arbeitstitel
forum für leipziger promovierende

Arbeitstitel – Forum für Leipziger Promovierende // Gegründet 2009
Herausgegeben von Stephanie Garling, Susanne Bunzel, Franziska Naether,
Christian Fröhlich, Felix Frey
Meine Verlag, Magdeburg

Wie man zwischen den Zahlen liest. Data-Mining und computergestützte Vorhersagen am Beispiel Bioinformatik

Thomas Schmid

Zitationsvorschlag: Thomas Schmid: Wie man zwischen den Zahlen liest. Data-Mining und computergestützte Vorhersagen am Beispiel Bioinformatik. In: Arbeitstitel – Forum für Leipziger Promovierende Bd 5, Heft 1 (2013). S. 13–29.

urn:nbn:de:bsz:15-qucosa2-170006

Abstract

– deutsch –

Egal, ob Facebook, Ratingagenturen oder Erbgutanalysen: Immer umfangreicher und detaillierter wird die Welt des 21. Jahrhunderts digitalisiert und so am Computer erforschbar. Doch je größer die „Datenberge“, desto schwieriger sind darin verborgene Zusammenhänge zu erkennen. Wie und unter welchen Umständen dies trotzdem gelingen kann, illustriert ein Projekt zur digitalen Erforschung menschlicher und tierischer Körperzellen.

– englisch –

Facebook, rating agencies, genome analyses: In the 21st century, the world gets more and more digitalized and thereby computationally explorable. But the bigger the data, the harder the underlying concepts are to be identified. A research project that investigates human and animal body cells computationally illustrates how and under which conditions this can be successful after all.

1. Einleitung

Im beginnenden 21. Jahrhundert befindet sich die Realität des Menschen in einem tiefgreifenden Transformationsprozess. Mit zunehmender Vernetzung und Rechenkapazität werden immer mehr Entscheidungsfindungsprozesse an technische Systeme ausgelagert, Datenbanken und Algorithmen sollen Fragen klären, auf die Menschen selten schnelle, dafür oft unzuverlässige Antworten finden: An welcher Straßenecke wird sich Werbung für welche Konsumprodukte am meisten lohnen? Zu welchem Zeitpunkt wird der Fahrzeugmotor das nächste Mal ausfallen? Welches Fahrtziel wird der einzelne Verkehrsteilnehmer als nächstes ansteuern? Insbesondere in den westlichen Industrienationen verlässt sich der moderne Mensch durch diese Auslagerung in seinem Urteil über die Umwelt und in seinem realen Handeln immer öfter nur noch auf ein virtuelles Abbild der Realität.

Dieses Abbild ist eine Art Destillat aus einer ständig wachsenden Menge an Zahlen und Daten, die durch eine immer kleinteiligere und höher auflösende Vermessung der Welt entstehen. Kaum ein Bereich, der nicht von einer „Datifizierung" (Kreissl 2013) erfasst wird. Wo früher Anzeigetafeln Piloten bei der Kontrolle lebenswichtiger Maschinen unterstützten, funken heute Flugzeugmotoren einen steten Datenstrom an tausende Kilometer entfernte Rechenzentren (Austin et al. 2005). Wo früher Schüler mit Stift und Papier die Anzahl passierender Fußgänger oder Autos erheben mussten, entstehen heute durch Mobilfunktelefone und satellitengestützte Navigationssysteme automatisch zehntausende hochauflösende Bewegungsdaten (Song et al. 2010).

Infolgedessen ist der Alltag in den westlichen Industrienationen heute durch Informationsdestillate einer regelrechten „Bewusstseinsindustrie" (Fischermann/Hamann 2012) geprägt. Internetdienste wie Google, Facebook oder Wolfram Alpha versprechen das Wissen der Welt „für alle zu jeder Zeit zugänglich"[1] oder sogar „berechenbar"[2] zu machen. Obwohl in Wahrheit stets nur ein Ausschnitt aller Inhalte des Internets, nur ein kleiner Teil aller Menschen der Erde und höchstens ein Bruchteil des vollständigen Wissens der Menschheit auf diese Weise zugänglich ist, sind diese Dienste für viele Menschen längst zu einem „Index ihrer Wirklichkeit" (Bunz 2011) geworden. Auch die Ordnung dieses Indexes durch das jeweilige Unternehmen kann naturgemäß nicht neutral sein (Kohl 2013) und stellt somit stets eine subjektive Interpretation der verfügbaren Datengrundlage dar[3]. Bei Google etwa wird dies besonders anschaulich dadurch, dass sich Suchergebnisse mittlerweile[4] nicht mehr nur regional (Meserve/Pemstein 2012; Zittrain/Edelman 2002; Urban/Quilter 2005), sondern sogar von Nutzer zu Nutzer individuell unterscheiden (Pariser 2011).

Für den Einzelnen noch undurchschaubarer, jedoch um so folgenreicher sind diejenigen Informationsdestillate, die sich Kreditinstitute und andere Unternehmen zu wirtschaftlichen Zwecken schaffen. Aus allen über einen potentiellen Kunden verfügbaren Eigenschaften wird dabei eine Art „Datendoppelgänger" (Kreissl 2013) erzeugt, dessen Eigenschaften analysiert und daraus wiederum Schlussfolgerungen über den Kunden selbst gezogen werden. Die private deutsche Wirtschaftsauskunf-

1 Eigendarstellung des Unternehmens Google (http://www.google.com/about/company/; abgerufen am 1.6.2013).

2 Eigendarstellung des Internetdiensts Wolfram Alpha (http://www.wolframalpha.com/about.html; abgerufen am 1.6.2013).

3 Nach amerikanischer Rechtsauffassung ist eine solche Anordnung beziehungsweise Gewichtung wahrscheinlich sogar als Meinungsäußerung des jeweiligen Unternehmens zu werten (Cohen 2012).

4 Anfangs basierte die Gewichtung von Internetseiten bei Google im Wesentlichen auf der Häufigkeit, mit der Links – nach aktuellem Kenntnisstand – auf die jeweilige Seite gesetzt wurden (Brin/Page 1998).

tei Schufa etwa destilliert diese Doppelgänger durch die Verknüpfung und Analyse strukturierter Daten, die sie entweder aus öffentlichen Datenbeständen oder den an ihr beteiligten Kreditinstituten und Versicherungsunternehmen bezieht[5]. Neuere Ansätze die Kreditwürdigkeit von Antragsstellern zu bewerten stützen sich dagegen auf öffentliche oder dafür freizugebende Informationen aus sozialen Netzwerken, etwa indem ein mit dem Google-Ranking für Webseiten vergleichbarer Wert berechnet wird (Leber 2012). In Kombination mit ortsbezogenen Daten wie GPS-Koordinaten, Geräteeigenschaften wie dem verwendeten Betriebssystem oder protokolliertem Kaufverhalten können so mehrere tausend Eigenschaften in die Erzeugung eines Datendoppelgängers einfließen[6].

