Es hat sich einiges getan in der Welt, seit wir uns zu den Gewalt-Schule-Medien Seminaren getroffen haben. Vor allem der NSA-Skandal rund um den Whistleblower Ed Snowden hat einiges an Staub aufgewirbelt. Man hat zwar auch vorher immer vermutet, dass es bis zu einem gewissen Grad eine systematische Überwachung durch öffentliche Stellen gibt – allerdings nicht in diesem Umfang und Ausmaß. Mittlerweile hat sich aber auch hier die Aufregung wieder weitestgehend gelegt. Geändert hat sich allerdings nichts. Daten werden noch immer im gleichen Umfang abgegriffen und verwertet.

Ich möchte in meinen Ausführungen heute die Snowden- und NSA-Debatte nur hin und wieder streifen und mich eher auf die kommerzielle Seite von „Big Data“ konzentrieren, da dies meiner Meinung nach mindestens ebenso große Auswirkungen auf uns als Konsumenten und auch Menschen haben wird, wie die Überwachung durch staatliche Einrichtungen. Ich möchte euch also einen kurzen Überblick über das Thema „Big Data“ im Allgemeinen geben und auch darauf einzugehen, was das für jeden einzelnen bedeutet.

Was ist Big Data

Im eigentlichen Sinn ist Big Data nichts anderes als – große Überraschung – der Begriff für eine große Datenmenge. Dabei spielt es für den Begriff als solchen mal keine Rolle, woher diese Daten kommen. Im Normalfall haben sie aber eins gemeinsam: Diese Datenmengen werden aufgrund ihrer Größe nicht ständig ausgewertet, sondern nur, wenn es einen spezifischen Anlass dazu gibt. Paradebeispiel dafür ist eine Log-Datei. Im Normalfall wird man nicht täglich sämtliche Seiten kontrollieren, die in einem Unternehmen oder einer Schule aufgerufen wurden. Erst, wenn es einen wirklichen Anlassfall gibt, beginnt man mit der Suche und somit der Auswertung.

Prinzipiell geht es also darum, dass die Datenmenge auf der ganzen Welt immer schon groß war, immer größer wird – jetzt aber erstmal Möglichkeiten vorhanden sind, durch den Einsatz von Computern, diese Daten auch im wirklich großen Stil zu verarbeiten.

BigData kommt auch in der Strafverfolgung vor. In Deutschland wurde fünf Jahre nach einem Heckenschützen gefahndet, der auf Autobahnen wahllos auf vorbeifahrende Fahrzeuge schoss. Erst, als die Polizei anfing, an sieben neuralgischen Abschnitten Kameras zu installieren, und **jedes einzelne Kennzeichen** zu erfassen, konnten wahrscheinliche Tatorte, bevorzugte Strecken und vor allem immer wiederkehrende Kennzeichen identifiziert werden, die schlussendlich zur Ergreifung des Täters führten. Dabei half dann auch der Abgleich mit den Funkzellendaten des Handys usw. Von Datenschützern wurde diese Vorgehensweise scharf kritisiert, weil für die Ermittlungen auch unzählige Aufnahmen von Personen gemacht wurden, die nichts damit zu tun hatten. Hier ist es also eine Frage der Abwägung, ob es legitim ist, für die Suche nach der Nadel im Heuhaufen die ganzen Grashalme zu registrieren. Bei solchen Fällen von BigData ist es immer sehr schwierig, die Legitimität zu bewerten. Es geht um den Schutz von Menschenleben und im Normalfall kann man davon ausgehen, dass die Daten, sobald sie nicht mehr benötigt werden, wieder gelöscht werden.
Um solche „staatlichen“ Fälle von BigData wird es heute aber wie gesagt nicht primär gehen. Wir konzentrieren uns eher auf die kommerzielle Seite. Dafür wird es aber leider unerlässlich, dass wir uns kurz einen Exkurs in die Daten-Theorie antun. Wir müssen uns kurz ansehen, wie Daten erstellt, interpretiert und verwendet werden können. Diese Informationen werden anfangs noch etwas „unzusammenhängend“ erscheinen, werden aber schlussendlich zu einem großen Ganzen.

