zu erkennen, wo dies bis dato nur mittels unverhältnismäßig hohem Aufwand möglich gewesen wäre. Durch die Anwendung von **statistischen und machine learning Algorithmen** können nun etwa Betrugsmuster erkannt oder Personalmanagementaspekte vermehrt optimiert werden. Dies wiederum hat **signifikante Auswirkungen auf die Kosten/Nutzen Aspekte des öffentlichen Bereichs**. All diese Aspekte bedürfen einer organisatorischen, rechtlichen und technologischen Anpassung in der öffentlichen Verwaltung, sodass alle Vorteile von Big Data optimal ausgenutzt werden können.

Um bei der Auswertung von umfassenden Datenbeständen **Privatsphäre und Anonymität** (Stichwort „**Gläserner Bürger**“) zu schützen, sind sowohl adäquate rechtliche Rahmenbedingungen, als auch technische Ansätze und Lösungen unerlässlich. Die Herausforderung besteht darin, technische Lösungen zu entwickeln, welche das Potenzial neuer Technologien innerhalb dieser notwendigen rechtlichen Schranken ausschöpfen. Zum Beispiel könnten technologieimmanente Ansätze angedacht werden, welche den NutzerInnen schon aus der Systemkonzeption heraus bestimmte Risikofaktoren aufzeigen und sie zum sicherheitsbewussten Handeln anleiten. Letzten Endes geht es darum, eine angemessene **Balance aus Sicherheit, Zuverlässigkeit, Rechtmäßigkeit und NutzerInnen-freundlichkeit** zu finden.

Die immense Bedeutung der Entwicklungen im Bereich Big Data wird mittlerweile auch von Seiten der politischen EntscheidungsträgerInnen wahrgenommen. Die **Digital Roadmap Austria**, als allumfassende Strategie, mit deren Hilfe die österreichische Verwaltung die aktuellen digitalen Herausforderungen und Chancen meistern soll, beschäftigt sich u.a. intensiv mit dem Thema. So sollen etwa im **Bildungsbereich** „Verstärkte Forschungstätigkeit (…) in den Bereichen Verarbeitung und Analyse von Daten, semantische Verarbeitung und kognitive Systeme erfolgen“.[77]

[77] https://www.digitalroadmap.gv.at/de/

Neben den sich ergebenden Möglichkeiten darf man aber auch **potenzielle Risiken** nicht außer Acht lassen. Auch in diesem Fall kann technischer Fortschritt beispielsweise zum Wegfall von Arbeitsplätzen in bestimmten Bereichen führen. Diesen Entwicklungen kann man aber durch entsprechende Maßnahmen entgegentreten, indem etwa gezielt neu entstehende Berufszweige und Berufsbilder unterstützt und gefördert werden.

Wissensvermittlung und Aufzeigen von gelungenen Anwendungen ist hier ausschlaggebend, denn die noch relative Neuartigkeit und Komplexität der Technologie sind für viele betroffene VerwaltungsmitarbeiterInnen Hemmfaktoren für die praktische Anwendung. Der Schlüssel liegt in der **Know-how Generierung und Vermittlung** bzgl. Anbieter, Verfahren und Werkzeugen durch innovative Lehrkonzepte. Notwendiges **Fachwissen** muss daher auch **innerhalb der Verwaltung durch Weiterbildungsangebote** oder die **Aufnahme fachkundiger MitarbeiterInnen** aufgebaut bzw. im Bedarfsfall zugekauft werden.[78]

Internationale Erfahrungen zeigen, dass **interdisziplinäre und verwaltungsübergreifende Teams** bei der Umsetzung von Big Data-Vorhaben besonders erfolgsversprechend sind. Auch eine verstärkte und womöglich institutionalisierte **Kooperation zwischen Verwaltung, Wissenschaft, Forschung und Wirtschaft** (Stickwort **Daten-Service-Ökosystem**), würde die weiteren Big Data Entwicklungen in Österreich maßgeblich unterstützen. Voraussetzung ist in jedem Fall eine interdisziplinäre Sicht- und Herangehensweise, sowohl für die Entwicklung von Big Data-Anwendungen, als auch für die Entwicklung der rechtlichen und gesellschaftlichen Rahmenbedingungen, in denen – bestenfalls tatsächlich auf der Technologie-Ebene wirksame – Risikominimierung beim Design von Big Data-Anwendungen integriert werden kann.[79]

Aus Verwaltungssicht sind auch jene Potenziale und Synergien, die sich aus der Verknüpfung von **Open (Government) Data**, **Cloud Services und Big Data** Technologien ergeben, von großer Bedeutung. Großes Potential besteht in der Verknüpfung von Daten. Datensilos, die in Form von nicht-standardisierten Daten verschiedener Anwendungsfelder bestehen, können und sollen geöffnet werden. Der freie Zugriff auf offene Verwaltungsdaten (**Open Government Data**), hat sich als **Innovationsmotor** sowohl für die **Wirtschaft** als auch für die **öffentliche Verwaltung** gezeigt, eine Vielzahl an neuen cloudbasierten Services hervorgebracht (**Linked Data/Apps**) und hat positive Effekte auf das Image der öffentlichen Verwaltung. Die Verbindung mit Big Data ist ein weiterer Schritt in die richtige Richtung, um die Position Österreich als Wissensstandort im IKT Bereich zu festigen.

Die **Digital Roadmap Austria** regt daher im Zusammenhang mit Big Data an bessere und **einfachere Lösungen für komplexe IKT-Systeme** zu finden: „Es ist davon auszugehen, dass in Zukunft Systeme, die in der Lage sind, auch bei Störungen und Veränderungen der Umwelt ihre grundlegende Organisationsweise zu erhalten, eine große Rolle spielen. Mit steigender Komplexität von

[78] https://www.digitalroadmap.gv.at/de/ „Die Einrichtung von Professuren für Data Science und Big Data soll die Ausbildung von wissenschaftlichen Datenfachkräften vorantreiben."
[79] Vgl. [BEST16], o.S.

