
2.1 Lücken trotz Sammelwut

Wie beschrieben werden Daten in wildem Eifer gesammelt. Jedoch besteht der Verdacht, dass dies nicht zwingend dort geschieht, wo dringende Informationen benötigt werden, sondern dort, wo es sich gut sammeln lässt. Deshalb werden auch heute noch bei jedem krisenhaften Szenario Datenlücken festgestellt und beklagt. Auch in der aktuellen Flüchtlingskrise in Deutschland wurden zum Beispiel Datenlücken zum Leerstand von Häusern und Wohnungen beklagt, trotz zahlreicher bereits existierender Datensammlungen zu Immobilien und Kauf- und Mietpreisen.

Ein weiteres Beispiel stammt aus der aktuellen Diskussion über die möglicherweise krebserregende Wirkung des zur Unkrautbekämpfung eingesetzten Breitbandherbizids Glyphosat. Bei dieser Diskussion stellte sich heraus, dass das Wissen über die geografische Verteilung von Krankheiten sehr lückenhaft ist. Daher können keine Untersuchungen durchgeführt werden über Korrelationen von (gehäuften) Erkrankungsfällen mit der Lage potenzieller Gefahrenquellen, wie zum Beispiel dem Einsatzgebiet gefährlicher Wirkstoffe in der Landwirtschaft bzw. an Bahnstrecken, den Standorten von Kraftwerken oder emissionsintensiven Fabriken sowie Verkehrsknotenpunkten. Dazu passend wäre auch die Frage, ob Studien zur Prävalenz von Hautkrebserkrankungen deutlich an Erkenntnis gewinnen könnten, wenn Daten aus der Medizin und der Meteorologie (Sonnenstrahlung, -intensität, Sonnenstunden nach geografischen Zuordnungen etc.) zusammengetragen würden.

Schließlich wurde bei den diversen Finanzkrisen von den neu geschaffenen internationalen Institutionen zur Überwachung der Finanzstabilität, zum Beispiel dem Financial Stability Board der G20-Staaten, eine ganze Serie von Datenlücken (*data gaps*) in den finanz- und realwirtschaftlichen Datensammlungen identifiziert.

2.2 Fehlende Ordnung im Datenuniversum

Für Datenanalysten gilt: Mit den Daten beginnt alles – sie sind die Bausteine, die Atome in unserem Universum und der Ausgangspunkt unser Arbeit. Damit können sie zugleich unser höchstes Gut oder der schrecklichste Fluch sein. Denn wenn nichts zusammenpasst, sind sie wertlos.

Von dieser Messlatte ist die in diesem Kapitel beschriebene explodierende Datenwelt sehr weit entfernt! Denn der Vergleich mit der Realität zeigt: Die Informationsindustrie hinkt bezüglich einer Standardisierung und Normung weit hinter anderen Industriezweigen oder Wissenschaftsdisziplinen hinterher: Denn weder gibt es ein Ordnungssystem für Daten und Informationen noch eine ausgeprägte Standardisierung und schon gar kein „Periodensystem der Elemente“ wie in der Naturwissenschaft. Nirgendwo finden wir auch nur Ansätze eines *Unique Identifiers*, eines Barcodes für Informationen. So könnte der Eindruck entstehen, Google sei das eigentliche Ordnungssystem. Diese fehlende Standardisierung ist themenübergreifend zu beklagen, so wird zum Beispiel auch die fehlende Standardisierung von Daten in der Pflanzen(-genom-)forschung beklagt (vgl. Div. 2011).

Diese Art des Wildwuchses lässt sich auf globaler Ebene mit der Unkontrolliertheit des Internets erklären, jedoch findet sie auch in dem sonst deutlich besser verwalteten Bereich der Industrie statt: Der Mangel an Ordnung zeigt sich in den Datenwelten fast aller Unternehmen und begründet dort die zahllosen Datenintegrations-, BI- oder Data-Warehousing-Projekte, neuerdings auch durch *Big-Data-Projekte*. Der enorme Bedeutungszuwachs von Daten zeigt sich auch in der vielfachen Ernennung von *Chief Data Officers*, deren Hauptaufgabe in der Regel darin besteht, eine Gesamtordnung in die Datenwelt des Unternehmens zu bringen.

