Thomas Lengauer

Macht, Suggestivität, Grenzen und Risiken der Datenanalyse in den Zeiten von Big Data

In: Grötschel, Martin u.a. (Hg.): Vision als Aufgabe : das Leibniz-Universum im 21. Jahrhundert. – ISBN: 978-3-939818-67-0. – Berlin: [2016], S. 125-145

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-26279

Thomas Lengauer

Macht, Suggestivität, Grenzen und Risiken der Datenanalyse in den Zeiten von Big Data

1 *Prolog: Datenanalyse zur Zeit von Leibniz*

Schon seit langer Zeit lernen wir aus Daten. Der generelle Prozess ist immer der gleiche: In einem ersten Schritt werden in möglichst systematischer Form Daten erhoben. Das kann durch reine Beobachtung des zu untersuchenden Systems geschehen, oder auch in einem kontrollierten Experiment, das systematisch Randbedingungen für die Beobachtungen und gezielte Eingriffe in das System festlegt. In den so erhobenen Daten wird dann in einem zweiten Schritt nach bedeutungsvollen Mustern gesucht. Diesen Prozess nennt man Datenanalyse. Die aufgedeckten Muster werden dann interpretiert; so werden Gesetzmäßigkeiten abgeleitet. Häufig endet der Wissensgenerierungsprozess an dieser Stelle. Im besten aller Fälle wird jedoch nach Aufdeckung der Gesetzmäßigkeiten in einem dritten Schritt nach kausalen Zusammenhängen gesucht, die die Gesetzmäßigkeiten erklären, das heißt, auf fundamentalere Prinzipien zurückführen.

Hinweise auf systematische Sammlung und Analyse finden sich bereits bei den Babyloniern (Grasshoff 2012). Betrachten wir an dieser Stelle ein Beispiel, das häufig als eine der ersten systematischen Datenanalysen der Neuzeit angesehen wird. Es geht um die Erkenntnisgewinnung über die Bahnen der Planeten im Sonnensystem und die diesbezüglichen naturwissenschaftlichen Grundlagen. Der Prozess begann mit einer mehrere Jahrzehnte währenden peniblen Aufzeichnung der Koordinaten von Planetenbahnen durch den dänischen Astronomen Tycho Brahe (1546–1601). Johannes Kepler, der im letzten Lebensjahr von Tycho Brahe sein wissenschaftlicher Assistent war, hat nach dessen Tod seine Datensammlung weiter bearbeitet und im Jahr 1627 in den Tabulae Rudolphinae (Kepler 1627) veröffentlicht. Er widmete sich auch der Datenanalyse. Das Ergebnis waren die drei sogenannten Keplerschen Gesetze in den ersten Jahrzehnten des 17. Jahrhunderts (siehe vor allem Kepler 1609, 1619), die die Geometrie der Umlaufbahn eines Planeten sowie die Zusammenhänge zwischen seiner Umlaufgeschwindigkeit und dem Abstand von seinem Zentralgestirn mathematisch formulieren. Es ist wichtig, zu betonen, dass die drei Keplerschen Gesetze keine Erklärung für die gefundenen Formeln darstellen. Deswegen sprechen wir im Zusammenhang dieses Aufsatzes lieber von „Gesetzmäßigkeiten" als von Gesetzen. Das sogenannte erste Keplersche Gesetz formuliert genau die Muster, die in den Daten von Tycho Brahe über die Form der Planetenbahnen und Umlaufgeschwindigkeiten aufzufinden waren. Keplers Begründung beruhte auf Renaissance-Vorstellungen über Kräfte und Seele, stellte aber einen Versuch zur Physikalisierung der Planetentheorie dar (Stephenson 1987). Wie wir im Folgenden sehen werden, ist schon die Auffindung solcher Muster in Daten höchst nützlich und kann in vielen Fällen die Grundlage für nachfolgende Entscheidungen und Strategien bilden. Im Fall der Planetenbahnen fand der Prozess der Wissensgewinnung hier (glücklicherweise und

bezeichnend für das Vorgehen in der Wissenschaft) kein Ende. Vielmehr hat Isaac Newton (1642–1727), der ein Zeitgenosse, Kollege und auch Kontrahent von Leibniz war, auf der Basis der Keplerschen Gesetze und anderer Beobachtungen sein Gravitationsgesetz abgeleitet und im Jahre 1687 in seinem Werk *Philosophae Naturalis Principia Mathematica*, kurz *Principia* genannt, veröffentlicht (Newton 1687). Das Gravitationsgesetz stellt eine fundamentale „Neuerklärung" für die Keplerschen Gesetze dar. Genauer gesagt lassen sich die Keplerschen Gesetze aus dem Newtonschen Gravitationsgesetz mathematisch herleiten. Das Newtonsche Gravitationsgesetz stellte in der Nachfolge die axiomatische Grundlage für die Himmelsmechanik dar. Das heißt, es wurde aus der Newtonschen Perspektive als gegeben und nicht weiter erklärungsbedürftig angenommen. Diese Lage hat sich erst mit der Entwicklung der allgemeinen Relativitätstheorie im Jahre 1915 (Einstein 1916; 2009) durch Albert Einstein (1879–1955) geändert, aus der sich nun wiederum das Newtonsche Gravitationsgesetz (als Grenzfall für langsame Geschwindigkeit im Verhältnis zur Lichtgeschwindigkeit und schwache Gravitation) ableiten lässt.

2 *Datenanalyse und Theoriebildung: Was ist anders bei Big Data?*

Der im Prolog geschilderte Prozess ist ein Paradebeispiel für akkurat durchgeführte Datenanalyse, aber als Tabellenwerk mit insgesamt etwas über 250 Seiten sicher keines für Big Data. Bezeichnender als der Umfang der Datensammlung ist die Tatsache, dass die relevanten Muster, mathematisch formuliert in den Keplerschen Gesetzen, manuell von Johannes Kepler durch Sichtung der Datensammlung abgeleitet werden konnten. Die Datensammlungen, die heutigen Big Data Analysen zugrunde liegen, sind zum einen wesentlich umfangreicher. Zum anderen sind die in ihnen enthaltenen Muster häufig sehr komplexer Natur. Aus diesen beiden Gründen ist die Auffindung relevanter Muster in solchen Datenbeständen ohne Zuhilfenahme ausgeklügelter Algorithmen und Einsatz von Computern praktisch nicht mehr möglich.

Ein zweites historisches Beispiel führt uns in den Bereich der Big Data-Analysen. Hier beginnt die Datensammlung mit der Gründung eines Hochtechnologie-Instituts, nämlich der Physikalisch-Technischen Reichsanstalt in Berlin im Jahr 1887. Das Institut wurde zum Zwecke der Entwicklung und des Einsatzes hochpräziser Messinstrumente für die Industrieproduktion geschaffen. Ein Anwendungsziel war die präzise Messung von Lichtspektren, die von erhitzten Festkörpern ausgesandt wurden, die so genannte Schwarzkörperstrahlung. Anwendungshintergrund war die Produktion von Glühbirnen mit kontrollierten Farbeigenschaften. Es dauerte knapp zehn Jahre, bis die Spektren mit einer Genauigkeit gemessen werden konnten, die Lücken im physikalischen Verständnis der zu Grunde liegenden Eigenschaften des Lichtes zu Tage treten ließen und damit die Geburtsstunde der Quantenmechanik einläuteten. Die Datensammlung wurde hier von Ingenieuren der Physikalisch-Technischen Reichsanstalt vorgenommen, die Datenanalyse unternahmen Physiker wie Wilhelm Wien und Max Planck. Max Planck führte zur Angleichung der Formel für das Spektrum der Schwarzkörperstrahlung an die gemessenen Spektren das Quantum ein (Planck 1900), das er selbst lange Zeit nicht erklären konnte. Hier handelt es sich also wiederum nicht um ein Gesetz, das das Muster im Sinne von kausalem Verständnis erklärt, sondern lediglich um eine Gesetzmäßigkeit im Sinne der Beschreibung eines beobachteten Musters. Es hat fast 30 Jahre gedauert, bis durch eine ganze Reihe von Physikern,

unter ihnen insbesondere Werner Heisenberg, Max Born und Paul Jordan auf der einen Seite (Heisenberg 1925; Born, Heisenberg und Pascual 1926; Born und Pascual 1925) sowie Erwin Schrödinger auf der anderen (Schrödinger 1926a, 1926b, 1926c, 1926d), im Jahr 1925 die theoretische Fundierung der Quantenmechanik bereitgestellt werden konnte. Seitdem haben wir in den Grundannahmen der Quantenmechanik eine axiomatische Grundlage für die inneratomaren physikalischen Vorgänge, aus denen sich die von Max Planck entdeckten Muster ableiten lassen.[1] Im Falle der Quantenmechanik setzte sich der Wissensgenerierungsprozess jedoch weiter fort. Jetzt übernahmen die Theoretiker die Führung, indem sie die Mathematik weiter trieben und mit Hypothesen über die subatomare Welt aufwarteten, die im Nachhinein durch aufwändige Messungen verifiziert oder falsifiziert werden konnten. So begann nach dem Zweiten Weltkrieg eine fieberhafte Suche nach Elementarteilchen, die durch theoretische Überlegungen hypothetisiert wurden und in zahlreichen Fällen mit dem Einsatz von Teilchenbeschleunigern nachgewiesen werden konnten.[2] Die Krönung dieses Prozesses ist derzeit wohl die Entdeckung des von Peter Higgs (und zeitgleich anderen Wissenschaftlern) im Jahr 1964 postulierten Elementarteilchens (Higgs 1964; Englert und Brout 1964; Guralnik, Hagen und Kibble 1964) im Jahr 2012 durch große Wissenschaftlerteams am Forschungszentrum CERN (Aad et al. 2012; Chatrchyan et al. 2012). Die umfangreiche Datensammlung und Datenanalyse, die zur Auffindung des Higgs-Teilchens führte, gehört ohne Zweifel in den Bereich von Big Data (Adam-Bourdarios et al. 2015).

In teilweiser Abweichung von den Kriterien, die landläufig für Big Data genannt werden,[3] möchte ich hier folgende Vorbedingungen für Big Data-Sammlungen und -Analysen nennen.

(i) Das Datenaufkommen muss sehr groß sein. Ich verlange hier nicht die Größe von Petabytes oder mehr, die häufig für den Begriff Big Data ins Feld geführt wird. Für die Zwecke dieses Aufsatzes reicht es, eine Größe anzunehmen, die eine Datenanalyse durch komplexe Computeralgorithmen erfordert.

(ii) Die Datenanalyse muss der Engpass bei der Wissensgenerierung sein, nicht die Datengenerierung. Es ist ein Charakteristikum von Big Data-Analysen, dass uns die Daten mit vergleichsweise geringem Aufwand zur Verfügung stehen. Dies bedeutet eine Umkehrung gegenüber klassischer Datenanalyse, bei der in der Regel die Datengenerierung wesentlich komplexer und teurer ist als die Datenanalyse. Dies war definitiv im Fall von Tycho Brahe und auch im Fall der Entwicklung der Quantenmechanik so. Heute dagegen fallen uns Daten über Hochdurchsatzexperimente in der Wissenschaft und über das Internet im täglichen Leben in einem Umfang zu, den man als lawinenartig bezeichnen kann. Damit wird die Datenanalyse zum Engpass des Prozesses der Wissensgenerierung.