Exkurs Datentheorie

Daten: Maschinell erstellt oder vom Menschen geschaffen

Maschinen können sehr schnell sehr viele Daten erstellen. Manchmal ist das sogar ihre einzige Aufgabe. Maschinen können beispielsweise problemlos eine Liste aller Telefonanrufe erstellen, die in Österreich getätigt werden. Menschen können auch ganz gut Daten erstellen. Ein Tagebuch oder einen Blog führen, Statuseinträge auf Facebook verfassen, Fotos auf Flickr hochladen – alles kein Problem und auch nicht sonderlich spannend.

Daten: Strukturiert oder Unstrukturiert

Eine ganz wichtige Unterscheidung bei Daten betrifft die Struktur. Ein Beispiel für strukturierte Daten ist zum Beispiel das soeben angesprochene Telefon-Listing. Wer hat wen um wie viel Uhr für wie lange angerufen. Diese Daten können sehr gut von Maschinen erstellt – und natürlich auch wieder verarbeitet werden. Unstrukturierte Daten wären beispielsweise, worüber die beiden Gesprächspartner geredet haben. Diese Form von Daten kann nur sehr schwer von Maschinen bearbeitet werden. Bisher. Glaubten wir. Oder glauben wir noch immer. Das ist ein wenig kompliziert, weil wir zwar relativ gut einschätzen können, wie weit die Erkennungstechnologie mittlerweile fortgeschritten ist – allerdings leider noch immer nicht, in wie weit und „wie groß“ sie zum Einsatz kommt. Die reine Texterkennung (also einen Computer in einem Gespräch nach Schlüsselwörtern wie „Bombe“ suchen lassen) geht schon sehr gut. Sprache ist aber mehr als die gesprochenen Worte. Da geht es um Sprechtempo, Betonung, Frequenzmodulationen usw. Sarkasmus beispielsweise kann ein Rechner nur sehr schwer erkennen. Auch ist es natürlich nicht sinnvoll, wenn ein Suchalgorithmus Alarm schlägt, nur weil ein

Gesprächsteilnehmer erwähnt, dass auf der Party gestern eine BOMBENstimmung gewesen ist. Oder wenn ein Journalist dem Chefredakteur mitteilt, dass sein Artikel einschlagen wird wie eine Bombe. Wirklich abstrus wird das Ganze dann aber, wenn man sich eine Liste ansieht, die 2012 veröffentlicht wurde und Schlüsselwörter aufführte, die die amerikanische Homeland-Securtity hellhörig werden lässt. Darauf befinden sich nämlich neben „verständlichen“ Alarmwörtern auch Wörtern, von denen man nicht versteht, warum sie als beachtenswert eingestuft sind. Eine kleine Auswahl gefällig? Wenn sie folgende Wörter in Mails oder sozialen Medien verwenden, sind sie schon „beachtenswert“: Initiative, Team, Landwirtschaft, Welle, Kommunikation, Zittern und natürlich das schlimmste von allen -> Schnee!!! Natürlich muss man nicht davon ausgehen, dass die Verwendung eines dieser Wörter dazu führt, dass ein Cobra Team vorbeischaut und die Bruchfestigkeit der eigenen Wohnungstür testet. Es kann aber dazu führen, dass sich ein tiefgreifenderer Algorithmus mit ihrem Profil beschäftigt.

