Daten – Logik, Interpretation und Wahrheit

Umgang mit der Komplexität großer Datenmengen

Deggendorfer Forum zur digitalen Datenanalyse e.V. (Hrsg.)

IDW VERLAG GMBH

Daten – Logik, Interpretation und Wahrheit

Umgang mit der Komplexität großer Datenmengen

Deggendorfer Forum zur digitalen Datenanalyse e.V. (Hrsg.)

IDW VERLAG GMBH

Das Thema Nachhaltigkeit liegt uns am Herzen:

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne vorherige schriftliche Einwilligung des Verlages unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verbreitung in elektronischen Systemen. Es wird darauf hingewiesen, dass im Werk verwendete Markennamen und Produktbezeichnungen dem marken-, kennzeichen- oder urheberrechtlichen Schutz unterliegen.

© 2022 IDW Verlag GmbH, Tersteegenstraße 14, 40474 Düsseldorf

Die IDW Verlag GmbH ist ein Unternehmen des Instituts der Wirtschaftsprüfer in Deutschland e. V. (IDW).

Satz: Reemers Publishing Services GmbH, Krefeld
Druck und Bindung: C.H.Beck, Nördlingen
KN 12050

Die Angaben in diesem Werk wurden sorgfältig erstellt und entsprechen dem Wissensstand bei Redaktionsschluss. Da Hinweise und Fakten jedoch dem Wandel der Rechtsprechung und der Gesetzgebung unterliegen, kann für die Richtigkeit und Vollständigkeit der Angaben in diesem Werk keine Haftung übernommen werden. Gleichfalls werden die in diesem Werk abgedruckten Texte und Abbildungen einer üblichen Kontrolle unterzogen; das Auftreten von Druckfehlern kann jedoch gleichwohl nicht völlig ausgeschlossen werden, so dass für aufgrund von Druckfehlern fehlerhafte Texte und Abbildungen ebenfalls keine Haftung übernommen werden kann.

ISBN 978-3-8021-2732-8

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://www.d-nb.de abrufbar.

Coverfoto: www.adobestock.com/Weissblick

www.idw-verlag.de

1 No Signal – verschwindet die Wahrheit in der Flut der Daten?

Von Ernst-Rudolf Töller

Ernst-Rudolf Töller war über 25 Jahre als IT-Prüfer tätig und hat sich in dieser Zeit intensiv mit Fragen der Datenanalyse beschäftigt. Er war in dieser Zeit für Wirtschaftsprüfungsgesellschaften tätig und befindet sich seit seinem Ausscheiden bei der BDO AG, Hamburg (Ende Februar 2016), im aktiven Ruhestand. Während seiner Tätigkeit als „Data scientist" war er sowohl im Rahmen von Jahresabschlussprüfungen wie auch bei der Untersuchung doloser Handlungen oder Projekten zur Unterstützung der internen Revision tätig. Seine Schwerpunkte lagen und liegen immer noch auf mathematisch-statistischen Modellen zur Datenanalyse sowie deren technische Umsetzung.

1.1 Einleitung

In der Informationstechnologie fehlt es aktuell nicht an beeindruckenden Beispielen für innovative Anwendungen und Lösungen:

- **Suchmaschinen**
  Jeder kann überall auf der Welt nahezu unbegrenzt Suchmaschinen nutzen, um Einzelheiten zu speziellen Fachthemen zu studieren, und keineswegs nur, um das Neueste zu werblich relevanten Themen oder Begriffen zu erfahren.
- **Übersetzungsprogramme**
  Texte können über Webportale automatisch und nahezu in Echtzeit übersetzt werden. Das Ergebnis einer Rückübersetzung des übersetzten Textes in die ursprüngliche Sprache zeigt die beeindruckende Qualität der hinter den Portalen eingesetzten Software.
- **Autonomes Fahren**
  Es ist heute schon selbstverständlich, dass ein Auto selbstständig rückwärts einparkt oder, speziell auf der Autobahn, selbstständig in der Spur fährt und den Sicherheitsabstand zum davor fahrenden Auto einhält. Die Weiterentwicklung solcher Funktionen bis hin zu einem möglicherweise vollständig autonomen Fahren erscheint hier gleichermaßen konsequent wie selbstverständlich.

Der Anteil der digitalen Technik an der schnellen Entwicklung von Corona-Impfstoffen ist ein weiteres aktuelles – und besonders beeindruckendes – Beispiel. Solche Leistungen prägen das Bild einer Technik, die offenbar heute schon über Optionen für die Lösung beliebig komplexer Probleme verfügt. Und wenn Lösungen heute noch nicht verfügbar sind, dann werden sie in Zukunft durch schrittweise Weiterentwicklung im Prinzip bereits vorhandener Ansätze mithilfe immer schnellerer Rechner und größerer Speicher verfügbar sein.

