Datengetriebene Forschung – Herausforderung für die Informatik

Auf allen Gebieten der Natur- und Ingenieurwissenschaften gewinnt eine Arbeitsmethode an Bedeutung, bei der die Analyse von sehr großen Mengen an Daten zu neuen Erkenntnissen führt. Mit welchen technischen und organisatorischen Maßnahmen kann die Informatik eine solche datengetriebene Forschung unterstützen?

Von Andreas Reuter

Derzeit vollzieht sich ein grundlegender Wandel in den Natur- und Ingenieurwissenschaften – und das gleich auf mehreren Ebenen. So ändern sich etwa als Folge des Internets und der darauf aufbauenden Dienste die Kommunikationsstrukturen innerhalb und zwischen den Fachgebieten ganz erheblich. Außerdem entwickeln sich neue Organisationsformen für wissenschaftliche Einrichtungen, die der rasch wachsenden Komplexität der Forschungsvorhaben Rechnung tragen. Das wiederum erfordert neue Finanzierungsmodelle für wissenschaftliche (Groß-)Vorhaben. Ferner wandelt sich die Position der Wissenschaft in der Gesellschaft: Ihr wird – zumindest in den westlichen Ländern – sehr viel mehr Transparenz und Rechtfertigung der Ziele und Methoden abverlangt als früher.

Aber auch die wissenschaftliche Methodik selbst befindet sich im Umbruch. In der öffentlichen Wahrnehmung, die sich in Filmen oder Reportagen widerspiegelt, erscheint Wissenschaft immer noch als Tätigkeit, der einzelne (vorzugsweise geniale) Forscher in der Abgeschiedenheit eines Labors nachgehen. Dort kommen sie nach zähem Ringen und diversen Geistesblitzen zu bahnbrechenden Erkenntnissen – oder aber erfinden eine fürchterliche Waffe, je nachdem, ob sie auf der Seite der Guten oder der Schurken stehen.

Dieses romantisierende Bild hat mit moderner Forschung wenig zu tun. Zum einen wird Wissenschaft in immer größeren, komplexeren Projekten und Projektverbünden organisiert – man denke etwa an den Large Hadron Collider (LHC) bei der Europäischen Organisation für Kernforschung CERN in Genf. Solche Vorhaben lassen sich ohne industrielle Methoden und eine hochgradig arbeitsteilige Vorgehensweise nicht bewältigen. Zum anderen stehen Wissenschaftler immer häufiger nur noch in sehr indirektem Kontakt mit den Gegenständen ihrer Untersuchung, seien es Zellen oder Galaxien.

Wenn wir die Entwicklung im Methodenvorrat der Naturwissenschaften einmal Revue passieren lassen (unterer Kasten auf S. 8), so gab es ganz am Anfang die empirische Beschreibung, die sich gelegentlich zu – gleichfalls empirisch abgeleiteten – Handlungsregeln verdichtete. Parallel dazu, aber doch mit einer merklichen Verzögerung, entwickelte sich die Theoriebildung. Ihr Ziel war, formalisierte Modelle der beobachteten Phänomene zu erstellen. Diese sind aber nur um den Preis idealisierender Annahmen möglich – beispielsweise durch Vernachlässigung der Reibung bei der Beschreibung von Bewegungsgesetzen.

Rasch anschwellende Datenflut

Der nächste Schritt bestand darin, den Grad der Idealisierung zu verringern, um auch komplexe Vorgänge wie etwa die Verformung eines Autos beim Aufprall oder die Flugeigenschaften eines Flugzeugs so realistisch beschreiben zu können, dass sich die Ergebnisse der Modellanalyse auf das tatsächliche System übertragen lassen. Dies führte zu sehr komplizierten Modellen, deren Gleichungen nicht mehr direkt lösbar waren. In solchen Fällen bleibt nur die Möglichkeit, mit Methoden der numerischen Mathematik Näherungslösungen zu bestimmen, was bei großen Problemen wie etwa der Crashsimulation im Automobilbau Gleichungssysteme mit Hunderttausenden oder Millionen von Unbekannten ergibt. Deren Handhabung ist überhaupt nur noch mit dem Computer möglich.

Heute gilt die Simulation vielfach als dritte Säule der Wissenschaft – neben Experiment und Theorie. Manche sprechen ihr zwar diesen Rang ab und betrachten sie nur als eine von mehreren möglichen Arten, theoretische Modelle auszuwerten. Tatsache aber ist, dass ohne Simulation viele Modelle »steril« bleiben würden, da es nicht möglich wäre, Ergebnisse daraus abzuleiten.

