MICHAEL SEEMANN

iRIGHTS media

DAS NEUE SPIEL

STRATEGIEN FÜR DIE WELT NACH DEM DIGITALEN KONTROLLVERLUST

Bildschirmauflösung und andere Daten an jede aufgerufene Website. In Summe, als verknüpftes Muster, verrät das, wer eine Seite besucht hat. Jede einzelne Information für sich ist harmlos. In der Zusammenführung ergibt sie ein Muster, das eindeutige Identifizierung ermöglicht, wie bei einem Fingerabdruck.

XKeyscore arbeitet mit dieser Verknüpfbarkeit von Daten. Das Programm verwandelt die riesigen Datenschätze, die die NSA zusammenträgt, aus totem gespeichertem Wissen in auswertbares Material; ähnlich wie wir es von Google kennen, nur mit mehr und präziseren Möglichkeiten. Die Suchen darin lassen sich beliebig einschränken und filtern, zum Beispiel nach bestimmten Schlagwörtern, Geschlecht, Uhrzeit und Ort der Kommunikation, hinsichtlich der verwendeten Sprache, ob Verschlüsselung eingesetzt wird oder nicht. Komplexe Abfragen wie „Zeige mir alle verschlüsselten Word-Dokumente in Iran“ können ohne weiteres generiert werden – oder auch „Gib mir alle Google-Suchanfragen der letzten 10 Tage nach ‚Islam‘ samt IP-Adresse, Sprache und verwendetem Browser in Deutschland und suche mir die Profile der betreffenden Nutzerinnen zusammen.“

Mit jedem zusätzlichen Datensatz wird einem anderen Datensatz neues Leben eingehaucht. Mit jeder Korrelation entstehen neue Such-Möglichkeiten, mit jeder Abfrage potenzielle neue Aussagen.

Big Data und das Ende der Anonymität

Das in der Debatte um die Digitalisierung herumgeisternde Schlagwort „Big Data“ bezeichnet im Grunde genau das beschriebene Prinzip: Erkenntnisgewinne durch die statistische Befragung großer Datenmengen. Empirische Forschung arbeitete bis vor kurzem ausschließlich mit kleinen, selbst zusammengesuchten Datenmengen, etwa aus der aufwendigen Befragung von circa tausend Leuten, um daraus ein repräsentatives Ergebnis abzuleiten. Seit einigen Jahren steht nun aber eine ganze Menge Daten zur Verfügung, die nicht aufwendig gesammelt werden müssen, sondern einfach „anfallen“; etwa die Verbindungsdaten von Handys, die Klickgewohnheiten auf Websites, die Angaben auf Facebook-Profilen oder die Bewegungsdaten von Menschen.

Chris Anderson, Herausgeber der Zeitschrift Wired, brachte Big Data einmal auf die Formel, es sei das „Ende der Theorie“ [13] – in Zukunft brauche niemand mehr eine Hypothese aufzustellen, stattdessen könne man die riesigen Datenmassen einfach direkt befragen. Das ist übertrieben. Dennoch verändert sich durch die Verfügbarkeit großer Datenmassen das wissenschaftliche Vorgehen. Daten können in einer Art Brainstormingphase korreliert und ausgehend davon

statistische Auffälligkeiten genauer unter die Lupe genommen werden. Dafür gibt es inzwischen jede Menge Beispiele.

Seinen Übersetzungsdienst „Translate“ hat das Unternehmen Google ohne große Kenntnis über Syntax und Grammatik so unterschiedlicher Sprachen wie Chinesisch und Arabisch entwickelt. Stattdessen konzentrierten sich die Google-Ingenieure auf die Suche nach genügend Texten, die in viele verschiedene Zielsprachen übersetzt worden sind. Aus diesem Rohmaterial „lernte“ Google Translate. Das funktioniert ausgehend von zehn oder hundert Texten nicht, aber bei einer Million Texten schon recht gut. Nach dem gleichen Prinzip erkennt Google auch, wie sich Grippeepidemien verbreiten. Die Kombination der entsprechenden Suchworte (zum Beispiel bestimmte Medikamente) mit dem Ort ihrer Abfrage erlaubt es, auf einer Landkarte in Echtzeit zu verfolgen, wohin die Grippe wandert.

Der Navigationsgeräte-Hersteller TomTom erkennt in Zusammenarbeit mit dem Mobilfunkprovider Vodafone Staus. Verändern sich die Standortdaten vieler Handys auf Autobahnen über einen längeren Zeitraum nur noch wenig, ist das ein sicheres Zeichen für zäh fließenden Verkehr. Per Mobilfunk kann das Navigationssystem dann „Stau“ auf den Geräten der TomTom-Kunden melden. TomTom versichert den Datenschützerinnen, dass die ausgewerteten Mobilfunkdaten für die Analyse natürlich anonymisiert werden. Das heißt, es werden keine Namen oder Telefonnummern in den Datensätzen verwendet. Doch wie anonym können Daten heute überhaupt sein?