Computersystemen erhöht sich auch die Herausforderung der Sicherstellung ihrer Korrektheit (…) Auf dem Gebiet komplexer IKT-Lösungen soll Forschung samt diesbezüglichem Know-how Transfer an relevante Stakeholder verstärkt werden“.[80]

Aus technologischer Sicht wird die Erarbeitung **flexiblerer Sicherheitslösungen** in Zukunft noch bedeutender. Diese sollen dabei helfen allfällige sich rasch ändernde **Bedrohungslandschaften, Systemeinbrüche, Datenabflüsse** und **Angriffe** zu erkennen und dadurch **Systemschäden** zu **minimieren**. Angesprochen sind dabei v.a. die Forschungsfelder Usable Security, Security Engineering, Identitätsmanagement und Verschlüsselungstechnologien, Pseudonymisierung/ Anonymisierung und Privacy Impact Assessment.

Als einer der ersten Staaten weltweit hat die australische Regierung eine **Big Data-Strategie** ausgearbeitet.[81] **Ziel** dieser Strategie ist es unter anderem, auf **Basis** des **Datenkapitals** die **Bereitstellung bestehender Services zu modernisieren, Innovationschancen zu kreieren** sowie **neue Ansätze für Services** zu schaffen.

Entsprechend der australischen Big Data Strategie ist laut **Digital Roadmap Austria** der Bundesregierung auch eine solche für Österreich geplant: „Die österreichische Bundesregierung wird die Ausarbeitung einer „**Big Data Strategie**“ unter Bezugnahme auf Chancen und Risiken prüfen.“[82] Das vorliegende Big Data Positionspapier kann hierfür als Grundlage herangezogen werden.

Zusammenfassend kann gesagt werden, dass der Einsatz von Big Data in der österreichischen Verwaltung, unter **Berücksichtigung des Spannungsverhältnisses zwischen Privacy-Aspekten und Innovation**, einen **wesentlichen Beitrag** zur **Steigerung der Lebensqualität,** zur **Absicherung** des **Wirtschaftsstandortes** und zur **Verbesserung** der **BürgerInnenorientierung** leisten kann.

[80] https://www.digitalroadmap.gv.at/de/
[81] Vgl. [AUSG13], o. S.
[82] https://www.digitalroadmap.gv.at/

Referenzen

[AUSG13]	Australian Government – Big Data Strategy Department of Finance and Deregulation (Australian Government Information Management Office): The Australian Public Service Big Data Strategy – Improved understanding through enhanced data-analytics capability, August 2013 http://www.finance.gov.au/sites/default/files/Big-Data-Strategy.pdf (15.03.2016)
[BEST16]	BEST-AT - Roadmap, Vertrauen rechtfertigen: sichere Systeme, 2016 (Univ. Wien, IDC, FH St. Pölten, OCG) http://best-at.ocg.at/?p=106 (30.3.2016)
[BIT13]	Management von Big Data Projekten - Leitfaden BITKOM, Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V.: Management von Big Data Projekten, 2013 http://www.bitkom.org/files/documents/LF_big_data2013_web.pdf (31.03.2016)
[BMVIT14]	#Big Data in #Austria Köhler, M./Meir-Huber, M.: Österreichische Potenziale und Best Practice für Big Data, April 2014 https://www.ffg.at/sites/default/files/allgemeine_downloads/thematische%20programme/IKT/big_data_in_austria.pdf (31.03.2016)
[BRZ13]	BRZ F&E-Projekt BI – NoSQL Technologien Höllwerth, Hans-Peter Dezember 2013
[BRZ14]	BRZ strategische Initiative E2 :Big Data Höllwerth, Hans-Peter

	Dezember 2014
[BRZ15]	BRZ White Paper – Big Data in der öffentlichen Verwaltung März 2015 https://www.brz.gv.at/presse/newsletter/2015-03-30_Big_Data_in_der_oeffentlichen_Verwaltung_v1.2_pub.pdf?4vufru (31.03.2016)
[BRZ16]	BRZ F&E-Projekt BI – NoSQL Technologien, und Java Magazin März 2016
[CLOUD12]	Cloud Positionspapier, 2012 http://reference.e-government.gv.at/fileadmin/_migrated/content_uploads/20120228_Cloud_Computing_Positionspapier_1.0.1.zip Rechtliche Checkliste zu Cloud Computing http://reference.e-government.gv.at/fileadmin/user_upload/Checkliste_Cloud_Computing_ChCC_1-0-0_20151014.pdf (31.03.2016)
[DatBank14]	Datenbanken 2014 Geisler, F.: Datenbanken, Grundlagen und Design. 5., aktualisierte und erweiterte Auflage. 2014
[DOAG15]	DOAG SOUG News Juni 2015
[Feiler/Fina13]	Big Data im Spannungsverhältnis zu datenschutzrechtlichen Grundsätzen Feiler, L./Fina, S.: Big Data im Spannungsverhältnis zu datenschutzrechtlichen Grundsätzen (österr. Datenschutzrecht) - Konflikt mit Grundsatz der Zweckbindung - wissenschaftliche Zwecke - Datensicherheitsmaßnahmen - Grenzen automatisierter Einzelentscheidungen - Durchsuchen großer Datenmengen zur Identifizierung von Verdächtigen (Suche nach Nadel im Heuhaufen), März 2013 http://www.medien-recht.ws/index.php?article_id=2213&id=1999 (31.03.2016)

