Was ist Big Data?

Was hat es mit meinem Leben zu tun?

Der Begriff „Big Data“ ist seit einiger Zeit in aller Munde. Doch was steckt hinter dem Begriff? In diesem Dossier wird BEE SECURE erklären, was es mit den „großen Daten“ auf sich hat, wie die Datenberge benutzt werden und was das mit unserem Alltag zu tun hat.

„Big Data“ heißt wörtlich übersetzt „große Daten“. Wir verstehen darunter Datenmengen, die zu groß, zu komplex, zu schnelllebig und nicht gut genug strukturiert für die Auswertung mittels manueller Datenverarbeitung (z.B. Auslesen eines Wertes aus einer Excel-Tabelle) sind.

Das „big“ in Big Data bezieht sich auf drei Dimensionen:

- **Volume** (Datenvolumen): Wie viel Daten sind es?
- **Velocity** (Geschwindigkeit): Mit welcher Geschwindigkeit werden Daten generiert, verschoben und wieder geändert?
- **Variety** (Vielfalt): Aus welchen unterschiedlichen Quellen stammen die Daten? Welche verschiedenen Datentypen gibt es?

Der deutschsprachige Begriff wäre „Massendaten“ – dieser wird jedoch in den kaum genutzt, weshalb wir nun auch den englischen Begriff Big Data verwenden werden.

Unter dem Schlagwort werden allerdings nicht nur die Datenmengen an sich verstanden, sondern auch die Technologien, die benötigt werden, um sie zu sammeln und zu verarbeiten: Etwa, um sie zu visualisieren oder anderweitig aufzubereiten, so dass sie für Menschen verständlich sind.

Der technologische Fortschritt hat es erlaubt, dass immer größere Datenmengen speicherbar wurden. Noch in den 1990ern waren Disketten mit nur 1,4 MB und Festplatten mit wenigen hundert MB Speicherkapazität Standard, während heute USB-Sticks mit mehreren Gigabyte als Werbegeschenke verteilt werden und Festplatten mit mehreren Terabyte in handelsüblichen Computern verbaut sind. Kurz gesagt: Heute ist es kein Problem mehr, auch noch so große Datenberge zu speichern – durch die Verbreitung sogenannter „Cloud Storage“ müssen die Festplatten auch nicht mehr physisch im eigenen Haus vorhanden sein, sondern können online überall auf der Welt angemietet werden. Gleichzeitig hat auch die Prozessorkapazität derart zugenommen, dass die Verarbeitung immer größerer Datenmengen kein Problem mehr darstellt.

Woher stammen die Daten in Big Data?

Grundsätzlich können Daten, die als „Big Data“ bezeichnet werden, aus jedem nur erdenkbaren Lebensbereich stammen, in dem Daten anfallen. Grundsätzlich müssen wir uns als User_innen (d.h. Nutzer_innen) vor Augen halten, dass sämtliche Daten, die gesammelt werden können, vermutlich auch gesammelt werden. Die nachfolgende Liste ist also auf keinen Fall vollständig, bietet aber einen groben Überblick der Datenquellen von Big Data:

Besuche von Webseiten: Standardmäßig werden Datum, Uhrzeit, IP-Adresse, ungefährer Ort des Internetanschlusses, verwendetes Betriebssystem (Windows, OS X, Linux, inklusive genaue Version) benutzter Browser, installierte Plug-Ins, Größe des Browserfensters, Bildschirmauflösung, vorher besuchte Webseite (falls der Besuch durch Anklicken eines Links erfolgte) und angeklickte Links "geloggt" (d.h. digital erfasst). Zu Marktforschungszwecken werden aber teilweise noch viel genauere Daten geloggt, zum Beispiel, wie weit eine Webseite heruntergescrollt wurde (z.B. um herauszufinden, ob ein Artikel ganz gelesen wurde), wo der Mauszeiger herumgekreist ist, welche Dinge angeklickt wurden, usw. Teilweise werden diese Informationen in Cookies gespeichert, um User_innen bei einem weiteren Besuch wiedererkennen zu können.
Eine spielerische Art und Weise herauszufinden, was alles getrackt (d.h. "verfolgt") werden kann, ist die Webseite clickclickclick.click – einfach mal auf den Link klicken, Lautsprecher einschalten und sich überraschen lassen, wie viel die Webseite über einen weiß!

elektronische Kommunikation: E-Mail, SMS, WhatsApp, Facebook-Messenger, usw. (Aufgezeichnet wird zumindest: „Wer, wann und mit wem?“)

