# Kumpels bei der Polizei

Data Mining, Rasterfahndung und der Überwachungsstaat

(216205, 165580141), (283864, 1134903170), (18153, 63245986), (321860, 89), (228991, 1134903170), (394149, 3), (183867, 433494437), (380014, 63245986), (305007, 144), (406727, 2) -- was ist an diesen Zahlenpaaren verdächtig?

Fragen dieser Art bewegen die Staatsgewalt der BRD spätestens seit Anfang der 1970er Jahre, als der damalige BKA-Chef Horst Herold anfing, seine Vision der Nutzung von Daten aller möglichen Sorten als „gesellschaftliches Diagnose- und Korrekturinstrument“ umzusetzen -- die Gesellschaft als Körper, der BKA-Chef als Arzt, der Kriminalität und Dissidenz als Krankheit diagnostiziert und heilt, will sagen eliminiert.

## Rasterfahndung und andere Fehlschläge

Herolds Vision selbst krankte damals an technischen Problemen: Bei weitem nicht genug Daten waren tatsächlich für seine Rechner zugänglich, und auch die Verfügbarkeit vieler Daten hätte angesichts der aus heutiger Sicht lächerlichen Leistung der Rechner nicht viel geholfen. Vor allem aber war die theoretische Fundierung schwach, denn das BKA war und ist eine Hochburg der Pseudowissenschaft Kriminologie, einer wirren Mischung von Versatzstücken aus Soziologie, Psychologie, Konspiratologie und viel Würze, nach Gusto des jeweiligen Gurus zusammengerührt zu einem meist reaktionären Brei.

Nimmt man große Datenbanken, empirische Befunde aus der Ermittlung und die „Erkenntnisse“ der Kriminologie zusammen, kommt die Rasterfahndung raus. Je nach Mischung von Ermittlungsergebnis und unfundierter Ideologie funktioniert sowas entweder gar nicht oder nur schlecht. Herold selbst konnte in der Frühzeit mit der Ergreifung von Rolf Heißler noch einen (seltenen) Erfolg der Rasterfahndung feiern, die angesichts der inzwischen traumhaften Datenfülle eigentlich viel aussichtsreichere Rasterfahndung nach Nineeleven hingegen siebte trotz Überprüfung von 8,3 Millionen Personen keinen einzigen Menschen heraus, den ein Gericht hätte verurteilen wollen.

## Überraschende Einsichten

Doch ist der Fortschritt unaufhaltsam -- die Rechner sind schnell, wir alle hinterlassen permanent tiefe Datenspuren, auf die sich der Staat immer weitgehenderen Zugriff schafft (erwähnt sei nur die Vorratsdatenspeicherung, vgl. RHZ 3/2006), was gemeinsam bewirkt, dass Ideologie durch Mathematik ersetzt werden kann. Die Wurzeln des mittlerweile als Data Mining firmierenden Ergebnisses einer solchen Substitution liegen im kommerziellen Bereich, in dem bereits seit den 1950er Jahren im Umfeld von Bonitätsprüfung oder Marketing versucht wurde, aus unübersichtlichen Datenhalden „wertvolle“ Information zu gewinnen. Dass aus Sicht der Mathematik kein wesentlicher Unterschied zwischen dem Finden von InteressentInnen für neue Bausparverträge und dem Finden von künftigen RevolutionärInnen besteht, sollte nicht überraschen.

Ob Bank oder Bullen: Das zu lösende Problem besteht darin, dass Menschen zwar viel besser als Rechner komplizierte Szenen überblicken oder eine vertraute Stimme im größten Lärm wiedererkennen können, mit großen, vieldimensionalen Datenmengen und am Schluss noch nichtlinearen Zusammenhängen aber gewaltige Schwierigkeiten haben. Es kann sich einfach niemand vorstellen, wie Punkte in einem tausenddimensionalen Raum verteilt sind und darin auch nur etwas so einfaches wie eine Linie erkennen -- für den Rechner ist zumindest das kein Problem.