[GUR14]	Open Data Now – The Big Promise of Open Data Gurin, J.: Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation, 2014 http://www.opendatanow.com/book-open-data-now/ (31.03.2016)
[IDC15]	International Data Corporation (IDC) FutureScape: Worldwide Big Data and Analytics 2016 Predictions https://www.idc.com/getdoc.jsp?containerId=259835 (5.4.2016)
[FRAF14]	Fraunhofer Fokus – Big Data Fraunhofer-Institut für offene Kommunikationssysteme Fokus: Big Data – Ungehobener Schatz oder Digitaler Albtraum, März 2014 https://www.oeffentliche-it.de/documents/10181/14412/Big+Data+ungehobene+Sch%C3%A4tze+oder+digitaler+Albtraum (31.03.2016)
[Java16]	Java Magazin 3.2016 Siprell, S./ Stroh, P.: Auswahl einer NoSQL-Lösung – Welche Datenbank passt zu mir? März 2016
[KIN11]	McKinsey Global Institute – Big Data: The next frontier for innovation, competition, and productivity, May 2011 http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation (31.03.2016)
[OECD15]	OECD Data-Driven Innovation - Big Data for Growth and Well-Being, October 2015 http://www.oecd.org/sti/data-driven-innovation-9789264229358-en.htm (7.4.2016)

[OECDDDI15]	OECD, DDI Reimsbach-Kounatze, C.: Data-driven innovation and the implications on jobs and skills, 26 May 2015 https://polcms.secure.europarl.europa.eu/cmsdata/upload/c5c2250c-d4c7-4f19-805e-78a9e6ab9cb3/DDI_Christian%20Reimsbach_26.05.2015.pdf (7.4.2016)
[OECDSTI15]	OECD STI Policy Paper Data-driven Innovation for Growth and Well-being What Implications for Governments and Businesses? http://www.oecd.org/sti/data-driven-innovation-9789264229358-en.htm (7.4.2016)
[SCHU14]	Doing Data Science Schutt, R./O'Neil, C.: Doing Data Science, 2014 https://it-ebooks24.com/ebook/doing-data-science (31.03.2016)
[SIE14]	Delivering on the Promise of Big Data Siegel, E.: Predictive Analytics: Delivering on the Promise of Big Data. IBM Government Analytics Forum, May 2014
[SmartGov15]	Smart Government. von Lucke, J. Prof.: Smart Government. Wie uns die intelligente Vernetzung zum Leitbild "Verwaltung 4.0" und einem smarten Regierungs- und Verwaltungshandeln führt. Whitepaper. Version vom 14.09.2015. https://www.zu.de/info-de/institute/togi/assets/pdf/ZU-150914-SmartGovernment-V1.pdf (7.4.2016)

Best Practice Beispiele

(1) Elektronische Volkszählung (Registerzählung)

Anwendungsbetreiber: Statistik Österreich
Kategorisierung: (Verwaltung/Privatwirtschaft): Verwaltung
Produktiv seit: 2011

_____

Ausgangssituation: Im Mai 2001 war man noch in 3,3 Millionen Haushalten mit der Beantwortung der Papierfragebögen beschäftigt und musste im Zuge dessen pro Familienmitglied bis zu 20 Fragen zu Geschlecht, Familien- und Bildungstand, zur beruflichen Tätigkeit, etc. beantworten. Es war notwendig, dass mehrere tausend Personen von Tür zu Tür gingen, die Fragebögen austeilten, bei der Beantwortung halfen und die Bögen wieder einsammelten. Die Gemeinden leiteten die Fragebögen anschließend an Statistik Austria weiter, wo sich mehr als 100 Personen mit deren Auswertung beschäftigten. Mit den dabei entstandenen 230 Tonnen an Fragebögen hätte man einen Papierstapel errichten können, der 27 Mal so hoch ist wie der Stephansdom oder fast so hoch wie der Großglockner. Die 46 Mio. A4-Blätter aneinandergereiht hätten eine Strecke von Wien bis zur australischen Ostküste ergeben.

_____

Kurzbeschreibung: Die elektronische Volkszählung zählt zu den ersten und wohl auch erfolgreichsten Projekten im Bereich der automationsgestützten Datenverarbeitung der letzten Jahre. Mit der Registerzählung wurden 2011 die Informationen erstmalig nicht von den BürgerInnen eingeholt, sondern den vorliegenden Verwaltungsregistern entnommen. Das Zentrale Melderegister, welches die Informationen der Haupt- und Nebenwohnsitze beinhaltet, bildet die Basis der Registerzählung. 14 weitere Registerbereiche (Daten des Hauptverbands der Sozialversicherungsträger, das Bildungsstandregister, das Gebäude- und Wohnungsregister, etc.) dienen der Vervollständigung und dem Abgleich der Daten. Bevor die Daten bei der Statistik Austria zusammenlaufen, werden diese anonymisiert und somit der Datenschutz gewährleistet.

Die Registerzählung umfasst die Themen Demographie, Bildung, Haushalte und Familien, Pendelverhalten und Erwerbsstatus, Arbeitsstätten, sowie Gebäude und Wohnungen. Im Gegensatz zu 2001 war 2011 nur noch ein 16-köpfiges Team notwendig, um die Daten zusammenzuführen, zu analysieren und auszuwerten. Die Ergebnisse sind vielfältig und vielfältig nutzbar. Beispielsweise regeln die Einwohnerzahlen der Länder und Gemeinden die Verteilung der öffentlichen Mittel. Die Arbeits- und Schulwege werden in der Pendlerstatistik ausgewertet und stehen für verkehrstechnische Planungen zur Verfügung. Die Ergebnisse der Gebäude- und Wohnungs-, sowie Arbeitsstättenzählung liefern wertvollen Informationen für wirtschafts- und sozialpolitische Entscheidungen. Ein denkbarer weiterer Entwicklungsschritt wäre die Registerzählung in Echtzeit, um aus dem automationsgestützten Datenverarbeitungsprozess mit zahlreichen Big Data Elementen eine „vollumfängliche“ Big Data Anwendung zu bauen.