Unter Wissenschaftlern ist Deanonymisierung inzwischen sowas wie eine Art Big-Data-Sport geworden. Am MIT in Cambridge extrahierten sie aus anonymisierten Mobilfunk-Zellendaten (ähnlich denen, mit denen TomTom arbeitet) nicht nur genaue Bewegungsprofile der einzelnen Handybesitzerinnen, sondern fanden heraus, dass lediglich vier Datenpunkte nötig waren, um diese mit 95-prozentiger Genauigkeit zu identifizieren. [14] Solche Datenpunkte können zum Beispiel Ortsdaten sein, wie Check-in-Daten auf Diensten wie Foursquare oder Facebook oder die Geo-Koordinaten in Fotos oder Tweets.

Die deanonymisierte Affäre

Vor der deanonymisierenden Macht verknüpfter Daten sind selbst Chefs von Geheimdiensten nicht sicher. General David Petraeus ist ein Mann, der sein Leben im Griff hat. Verheiratet, Kinder und erfolgreich im Job. Ein Mustersoldat: Seit 37 Jahren beim amerikanischen Militär, Vier-Sterne-General, ehemaliger

Kommandeur der amerikanischen Streitkräfte – erst im Irak, dann in Afghanistan, dann, nach dieser beispielhaften Karriere, freiwillig in den Ruhestand gegangen. Barack Obama persönlich hat ihn reaktiviert und auf den Chefsessel der CIA gesetzt.

Auch in Paula Broadwells Leben verläuft offensichtlich alles nach Plan. Sie hat selbst eine militärische Karriere hinter sich, unter anderem in einer Spezialeinheit. Sie ist verheiratet und gilt als *hockey mum*, als außerordentlich engagierte Mutter, die ihre Kinder jeden Morgen persönlich zum Schulbus bringt. Nebenher engagiert sie sich ehrenamtlich für Kriegsveteranen. Eine amerikanische Vorzeige-Superfrau. Paula Broadwells Interesse an Petraeus war zunächst ein journalistisches: „Der Mustersoldat" war der Arbeitstitel der Biografie, die sie über ihn schreiben wollte. Jahrelang begleitete sie ihn; auch in den Irak und nach Afghanistan, überall war sie mit dabei. Die Öffentlichkeit bekam nichts davon mit, dass sich die beiden auch abseits des Beruflichen näherkamen.

Eine Affäre mit einem amtierenden CIA-Chef geheim zu halten, ist nichts für Anfänger. Doch Broadwell passte gut auf. Nie machte sie den Fehler, Intimes mit Petraeus von ihrer persönlichen Handynummer oder E-Mail-Adresse aus zu kommunizieren. Die beiden legten einen gemeinsamen, anonymen E-Mail-Account bei einem freien Webmailer an. Wenn Broadwell Petraeus etwas mitteilen wollte, schrieb sie ihm eine E-Mail – doch statt sie abzuschicken, speicherte sie sie in den Entwürfen. Petraeus, der ebenfalls das Passwort zu dem Account hatte, konnte ihre Nachricht dort lesen und antworten. Broadwell war nie so dumm, sich von ihrem heimischen Internetanschluss aus in den Account einzuloggen. Sie nutzte ausschließlich öffentliche Internetzugänge, um mit Petraeus zu kommunizieren.

Jeder Internetanschluss ist identifiziert durch eine IP-Adresse. Sie ist einmalig im Internet, aber zunächst nicht direkt an eine Person gebunden. Doch der Internetprovider weiß, welche IP-Adressen welchen Kundinnen zugeordnet sind.

Als das FBI in einem Fall von Stalking in General Petraeus' Umfeld ermittelte, stieß es auf den anonymen E-Mail-Account. Mit den IP-Adressen, die auf den Account zugriffen, konnten die FBI-Agenten kaum etwas anfangen; dahinter befanden sich nur öffentliche Cafés in verschiedenen Städten sowie verschiedene Hotels. Die Hotels wurden Broadwell zum Verhängnis. Anhand der Check-in-Informationen aller Hotels, deren IP-Adressen auf das Konto zugegriffen hatten, konnten die FBI-Agenten die Hoteldaten untereinander abgleichen und analysieren. Gab es einen Namen, der in allen diesen Hotels zu

den fraglichen Zeiten eingecheckt war? Es genügten wenige übereinstimmende Datenpunkte, um zu Paula Broadwell zu führen. Das FBI wartete noch ein paar Monate, bis nach der Wiederwahl von Barack Obama, bevor es die Sache auffliegen ließ. General David Petraeus legte am 7. November 2012 sein Amt als CIA-Chef nieder – gestürzt über Datenanalyse.