Eine simple Illustration dessen ohne Star Trek-Referenzen: Als Paul Graham vor einigen Jahren den

ersten bayesianischen Spamfilter (das sind die mittlerweile üblichen Programme, die aufgrund der in einer Mail vorgefundenen Wörter entscheiden, ob es sich um Spam handelt) entwickelte, war er sehr überrascht, dass das „Wort“, das in seiner Mailsammlung am allerklarsten die Unterscheidung von Spam und „guten“ Mails erlaubte, FF0000 war. Im Nachhinein ist der Grund leicht zu erkennen -- die Zeichenfolge ist eine der Möglichkeiten, in HTML roten Text zu erzeugen --, aber a priori hätte das wohl nicht mal der/die Spamgeplagteste erwartet.

Geht es nun um Kommunistenhatz statt Spamfilterung, macht nur die komplexere Datenbasis das Problem schwieriger -- statt einer Verteilung von Worthäufigkeiten muss die Staatsgewalt nämlich neben ihren gewohnten Delinquenten- und Verdächtigendaten aus INPOL und Co etwa Sammlungen von Kommunikations- und Konsumprofilen oder Kreditkarten- und Bewegegungsdaten haben und modellieren.

## Überwacht oder nicht

Bei der Klassifikation von Menschen kann man dabei zwei Wege beschreiten: Entweder, man nimmt bekannte StaatsfeindInnen und lässt den Rechner die Merkmale herausfiltern, die diese am deutlichsten von Nicht-StaatsfeindInnen unterscheiden -- so etwas heißt „supervised“ oder „überwacht“. Im Gegensatz zur klassischen Rasterfahndung wird der Rechner dabei nur von den Daten geleitet, was die Zielgenauigkeit der Suche massiv verbessern dürfte. Das ist nicht nur deshalb zu erwarten, weil Überraschungseffekte des Typs FF0000 in komplexeren Datensammlungen die Regel und nicht die Ausnahme sind, sondern wird auch von der Erfahrung der privaten Miner nahegelegt, deren Programme Kreditwürdigkeit oder Konsumbereitschaft ihrer KundInnen schneller, billiger und meist auch besser beurteilen als ihre SachbearbeiterInnen oder Marketingheiner.

Die überwachten Verfahren haben aus Sicht der Herold'schen Vision eine große Schwäche: Die „Krankheit“ muss schon ausgebrochen sein, man muss z.B. schon Staatsfeinde kennen, bevor man sie diagnostizieren kann. „Präventive Sozialhygiene“, wie sie Herold und seinen Erben vorschwebt(e), braucht ein Frühwarnsystem, das Entwicklungen wahrnimmt, bevor sie in Staatsgefährdung oder Kriminalität münden. Dazu könnte dann die nicht-überwachte Klassifikation dienen, bei der der Rechner einfach mal aufs Geratewohl Gruppen innerhalb der Datensätze bildet, in der Hoffnung, dabei auf interessante Strukturen („Parallelgesellschaften“) zu stoßen, die einer verschärften Kontrolle oder Repression bedürfen. Es ist klar, dass diese Verfahren um Größenordnungen aufwändiger sind, aber auch sie werden bereits kommerziell etwa zum Aufspüren neuer Absatzmöglichkeiten eingesetzt.

## Realität

Während Data Mining im privaten Bereich Standard ist und Unternehmen wie der österreichische Mobilfunker tele.ring ihren marktanteiligen Erfolg öffentlich auf besonders geschickte Nutzung ihrer Kundendaten zurückführen und Läden wie Google und Payback praktisch nur vom Verkaufen von Rohstoffen oder Fertigprodukten im Bereich Data Mining leben, sind die entsprechenden Möglichkeiten auf staatlicher Seite vorerst noch eher beschränkt. Einfache Anwendungen wie etwa die Bekämpfung von, nun, Betrug bei Transferleistungen (BAföG, ALG II) finden zwar statt, aber die große Datenkonzentration, ohne die Data Mining keinen Spaß macht, unterblieb in der BRD bislang, es sei denn, irgendwer hätte weit hinter den Kulissen den großen Hattrick gelandet.