Auch staatliche Stellen suchen in ihren an Umfang und Komplexität immer schneller wachsenden Datenbeständen zunehmend nach Erkenntnisgewinn. In Deutschland setzen beispielsweise das Bundesamt für Verfassungsschutz, das Bundeskriminalamt und die Bundespolizei seit Jahren einschlägige Datenanalyse-Software zur Suche nach Mustern und Netzwerk-Beziehungen innerhalb ihrer Datenbanken ein.[7] Auch eine großflächige Auswertung privatwirtschaftlicher Daten[8] ist mittlerweile ein „alltägliches Ermittlungsinstrument“ (Beauftragter für Datenschutz und Informationsfreiheit 2013) deutscher Polizeibehörden, um „Personen als Beschuldigte zu identifizieren“ (Henrichs/Wilhelm 2010). Mehrere Softwareanwendungen zur Fahndung nach Verdächtigen, Zeugen oder Flüchtlingen beziehen zusätzlich auch Informationen aus Sozialen Netzwerken mit ein.[9] Noch komplexere Softwarepakete wie das vom US-amerikanischen Konzern IBM entwickelte Vorhersagesystem „Blue C.R.U.S.H.“ werden sogar zur zeitlichen und räumlichen Vorhersage von Verbrechen eingesetzt[10].

5 Laut Eigendarstellung speichert die Schufa pro Person sowohl personenbezogene Daten (wie Namen, Geburtsdatum, Geburtsort, aktuelle und frühere Anschriften) als auch Informationen über Bankkonten, Kreditkarten, Leasingverträge, Mobilfunkkonten, Versandhandelskonten, Ratenzahlungsgeschäfte, Kredite und Bürgschaften sowie etwaige Zahlungsausfälle (vgl. http://www.schufa.de/de/private/wissenswertes/faq/faq.jsp; „Welche Daten speichert die Schufa?“; abgerufen am 1.6.2013). Eine vom Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz in Auftrag gegebene Studie ergab 2009, dass aus einer Stichprobe von 100 Kunden-Datensätzen 45 Datensätze fehlerhafte, unvollständige oder falsche Daten enthielten (Korczak/Wilken 2009).

6 Solche Informationen stehen auf modernen Mobilfunkgeräten, insbesondere so genannten Smartphones, standardmäßig zur Verfügung. Das deutsche Unternehmen Kreditech etwa bezieht in die Bewertung von Kreditbewerbern laut Eigendarstellung (http://www.kreditech.com/about-us/; Folie 8 von 16; abgerufen am 1.6.2013) bis zu 8000 solcher Eigenschaften ein.

7 Einer Antwort der Bundesregierung auf eine Kleine Anfrage der Bundestagsfraktion DIE LINKE zufolge nutzen diese Behörden seit 16 bzw. 12 bzw. 10 Jahren das durch das US-amerikanische Unternehmen IBM vertriebene Software-Paket „Analyst‘s Notebook“ zur Datenanalyse (Bundestag 2012).

8 Im Rahmen von Protestkundgebungen gegen Rechtsextremismus forderte das Landeskriminalamt Sachsen allein im Februar 2011 innerhalb weniger Tagen hunderttausende Kommunikations- und zehntausende Kundendatensätze von Mobilfunkbetreibern an, um auf dieser Basis eine so genannte Funkzellenauswertung vorzunehmen (Landtag 2011).

9 Etwa das Softwarepaket „Accurint for Law Enforcement“ des US-amerikanischen Unternehmens LexisNexis (vgl. http://www.lexisnexis.com/government/solutions/investigative/accurint-le.aspx; abgerufen am 1.6.2013).

10 Blue C.R.U.S.H. wurde 2005 erstmals in Memphis im US-Bundesstaat Tennessee getestet, mittlerweile nutzen weitere US-Polizeibehörden die Software (Boeing et al. 2011). Eine mögliche Nutzung in durch das Bundeskriminalamt in Deutschland wird derzeit in Kooperation mit der Albert-Ludwigs-Universität Freiburg geprüft (Bundestag 2013). Auch Konkurrenzprodukte wie die von Forschern der Universität von Kalifornien entwickelte Software „PredPol“ werden bereits in mehreren amerikanischen Städten eingesetzt, etwa in Seattle im US-Bundesstaat Washington (Talbot 2013).

Die weite Verbreitung computergestützter Methoden zur Bergung verborgenen Wissens belegt ein tiefes Vertrauen in die Objektivität von Daten und Algorithmen. Doch Kritiker warnen vor zu viel Vertrauen: Überzeugende computergestützte Vorhersagen jeder Art – ob Fahrtziel, Kreditwürdigkeit oder Kriminalitätsrate – beruhten nicht allein auf objektiven Verfahren und fortgeschrittener Mathematik, sondern auch auf menschlicher Intuition und praktischem Verständnis des Anwendungsgebiets, sagt der Gründer einer vielgenutzten Internet-Plattform für Datenanalyse-Wettbewerbe (Fischermann/Hamann 2013). Und der US-amerikanische Technik-Historiker Kranzberg wurde berühmt mit seinem Postulat, Technologie als solche sei grundsätzlich weder gut noch schlecht noch neutral (Kranzberg 1986). Infolgedessen ist bei massenhafter Datenauswertung Stereotypisierung „unvermeidbar" (Andrews 2013); und die weitverbreitete Vorstellung, eine Internet-Suchmaschine präsentiere neutrale Suchergebnisse halten Fachleute schlicht für „Fiktion" (Fischermann/Hamann 2012).

Jedes der zuvor beschriebenen Beispiele nutzt ein individuelles virtuelles Abbild der Realität. Allen diesen Abbildern ist gemein, dass Außenstehende weder deren konkrete Eigenschaften nachvollziehen können noch den Einfluss dieser Eigenschaften auf die daraus gezogenen Schlüsse. Schwierigkeiten in der Datenanalyse und bei computergestützten Vorhersagen werden daher erst dort erkennbar, wo Modelle und Hypothesen schon *per definitionem* kritisch beäugt werden: in der Wissenschaft. In wirtschaftswissenschaftlichen Arbeiten zur Vorhersage von Kreditausfällen etwa liegen die Trefferraten selten über 80 Prozent (Khandani et al. 2010); im direkten Vergleich zu individuell definierten Expertensystemen für Kreditwürdigkeitsbewertung schneiden computergestützte Vorhersagen überwiegend schlechter ab (Aldrich 1995). Auch zur Analyse von Erbgut-Daten gibt es keine anerkannten Methoden mit hundertprozentiger Treffsicherheit – trotz zahlreicher technisch anspruchsvoller Ansätze (z.B. Eichner et al. 2011) sowie einer ständig wachsenden Zahl sequenzierter, das heißt zur digitalen Auswertung verfügbarer Genome[11].

Viele wissenschaftliche Anwendungen für Data-Mining und Vorhersagen zielen auf die Diagnostik schwerer und schwerster Krankheiten ab. Obwohl dabei bislang häufig auf Vererbungswege oder einfache Ja-Nein-Diagnosen fokussiert wird, bieten sich diese Methoden auch zur Verbesserung etablierter Messverfahren an. Ein Beispiel dafür ist die Vorhersage schwer messbarer elektrischer Eigenschaften von Epithel-Gewebe, auf welche das im Folgenden vorgestellte bioinformatische Forschungsprojekts abzielt. Eigenschaften wie die elektrische Leitfähigkeit[12] sind von großem medizinischem Interesse, weil sie direkte Rückschlüsse auf mögliche Fehlfunktionen dieser Barriere-bildenden Zellschichten und dem von ihnen regulierten Stoffaustausch (zwischen dem Inneren eines Organismus und seiner Umgebung) erlauben. Um Epithelgewebeproben und -zellkulturen zu charakterisieren, werden in der klinischen Forschung und Diagnostik häufig Wechselstrommessungen durchgeführt (Günzel et al. 2012). Die daraus resultierenden diskreten Messkurven erlauben allerdings zunächst nur eine grobe Abschätzung einiger weniger Eigenschaf-

11 Alleine in der „GenBank" des US-amerikanischen National Center for Biotechnology Information (NCBI) stieg die Zahl der gespeicherten Gensequenzen zwischen Dezember 2012 und Februar 2013 von rund 162 auf rund 163 Millionen, die der darin enthaltenen Basen von rund 148 auf rund 150 Milliarden; vgl. http://www.ncbi.nlm.nih.gov/genbank/statistics (abgerufen am 1.6.2013);

12 Elektrische Leitfähigkeit und elektrischer Widerstand sind rechnerisch in einander überführbare Größen (Kehrwert). Im Folgenden wird nur die Bezeichnung Widerstand verwendet.

ten (Fromm et al. 1985; Krug et al. 2009), und dies auch nicht unter allen physiologisch gegebenen Bedingungen (Schmid et al. 2013b).

