Dr. Christian Folini

Big Data – Eine Einführung

Vortrag im Rahmen einer Veranstaltung des Datenschutz-Forums Schweiz am 6. Mai 2014

Ich freue mich sehr den Abend eröffnen zu dürfen. Ich freue mich deshalb, weil Sie - so nehme ich an - alles Datenschützer und Datenschützerinnen sind. Ich wäre nämlich auch gerne Datenschützer geworden. Im Laufe meines Studiums fiel mir aber auf, dass ein Eintauchen in die mittelalterlichen Geschichte kein typischer Einstieg in eine Datenschützerkarriere ist. Ich bin also statt dessen Geisteswissenschafter und Informatiker geworden. Das ist zwar auch nicht typisch, führt aber zu dankenswerten Einladungen wie dieser heute Abend, wo ich gebeten wurde, einige geistreiche und hoffentlich erfrischende Gedanken zum Thema Big Data zum Besten zu geben. Es handelt sich hier also um einen Eröffnungsvortrag und der soll auf keinen Fall langweilen. Da scheint es mir gelegen, einige Thesen etwas zuzuspitzen und die intellektuelle Vorsicht, die andernfalls im Vordergrund stehen sollte, etwas zurückzunehmen. Bitte sehen Sie es mir nach, wenn ich es damit etwas übertreibe.

Lassen Sie mich mit einem Gedankenexperiment beginnen. Wir Geisteswissenschaftler lieben ja Gedankenexperimente, weil sie es uns erlauben, an einem gesellschaftlichen Diskurs teilzunehmen, ohne den Elfenbeinturm verlassen zu müssen: Stellen Sie sich vor, sie würden jeden Artikel über Big Data, der in den letzten Jahren in der Schweiz gedruckt oder sonstwie publiziert worden ist, aus der Zeitung oder dem Magazin herausreissen, über den Drucker ausgeben oder vom Radio transkribieren. Und dann würden Sie alle diese verschiedenen Artikel stapeln. Um das Experiment zu erleichtern bitte ich sie, die Artikel vor Ihrem inneren Auge nicht feinsäuberlich aufzuschichten, sondern legen Sie die Papiere einfach mehr oder weniger bündig aufeinander. Da darf ab und zu auch mal ein Eselsohr herausschauen. Angesichts der Masse der Artikel über Big Data, die in den letzten Jahren hier und anderswo erschienen sind, müsste dieser Stapel eine beachtliche Höhe erreichen. So ein paar hundert Meter dürften es wohl sein. Und nun beginnen wir Daten

dazuzulegen: Wir legen die Profile der Autoren der Artikel hinzu. Dann die Privatadressen der Leserschaft der Artikel und etwas Informationen zu ihrer Umgebung. Wenn der Artikel online gelesen oder gehört wurde, dann haben wir natürlich die Möglichkeit, die Zugriffsdaten mitzuschreiben, einen Fingerprint des Browsers abzubilden und weitere Informationen zum Leseverhalten der Konsumenten zu erruieren. Den Ideen sind ja - so will es scheinen - keine Grenzen mehr gesetzt. Wenn wir all diese Daten auch ausdrucken, dann wächst unser Datenberg mit dem Tempo eines sehr aktiven Vulkans. Eigentlich dem Tempo eines explodierenden Vulkans. Ich behaupte, alle Schweizer Artikel zu Big Data und sämtliche Metadaten zur Produktion und zum Konsum dieser Publikationen aufeinander aufgeschichtet ergäbe einen Datenberg, der nur noch von einem Bergsteiger zu meistern wäre und das Matterhorn in den Schatten stellen würde.

Inhaltlich haben wir damit noch nichts gewonnen. Wer den Diskurs zu Big Data etwas verfolgt hat, dem dürfte aufgefallen sein, dass die meisten der Artikel und Interviews, Essays und Vorträge in etwa dieselben Thesen und Behauptungen aufstellen und wir können Annehmen, dass die Höhe unseres Berges grösser ist als die Vielfalt der vertretenen Meinungen. Wir dürften aber auch damit rechnen, dass der eine oder andere Artikel, sagen wir alle paar Dutzend Meter einer, eine originelle Frage aufwirft oder einen Aspekt beleuchtet, der den anderen Publikationen bis dahin entgangen ist. Ferner dass bei den hinzugefügten Metadaten ein paar originelle Erkenntnisse zu gewinnen wären. Erkenntnisse, die wir dann in der Zeitung unter den vermischten Meldungen publizieren könnten.

