Big-Data Analytics transformiert die Lebenswissenschaften

Ivo F. Sbalzarini

Die Lebenswissenschaften – Biologie, Medizin und Psychologie – sind im Unterschied zu anderen Naturwissenschaften wie Physik oder Chemie nicht theoriegetrieben. In der Physik ist die Theorie den Experimenten oft Jahrzehnte voraus, z. B. in der Quantenphysik oder der Relativitätstheorie, und auch in der Chemie ist seit der Entdeckung des Periodensystems der Elemente, und insbesondere seit dem Verständnis der Nuklear- und Quantenphysik Mitte des 20. Jahrhunderts, eine theoretische Grundlage vorhanden, welche die Wissenschaft von der Alchemie zur modernen Chemie transformierte. Aufgrund dieser theoretischen Grundlagen können Physik und Chemie die Dynamik komplexer Systeme mathematisch modellieren und in numerischen Rechnersimulationen dieser Modelle Verhalten vorhersagen und Systeme optimieren. Derart grundlegende und prädiktive Theorien sind in den Lebenswissenschaften praktisch nicht vorhanden.

Die Gegenbeispiele, wo solche Theorien doch vorhanden sind, zeigen aber, dass sie grundsätzlich möglich sind. Darwins Evolutionstheorie ist so ein Gegenbeispiel. Vor Darwin erschien die Vielfalt der Arten und Lebewesen verwirrend, ihre Entstehung mysteriös und Beziehungen zwischen ihnen waren unbekannt. Durch die einfache Idee der Merkmalevolution („survival of the fittest“) brachte Darwins Theorie jedoch Struktur und Klarheit, noch vor der Entdeckung der molekularen Grundlagen wie DNS, Gene und Aminosäuren. Dank Darwins Theorie können wir mathematische Modelle der Evolution aufstellen und Stammbäume der Arten algorithmisch errechnen, auch rückwärts in der Zeit. Ähnlich verhält es sich in der Genetik, wo seit Mendels Theorie der Vererbung, ebenfalls vor der Entdeckung von DNS aufgestellt, die Entwicklung von Merkmalen z. B. in der Tier- und Pflanzenzucht planbar und berechenbar wurde. Die Gemeinsamkeit dieser Theorien ist, dass sie biologische *Prozesse* modellieren, nicht jedoch biologische *Systeme*. Sie machen Aussagen darüber, wie sich Arten oder Populationen entwickeln, jedoch nicht darüber, wie das einzelne Lebewesen im Innern funktioniert. Systemische Theorien in der Zellbiologie, der Entwicklungsbiologie, der Medizin oder der Psychologie sind weitestgehend unbekannt.

Die Gründe dafür sind vielfältig. Erstens sind lebende Systeme Nichtgleichgewichtssysteme. Dies ist insbesondere seit dem 1944 veröffentlichten Buch von Erwin Schrödinger *What Is Life? The Physical Aspect of the Living Cell* bekannt. Thermodynamisches Gleichgewicht ist nur in lebloser Materie möglich. Lebende Systeme sind offene Systeme, die ständig Masse und Energie aufnehmen, umsetzen und abgeben. Dies erschwert die Formulierung prädiktiver physikalischer Theorien lebender Systeme, denn viel der bekannten Physik basiert auf Gleichgewichtsannahmen. Zweitens sind lebende Systeme oft über viele Längenskalen gekoppelt, vom Molekül zum Organismus, und für eine vorhersagekräftige Beschreibung kann keine der Skalen sinnvoll weggelassen oder vernachlässigt werden. Mathematische

https://doi.org/10.1007/s00287-019-01227-5

Ivo F. Sbalzarini
Technische Universität Dresden, Fakultät Informatik, Institut für Künstliche Intelligenz, Professur für Wissenschaftliches Rechnen in der Systembiologie,
Nöthnitzer Str. 46, 01187 Dresden