Um die medizinische Diagnostik für Epithel-Erkrankungen wie Colitis ulcerosa oder Morbus Crohn zu verbessern, sollen mit dem vorgestellten Forschungsprojekt diejenigen Messkurveneigenschaften identifiziert werden, die die schnellste, präziseste und verlässlichste Vorhersage elektrischer Zelleigenschaften unter verschiedensten physiologischen Bedingungen ermöglichen. Die in der Praxis zu erwartenden Messdaten werden dazu möglichst detailgetreu nachgebildet (Modellierung) und mithilfe so genannter Data-Mining-Methoden auf vorhersage-geeignete Kurveneigenschaften untersucht (Hypothesenbildung). Diese potenziell nützlichen Eigenschaften werden dann in einem zweiten Schritt mit Methoden des so genannten maschinellen Lernens auf ihre tatsächliche Vorhersagekraft hin untersucht. Die so gewonnenen Erkenntnisse sollen letztlich zur Entwicklung eines Software-basierten Vorhersagesystems für die medizinische Forschung und Diagnostik verwendet werde

2. Computergestützte Hypothesen bildung

Wie bereits eingangs beschrieben, herrscht in modernen Data-Mining-Projekten an einbeziehbaren Eigenschaften – bei Wechselstrom-Messungen an Epithel-Gewebe zum Beispiel bis zu 100 Einzelinformationen pro Messung – typischerweise kein Mangel. Anschaulich entspricht eine große Anzahl an verfügbaren Eigenschaften einer sehr breiten Wertetabelle, während eine große Anzahl an Datensätzen eine sehr lange Wertetabelle bedeutet. Bei umfangreichen Projekten ist eine Beschränkung auf relevante Eigenschaften daher oft nicht nur aus wissenschaftlichen, sondern schon allein aus praktischen Gründen geboten. Die Reduzierung eines Modells auf relevante Eigenschaften wird typischerweise durch Reihung beziehungsweise Auswahl der Eigenschaften aufgrund ihrer Relevanz[13] erreicht und stellt eine grundlegende Form der Hypothesenbildung dar.

Unter einer Hypothese versteht man im Allgemeinen eine Aussage, deren Gültigkeit möglich, aber nicht bewiesen ist. In der Datenanalyse ist unter einer Hypothese ein vermuteter, mathematisch beschreibbarer Zusammenhang zwischen gegebenen Eingangsvariablen und einer oder mehreren Ausgangsvariablen zu verstehen; Eingangsvariablen besitzen dabei typischerweise einen numerischen Wert, eine Ausgangsvariable kann je nach Fragestellung sowohl einen numerischen (Regression) als auch einen kategorischen (Klassifikation) Wert aufweisen. Eingangsvariablen werden üblicherweise als Features, Ausgangsvariablen als Targets (Regression) oder Labels (Klassifikation) bezeichnet (Witten 2011).

Für das computergestützte Herausarbeiten einer Hypothese aus einer Wertetabelle oder Datenbank hat sich der Begriff „Data-Mining“[14] etabliert. Etwas konkreter kann Data-Mining als automatisierte (oder halb-automatisierte) Suche nach Mustern in Daten definiert werden (Witten 2011), bei der Konzepte und Algorithmen sowohl aus dem Bereich des maschinellen Lernens als auch aus Statistik, künstlicher Intelligenz und Datenmanagement zur Anwendung kommen (Harding et al 2006).

In der Regel wird zwischen drei verschiedenen Ansätzen zur Auswahl von Features (engl. feature selection) beziehungsweise zur Exklusion von Features (engl. dimension reduction) unterschieden (Mladenic

13 Es existieren mehrere, unterschiedliche Definitionen für den Begriff Relevanz (vgl. Kohavi/John 1997); hier genügt es, festzuhalten, dass mathematische Kriterien hierfür existieren.

14 In Anlehnung an das Schürfen nach Rohstoffen im Bergbau (engl. mining).

2006; Guyon/Elisseeff 2003). Im einfachsten Fall werden Features einzeln anhand ihrer Korrelation mit dem Zielwert gereiht und ausgewählt (Hall 1999). Sofern computergestützte Vorhersagen optimiert werden sollen, werden stattdessen häufig die zu verwendenden Vorhersagesysteme zur Evaluation von Features oder Feature-Subsets verwendet (Kohavi/John 1997). Ein dritter Ansatz kombiniert Feature Selection und Training des Vorhersagesystems unmittelbar (engl. embedded feature selection; Guyon/Elisseeff 2003).

Eine Feature-Selection-Hypothese, also eine Benennung für einen bestimmten Zielparameter ausschlaggebender Attribute, kann auch als Modell betrachtet werden. Darunter ist nach der Allgemeinen Modelltheorie von Stachowiak ein vereinfachtes Abbild der Wirklichkeit zu verstehen, das diese für bestimmte Subjekte, Zeiträume und Aufgaben ersetzt (Stachowiak 1973). Ein Modell zeichnet sich demnach insbesondere durch das Ignorieren zahlloser Eigenschaften der Wirklichkeit aus[15]. Nach Stachowiak kann Erkenntnis jeder Art ausschließlich in Form von Modellen oder durch Modelle erfolgen (Stachowiak 1973).

3. Computergestützte Vorhersagen

Aus der Anwendung von Feature-Selection-Methoden ergibt sich eine Nominierung von Eingangsvariablen, die potentiell zur Vorhersage der Ausgangsvariablen geeignet sind. Ein Modell dieser Art ist aber offensichtlich nicht ausdifferenziert genug, um damit konkrete Zielwerte vorhersagen zu können. Einerseits, weil man für die nominierten Eingangsvariablen keinen expliziten mathematischen Zusammenhang zur Ausgangsvariablen erhält, sondern höchstens eine Gewichtung oder Reihung (engl. ranking). Andererseits ist aber zum Beispiel bei Ergebnissen korrelationsbasierter Feature-Selection-Verfahren wie Filter-Methoden auch zu beachten, dass Korrelation nicht Kausalität impliziert (Aldrich 1995; Pearl 2009).

Der klassische Weg ein solches durch Feature-Selection gewonnenes Modell auszudifferenzieren, ist die so genannte Regressionsanalyse beziehungsweise multivariate Regressionsanalyse (im Falle mehrerer Zielparameter). Bei diesem in der Statistik häufig genutzten Verfahren wird eine mathematische Funktion ermittelt, deren Abweichung für gegebene Datenpunkte und Zielwerte minimal ist; die grundsätzliche Form der Funktion (linear, logistisch, exponentiell, etc.) ist dabei oft bereits durch Vorwissen aus dem Anwendungsgebiet gegeben.