Von der Simulation führt der Weg schließlich zur datengetriebenen Wissenschaft. Auf den ersten Blick scheint sie nichts grundsätzlich Neues zu bieten; schließlich geht es nur um die Zusammenführung von Experiment (Messung), Theoriebildung und Simulation zu einem kohärenten Methodenvorrat. Das eigentlich Interessante ist jedoch der Grund, der diese Zusammenführung notwendig macht: die rasch wachsende Menge von Daten, die von Messgeräten (wie Satelliten, Teleskopen, Sequenziermaschinen und Microarrays) oder aus Simulationen (etwa Klimavorhersagen und Szenarienanalysen) stammen. Denn die Menge neu erzeugter und gespeicherter Daten verdoppelt sich jedes Jahr, oder anders ausgedrückt: In jedem einzelnen Jahr fallen mehr experimentelle oder Simulationsdaten an als in allen Vorjahren zusammen. Am Anfang sieht eine solche exponentielle Wachstumskurve noch relativ harmlos aus, und tatsächlich konnten Forscher immerhin bis ins 20. Jahrhundert hinein Messergebnisse durch Sichten und Darübernachdenken analysieren.

Mit zunehmender Automatisierung der Messgeräte und dem breiteren Einsatz von Simulationsmethoden sind die Daten in vielen Projekten jedoch schon längst auf einen Umfang angewachsen, der es völlig unmöglich macht, sie im herkömmlichen Sinn direkt in Augenschein zu nehmen. Hierzu nur zwei

Beispiele: Das LHC-Experiment des CERN wird im Vollbetrieb 15 Petabytes (PB) pro Jahr erzeugen, und beim Square Kilometer Array (einem für 2024 geplanten System von Radioteleskopen) soll es sogar 1 PB pro Tag sein. Liegen die Milliarden der aktuellen Finanzkrise schon jenseits der menschlichen Vorstellungskraft, so verhält es sich mit den Petabytes noch eine Million Mal schlimmer (oberer Kasten auf S. 8). Wenn ein Mensch 80 Jahre lang ohne Unterbrechung nichts anderes täte, als sich 1 PB an Ergebnissen »anzusehen«, müsste er pro Sekunde 320 000 Buchstaben (ein Taschenbuch) lesen, um ganz durchzukommen.