Da die wiederholten Versuche, die eigene (!) Datenlandschaft begehbar und beherrschbar zu machen, meist nur von mäßigem Erfolg gekrönt sind, zeigt sich das Phänomen fehlender Gesamtordnung instituts-, branchen- oder länderübergreifend noch stärker. Lediglich in Spezialgebieten mit entsprechenden kommerziellen Interessen liegen gut aufbereitete Datenwelten vor. So etwa bei der Suche nach Gebrauchtwagen, Hotelzimmern, Flugverbindungen und Wohnungen. Hinter den bekannten Scout- und Preisvergleichswebseiten stecken nicht etwa Big-Data-Lösungen, die die Gebrauchtwagenanzeigen des Internets per *Text-Mining* durchforsten und mithilfe ihrer vernetzten Intelligenz daraus die benötigten Informationen ermitteln. Nein, diese Daten wurden zuvor „in Ordnung gebracht“, durch eine durchgängige Klassifikation (zum Beispiel Marke, Typ, Jahr der Zulassung, Postleitzahl des Anbieters, Kilometerstand) und einen durchgängigen Satz an Attributen (zum Beispiel Vorhandensein von Klimaanlage, Anhängerkupplung).

2.3 Nutzung der IT-Technologie nicht ohne fachliche Expertise möglich

Für die neuen aus Mikrodaten und deren Verknüpfung gebildeten hochwertigen Datensammlungen ergibt sich eine Änderung der Arbeitsweise: Bei der Masse, Vielfalt und

Komplexität der Daten kann a priori gar nicht festgelegt werden, welche Fragestellungen mit diesem Datenmaterial überhaupt beantwortet werden sollen. Das heißt, es ergibt sich eine stark erhöhte Volatilität der Auswertungswünsche. Damit ist die Informationsgewinnung nicht mehr als klassische, geradlinige Statistikproduktion abbildbar, stattdessen ist die Umsetzung eines Konzepts der Datenanalyse *on demand* erforderlich. So wird etwa eine granular aufgebaute Statistik über Wertpapierinvestments (*security-by-security*) natürlich mit Blick auf standardisierte, stets wiederkehrende Auswertungen konzipiert. Doch ein wachsender Anteil an Analyseanforderungen widmet sich Fragestellungen, die über Nacht brisant wurden, so etwa: Wie sieht international die Halterstruktur für Staatsanleihen eines bestimmten europäischen Landes aus?

Gerade mit der Verknüpfung mehrerer Datenquellen werden schnell neue Datensammlungen mit mehr als 30 Dimensionen (Identifikationsmerkmalen für einen Datenpunkt) gebildet, die eine gigantische Vielfalt an Auswertungsmöglichkeiten bieten. Aber welcher Analytiker oder Wissenschaftler kann (und möchte) schon mit 30 Dimensionen umgehen und solche Analysen *on demand* formulieren? In der Praxis wird man häufig an drei, vier oder fünf „Schräubchen drehen“ oder mit entsprechend vielen „Bällen jonglieren“, aber dann ist die Grenze bald erreicht. Die Informationssysteme müssen also die tiefe fachliche Expertise durch geeignete Datenanalysen unterstützen, statt umgekehrt den Experten in einem nicht mehr beherrschbaren Datensumpf ertrinken zu lassen. Sonst besteht die Gefahr, dass wir zwar datentechnisch korrekt operieren, aber letztlich Äpfel mit Birnen vergleichen. Dies gilt auch für die häufig praktizierte Technik des *Data-Mining*. Hier werden roboterartig ganze Datenschungel durchforstet, alle möglichen Permutationen der oben genannten 30 Dimensionen gebildet und auf signifikante Ausprägungen der Messgrößen untersucht. Hier ist jedoch genauso wie bei der im nachfolgenden Abschnitt beschriebenen *Big-Data-Technik* die fachliche Expertise zur Bewertung der Ergebnisse von größter Wichtigkeit.

Literatur

Div. (2011) The iPlant collaborative: Cyberinfrastructure for plant biology. Front Plant Sci (Review Article). doi: [10.3389/fpls.2011.00034](https://doi.org/10.3389/fpls.2011.00034)

Die Vermessung des Datenuniversums

Datenintegration mithilfe des Statistikstandards SDMX

Stahl, R.; Staab, P.

2017, XIII, 108 S. 39 Abb. in Farbe., Softcover

ISBN: 978-3-662-54737-3