Daten aus sozialen Netzwerken: Dazu gehören nicht nur Postings, sondern auch Angaben über den Gefühlszustand, Beziehungen, Einstellungen und Vorlieben

Bezahlen: die Nutzung von Debit- („Bankomat), Kredit- und Kundenkarten

Sport und Bewegung: Aufzeichnung der Bewegungsdaten durch das Smartphone, Fitnessarmbänder oder anderer „Wearables“

Reisen: Flüge und Zugreisen, durch Nummernschilderkennung ist auch die Erfassung von Autoreisen technisch möglich

Automobil: durch das „vernetzte Fahrzeug“ fallen große Datenmengen an

Smart Home: das „Internet of Things“ sowie sogenannte „Smart Meters“ (elektronische Stromzählgeräte) sammeln ebenfalls Daten

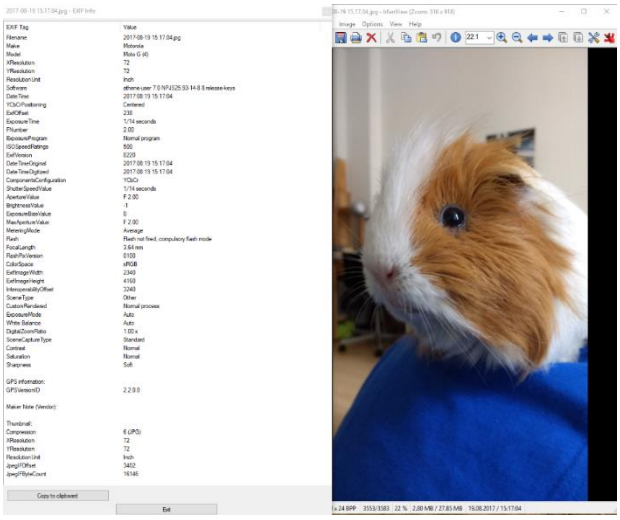
biometrische Daten: Fingerabdrücke, Iris-Scan, Gesichtserkennung und ähnliche Techniken werden immer häufiger von Smartphones, Laptops und anderen Geräten zum „einfachen“ Entsperren verwendet. Auch die Regierungen sammeln (z.B. für Pässe) solche Daten. Biometrische Daten sind sehr sensibel, dass sie anders als ein Passwort im Falle eines Diebstahls nicht geändert werden können.

Überwachung: Videokameras (von Polizei oder Privat), Gesichtserkennung, „Vorratsdatenspeicherung“ der elektronischen Kommunikation

Inhaltsdaten und Metadaten

Bei all diesen Daten wird grundsätzlich zwischen zwei Arten von Daten unterschieden: **Inhaltsdaten und Metadaten**. Der Unterschied lässt sich leicht anhand einer E-Mail erklären: Die **Inhaltsdaten** sind in diesem Fall der Text der E-Mail (und eventuelle Anhänge wie Fotos). Die **Metadaten** sind zumindest Absender_in, Empfänger_in und Betreff. Meistens werden aber noch viele weitere Daten im sogenannten Mail-Header mitgeschickt: Versendezeitpunkt (Datum und Uhrzeit), verwendetes Mailprogramm, verwendete Sprache. Oft hat jede E-Mail eine einzigartige Identifikationsnummer. Handelt es sich um eine Antwort auf eine Mail, wird auch deren ID mitgeschickt.

Auch in Fotos sind Metadaten enthalten, was vielen Nutzern oft nicht bewusst ist. Bei einem JPG-Foto sind in den sogenannten EXIF-Daten viele Informationen über das Foto und das Gerät, mit dem es aufgenommen wurde, versteckt. So finden sich in einem Foto, das mit einem Smartphone aufgenommen wurde, das Modell, der Hersteller und die genaue Version des Betriebssystems. Fototechnische Informationen wie Datum und Uhrzeit, Ausrichtung der Kamera, Brennweite, Belichtungszeit, Blendeneinstellung, Blitz-Zeit und ISO-Wert sind ebenfalls vermerkt. Auch ein Vorschaubild (Thumbnail) kann gespeichert sein – manchmal bleibt dieses auch bei Änderungen des Bildes erhalten. Außerdem können geographische Koordinaten gespeichert sein, wie eine genaue Ortung ermöglichen. Die meisten dieser Informationen speichert auch eine normale Kamera. Unser Beispielfoto zeigt (neben einem sehr putzigen Meerschweinchen) die Metadaten, die mit einem ganz normalen Smartphone aufgenommen wurden.