Auch in den USA, in denen nach Nineeleven rechte Intriganten vom Schlage des Iran-Contra-Masterminds John Poindexter eine solche Datensammlung unter dem wirklich extrem undiplomatischen Arbeitstitel „Total Information Awareness“ auf den Weg hatten bringen wollen, ist wohl von einer Datensammlung, die reich genug ist, um etwa präventive Analyse lohnend schein zu lassen, nicht viel zu sehen, auch wenn die NSA den deutschen Behörden wohl etliche Nasenlängen voraus ist, vor allem, weil sie sich als Geheimdienst liberal bei Finanzdienstleistern wie SWIFT und den diversen Telekommunikationsunternehmen bedienen konnte und gleichzeitig über entscheidend mehr einschlägigen Sachverstand verfügt als die Nachfolgeorganisation der NS-Aufklärung „Fremde Heere Ost“ oder deren Geschwister.

Die Polizei, die im BKA wenigstens potenziell über ausreichend Sachverstand verfügt, wird indes bei allem alltäglichen Rechtsbruch doch noch vom Datenschutzrecht gezähmt. Nach wie vor gilt bei staatlich erhobenen Daten eine Zweckbindung, und trotz

aller Versuche, diese auszuhebeln, ist fürs nächste Jahr wohl nicht damit zu rechnen, dass das BKA etwa Telekommunikations- und Toll-Collect-Daten zum freien Zugriff bei sich sammeln darf. Stattdessen darf die Polzei Daten dieser Art -- wenn denn überhaupt -- nur bei Vorliegen eines konkreten Verdachts abfragen. Auch der Zugriff auf Kontodaten ist zwar bequem, aber doch nicht miningtauglich.

Dennoch betreibt das BKA natürlich Data Mining, etwa gemeinsam mit dem Schweizer Unternehmen semantic system. Der große Spaß ist das zwar nicht, aber bereits die Daten, die im BKA-eigenen INPOL-System lagern, halten viele Schätze bereit. Ihre Zweckbindung ist de facto ohnehin weitgehend aufgehoben, ohne dass dies bisher zu einem großen Aufschrei geführt hätte.

In der Tat war in der ursprünglichen Planung zur Überarbeitung von INPOL in den 90er Jahren von einem „dispositiven" Teil von INPOL-neu die Rede gewesen (im Gegensatz zum „operativen", bei dem einfach nur Fragen wie „Kennen wir einen Karl Marx?" beantwortet werden). Dieser dispositive Teil hätte Fragen wie „Was hat diese Handtasche mit diesem Haus zu tun?" durch etwas wie „Die Handtasche wurde von Herrn X entwendet, der der Bruder von Frau Y ist, die die Bank von Herrn Z augeraubt hat, die wiederum dem Eigentümer des Hauses einen Kredit gegeben hat" beantworten sollen. Angeblich soll dieser dispositive Teil schon weitgehend funktioniert haben, soll dann aber der Ausbootung von T-Systems bei der INPOL-Entwicklung zum Opfer gefallen sein. Die Wahrheit dazu wissen aber wohl nur wenige innerhalb des BKA und des Bundesinnenministeriums -- an dispositiven Komponenten des „neuen" INPOL-neu wird aber mit Sicherheit weiter gearbeitet.

Ein anderes Stichwort in diesem Zusammenhang ist ViCLAS, das Violent Crime Linkage Analysis System, bei dem es allerdings weniger um DissidentInnen als um SerientäterInnen geht -- das Ganze ist ein System wie aus dem US-Profilerkrimi (auch wenn es in Kanada entwickelt wurde), gedacht zum Auffinden der verborgenen Verbindungen zwischen Gewaltverbrechen. Aufgrund seiner eher starren Architektur gehört es aber nur am Rande hierher.