Eine moderne und häufig genutzte Alternative zur Regressionsanalyse stellt das so genannte maschinelle Lernen (engl. machine learning) dar (Paliwal / Kumar 2009). Kennzeichen solcher lernenden Verfahren ist, dass der jeweils verwendete Algorithmus eine mathematische Funktion aus vorgegebenen Beispielen verallgemeinern kann. Obwohl dies üblicherweise in Software realisiert wird, ist es dafür nicht erforderlich, die zu erlernende Funktion zuvor explizit im Programmcode zu definieren; stattdessen wird nur das Lernverfahren selbst einprogrammiert (Samuel 1959). Machine-Learning-Verfahren existieren sowohl für Klassifikations- als auch Regressionsaufgaben (Alpaydin 2004, Kapitel 1). Bei einigen dieser Algorithmen ist das gelernte Wissen explizit dargestellt[16], viele erzeugen jedoch keine für Menschen unmittelbar interpretierbare Darstellung.

Die meisten etablierten Machine-Learning-Methoden erfordern Vorannahmen über die mathematische Form der zu er-

15 Dem britischer Statistiker George E. P. Box (geboren 1919) etwa wird die Bemerkung zugeschrieben, dass zwar jedes Modell falsch, manches Modell aber zumindest nützlich sei.

16 Bekanntestes Beispiel sind so genannte Entscheidungsbäume, engl. decision trees (vgl. Kothari / Dong 2002). Für Anwendungsbeispiele in der Bioinformatik siehe auch Chen et al. 2011.

lernenden Funktion und können jeweils überhaupt nur bestimmte Funktionsarten erlernen. Eine bedeutende Ausnahme stellen so genannte künstliche neuronale Netze dar, die theoretisch jede erdenkliche mathematische Funktion erlernen können (Hornik et al. 1989) und sich daher besonders für Regressionsaufgaben mit unbekannter Regressionsfunktion bewährt haben[17]. Ähnlich wie bei biologischen neuronalen Netzen handelt es sich um komplexe hierarchische Netzwerke, die sich aus einzelnen Recheneinheiten zusammensetzen (Krogh 2008). Diese Recheneinheiten entsprechen mathematischen Funktionen, die mittels numerischer Gewichtungen miteinander verbunden sind. Während eines Trainings mit Beispielen werden diese Gewichtungen verändert, um vorgegebene Zielwerte besser zu approximieren (Alpaydin 2004, Kapitel 2).

Wie gut ein Machine-Learning-Verfahren eine Funktion erlernt hat, wird anhand von Beispielen evaluiert, die dem Verfahren unbekannt sind. Dazu wird in der Regel eine vorhandene Menge an Beispielen in einen größeren Trainings- und einen kleineren Testdatensatz aufgeteilt. Sind nur wenige Beispiele verfügbar, wird diese Aufteilung zumeist mehrfach (überschneidungsfrei) wiederholt, um Überanpassung zu vermeiden (Browne 2000); dieses Vorgehen wird als Kreuzvalidierung bezeichnet. Da der Anwender die Zielwerte des Testdatensatzes kennt, können diese mit den Vorhersagen des Verfahrens verglichen werden. Als Fehlermaß kommt typischerweise entweder der durchschnittliche absolute Fehler, der durchschnittliche relative Fehler oder die Wurzel aus dem Durchschnitt der Fehlerquadrate (engl. root mean squared error, RMSE) zur Anwendung.

17 Für Anwendungsbeispiele in der Biotechnologie, Biochemie und Mikrobiologie siehe auch Montague / Morris 1994.

4. Vom Epithel-Gewebe zum Epithel-Modell

Epithelien sind Zellen, die in jedem vielzelligen tierischen Organismus zahlreich und in zahlreichen Variationen zu finden sind (Al-Awqati 2011). Sie bilden eine oder mehrere Schichten, die das Innere des Organismus von der Umgebung trennen und den Austausch von Molekülen mit ihr regulieren (Groschwitz / Hogan 2009). Das so entstehende Epithel kann aus einer Schicht oder mehreren Schichten von Zellen bestehen; in der Regel befindet sich darunter die so genannte Basalmembran (engl. basement membrane, auch: subepithelium), das eine weitere, allerdings deutlich dünnere Barriere darstellt (Turner / Madara 2009). Im Folgenden werden einschichtige Epithele (und davon abgeleitete Zellkulturen) betrachtet, wie sie in Darm (Kato/Owen 1999) und Niere (Zimmermann 1911) vorkommen. Fehlfunktionen des Epithels im Darm können zu Diarrhöe oder chronisch entzündlichen Darmerkrankungen wie Colitis ulcerosa oder Morbus Crohn führen (Herrlinger 2009).

Welche und wie viele Moleküle über ein Epithel ins Innere oder nach außen gelangen, hängt von zahlreichen Faktoren ab. Grundsätzlich können Wasser, Ionen oder Nährstoffe sowohl durch Zellen hindurch (transzellulär) als auch durch Zellzwischenräume (parazellulär) transportiert werden (Sachs et al. 1973; vgl. Abb. 1). Ein Transport durch die Zellzwischenräume erfolgt passiv, also durch äußere chemisch-physikalische Kräfte getrieben, und wird primär durch Tight Junctions genannte Zell-Zell-Kontakte reguliert (Farquhar/Palade 1963). Ein Transport durch die Zellen kann sowohl aktiv, also unter Energieverbrauch, als auch passiv erfolgen.

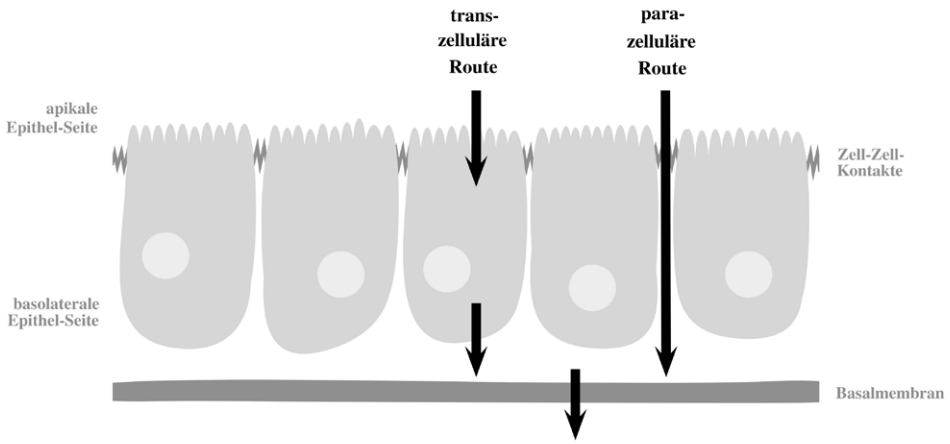

Abbildung 1: Schematische Zeichnung eines einschichtigen Epithels. Die Zellen des Epithels sind durch Zell-Zell-Kontakte (engl. tight junctions) miteinander verbunden. Diese trennen die dem Körperäußeren zugewandte (apikale) und die dem Körperinneren zugewandte (basolaterale) Membranen, welche sich in ihrer Zusammensetzung unterscheiden. Ein Transport von Stoffen über das Epithel kann sowohl durch die Zellen hindurch (transzellulär) als auch durch die Zellzwischenräume (parazellulär) erfolgen; in beiden Fällen müssen Stoffe danach die so genannte Basalmembran überwinden, die sich direkt unterhalb des Epithels anschließt.