Max-Planck-Institut für Molekulare Zellbiologie und Genetik, Zentrum für Systembiologie Dresden,
Pfotenhauerstr. 108, 01307 Dresden
E-Mail: sbalzarini@mpi-cbg.de

und physikalische Modelle hingegen werden meist auf einer bestimmten Skala formuliert, z. B. die Schrödingergleichung auf der atomaren Skala, die Fokker-Planck-Gleichung auf der zellulären Skala und die Theorien der Kontinuumsmechanik auf der Gewebe- und Organskala. Keine dieser Theorien alleine reicht also aus, Prozesse des Lebens zu erklären, und Kopplungen zwischen verschiedenen Theorien sind Gegenstand aktueller Forschung. Drittens haben lebende Systeme sehr viele Freiheitsgrade. Dank Newtons Theorie der Mechanik können wir z. B. die Flugbahn eines Tennisballs oder eines Artilleriegeschosses vorhersagen und optimieren. Wir können aber nicht die Flugbahn eines Vogels erklären oder vorhersagen, denn der Vogel hat ein Gehirn mit Millionen von Nervenzellen, jede mit hunderten von Freiheitsgraden, die zusammenwirken um Entscheidungen über die Flugrichtung zu treffen. Nur wenn wir alle diese inneren Freiheitsgrade modellieren und simulieren könnten, wären wir in der Lage die Flugbahn des Vogels vorherzusagen.

Das Formulieren und Validieren derartiger Modelle setzt jedoch zuerst einmal voraus, dass Messdaten über die Dynamik einer genügend großen Anzahl dieser Freiheitsgrade vorhanden sind. Um die letzte Jahrtausendwende entwickelte sich daher der neue Wissenschaftszweig der Systembiologie. Erstes Ziel der Systembiologie war die Entwicklung neuer Mess- und Experimentierverfahren, welche es erlauben, möglichst viele interne Freiheitsgrade eines biologischen Systems systematisch zu erfassen. Dazu gehörten Methoden zur Sequenzierung vollständiger Genome, insbesondere das Human-Genom-Projekt („Genomik"), Methoden zur Bestimmung aller Proteine in einer Zelle („Proteomik"), Methoden zur Messung aller chemischen Konzentrationen in einer Zelle („Metabolomik"), etc. Ziel dieser sogenannten „-omik"-Ära war es, hinreichend viele Daten zu sammeln, um ein integriertes Bild der Prozesse – vom Genom über die Chemie der Zelle bis zur Mechanik von Geweben – zu erlangen, welche in ihrem Zusammenspiel „Leben" erzeugen.

In Alan Turings berühmtem Artikel *The chemical basis of morphogenesis* (1952) [15] wurde postuliert, dass Leben das Resultat komplexer chemischer Reaktionsnetzwerke mit zehntausenden verschiedener Substanzen ist. In der vergangenen Dekade wurde jedoch klar, dass das Bild nur unter Hinzunahme einer eng und bidirektional an die Chemie gekoppelten Mechanik stimmen kann [7], und heute wird auch die Informationsverarbeitung und die Algorithmik hinzugenommen [12]. Die Systembiologie integriert somit Konzepte der Biophysik und der Informatik mit den Daten der „-omik"-Experimente.

Aus Informatiksicht ist diese Integration insbesondere aus zwei Aspekten spannend: Zum einen können Informatikmethoden diese Integration ermöglichen oder beschleunigen, z. B. Methoden der Datenwissenschaften, der Rechnersimulation, der Bildverarbeitung oder des maschinellen Lernens. Zum anderen kann ein lebendes System, z. B. ein Gewebe oder ein Organ, auch selbst als „Rechner", als informationsverarbeitendes System verstanden werden. In dieser Sichtweise ist jede Zelle eines Gewebes ein „Prozessor", welcher mittels interner Signalverarbeitung Berechnungen ausführt und mit anderen Prozessoren/Zellen kommuniziert. Während uns diese Beschreibung v. a. für Nervensysteme intuitiv ist, weil die Informationsverarbeitung dort über elektrische Signale geschieht, womit eine direkte Analogie zu elektronischen Rechnern besteht, ist sie auch für andere Gewebe anwendbar. Dort findet die Informationsverarbeitung jedoch oft chemisch statt und chemische Reaktionsnetzwerke nehmen die Rolle elektronischer Schaltkreise ein. Auch ist die Kommunikation zwischen Zellen nicht auf elektrische Signale beschränkt, sondern beinhaltet chemische und mechanische Signale sowie direkte Zell-Zell-Kontakte über Membranproteine. Ein Gewebe ist somit ein massiv paralleler Rechner mit Millionen oder Milliarden von Prozessoren, welche über ein topologisch selbstorganisiertes Netzwerk miteinander verbunden sind. Die „Hardware" dieses Rechners sind die Materialien, aus denen die Zellen aufgebaut sind, also Proteine, Lipidmembranen, Nukleotide und andere Moleküle. Die „Software" ist das Genom, d. h. der in der DNS gespeicherte Buchstabencode. Dank der Resultate aus Jahrzehnten biologischer, biochemischer und systembiologischer Forschung wissen wir sehr viel über die Hardware und ihren Aufbau, und wir kennen auch den vollständigen genetischen Code einer exponentiell wachsenden Zahl von Arten. Wir haben jedoch bisher fast kein Verständnis der Algorithmen, welche dieser Code auf dieser Hardware implementiert.