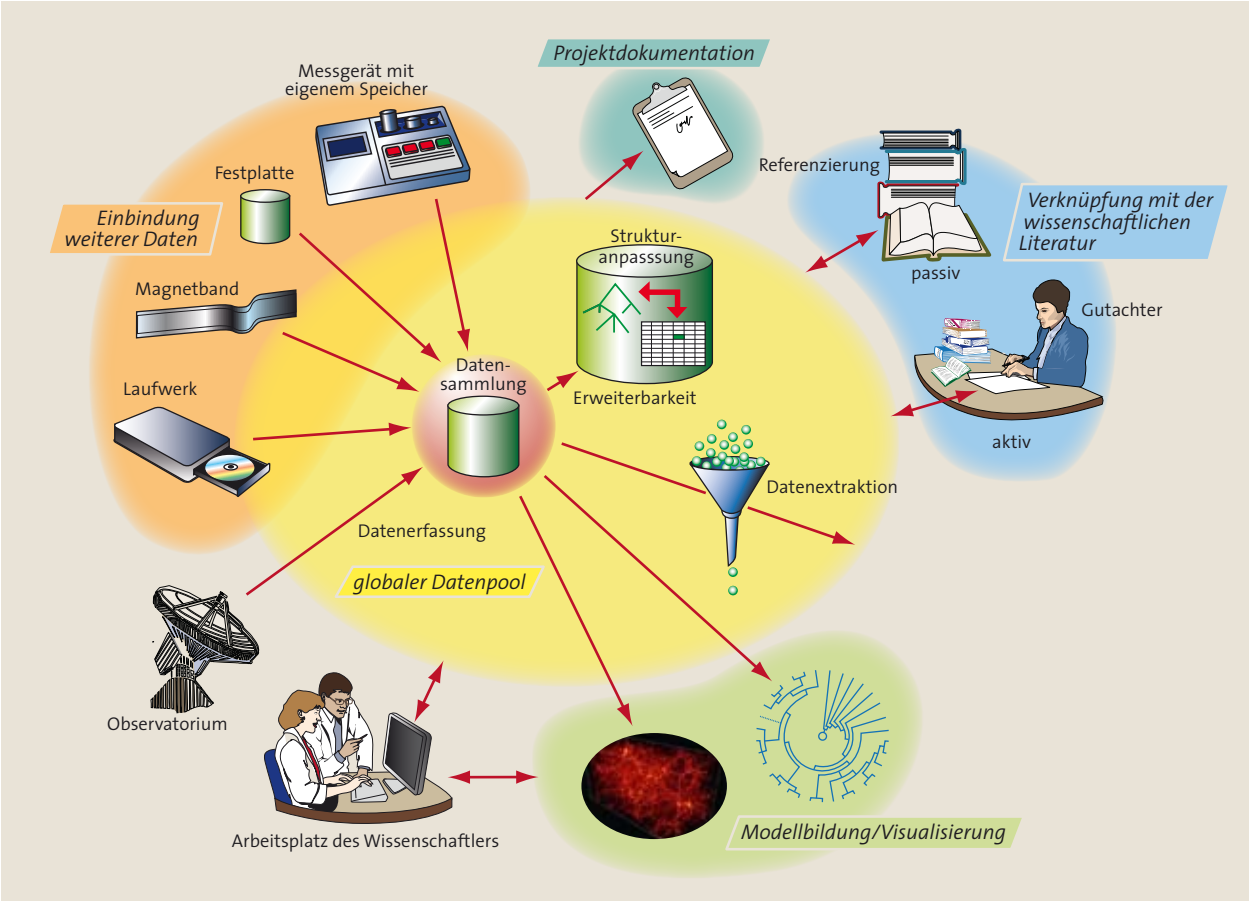
Es bleibt also nichts anderes übrig, als die experimentellen Daten zunächst von Software unterschiedlichster Art aufbereiten zu lassen. Das Datenvolumen muss durch Verdichtung, Selektion, statistische Analyse, Visualisierung und andere Verfahren so weit reduziert werden, dass das Ergebnis für den Menschen wieder aufnehmbar ist. Vor 50 Jahren haben Wissenschaftler noch unmittelbar durch die Teleskope oder Mikroskope geschaut, selbst die Messgeräte abgelesen und die Vorgänge im Reagenzglas beobachtet. Heute kommen sie mit den Experimenten oft erst durch das in Berührung, was auf dem Bildschirm ihres PCs erscheint, nachdem es über viele Stufen hinweg gefiltert, komprimiert und visualisiert worden ist.

Forschung im heutigen Sinn besteht also großteils in der durch Computer, Datenbanken und viele andere Softwarewerkzeuge unterstützten Verarbeitung sehr großer Mengen von Daten, die aus einer Vielzahl von Quellen stammen. Und damit stellt sich die Frage, was die Informationstechnik (IT) dazu beitragen kann und muss. Die offensichtliche Antwort lautet: Ihre Aufgabe ist es, Hardware- und Softwaresysteme zur Verfügung zu stellen, die es den Wissenschaftlern ermöglichen, alle für ihre Fragestellung erforderlichen Auswertungen effizient durchzuführen, ohne sich dabei um IT-spezifische Aspekte kümmern zu müssen. Ein Biologe will schließlich Biologie betreiben und nicht programmieren. Aber was heißt das konkret? In den folgenden Abschnitten skizziere ich die wichtigsten Forderungen an die IT (Übersicht im Kasten unten).

Zunächst müssen die von den Experimenten oder Simulationen kommenden Daten zuverlässig gespeichert werden – und dies unter Umständen mit enormer Geschwindigkeit, wenn man an die oben zitierten Beispiele denkt. Es darf keine Unterbrechungen geben, weil viele Versuche nicht wiederholbar sind. Ferner gilt es, die Daten schon beim Erfassen zu prüfen, zu filtern und für die langfristige

Computergestützter Umgang mit riesigen Datenmengen

Zum Anlegen, Verwalten und Nutzbarmachen eines globalen Datenpools braucht es Software, die vielerlei Anforderungen erfüllen muss. Einige wesentliche sind hier in der Grafik veranschaulicht.



SPEKTRUM DER WISSENSCHAFT / BUSKE-GRAFIK, NACH ANDREAS REUTER

Speicherung aufzubereiten, was weitere hohe Anforderungen an die Leistungsfähigkeit der Hard- und Software stellt.