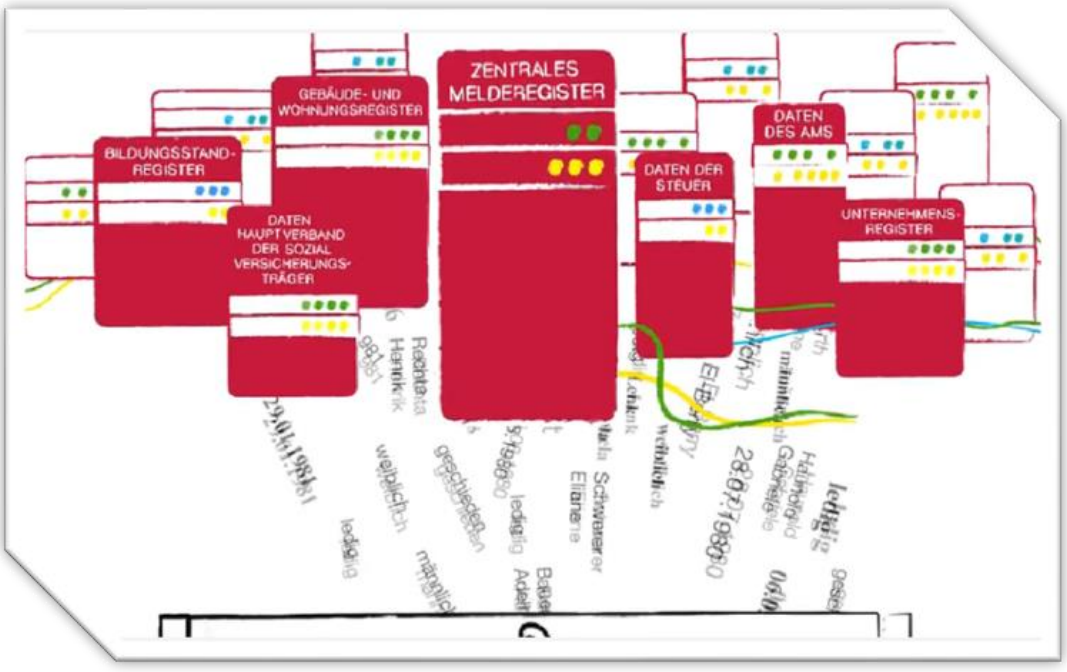
_____

Outcome/Wirkung:

- Entlastung der BürgerInnen
- Registerzählungen rascher und kostengünstiger als konventionelle Volkszählungen (9,9 Mio. Euro statt 72 Mio. Euro)
- Registerzählungen können öfter (alle 5 Jahre) durchgeführt werden und bieten somit die Möglichkeit auf „Fehlentwicklungen“ schneller reagieren zu können
- Einsparung von ca. 230 Tonnen (46 Mio. A4 Blätter) verglichen mit der Volkszählung
- Vielfältige Auswertungsmöglichkeiten als Basis für strategische Planungen (z. B. verkehrstechnische oder sozialpolitische Planungen)

Kontakt: Statistik Austria; info@statistik.gv.at
Web: http://www.statistik.at/web_de/frageboegen/registerzaehlung/index.html bzw. https://www.youtube.com/watch?v=4jXLuM-ta44

Abbildung 11 Schematische Darstellung Registerzählung



Quelle: Statistik Austria

(2) Kriminalitätsprävention

Anwendungsbetreiber: Bundesministerium für Inneres - Bundeskriminalamt
Kategorisierung: (Verwaltung/Privatwirtschaft): Verwaltung
Produktiv seit: 2011

Ausgangssituation: Ein wichtiges Instrument für die kriminalpolizeiliche Führung Österreichs stellen zuverlässige kriminalstatistische Auswertungen der Kriminalitätsentwicklung dar.

Kurzbeschreibung: Predictive Policing setzt sich mit dem Schlüsselfaktor der computergestützten Auswertung und Visualisierung von Daten auseinander. Diese Art der Datenverarbeitung ermöglicht eine Bewertung und Beurteilung der vorliegenden Informationen in Echtzeit.
Bei dieser Analyse werden die gesammelten Informationen von z.B. vergangenen Einbrüchen ausgewertet um Muster erkennbar zu machen und Prognosen aufstellen zu können. Anschließend werden diese Vorhersagen benutzt um präventive Maßnahmen wie z.B. verstärkte Präsenz durch Polizeistreifen in den berechneten Hotspots zu setzen.

Ähnliche präventive Anwendungen gibt es in Städten der USA und Großbritannien. Darüber hinaus gibt es inzwischen Pilotprojekte in Deutschland, der Schweiz und China.

Outcome/Wirkung:

- Effektivere Kriminalitätsprävention durch in Echtzeit erstellte Prognosen und Risikoabschätzungen
- Geringere Kriminalitätsfurcht in der Bevölkerung
- Kosteneinsparungen im Zusammenhang mit geringerer Kriminalität (Schaden an Leib und Leben, Sachschäden, Versicherungskosten,…)

Kontakt: Bundesministerium für Inneres Bundeskriminalamt (BMI-II-BK-SPOC@bmi.gv.at)

Web: http://www.bmi.gv.at

(3) Kriminalitätsbekämpfung

Anwendungsbetreiber: Bundesministerium für Inneres - Bundeskriminalamt
Kategorisierung: (Verwaltung/Privatwirtschaft): Verwaltung

Produktiv seit: 2015

Ausgangssituation: Kriminalitätsbekämpfung umfasst auch die Planung und Umsetzung von Maßnahmen für spezielle Großlagen, wie zB. (Terror) Anschläge. Eine Aufklärung kann nur unter dem Einsatz moderner Technologie, wie Big Data Methoden es sind, rasch und effektiv durchgeführt werden.

Kurzbeschreibung:
Ein Anwendungsbereich der Kriminalitätsbekämpfung mit Big Data ist die Aufklärung von Verbrechen. Hierbei steht die Datenanalyse Upload Plattform nicht nur vor der Herausforderung im Ereignisfall viele, unstrukturierte Daten in kurzer Zeit verarbeiten zu müssen, sondern auch vor dem Problem der großen, qualitativen Unterschiede zwischen den eingebrachten Daten. Die Auswertung und Visualisierung stellt dabei einen wesentlichen Faktor dar und trägt potentiell zur Aufklärung des Ereignisses und der Ausforschung der Täter bei.