(iii) Datenanalyse ist von Hand nicht mehr möglich. Dies trifft auf die beiden oben genannten Beispiele, die Ableitung der Himmelsmechanik und der Quantenmechanik, nicht zu.

(iv) Durch mächtige statistische Verfahren, die man häufig auch als Data Mining oder maschinelles Lernen bezeichnet, sind mit Hilfe des Computers auch komplexe Muster in den Daten auffindbar. Damit ist die Anwendung statistischer Verfahren der wesentliche Schritt der Wissensgenerierung bei Big Data.

Big Data begegnet uns heute in allen Bereichen des Lebens und der Wissenschaft. Im täglichen Leben finden seit der Entstehung des Internets und durch die zunehmende Vernetzung der

Technologien massive Datensammlungen in den verschiedensten Bereichen statt. Daten werden gesammelt, wenn immer wir mit unserem Rechner ins Internet gehen, wenn wir fernsehen, wenn wir Auto fahren, wenn wir unser Handy benutzen, wenn wir Bankgeschäfte betreiben oder einkaufen. Im öffentlichen Leben werden unsere Spuren von Webcams erfasst. Haushaltsgeräte enthalten zunehmend vernetzte Intelligenz, und Vernetzung und Datensammlung setzen sich auch bei der Energieversorgung durch.

Desgleichen ist Big Data zunehmend ein wesentlicher Bestandteil von praktisch allen Wissenschaftsdisziplinen. Die Elementarteilchenphysik wurde bereits erwähnt. In der Astronomie werden detaillierte dreidimensionale Modelle des gesamten Universums und der in ihm enthaltenen Sterne und Galaxien entwickelt (Ivezić et al. 2012). Die Geo-und Umweltwissenschaften (Vitolo et al. 2015; Sellars et al. 2013; Guo, Zhang und Zhu 2015) sammeln eine Vielzahl von Daten über divergente Aspekte des Zustandes unseres Planeten und projizieren diese Daten in die Zukunft und in die Vergangenheit. In der Chemie werden sowohl umfassende Datenbanken über chemische Verbindungen und deren Eigenschaften gesammelt als auch durch quantenmechanische Berechnungen entstehende umfassende Datensätze weltweit zugänglich gemacht (Ghiringhelli et al. 2015; Lusher et al. 2014; Rajan 2015). Die Wirtschafts- und Sozialwissenschaften legen ihren Untersuchungen umfangreiche und vielseitige Datensammlungen zu Grunde (Hesse, Moser und Riley 2015; Levin und Einav 2014). Und der Zugang zu vollständiger genomischer Information war eine wesentliche Voraussetzung für die Wandlung von Biologie und Medizin in in hohem Maße durch molekulare Daten getriebene Wissenschaften (Pennisi 2010; Bernstein et al. 2012).

Über Big Data ist schon viel gesagt und geschrieben worden. Wir wollen uns auf einen besonderen wissenschaftsrelevanten Aspekt von Big Data-Analysen konzentrieren, nämlich das Spannungsfeld zwischen der Aufdeckung von Gesetzmäßigkeiten, die Muster beschreiben und von Gesetzen, die Muster erklären. Nach der Einführung des für diesen Diskurs grundsätzlichen Modellbegriffs werden wir dies an zwei Beispielen von modernen Datenanalysen in Biologie und Medizin tun.

3 *Mathematische Modelle: Vorhersagen und Erklärungen*

Gehen wir ein wenig genauer auf den Prozess der Generierung quantitativen Wissens aus Daten ein. Der zentrale Begriff in diesem Zusammenhang ist der des mathematischen Modells (Imboden und Koch 2008). Ein mathematisches Modell ist eine mathematische Vorschrift, mit der unter Bezugnahme auf wesentliche Parameter über einen begrenzten Ausschnitt der Wirklichkeit Vorhersagen über diesen Ausschnitt der Wirklichkeit getroffen werden können. Diese Vorhersagen sind in der Regel mit Ungenauigkeiten behaftet, die jedoch sehr unterschiedliche Größenordnungen annehmen können. Grundsätzlich gibt es für mathematische Modelle zwei Qualitätskriterien: Vohersagegenauigkeit und Erklärungsvermögen. Das wesentliche Kriterium ist das der Vorhersagegenauigkeit. Ein Modell, das genaue Vorhersagen ermöglicht, ist einem Modell überlegen, das dies nicht tut. Wir streben nach Modellen mit hoher Vorhersagegenauigkeit und knüpfen die Tiefe unseres Verständnisses des betrachteten Ausschnittes der Wirklichkeit an die Vorhersagegenauigkeit der Modelle.

Ich möchte jedoch als zusätzliches Kriterium das des Erklärungsvermögens hinzufügen. Nicht jedes Modell, das eine hohe Vorhersagegenauigkeit hat, erklärt auch viel. Wie wir sehen werden, sind die meisten Modelle, die sich aus Datenanalysen im Big Data-Bereich ergeben, statistischer Natur: Sie treffen Wahrscheinlichkeitsaussagen über den Zustand eines Systems, gegründet auf seiner Beschreibung durch die verfügbaren Werte der im Allgemeinen sehr zahlreichen betrachteten Parameter. Zum Beispiel würde ein solches Modell auf der Basis von über einen Patienten erhobener medizinischer Information, darunter komplexer Information über sein Genom, eine Aussage über die Wahrscheinlichkeit der Wirksamkeit eines Medikaments treffen. In vielen Fällen kann das Modell sogar eine Aussage über die Zuverlässigkeit seiner eigenen Vorhersage treffen. Bei komplexen Systemen ist die Vorhersage jedoch immer mit Unwägbarkeiten behaftet, und das Modell ist auch schwer zu verstehen – sein Erklärungsvermögen ist gering. Anders formuliert: Die Frage, warum das Modell im gegebenen Fall genau diese Vorhersage macht, bleibt oft unbeantwortet. Beides, die geringe Vorhersagegenauigkeit und das geringe Erklärungsvermögen, liegen daran, dass statistische Datenanalysen nur Assoziationen zwischen den zahlreichen Systemparametern – also die Häufung gemeinsam auftretender Werte verschiedener Parameter – aufdecken und keine kausalen Zusammenhänge nachweisen. Die Statistik stellt die Behelfslösung dar, mit der wir den Mangel an Verständnis des betrachteten Systems auszugleichen suchen. Wir behandeln nicht verstandene Aspekte als Streuung oder Rauschen, und das erhöht die Ungenauigkeit in unseren Vorhersagen und mindert das Erklärungsvermögen.

Wenn man bekannte kausale Abhängigkeiten im System in das mathematische Modell einbringen kann, kann man in der Regel die Genauigkeit der Vorhersage wesentlich erhöhen. Darüber hinaus bietet das Verständnis kausaler Zusammenhänge auch eine tiefergehende Erklärung der modellierten Sachverhalte. Kausale Zusammenhänge sind jedoch mit statistischen Methoden bis heute nur recht schwer nachzuweisen. Und sie sind in komplexen Systemen allgemein nur schwer aufzudecken. Wie wir noch sehen werden, ruht hierin eine der wesentlichen Problematiken der Datenanalyse bei Big Data.

Illustrieren wir diese Begriffe nun an dem Beispiel Brahe–Kepler–Newton. Der betrachtete Ausschnitt der Realität ist die Himmelsmechanik. Tycho Brahe stellte die Datensammlung zur Verfügung. Johannes Kepler entwickelte das erste mathematische Modell – die Keplerschen Gesetze. Dieses Modell ist nicht statistisch. Das Universum ist so „einfach", dass hier tatsächlich im ersten Schritt deterministische Gleichungen eine sehr hohe Vorhersagegenauigkeit erbrachten – es trat kein wesentliches Rauschen auf. Dennoch sind die Keplerschen Gesetze eher beschreibend als erklärend: Es bleibt unklar, was die treibende Kraft der Himmelsmechanik ist. Diese Frage hat Newton beantwortet, indem er die Keplerschen Gesetze auf das von ihm formulierte Gravitationsgesetz zurückführte. Dieses Gesetz hat eine weit umfassendere Anwendbarkeit als die Keplerschen Gesetze und daher eine größere Erklärungskraft. Die Genauigkeit der Berechnung der Planetenbahnen erhöhte sich damit auch, weil der Einfluss mehrerer Körper aufeinander berücksichtigt werden kann – etwas, das die Keplerschen Gesetze nicht bieten. In einem weiteren Schritt hat Einstein mit seiner allgemeinen Relativitätstheorie sowohl die Genauigkeit der Vorhersagen erhöht als auch ihren Anwendungsbereich erweitert. Es ist interessant, zu bemerken, dass er dies im Wesentlichen ganz ohne komplexe Datenanalyse getan hat. Er nahm lediglich die Konstanz der Lichtgeschwindigkeit zum Ausgangspunkt für seine Gedankenexperimente.

In den weiteren Abschnitten werden wir das Wechselspiel zwischen Datenanalyse und Modellbildung an zwei Beispielen betrachten und dabei auf die in diesem Abschnitt eingeführten Begriffe und ihre Zusammenhänge zurückkommen.

4 *Biologie – Big Data vs. Naturgesetze: Wann und warum ist Big Data erfolgreich?*

Die Biologie ist ein für die Untersuchung unserer Fragestellung besonders geeignetes Gebiet, handelt es sich hierbei doch um eine Wissenschaft mit Zwittercharakter. Ziel der Biologie ist das Verständnis von lebenden Organismen in ihrer Funktion und ihren Wechselwirkungen. Zum einen ist dies eine Naturwissenschaft. Letztlich ist Leben ein molekularer Prozess. Lebende Organismen sind komplexe molekulare Systeme. Das Leben manifestiert sich in der Struktur und Dynamik dieser Systeme. Moleküle aber gehorchen den Naturgesetzen, die durch Physik und Chemie gegeben sind. Und in diesem Bereich gibt es ein hohes Maß an Theorie. Die entsprechenden Naturgesetze sind bereits sehr gut verstanden. Also ist die Biologie eigentlich nichts weiter als eine Chemie von speziellen hochkomplexen molekularen Systemen, die den bekannten Naturgesetzen folgen. Leider sind die betrachteten molekularen Systeme aber so komplex, dass das theoretische Verständnis nicht zu genauen Vorhersagen führt: Die Theorie anzuwenden übersteigt die verfügbaren Rechenressourcen bei weitem.