Für uns stapelnde Bergsteiger wäre es aber sehr schwierig, inhaltliche Fragen zu Big Data zu beantworten. Wir haben die ganzen Daten zu Big Data nämlich einfach mal locker stapelnd gesammelt: Unstrukturiert gesammelt! Wir haben grosszügig alle Erzeugnisse zum Schlüsselbegriff

Big Data ohne weitere Analyse zusammengetragen und in ihrem Umfeld alle Metadaten zusammengerafft, die einfach greifbar waren. Das wurde in den verschiedensten Datenformaten gespeichert, zusammengefügt und ohne den geringsten strukturierenden Formalisierungsprozess aufgetürmt. Und das ist das bahnbrechend neue an Big Data. Es ist uns schlicht und einfach egal was wir da gesammelt haben, Hauptsache wir haben es beisammen und können den unstrukturierten Datenbestand zukünftig irgendwann analysieren.

Ich halte die Strukturierung von Information für eine kulturgeschichtliche Errungenschaft. Es handelt sich dabei um ein menschliches Bedürfnis. Ein Bedürfnis, das natürlich viel älter ist als der Computer selbst. Ja die gegenseitige Bedingtheit von Wunsch nach Strukturierung und Computer geht sogar in die andere Richtung: Erst in den 1980er Jahren, als der Computer dank relationalen Datenbanken und der Tabellenkalkulation in der Lage war, die Strukturierung von Daten für den normalen Büromitarbeiter zu erleichtern, erst zu diesem Zeitpunkt war er reif für seinen Siegeszug und die rasche Durchdringung der Arbeitswelt. Machen wir uns nichts vor: die Firmen haben nicht Computer angeschafft, damit die Mitarbeiter schöne Photos von Bergen mit Büsis im Vordergrund online austauschen können. Es ging darum, Bestellungen in Tabellen zu erfassen und den Computer ausser dem Einkaufs- und dem Verkaufspreis nebenher noch die Bruttomarge berechnen zu lassen. Rasch kam die Kundendatenbank dazu und im Nu hatten die Computer die komplette Buchhaltung übernommen. Erst viel später kam das Internet als Massenphänomen auf und der Computer wurde mehr und mehr zu einem Kommunikationsmittel - wenn wir vom einen oder anderen bereits in den 80er Jahren in Wordperfect geschriebenen Geschäftsbrief absehen. Zentral für den Gebrauch des frühen PCs ist die Strukturiertheit der Daten. Unstrukturierte Datenbestände gab es eigentlich im Computer

drinnen noch gar nicht und es hätten auch die Möglichkeiten gefehlt diese etwas wild anmutenden Daten softwaremässig zu verarbeiten. Und die Software wiederum gab es nicht, weil erstens das Bedürfnis noch nicht da war und zweitens das Speichern von Daten viel zu teuer war, als dass man Daten ohne vorherigen Formalisierungsprozess hätte ablegen können. Denn Formalisierung und Strukturierung bedeutet ja immer auch Abstraktion und Reduktion und damit das Sparen von Platz und Geld.

Mit dem Aufkommen des Internets als Geschäftskanal und Plattform für die Freizeitgestaltung geschah etwas unerwartetes: Die Computer erhielten plötzlich die Möglichkeit, die Menschen beim Leben zu beobachten. Es gibt da diese App die einem hilft, Berge zu identifizieren und mit ihrem Namen zu benennen. Sie halten Ihr Telefon in Richtung Horizont und die App macht dannso ein Panoramabild und beschriftet jeden Berg. Diese App soll in der Schweiz sehr beliebt sein.

Worauf ich hinaus will ist dies: Es ist sehr schwierig schweizweit sämtliche Berg- und Papierstapelbesteigungen zu bemerken und mitzuzählen. Es ist für den Betreiber der Bergbeschriftunsapp aber sehr einfach, jeden Berg mitzuzählen, den die Schweizer und Schweizerinnen sich jeden Tag so anschauen. Hat man Zugang zu den Daten der besagten Bergbeschriftungsapp und zu den Daten von Facebook, so dürfte es ein leichtes sein, bei einer chronologischen Abfolge von Bergpanoramabeschriftung und Facebook-Photo-Posting auf ein alpines Motiv zu schliessen und davon Bergbesteigungen abzuleiten. Und ich bin sogar ziemlich sicher, dass das bereits jemand macht und jemand anderes wüsste, wie man mit diesen Daten Geld verdienen könnte.