Data Mining

Der Begriff heißt wörtlich übersetzt so viel wie „Daten-Bergbau“. Im Grunde geht es darum, in einem großen Haufen von Daten systematisch statistische Methoden anzuwenden. Ziel ist es dabei, neue Muster zu erkennen. Es geht also – konträr zum eigentlichen Begriff – weniger um die Gewinnung von Daten, als um „Wissen“. Kehren wir zu dem Beispiel mit der Liste aller Telefonanrufe zurück. In Österreich werden jeden Tag sehr viele Telefonate geführt. Und es gab schon sehr viele Telefonate seit der Einführung des Telefons 1881 in Österreich (auch wenn damals mit Sicherheit noch nicht alle protokolliert wurden). Ein Data-Mining-Algorithmus könnte jetzt aber beispielsweise herausfinden, dass eine bestimmte Person eine andere bestimmte Person immer am selben Wochentag um dieselbe Zeit anruft. OK: Meist ist das zwar nur der wöchentliche Kontrollanruf von Mama, wenn der Junior dann doch mal zuhause ausgezogen ist – per se kann man das aber zu dem Zeitpunkt noch nicht wissen. Theoretisch könnte es ja auch der Kontroll-Anruf einer Terror-Schläfer-Zelle sein.
Wir sehen also, dass Data-Mining bei strukturierten Daten relativ problemlos abläuft. Da braucht man eigentlich nur die entsprechende Rechenpower und einen guten Algorithmus. Schwieriger wird es aber bei unstrukturierten Daten. Man könnte zwar schon alle Telefongespräche aufzeichnen – diese zu interpretieren ist aber – wie erwähnt – nicht so leicht.

Was wir von diesem Part mitnehmen sollen ist auf jeden Fall mal die Information, dass man auch aus großen Datenbeständen echtes Wissen herauskitzeln kann. Sei es nun maschinell, oder durch den Einsatz von Manpower…

Internet der Dinge

Anfangs gingen wir alle mit unseren Computern ins Internet. Irgendwann dann auch mit unseren Handys und Tablets. Aktuell sind es die Fernseher und DVD-Player. Bald schon sollen es aber auch der Kühlschrank, die Mikrowelle, die Waschmaschine und die Verkehrsampel sein. Das Internet der Dinge wird langsam Realität. Dabei geht es allerdings nicht darum, mit der Mikrowelle im Internet zu surfen, sondern eher um die Steuerung der Geräte. Auch dadurch entsteht eine riesig große Menge an Daten, da man jetzt schon davon ausgehen kann, dass bestimmte Nutzerdaten regelmäßig übertragen werden, um das Benutzerverhalten zu analysieren.

Kundenkarten

Es gibt da ein schönes Beispiel aus den USA. Dort bekam eine junge Frau Werbung von einem Supermarkt, die sich sehr stark auf Schwangerschafts- und Babyartikel konzentrierte. Daraufhin beschwerte sich der Vater des Mädchens, was denn das soll. Seine Tochter gehe noch zur Schule und er findet es nicht gut, wenn der Supermarkt hier für das Babykriegen bei so einem jungen Mädchen wirbt. Was er jedoch nicht wusste, der Supermarkt aber schon: das Mädchen war tatsächlich schwanger. Die Vergangenheit hatte nämlich gezeigt, dass schwangere Frauen in den ersten 20 Wochen vermehrt Spurenelemente kaufen und im zweiten Drittel verstärkt zu unparfümierten Körperlotionen greifen. So konnte man 25 Produkte identifizieren, die auf Schwangerschaften hindeuten. Nachdem das Mädchen eine Kundenkarte des Supermarkts besaß, konnten ihre veränderten Kaufgewohnheiten beobachtet werden, woraufhin der Supermarkt die entsprechende Werbung an sie verschickte.