Diese Algorithmen zu entdecken, zu verstehen, wie sie Information verarbeiten und wie sie im Genom codiert sind, ist die große Herausforderung,

vor der die Systembiologie heute steht. Wie treffen Zellen Entscheidungen? Wie weiß eine Zelle in einem Embryo, wann und in welche Richtung sie sich zu teilen hat, wie groß sie wachsen soll, oder wohin sie migrieren soll, sodass im Konzert mit den Millionen oder Milliarden anderer Zellen ein funktionsfähiges Lebewesen entsteht und nicht ein strukturloser Zellklumpen? Könnten wir die Algorithmen entschlüsseln und verstehen, nach denen dies abläuft, und deren Implementierung in molekularen Netzwerken sowie deren Codierung im Genom verstehen, so wären wir in der Lage, Fehlfunktionen dieser Algorithmen zu erkennen und zu beheben, sie umzuprogrammieren oder neu einzustellen. Die Implikationen für Medizin, Landwirtschaft, Pharmakologie und Psychologie wären kaum absehbar, und es wären auch ethische und humanitäre Aspekte zu beachten und neu zu bewerten. Es könnte dann z. B. möglich werden, die Zellen an der Wunde eines abgetrennten Fingers so umzuprogrammieren, dass sie nicht die Wunde schließen und eine Narbe bilden, sondern einen neuen Finger wachsen lassen, so wie es im Embryo auch mal geschah. Auch wird es dann denkbar, Transplantationsorgane im Labor aus körpereigenen Zellen nachzuzüchten, indem diese Zellen umprogrammiert werden und die Kommunikation der fehlenden umliegenden Zellen von einem elektronischen Rechner emuliert wird, wie dies in den cyberbiologischen Systemen im Labor von Prof. Khammash (ETH Zürich) bereits an Einzelzellen funktioniert [10]. Die Möglichkeiten in Diagnostik und Therapie sind schier grenzenlos, und eine Star-Trek-Medizin könnte Realität werden.

Um diese Herausforderung zu meistern, bedarf es neuer physikalischer Theorien, wie z. B. der Ungleichgewichtstheorie der aktiven Materie, neuer Mess- und Beobachtungsmöglichkeiten, wie z. B. KI-gesteuerter automatisierter 3D-Mikroskope und neuer Informatikmethoden. Letzteres beinhaltet Algorithmen zur numerischen Simulation der neuen physikalischen Theorien, Visualisierungs- und Interaktionsmodi für multimodale Daten, Datenbanken zum Ablegen, Strukturieren und Wiederfinden heterogener Daten, sowie Big-Data Analytics zur Auswertung der Daten und zum Entdecken von Modellen und Mechanismen.

Die Datenwissenschaften nehmen dabei eine zentrale Rolle ein. Dies zum einen weil dank „-omik" große Datenmengen vorhanden sind und stets weiter wachsen, zum anderen weil biophysikalische Modelle immer nur Teilsysteme beschreiben, z. B. Blutströmung, Gewebemechanik oder die Kinetik chemischer Reaktionsnetzwerke, die über Daten gekoppelt werden müssen und außerdem typischerweise viele unbekannte Parameter und Koeffizienten haben, welche aus Daten geschätzt werden müssen.