Ein Epithel weist viele Kontaktstellen zwischen benachbarten Zellen und typischerweise vergleichsweise kleine Zellzwischenräume auf (Marchiando et al. 2010). Die dem Äußeren zugewandte (apikale) Seite wird dadurch strukturell und funktionell von der dem Inneren zugewandten (basolateralen) Seite getrennt (Rodriguez-Boulan / Nelson 1989). Da sich die apikalen Zellmembranen hinsichtlich der enthaltenden Lipide und Proteine meist deutlich von den basolateralen Membranen unterscheiden, werden Epithel-Zellen auch als polare Zellen bezeichnet (Handler 1989).

Die Modellierung von Zellen und Zellschichten kann grundsätzlich auf zwei unterschiedliche Arten erfolgen: entweder einem Bottom-Up- oder einem Top-Down-Ansatz folgend. In Bottom-Up-Ansätzen versucht man auf Basis physikalischer (etwa über Teilchenbewegungen oder elektrische Ladungen) und chemischer Grundannahmen (etwa Atommasse) das Wechselspiel von Molekülen unter zell- bzw. organismusähnlichen Bedingungen zu modellieren bzw. vorherzusagen. Mit so genannten Molecular-Dynamics-Simulationen (Allen 2004) wird beispielsweise versucht, das Verhalten von Proteinen nachzuahmen und zu analysieren (Klepeis et al. 2009). Aufgrund des damit verbundenen Rechenaufwands ist dies in der Regel allerdings entweder auf die Ebene einzelner Membranen oder auf Zeitspannen limitiert, die kaum Rückschlüsse auf das makroskopische Verhalten erlauben. Darüber hinaus hat sich insbesondere die realistische Simulation makroskopischer Membraneigenschaften wie der elektrischen Kapazität oder der dielektrischen Konstante als problematisch erwiesen (Nymeyer / Zhou 2008; Stern / Feller 2003; Zhou / Schulten 1995).

In der Physiologie wird dagegen traditionell eher ein Top-Down-Ansatz zur Modellierung verwendet. Dieser beschränkt sich auf elektrische Eigenschaften von Zellen: In elektrophysiologischen Messungen gemachte Beobachtungen werden häufig erklärbar, indem Zellschichten als elektrische Schaltkreise betrachtet werden (Günzel et al. 2012). Insbesondere werden Zellmembranen als so genannte RC-Glieder betrachtet, bestehend aus je

einem Widerstand (R, von engl. resistor) und einem Kondensator (C, von engl. capacitor). Je nach Beobachtung und Zelltyp werden elektrische Schaltkreise unterschiedlicher Komplexität angenommen. Der einfachste Schaltkreis, der explizit berücksichtigt, dass sich die nach innen und die nach außen gerichtete Seite eines Epithels deutlich unterscheiden können, besteht aus zwei in Reihe geschalteten RC-Gliedern und einem parallelem Widerstand (der die Zellzwischenräume repräsentiert); ein weiterer, dazu in Reihe geschalteter Widerstand repräsentiert die vorgelagerte Basalmembran (Günzel et al. 2012; vgl. Abb. 2). Im Folgenden wird entsprechend der Gepflogenheiten in der Physiologie davon ausgegangen, dass Messungen an Epithel-Gewebe und an einem Epithel-äquivalenten Schaltkreis zu identischen Messergebnissen führen.

5. Data-Mining und Vorhersagen für Epithel-Modelle

Ziel des hier vorgestellten Forschungsprojekts ist es, aus Wechselstrommessungen an Epithel-Gewebe verlässliche Vorhersagen über die Komponenten eines zugrunde gelegten elektrischen Schaltkreises treffen zu können. Dazu sollen so genannte impedanzspektroskopische Messungen an zwei Referenzzelltypen ausgewertet werden, die von Kooperationspartnern untersucht werden. Da die wahren Zielwerte naturgemäß nicht bekannt sind[18], werden stattdessen zunächst auf Basis des zuvor beschriebenen sechselementigen Schaltkreises (vgl. Abb. 2) künstliche Messdaten erzeugt. Diese modellierten Messkurven werden auf Muster untersucht; insbesondere sollen Features dieser Wechselstrommessungen identifiziert werden, die eine zuverlässige Vorhersage der Werte einzelner Elemente des Schaltkreises erlauben.

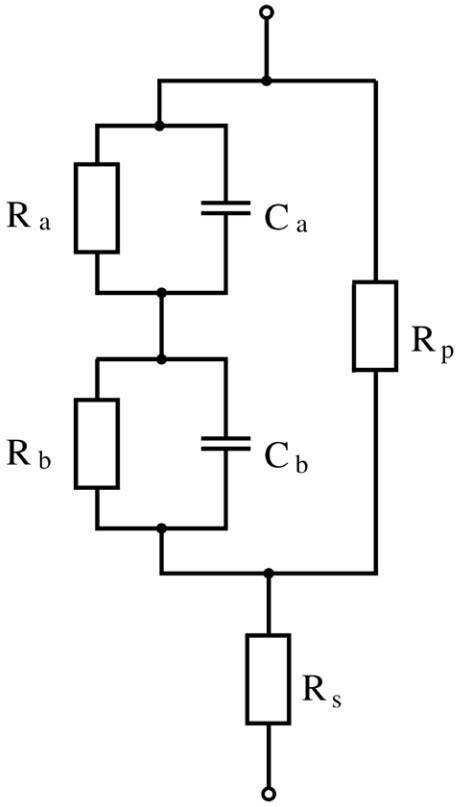

Abbildung 2: Elektrisches Ersatzschaltbild für einschichtige Epithelien. Die parazelluläre Transportroute wird durch den parazellulären Widerstand (R_p) beschrieben, die parallel zur transzellulären Transportroute verläuft; diese gliedert sich in je ein RC-Glied für die Eigenschaften der apikalen Seite (R_a, C_a) und der basolateralen Seite (R_b, C_b). Der in diesen Bauteilen in Reihe nachgeschaltete subepitheliale Widerstand (R_s) repräsentiert die elektrischen Eigenschaften der Basalmembran.

Um Zusammenhänge zwischen Schaltkreis-Komponenten und diesen Wechselstrommessungen finden zu können, ist zunächst eine ausdifferenzierte Modellierung der Schaltkreis-Eigenschaften erforderlich. Die sechs Elemente des hier für Epithel-Gewebe angenommenen Ersatz-Schaltkreises können je nach Zelltyp und physiologischer Umgebung unterschiedliche Werte annehmen. Zelltypspezifische Grenzwerte für jedes Element sind aufgrund der damit verbundenen aufwendigen Messungen allerdings nicht für alle Zelltypen verfügbar. Für die von Darmzellen

18 In der Messtechnik wird der wahre Wert einer Messgröße als ideeller Wert betrachtet, der nicht exakt bekannt ist (Norm DIN 1319-1 1995).

abgeleiteten Zellkulturen HT-29/B6 und IPEC-J2 sind entsprechende Ober- und Untergrenzen der angenommenen Widerstände und Kondensatoren sowohl unter natürlichen Bedingungen als auch unter Einfluss bestimmter medizinischer Wirkstoffe benennbar (Schmid et al. 2013b); sie dienen hier als Referenzzelltypen.