Outcome/Wirkung:

- Effektivere Kriminalitätsbekämpfung durch schnelle Datensammlung und -auswertung
- Höheres Sicherheitsgefühl und damit höhere Lebensqualität der Bevölkerung.
- Positives Bild Österreichs im Ausland aufgrund gut ausgebauter Maßnahmen
- Kosteneinsparungen im Fall einer speziellen Großlage

Kontakt: Bundesministerium für Inneres –Bundeskriminalamt (BMI-II-BK-SPOC@bmi.gv.at).

Web: http://www.bmi.gv.at

(4) Gesundheit

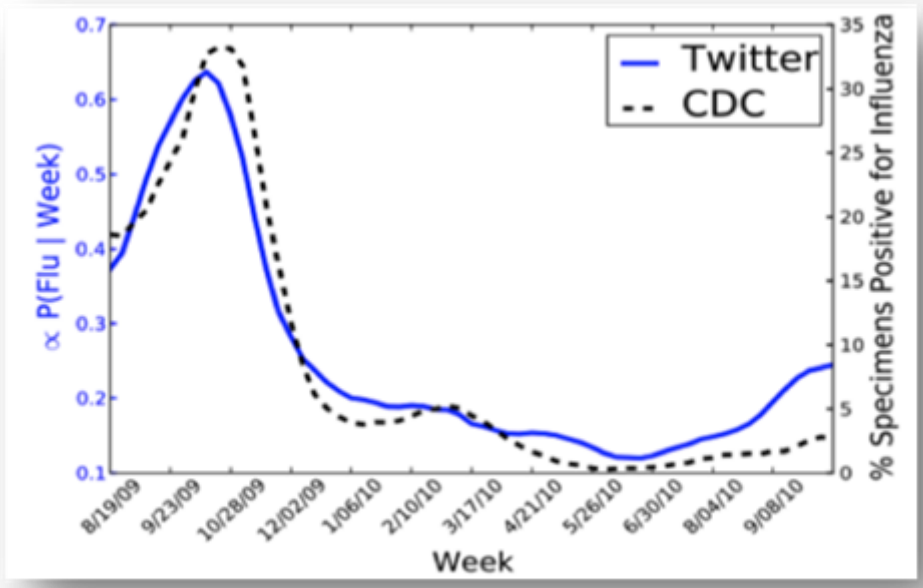
Anwendungsbetreiber: Gesundheitssektor
Kategorisierung: (Verwaltung/Privatwirtschaft): Verwaltung und Privatwirtschaft
Produktiv seit: 2009

Ausgangssituation: Besonders vielfältig sind die Einsatzmöglichkeiten im Gesundheitsbereich: Sie reichen von der Prävention (etwa der Erkennung von Epidemien) über die Diagnostik bis hin zur Therapie und Medikation.

Kurzbeschreibung:
Eine viel diskutierte Möglichkeit liegt in der Erkennung von Epidemien durch die Beobachtung von Aktivitäten in Social Media oder Suchmaschinen.

So konnte bereits in einer Untersuchung von August 2009 bis Oktober 2010 eine Korrelation zwischen der Erwähnung von Grippesymptomen in Twitter-Meldungen und dem tatsächlichen Auftreten von Grippeerkrankungen erkannt werden. Dies ist in nachfolgender Abbildung ersichtlich.[83]

Abbildung 12 Korrelation bei Grippeerkrankungen - Twitter vs. Messung durch CDC FluView

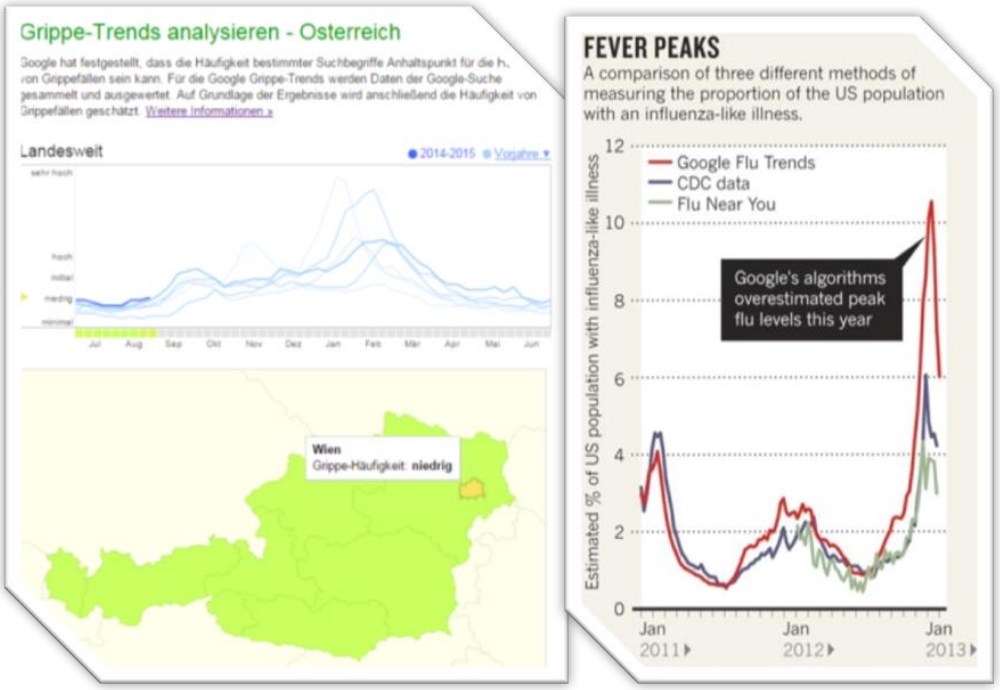


Quelle: Vgl. [BRZ15], Seite 11

[83] You Are What You Tweet: Analyzing Twitter for Public Health, M. J. Paul and M. Dredze, 2011. http://www.cs.jhu.edu/%7Empaul/files/2011.icwsm.twitter_health.pdf

Für mediales Aufsehen sorgt in diesem Zusammenhang immer wieder das Projekt „Google Flu Trends“, das den Anspruch stellt, auf Basis von Suchbegriffen Grippeepidemien vorauszusagen.[84] Allerdings zeigen die angewendeten Algorithmen auch die Grenzen solcher Voraussagen auf. So zeigt nachfolgend abgebildete Analyse, dass es auch deutliche Abweichungen geben kann.[85] Dennoch hat Google die diesbezüglichen Aktivitäten mittlerweile ausgeweitet und erstellt auch Analysen zum Dengue Fieber.