Hier ist ein ganzes Informatikökosystem zu entwickeln, wie in Abb. 1 schematisch abgebildet. Am Anfang stehen große Datenmengen, welche oft in Echtzeit aus Laborgeräten strömen. Nebst den „klassischen" Daten wie Genomsequenzen, Massenspektra von Molekülen oder mechanischen Messungen sind dies heute vermehrt Bilddaten von hochauflösenden 3D-Mikroskopen, z. B. Lichtblattmikroskopen. Diese Mikroskope liefern zeitaufgelöste Videos, z. B. von der Entwicklung eines Embryos, mit subzellulärer Auflösung über Stunden oder gar Tage [8]. Sie erlauben es uns erstmalig, Gewebebildung mit einer Genauigkeit zu beobachten, welche die Entscheidungen und das Verhalten einzelner Zellen erkennen lässt. Die dabei anfallenden Datenmengen und -raten sind jedoch eine Herausforderung. Moderne Mikroskope liefern zwischen 1 und 5 GB/s an Bilddaten. Ein Video der Entwicklung eines Zebrafischembryos über 72 Stunden bis zum Larvenstadium, hat dann 1,3 TB. Diese Datenströme müssen in Echtzeit verarbeitet (z. B. komprimiert, entrauscht, etc.), gespeichert und visualisiert werden.

Da die Daten dreidimensional sind, erfolgt die Visualisierung idealerweise ebenfalls dreidimensional, z. B. mittels Techniken der virtuellen Realität (VR) oder der erweiterten Realität (AR). Mit gängigen WUXGA VR-Systemen erfordert dies 1–2 Gpix/s an Renderingleistung, welche verteilt auf einem Cluster von Grafikkarten erbracht werden muss. Geschieht das in Echtzeit, so wird es möglich, ins Spezimen „einzutauchen" und z. B. in einem sich entwickelnden Embryo spazieren zu gehen und sich umzuschauen, auch im Inneren, wo man bei einer normalen, projektiven Darstellung nicht hineinsieht. Noch besser wird es, wenn die Wissenschaftlerin oder der Wissenschaftler direkt über natürliche Benutzerschnittstellen mit dem Spezimen interagieren kann. Dies könnte z. B. ermöglichen, dass die Bewegung oder das Verhalten einer Zelle durch bloßes Ansehen der Zielzelle (mittels Pupillentracker) aufgezeichnet werden kann, dass Laserschnitte im Gewebe durch Handgesten ausgeführt werden können, oder dass Gene in einzelnen Zellen durch

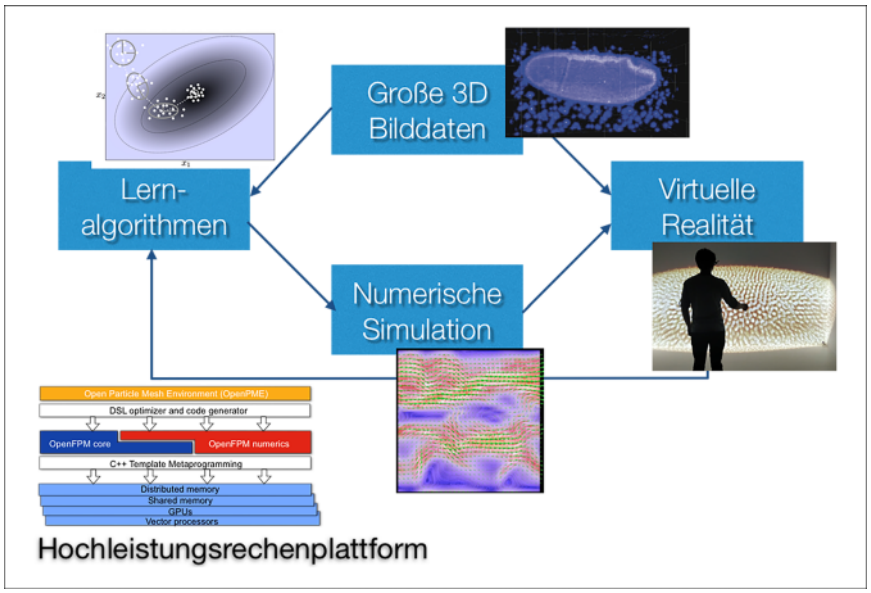

Abb. 1 Schematische Darstellung des Arbeitsablaufs. Hochauflösende 3D Bilder von Mikroskopen zeigen in Echtzeit die Entwicklung eines Embryos (hier im Bild: ein Fruchtfliegenembryo). Diese Bilder werden immersiv in virtueller Realität dargestellt und mit Resultaten numerischer Computersimulationen der vermuteten biologischen und physikalischen Prozesse überlagert, um zu prüfen, ob die Vermutungen hinreichend zur Erklärung des Embryoverhaltens sind. Die Simulationsmodelle werden von Lernalgorithmen aus Nutzerinteraktionen in der virtuellen Realität einerseits und, mittels datengetriebenen maschinellen Lernens direkt auf den Rohbildern andererseits, ermittelt und verbessert. Um Datenraten und Modellkomplexitäten zu erreichen, die zum Verständnis biologischer Systeme notwendig sind, laufen alle Prozesse auf einer Hochleistungsrechenplattform in einer einheitlichen Softwareumgebung