Das Abspeichern hat dabei so zu erfolgen, dass die Bestände wachsen können, unter Umständen um mehrere Größenordnungen. Außerdem muss jederzeit eine Erweiterung um neue Informationskategorien und Datenstrukturen möglich sein.

Verwandt damit ist die Forderung, Daten aus verschiedenen Projekten und Disziplinen miteinander verknüpfen zu können, um übergreifende Fragen zu untersuchen. Zum Beispiel müssen in der Klimaforschung meteorologische, ozeanografische, geografische, statistische und etliche weitere Datensammlungen zueinander in Beziehung gesetzt werden. Das scheitert heute oft an ihrem unterschiedlichen Aufbau. So verwenden die einzelnen Disziplinen häufig andere Begriffe und Einheiten oder nicht einmal dasselbe Koordinatensystem. Da jedes Fachgebiet zudem seine eigenen Modellierungsmethoden einsetzt, muss es möglich sein, die Daten flexibel in der dafür erforderlichen Struktur bereitzustellen.

Zur Verarbeitung der Rohdaten gehört auch, sie zu verdichten; denn nur in komprimierter Form kann der Forscher die enthaltene Information aufnehmen. Die Software sollte möglichst verschiedene Arten der Verdichtung erlauben, so dass sich im Einzelfall diejenige Methode auswählen lässt, die am besten zu den jeweiligen Daten und Modellen passt. Von besonderer Bedeutung ist dabei die visuelle Darstellung.

Meist müssen Datenbestände für verschiedene Auswertungen immer wieder durchsucht und verarbeitet werden. Wenn sie sehr groß sind, beansprucht das viel Zeit. Die Geschwindigkeit des Zugriffs auf gespeicherte Daten beträgt heute bestenfalls 10^{12} Bytes (1 Terabyte) pro Sekunde; 10 PB zu durchsuchen, dauert somit rund drei Stunden. Um übermäßige Wartezeiten zu vermeiden, sollte man deshalb den Daten Indexstrukturen überstülpen können, die es erlauben, jederzeit gezielt relevante Teilmengen auszuwählen.

Wenn Forschungsarbeiten auf der Auswertung verschiedener Datensammlungen beruhen, ist es zudem unabdingbar, dass die entsprechenden Publikationen eindeutig auf die zu Grunde liegenden Datenbestände verweisen. Dabei müssen Bestände und Software zur Auswertung auch für die Gutachter und andere Leser der Artikel zugänglich sein, weil eine Beurteilung solcher Veröffentlichungen anders nicht möglich ist.

Schließlich ist zu berücksichtigen, dass wissenschaftliche Projekte immer öfter gemeinsam von mehreren Instituten und Arbeitsgruppen durchgeführt werden. Jede Einrichtung erzeugt oder verarbeitet in diesem Fall einen Teil der Daten, wobei andere Kooperationspartner eventuell auf ihre Ergebnisse zugreifen. Da auch Urheberrechte und Fragen der wissenschaftlichen Priorität eine Rolle spielen, muss gewährleistet sein, dass keine Gruppe Daten einer anderen sehen kann, die diese nicht zur gemeinsamen Nutzung freigegeben hat. Eng damit verwandt ist die Forderung, dass alle Interaktionen der Wissenschaftler mit den Datenbeständen – wie Modelldefinitionen, Auswertungen, Veröffentlichungen und so weiter – automatisch zu einer Projektdokumentation zusammengeführt werden.

Allerdings sollen die Schutzvorkehrungen die Zusammenarbeit nicht behindern. Tat-

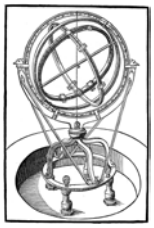
Größenvergleich

1 Petabyte = 10^{15} Bytes = 1 000 000 000 000 000 Bytes
Buch mit 330 Seiten: 1 Million = 10^6 Buchstaben (1 Buchstabe entspricht 1 Byte)
Library of Congress: Rund 31 Millionen Bücher (ohne Handschriften, Fotos und so weiter); 1 PB entspricht also dem Umfang von 10 Millionen Kongressbibliotheken.
Schnelle DSL-Leitung: 50 Mbit/Sekunde ≈ 8×10^6 Bytes/Sekunde
Transfer von 1 PB über diese Leitung: $1{,}25 \times 10^8$ Sekunden ≈ 1448 Tage ≈ 4 Jahre

Entwicklung der wissenschaftlichen Vorgehensweise

Bis vor rund 300 Jahren: Empirie

Wissenschaft beschränkt sich auf die empirische Beschreibung der Naturphänomene. Gelegentlich werden auch (empirisch abgeleitete) Rechenregeln entwickelt, etwa zum Erstellen von Kalendern.