Abbildung 13 Google Grippe-Trend für Österreich und Grenzen von Google Flu



Quelle: Google

Outcome/Wirkung:

- Bessere Prävention (etwa der Erkennung von Epidemien), Diagnostik, Therapie und Medikation
- Neben der Erkennung von Epidemien und dem Potenzial, rechtzeitig Maßnahmen zu treffen, wird für diese und ähnliche Ansätze auch Potenzial in anderen Bereichen gesehen. So könnten etwa in sozialen Medien Trends hinsichtlich Drogen und deren Missbrauch identifiziert und diesen präventiv begegnet werden.
- Kosteneinsparungen im Gesundheitssektor bzw. höhere Kosteneffizienz: durch Bereitstellung vorhandener Daten werden unnötige Mehrfachuntersuchungen verhindert

Kontakt: Google Flu Trends http://www.google.org/flutrends/

[84] http://www.google.org/flutrends/
[85] http://www.nature.com/news/when-google-got-flu-wrong-1.12413

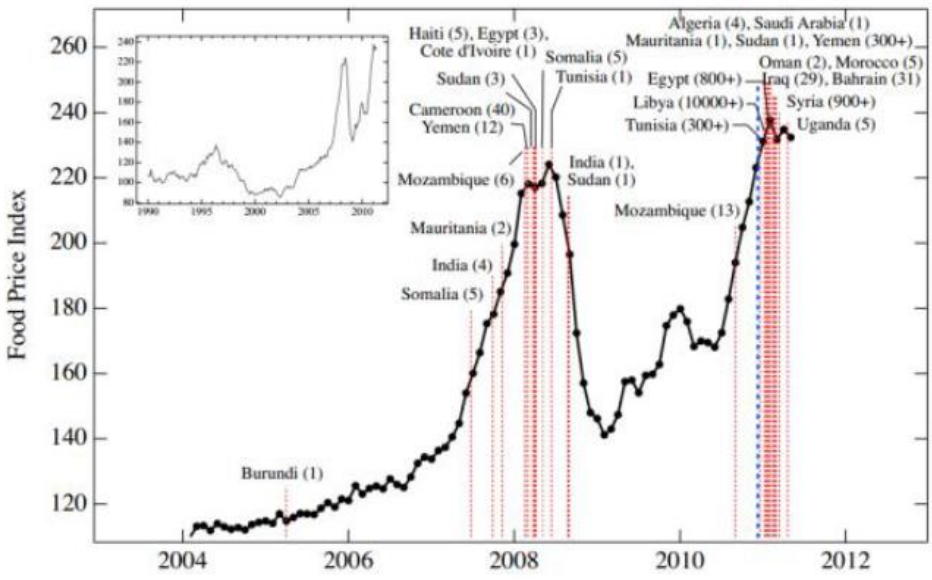
(5) Sozialer Bereich

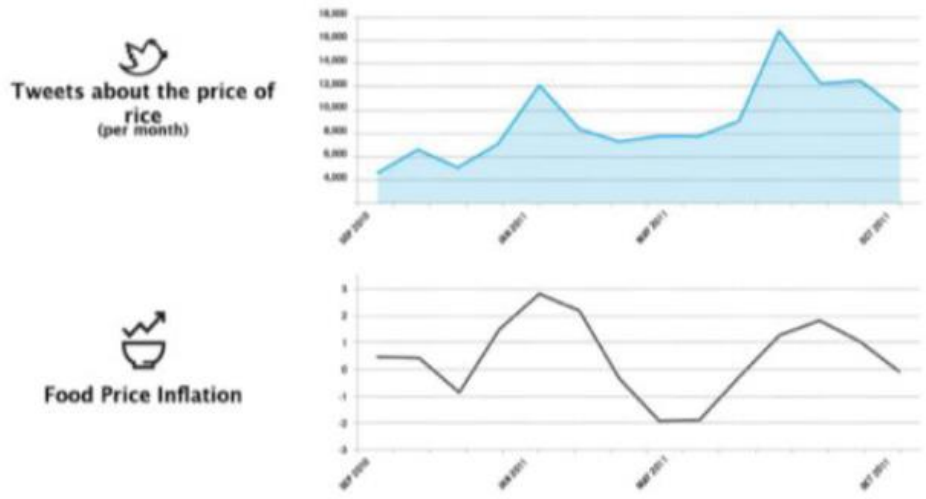
Anwendungsbetreiber: Sozialer Sektor
Kategorisierung: (Verwaltung/Privatwirtschaft): Verwaltung
Produktiv seit: 2011

Ausgangssituation:
Auch in der Bekämpfung sozialer Probleme wird Potenzial anhand von Big Data-Technologien und Predictive Analytics gesehen.

Kurzbeschreibung:
In einer Studie des New England Complex Systems Institute wurde 2011 ein Zusammenhang zwischen globalen Nahrungsmittelpreisen und gewaltsamen Protesten in Nordafrika und dem Mittleren Osten festgestellt, der in nachfolgender Darstellung dargestellt ist.