Antippen mit dem Finger in der virtuellen Realität über optogenetische Schalter an- und ausgeschaltet werden können. Durch die direkte Beobachtung der raumzeitlichen Reaktion des Gewebes auf derartige Perturbationen lassen sich dann Rückschlüsse ziehen über die Kommunikation zwischen Zellen und über die Funktion einzelner Zellen oder Gene. Dadurch entsteht eine ganz neue Qualität von Verständnis, und es entstehen im Kopf der Betrachterin oder des Betrachters unweigerlich Hypothesen über die Funktionsweise des „zellulären Rechners".

Diese Hypothesen können dann, in mathematischen Modellen formalisiert, in numerischen Simulationen getestet werden. Das heißt, durch numerische Vorhersage des Gewebeverhaltens, welches aus einem hypothetischen Modell resultieren würde, kann gezeigt werden, ob oder dass das Modell hinreichend ist zur Erklärung des beobachteten Verhaltens. Spannend sind auch immer die Fälle oder die Teilbereiche der Daten, in denen die Simulationsvorhersage nicht passt. Sind Fehlermaße bekannt, so kann dies rechnerisch ermittelt werden. Sind sie nicht bekannt, wie oft in der Entwicklungsbiologie, wo sich auch zwei genetisch identische Embryos leicht unterschiedlich entwickeln, so kann auch die Wissenschaftlerin oder der Wissenschaftler in der virtuellen Realität mittels Handgesten entsprechende Bereiche markieren. Diese Information wird dann einem Lernalgorithmus zur Verfügung gestellt, welcher daraus das Simulationsmodell oder dessen Parameter optimiert.

Mathematische Modelle können aber auch, sozusagen hypothesenfrei, direkt aus Daten gelernt werden. Für raumzeitliche Prozesse könnte man z. B. versuchen, eine mathematische Beschreibung der Dynamik des Prozesses direkt aus Beobachtungsdaten zu lernen. Es wäre dann möglich, die physikalischen Modelle der Strömungsmechanik, die Navier-Stokes-Gleichungen, direkt aus Videoaufnahmen von Flüssigkeitsströmungen zu lernen, ohne physikalische Annahmen oder Verständnis der zugrunde liegenden Prinzipien. Diese Prinzipien könnten dann aber durch Analyse der gelernten mathematischen Modelle entdeckt werden. Dass dies grundsätzlich geht, zeigte eine Arbeitsgruppe um Prof. Nathan Kutz von der University of Washington. Sie entwickelten den PDE-FIND-Algorithmus [13], welcher interpretierbare mathematische Differenti-

algleichungsmodelle dynamischer Prozesse direkt aus raumzeitlichen Beobachtungsdaten lernt. Eine Übertragung dieser Idee auf Mikroskopievideos in den Lebenswissenschaften hätte großes Potenzial, denn die zugrunde liegenden physikalischen Prinzipien und deren mathematische Modelle sind hier oft (noch) unbekannt.

Schließlich läuft das ganze Labor der Zukunft auf einer Hochleistungsrechenplattform, sodass die Resultate der numerischen Simulationen in unter einer Sekunde zur Verfügung stehen, die Datenströme in Echtzeit verarbeitet werden und die Lernalgorithmen in der Schleife mitlaufen. Damit das System auch erweiterbar, flexibel und neuen biologischen und physikalischen Modellhypothesen anpassbar ist, sollte diese Hochleistungsrechenplattform auf einer modularen und skalierbaren Software basieren und idealerweise über eine oder mehrere domänenspezifische Hochsprachen und einen optimierenden Compiler verfügen, um Code-Entwicklungszeiten zu reduzieren. Soweit die Wunschliste.