Nun Facebook, schlaue Natels und Bergbeschriftungsapps gab es im Jahre 2000 noch nicht. Den Online-Datenverkehr konnte man aber damals bereits

erheben und auch der SAC, der Schweizer Alpenclub, hatte mindestens seit 1998 eine Homepage. Ich habe extra nachgeschaut. Ein Teil der Metadaten zu den Zugriffen auf eine Homepage war also bereits da - namentlich IP-Adresse, Zeitstempel, Navigationsverhalten auf der Homepage, Informationen zum Browser - aber man hat davon abgesehen, die Metadaten wirklich mitzuschreiben respektive sie konsequent zusammen abzuspeichern. Vielmehr musste man eine Vorauswahl treffen und diese gefilterten Daten dann strukturiert ablegen - auf verhältnismässig teuren Harddisks oder den etwas billigeren, aber sehr langsamen Magnetbändern. Meist hat man sie nach ein paar Monaten wieder gelöscht, weil sich ohnehin niemand für die Daten interessierte.

Die letzten Jahre nun schritt die Erosion der Preise bei den Speichermedien zügig voran. Man spricht da ja auch von Moore's Law, was zwar nicht dasselbe meint, aber das ökonomische Phänomen technisch begründet. Dazu kommen die Möglichkeiten der Cloud, so dass wir uns heute in der Lage befinden, dass eine Online-Firma für wenig Geld und mit minimalem Aufwand all das speichern kann, was sie möchte, es sei denn der Datenschutz funkt dazwischen.

Noch nicht gesagt ist damit, dass die Firma wissen muss, was sie mit den ganzen Daten anfangen will. Muss sie auch nicht. Die Daten werden in ihrer Rohform oder eben unstrukturiert abgelegt. Die Überlegung, was die Firma mit den Daten und wir mit unseren Zeitungsausschnitten und Metadatenberg zu Big Data anfangen möchten, diese Überlegung wird bequem in die Zukunft verschoben.

Das führt zu einer iterativen Verstärkung. Denn wenn ich nicht weiss, was ich zukünftig wissen möchte, dann werde ich sicherheitshalber alles speichern, wessen ich habhaft werden kann. Mir geht es da genauso:

Von Haus aus bin ich ja Historiker und als Historiker krankt man immer an einem Mangel an Quellen. Wenn ich also die Möglichkeit habe, mehr Archivmaterial anzulegen, dann tue ich das unbedingt. Ich kreiere damit Big Data. Daten in grossen Mengen. Mehr unstrukturierte Daten, als dass ich sie jemals werde strukturieren und intellektuell durchdringen können. Und ich glaube ich befinde mich da mit Google, Facebook und der NSA in illustrer Gesellschaft.

Meiner Meinung nach mussten zwei Bedingungen erfüllt sein, um Big Data entstehen zu lassen:

1. - Der Preis für den Speicherplatz musste extrem niedrig werden.
2. - Die Virtualisierung des Lebens oder eigentlich die computergestützte Durchdringung des menschlichen Lebens musste einen Stand erreichen, der es interessant machte, die ganzen Datenbestände überhaupt erst anzulegen und damit einen nennenswerten Ausschnitt des menschlichen Lebens erfassen zu können.

Das Resultat sind nun riesige Bestände von Daten. Alle möglichen Daten, mehr oder weniger wild gemischt und formlos aufeinander gestapelt. Verstehen Sie mich nicht falsch: Die Einzeldaten für sich besitzen natürlich eine innere Struktur. Aber diese Struktur zerfliesst, wenn der Big Data Stapel höher und höher wird und immer mehr verschiedene Datentypen zueinander gelegt werden: Wenn Zeitschriftenartikel, AHV-Nummern von Journalisten, IP-Adressen von Lesern, Bildschirmauflösungen von Natels und Virensignaturen von Notebooks nebeneinander abgelegt werden, dann wird es sehr schwierig, sich in diesem Wust noch zurecht zu finden.

Ist Ihnen aufgefallen, dass man vor 10 Jahren noch von Data Mining

gesprochen hat und inzwischen die Analyse hinter das Sammeln zurückgetreten ist? Ich behaupte, wir haben inzwischen so viele Daten, dass ein Mineur nicht mehr weiter kommt. Deshalb werden auch nur noch wenige Stellen unter dem Begriff Data Mining ausgeschrieben. Vielmehr bieten Universitäten Weiterbildungen in Data Science an. Es geht dabei nicht mehr länger darum, einzelne Informationsstücke, gleichsam Diamanten in der Mine, ans Tageslicht zu fördern, sondern mit wissenschaftlichen Methoden aus der linearen Algebra und vor allem der Statistik, der Leitwissenschaft der Datenbergsteiger, erstens die Datensammlungen zu sichten, zweitens minimale signifikate Beziehungen zu erkennen und drittens ökonomisch verwertbare Zusammenhänge knapp über der Signifikanzgrenze zu rapportieren.