Abbildung 14 Korrelation Nahrungsmittelpreise und Proteste in Nordafrika und dem Mittleren Osten





Quelle: [BRZ15], Seite 13

Ein gemeinsames Projekt von United Nations Global Pulse und SAS zeigte außerdem einen Zusammenhang zwischen Twitter-Meldungen über den Preis von Reis und dem tatsächlichen Preis in Indonesien.[86]

Outcome/Wirkung:

- Zukünftig könnten zeitnahe Präventionsmaßnahmen im Falle hoher Nahrungsmittelpreise und in diesem Zusammenhang drohender gewaltsamer Ausschreitungen abgeleitet werden.[87]

Kontakt: New England Complex Systems Institute (NECSI) web@necsi.edu bzw. United Nations Global Pulse http://www.unglobalpulse.org/contact

[86] Finding proxy indicators (Global Pulse and SAS project) http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress

[87] Modeling & predicting food riots: http://necsi.edu/research/social/food_crises.pdf

(6) Crowd-basiertes Smart Parking Service

Anwendungsbetreiber: Parkbob GmbH
Kategorisierung: (Verwaltung/Privatwirtschaft): Privatwirtschaft
Produktiv seit: 2015

Ausgangssituation: Die Nutzung eines Autos in Städten ist oft eine mühsame Angelegenheit. Vor allem der letzte Teil einer Reise mit dem Pkw, das Parken kostet Zeit und Geld. Autofahrer (vor allem Pendler) verwenden im Stadtgebiet durchschnittlich 10 Minuten für die Parkplatzsuche. Die aktuelle Situation ergibt sich aus einem Informationsdefizit zwischen Autofahrern und der Verfügbarkeit von öffentlichen Parkflächen. Der Autofahrer sucht meist rein zufällig nach einer Parkfläche in der unmittelbaren Umgebung der Wunschadresse, vor allem in Bereichen, wo er auf wenig persönliche Erfahrung zurückgreifen kann. In vielen Fällen kennt er die lokalen Einschränkungen vor Ort wie Anrainerparken oder zeitliche Beschränkungen der Kurzparkzonen nicht. Dieses Problem wird derzeit auch von Navigationssystemen nicht gelöst. Studien bestätigen, dass bis zu 25 % des innerstädtischen Verkehrs von parkplatzsuchenden Verkehrsteilnehmern produziert werden, was neben der Lärm- und Feinstaubentwicklung auch zu einem nicht unerheblichen CO2-Ausstoß führt.

Kurzbeschreibung: In den letzten Jahren wurde eine große Anzahl von Vorhersagemodellen für den fließenden Verkehr entwickelt und zur Anwendung gebracht. Überraschenderweise gilt dies nicht für den ruhenden Verkehr. Dieser erweist sich bis auf wenige Ausnahmen als blinder Fleck im Bereich Mobilität. Parkbob hat es sich zur Aufgabe gemacht ein generisches Vorhersagemodell für den ruhenden Verkehr zu entwickeln und als Service anzubieten. Da es sich um ein multidimensionales Modell mit dutzenden Einflussfaktoren handelt, verwendet Parkbob einen Top-Down Ansatz mit folgenden Phasen:

- Entwicklung eines statischen Vorhersagemodells für Parken auf Basis Parktransaktionen (mehrere 100 Millionen Datensätze, Quellen OGD sowie Mobilitätsanbieter).
- Einbeziehung von Echtzeitinformation von fließendem Verkehr bis zu Wetter.
- Integration von Echtzeitdaten, die in einem dynamischen Sensornetzwerk erhoben werden.

Der Zündfunke für die Entwicklung waren die Erfolge in der automatischen Erkennung von Parkvorgängen. Im Juni 2014 wurde der erste Prototyp der Recognition Engine entwickelt, der ohne Nutzerinteraktion einen Parkvorgang (Ein- oder Ausparken) mit hoher Genauigkeit erkennt. Dabei kommt die sogenannte Sensor-Fusion Technologie zum Einsatz bei der Echtzeitdaten von mehreren Hardware-Sensoren eines Smartphones verwendet werden. In einem zweiten Schritt werden diese Daten durch Machine-Learning Prozesse weiterverarbeitet. Dieser Ansatz stellt eine Alternative zum Einbau von Sensorhardware in den Asphalt dar und vermeidet die im Vergleich sehr hohen Installations- sowie Wartungskosten.

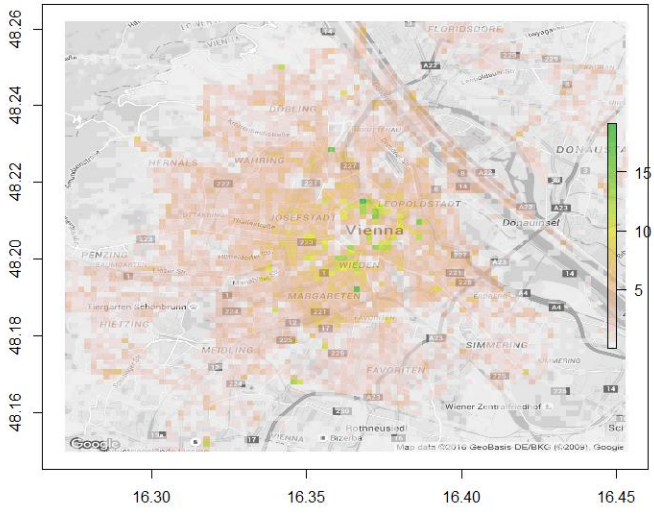
Erst die auf Big Data Ansätzen beruhende Veredelung der Echtzeitdaten erlauben die Schaffung einer Dienstleistung, die den Erwartungen der BürgerInnen entspricht. Gleichzeitig können die gewonnenen Daten sowohl für strategische als auch operative Planungs- und Steuerungsaufgaben der öffentlichen Hand verwendet werden (z.B. Bedarfsbasierte Einführung von Anrainerparkplätzen auf Basis harter Daten).

Outcome/Wirkung:

- Schnelles und einfacheres Finden von freien Parkplätzen im öffentlichen Raum.
- Reduktion der durchschnittlichen Parksuchzeit um 50%.
- Einsparung von CO2 und anderen Emissionen durch verminderten Parkplatzsuchverkehr (> 10.000 Tonnen CO2 nach Einführung Service mit 80.000 Nutzern).
- Steuerungsfunktion erlaubt bessere Verteilung von öffentlichem Parken und Parkgaragen.
- BürgerInnenfreundliche begleitende Maßnahme zur Abfederung von Maßnahmen der Reduktion von Parkraum.
- Datenorientierte Entscheidungsbasis für strategische und operative Planungsaufgaben für den ruhenden Verkehr.

