# Big Data



Gabriele Gramelsberger

2008 lanciert Google mit *Google Flu Trends* einen neuen Webservice, der anhand der Häufigkeit bestimmter Suchbegriffe Rückschlüsse auf Grippeepidemien zieht. Jeremy Ginsberg, einer der Entwickler, beschreibt 2009 in seinem Artikel „Detecting Influenza Epidemics Using Search Engine Query Data" in der Zeitschrift *Nature* die Vorgehensweise der Programmierer. In mehreren hundert Billionen Suchanfragen zwischen 2003 und 2008 wurde nach Mustern der Häufigkeit bestimmter Suchbegriffe verteilt nach Regionen gesucht, so dass eine Wahrscheinlichkeitsaussage über eine mögliche Grippeinfizierung in einer Region zu einem bestimmten Zeitpunkt getroffen werden konnte. Nach eigenen Angaben wurden diese Aussagen anhand staatlich erhobener Grippedaten evaluiert, und es zeigte sich, dass *Google Flu Trends* die Ausbreitung von Grippeepidemien ebenso schnell und treffend vorhersagen konnte wie die staatlichen Gesundheitsbehörden, die auf Meldungen konkreter Fälle durch Mediziner angewiesen sind. In den stolzen Worten von Ginsberg und seinen Kollegen: „Whereas traditional systems require 1–2 weeks to gather and process surveillance data, our estimates are current each day."
Big Data wird dieses neue Phänomen genannt, das zum einen die quantitative Explosion der Daten in den Exabyte Bereich – in Worten: eine Trillion (1018) Bytes respektive eine Milliarde Gigabytes – durch die immer leistungsfähiger werdenden Supercomputer und die zunehmend schnelleren Datenleitungen meint; zum anderen auf die Möglichkeiten verweist, die diese Unmengen an Daten durch geschickte Analysemethoden wie eben *Google Flu Trends* ermöglichen. Nicht nur Grippeepidemien lassen sich so in Echtzeit lokalisieren, sondern die digitalen Spuren eines Surfers ergeben ein Profil, das interessante Aufschlüsse über dessen Vorlieben und Verhalten liefert. An diese Vorlieben angepasste Vorschläge generieren mittlerweile alle großen Online-Shops und selbst die Resultate von Suchanfragen sind auf den Einzelnen hin „profiliert". Adaptive Hypermedia wird dieser User-zentrierte Umgang mit den digitalen Medien genannt, und in Kombination mit den persönlich erstellten Profilen in sozialen Netzwerken und in App-kontrollierten Mobiltelefonen sowie Überwachungsdaten öffentlicher Räume ergibt dies eine explosive Mischung, die sich nicht einmal George Orwell hätte erträumen können – nur dass heute neben Diktatoren auch Firmen und Behörden demokratischer Staaten dieses Potential weidlich nutzen. Das Interessante daran ist, dass dies alles nur möglich ist, weil User – ja, auch Sie! – freiwillig, wenn auch unbedacht, eine Datengenerosität an den Tag legen, die atemberaubend ist. So wurden Facebook bis heute 40 Milliarden Fotos von 1,2 Milliarden Mitgliedern gespendet. Google kann täglich etwa 3,5 Milliarden Suchanfragen nach Wünschen, Sehnsüchten und zwischenmenschlichen Verbindungen auswerten und über 1 Milliarde Android-Geräte betreuen. Hinzukommen vernetzte Rauchmelder und Thermostate, Roboter, und andere nützliche Gadgets. Die Eigeninitiative dieser Konglomerate an Betriebssystemen, Apps und Webdiensten in puncto Datenbeschaffung ist beachtlich. Insgesamt, so wird geschätzt, werden jeden Tag 2,5 Exabyte Daten neu generiert.

Exa (εξ [héx] = sechs) klingt lapidar, doch es ist die Steigerung des all-umfassenden Peta (πεταννυαι = alles umfassen) und des ungeheuerlichen Tera (τερας = Ungeheuer). Auch wenn die antike Welt und ihre an den Menschen orientierten Maßeinheiten weiter denn je unter die Datenschichten versunken scheinen, so ist die Hybris der antiken Tragödie nicht weit entfernt. Was *Nature* 2009 als Revolution der Big-Data-Gesellschaft feierte, erklärten David Lazer und Kollegen in ihrem 2014 in *Science* erschienenen Aufsatz „The Parable of Google Flu: Traps in Big Data Analysis" zur Datenhybris. Eine Reihe peinlicher Fehlprognosen von *Google Flu Trends* führte schließlich zum Fall der selbsternannten Datengötter. Die Datenhybris bestehe darin, so die Autoren der Parabel der schönen neuen Datenwelt, dass soziale Daten beliebige bis bizarre Korrelationen enthalten. So war eines der am meist korrelierten Suchbegriffspaare das nach „Grippe" und „High School Basketball". Zudem ändert Google seinen Suchalgorithmus kontinuierlich, so dass dieselbe Abfrage Monate später andere Resultate zu Tage fördert. Eine von der Abfrage abhängige Datenbasis kann aber kaum Grundlage fundierter Prognosen sein. Und nicht unwesentlich, Big Data sind dynamisch und längst wird eine nicht-intendierte Feedbackschleife mitpropagiert: Erfahrene User, Abhördienste und Unterhaltungschefs großer Fernsehsender wissen, wie sie Resultate von Umfragen, Bewertungen oder Prognosen manipulieren können – sei dies auf Twitter, Google, Facebook oder anderswo. Frei nach Friedrich Nietzsches *Die Geburt der Tragödie aus dem Geiste der Musik* (1872) ist die schöne neue Datenwelt eventuell weniger licht als gedacht und macht investorenunfreundliche Mucken. Noch sträuben sich die dionysischen Datenfluten und -abgründe ein bisschen im Kampf gegen die absolute apollinische Harmonie der Algorithmenanalysen und profilierten Datenaggregate, und wir sind noch nicht ganz und gar ausspähbar und taxierbar. Aber die Logik des Digitalen lautet: Alles oder Nichts! Und wie uns bereits die CBS-Serie *Person of Interest* vor Edward Snowden lehrte: Yes, we (s)can!