Die Frage hier ist: Was bedeutet es eigentlich für die digitale Analyse von Daten, wenn der Grad der Durchdringung komplexer Themen durch digitale Technik so hoch ist, wie die Beispiele zeigen?

Beispiel

- Suchmaschinen können Anfragen offenbar **deuten** und fachlich qualifiziert beantworten
- Übersetzungsprogramme durchdringen einen Text und können ihn insofern **verstehen**
- Fahrzeuge können zumindest einzelne reale Situationen vollständig **begreifen** und sich insoweit autonom bewegen

Alle genannten Anwendungen sind nicht ohne **die künstliche Intelligenz** denkbar. Eine besondere Rolle spielen dabei die sog. **neuronalen Netze**. An diese Technik knüpfen sich auch bei der digitalen Analyse von Unternehmensdaten vielfältige Erwartungen.

1.2 Die Welt und die Welten

Die Welt wird heute vielfach als uneinheitlich wahrgenommen. Offensichtlich gibt es keine Sichtweise, keine Theorie, die alles gleichermaßen gut erklären kann. Auch die Naturwissenschaften sind offenbar nicht in der Rolle des **Alles-Erklärers**. Es gibt keinen **Wahlwissenschaftler**, der das Ergebnis einer Wahl aus welchen Parametern oder Daten auch immer zuverlässig im Voraus berechnen kann. Ähnliches gilt in Unternehmen für die Daten eines Geschäftsjahres, daran können auch methodisch sauber ermittelte Planzahlen nichts ändern. In beiden Welten sind alle im Voraus berechneten Zahlen nur Prognosen und alle Beteiligten sind gut beraten, die Grenzen dieser Prognosen zu kennen und in ihrem Handeln zu berücksichtigen.

Vielleicht ist es in diesem Zusammenhang hilfreich, sich die Welt als Summe einzelner *Welten* zu vergegenwärtigen, die sich untereinander ein Stück überlappen können, aber jeweils unterschiedlichen Regeln und Gesetzmäßigkeiten unterliegen. Diese Regeln oder besser die jeweils *bekannten* Regeln bestimmen in den einzelnen Welten *grundsätzlich* aber nicht immer bis ins Detail deterministisch, was dort geschieht. Der Wahlforscher muss den Wahlgang abwarten, das Unternehmen kennt die Ist-Zahlen erst am Ende eines Geschäftsjahres.

Die Welt ist **pluralistisch**, nicht einfach. Das Neben- und Miteinander der verschiedenen Welten bedingt ein hohes Maß an Komplexität. Das könnte den Schluss nahelegen, dass es angesichts einer so vielfältigen Welt sowieso nur subjektive Meinungen und Ansichten gibt und geben kann und dass sich diese Welt anders gar nicht beschreiben lässt. Unwidersprochen regieren vielfach nur noch herrschende Meinungen, abgesehen vielleicht von einigen Welten, die das Refugium exakter Wissenschaften sind. Es drängt sich sogar der Eindruck auf, dass einzelne Welten überhaupt nur die Summe subjektiver Äußerungen sind, die sich dort in einer schillernden Mischung aus Sprache, Bildern, Symbolen und Logos präsentieren.

1.3 Daten und Meinungen

Die Pluralität der Welt geht nicht so weit, dass die Wirklichkeit nur noch Meinung und Einbildung ist. Begrenzt wird der extreme Meinungspluralismus immer dann, wenn Daten vorliegen, die wesentliche Bestandteile einer bestimmten Welt dokumentieren. Die Analyse dieser Daten determiniert die möglichen Aussagen über diese Welt. Unternehmen stellen aufgrund der Art und des Umfangs der mittlerweile verfügbaren Daten genau so eine Welt dar.

Aktuell wächst die Menge der weltweit erzeugten Daten pro Jahr um mehr als 25%[1]. Die höchsten Zuwächse gibt es dabei in den Sektoren Produktion und Handel. Die globale Datenmenge wächst offenbar in Zusammenhang mit der fortschreitenden Digitalisierung von Unternehmen bei zentralen Prozessen wie Absatz oder Beschaffung. Die Aufzeichnung von Prozessen in den Daten des Unternehmens erfolgt

[1] Vgl. https://www.iwd.de/artikel/datenmenge-explodiert-431851/ (abgerufen am 24.02.2022).

mit immer höherer Auflösung und auch nicht mehr ausschließlich über manuelle Eingaben von Mitarbeitern. Die Erfassung der Daten selbst vollzieht sich bereits automatisiert über Scanner, Sensoren (RFID) oder Techniken wie das Internet der Dinge.