Die Tiefe der Analyse verhält sich dabei umgekehrt proportional zur Datenmenge. Oder anders gesagt: Wo jeder in der Datenlawine schwimmt, da wird nicht mehr weiter getaucht; da versucht jeder und jede nur noch schwimmend an der Oberfläche zu bleiben - auch auf Empfehlung des SAC zum Verhalten bei Lawinenniedergängen.

Wir haben festgestellt, dass die Menge der Daten und vor allem der Metadaten so immens ist, dass alles zusammen einen sehr unstrukturierten Charakter annimmt. Der fehlende Formalisierungsprozess, der fehlende Review der Daten, führt zu einer sehr niedrigen Datenqualität. Die physikalische Welt und die elektronischen Sensoren als Eingang in die Onlinewelt der Datensammlungen besitzen zahlreiche Berührungspunkte, aber stets sind die Berührungen ungenau und die Zuordnungen unsicher. Von einem Endgerät auf einen Halter des Geräts zu schliessen scheint zwar vordergründig plausibel, in der Praxis ergeben sich im Einzelfall aber enorme Unsicherheiten - und sei es nur, dass ein Kind sich das Telephon des Vaters schnappt und damit heimlich im Internet einen

Film über Bergsteigen anschaut.

Wohl können sie den elektronischen Fingerprint von jedem aktiven oder schlafenden Natel notieren, das über eine der Rolltreppen im Bahnhof Zürich fährt. Aber können Sie vom Fehlen eines Fingerabdruckes aber auch darauf schliessen, dass eine bestimmte Person zu diesem Zeitpunkt nicht Rolltreppe im Bahnhof gefahren ist? Oder war der Akku im Eimer, das Telefon im Tram vergessen worden oder das Gerät so zwischen in einer Einkaufstüte verstaut, dass die Selbst-Identifikation des Gerätes nicht mehr sauber gelesen werden konnte? Verlass ist auf solche Daten also nur bedingt oder allenfalls im Rahmen einer bestimmten, klar umrissenen Fragestellung. Eine Fragestellung freilich, welche zum Zeitpunkt der Datensammlung aber noch gar nicht bestand.

Es ist natürlich unglaublich, dass wir diese Daten heute überhaupt kreieren oder zulassen, dass sie kreiert werden. Ein Sicherheitsexperte, ich glaube, es war Bruce Schneier, stellte vor Jahren die These auf, Datenspuren unserer Kommunikation im Internet brächten sehr ähnliche Eigenschaften mit sich, wie die Umweltverschmutzung:

1. - Sie werden beiläufig erzeugt
2. - Wir nehmen die Erzeugung der Daten in Kauf um dafür kleine Gefälligkeiten zurückzuerhalten (ancillary benefits lautet der englische Fachbegriff hiezu)
3. - Sind die Daten erst einmal erzeugt, entziehen sie sich unserem Zugriff und es ist extrem schwierig, sie zu verfolgen oder sogar zu löschen.

Zunächst hielt ich die These vom Vergleich der Internet-Verkehrsspuren und der Umweltverschmutzung für überrissen. Inzwischen denke ich aber,

dass da etwas Wahres dran ist.

Wir nehmen das beiläufige Hinterlassen von Datenspuren also gerne in Kauf. Ganz sicher verschwenden wir keinen Gedanken an die sich hinter unserem Rücken stapelnden Datensätze wenn wir dafür mit dem Bild eines Datenberges mit einem niedlichen Büsi davor belohnt werden. Unsere Bequemlichkeit untergräbt da jede Vorsicht und verführt uns zu dauerhafter Fahrlässigkeit. Wären Sie nun nicht alles Datenschützer und Datenschützerinnen, dann könnten wir ebenso liederlich wie der Grossteil der Bevölkerung sagen, dass wir nichts zu verbergen haben. Wir könnten uns in der Sicherheit wiegen, dass unser Surfverhalten keine Rückschlüsse auf unseren Gesundheitszustand zulässt, wir könnten mit Inbrunst behaupten, dass eine Bewegungskarte unseres Telefons uns nicht der automobilen Geschwindigkeitsübertretung überführen könnte und bedenkenlos einwilligen, dass die Krankkasse die Daten unserer Cumuluskarte auf eine ausgewogene Ernährung hin überprüft. Nun sind Sie aber Datenschützer und so nehme ich an, dass Sie das alles berufshalber interessiert. Es muss Sie aber zukünftig noch viel mehr interessieren, denn der Schweiz wird in diesem Zusammenhang eine Schlüsselrolle zukommen.