Auf der anderen Seite hat die heute zu beobachtende Biosphäre einen immens langen und komplexen Entwicklungsprozess hinter sich. Die heutigen Organismen sind evolutiv über Milliarden von Jahren entstanden. Dabei wurden extrem große Populationen einer hochkomplexen Umwelt ausgesetzt. Im Wechselspiel zwischen Variation und Selektion entstand die heutige Ausprägung des Lebens. Dieser evolutive Charakter des Lebens ist so grundlegend, dass er als eine eigene formative Kraft angesehen werden muss, deren Verständnis sich letztlich nicht auf rein physikochemische Prozesse reduzieren lässt. Und für die Evolution fehlt es uns noch hinlänglich an Theorie. Der Erfolg der Datenanalyse in der modernen Biologie begründet sich vor allem darin, dass sie in Abwesenheit von stark ausgebildeten theoretischen Grundlagen ein sehr wirksames und eigentlich das einzig verfügbare Instrument für Wissensgenerierung darstellt.

Illustrieren wir dies kurz an dem Problem der Bestimmung der dreidimensionalen Struktur von Proteinen, also Eiweißmolekülen. Proteine sind die molekularen Maschinen unseres Körpers, im Speziellen, und von lebenden Organismen, im Allgemeinen. Proteine sind langkettige Moleküle, die sich aus Abfolgen von Aminosäuren zusammensetzen. Das irdische Leben verwendet zwanzig Aminosäuren. Sie stellen quasi einen universellen chemischen Bausatz aus kleinen Molekülen dar, mit dem das Leben durch Verkettung komplexe Moleküle mit hochspezifischen Funktionen bildet, die Proteine. Im menschlichen Körper gibt es schätzungsweise eine Million verschiedene Proteine. Die Funktion eines Proteins manifestiert sich in seiner dreidimensionalen Struktur, und Proteine sind in ihren Strukturen extrem divers (siehe Abb. 1). Viele unserer Proteine nehmen nach ihrer Synthese spontan eine eindeutige dreidimensionale Struktur an. Proteinstrukturen sind bis heute sehr schwer zu vermessen. Daher ist es sehr hilfreich, wenn man aus der Kenntnis der Proteinsequenz mit Computeralgorithmen ein Modell der Proteinstruktur ableiten kann. Die Struktur eines Proteins ist letztlich in der Abfolge der Aminosäuren im

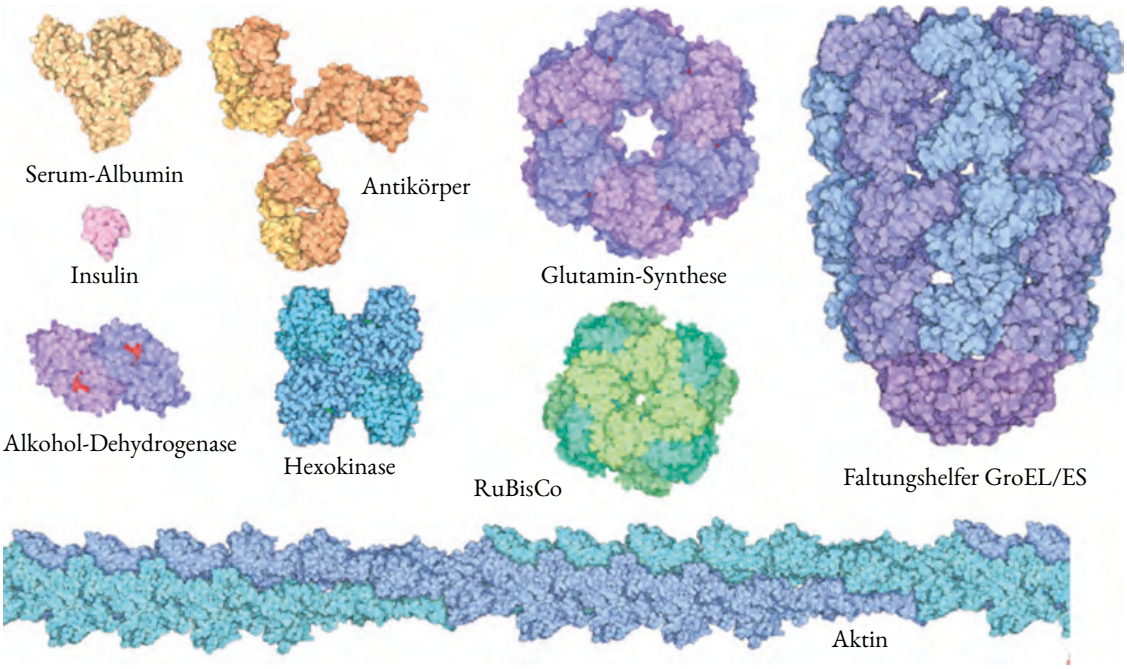


Abbildung 1. Dreidimensionale Strukturen diverser Proteine. Die angegebenen Kürzel identifizieren die Einträge der Proteine in der Protein Structure Database. (Auswahl aus „Molecular Machinery: A Tour of the Protein Data Bank" by David S. Goodsell [http://mm.rcsb.org])

Protein, also seiner Sequenz, begründet und sollte sich daher auch prinzipiell rechnerisch aus der Proteinsequenz ableiten lassen. Dieses so genannte Proteinstrukturvorhersage-Problem hat genau den Zwittercharakter, den wir eben dem gesamten Feld der Biologie zugesprochen haben (Dill und MacCallum 2012; Wolynes 2015).

Zum einen gehorcht das Protein bei seiner Faltung, also dem Prozess, mit dem es seine dreidimensionale Struktur annimmt, den Gesetzen der Thermodynamik – und die besagen, dass es sich in den energetisch günstigsten Zustand faltet. Die entsprechenden Formeln für die Berechnung der Energie eines eine bestimmte Struktur annehmenden Proteins sind seit der Entwicklung der Thermodynamik und Quantenmechanik bekannt. Das Protein ist jedoch ein so komplexes Molekül – es kann viele tausend Atome enthalten – dass die Formeln in der Praxis nicht berechnet werden können. Wir sind hier also in der misslichen Lage, dass wir eigentlich alle Theorie für die Berechnung der Proteinstruktur zur Verfügung haben, sie aber aufgrund der Komplexität des zu betrachten Moleküls nicht anwenden können.

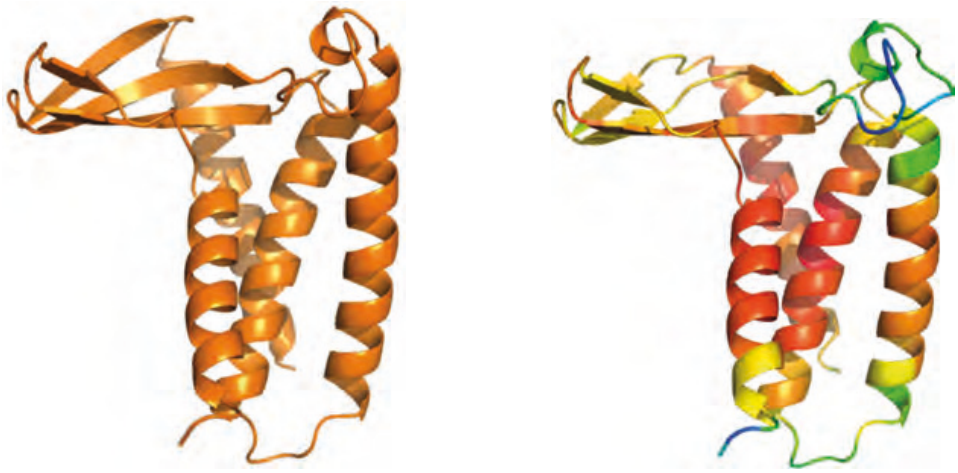
Auf der anderen Seite sind die Proteine, die das Leben heute bereitstellt, durch den langen Prozess der evolutionären Entwicklung gegangen. Die Gesamtheit aller vom Leben hervorgebrachten Proteine ist mitnichten zufällig zusammengewürfelt. Vielmehr ergeben sich tiefgehende und oft komplexe Verwandtschaften zwischen Proteinen in verschiedenen Organismen und auch zwischen verschiedenen Proteinen in demselben Organismus. Diese Verwandtschaftsbeziehungen führen dazu, dass etwa Proteine, deren Aminosäureabfolge sich in weniger als etwa 65–70 % der Aminosäuren unterscheidet, mit sehr hoher Wahrscheinlichkeit fast identische

dreidimensionale Strukturen haben (Sander und Schneider 1991; Rost 1999). Wir können also die Struktur eines Proteins, dessen Aminosäuresequenz wir kennen, dann mit hoher Genauigkeit ableiten, wenn wir ein weiteres in der Aminosäuresequenz ähnliches Protein kennen, dessen Struktur uns bereits bekannt ist (Fiser et al. 2002; Dunbrack 2007). Dieses Protein wird in diesem Zusammenhang als strukturelles Templat bezeichnet. Genau das ist die Grundlage von Algorithmen, die aus der Aminosäuresequenz eines Proteins eine dreidimensionale Struktur ableiten. Solche Algorithmen greifen auf Datenbanken von Proteinen zurück, deren Sequenz und Struktur wir kennen. Dies gilt heute für etwa 110 000 Proteine, die in der *Protein Data Bank* (www.pdb.org) zusammengefasst sind. Dann wird mit einer Data Fitting-Prozedur die Struktur des zu modellierenden Proteins aus der Struktur des Templats abgeleitet. Der letzte Verfeinerungsschritt dieser Methode benutzt dann häufig tatsächlich theoretische Grundlagen von Mechanik und Thermodynamik.

Das geht immer dann gut, wenn wir in unseren Datenbanken tatsächlich ein Struktur-Templat für unser Protein finden. Vor einigen Jahren konnte dieses Verfahren sogar auf den Fall erweitert werden, dass kein Struktur-Templat für unser Protein vorliegt. Es stellte sich nämlich heraus, dass, wenn wir nur genug sequenzähnliche Proteine haben, in den Beziehungen zwischen deren Aminosäuresequenzen untereinander subtile Signale über Aspekte der dreidimensionalen Struktur des Proteins enthalten sind, etwa darüber, ob sich zwei Aminosäuren im Raum nah oder fern stehen. Wenn man diese Signale geeignet auswertet und wenigstens 1000 Sequenzen unserem Protein sequenzähnlicher Proteine zur Verfügung stehen, ist es auch ohne ein Struktur-Templat möglich, unser Protein zu modellieren, wenn auch mit geringerer Zuverlässigkeit (Mars et al. 2011; Kamisetty, Ovchinnikov und Baker 2013). Schwede (2013) gibt eine Übersicht über den derzeitigen Stand der algorithmischen Modellierung von Proteinstrukturen.