Die Zahlen eines Wahlergebnisses begrenzen die möglichen neuen Koalitionen und Regierungen. Weitere Entscheidungen bis zur Bildung einer neuen Regierung können nur noch abhängig von den aktuellen Ergebnissen erfolgen. Genauso liefert die Analyse der Daten in Unternehmen entscheidende Sichten auf wichtige Sachverhalte und bildet so den Rahmen für rationale Entscheidungen bei der Unternehmenssteuerung. Entsprechendes gilt für die Bilanzierung von Unternehmen oder für Bewertungsfragen.

Wähler können bei der Wahl frei entscheiden, Kunden und Lieferanten eines Unternehmens haben ebenfalls beträchtliche Handlungsspielräume. Trotzdem entwickeln sich die Dinge nicht einfach willkürlich, wenn man ein Wahlergebnis oder ein Unternehmen im Ganzen betrachtet. So kann man beim Vergleich von Wahlen zwischen verschiedenen Ländern Grundmuster und Strukturen feststellen, die offenbar über längere Zeiträume erhalten bleiben.

Beispiel

Beispiele für solche Grundmuster und Strukturen:

- Das Gesamtergebnis einer Wahl spiegelt sich jeweils in jeder einzelnen Wahlurne mit eher geringen Abweichungen wider (einheitliches Wählerverhalten, typisch für Frankreich)
- Die Ergebnisse in den einzelnen Wahlurnen sind grundverschieden. Das Gesamtergebnis der Wahl ist ein Mix aus ziemlich divergenten Einzelergebnissen (uneinheitliches Wählerverhalten, typisch für die Schweiz)[2]

[2] Weitere Details zur digitalen Analyse von Wahlergebnissen und Aufdeckung von Wahlbetrug vgl. Klimek, Peter u.a., Statistical detection of systematic election irregularities, in: Proceedings of the National Academy of Sciences, Vol. 109, Issue 41 (2012), S. 16469–16473, online archiviert unter https://www.pnas.org/content/pnas/109/41/16469.full.pdf (abgerufen am 24.02.2022).

Solche strukturellen Analysen haben nicht nur im Fall von Wahlentscheidungen ihre Bedeutung. Die Daten eines Unternehmens erlauben ebenfalls mehr als nur die Aggregation auf Ebene von Bilanz- und GuV-Zahlen. Strukturelle digitale Analysen in Bezug auf Kunden, Lieferanten, Artikel, Niederlassungen oder andere betriebswirtschaftliche relevante Einheiten sind auch mit Unternehmensdaten möglich. Das schließt entsprechende Periodenvergleiche genauso ein wie digitale Analysen, die zur Aufdeckung doloser Handlungen dienen können.

Es bleiben natürlich Fragen: Welchen Stellenwert haben solche Analysen? Geht es um **Wahrheit**, weil den Analysen Daten also **objektive Informationen** zugrunde liegen, oder stellen die Ergebnisse digitaler Analysen am Ende auch nur Meinungen dar? Können verschiedene Analysen unterschiedliche, vielleicht sogar einander widersprechende Ergebnisse zeigen? Im Folgenden sollen diese Fragen näher untersucht werden, ohne gleich in allen Fällen endgültige Antworten zu geben.

1.4 Computer verstehen. Verstehen Computer?

Ein wichtiger Faktor zur richtigen Einschätzung der Leistung von IT-Systemen, insbesondere von Spitzenleistungen, wie sie am Anfang aufgelistet sind, ist die Frage nach der Semantik von Daten, genauer: die Frage, ob Rechner in der Lage sind, Daten tiefer zu **verstehen**, Schlüsse aus der Analyse von Daten zu ziehen und auch entsprechend zu agieren.