Diese These stammt nicht von mir. Vielmehr vertritt Franz Grüter, Mitbesitzer einer ganzen Reihe von immer grösseren Rechenzentren in der Schweiz diese These seit drei, vier Jahren mit grosser Vehemenz. Die These besagt, dass die Schweiz als sicheres Alpenland der natürliche Standort für digitale Berge, für Big Data also ist. Die Datenbestände werden immer grösser, die steigenden Netzkapazitäten machen sie mobil und der Trend geht dazu, sie auf politisch stabilen Boden zu stellen.

Die These wurde schon verbreitet, als Edward Snowden noch als Maulwurf

Sandburgen in Hawaii baute. In der Zwischenzeit hat er uns aber einen ansehnlichen Datenhaufen mit Informationen zur Erzeugung, Beschaffung und Überwachung von Big Data durch die NSA gezeigt. Und es will scheinen, dass der Zwang zur Datensammlung stärker ist als jede verfassungsgemässe Vorsicht in dieser Beziehung.

Wie länger Amerika sich aber dagegen sträubt, die immensen Fangarme seiner staatlichen Datenkrake zu stutzen; je deutlicher wird, dass die amerikanischen Datenfirmen und Cloudbetreiber nicht in der Lage sind, unsere Daten vor dem Zugriff des amerikanischen Geheimdienstes zu schützen - schon rechtlich, geschweige denn technisch - so lange Google, Amazon, Facebook und Co also nicht in der Lage sind, unsere Daten sicher aufzubewaren, so lange steigt das Bedürfnis nach einer seriösen Alternative.

Nun hat Herr Grüter erkannt, dass der traditionelle Schweizer Bankier, der sich früher diskret, seriös und mit einer unauffälligen Professionalität um ihm anvertraute Vermögen gekümmert hat, dass dieser Bankier dieselben, ich möchte sagen calvinistisch protestantischen Tugenden mitbrachte, wie sie auch der zukünftige Schweizer Datenverwahrer mitbringen muss. Dazu kommt ein Rechtsstaat, der sich nicht heimlich an den Daten vergreift, eine Infrastruktur die den Rechenzentren einen hochverfügbaren Betrieb garantiert und schliesslich eine politische Kultur, welche dem aussenstehenden Beobachter als langweilige Sozialpartnerschaft und monotone politische Stabilität erscheint.
Die Schweiz ist das Land und ich behaupte: das einzige Land, das die geforderten Standortqualitäten mitbringt und deshalb wird Big Data in zunehmendem Mass den Weg in unsere wachsenden Rechenzentren finden.

Von Ihnen, meine sehr geehrten Damen und Herren wird es abhängen, dass

der Schutz dieser Daten auf sicherem Grund verankert bleibt. Denn natürlich sind die festen Mauern unserer Rechenzentren und Reduit-Daten-Bunker nicht ohne Risse. Das neue Nachrichtendienstgesetz (NDG) und das revidierte Bundesgesetz zur Überwachung des Post- und Fernmeldeverkehrs (BÜPF) erscheinen als mögliche Breschen in diesem Schutz- und Verkaufskonzept. Aus dieser Perspektive kommt den beiden Gesetzesvorhaben entscheidende Bedeutung zu und Ihre geplante Verabschiedung im Parlament kommt in einem ungünstigen Moment, der Vorbehalte gegen eine Schweizer Datenhaltung aufkommen lassen könnnte.

Es ist aber nicht angebracht, dass ich mich hier zu sehr in Richtung des Datenschutzes bewege. Da kennen Sie sich besser aus. Lassen Sie mich vielmehr zum Matterhorn und dem daneben errichteten Datenberg zurückkehren und meine Thesen nochmals zusammenfassen:

Die Menge an unstrukturierten Daten hat unsere intellektuelle Kapazität überstiegen und die Fähigkeit der Analyse ist der Datenmenge nicht mehr länger gewachsen. Die Metadaten machen im Wust der Daten einen immer grösseren Anteil aus und so geht es je länger je weniger um das Finden von Dateneinzelstücken, inhaltlichen Fakten, sondern um die statistische Auswertung von Datenbeständen. In fahrlässiger Art und Weise hinterlassen wir bei Schritt und Tritt Datenspuren, die wir so wenig ungeschehen machen können, wie die Umweltverschmutzung, die durch unseren ökologischen Fussabdruck belegt wird. Und schliesslich: Die Besitzer der riesigen Datenbestände fürchten sich vor den Vereinigten Geheimdiensten von Amerika und über kurz oder lang werden unermesslich grosse Datenbestände in der Schweiz zu riesigen Bergen anwachsen, welche selbst die höchsten Alpengipfel überschatten werden. Die dazu nötigen Rechenzentren entstehen laufend in diesem Land.