Dazu gibt es noch allerlei kleinere Versuche mit Data Mining, auch auf Länderebene. Solange aber nicht mal wieder per Rasterfahndung die große Datenbonanza bei den Polizeien anfällt, leiden sie alle vorläufig noch unter dem, was in der Mining-Community als „data sparseness problem" bekannt ist: ohne die Daten der Kundenkarten, Mailverkehr, Stromrechnung und Bewegungsprofil läuft die ganze Mathematik weitgehend ins Leere, und häufig wird bei der Minerei kaum mehr als „alle, die in unserer Datenbank sind, hatten mal Kontakt mit der Polizei" herauskommen. Allein das Leiden an diesem Umstand mag viel von der ungezügelten Gier auf mehr Daten erklären, die Polizei, Dienste und Regierung derzeit an den Tag legen.

## Was tun?

Data Mining ist Herrschaftstechnologie: Sie hilft den Herrschenden, ihre Untertanen ebenso wie die weniger erwünschten Effekte ihrer Herrschaft zu kontrollieren. Dass sie staatlicherseits im Augenblick noch recht eingeschränkt eingesetzt wird, heißt nicht, dass dies beim eingeschlagenen Weg in eine immer autoritärere Gesellschaft so bleibt. Ohnehin ist sicher, dass eine Krise, die auch nur annährend herrschaftsgefährend scheint, einen massiven Einsatz von Data Mining zur Unterdrückung von Dissidenz nach sich ziehen wird.

Die sich aus diesen Einsichten ergebenden Konsequenzen dürften schon fast langweilen:

- Datenspuren vermeiden -- der Hit am data mining ist, dass potenziell noch das kleinste Datenfitzelchen aus einem Meer von uninteressantem „Rauschen" herausidentifiziert werden kann. Der schlichte Kauf einer Bahnfahrkarte oder auch deren Nicht-Kauf kann ein relevantes Datum sein. Deshalb: Im Zweifel für die Anonymität.

- Record linking erschweren -- Data Miner kämpfen grundsätzlich mit dem Problem, ob zwei Datensätze auf dieselbe Person oder Sache bezogen sind. Ist „Eberhard G. Knüller" mit „Eberhart-Günther Knueller" identisch? Kleine Verschreiber oder Undeutlichkeiten helfen viel und sind im Zweifel ein Versehen. Umgekehrt sind abgesicherte Zahlen (Personalausweis-, Rentenversicherungs-, Kreditkartennummer) der Traum von RecordlinkerInnen und demnach wo immer möglich zu vermeiden.

- Uphold the law -- so weh es tut, mit Teilen der FDP auf einer Seite zu stehen: laut ständiger Rechtssprechung des Bundesverfassungsgerichts braucht Demokratie Dissidenz, und deswegen hat die informationelle Selbstbestimmung Verfassungsrang. Klassenjustiz hin oder her, die Widersprüche, die die Herrschenden untereinander haben, können helfen, den permanenten Verfassungsbruch durch Polizei und Dienste ebenso zu beschränken wie die immer dreister werdende Demontage der informationellen Selbstbestimmung durch die RepräsentantInnen des Volkes. Daten, die jetzt gelöscht werden, können später nicht für den großen Schlag verwendet werden.

Oh, ach ja: Eine Lösung des Rätsels am Anfang wäre, dass die zweiten Teile der Paare alles Fibonacci-Zahlen sind.

---

## Nachtrag

In RHZ 4/2005 hatten wir über ein Urteil des Amtsgerichts Frankfurt berichtet, nach dem ein Mausklick bereits die zur Vollendung des Tatbestands der Nötigung erforderliche Gewaltanwendung darstellt. Deshalb war ein Veranstalter einer Online-Demo gegen die Abschiebepraxis der Lufthansa verurteilt worden. Diese Argumentation war so lächerlich, dass das Urteil am 22.5.2006 vom OLG Frankfurt kassiert wurde und das Verfahren mit einem Freispruch endete. Schadensersatzansprüche der Lufthansa wollte das OLG aber nicht ausschließen.

Datenschutzgruppe der Roten Hilfe Heidelberg

datenschutzgruppe@rotehilfe.de

PGP Fingerprint: a3 d8 44 54 2e 04 68 60 0a 38 a3 5e d1 ea ec ce f2 bd 13 2a

http://www.datenschmutz.de