Der ganz, ganz große Problemfall ist das jetzt natürlich nicht. Klar: für den werdenden Großvater hätte es vielleicht bessere Möglichkeiten gegeben, von der Schwangerschaft seiner Tochter zu erfahren, dass allerdings Supermärkte das Verkaufsverhalten ihrer Kunden analysieren ist nicht neu. Schon früher wussten Marktleiter, dass im Sommer bei schönem Wetter mehr Sachen für Picknicks und Ausflüge gekauft wurden, bei Kälte, Schnee und Regen mehr Sachen für einen gemütlichen Filmabend. Jetzt können diese Sachen halt besser analysiert werden…

Verbindung von großen Datenmengen

Ob man es will oder nicht: bewegt man sich als „Normaluser“ im Internet, hinterlässt man Spuren. Wir hinterlassen beispielsweise auf Facebook eine mehr als detaillierte Übersicht, wen wir kennen. Wir setzen Tweets ab und bewerten dabei Tagesgeschehen und Meinungen. Wir dokumentieren in Xing oder LinkedIn unsere Lebensläufe und beruflichen Werdegänge. Bei jeder Google-Suche hinterlassen wir Spuren allein dadurch, dass wir nach gewissen Dingen suchen. Wir sehen uns Youtube-Videos an – meist nicht nur ein sondern wir hinterlassen eine Spur, von welchen Videos wir zu welchen Videos weiterklicken. Wir lesen Online-Zeitungen und hinterlassen durch die Auswahl dieser Zeitungen vielleicht einen Rückschluss auf unsere politische Richtung. Jede diese Informationen für sich allein ist meistens kein Problem. Manche der Informationen können nur schwer uns als echten Personen zugeordnet werden (nur „einer Person vor diesem Gerät“), andere wiederum sind an sich ja harmlos. Das ich vor 12 Jahren mal einen Spanisch-Kurs besucht habe, ist wohl für niemanden interessant. Jetzt ist es aber endlich soweit, dass wir auf ein echtes Problem von diesem „Big Data“ kommen. Das Problem, was passiert, wenn Informationen aus unterschiedlichsten Quellen verbunden werden. Hier kommt jetzt der nächste Fachbegriff mit ins Spiel.

Data Warehouses

Ein Data-Warehouse – auf gut deutsch Datenlager – ist eine Datenbank, in der Daten aus unterschiedlichsten Quellen zusammengefasst werden. Ursprünglich war es ein Ansatz aus der Betriebswirtschaft, um Führungsebenen einen möglichst ganzheitlichen Blick zu verschaffen und somit die Entscheidungsfindung zu erleichtern.

Selbstverständlich ist das nicht die einzige Einsatzmöglichkeit für ein Data-Warehouse. Auch das PRISM-Projekt, das von Edward Snowden aufgedeckt wurde, ist im Großen und Ganzen ein Data-Warehouse, da dabei ja auch aus den verschiedensten Online-Kommunikationsformen ein komplettes Bild der überwachten Person gewonnen wurde.
Durch solche Warehouses wird das Thema Big Data meines Erachtens dann etwas gruselig. Und zwar aus folgenden Gründen:
Wenn wir alle Informationsquellen die ich in meinen Ausführungen schon erwähnt habe (und noch viele, viele mehr) miteinander in Verbindung bringe, entsteht ein ziemlich detailliertes Bild eines Menschen. Und dieses Gesamtbild kann in den Händen der falschen Personen zum Nachteil ausgelegt werden. Ich versuche, hier mal ein paar Beispiele zu beschreiben, die dadurch möglich WÜRDEN. Das bedeutet nicht, dass die bereits passieren.

- Sie bestellen sich irgendetwas im Internet, das nicht funktioniert. Darauf kommt ein großes Hin und Her mit dem Händler wegen der Garantie zustande. Andere Händler weigern sich daraufhin, ihnen etwas zu liefern, weil sie als Problemkunde eingestuft werden.

- Auch die Kreditvergabe von Banken läuft über verschiedenste Algorithmen. Diese „Credit Scores“ werden dabei von jeder Bank anders berechnet. Das kann dann schlussendlich dazu führen, dass man einen Kredit nicht bekommt, weil man in der falschen Straße wohnt, den falschen Vornamen hat usw.

- Ihr Kühlschrank registriert, was sich in ihm befindet und wie oft welche Waren gegessen und nachgefüllt werden. Diese Daten gleicht er mit ihrer Körperwaage ab. Krankenversicherungen weigern sich daraufhin, ihnen

eine Zusatzversicherung zu geben, weil sie als Risikokunde eingestuft werden müssen.

