Große Daten – große Herausforderungen

Physikalische Großprojekte liefern riesige Datenmengen und -raten. Diese zu meistern, ist eine Herausforderung im Bereich Big Data.

Andreas Heiss

Big Data – ist das nicht nur ein Hype um Dinge, die wir längst beherrschen? Dies mögen sich viele Physikerinnen und Physiker fragen und dabei an die riesigen Datenmengen der LHC-Experimente denken. Der Begriff „Big Data" bezeichnet allgemein den Umgang mit großen Datenmengen, Datenraten oder besonders komplizierten bzw. unstrukturierten Daten. In der Wirtschaft geht es dabei beispielsweise darum, die Kundendaten eines Online-Händlers mit Daten aus sozialen Netzwerken zu verknüpfen, um daraus wertvolle Informationen für das Marketing zu gewinnen. In der Wissenschaft ist es oft schon schwierig, die stetig steigenden Datenmengen und -raten an sich zu verarbeiten.

Noch vor wenigen Jahren war es undenkbar, die von heutigen Experimenten und Apparten erzeugten Datenmengen zu beherrschen. So können bei der Genom-Sequenzierung, bei Hochdurchsatzmikroskopen oder bei Hochgeschwindigkeitskameras viele Terabytes an Daten pro Tag entstehen. Mit den noch vor wenigen Jahren üblichen „Hausmitteln" wie USB-Festplatte, Desktop-PC oder Laptop können die Nutzer dieser Instrumente die anfallenden Datenmengen nicht mehr speichern und verarbeiten. Stattdessen sind großskalige Datenmanagement- und Computing-Systeme, wohldurchdachte Workflows und komplexe Software notwendig.

Eine weitere Herausforderung besteht darin, diese Flut an wissenschaftlichen Daten lange Zeit zu archivieren. Denn unter anderem atmosphärische oder geologische Messungen lassen sich nicht einfach wiederholen, falls man die gemessenen Daten verloren hat oder sie nicht mehr lesen kann.

Bei beiden Aspekten gibt es in der Physik noch viel zu tun. Trotz langjähriger Erfahrung im Umgang mit riesigen Datenmengen ist es wichtig, die Computing-Anforderungen der kommenden Experimente ernstzunehmen. In einigen Jahren wird der High-Luminosity-LHC jährlich eine Datenmenge von annähernd einem Exabyte liefern, und beim Square Kilometre Array werden mehr als 100 Terabytes pro Sekunde anfallen. Diese Daten gilt es, vor Ort und quasi online auf einem High-Performance-Computing-System zu prozessieren, bevor der daraus resultierende Datenstrom zur weiteren Verarbeitung zu Datenzentren weltweit geleitet wird. Ohne weitere Optimierung von Computing-Modellen, Algorithmen und Software dürfte es voraussichtlich nicht möglich sein, diese speziellen Anforderungen zu erfüllen oder überhaupt zu finanzieren.

Es wäre fatal, wenn die Wissenschaft es im Konkurrenzkampf mit der freien Wirtschaft nicht schaffen würde, Big-Data-Experten zu binden.

Die Wissenschafts-Ministerien und Forschungsorganisationen in Deutschland sowie die EU haben die Herausforderungen erkannt und fördern Big-Data-Projekte und -Infrastrukturen – insbesondere solche, die auf eine disziplinübergreifende Zusammenarbeit und Nutzung von IT-Ressourcen wie Datenspeicher und Rechner abzielen. Beispiele sind das European Open Science Cloud-Projekt oder die Helmholtz Data Federation.

Gerade in der Physik gibt es viele junge Kolleginnen und Kollegen, die das notwendige Wissen und Interesse haben, in solchen Projekten an der Grenze zwischen Physik und Informatik zu arbeiten und die anstehenden Herausforderungen anzupacken. Und doch fehlt es weitgehend an attraktiven Karrieremöglichkeiten in diesem Arbeitsgebiet.

Meinung von **Dr. Andreas Heiss**, Leiter der Abteilung Scientific Data Management am Steinbuch Centre for Computing des Karlsruher Instituts für Technologie

Für diese computeraffinen Physikerinnen und Physiker ist es schwierig, eine klassische Wissenschaftskarriere in der Physik oder der Informatik erfolgreich zu durchlaufen. Ihre Forschung ist für die Physik meist zu wenig physikalisch und für die Informatik zu angewandt oder zu spezifisch. Häufig ist es bereits schwierig, entsprechende Forschungsarbeiten zu publizieren, da es kaum passende und etablierte Fachzeitschriften und Konferenzen gibt.

Viele Big-Data-Talente gehen daher lieber in die Industrie, als auf eine der wenigen dauerhaften Mittelbaustellen an Universitäten oder bei den Forschungsorganisationen zu hoffen. Große Unternehmen und verschiedenste Start-Ups suchen derzeit händeringend nach Big-Data-Experten, Data Scientists oder Entwicklern und bieten interessante Jobs, gerade auch für Physikerinnen und Physiker an. Das Wissen, das diese Experten in die Industrie mitnehmen, geht der Wissenschaft aber verloren.

Es wäre fatal, wenn die Wissenschaft es im Konkurrenzkampf mit der freien Wirtschaft nicht schaffen würde, Big-Data-Experten zu binden. Das ist vielleicht sogar eine der größten Herausforderungen in der Physik!