AUS: TYCHO BRAHE, MECHANICA, 1602

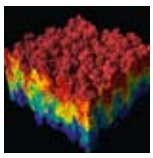
Seit 300 Jahren: Theorie

Forscher gehen dazu über, Naturphänomene zu generalisieren und in Form von (mathematischen) Modellen theoretisch erklärbar zu machen.

$$\left(\frac{\dot{a}}{a}\right) = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

Seit etwa 50 Jahren: Simulation

Naturphänomene wachsender Komplexität lassen sich mit zunehmender Genauigkeit auf Computern simulieren – oft unter Rückgriff auf mathematische Modelle.

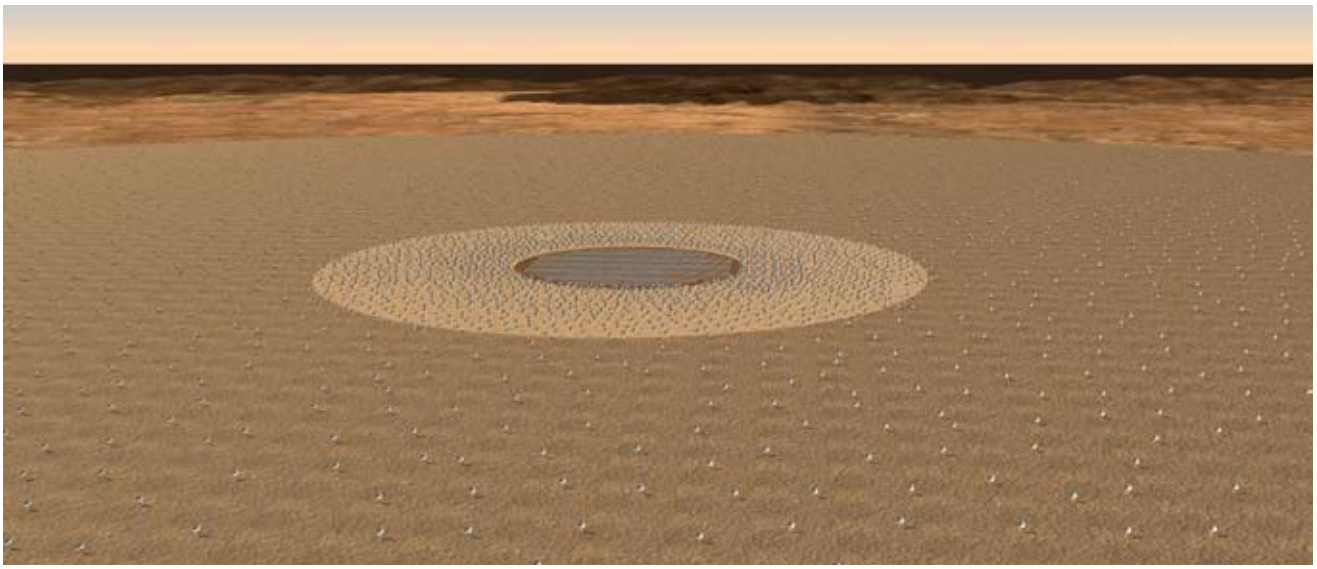

LLNL.GOV

Heute: Datengetriebene Wissenschaft

Experiment, Theoriebildung und Simulation werden zusammengeführt:

- Geräte und Simulationen erzeugen sehr große Mengen von Daten.
- Diese Daten werden durch Software aufbereitet.
- Die Daten und die daraus abgeleiteten Informationen werden in Computern gespeichert.
- Die Wissenschaftler analysieren die Datensammlungen mit Hilfe von Suchverfahren, statistischen Methoden, Visualisierungsverfahren und so weiter.


KLEINES FOTO: ESO, STÉPHANE GUISARD; RECHTS: BESSELFUNCTIONS, CC-BY-2.5



SKA / KILO STUDIOS

Das Square Kilometer Array, ein für 2024 geplantes System von Radioteleskopen, wird 1 Petabyte (PB) an Daten pro Tag liefern. Solch riesige Datenmengen lassen sich nicht mehr ohne äußerst leistungsfähige Computer und ausgefeilte Software auswerten.

sächlich scheuen viele Wissenschaftler immer noch davor zurück, ihre Ergebnisse in eine gemeinsam mit anderen genutzte Datenbank zu stellen, auch wenn es strikte Zugriffskontrollen gibt. Oft schicken dieselben Forscher ihre Daten freilich bedenkenlos per E-Mail an Kollegen, obwohl im Prinzip jeder deren Inhalt während der Übertragung mitlesen kann.