Im Folgenden zwei konträre Standpunkte in dieser Sache:

Standpunkt 1: Rechner können Daten nicht verstehen

Rechner, wie sie heute verfügbar sind, können Daten nur syntaktisch verarbeiten, d..h., sie führen arithmetische Operationen mit Speicherplätzen aus oder übertragen den Inhalt eines Speicherplatzes in einen anderen usw. Ob das Ergebnis einer arithmetischen Operation in dem jeweiligen Kontext Sinn ergibt oder ob das Übertragen von Feldern aus einer gespeicherten Adresse dazu führt, dass ein Adressfenster nach dem Ausdrucken einen für die Zustellung brauchbaren Inhalt zeigt, wird prinzipiell nicht berücksichtigt. Im jeweiligen Kontext sinnvolle oder notwendige Kontrollen müssen in allen Details explizit programmiert werden oder finden gar nicht satt. Ebenso müssen alternative oder modifizierte Verarbeitungen bis ins Detail codiert sein, um wirksam zu werden. Rechner führen einen Code nur aus, sie sind nicht in der Lage,

einen gespeicherten oder gerade ausgeführten Code zu verstehen oder aus diesem Code selbstständig etwas abzuleiten.

Die Existenz von IT-Systemen mit sog. künstlicher Intelligenz scheint dem zu widersprechen. Aber gerade der Umgang mit natürlicher Sprache zeigt, welche Grenzen der heutigen Technik hier gesetzt sind.

Die Probleme und Besonderheiten bei der Übersetzung von Literatur in eine andere Sprache zeigen diese Grenzen. Grundsätzliche Probleme im Umgang mit Sprache, die ein erfahrener Übersetzer kennen und berücksichtigen muss, lassen sich nicht immer einfach lösen. Es setzt viel Erfahrung und Wissen voraus, um gerade in solchen Fällen ein gutes Resultat oder zumindest einen gelungenen Kompromiss als Übersetzung zu liefern, in denen es in der Zielsprache keine offensichtliche Entsprechung für ein Wort oder eine Phrase aus der Quellsprache gibt. Die reine Übersetzung gelingt Programmen heute schon erstaunlich gut, auch dann, wenn es um die nicht immer einfache Sprache literarischer Texte geht. Stellt man aber die Ergebnisse einer maschinellen Übersetzung dem Text eines professionellen Übersetzers gegenüber, wird man schnell auf erhebliche Unterschiede stoßen. Ein professioneller Übersetzer **überträgt** den Text in eine andere Sprache, er bleibt nicht bei einer **Übersetzung** im engeren Sinne stehen.

Dass Literatur so übersetzt wird, hat im einzelnen Fall vielfache Gründe, die hier nicht weiter erörtert werden sollen. Die Übersetzung von Literaturtexten soll das grundsätzliche Themenfeld veranschaulichen. Die Probleme, einen solchen Text **angemessen** in eine andere Sprache zu **übertragen**, stellen sich genauso in ganz anderen Fällen:

- Juristische Texte können neben dem Verständnis des fachlichen Themas auch die Frage grundsätzlicher Unterschiede zwischen verschiedenen Rechtssystemen aufwerfen, die bei einer Übersetzung ebenfalls berücksichtigt werden müssen, besonders wenn es darauf ankommt, folgenschwere Missverständnisse zu vermeiden.
- Technische Beschreibungen (Gebrauchsanweisungen) können sachlogisch extrem zugespitzt sein. Es kommt hier im wahrsten Sinne auf jedes Wort an. Einwandfreie Übersetzungen können auch hier **Übertragungen** beinhalten, um alle wichtigen Details in der Zielsprache zweifelsfrei zu dokumentieren. In einer globalen Welt geht es dabei für die Hersteller medizinischer Systeme, für die Luftfahrtindustrie und diverse weitere Industriezweige um ein essentielles Thema (Haftungsfragen).

Auch wenn speziell neuronale Netze in den letzten Jahren einen deutlichen Fortschritt bei der maschinellen Übersetzung von Texten mit sich gebracht haben, bleibt zu fragen, ob Rechnern alle wichtigen Schichten der Bedeutung (Semantik) von Texten einschließlich der damit zusammenhängenden tieferen Themen überhaupt zugänglich sind oder sein können. Eine brauchbare Rückübersetzung (s.o.) ist für sich allein genommen kein Maßstab für die Qualität einer Übersetzung.

Das Problem der Semantik stellt sich bei der Analyse von Daten in gleicher Weise: Auch dort sind Rechnern, wie sie heute zur Verfügung stehen, sehr enge Grenzen gesetzt, wenn es um das **Verstehen** von Daten geht, auch dort geht es darum, was die relevanten Schichten der Semantik der betreffenden Daten sind und was, z.B. in Unternehmensdaten, aus fachlicher Sicht (z.B. Rechnungslegung und Prüfung) darin **wirklich** enthalten ist.

Soweit der Standpunkt eines **prinzipiellen Skeptikers** in der Frage, ob Rechner in der Lage sind, zu verstehen. Dieser kritischen Sicht soll nun die Position eines **vorsichtigen Optimismus** in dieser Frage gegenübergestellt werden.