Weiter müssen die für den jeweiligen Zelltyp zu erwartenden Messkurven möglichst präzise modelliert werden. Bei impedanzspektroskopischen Messungen wird nacheinander Wechselstrom mit variierenden Frequenzen auf ein Untersuchungsobjekt angewendet. Das Ergebnis ist eine Reihe komplexwertiger Zahlen, so genannter Impedanzen. Eine Impedanz lässt sich für Bauteile wie Widerstand oder Kondensator in Abhängigkeit von der angewendeten Frequenz eindeutig beschreiben. Auch die sich für den hier betrachteten sechselementigen Schaltkreis theoretisch ergebende Impedanz lässt sich somit berechnen, falls die Werte aller zugrundeliegenden Bauteile bekannt sind (Schmid et al. 2010). Wird die Impedanz dann für jede verwendete Frequenz berechnet, erhält man eine idealisierte Messkurve; in impedanzspektroskopischen Messungen werden typischerweise einige Dutzend Frequenzen zwischen 1 Hz und 100 kHz angewendet (Günzel et al. 2012). Variiert man die zur Berechnung angenommenen Werte der Schaltkreis-Elemente innerhalb der Grenzwerte der Referenzzelltypen, erhält man insgesamt ein beliebig skalierbares Abbild der erwartbaren Messergebnisse.

Grundsätzlich ist die Vorhersage verschiedener Schaltkreis-Eigenschaften aus impedanzspektroskopischen Messkurven vorstellbar; hier sollen zunächst nur der epitheliale Widerstand sowie der Widerstand der Basalmembran (im folgenden subepithelialer Widerstand genannt) vorhergesagt werden. Wie zuvor beschrieben, geht dem Einsatz von Vorhersagesystemen typischerweise eine Form der Hypothesenbildung voraus. Da die komplexwertigen Impedanzen einer Messung häufig in einem zweidimensionalen kartesischen Koordinatensystem aufgetragen werden, bietet sich hier statt eines echten Feature-Selection-Verfahrens ein vergleichsweise triviales Vorgehen an: Die vorherzusagenden Widerstände sind identisch mit den beiden hypothetischen Schnittpunkten der diskreten Messkurven mit der x-Achse (Günzel et al. 2012), weshalb intuitiv ein aussagekräftiger Zusammenhang zwischen den kartesischen Koordinaten der ersten zehn Impedanzen und dem epithelialen Widerstand sowie den letzten zehn Impedanzen und dem subepithelialen Widerstand unterstellt werden kann (Schmid et al. 2013b).

Um diesen Zusammenhang zu testen, wurden lernfähige Algorithmen mit mehreren tausend Modellmesskurven trainiert, die jeweils entsprechend der Annahmen über die Referenzzelltypen erzeugt wurden. Es konnte gezeigt werden, dass die hier intuitiv angenommenen Zusammenhänge bereits ausreichen, um bessere Vorhersagen der Zielwerte machen zu können als mit herkömmlichen Verfahren (Schmid et al. 2013b). Gleichzeitig konnte gezeigt werden, dass auf diese Art Modellmesskurven erzeugt werden können, die sowohl in Form als auch Zielwerten tatsächlich gemessenen impedanzspektroskopischen Kurven entsprechen (Schmid et al. 2013b). Um die Zielwerte von modellierten und gemessenen Kurven zu vergleichen, wurde auf einen indirekten Vergleich zurückgegriffen.

Nachteil dieses biologisch motivierten Ansatzes ist, dass die Erzeugung der zum Training genutzten künstlichen Messdaten konkrete Grenzwert-Annahmen über die zelltypäquivalenten Schaltkreise erfordert. Das bedeutet insbesondere, dass Trainingsdaten für jeden Zelltyp neu angepasst werden müssen. Die Vorhersagemethode ist also direkt abhängig vom zu untersuchenden Zelltyp. Gleichzeitig haben vorhergesagte Zielwerte kaum Aussagekraft,

wenn beim untersuchten Gewebe eine oder mehrere Schaltkreis-Komponenten Werte außerhalb der vorausgesetzten Grenzwerte annehmen können. Dieses Vorgehen ist folglich ausschließlich bei Zelltypen zuverlässig anwendbar, die gut untersucht und für die detaillierte Grenzwerte bekannt sind.

Eine entscheidende Frage ist daher, ob sich Zusammenhänge finden lassen zwischen Zielwerten und Messkurveneigenschaften, die nicht zelltypspezifisch sind. Dies wurde bereits an einem konkreten Beispiel untersucht: Mit Data-Mining-Verfahren wurden globale Messkurveneigenschaften gesucht, die Aussagen über den epithelialen Widerstand erlauben (Schmid et al. 2013a). Dazu wurden für die beiden Referenzzelltypen künstliche Messkurven erzeugt und jeweils globale Eigenschaften jeder diskreten Kurve berechnet, also etwa Minimal- oder Maximalwerte. Besonders aussagekräftig hinsichtlich des zu erwartenden epithelialen Widerstands erschien dabei die maximale Magnitude einer impedanzspektroskopischen Messung. Dieser Zusammenhang wurde dann durch Training künstlicher neuronaler Netze mit dieser und weiteren Eigenschaften und anschließender Evaluation der dadurch möglichen Vorhersagen bestätigt (Schmid et al. 2013a).

6. Einschränkungen und Heraus forderungen

Die hier skizzierten Vorhersagemöglichkeiten elektrischer Eigenschaften von Epithel-Gewebe aus impedanzspektroskopischen Messkurven unterliegen verschiedenen Einschränkungen. Diese resultieren vor allem aus der Modellierung von Gewebe und Messkurven, die dem Vorhersage-Verfahren zugrunde liegt.

Eine eher praktische Einschränkung entsteht durch die unterschiedlichen Frequenzspektren, die für impedanzspektroskopische Messungen verwendet werden können. Nutzt das Vorhersagesystem ein anderes Frequenzspektrum als das Gerät, mit dem zu testende Messkurven aufgezeichnet wurden, sind sinnvolle Vorhersagen von vornherein ausgeschlossen. Diese Einschränkung ist jedoch durch Nutzung standardisierter Geräte beziehungsweise Geräteeinstellungen oder durch Erzeugung mehrerer Vorhersagesysteme mit unterschiedlichen Frequenzspektren umgehbar.

Eine grundsätzlichere Einschränkung entsteht durch die praktische Notwendigkeit den gerätespezifischen Messfehler zu modellieren. Sofern Vorhersagen nicht ausschließlich für idealisierte Messkurven (z.B. innerhalb reiner Simulationen), sondern auch für tatsächlich gemessene Messkurven getroffen werden sollen, ist dies unverzichtbar. Wird der Gerätefehler jedoch unrealistisch modelliert (z.B. kein Fehler, zu grobe Schätzung oder Verwendung des Gerätefehlers eines anderen Gerätes), können infolgedessen auch sämtliche damit modellierten Messkurven nicht als realistisch erachtet werden. Entsprechend sind sämtliche nachfolgenden Feature-Selection-Ergebnisse und Vorhersagen korrumpiert.

Noch grundlegender ist die Einschränkung, die durch die Verwendung des sechselementigen Schaltkreises als Epithel-Modell entsteht. Für die beiden hier verwendeten Referenzzelltypen sehen Physiologen dieses Modell im Allgemeinen als adäquat an (Günzel et al. 2012), da die wichtigsten physiologischen Eigenschaften abgebildet sind. Für noch komplexeres, beispielsweise mehrschichtiges Epithel-Gewebe, eignet sich dieses Ersatz-Schaltbild dagegen nicht. Die physiologische Plausibilität eines für Modellierung und Data-Mining verwendeten Schaltkreises muss daher zwingend gegeben sein, damit darauf basierende Vorhersagen als plausibel angenommenen werden können.