Schließlich muss sichergestellt sein, dass relevante Daten nicht durch Hardwareausfälle oder Bedienfehler verloren gehen können. Viele Fördereinrichtungen für Forschungsprojekte verlangen Mindestaufbewahrungsfristen für alle projektbezogenen Daten und Ergebnisse.

Das Ende pragmatischer Schnellschüsse

Heute werden die genannten Probleme oft in jedem Institut oder für jedes Projekt durch Rückgriff auf etwas halbwegs Brauchbares immer wieder von Neuem gelöst. Diese Ad-hoc-Lösungen sind in der Regel aber so spezifisch, dass sie schon für das nächste Projekt nicht mehr taugen (jedenfalls nicht vollständig). Außerdem legt jedes Labor und jede Projektgruppe eigene Regeln und Konventionen fest. Das macht die Übertragbarkeit der Daten oft schwierig bis unmöglich. Statt pragmatischer Schnellschüsse müssen in Zukunft also generische Lösungen her, die sich für eine große Klasse von Problemen und für unterschiedliche Auswertungsbedürfnisse eignen. Auf sie hinzuwirken, ist auch eine Aufgabe der nationalen und supranationalen Fördereinrichtungen. Anderenfalls wäre eine datenzentrierte Kooperation über verschiedene Disziplinen hinweg zum Scheitern verurteilt.

Im Zusammenhang mit der computergestützten Wissenschaft sind aber nicht nur methodische und Informatikprobleme zu lösen. So erfordert etwa die Möglichkeit zur Integration von Datenbeständen über die Grenzen von Projekten und Disziplinen hinweg die Definition von Standards möglichst großer Reichweite. Außerdem können Zentren zum Verwalten umfangreicher Datenbestände sowie die Hochleistungsrechner zu deren Bearbeitung nicht an jedem Institut oder auch nur an jeder Universität eingerichtet werden – das wäre viel zu teuer. Sinnvoll ist eine hierarchische Organisation mit wenigen Supercomputerzentren an der Spitze, einigen »großen« Zentren darunter und vielen Institutsservern auf der dritten Stufe.

Der Aufbau solcher nationalen oder besser noch internationalen Kooperationsstrukturen ist naturgemäß auch ein politisches Thema, in das Standortpräferenzen und Prestigefragen hineinspielen. Immerhin laufen bereits die erforderlichen Abstimmungsprozesse in Deutschland, Europa, den USA, Australien oder China. Das nächste Ziel für die Spitze der Hierarchie ist jedenfalls schon definiert: ein Rechner, der rund 1000-mal so schnell arbeitet wie der heutige Rekordhalter, also eine Leistung im Bereich von Exaflops (10^{18} Rechenoperationen pro Sekunde) erbringt.

Die Informationstechnologie hat somit eine ganze Reihe von Problemen zu lösen, um der modernen, datengetriebenen Wissenschaft gerecht zu werden –, und eines der schwierigsten, die Parallelverarbeitung auf Millionen von Rechenknoten, habe ich nicht einmal angesprochen. Wichtig ist, dass die Werkzeugentwicklung auf Seiten der Informatik Hand in Hand mit Methodenentwicklung auf Seiten der Wissenschaft geht. Denn nur so funktioniert jenes Wechselspiel, das seit jeher Triebfeder des wissenschaftlichen Fortschritts war: Neue Methoden stellen neue Anforderungen, und neue technische Möglichkeiten eröffnen den Weg zu neuen Methoden.

DER AUTOR

Andreas Reuter ist Professor für Informatik an der Universität Heidelberg und Geschäftsführer des Heidelberger Instituts für Theoretische Studien (HITS).

QUELLEN

Bell, G. et al.: Petascale Computational Systems: Balanced Cyber-Infrastructure in a Data-Centric World. Letter to NSF Cyberinfrastructure Directorate. In: IEEE Computer 39, S. 110–112, 2006

Hey, T. et al.: The Fourth Paradigm – Data-Intensive Scientific Discovery. Microsoft Corporation, 2009