Die Arbeitsgruppe des Autors am Zentrum für Systembiologie Dresden und an der Technischen Universität Dresden, sowie am Max-Planck-Institut für Molekulare Zellbiologie und Genetik in Dresden und am DFG-Exzellenzcluster „Physik des Lebens“, entwickelt seit über 10 Jahren die theoretischen, algorithmischen und technologischen Grundlagen dieser Vision. Dazu gehören neuartige numerische Simulationsverfahren, mit denen es möglich wurde, die physikalische Theorie der aktiven Materie zu simulieren, um z. B. vorhersagen zu können, zu welcher Form ein Gewebe heranwächst. Dazu gehört auch ein VR/AR-System zur immersiven Echtzeitdarstellung großer 3D-Bild- und Videodaten, inkl. der entsprechenden Open-Source-Software *scenery* (*github.com/scenerygraphics*). Dazu gehört die Entwicklung und Implementierung einer Softwareplattform für skalierbares paralleles Hochleistungsrechnen auf CPUs und GPUs namens *OpenFPM* (*openfpm.mpi-cbg.de*), sowie eine dazu passende domänenspezifische Hochsprache. Diese ermöglicht es, verteilte Anwendungen für numerische Simulation und Datenanalyse in wenigen Stunden oder Tagen zu implementieren, was vorher oft eine gesamte Doktorarbeit von mehreren Jahren in Anspruch nahm.

Diese algorithmischen Fortschritte beruhen auf einer Reihe von theoretischen Vorarbeiten, welche diese Technologie erst ermöglichen. Dazu gehört die Entwicklung einer Theorie zur konsistenten Approximation beliebiger Differenzialoperatoren auf irregulär verteilten Datenpunkten [14], die Verallgemeinerung der klassischen Newton-Interpolation auf hochdimensionale Räume [6], Beweise von Fehlerschranken in der Funktionsapproximationstheorie [3] und die Entwicklung verteilter Graphenzerlegung [1]. Dadurch wurde es möglich, wichtige Probleme in den Datenwissenschaften anzupacken, deren Lösung für die Anwendungen der Lebenswissenschaften essenziell ist. Drei Beispiele seien genannt:

(1) Das Lernen von Modellen aus Daten wird typischerweise als Optimierungsproblem verstanden. So ist es die übliche Herangehensweise im maschinellen Lernen, ein Modell zu trainieren, welches die Trainingsdaten so gut wie möglich wiedergibt. Dazu wird eine Kostenfunktion definiert, welche den Regressionsfehler misst, und die dann minimiert wird, um das Modell oder dessen Parameter zu lernen. Wenn sehr viele Datenpunkte vorhanden sind, so funktioniert das sehr gut, wie z. B. im Deep Learning. In den Lebenswissenschaften sind aber meist nicht Millionen von Datenpunkten verfügbar, sondern „nur“ ein paar hundert bis tausend. Jeder einzelne Datenpunkt ist aber sehr groß. Deep Learning und optimierungsbasierte Verfahren im Allgemeinen neigen in diesen Fällen zu Überanpassung („overfitting“), wobei die Trainingsdaten im Wesentlichen auswendig gelernt werden mit sehr geringer Vorhersagekraft auf neue, im Training ungesehene Daten. Genau dies ist aber Sinn und Zweck der Inferenz. Bereits seit den grundlegenden Arbeiten von Gregor Kjellström in den 1970er-Jahren ist jedoch bekannt, dass es eine alternative Formulierung des Problems gibt, bei der Überanpassung nicht vorkommen kann: Design Centering [9]. Hier wird davon ausgegangen, dass es gerade bei komplexen Systemen und begrenzten Daten viele Modelle geben kann, die akzeptabel gut passen. Was akzeptabel ist, wird z. B. durch die Messfehler in den Daten definiert. Alle Modelle, welche die Daten innerhalb der Messungenauigkeit wiedergeben, sind akzeptabel. Im Allgemeinen ist man frei, beliebige Akzeptanzkriterien zu definieren, welche z. B. auch die Modellkomplexität, die Rechenkosten, o. ä. berücksichtigen. Es muss lediglich gewährleistet sein, dass die Kriterien für ein gegebenes Modell überprüft werden können. Im Raum aller möglichen Modelle bilden die akzeptablen