Wir haben es hier also mit einer Methode der Datenanalyse zu tun. Es wird kein Naturgesetz bemüht, sondern die Struktur wird assoziativ aus Daten über ähnliche Proteine abgeleitet. Obwohl der Umfang der Daten, die benutzt werden, nicht an das Volumen heranreicht, das sonst im Big Data-Kontext zu beobachten ist, rechne ich die Methode doch zu Big Data, weil sie die drei in Abschnitt 1 aufgeführten Kriterien erfüllt. Es ist eine ganz wesentliche Bemerkung, dass die mit solchen Methoden abgeleiteten Strukturmodelle, die allesamt statistischer Natur sind, mit nicht zu vernachlässigenden Unsicherheiten und Ungenauigkeiten behaftet sind. Manchmal sind sie einfach grundfalsch. Wie in Abschnitt 2 ausgeführt, liegt das daran, dass unsere Data Mining-Methode zur Modellierung von Proteinstrukturen keine Kausalzusammenhänge, sondern nur assoziative Muster auswertet. Deshalb ist es von großer Wichtigkeit, dass eine Methode zur Modellierung von Proteinen mit dem abgeleiteten Modell auch eine Schätzung für die Zuverlässigkeit dieses Modells zur Verfügung stellt. In der Tat gibt es solche Schätzungen sowohl für das Strukturmodell als Ganzes als auch für lokale Bereiche des Strukturmodells, etwa die besonders wichtigen Bindungsstellen eines Proteins, an die andere Moleküle andocken (Kihara, Chen und Yang 2009), siehe Abb. 2.

Aufgrund der Tatsache, dass lebende Systeme so komplex sind, dass sie sich praktisch nie durch direkte Anwendung der physikochemischen Grundlagen analysieren lassen, ist die Bioinformatik, die molekularbiologische Daten zum Verständnis von lebenden Systemen mit Computern auswertet, im Wesentlichen eine datengetriebene Wissenschaft, d. h. sie gehört in den Bereich von Big Data (Lengauer, Albrecht und Domingues 2012).



(a) (b)

Abbildung 2. Nach einem Templat modellierte Proteinstruktur. Sowohl das Templat als auch das modellierte Protein sind für die Entwicklung von Antibiotika relevante bakterielle Proteine (Lipoprotein Signalpeptidase II) mit ähnlicher Funktion, die aber in verschiedenen Organismen vorkommen. Die Proteine sind in der „Cartoon"-Version dargestellt, die eine geglättete Form des Proteinrückgrates angibt. (a) Struktur des Templatproteins in konstant oranger Farbe. (b) Nach Zuverlässigkeit eingefärbtes Strukturmodell des Proteins. Zuverlässige Strukturteile des Modells sind rot, weniger zuverlässige blau eingefärbt. Die von der Templatstruktur weiter abweichenden Teile des Modells haben generell eine geringere Zuverlässigkeit. Die Modellierung und Bewertung der Zuverlässigkeit des Modells wurden mit dem SwissModel Server durchgeführt (Benkert, Tosatto und Schomburg 2008; Biasini et al. 2014). Bei Drucklegung lag keine experimentell vermessene Struktur dieses Proteins vor. (Modellierung von Olga Kalinina)

5 *Medizin – Big Data zum Wohl des Patienten: Warum müssen wir nicht immer die Ursachen verstehen?*

Die Medizin ist wie die Biologie ein Teilgebiet der Lebenswissenschaften und teilt mit ihr die im letzten Beitrag beschriebenen Zwittereigenschaften. Sie beschäftigt sich mit den Unterschieden zwischen gesunden und kranken menschlichen Körpern. Körperliche Krankheiten haben in aller Regel eine konkrete molekulare Grundlage. Sie manifestieren sich in Deregulierungen von hochkomplexen molekularen Wechselwirkungsnetzen in oder zwischen Zellen unseres Körpers. Eine physikochemische Analyse dieser Grundlagen auf der Basis der Anwendung von Naturgesetzen ist jedoch in aller Regel nicht möglich. Das heißt, wir müssen auch hier wieder datengetrieben vorgehen. Dies ist an sich nichts Neues. Medizin war schon immer datengetrieben. Der Arzt hat seine Diagnose und Therapie aus einer Sammlung von Informationen über die Krankheitsgeschichte des Patienten, Laborwerten und seiner beruflichen Erfahrung abgeleitet. In gewisser Weise hat er auf der Basis der ihm zur Verfügung stehenden Information ein Modell des Patienten abgeleitet, aufgrunddessen er therapeutische Entscheidungen trifft. Jedoch sind weder das Modell noch der Prozess seiner Ableitung mathematischer oder durchgängig systematischer Natur. Im Zeitalter von Big Data erfährt diese Vorgehensweise jetzt ein hohes Maß an Mathematisierung und Systematisierung. Dabei gewinnt der Aspekt der gemessenen Laborwerte, insbesondere solcher mit genomischem Charakter, und ihrer mathematischen Interpretation mit Computerhilfe zunehmende Bedeutung.

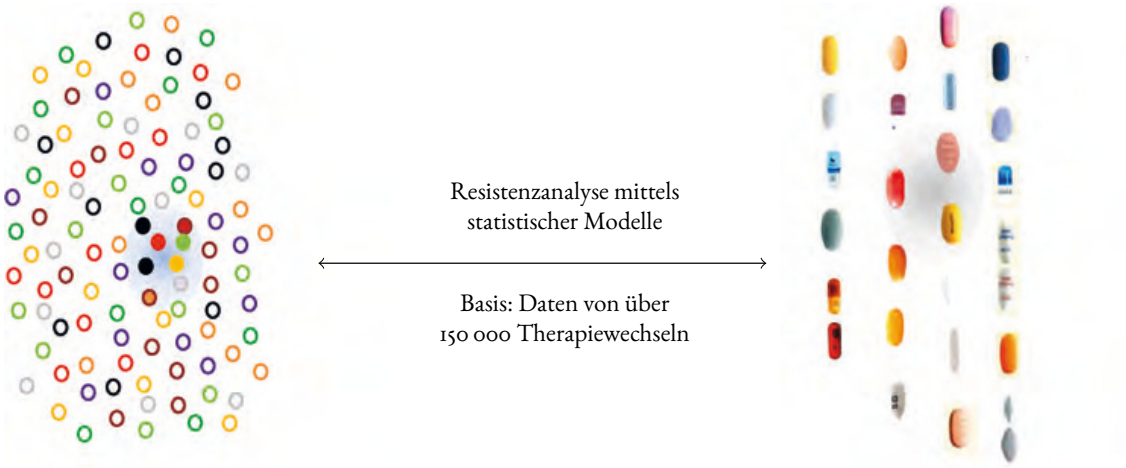


Abbildung 3. Mit Hilfe von Data Mining in großen Datenbanken zur Resistenz von HIV gegen Medikamente werden aus einer Menge von über zwei Dutzend HIV-Medikamenten (rechts) Medikamentenkombinationen (grau hinterlegt) vorgeschlagen, die gegen die im Patienten vorherrschende Population von HI-Viren (links, grau hinterlegt) wirksam sind.

Wir illustrieren dies am Beispiel der HIV-Therapie. Das HI Virus verändert sich im Körper des infizierten Patienten äußerst schnell. Hier findet dieselbe Art von Prozess statt, die wir auch von der Entwicklung von Antibiotika-Resistenz bei Bakterien kennen. Nur können wir ihn in diesem Fall nicht nur in der Gesamtheit aller Infizierten, sondern im einzelnen Patienten und über wesentlich kürzere Zeiträume, nämlich Tage, Wochen oder Monate beobachten. Ein HIV-Patient beherbergt eine ganze Vielfalt von ähnlichen, aber unterschiedlichen HI-Viren, und diese Viren verändern sich ständig, um den Angriffen des Immunsystems des Patienten und der bestehenden Medikamenten-Therapie zu entgehen. Aus diesem Grund gibt es heute über zwei Dutzend verschiedene Medikamente gegen HIV, die dem Patienten in einer Kombination von etwa drei Medikamenten verabreicht werden. Diese Kombination führt zu über tausend möglichen Therapieoptionen. Ob ein Virus resistent gegen ein bestimmtes Medikament ist, ist im viralen Genom codiert, allerdings auf eine Art und Weise, die nicht ohne weiteres für das menschliche Auge ersichtlich ist, und die im Labor auch schwer zu messen ist. Wir haben es also mit einem typischen Datenanalyse-Problem zu tun (Lengauer und Kaiser 2009). Die Eingabe für das Problem besteht aus der Sequenz des im Patienten vorrangig zu findenden Virus. (Seit neuestem kann man auch die Gesamtheit der im Patienten vorkommenden Viren mit hoher Genauigkeit vermessen [Thielen und Lengauer 2012].) Die Ausgabe ist eine Liste der geschätzten Resistenzen des Virus gegen das Sortiment der verfügbaren HIV-Medikamente (siehe Abb. 3). Die Datenbank (der Big Data-Aspekt des Problems) besteht in diesem Fall aus einer internationalen Sammlung von Daten über mehr als 150 000 Therapieepisoden von HIV-Patienten (Zazzi et al. 2012), siehe auch www.euresist.org. Aus diesen Daten werden mit Data Mining-Methoden, auf die wir im nächsten Abschnitt kurz eingehen werden, statistische Modelle abgeleitet, die, nach Eingabe einer viralen Sequenz, die Resistenz des Virus gegenüber den verfügbaren Medikamenten schätzen.

Wir stellen ein solches Beratungssystem für die HIV-Therapie unter www.geno2pheno.org frei über das Internet zur Verfügung (Beerenwinkel et al. 2002; Lengauer und Sing 2006). Das System wird über Deutschland hinaus zur Behandlung von HIV-Patienten eingesetzt. Natürlich sind auch die Schätzungen dieses Systems nicht fehlerfrei, da sie ja einer Datenanalyse entspringen und keine Kausalzusammenhänge auswerten. In einigen unserer Analysearten bieten wir daher auch Zuverlässigkeitsschätzungen an.

Dies ist die geeignete Stelle, um über Nutzen und Grenzen datengetriebener Methoden zu sprechen. Rufen wir uns das Bild des Prozesses zur Wissensgenerierung ins Gedächtnis, das mit Datensammlung beginnt, auf die Datenanalyse und Modellbildung sowie schließlich Theoriebildung, das heißt die Aufdeckung kausaler Zusammenhänge folgen. Wir erkennen, dass der dritte Schritt, nämlich der der Theoriebildung, sowohl bei unserem biologischen Beispiel aus dem letzten Abschnitt als auch bei diesem medizinischen Beispiel fehlt. Während unsere Vorhersagen – der Proteinstruktur oder der Resistenz des Virus gegen Medikamente – durchaus einen hohen Genauigkeitsgrad haben können, können wir über die kausalen Zusammenhänge in beiden Fällen nur Vermutungen anstellen. Dies manifestiert die Grenzen der Datenanalyse. Warum aber sollte eine Methode, die lediglich auf Datenanalyse beruht und keine kausalen Zusammenhänge zu Medikamentenwirkung aufdeckt oder verwendet, medizinisch nützlich sein?

Hier werden wir uns der Macht der Datenanalyse bewusst. Sie besteht in der Fähigkeit, auch bei Systemen von hoher Komplexität, die wir theoretisch nicht durchdrungen haben, zu Vorhersagen zu kommen. Der Nutzen dieser Vorhersagen beruht vor allem darauf, dass sie in vielen Fällen akkurat sind. Gerade, wenn wir kein theoretisches Verständnis haben, ist die Datenanalyse das einzige Mittel zur Entscheidungsunterstützung. Ein aus einer Datenanalyse resultierendes Modell, das eine hohe Vorhersagegenauigkeit hat, reicht aus, auch wenn es nicht direkt verständlich ist.