- Auf Facebook kommunizieren Sie mit einem Freund und beschreiben ihm, welche Probleme Sie mit ihrem Nachbarn haben. Später an diesem Tag registriert ihre Tankstellen-Kundenkarte, dass Sie sich einen Kanister, 10 Liter Benzin und einen Schokoriegel gekauft haben. Als Tags darauf das Haus des Nachbarn in Flammen steht, werden Sie zum Hauptverdächtigen. Dabei haben Sie sich nur Ihr Auto leergefahren…

Die vier Beispiele sollen nur ganz kleine Denkanstöße darstellen und ich kann und will auch gar nicht einschätzen, wie realistisch das heute, morgen oder übermorgen ist. Ich will auch gar nicht sagen, dass man keine sozialen Medien oder Kundenkarten oder intelligente Haushaltsgeräte mehr verwenden soll. Das Problem ist einfach mittlerweile, dass es nach den ganzen Datenaffären und Publikationen zu diesen Themen immer schwieriger wird, nicht paranoid zu sein …
Auch ist es ein großer Irrtum, zu glauben, dass es hilft, sich vor all diesen Dingen zu verschließen. Ganz im Gegenteil. Wenn jemand gar keine Spuren im Internet hinterlässt, macht er es zwar der kommerziellen Seite schwierig an seine Daten zu kommen – allerdings macht er sich dadurch auch wieder irgendwie verdächtig. Irgendwas hat er ja offensichtlich zu verbergen, weil man so gar nichts über ihn weiß...

Conclusio

Die Aufbereitung von großen, ja sehr großen Datenmengen wird in Zukunft immer häufiger passieren. Sei es für Geheimdienstarbeit, für die Strafverfolgung, im Business-Bereich und im Marketing oder auch in der Forschung und er Medizin. Das Problem dabei wird sich vor allem dort ergeben, wo unwissentlich Datenmengen in Warehouses zusammengefasst werden und vor allem auch dort, wo die Algorithmen, die zum Data-Mining verwendet werden nicht richtig funktionieren oder eine unzuverlässige Datenbasis verwenden. In all diesen Fällen kann es nämlich schlussendlich passieren, dass extreme Nachteile für einzelne Menschen geschaffen werden.

Aber es ist nicht so, dass es nur Nachteile gibt bzw. man diesem System hilflos ausgeliefert ist. BigData-Analysen helfen uns beispielsweise tagtäglich, wenn es um Navigationshilfen und Stauwarnungen geht. BigData hilft uns, wenn es um die Vorhersage von Grippe-Epidemien etc. geht. BigData hat auch schon Verbrechen und Terroranschläge vereitelt (auch wenn hier Datenschützer immer laut aufschreien).
Und auch jeder einzelne kann sich selbst bis zu einem gewissen Grad schützen. Was man auf Facebook, Twitter oder Instagramm postet kann ruhig mal etwas objektiv betrachtet und bewertet werden. Oft stellt man auch fest, dass die Angebote durch Kundenkarten gar nicht so wahnsinnig toll sind. Ich bekam neulich eine von einer Tankstelle angeboten. Wenn ich dort um 2.000 Euro getankt hätte, hätte ich zwei Handtücher um 9,90 bekommen. So der Hammer ist dieses Angebot jetzt auch nicht, als dass ich dafür mein Tank- und Einkaufsverhalten mitprotokolliert haben will.

Es ist also wieder einmal so, wie fast immer wenn es um Innovation geht. BigData kann je nach Einsatzgebiet zu unser aller Wohl eingesetzt werden, oder und ganz massiv in unseren Freiheiten und Bürgerrechten beschränken. Es geht auch bei dieser Entwicklung – so wie bei fast allen Revolutionen die wir als Gesellschaft schon erlebt haben und noch erleben werden – darum, was wir daraus machen…