Big Data – ganz gross?

Cyrill Helg, BSc ZFH EE

Eine abschliessende und allgemein akzeptierte Definition von Big Data ist nicht ausfindig zu machen. Trotzdem fehlt es nicht an vollmundigen Versprechen, wie Big Data unsere Welt verbessern wird: Mit Leichtigkeit könnten Firmen aus riesigen Datensammlungen unvorhersehbare Erkenntnisse und Wettbewerbsvorteile gewinnen. Sie müssten nur im grossen Stil in eine der angesagten Technologien investieren und schon würden sie in wenigen Monaten die Konkurrenz weit hinter sich lassen.

Wir haben die wiederkehrenden Aussagen zu Big Data zusammengetragen und kritisch hinterfragt.

Die Glaskugel ist endlich Realität

Einmal mehr glaubt man, das Werkzeug für den Blick in die Zukunft gefunden zu haben. Auch wenn gewisse Prognosen exakter werden, darf man nie aus den Augen verlieren, woher die Daten stammen: aus der Vergangenheit. Zudem bilden sie ein Ereignis nur indirekt ab. Zum Beispiel verhält sich die Zahl der Kaufabschlüsse direkt proportional zur Anzahl Kurznachrichten auf Twitter oder Suchanfragen bei Google. Dieser Zusammenhang mag korrekt sein, aber nicht zwingend, nicht in allen Fällen und auch nicht für alle Zeiten. So ist bis heute jede Firma gescheitert, die versucht hat, Börsenkurse aufgrund solcher scheinbarer Zusammenhänge vorauszusagen.

Mehr Daten liefern genauere Ergebnisse

Diese Annahme setzt irrtümlicherweise Daten mit Information gleich. Bei Big Data wird davon ausgegangen, dass mit jedem zusätzlichen Datensatz auch wirklich mehr Information hinzukommt. Meist ist dies aber nicht der Fall. Zu Beginn kommt mit jedem eindeutigen Sample auch Neues dazu, aber mit zunehmender Anzahl Stichproben steigt auch die Wahrscheinlichkeit einer Überlappung zu Vorhandenem. Man denke nur an die Datensammlung zu Hause auf dem eigenen Computer: Oft speichern wir dieselben Dokumente an unterschiedlichen Orten, erstellen Backups, versenden Daten an weitere Personen und legen Kopien auf portablen Speichermedien ab. Obschon die totale Menge an Daten zunimmt, enthält sie nicht zwingend mehr Information. Dies entkräftet die typische Big-Data-Aussage, aufgrund riesiger Datenmengen könne auch Unvorstellbares zu Tage gefördert werden.

N = alle

Eine der Schlüsselannahmen bei Big Data besagt: Anstelle einer kleinen Stichprobe werden alle verfügbaren Daten verwendet und daher wird bei den Analysen davon ausgegangen, dass N = alle gilt. Dies ist insbesondere bei jenen Erhebungen, die auf Daten aus dem Web und den sozialen Netzwerken beruhen, ein Trugschluss. Unterschiedliche Bevölkerungsgruppen sind unterschiedlich stark vertreten – und schlichtweg nicht alle

Personen sind auf den Plattformen oder gar im Internet aktiv.

Das Ende der exakten Modelle

Big-Data-Analysen liefern keine Erklärung für die Zusammenhänge, die sie zu Tage fördern – sie stehen lediglich für das „Was“ und nicht für das „Warum“ eines Ergebnisses. Es wird sogar behauptet, das klassische Bild der Wissenschaft sei bedroht: Eine exakte Methode und Grundlagenforschung werde mit Big Data hinfällig und der Frage nach Kausalität müsse nicht mehr Rechnung getragen werden. Dabei wird eines vergessen: Ohne Erfahrung und ohne eine fundierte Beurteilung der neuen Zusammenhänge lassen sich die Ergebnisse nicht in die Praxis umsetzen. Es braucht, insbesondere für zukunftsweisende Entscheidungen, ein kausales Verständnis – und dafür ist das „Warum“ unerlässlich.