In der Wissenschaft hat eine Vorhersage, die nicht auf erkennbaren Kausalzusammenhängen basiert, einen besonders schlechten Ruf. Im täglichen Leben gehen wir jedoch fast immer so vor. Praktisch all unsere Entscheidungen, zum Beispiel die darüber, wie wir mit unserem menschlichen Gegenüber kommunizieren, alle Hypothesen, die wir über die Verfassung und Einstellung des Dialogpartners aufstellen, ergeben sich daten- oder hier besser erfahrungsgetrieben aus der Beobachtung von Körperhaltung, Mimik, Tonfall oder der Kenntnis von relevanten Fakten über unseren Dialogpartner. Wir tun dies alles sogar noch größtenteils unbewusst. Das ist nur ein kleines Beispiel für das datengetriebene Vorgehen, das wir ganz überwiegend einsetzen, um uns im täglichen Leben zurechtzufinden. Dies gilt im Übrigen auch für alle anderen Spezies. So wird deutlich, in welch großem Umfang sich die Natur der Datenanalyse bedient. In der Tat ist das Konzept der Kausalität erst durch die kognitiven Fähigkeiten des rational denkenden Menschen in die Welt gekommen.

Aber zurück zur Medizin: Wir haben auch hier in den meisten Fällen, jedenfalls bis jetzt, keine andere Wahl, als zu datengetriebenen Methoden zu greifen. Das wird deutlich, wenn man sich das Prinzip der klinischen Studien im Zusammenhang mit der Zulassung neuer Medikamente oder Therapien vor Augen führt – eine rein datengetriebene Prozedur. Dabei wird zum Beispiel durch den sorgfältigen Entwurf einer randomisierten kontrollierten Studie (Vineis 2003) oder mittels des Instruments der Mendelschen Randomisierung (Davey Smith und Hemani 2014) sorgfältig darauf geachtet, dass Störfaktoren (siehe auch Abschnitt 5) das Studienergebnis

so wenig wie möglich beeinflussen. Danach wird die Assoziation des Therapieerfolgs mit der Gruppenzugehörigkeit des Patienten in der Studie kausal interpretiert (Kottas et al. 2011). Dies ist jedoch nach unserer strengen Auslegung kein wirkliches Verständnis des kausalen Zusammenhangs, in demselben Sinne, wie wir auch Keplers Gesetze nicht als erklärend verstehen, obwohl sie sehr genau die Planetenbahnen beschreiben. Vielmehr steht der statistische Nachweis der Überlegenheit einer Therapie gegenüber alternativen Therapien, das heißt die Vorhersagegenauigkeit, nach wie vor im Vordergrund. Das wird in der Regel als ein ausreichender Nachweis für Kausalität angesehen, auch wenn die molekularen Prozesse nicht im Einzelnen aufgeklärt sind.

6 *Methoden bei statistischen Analysen: Von Mathematik und Ethik*

In diesem Abschnitt möchte ich sehr kurz auf den Charakter und die Form der mathematischen Methoden zur Datenanalyse eingehen und zwei der wichtigsten Probleme diskutieren, denen man sich bei der Datenanalyse stellen muss. Einleitend sei erwähnt, dass Leibniz einer der frühen Proponenten von Inferenz von Wissen aus Daten war (Gigerenzer et al. 1990; Wagner 2008).

In der journalistischen Presse wird gelegentlich fälschlicherweise den Algorithmen, die zur Datenanalyse verwendet werden, eine ethische Qualität zugesprochen. Dies beruht meiner Ansicht nach auf einem Missverständnis. Algorithmen, also mathematische Verfahren zur Berechnung eines Ergebnisses, können nicht gut oder schlecht sein, genauso wie die Zahl 5 oder der Satz von Pythagoras nicht gut oder schlecht sein können. Das mathematische Verfahren an sich verwendet immer eine gegebene Eingabe, in unseren beiden Beispielen die Proteinsequenz bzw. die Sequenz des viralen Genoms, in einer angemessenen mathematischen Kodierung. Diese Kodierung erfolgt nicht automatisch, sondern wird von dem Entwickler der Datenanalysemethode entworfen. Sie transzendiert also den Algorithmus. Es ist durchaus so, dass die Kodierung entscheidenden Einfluss auf das Resultat hat. In der Regel wird der Entwickler jedoch die Kodierung so wählen, dass das Analyseverfahren eine größtmögliche Vorhersagegenauigkeit aufweist. Also auch hier ist zunächst ein ethischer Wert noch nicht vorrangig auszumachen. Auf der Grundlage der entsprechend kodierten Eingabe versucht jeder Algorithmus, eine Ausgabe zu erzeugen, die in einem präzise definierten Sinne optimal ist. Bei der Proteinstrukturvorhersage wollen wir ein Proteinmodell haben, das der tatsächlichen Proteinstruktur am nächsten kommt. Bei der HIV-Therapie suchen wir nach einer Schätzung der Resistenz des Virus gegen den Wirkstoff, die der tatsächlichen Resistenz am nächsten kommt. Dieses Qualitätskriterium, das sich in einer Kostenfunktion manifestiert, die das algorithmische Verfahren zu optimieren sucht, ist ein ganz wesentlicher Bestandteil der Entwicklung einer Datenanalysemethode. Beispiele für Kostenfunktionen sind: (i) die gegenwärtige Wirksamkeit einer Therapie, (ii) die Länge der Wirksamkeit einer Therapie, (iii) die Anzahl und geschätzte Wirksamkeit möglicher Anschlusstherapien sobald die gegebene Therapie versagt, (iv) die Kosten einer Therapie – oder Kombinationen von diesen Kriterien. Es ist offensichtlich, dass an dieser Stelle auch ethische Belange Einzug halten. Wir haben hier jedoch die Domäne der Mathematik verlassen und befassen uns mit der Spezifikation unserer Methode.

Mathematisch geht der allergrößte Teil der Datenanalysemethoden wie folgt vor. Die Kodierung der Eingabe ergibt gewöhnlich für jeden Datenpunkt einen Punkt in einem in der Regel

hochdimensionalen euklidischen Raum. Zur Illustration verwenden wir hier einen Beispieldatensatz,[4] der Auskunft über den Benzinverbrauch von Kraftfahrzeugen in Abhängigkeit von deren Baujahr und Gewicht gibt. Da die Daten aus den USA stammen, ist der Benzinverbrauch in Meilen pro Gallone (mpg) und das Gewicht in englischen Pfund (lbs) angegeben. Wir haben also zwei Eingaben (Gewicht und Jahr) – unser Datenraum ist damit eine Ebene – und eine Ausgabe (mpg). In vielen Fällen kann die Anzahl der Dimensionen des Datenraums durchaus in die Hunderttausende gehen.

Wir konzentrieren uns hier auf das sogenannte überwachte Lernen, die vornehmlich für die Berechnung von Vorhersagen verwandte Methode. Hier gibt es zwei Formen der Datenanalyse (Abb. 4).

(i) Klassifikation: Hier gehört jeder Datenpunkt einer Klasse an. Die Anzahl der Klassen ist endlich und in der Regel klein. Im Beispiel von Abb. 4a und b haben wir zwei Klassen gewählt. Die Autos werden in die Klassen niedriger Verbrauch (mpg $\geq$ 20, blau) und hoher Verbrauch (mpg $<$ 20, orange) unterteilt. Das Ziel der Datenanalyse ist, den Datenraum – hier also die Ebene – in Bereiche einzuteilen, die den Klassen zugeordnet werden. Die Klassifikation eines neuen Punktes geschieht dann durch die Zuordnung der Farbe des Bereiches, in dem der Punkt liegt. Die Abbildung zeigt zwei solche Modelle bei einer gegebenen Punktemenge, ein lineares (Abb. 4a, geradlinige Klassengrenze) und ein nichtlineares (Abb. 4b, gekrümmte Klassengrenze). Wie man sieht, machen beide Modelle Fehler, gekennzeichnet durch Punkte, deren Farbe sich von der ihres Hintergrundes unterscheidet. Dabei macht das nichtlineare Modell auf dem gegebenen Datensatz weniger Fehler (2 aus 392) als das lineare (30 aus 392).

(ii) Regression: Hier ordnen wir die Datenpunkte keinen Klassen zu, sondern weisen ihnen eine Zahl als Markierung zu. Bei unserem Beispieldatensatz schätzen wir den Benzinverbrauch in Meilen pro Gallone. Wieder haben wir es mit zwei Eingabevariablen zu tun (Gewicht, Jahr), die die Ebene aufspannen, über die sich nun eine senkrechte Achse erhebt, die die Markierung „Verbrauch" trägt. Die geschätzten Verbrauchswerte sind in Abb. 4c und 4d durch Oberflächen dargestellt. Die tatsächlichen Datenpunkte umgeben diese Fläche als Punktwolke. Wieder zeigen wir ein lineares (Abb. 4c) und ein nichtlineares (Abb. 4d) Modell. Die senkrechten Linien, die von den Datenpunkten auf die Oberflächen gezogen sind, repräsentieren die Fehler, die das Modell bei der Schätzung der einzelnen Verbrauchswerte macht. Auch hier ist der Fehler des linearen Modells größer als der des nichtlinearen Modells.

Die Kunst der problemgerechten Datenanalyse besteht in der geeigneten Kodierung der Eingabe, der richtigen Auswahl der Fehlerfunktion und der geeigneten Auswahl des Modells, etwa der richtigen Entscheidung zwischen einem linearen (einfachen) und einem nichtlinearen (komplexen) Modell. Dabei haben einfache Modelle den Nachteil, dass sie die gegebenen Daten unter Umständen nicht mit hoher Genauigkeit wiedergeben. Wenn ein Modell zu komplex ist, dann besteht dagegen die Gefahr, dass es zwar die gegebenen Daten sehr genau wiedergibt, aber nicht hinreichend auf zukünftige Daten verallgemeinert. Ein solches Modell nennt man übertrainiert. Abb. 4b zeigt ein Beispiel eines übertrainierten Modells für den Beispieldatensatz. Dass dieses Modell übertrainiert ist, ist leicht an den komplexen Klassengrenzen zu erkennen, die ganz