Modelle einen Unterraum, eine Punktwolke. Design-Centering-Algorithmen lösen dann das Problem, die größte zusammenhängende Punktwolke akzeptabler Modelle in diesem Raum zu finden und deren geometrisches Zentrum zu ermitteln. Das Modell, das diesem Zentrum entspricht, hat größtmögliche Robustheit, d. h. größtmögliche Wahrscheinlichkeit noch immer akzeptabel zu sein, wenn die Daten oder das Modell zufällig variieren. Dieses Verfahren trainiert auch auf begrenzten Datenmengen robuste und generalisierbare Modelle und liefert auch gleich Robustheitsmaße, welche zur Modellselektion oder Risikoabschätzung verwendet werden können. Leider ist Design Centering NP-hart und kann daher auf deterministischen Rechnern nicht effizient gelöst werden. Seit 2017 gibt es jedoch einen effizienten randomisierten Algorithmus, welcher beliebige Design-Centering-Probleme näherungsweise löst [2]. Damit können nun datengetriebene Modelle robust und generalisierbar gelernt werden, ohne Überanpassung und mit geschätzten Robustheiten.

(2) Daten sind nicht gleich Information. Viele Rohdaten in den Lebenswissenschaften sind dünn besetzt, d. h. die pro Byte Daten enthaltene Informationsmenge ist viel kleiner als ein Byte. Dies trifft insbesondere bei Fluoreszenzmikroskopiebildern zu, wo viele Pixel zur Repräsentation des schwarzen Hintergrunds verwendet werden oder ein uniform helles großes Objekt genau gleich hoch aufgelöst wird wie ein dunkles kleines Objekt. Dies ist ein Artefakt der Wahl, digitale Bilder als uniforme Gitter quadratischer Pixel darzustellen. Fasst man ein Bild als mathematische Funktion auf, welche an jedem Punkt im Raum die dortige Farbe oder Helligkeit definiert, so muss diese Funktion aber nicht zwingend auf einem regulären Gitter digitalisiert werden. Die Stützstellen, also die Punkte, an denen der Funktionswert abgetastet und gespeichert wird, können auch „intelligent" verteilt werden, so wie es z. B. der Bildinhalt erfordert. Gegeben ein konkretes Pixelbild, ist also die Frage zu beantworten, wie viele Stützstellen zu dessen Darstellung innerhalb einer vorgegebenen Fehlerschranke notwendig sind, und wo diese platziert werden müssen. Dieses Problem optimal zu lösen, ist sehr rechenintensiv. Schränkt man die möglichen räumlichen Abstände zwischen Stützstellen aber auf Zweierpotenzen ein, so wurde ein sehr schneller und unter dieser Annahme auch exakter Algorithmus gefunden. Es entsteht dabei eine neue, inhaltsadaptive Darstellung raumzeitlicher Daten mit mehr Information pro Dateneinheit, die sogenannte *Adaptive Particle Representation (APR)* [3]. Damit können große Bilddaten 50- bis 5000-mal kompakter dargestellt, gespeichert und verarbeitet werden als auf Pixeln. Direkte Bildverarbeitung auf APR-Bildern und direkte Visualisierung von APR-Bildern ist möglich, sodass nie zu Pixeln zurückgegangen werden muss. Mit *APRnet* ist auch eine neuronale Netzarchitektur in Entwicklung, welche direkt auf APR-Repräsentationen lernt. Die zusätzliche Information über die Positionierung und den Abstand der Stützstellen führt zudem dazu, dass APRnets besser und schneller lernen als pixelbasierte neuronale Netze.

(3) Das direkte Lernen von mathematischen Gleichungsmodellen aus raumzeitlichen Daten ist gerade in den Lebenswissenschaften eine attraktive Perspektive, da physikalisch basierte Modelle oft nicht bekannt sind. So ist es z. B. denkbar, direkt aus einem Mikroskopievideo eines entwicklungsbiologischen Prozesses ein interpretierbares Gleichungsmodell zu lernen, dessen Lösung der beobachteten Gewebedynamik entspricht. Dass dies möglich ist, ist seit PDE-FIND bekannt [13]. Allerdings ist PDE-FIND sehr anfällig auf Rauschen und Messfehler in den Daten und hat algorithmische Parameter, welche manuell eingestellt werden müssen. Je nach Einstellung kommen andere gelernte Modelle heraus. Mittels Verfahren der robusten Statistik ist es jedoch gelungen, einen alternativen Algorithmus zu entwickeln, PDE-STRIDE, welcher mit Messfehlern besser klarkommt und keine einstellbaren Parameter besitzt [11]. Damit war es dann z. B. möglich, die bei der frühen embryonalen Entwicklung des Fadenwurms *C. elegans* beteiligten Proteinnetzwerke und physikalischen Transportprozesse aus Videoaufnahmen automatisch zu lernen, was zuvor Jahre an biologischer Laborarbeit brauchte [4].