Kontakt: Parkbob GmbH; hi@parkbob.com
Web: http://www.parkbob.com bzw. https://www.youtube.com/watch?v=pHtIF_0ZxGI

Abbildung 15 Darstellung Verteilung Parkfrequenz
Average daily distribution



Quelle: Parkbob

(7) Großstrafverfahren im Bereich Wirtschaftskriminalität

Anwendungstreiber: Strafverfahren gem. StPO, BMJ
Kategorisierung (Verwaltung/Privatwirtschaft): Verwaltung/Justiz
Produktiv seit: (Plan: Q4/2016 – Q1/2017)

_____

Ausgangssituation: Das Bundesministerium für Justiz hat mit der Wirtschafts- und Korruptionsstaatsanwaltschaft (WKStA) für Großstrafverfahren im Bereich Wirtschaftskriminalität eine zentrale Kompetenzstelle geschaffen, an der das technische und organisatorische Know-How zusammengefasst ist. Der Staatsanwaltschaft obliegt dabei gem. StPO die Leitung des Ermittlungsverfahrens und arbeitet dabei eng mit den Stellen des Bundesministeriums für Inneres (BK, BAK, LKA's) sowie externen Sachverständigen zusammen. In diesem Zusammenhang betrachtete Großstrafverfahren sind im Wesentlichen durch viele Verfahrensbeteiligte und/oder zumeist große Mengen an gesicherten (analogen und digitalen) Unterlagen gekennzeichnet. Dies führt zwangsläufig zu erheblich höherem Ermittlungsaufwand, da die Sichtung des Materials ohne zu Hilfenahme von IKT nur noch schwer zu bewerkstelligen ist.

_____

Kurzbeschreibung:
In einem typischen Großstrafverfahren werden durch BMI/Polizei erhebliche Mengen von Akten und sonstigen Unterlagen gesichert. Mit Fortschreiten der Digitalisierung in der Wirtschaft sind die in diesem Zusammenhang sichergestellten Daten zunehmend nur noch auf digitalen Datenträgern vorhanden, wobei Datenträger überwiegend Festplatten von Bürocomputern und Fileservern, aber auch Mobiltelefone und sonstige Speichermedien darstellen. Eine handelsübliche Festplatte hat heute durchschnittlich etwa 500GB Kapazität, auch Smartphones bringen es leicht auf Kapazitäten von 100GB und mehr. Es kommen so schnell Datenmengen von 100TB alleine im Ermittlungsverfahren zustande. Die Aufgabe von Polizei und Staatsanwaltschaften ist nun, diese Daten nach be- und entlastenden Fakten zu sichten. Vergegenwärtigt man sich, dass 100TB Daten ausgedruckt etwa einen Stapel Papier von London nach New-York ergeben, ist deutlich, dass diese Datenmengen nur noch mit Hilfe von Maschinen gesichtet werden können, beziehungsweise eine manuelle Sichtung bestenfalls den Charakter von Stichproben haben kann. Selbst wenn man annimmt, dass lediglich die darin hypothetisch enthaltenen 200 Email-Konten mit durchschnittlich 2500 Emails gesichtet werden müssen, wären das eine halbe Million Emails und damit zu viel um diese E-Mails alle manuell inhaltserfassend zu lesen.

Es wird daher an Big Data -Verfahren gearbeitet, die anfallenden Datenmengen nicht bloß lexikalisch durchsuchbar machen, sondern nach Möglichkeit auch den Schritt zum Data-Understanding mit Big Data (BD) erlauben.

Dazu existieren verschiedene Ansätze. Allen gemein ist, dass sie retrospektive BD Anwendungen sind, das heißt, sie beziehen sich immer auf einen (forensisch) gesicherten Datenbestand, welcher abhängig des Verlaufs des

Ermittlungsverfahrens (z. B. durch zusätzlich gesichertes Material aus weiteren Hausdurchsuchungen) weiter anwächst. Die Lebensdauer eines solchen „Big Data-Verfahrens“ erstreckt sich somit zumeist über mehrere Jahre und erfordert ein ständiges Anwenden rechenintensiver Analyseläufe. Eine besondere Herausforderung in diesem Zusammenhang ist allerdings, dass jedes neue Ermittlungsverfahren einen in der Regel völlig anders gearteten Datenbestand zur Grundlage hat; bereits der erste Schritt der Datenaufbereitung benötigt daher bereits BD-Methoden. In der Folge gilt es, aus den aufbereiteten Daten strafrechtlich relevante Tatsachen zu extrahieren. In diesen beiden Bereichen (Describe + Discover) werden zurzeit einige vielversprechende Softwarelösungen evaluiert sowie neue wissenschaftliche Ansätze erarbeitet.

Parallel dazu wird auch eine zentrale Hardwareplattform für die BD-Anwendungen spezifiziert, eventuell in Kooperation mit BMI und BMF, um eine gleichmäßige Auslastung zu erreichen, Synergieeffekte zu nutzen und somit allgemein Kosten einzusparen.

Outcome/Wirkung:

- Schnellere Analyse von strafrechtlich relevanten Daten und Erkenntnisse zu be- und entlastenden Fakten in Großverfahren.
- Anfallende Datenmengen nicht bloß lexikalisch durchsuchbar, sondern nach Möglichkeit auch ein Schritt in Richtung Data-Understanding.
- In Bereichen „Describe and Discover“ werden zurzeit vielversprechende Softwarelösungen evaluiert.

Kontakt **Bundesministerium für Justiz**
Web: https://www.justiz.gv.at/

Wenn auch Sie Ihre Big Data Lösung über den Best Practice Katalog präsentieren möchten, kontaktieren Sie bitte i11@bka.gv.at