Es funktioniert überall

Auch wenn Big Data in einem Bereich einen konkreten Nutzen gebracht hat, lässt sich der Nutzen nicht automatisch auf andere Gebiete übertragen. Die neu entdeckten Infektionssymptome bei Frühgeborenen, die ohne Hypothesen von medizinischen Forschern aus Big-Data-Korrelationen und anschliessender Bewertung erarbeitet wurden, sind ein häufig zitiertes und erfreuliches Beispiel. Es ist aber verkehrt anzunehmen, dass genau dieses Vorgehen aus der Medizin auch für die Analyse von Märkten, das Verhalten von Konsumenten oder die Bewertung von Unternehmen nützlich ist. Denn die zugrundeliegenden Daten sind jeweils vom Kontext abhängig – und somit auch die darauf basierenden Folgerungen.

Ein Echtzeit-Selbstläufer

Die lokale Vorhersage der Grippewelle durch Googles „Flu Trend“ hat zwar kurz nach der Freigabe beeindruckende Resultate geliefert; sie liegt aber nach jüngsten Auswertungen bis zu 100% daneben. Dies zeigt: Die aufwändige Auswertung von Millionen von Korrelationen war zwar für einen bestimmten Zeitraum richtig, aber die Variable der Zeit konnte nicht in das finale Modell einfliessen. Deshalb müssen Unternehmen Massnahmen aufgrund von Big-Data-Auswertungen stetig hinterfragen und neu beurteilen. Dafür braucht es einen Entscheider, also eine Person, welche die Verantwortung trägt und die Beschlüsse umsetzt.

Fehlende Daten sind leicht zu beschaffen

Behördendaten wie Einwohnerstatistiken, Wahlresultate oder Strassenkarten mit Unfallhäufigkeiten werden in der Schweiz online zur Verfügung gestellt. Das Bundesamt ist bestrebt, die Menge und Vielfalt an Datensätzen weiter zu erhöhen und erhofft sich davon einen transparenteren und effizienteren Staat sowie einen Beitrag an die Innovationskraft der Schweizer Wirtschaft. Die Nutzungsbedingungen für den kommerziellen Gebrauch sind je nach Art der Daten verschieden, aber grundsätzlich bewilligungspflichtig und mit einer Nutzungsgebühr versehen. Für Unternehmen gestaltet sich die Verwendung von Kundendaten hingegen viel komplizierter: Das Sicherstellen von Datenschutz, Anonymisierung und dem Schutz vor Manipulation ist eine grosse Herausforderung. Den Aufwand dafür zu rechtfertigen ist ebenfalls heikel, denn niemand kann voraussehen, was bei der Analyse dieser externen Daten in Kombination mit den Daten aus dem

eigenen Unternehmen herauskommt. Dieser Umstand widerspricht zudem der grundliegenden Idee des Datenschutzes: Die einwilligende Person soll genau wissen, wie ihre Daten verwendet werden.

Befreit von den diskutierten Mythen, lässt sich das Potenzial von Big Data für ein Unternehmen nüchtern betrachten. Dabei steht nicht die Datenmenge oder eine komplexe IT-Lösung im Vordergrund, sondern die Fähigkeit eines abteilungsübergreifenden, schlanken Projektteams, das sich einer klaren und geschäftsrelevanten Fragestellung annimmt. Das Team muss in der Lage sein, unerwartete Resultate zu beurteilen und in den richtigen Kontext zu setzen. Weitaus anspruchsvoller zeigt sich das Identifizieren relevanter Daten und deren Strukturierung, denn es ist an der Tagesordnung, wichtige Geschäftsdaten in Spreadsheets verteilt über historisch gewachsene Ordnerstrukturen zu speichern. Der erste Schritt besteht also darin, herauszufinden, wo, wie und wie häufig die unerlässlichen Daten abgelegt werden, um sie dann in eine Form zu bringen, die es erlaubt, vernünftige und kosteneffiziente Auswertungen durchzuführen. Haben Sie diesen Grundstein gelegt und Ihre Daten im Griff?