Standpunkt 2: Die Semantik von Daten ist nicht prinzipiell unberechenbar

Wirklich tiefgreifende Umwälzungen geben keine neuen Antworten auf alte Fragen, sondern stellen zuallererst die Fragen neu. Das gilt auch heute, wenn die Digitalisierung das Verhältnis von **Syntax** und **Semantik** bei Daten und Texten neu bestimmt. Die Entwicklung hin zur heutigen digitalen Technik hat noch kein *System* hervorgebracht, das Daten oder Texte wirklich **verstehen** kann, aber diese Entwicklung hat gleichwohl *unser* Verständnis der Welt grundlegend verändert.

Der Stand der Technik erlaubt heute auch die Analyse (sehr) großer Datenmengen. Schon einfache Reports bieten in Verbindung mit Expertenwissen tiefere Einsichten in die Semantik der untersuchten Daten:

- in einer Dimension: Schichtungen (Beispiel: Umsätze nach Artikelgruppen)
- in zwei Dimensionen: Kreuztabellen (Beispiel: Umsätze nach Artikelgruppen vs. Regionen)

Die Analyse von Daten nach den einzelnen Ausprägungen eines bestimmten Attributs (mit Anzahl und Beträgen) kann wichtige Rückschlüsse auf die wirtschaftliche Situation oder Entwicklung eines Unternehmens zulassen. Rechner verlängern die fachliche Expertise von Mitarbeitern und bieten **neue, (beliebig) skalierbare Sichten** auf Unternehmensdaten („der Wald und die Bäume"). Dabei ist *nicht* vorausgesetzt, dass Rechner die Daten oder die Bedeutung einzelner Attribute/Felder und deren Ausprägung wirklich verstehen.

Die neue Qualität des Umgangs mit Unternehmensdaten entsteht letztlich nur im Kopf des Betrachters durch die Verbindung seiner Fachlichkeit mit der digitalen Technik. Der Wert eines Analyseergebnisses beruht auf der fachlich richtigen Kombination aus:

- Auswahl der relevanten Daten
- Bestimmung der kritischen Dimensionen (Attribute/Felder) in diesen Daten
- Wahl des Analyseansatzes und der Werkzeuge
- Qualifizierter Interpretation der Ergebnisse

Die Bill Gates zugesprochene Aussage „Wir überschätzen immer die Veränderungen, die in den nächsten zwei Jahren stattfinden werden, und unterschätzen die Veränderungen, die in den nächsten zehn Jahren stattfinden werden" gilt sicher auch für die Frage, ob Rechner Texte und Daten jemals wirklich verstehen.

Wenn es um die kurzfristige Entwicklung der künstlichen Intelligenz geht, werden heute viele in der Regel wenig überzeugende Debatten geführt. Übertriebenen, unrealistischen Erwartungen auf der einen Seite stehen oft grundlose Befürchtungen auf der anderen Seite gegenüber. Die langfristige Entwicklung der künstlichen Intelligenz realistisch zu beurteilen ist sehr schwierig. Es ist nicht möglich, vorherzusagen, ob eine stetige Weiterentwicklung bestehender Konzepte wie „**neuronale Netze**" weitere entscheidende Fortschritte bringt oder ob ganz neue Paradigmen erforderlich sind, um größere Fortschritte in der Datenanalyse, z.B. beim Thema Semantik von Texten und Daten, zu machen.

Daten sind das Öl der heutigen Zeit. Überall werden sie generiert, erhoben und gespeichert. Der wahre Schatz liegt jedoch in der sinnvollen und korrekten Aufbereitung, Auswertung und Interpretation der Datenmengen.

Der aktuelle Tagungsband des Deggendorfer Forums beschäftigt sich mit dieser Thematik und geht dabei auf folgende Aspekte aus den Bereichen Wirtschaftsprüfung, Steuerberatung, Finanzwesen u.a. ein:

- No Signal – verschwindet die Wahrheit in der Flut der Daten?
- Verantwortungsvolle Künstliche Intelligenz
- Discover Cybersecurity Vulnerabilities based on Social Media Data for Internal Audit Analytics
- Statistik ...?! Zu Recht!
- Datenanalysen und Werkzeuge im Prüfungsansatz der Abschlussprüfung
- Visualisierung und Benchmarkstudien bei Verrechnungspreisen
- Künstliche Intelligenz – Eine Lösung für die Wirtschaftsprüfung?

IDW VERLAG GMBH

ISBN 978-3-8021-2732-8
29,00 (D)
www.idw-verlag.de