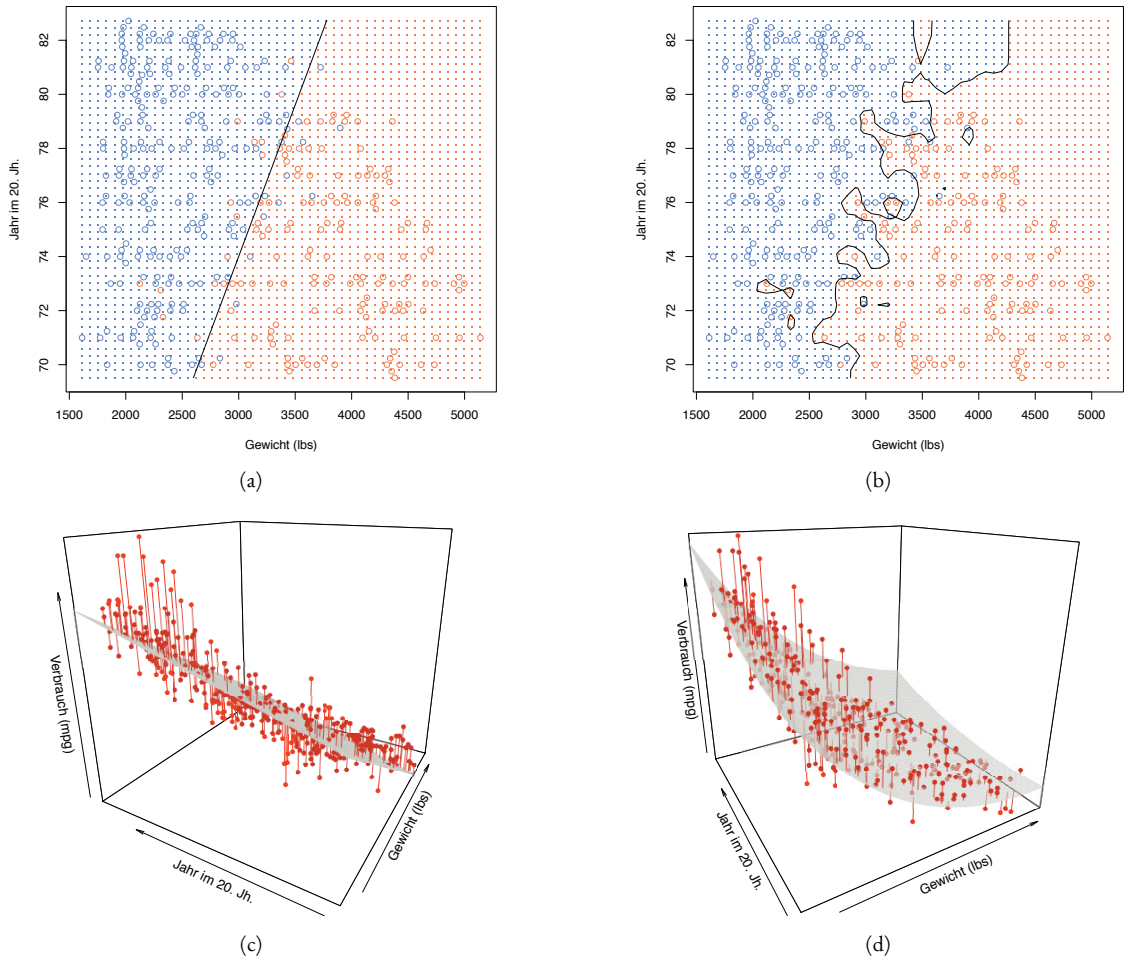


Abbildung 4. Statistische Verfahren zur Datenanalyse. (a) Klassifikation, lineares Modell. (b) Klassifikation, nichtlineares Modell. Die Datenpunkte sind blau (niedriger Verbrauch) bzw. orange (hoher Verbrauch) eingefärbt. Die schwarzen Linien zeigen die mit je einem linearen bzw. nichtlinearen Verfahren abgeleiteten Klassengrenzen. (c) Regression, lineares Modell. (d) Regression, nichtlineares Modell. Die Datenpunkte sind rot gefärbt. Die grauen Oberflächen repräsentieren die geschätzten Verbrauchswerte. Die senkrechten roten Linien geben die Abweichungen der tatsächlichen von den geschätzten Werten an.

offensichtlich auf die spezifisch gegebene Datenmenge angepasst sind und nicht die wirkliche Beziehung zwischen hohen/niedrigem Verbrauch in Abhängigkeit von Baujahr und Gewicht wiedergeben. Das Modell macht deshalb nur wenige Fehler auf den gegebenen Daten. Es ist aber nicht zu erwarten, dass das Modell auf zukünftigen Daten genaue Aussagen macht. Allerdings ist die Beziehung zwischen dem Benzinverbrauch und den Eingaben Baujahr und Gewicht offensichtlich nichtlinear: Wie sich nachrechnen lässt, gibt das nichtlineare Regressionsmodell in Abb. 4d sie besser (mit geringerem Fehler und höherer Vorhersagekraft) wieder als das lineare Modell in Abb. 4c. Die richtige Modellkomplexität zu wählen ist ein zentrales Problem bei der Datenanalyse. (Eine ausführliche Einführung in die methodischen Grundlagen der statistischen Datenanalyse geben James et al. (James et al. 2014).

7 *Fallen bei statistischen Datenanalysen: Nicht alles ist so, wie es scheint*

In diesem Abschnitt wollen wir unsere Diskussion der Grenzen von Datenanalyse noch etwas vertiefen. In dem Titel dieses Artikels gibt es noch zwei Kernbegriffe, nämlich „Suggestivität" und „Risiken". Sie betreffen zwei Fallen, die sich bei Datenanalysen auftun können und die es unter allen Umständen zu umschiffen gilt.

Beginnen wir mit dem Risiko: Betrachten wir ein aktuelles Beispiel aus der modernen genombasierten Medizin. Hier versucht man, Zusammenhänge zwischen Krankheitsbildern und Genomvarianten des Patienten aufzudecken. Kurz gesagt, man sucht nach Krankheitsgenen. Der Begriff „Krankheitsgen" ist recht irreführend, denn in der Regel hat jeder Mensch das betreffende Gen. Dabei können sich die Varianten des Gens zwischen verschiedenen Menschen unterscheiden. Nur manche Varianten sind mit Krankheiten aussoziiert. Wonach man sucht, sind Varianten von Genen, die bei bestimmten Krankheitsbildern gehäuft auftreten. Dieser zunächst rein statistische Zusammenhang legt einen kausalen Zusammenhang zwischen der betreffenden Genvariante und dem Krankheitsbild nahe, beweist ihn aber nicht. Solche statistischen Zusammenhänge werden mit so genannten genomweiten Assoziationsstudien (GWAS) untersucht. Hier bedient man sich einer großen Kohorte von in der Regel tausenden oder zigtausenden Probanden, gesund und krank. Man liest deren Genome bzw. diejenigen Teile des Genoms ab, die man für krankheitsrelevant hält. Und dann untersucht man mit geeigneten statistischen Verfahren, ob an gewissen Orten im Genom (Genorten, Gen loci) genomische Varianten unter den Patienten mit einem bestimmten Krankheitsbild gehäuft auftreten. Mit GWAS hat man schon vielfache Hinweise auf die Assoziationen von Genorten zu Krankheitsbildern gefunden (Burton 2007; siehe auch den GWAS Catalog unter www.ebi.ac.uk/gwas/). Als Studienresultate ergeben sich Diagramme wie in Abb. 5 dargestellt. Hier sind auf der waagerechten Achse die 22 Chromosomen des menschlichen Genoms aufgetragen (das Geschlechtschromosom wird nicht betrachtet). Auf

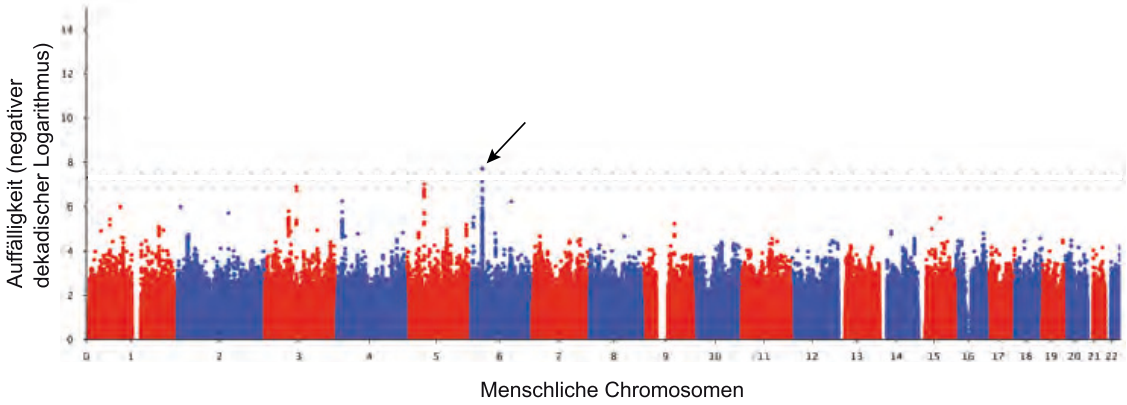


Abbildung 5. Ergebnisse einer genomweiten Assoziationsstudie zur Aufdeckung von mit der Intensität der HIV-Infektion assoziierten Genomorten. Nur der mit dem Pfeil gekennzeichnete Genort ist nach Bereinigung der Resultate unter Berücksichtigung des Multiplen-Testphänomens signifikant mit dem Krankheitsbild assoziiert. Für diesen Genort, an dem ein Protein unseres Immunsystems kodiert ist, ist auch die biologische Grundlage dieser Assoziation bekannt. (Abbildung adaptiert aus (Bartha 2013) unter Creative Commons License CC BY 4.0.)

der senkrechten Achse ist ein Maß für die Auffälligkeit des Genortes abgetragen. Auffälligkeit bedeutet hier das gehäufte Auftreten einer Variante des Gens bei Menschen mit der betrachteten Krankheit (Sham und Purcell 2014). Größere Zahlen bedeuten höhere Auffälligkeit. Man sieht, dass sich aus der allgemeinen Wolke von Punkten, die Genorten entsprechen, ein paar Dutzend Genorte nach oben ablösen.

Die hier konkret genannte Studie betrachtet die Virenkonzentration im Blut bei einer HIV-Infektion (Bartha et al. 2013). Die Betrachtung von Abb. 5 würde nahe legen, dass etwa zwei Dutzend Genorte eine mögliche Assoziation mit dem Krankheitsbild haben, nämlich diejenigen, die sich aus dem Ozean der Punkte in der Abbildung nach oben herausheben. Dem ist jedoch nicht so. Nach einer stringenteren Signifikanzanalyse weist nur ein Genort, nämlich der mit dem Pfeil gekennzeichnete, tatsächlich eine statistisch signifikante Assoziation mit dem Krankheitsbild auf. Sind wir nicht stringent genug, dann unterstellen wir vielen Genen eine Assoziation mit der Krankheit, die diese tatsächlich gar nicht aufweisen. Wie kommt das? Der Grund hierfür liegt darin, dass wir sehr viele Genorte untersuchen. Dass ein einzelner Genort rein zufällig, das heißt, ohne biologische Basis auffällig ist, ist möglich, aber sehr unwahrscheinlich – genauso, wie ein Sechser im Lotto möglich, aber sehr unwahrscheinlich ist. Aber wenn man zigtausend Genorte untersucht, können sich Auffälligkeiten auch rein zufällig ergeben. Wenn viele Leute Lotto spielen, wird es mit hoher Wahrscheinlichkeit einen Gewinner geben. Alle in Abb. 5 auffällig erscheinenden Genorte, außer dem mit dem Pfeil gekennzeichneten, sind solche Lottogewinner – sie sind auffällig ohne erkennbare biologische Grundlage. Einen solchen falschen Hinweis auf eine Auffälligkeit nennt man ein Falsch-Positiv. Das Risiko von Falsch-Positiven ist besonders dann groß, wenn die Analyse viele gleichartige Tests umfasst (Multiples Testen [Sham und Purcell 2014]). Falsch-Positive sind ein großes Ärgernis bei Datenanalysen. Sie treten auch im täglichen Leben auf, sei es bei einem medizinischen Test, der einem ein Krankheitsrisiko bescheinigt hat, das sich bei einer Nachuntersuchung nicht bestätigt, oder bei einem Verdacht der Beteiligung bei einem Vergehen, der sich im Nachhinein nicht bestätigt.