7. Zusammenfassung und Ausblick

Sowohl in Wirtschaft als auch Wissenschaft sind Data-Mining und computergestützte Vorhersagen heute weit verbreitet. Im öffentlichen und selbst im fachlichen

Diskurs wird dabei oft übersehen, welchen Einfluss Modellierung und anwendungsspezifische Gegebenheiten auf Erfolg und Zuverlässigkeit solcher Verfahren haben.

Das hier beschriebene Forschungsprojekt illustriert exemplarisch typische Probleme. Bei der Modellierung von Messkurven etwa wird ersichtlich, dass unrealistische Annahmen über das Verhalten der zu untersuchenden Zellen das gesamte Vorgehen korrumpieren: Unrealistische Annahmen führen nicht nur zu unrealistisch modellierten Messkurven, sondern im weiteren Verlauf auch zu schlecht ausgerichteten Vorhersagensystemen. Die bisherigen Ergebnisse des Projekts zeigen, dass diese Fallstricke bei der Vorhersage von epithelialem und subepithelialem Widerstand weitgehend vermieden werden konnten und das Verfahren plausible Vorhersagen ermöglicht.

In künftigen Arbeitsschritten soll untersucht werden, ob für den sechselementigen Ersatz-Schaltkreis (vgl. Abb. 2) das Verhältnis der Zeitkonstanten der beiden RC-Glieder aus einer impedanzspektroskopischen Messung vorhergesagt werden kann. Insbesondere soll untersucht werden, wie viele und welche Messkurveneigenschaften für diese Aufgabe mindestens erforderlich sind; als Arbeitshypothese wird angenommen, dass das Verhältnis der Zeitkonstanten eng mit der Form der impedanzspektroskopischen Messkurven verbunden ist. Eine zuverlässige Bestimmung wird auch für die Vorhersage weiterer Eigenschaften wie der epithelialen Kapazität hilfreich sein.

Mittelfristig soll außerdem untersucht werden, ob Vorhersagen ohne vorherige Prägung des Vorhersagesystems auf konkrete Frequenzen getroffen werden können. Dazu ist beispielsweise eine Gruppierung der in den konkreten impedanzspektroskopischen Messungen verwendeten Frequenzen in Frequenzbereiche vorstellbar; Vorhersagemethoden könnten sich dann auf Frequenzbereiche statt auf einzelne Frequenzen beziehen. Zu prüfen wird dabei insbesondere sein, ob ein solcher Ansatz eine ähnliche Vorhersagegenauigkeit aufweist wie ein frequenzspezifischer.

Literatur

Al-Awqati, Qais (2011): Terminal Differentation in Epithelia: The Role of Intergrins in Hensin Polymerization. In: *Annual Review of Physiology* 73, S. 401–412.

Aldrich, John (1995): Correlations genuine and spurious in pearson and yule. In: *Statistical Science* 10, Nr. 4, S. 364–376.

Allen, Michael P. (2004): Introduction to Molecular Dynamics Simulation. In: Attig, Norbert et al. (Hrsg.): Computational Soft Matter: From Synthetic Polymers to Proteins, NIC Series 23, S. 1–28.

Alpaydin, Ethem (2004): Introduction to Machine Learning. MIT Press.

Andrews, Lori (2013): Wie die Datensammel-Industrie hinter Facebook und co. funktioniert. In: *Süddeutsche Zeitung*, 10. Februar 2013.

Austin, Jim et al. (2005): DAME: Searching large data sets within a grid-enabled engineering application. In: *Proceedings of the IEEE* 93, Nr. 3, S. 496–509.

Ben-David, Arie / Frank, Eibe (2009): Accuracy of machine learning models versus "hand crafted" expert systems – a credit scoring case study. In: *Expert Systems with Applications* 36, Nr. 3, S. 5264–5271.

Boeing, Niels et al. (2011): Tatort Zukunft. In: Technology Review, Nr. 3, S. 46–51.

Brin, Sergey / Page, Lawrence (1998): The anatomy of a large-scale hypertextual web search engine. In: *Computer networks and ISDN systems* 30, Nr. 1, S. 107–117.

Beauftragter für Datenschutz und Informationsfreiheit, Berliner (2013): Jahresbericht 2012, Kapitel 2.1, S. 23–27.

Browne, Michael W. (2000): Cross-Validation Methods. In: *Journal of Mathematical Psychology* 44, S. 108–132.

Bundestag, Deutscher (2012): Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Ulla Jelpke, Andrej Hunko, Christine Buchholz, weiterer Abgeordneter und der Fraktion DIE LINKE – Automatisierte Strafverfolgung, Data Mining und sogenannte erweiterte Nutzung von Daten in polizeilichen Informationssystemen, Drucksache 17/11582 vom 22. November 2012.

Bundestag, Deutscher (2013): Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Andrej Hunko, Annette Groth, Harald Koch, weiterer Abgeordneter und der Fraktion DIE LINKE – Entwicklung einer Meta-Suchmaschine für internationale, europäische und nationale Polizeidatenbanken durch EUROPOL, Drucksache 17/13441 vom 10. Mai 2013.

Bunz, Mercedes (2011): Das Denken und die Digitalisierung. In: *Frankfurter Allgemeine Zeitung,* Nr. 18, S. Z1, 22.1.2011.

Chen, Xiang et al. (2011): The use of classification trees for bioinformatics. In: *WIREs Data Mining and Knowledge Discovery* 1, S. 55–63.

Cohen, Noam (2012): Professor Makes the Case That Google Is a Publisher. In: *New York Times*, S.B3, 21.5.2012.

Eichner, Johannes et al. (2011): Support vector machines-based identification of alternative splicing in arabidopsis thaliana from whole-genome tiling arrays. In: *BMC Bioinformatics* 12, Nr. 55, o.S.

Farquhar, Marilyn G. / Palade, George E. (1963): Junctional complexes in various epithelia. In: *Journal of Cell Biology* 17, S. 375–412.

Fischermann, Thomas / Hamann, Götz (2012): „Neutrale Suchergebnisse sind eine Fiktion". In: *Die Zeit*, Nr. 38, 13.9.2012.

Fischermann, Thomas / Hamann, Götz (2013): Wer hebt das Datengold? In: *Die Zeit*, Nr. 2, 3.1.2013.

Fromm, Michael et al. (1985): Epithelial and subepithelial contributions to transmural electrical resistance of intact rat jejunum, in vitro. In: *Pflügers Archiv* 405, S. 400–402.

Groschwitz, Katherine R. / Hogan, Simon P. (2009): Intestinal barrier function: Molecular regulation and disease pathogenesis. In: *Journal of Allergy and Clinical Immunology* 124, Nr. 1, S. 3–20.

Guyon, Isabelle / Elisseeff, André (2003): An introduction to variable and feature selection. In: *Journal of Machine Learning Research* 3, S. 1157–1182.

Günzel, Dorothee et al. (2012): From TER to trans- and paracellular resistance: lessons from impedance spectroscopy. In: *Annals of the New York Academy of Sciences* 1257, S. 142–151.

Hall, Mark A. (1999): Correlation-based feature selection for machine learning. PhD thesis, University of Waikato.

Handler, Joseph S. (1989): Overview of epithelial polarity. In: *Annual Review of Physiology* 51, S. 729–740.