Die Metadaten eines Fotos können mit Bildbearbeitungsprogrammen (wie dem kostenlosen IrfanView) oder dem Browser ausgelesen werden. Allerdings lassen sie sich auch entfernen, wie diese Anleitung erklärt.

Im Fall der sogenannten „Vorratsdatenspeicherung“ (frz. *Conservation des données*) werden meistens nur die Metadaten von Kommunikation gespeichert. Das mag harmlos klingen, in Wirklichkeit lassen sich aber alleine aus den Metadaten enorm genaue Profile erstellen. Im Selbstversuch hat der deutsche Grünen-Poltiker Malte Spitz sechs Monate seiner Vorrratsdaten eingeklagt und mit Hilfe der „ZEIT“ aufbereitet. Die entsprechende Visualisierung zeichnet ein erschreckend genaues Bild des Lebens des Politikers. Es wird deutlich: Wenn vermeintlich harmlose oder „nichtssagende“ Daten zusammengeführt werden, kann ein sehr genaues Bild einer Persönlichkeit entstehen. Bei jeder Aktivität im Internet – und durch vernetzte Geräte auch abseits von Browsern – entstehen Metadaten, die gesammelt werden können. Und wir sollten davon ausgehen, dass, sobald Daten anfallen, auch jemand sie sammelt, ordnet und analysiert.

Anwendungen und Datenverkauf

Natürlich sammeln nicht nur Regierungen (Meta)daten, sondern auch Firmen. Die Internetgiganten wie Amazon, Google, Facebook, usw. erstellen Profile ihrer Nutzer_innen, nach eigenen Angaben, um ihre Produkte zu verbessern – meistens heißt das, den Nutzer_innen möglichst passende Angebote zu machen. Das Internetkaufhaus versucht, unsere Kaufwünsche vorherzusehen, bevor sie uns selbst bewusst werden, die Suchmaschine und das soziale Netzwerk wollen uns jene Werbung anzeigen, die genau zu unserer Persönlichkeit passt. Mit den gesammelten Daten wird teilweise auch entschieden, wie wir als Kund_innen behandelt werden oder ob wir kreditwürdig sind.

Neben den selbst gesammelten Daten gibt es für Unternehmen natürlich auch immer die Möglichkeit, Daten zu kaufen. Auf Plattformen wie big.exchange werden offen Nutzerdaten angeboten – und Webseitenbetreiber_innen haben die Möglichkeit, die Daten ihrer Nutzer zu verkaufen. Unroll.me, ein Dienst, mit dem sich Nutzer von Newslettern abmelden können, wurde im April 2017 dabei erwischt, Daten an den Taxi-Dienstleister „Uber" verkauft zu haben – ausgerechnet von jenen Nutzern, die den Newsletter des Uber-Konkurenten „Lyft" nicht mehr lesen wollten. Viele dieser Daten-Deals laufen im Verborgenen und sind für die Nutzer nicht unbedingt nachvollziehbar – auch wenn sie oft durch die AGBs der Dienste gedeckt sind.

Wir sollten uns also bewusst sein: **Wenn wir einen Dienst „kostenlos" nutzen können, zahlen wir im Regelfall trotzdem – nicht mit Geld, sondern mit unseren Daten.**

Wer selbst einmal ausprobieren will, wie das Datensammeln und -Verkaufen funktioniert, kann das Online-Game „Data Dealer" ausprobieren. Spielerisch wird hier gezeigt, mit welchen Methoden die Datensammler zu unseren Daten kommen und wer sich für welche Daten interessieren könnte.

Big Data hat aber nicht nur Anwendungen in der Werbung oder im Verkauf: Auch Bereiche wie Verkehr, Medizin, Bildung, Wissenschaft und Journalismus können von der Verarbeitung enormer Datenberge profitieren und so unser Leben angenehmer gestalten. Neben Gefahren für unsere Privatsphäre birgt Big Data eben auch enorme Chancen: So können Infektionsherde ausgemacht, Transportwege optimiert, neue Zusammenhänge gefunden und Skandale aufgedeckt werden.

Wie jede Technologie ist auch Big Data weder inhärent gut oder böse, noch neutral. Sie ist ein Werkzeug – und wie jedes Werkzeug kann auch Big Data für die verschiedensten Vorhaben eingesetzt werden. Leider ist es für uns als Nutzer_innen meistens nicht ganz klar, welche unserer Daten zu welchen Zwecken verwendet werden – oft wissen wir überhaupt nicht, welche Daten wir wo hinterlassen.