Die Nadel im Big Heuhaufen

Auch die Datenbanktechnologie der NSA ist weit fortgeschritten. Sie beruht auf der Datenbanksoftware Accumulo, einer Weiterentwicklung von Googles Software Big Table. Mit ihr lassen sich Mustererkennungsanalysen bewerkstelligen. In großen Datenmassen können sich wiederholende Strukturen gefunden und erkannt werden.

Das Interessante dabei sind aber oft gar nicht die Muster, sondern die Abweichungen davon. Wo eine Nadel im Heuhaufen gesucht wird, ist normalerweise jeder Halm einer zu viel. Big Data dagegen mag Heu. Jeder Halm ist anders als alle anderen, deswegen will Big Data möglichst viele von ihnen kennenlernen. Denn je besser das Verständnis des Computers für Heu ist, desto schneller findet er darin die andersartige Nadel. Die NSA braucht darum eine Menge Kommunikationsdaten: Je besser der Computer versteht, was „normale Kommunikation“ (Heu) ist, desto eher findet er die „verdächtige Kommunikation“ (Nadel).

Es liegt außerdem nahe, sich mithilfe der Analyse der Kommunikations-Metadaten ein Bild davon zu machen, wer mit wem kommuniziert und auf welche Weise einzelne Gruppen untereinander vernetzt sind. Die sogenannte Graphen-Analyse ist heute ein gängiges Verfahren, um versteckte Zusammenhänge zwischen Personen oder Fakten in großen Datenmengen zu finden. Accumolo ist darauf spezialisiert.

Daten, die viele von uns sorglos veröffentlichen, weil sie keiner sensiblen Information verdächtig sind, erlauben Rückschlüsse auf durchaus sensible Daten. 2008 zeigten Studierende an der Technik-Uni MIT, dass sie mithilfe einer Analyse von Facebook-Freundschaften errechnen konnten, mit welcher Wahrscheinlichkeit jemand homosexuell ist. Die Idee des Projekts „Gaydar“ ist einfach: Manche Menschen haben ein engeres Verhältnis zu bestimmten Menschengruppen als andere Menschen. In jedem sozialen Netzwerk lassen sich also besonders eng vernetzte Gruppen erkennen – das nennt sich Clustering. Homosexuelle stehen oft in Kontakt zu anderen Homosexuellen. Lässt sich eine

Person einem Cluster mit vielen bekennenden Homosexuellen zuordnen, lässt sich davon mit einer gewissen Wahrscheinlichkeit auf ihre sexuelle Orientierung schließen. Die Genauigkeit lag im Fall des MIT-Experimentes bei 86 Prozent. [15]

Als die Idee des Datenschutzes geboren wurde, hatte man die Daten im Sinn, die nach damaligem Verständnis gelesen und entziffert werden konnten. Wenn bekannt ist, welche Daten von einem selbst existieren und was sie aussagen, kann man versuchen, den Zugriff darauf zu kontrollieren. Die „informationelle Selbstbestimmung“, wie sie das Bundesverfassungsgericht 1983 anerkannte, räumt jedem das Recht ein, über den Zugang zu seinen Daten und ihre Verwendung bestimmen zu dürfen. Auch wenn schon in den 1980ern bekannt war, dass sich Mess- und Analysemethoden kontinuierlich verbessern und dass es Techniken zur Verknüpfung von Daten gibt: Es sah noch aus, als ob Daten das bleiben würden, was sie zur Zeit der Speicherung waren. Wir glaubten noch zu wissen, dass eine Spur zu hinterlassen und sogar, einen „Write“ in eine Datenbanktabelle auszuführen ein endgültiger Vorgang sei, der das Feld seiner Interpretation von vornherein absteckt. Aber wir haben uns geirrt.

Der dritte Treiber des Kontrollverlusts besteht in den immer weiter wachsenden Möglichkeiten zur Verknüpfung von Datensätzen. Die Aussagefähigkeit von Daten wird damit in eine unbekannte Zukunft katapultiert. Weder wissen wir heute, was morgen Daten sein werden, noch wissen wir, was Daten von heute schon morgen aussagen können.

Wir haben die Kontrolle über die Daten also auf dreifache Weise verloren: Wir wissen nicht mehr, welche Daten zu welcher Zeit erhoben werden können, weil die ganze Welt durch die allgegenwärtige Verbreitung von Sensoren digitalisiert wird. Wir bestimmen nicht selbst, was mit diesen Daten geschieht, wo sie gespeichert werden, wo sie hinkopiert werden, wer darauf Zugriff hat. Und wir können nicht ermessen, welche Dinge diese Daten potenziell aussagen. Kurz: Daten, von denen wir nicht wussten, dass es sie gibt, finden Wege, die nicht vorgesehen waren, und offenbaren Dinge, auf die wir nie gekommen wären.