Zusammenfassend ist es wohl legitim zu sagen, dass die Datenwissenschaften und insbesondere Big-Data Analytics die Lebenswissenschaften von einer heuristisch-beobachtenden zu einer prädiktiv-mechanistischen Disziplin transformieren. Da die Prozesse und Systeme des Lebens über viele Skalen gekoppelt (d. h. *multi-scale*), nichtlinear und hochdimensional (d. h. viele Freiheitsgrade) sind, werden geschlossene Theorien eine untergeordnete Rolle spielen. Die 1960 von Eugene Wigner im berühmten Artikel *The Unreasonable Effectiveness*

of Mathematics in the Natural Sciences [16] diskutierte Vorhersagekraft mathematischer Theorien in der Physik wird in den Lebenswissenschaften wohl eher von Daten gespielt werden, oder von einer Verbindung aus Daten und Theorie. Es hält sich hier also mit dem 2009 von drei Google-Ingenieuren verfassten Artikel *The Unreasonable Effectiveness of Data* [5]. Um die Effektivität der Daten aber nutzen zu können, sind neue Informatikmethoden notwendig, sowohl in der Theorie und der Algorithmik als auch in der praktischen Softwaretechnologie, der Computergrafik und den Benutzerschnittstellen. Die Systembiologie ist somit zu einem wichtigen Technologietreiber für die Informatik geworden. Aufgrund der Komplexität lebender Systeme und des aktuellen Fortschritts in der Systembiologie wird sich dies in Zukunft wohl noch weiter verstärken.

Literatur

1. Afshar Y, Sbalzarini IF (2016) A parallel distributed-memory particle method enables acquisition-rate segmentation of large fluorescence microscopy images. PLoS One 11(4):e0152528
2. Asmus J, Müller CL, Sbalzarini IF (2017) Lp-Adaptation: Simultaneous design centering and robustness estimation of electronic and biological systems. Sci Rep 7(1):6660
3. Cheeseman BL, Günther U, Gonciarz K, Susik M, Sbalzarini IF (2018) Adaptive particle representation of fluorescence microscopy images. Nat Commun 9:5160
4. Goehring NW, Khuc Trong P, Bois JS, Chowdhury D, Nicola EM, Hyman AA, Grill SW (2011) Polarization of PAR proteins by advective triggering of a pattern-forming system. Science 334(6059):1137–1141
5. Halevy A, Norvig P, Pereira F (2009) The Unreasonable Effectiveness of Data. IEEE Intell Syst 24(2):8–12
6. Hecht M, Hoffmann KB, Cheeseman BL, Sbalzarini IF (2018) Multivariate Newton Interpolation. arXiv:1812.04256
7. Howard J, Grill SW and Bois JS (2011) Turing's next steps: the mechanochemical basis of morphogenesis. Nat Rev Mol Cell Biol 12:392–398.
8. Huisken J (2012) Slicing embryos gently with laser light sheets. Bioessays 34(5):406–411
9. Kjellström G, Taxen L (1981) Stochastic optimization in system design. IEEE Trans Circ Syst 28(7):702–715
10. Lillacci G, Benenson Y, Khammash M (2018) Synthetic control systems for high performance gene expression in mammalian cells. Nucl Acids Res 46(18):9855–9863
11. Maddu S, Cheeseman BL, Sbalzarini IF, Müller CL (2019) Stability selection enables robust learning of partial differential equations from limited noisy data. arXiv:1907.07810
12. Navlakha S and Bar-Joseph Z (2011) Algorithms in nature: the convergence of systems biology and computational thinking. Nat EMBO Mol Syst Biol 7:546
13. Rudy SH, Brunton SL, Proctor JL, Kutz JN (2017) Data-driven discovery of partial differential equations. Sci Adv 3(4):e1602614
14. Schrader B, Reboux S, Sbalzarini IF (2010) Discretization correction of general integral PSE operators in particle methods. J Comput Phys 229:4159–4182
15. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans R Soc London B237:37–72
16. Wigner EP (1960) The Unreasonable Effectiveness of Mathematics in the Natural Sciences. Commun Pure Appl Math 13:1–14