In heutiger Zeit, in der sehr viele Hypothesen auf der Basis von Big Data-Analysen aufgestellt werden, werden Falsch-Positive von einem Risiko zu einer wirklichen Gefahr. Wird die Datenanalyse unsachgemäß vorgenommen oder werden Schlussfolgerungen aus ihr vorschnell gezogen, so kann dies zur Bildung von Vorurteilen, Ausgrenzung und Stigmatisierung führen.

Die zweite Falle bei statistischen Analysen fassen wir hier unter dem Kernbegriff „Suggestivität" zusammen. Damit meinen wir, dass die Resultate von Datenanalysen Kausalzusammenhänge dort suggerieren können, wo sie nicht nachgewiesen und häufig auch nicht einmal vorhanden sind.

Betrachten wir folgende Beispiele:

(i) Eine genomweite Assoziationsstudie deckt die Auffälligkeit eines Gens bezüglich eines Krankheitsbildes auf. Ist damit erwiesen, dass das Gen ursächlich für die Krankheit ist?
(ii) Eine statistische Studie zeigt, dass bei Leuten, die in der Nähe eines Kernkraftwerks wohnen, ein bestimmter Krebs gehäuft auftritt. Ist damit erwiesen, dass das Kernkraftwerk Strahlung abgibt, die die Ursache für das gehäufte Auftreten des Krebses ist?
(iii) Eine Studie zeigt, dass Raucher eine verringerte Häufigkeit von Parkinson haben. Wie verhält es sich hier mit den Kausalbeziehungen?

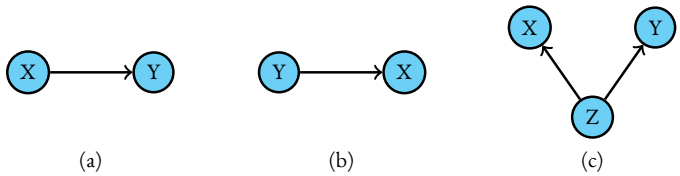


Abbildung 6. Drei Möglichkeiten für kausale Abhängigkeiten zwischen assoziierten Variablen

Nennen wir die zwei in Assoziation stehenden Größen X und Y. Generell legen Assoziationen zwischen X und Y einen kausalen Zusammenhang zwischen den beiden Größen nahe. Was aber hier Ursache und was Wirkung ist, bleibt unklar (Aalen und Frigessi 2007). Abb. 6 zeigt drei Möglichkeiten des kausalen Zusammenhangs zwischen X und Y. Im komplexesten Fall (c) geht die Assoziation zwischen X und Y auf eine dritte Größe Z zurück, die zu den beiden Größen X und Y im ursächlichen Zusammenhang steht. In vielen Fällen gibt es so genannte Störgrößen, Größen die im Rahmen der statistischen Studie nicht gemessen werden, aber mehrere gemessene Größen beeinflussen. Im Falle (ii) hat sich bei einer tatsächlichen Studie herausgestellt, dass die radioaktive Strahlung, die aus dem Kernkraftwerk entweicht, so gering ist, dass sie keine Ursache für gehäufte Krebsfälle sein kann. Es gibt jedoch andere Größen, etwa solche betreffend die demographische Zusammensetzung der Bevölkerung in der Nähe eines Kernkraftwerks, die auf die Krebsrate sehr wohl einen Einfluss haben. Eklatante Beispiele wie die Tatsache, dass die Anzahl von Storchenpaaren in einer Region mit der dortigen Geburtenrate assoziiert ist (Matthews 2000), zeigen, mit welcher Vorsicht man an Interpretationen der Resultate von Datenanalysen herangehen muss.

In jüngster Zeit hat gerade das Aufkommen von Big Data in der Statistik eine dynamische Entwicklung eingeleitet, unter anderem mit dem Ziel der Ableitung von kausalen Zusammenhängen aus statistischen Assoziationen (Pearl 2009). Wir stehen hier noch ganz am Anfang. Aber erste Resultate gibt es bereits, wenn sie auch bisher nur in eingegrenzten Szenarien anwendbar sind. Bernhard Schölkopf vom Max-Planck-Institut für Intelligente Systeme in Tübingen hat für seine Arbeiten auch auf diesem Gebiet (Mooij 2015; Janzing et al. 2014) im Jahr 2012 den Akademiepreis der Berlin-Brandenburgischen Akademie der Wissenschaften erhalten.

8 *Epilog: Von Big Data zur Theorie*

Was ist also unser Fazit? Ist, wie viele Leute meinen, Big Data der Motor der Innovation in unserem heutigen Zeitalter? Nach dieser Ansicht beherbergen die Daten und insbesondere, die in ihnen enthaltenen Muster, alle Weisheit des Universums. Es ist völlig ausreichend, dass die Daten die Ableitung von Modellen mit hoher Vorhersagegenauigkeit erlauben. Es ist nicht notwendig, dass die Modelle im Sinne von kausalen Zusammenhängen auch verstanden werden. Kausalität ist sowieso etwas, was nur durch die kognitiven Fähigkeiten des Menschen in das Universum Eingang gefunden hat. Bis es uns gab, kam die Natur auch ganz gut ohne diesen Begriff aus und hat es mit lediglich den Anpassungs- und Lernmechanismen, die auch der

Datenanalyse unterliegen, sehr weit gebracht. Oder ist Big Data der Beginn einer Veränderung der Gesellschaft, die zunehmend mehr auf Assoziationen vertraut als auf das aus sorgfältiger Ableitung von kausalen Zusammenhängen resultierende Verständnis – mit allen erwähnten Begleiterscheinungen wie Verdachten und Vorurteilen, die aus Datenanalysefehlern resultieren, und mit der Illusion von Kausalzusammenhängen, die nicht wirklich überprüft wurden?

Ich meine, dass Datenanalyse im Zeitalter von Big Data ein mächtiges Instrument ist, das mit großer Sorgfalt verwendet werden muss. Man braucht sicher eine Kombination von technischen Lösungen und gesellschaftlichen Regeln, um dieses Instrument kontrolliert und verantwortungsbewusst eizusetzen.

Generell bin ich gegen Schwarzweißmalerei, was Big Data betrifft. Aus wissenschaftlicher Sicht schließen sich Datenanalyse und Theoriebildung nicht aus. Vielmehr ist die Datenanalyse als ein Vorfilter für Untersuchungen zu verstehen, die Kausalzusammenhänge aufdecken. Früher diente allein der Geist des Forschers als Instrument, eine plausible Hypothese aufzustellen, die dann systematisch durch Experimente validiert oder falsifiziert werden konnte. Heute können wir das Instrument der Datenanalyse aus Big Data einsetzen, um aus einer zunächst unüberschaubaren Hypothesenvielfalt systematisch eine begrenzte Menge von vielversprechenden Hypothesen auszuwählen, die dann mit Methoden, die sich in der Wissenschaft seit Jahrhunderten bewährt haben, überprüft werden können. In einem solchen Szenario wird die Datenanalyse, wo immer möglich, durch eine auf die Aufdeckung von Kausalzusammenhängen und auf Theoriebildung abzielende Nachuntersuchung ergänzt. Dies findet in der Forschung vor allem in den Naturwissenschaften umfänglich statt. Den Vorschlägen, die sich aus der Datenanalyse ergeben, wird hierbei die notwendige kritische Vorsicht entgegengebracht. Als Instrument zur Eingrenzung von Hypothesenvielfalt ist die Datenanalyse in einem solchen Szenario von enormem Nutzen und oft in der Tat unverzichtbar.

Danksagung

Ich danke Markus Löffler und Jürgen Renn für ausführliche und sehr hilfreiche Hinweise zu diesem Artikel und Christian Lengauer und Marcel Schulz für eine kritische Durchsicht des Manuskripts. Die Fallbeispiele in Abb. 3, 4, bzw. 5 wurden von Olga Kalinina, Anna Feldmann, bzw. Nico Pfeifer beigesteuert, die auch anderweitige wertvolle Hinweise gegeben haben. Auch ihnen gilt mein aufrichtiger Dank.

Anmerkungen

1. Einen hervorragenden Einblick in die Geschichte der Quantenmechanik gibt Kumar (2009). Siehe auch Golstein und Hon (2005), Kuhn (1978), Kragh (1999).
2. Jörg Resag (2013) gibt eine allgemeinverständliche Einführung in die diesbezüglichen Hintergründe.
3. Häufig wird Big Data durch die fünf *V* charakterisiert: Volume (große Datenmenge, entsprechend unserem Kriterium (i)), Velocity (Schnelligkeit der Datengenerierung, entsprechend unserem Kriterium (ii), Variety (Heterogenität von Daten, d. h. unterschiedlicher Ursprung; dieses Kriterium steht in unserer Betrachtung nicht im Vordergrund), Veracity (Zuverlässigkeit von Daten; siehe Abschnitt 5), Value (ein ökonomisches Kriterium, das wir hier nicht betrachten).
4. Der Datensatz ist unter www-bcf.usc.edu/~gareth/ISL/data.html zugänglich

Abbildungsnachweise

Abb. 1: Protein Data Bank (www.pdb.org)
Abb. 2a, b: Max-Planck-Institut für Informatik
Abb. 3: Max-Planck-Institut für Informatik
Abb. 4a–d: Max-Planck-Institut für Informatik
Abb. 5: Entnommen aus und bearbeitet: Bartha I et al. (2013) A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. Elife 2:e01123/CC BY 4.0: https://creativecommons.org/licenses/by/4.0/

Literatur

Aad, Georges et al. (2012). "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716 (1), S. 1–29.

Aalen, Odd O. und Frigessi, Arnoldo (2007). "What can statistics contribute to a causal understanding?", In: *Scandinavian Journal of Statistics* 34 (1), S. 155–168.

Adam-Bourdarios, Claire et al. (2015). "The Higgs Machine Learning Challenge". In: *Journal of Physics: Conference Series* 664 (7), 072015.

Bartha, Istvan et al. (2013). "A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control". In: *Elife* 2, e01123.

Beerenwinkel, Niko et al. (2002). "Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype". In: *Proceedings of the National Academy of Sciences of the United States of America* 99 (12), S. 8271–8276.