Die Kontrolle behalten

Neben dem oben erwähnten Datenhandel gibt es auch die Gefahr von Datenklau: Hacks, Leaks und Sicherheitslücken betreffen uns alle. Kein Dienst kann 100-prozentige Sicherheit garantieren. Demnach lohnt es sich, möglichst wenige Spuren im Netz zu hinterlassen.

Gute Entscheidungen treffen

Informieren Sie sich, bevor Sie sich bei einer App, einem sozialen Netzwerk, einem Spiel oder einem sonstigen Dienst anmelden. Seiten wie tldrlegal.com können dabei helfen.

Die Kunst des Weglassens

Gar nichts von sich preiszugeben ist heute fast nicht mehr möglich. Das heißt aber nicht, dass jeder Dienst alles von Ihnen wissen muss! Überlegen Sie sich gut, welche Informationen Sie wo eintragen. Oft ist es besser, Lücken zu hinterlassen. Vor dem Posten von Fotos kann man die Metadaten entfernen.

Einstellungen im Blick halten:

Behalten Sie die Einstellungen Ihres Browers und Ihrer sozialen Netzwerke im Blick, um möglichst wenige Daten preiszugeben.

- Grundsätzlicher Check der Browser-Einstellungen: Is your Browser safe from tracking?
- Mozilla Firefox: Einstellungen für Privatsphäre, Surfchronik und Nicht-Verfolgen-Funktion
- Google Chrome: Whitepaper zum Datenschutz
- Apple Safari: Datenschutz-Einstellungen
- Microsoft Edge: Datenschutz bei Edge
- Facebook: Privacy Settings & Tools
- Twitter: Tipps zum Datenschutz
- Instagram: Welche Einstellungen gilt es zu beachten?
- Snapchat: Risiken bei Snapchat

Digitale Selbstverteidigung

Neben den Einstellungen gibt es auch die Möglichkeit, proaktiver vorzugehen und die sogenannte „digitale Selbstverteidigung“ zu praktizieren.
Tipps dazu gibt es bei digital-selfdefense.com oder der Electronic Frontier Foundation. Es gibt auch einige Plugins, die dabei helfen, nicht zu viele Daten preiszugeben:

- Privacy Badger
- Ghostery
- disconnect.me

Der TOR-Browser ermöglicht anonymes Surfen über das TOR-Netzwerk – der Browser ist so konfiguriert, dass er möglichst wenige Spuren hinterlässt.
Wer wirklich auf Nummer sicher gehen will, kann die Linux-Distribution „Tails“ verwenden. Das System kann von einem USB-Stick aus gebootet werden und verspricht, keine Spuren zu hinterlassen.

Weitere Informationen

BEE SECURE veröffentlicht laufend weitere Informationen zum Thema Big Data. Sie können uns auf Facebook, Twitter und Instagram folgen und die Hashtags #BIGDATA #fettdonneeen #BEESECURE verfolgen.

Bei Fragen zum Thema BIG DATA oder generell zur Internetnutzung können Sie jederzeit die BEE SECURE HELPLINE unter der Telefonnummer 8002-1234 kontaktieren.

Wenn Sie juristische Beratung suchen oder eine Beschwerde wegen Datenmissbrauchs einreichen wollen, können Sie das bei der nationalen Datenschutzkommission Commission nationale pour la protection des données machen.

Quellen:

- Gabler Wirtschaftslexikon
- Wikipedia
- Timeline der Entwicklung von Speichermedien
- Webseite-Tracking:
- Guardian: Tracking the Trackers
- clickclickclick.click
- Datenverkauf:
- big.exchange
- New York Times: Unroll.me Service Faces Backlash Over a Widespread Practice: Selling User Data
- Data Dealer

Bei Fragen bezüglich des Internetbetrugs
oder der Nutzung des Internet im Allgemeinen,
wenden Sie sich bitte an die BEE SECURE Helpline:

BEE SECURE HELPLINE BERATUNG Tel: 8002 1234

Herausgeber: BEE SECURE · B.P. 707 · L-2017 Luxemburg
Tel.: (+(+352) 247-86427 · Fax.: Fax: (+ 352) 46 41 86
bee-secure@snj.luwww.bee-secure.lu

THE GOVERNMENT OF THE GRAND DUCHY OF LUXEMBOURG
Co-financed by the European Union Connecting Europe Facility