Harding, Jenny A. et al (2006): Data mining in manufacturing: a review. In: *Journal of Manufacturing Science and Engineering* 128, Nr. 4, S. 969–976.

Henrichs, Axel / Wilhelm, Jörg (2010): Funkzellenauswertung – Teil 2. In: *Die Kriminalpolizei*, Nr. 6, o.S.

Herrlinger, Klaus et al. (2009): Chronisch entzündliche Darmerkrankungen. In: *Der Internist* 50, Nr. 10, S. 1229–1248.

Hornik, Kurt et al. (1989): Multilayer feedforward networks are universal approximators. In: *Neural Networks* 2, Nr. 5, S. 359–366.

Kato, Tomohiro / Owen, Robert L. (1999): Structure and function of intestinal mucosal epithelium. In: *Mucosal Immunology.* 2. Aufl. Academic Press, San Diego, CA.

Khandani, Amir E. et al. (2010): Consumer credit-risk models via machine-learning algorithms. In: *Journal of Banking & Finance* 34, Nr. 11, S. 2767–2787.

Klepeis, John L. et al. (2009): Long-timescale molecular dynamics simulations of protein structure and function. In: *Current Opinion in Structural Biology* 19, Nr. 2, S. 120–127.

Kohavi, Ron / John, George H. (1997): Wrappers for feature subset selection. In: *Artificial intelligence* 97, Nr. 1, S. 273–324.

Kohl, Uta (2013): Google: the rise and rise of online intermediaries in the governance of the Internet and beyond (Part 2). In: *International Journal of Law and Information Technology* 21, Nr. 2, S. 187–234.

Korczak, Dieter / Wilken, Michael (2009): Verbraucherinformation Scoring. Bundesministerium für Landwirtschaft, Ernährung und Verbraucherschutz.

Kothari, Ravi / Dong, Ming (2002): Decision trees for classification: a review and some new results. In: Pal SK / Pla A (Hrsg.): Pattern recognition from Classical to Modern Approaches. Singapore: World Scientific Publishing Companys, S. 169–186.

Kranzberg, Melvin (1986): Technology and history: "Kranzberg's laws". In: *Technology and Culture* 27, Nr. 3, S. 544–560.

Kreissl, Reinhard (2013): Komplette Umkehr der Beweislast. In: *New Scientist*, 22. Februar 2013.

Krogh, Anders (2008): What are artificial neural networks? In: *Nature Biotechnology* 26, Nr. 2, S. 195–197.

Krug, Susanne M. et al. (2009): Two-path impedance spectroscopy for measuring paracellular and transcellular epithelial resistance. In: Biophysical Journal 97, S. 2202–2211.

Landtag, Sächsischer (2011): Bericht zu den nichtindividualisierten Funkzellenabfragen und anderen Maßnahmen der Telekommunikationsüberwachung durch Polizei und Staatsanwaltschaft Dresden in Bezug auf den 13., 18. und 19. Februar 2011 in Dresden (Drucksache 5/6787).

Leber, Jesscia (2012): Can a credit score be crowdsourced? In: *Technology Review,* 7.6.2012.

Marchiando, Amanda M. et al. (2010): Epithelial Barriers in Homeostasis and Disease. In: *Annual Review of Pathology: Mechanisms of Disease* 5, S. 119–144.

Meserve, Stephen A. / Pemstein, Daniel (2012): Google politics: The political determinants of internet censorship. In: 2nd Annual General Conference of the European Political Science Association.

Mladenić, Dunja (2006): Feature selection for dimensionality reduction. In: *Lecture Notes in Computer Science* 3940, S. 84–102.

Montague, Gary / Morris, Julian (1994): Neural-network contributions in biotechnology. In: *Trends in Biotechnology* 12, Nr. 8, S. 312–324.

Norm DIN 1319-1 (1995): Grundlagen der Meßtechnik – Teil 1: Grundbegriffe.

Nymeyer, Hugh / Zhou, Huan-Xiang (2008): A method to determine dielectric constants in nonhomogeneous systems: application to biological membranes. In: *Biophysical Journal* 94, Nr. 4, S. 1185–1193.

Paliwal, Mukta / Kumar, Usha A. (2009): Neural networks and statistical techniques: A review of applications. In: *Expert Systems with Applications* 36, S. 2–17.

Pariser, Eli (2011): The Filter Bubble: What the Internet Is Hiding from You. Penguin Press HC.

Pearl, Judea (2009): Causal inference in statistics: An overview. In: *Statistics Surveys* 3, S. 96–146.

Rodriguez-Boulan, Enrique / Nelson, W. James (1989): Morphogenesis of the Polarized Epithelial Cell Phenotype. In: *Science* 245, Nr. 4919, S. 718–725.

Sachs, George et al. (1973): Conductance pathways in epithelial tissues. In: *Experimental Eye Research* 16, Nr. 4, S. 241–249.

Samuel, Arthur L. (1959): Some studies in machine learning using the game of checkers. In: *IBM Journal of Research and Development* 3, Nr. 3, S. 210–229.

Schmid, Thomas et al. (2010): Using an artificial neural network to determine electrical properties of epithelia. In: *Lecture Notes in Computer Science* 6352, S. 211–216.

Schmid, Thomas et al. (2013a): Efficient prediction of x-axis intercepts of discrete impedance spectra. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Brügge/Belgien, S. 185–190.

Schmid, Thomas et al. (2013b): Discerning apical and basolateral properties of HT-29/B6 and IPEC-J2 cell layers by impedance spectroscopy, mathematical modeling and machine learning. In: *PLOS ONE* 8, Nr. 7, o.S.

Song, Chaoming et al. (2010): Modelling the scaling properties of human mobility. In: *Nature Physics* 6, Nr. 10, S. 818–823.

Stachowiak, Herbert (1973): Allgemeine Modelltheorie. Springer.

Stachowiak, Herbert (1983): Modelle – Konstruktion der Wirklichkeit. Wilhelm Fink.

Stern, Harry A. / Feller, Scott E. (2003): Calculation of the dielectric permittivity profile for a nonuniform system: Application to a lipid bilayer simulation. In: *Journal of Chemical Physics* 118, Nr. 7, S. 3401–3412.

Talbot, David (2013): Crime Software: Still Awaiting a Verdict. http://www.technologyreview.com/view/512121/ (abgerufen am 1.6.2013).

Turner, Jerrold R. / Madara, James L. (2009): Epithelia: biological principles of organization. In: Yamada, Tadataka (Ed.): Textbook of Gastroenterology 1. 5. Aufl. Wiley-Blackwell, S. 169–186.

Urban, Jennifer M. / Quilter, Laura (2005): Efficient process or chilling effects-takedown notices under section 512 of the digital millennium copyright act. In: *Santa Clara Computer & High Tech. LJ* 22, S. 621ff.

Witten, Ian H. (2011): Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Zhou, Feng / Schulten, Klaus (1995): Molecular Dynamics Study of a Membrane-Water Interface. In: *Journal of Physical Chemistry* 99, Nr. 7, S. 2194–2207.

Zimmermann, K.W. (1911): Zur Morphologie der Epithelzellen der Säugetierniere. In: *Archiv für mikroskopische Anatomie* 78, Nr. 1, S. 199–231.

Zittrain, Jonathan / Edelman, Benjamin (2002): Localized google search result exclusions. a statement of issues and call for data. http://cyber.law.harvard.edu/filtering/google/ (abgerufen am 1.6.2013).