Benkert, Pascal, Tosatto, Silvio C. und Schomburg, Dietmar (2008). "QMEAN: 'A comprehensive scoring function for model quality assessment' ". In: *Proteins* 71 (1), S. 261–277.

Biasini, Marco et al. (2014). "SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information". In: *Nucleic Acids Research* 42 (Web Server issue): W252–258.

Born, Max, Heisenberg, Werner und Jordan, Pascual (1926). „Zur Quantenmechanik. II". In: *Zeitschrift für Physik A Hadrons and Nuclei* 35 (8), S. 557–615.

Born, Max und Jordan, Pascual (1925). „Zur Quantenmechanik". In: *Zeitschrift für Physik A Hadrons and Nuclei* 34 (1), S. 858–888.

Burton, Paul R. et al. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447 (7145), S. 661–678.

Chatrchyan, Serguei et al. (2012). "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716 (1), S. 30–61.

Bernstein, BE et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489 (7414), S. 57–74.

Davey Smith, George und Hemani, Gibran (2014). "Mendelian randomization: genetic anchors for causal inference in epidemiological studies". In: *Human Molecular Genetics* 23 (R1), R89–98.

Dill, Ken A. und MacCallum, Justin L. (2012). "The protein-folding problem, 50 years on". In: *Science* 338 (6110), S. 1042–1046.

Dunbrack, Jr. Roland (2007). "Homology modeling in biology and medicine". In: *Bioinformatics – From Genomes to Therapies*. Hrsg. von Thomas Lengauer. Weinheim: WILEY-VCH Verlag, S. 297–350.

Einstein, Albert (1916). „Die Grundlage der allgemeinen Relativitätstheorie". In: *Annalen der Physik* 49 (7), S. 769–822.

— (2009). *Über die spezielle und die allgemeine Relativitätstheorie*. New York: Springer.

Englert, François & Brout Robert (1964). "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Physical Review Letters* 13 (9), S. 321–323.

Fiser, Andràs, Feig, Michael, Brooks, Charles L., 3rd und Sali, Andrej (2002). "Evolution and physics in comparative protein structure modeling". In: *Acc Chem Res* 35 (6), S. 413–421.

Ghiringhelli, Luca M., Vybiral, Jan, Levchenko, Sergey V., Draxl, Claudia und Scheffler, Matthias (2015). "Big Data of Materials Science: Critical Role of the Descriptor". In: *Physical Review Letters* 114 (10), S. 105503.

Gigerenzer, Gerd et al. (1990). *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge, Großbritannien: Cambridge University Press.

Goldstein, Bernhard R. und Hon, Giora (2005). "Kepler's Move from Orbs to Orbits: Documenting a Revolutionary Scientific Concept". In: *Perspectives on Science* 13 (1), S. 74–111.

Grasshoff, Gerd (2012). *Globalization of Ancient Knowledge: From Babylonian Observations to Scientific Regularities. The Globalization of Knowledge in History*. Berlin: Edition Open Access.

Guo, Hua-Dong, Zhang, Li und Zhu, Lan-Wei (2015). "Earth observation big data for climate change research". In: *Advances in Climate Change Research* 6 (2), S. 108–117.

Guralnik, Gerald S., Hagen Carl R. und Kibble, Thomas W. B. (1964). "Global Conservation Laws and Massless Particles". In: *Physical Review Letters* 13 (20), S. 585–587.

Heisenberg, Werner (1925). „Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen". In: *Zeitschrift für Physik A Hadrons and Nuclei* 33 (1), S. 879–893.

Hesse, Bradford W., Moser, Richard P. und Riley, William T. (2015). "From big data to knowledge in the social sciences". In: *The ANNALS of the American Academy of Political and Social Science* 659 (1), S. 16–32.

Higgs, Peter W. (1964). "Broken Symmetries and the Masses of Gauge Bosons". In: *Physical Review Letters* 13 (16), S. 508–509.

Imboden, Dieter und Koch, Sabine (2008). *Systemanalyse: Einführung in die mathematische Modellierung natürlicher Systeme*, 2. Auflage. Heidelberg: Springer-Verlag.

Ivezić, Željiko et al. (2012). "Galactic Stellar Populations in the Era of the Sloan Digital Sky Survey and Other Large Surveys". In: *Annual Review of Astronomy and Astrophysics* 50 (1), S. 251–304.

James, Gareth, Witten, Daniela, Hastie, Trevor und Tibshirani, Robert (2014). *An Introduction to Statistical Learning with Applications in R*. New York: Springer Verlag.

Janzing, Dominik, Steudel, Bastian, Shajarisales, Naji und Schölkopf, Bernhard (2014). "Justifying information-geometric causal inference". USA, Cornell: arXiv.

Kamisetty, Hetunandan, Ovchinnikov, Sergey und Baker, David (2013). "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era". In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (39), S. 15674–15679.

Kepler, Johann (1609). *Astronomia Nova seu physica coelestis, tradita commentariis de motibus stellae martis*. Heidelberg: Voegelin.

— (1619). *Harmonices Mundi Libri V*. Francofurti: Tampachius.

— (1627). *Tabulae Rudolphinae, quibus astronomicae scientiae, temporum longinquitate collapsae restauratio continentur*. Ulm: Jonas Saur.

Kihara, Daisuke, Chen, Hao und Yang, Yifeng D. (2009). "Quality assessment of protein structure models". In: *Curr Protein Pept Sci* 10 (3), S. 216–228.

Kottas, M., Marchlewski, M., Polte, C., Strüver, V. und Witte, K. (2011). „Konzeptionelle Überlegungen zum Nachweis einer Ursache-Wirkungs-Beziehung in klinischen Studien". In: *Deutsche Zeitschrift für Klinische Forschung* 11 (12), S.77–81.

Kragh, Helge (1999). *Quantum Generations: A History of Physics in the Twentieth Century*. Princeton: Princeton University Press.

Kuhn, Thomas S. (1978). *Black-Body Theory and the Quantum Discontinuity, 1894–1912*. New York: Oxford University Press.

Kumar, Manjit (2009). *Quanten: Einstein, Bohr und die große Debatte über das Wesen der Wirklichkeit*. Berlin: Berlin Verlag.

Lengauer, Thomas, Albrecht, Mario und Domingues, Francisco S. (2012). "Computational Biology". In: *Systems Biology. Advances in Molecular Biology and Medicine*, ed. Meyers RA Heidelberg: Wiley-VCH, S. 277–348.

Lengauer, Thomas und Kaiser, Rolf (2009). „Computerjagd auf das AIDSvirus". In: *Spektrum der Wissenschaft* (August), S. 62–67.

Lengauer, Thomas und Sing, Tobias (2006). "Bioinformatics-assisted anti-HIV therapy". In: *Nat Rev Microbiol* 4 (10), S. 790–797.

Levin, Jonathan D. und Einav, Lirian (2014). "The data revolution and economic analysis". In: *NBER Innovation Ploicy and the Economy* 14, S. 1–24.

Lusher, Scott J., McGuire, Ross, van Schaik, René C., Nicholson, C. David und de Vlieg, Jacob (2014). "Data-driven medicinal chemistry in the era of big data". In: *Drug Discov Today* 19 (7), S. 859–868.

Marks, Debora S. et al. (2011). "Protein 3D Structure computed from evolutionary sequence variation". In: *PLoS ONE* 6 (12), e28766.

Matthews, Robert (2000). "Storks deliver babies ($p = 0.008$)". In: *Teaching Statistics* 22 (2), S. 36–38.

Mooij, Jorris M., Peters, Jonas, Janzing, Dominik, Zscheischler, Jakob und Schölkopf, Bernhard (2015). "Distinguishing Cause and effect using observational data: Methods and benchmarks." USA, Cornell: arXiv.

Newton, Isaac (1687). *Philosophiæ naturalis principia mathematica.* Londini: Jussu Societatis Regiæ ac Typis Josephi Streater.

Pearl, Judea (2009). *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge: Cambridge University Press.

Pennisi, Elizabeth (2010). "Genomics. 1000 Genomes Project gives new map of genetic diversity". In: *Science* 330 (6004), S. 574–575.

Planck, Max (1900). *Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum. Verhandungen. der deutschen Physikalischen Gesellschaft* 2 (17), S. 237–245.

Rajan, Krishna (2015). "Materials informatics: The materials 'gene' and big data". In: *Annual Review of Materials Research* 45 (1), S. 153–169.

Resag, Jörg (2013). *Die Entdeckung des Unteilbaren: Qaunten, Quarks und die Entdeckung des Higgs-Teilchens.* Berlin/Heidelberg: Springer Spektrum Verlag.

Rost, Burkhard (1999). "Twilight zone of protein sequence alignments". In: *Protein Eng* 12 (2), S. 85–94.

Sander, Chris und Schneider, Reinhard (1991). "Database of homology-derived protein structures and the structural meaning of sequence alignment". In: *Proteins* 9 (1), S. 56–68.

Schrödinger, Erwin (192a). „Quantisierung als Eigenwertproblem 2". In: *Annalen der Physik* 384 (6), S. 489–527.

— (1926b). „Quantisierung als Eigenwertproblem 3". In: *Annalen der Physik* 385 (13), S. 437–490.

— (1926c). „Quantisierung als Eigenwertproblem 4". In: *Annalen der Physik* 386 (18), S. 109–139.

— (1926d). „Quantisierung als Eigenwertproblem 1". In: *Annalen der Physik* 384 (4), S. 361–376.

Schwede, Torsten (2013). "Protein modeling: What happened to the 'protein structure gap'?" In: *Structure* 21 (9), S. 1531–1540.

Sellars, Scott et al. (2013). "Computational Earth Science: Big Data Transformed Into Insight". In: *Eos, Transactions American Geophysical Union* 94 (32), 277–278.

Sham, Pak C. und Purcell, Shaun M. (2014). "Statistical power and significance testing in large-scale genetic studies". In: *Nat Rev Genet* 15 (5), S. 335–346.

Stephenson, Bruce (1987). *Kepler's physical astronomy.* New York: Springer.

Thielen, Alexander und Lengauer, Thomas (2012). "Geno2pheno[454]: A web server for the prediction of HIV-1 coreceptor usage from next-generation sequencing data". In: *Intervirology* 55 (2), S. 113–117.

Vineis, Paolo (2003). "The randomized controlled trial in studies using biomarkers". In: *Biomarkers* 8 (1), S. 13–32.

Vitolo, Claudia, Elkhatib, Yehia, Reusser, Dominik, Macleod, Christopher J. A. und Buytaert, Wouter (2015). "Web technologies for environmental Big Data". In: *Environmental Modelling & Software* 63, S.185–198.

Wagner, Gert G. (2008). „Leibniz und die (Amtliche) Statistik". In: *Working Papers des Rats für Sozial- und Wirtschaftsdaten* 39, S. 1–7.

Wolynes, Peter G. (2015). "Evolution, energy landscapes and the paradoxes of protein folding". In: *Biochimie* 119, S. 218–230.

Zazzi, Maurizio et al. (2012). "Predicting response to antiretroviral treatment by machine learning: The EuResist project". In: *Intervirology* 55 (2), S. 123–127.