GFZ Helmholtz-Zentrum Potsdam

Aus: System Erde. GFZ-Journal (2018) 8, 1
„Data Science in den Geowissenschaften“
Alle Artikel verfügbar im Internet: http://systemerde.gfz-potsdam.de

Helmholtz-Zentrum Potsdam
Deutsches GeoForschungsZentrum

Data Science in den Geowissenschaften – eine Einführung

Doris Dransch
Deutsches GeoForschungsZentrum, Potsdam

Data Science aims at extracting information and knowledge from data. It comprises methods, processes, algorithms and technologies for management, processing and analysis of data. It is an interdisciplinary field applying concepts and techniques from mathematics, statistics, information technology, and application domains such as geosciences. Data Science is gaining more and more importance for science. It complements the traditional scientific approaches of empirical, theoretical and computational research with explorative strategies; developing reasoning-based hypotheses is complemented by suggesting correlations and complex interrelationships based on data alone. Data Science can also help to exploit the information contained in large data sets based on existing understanding of the processes and known gaps in our understanding. In this way data science may lead to novel, unexpected and valuable knowledge of complex natural or man-made systems.

System Erde (2018) 8,1 | DOI: 10.2312/GFZ.syserde.08.01.1

Data Science, Datenwissenschaft, befasst sich mit der Extraktion von Information und Wissen aus Daten. Sie umfasst Methoden, Prozesse, Algorithmen und Technologien zu Management, Verarbeitung, Aufbereitung und Analyse von Daten. Data Science ist ein interdisziplinäres Wissenschaftsfeld, das Konzepte und Techniken aus den Fächern Mathematik, Statistik, Informationstechnologie sowie Anwendungsdisziplinen, wie z. B. Geowissenschaften, verwendet. Der Begriff Data Science ist nicht neu, er wurde vor allem in der Statistik mit dem Aufkommen der computergestützten Datenverarbeitung und Datenanalyse verwendet. Eigene wissenschaftliche Schriftenreihen entstanden wie *Data Science Journal* und *The Journal of Data Science*. Data Science geht inzwischen über die Statistik hinaus und umfasst einen viel breiteren Ansatz zur Extraktion von Wissen aus Daten. Dazu gehören Aufbau von Datenbanken und Dateninfrastrukturen, Datenintegration, Datenanalytik und Datenexploration, Informationsvisualisierung sowie die Entwicklung von Informationstechnologien für die Verarbeitung von großen Datenmengen – „Big Data".

Neuartige technologische Entwicklungen in den Bereichen analytische Datenbanken, Analysemethoden, Cloud Computing, paralleles Rechnen und Software-as-a-Service schufen die Basis für die breite Anwendung von Data Science. Sie befassen sich mit den besonderen Herausforderungen an die Verarbeitung, Aufbereitung und Analyse von großen, komplexen Datenmengen. Vor allem Entwicklungen im Bereich Data Mining und maschinellem Lernen eröffnen neue Wege der Informationsextraktion aus Daten. Sie ermöglichen es, Relationen und Muster zu extrahieren und damit Schlüsse aus Daten zu ziehen, die über die Datenanalyse mit Tabellenkalkulationsprogrammen weit hinausgehen.

Data Science steht in enger Wechselbeziehung zur digitalen Transformation, die derzeit viele Lebensbereiche in unserer modernen Gesellschaft erfasst. Die digitale Transformation bezeichnet „Veränderungen des Alltagslebens, der Wirtschaft und der Gesellschaft durch Verwendung digitaler Technologien und Techniken sowie deren Auswirkungen" (Enzyklopädie der Wirtschaftsinformatik, http://www.enzyklopaedie-der-wirtschaftsinformatik.de). Der breite Einsatz digitaler Technologien ist Basis für die systematische Sammlung und Auswertung von Daten, umgekehrt liefert Data Science die Voraussetzung und Werkzeuge dafür, Informationen und Wissen aus den gewonnenen Daten zu extrahieren. Data Science geriet durch das Thema Big Data zunehmend auch in den Fokus der Wirtschaft. Die Verknüpfung von Daten und ihre umfassende Analyse ermöglichten es, vielfältige Information, z. B. über das Kaufverhalten von Kunden, zu gewinnen. Unter den Begriffen „Business Intelligence" und „Business Analytics" etablierten Unternehmen Verfahren und Prozesse zur systematischen Sammlung, Auswertung und Darstellung von Geschäftsdaten, um operative und strategische Entscheidungen zu unterstützen (https://en.wikibooks.org/wiki/Data_Science:_An_Introduction; https://de.wikipedia.org/wiki/Data_Science).

Data Science in der Wissenschaft

Data Science und Digitalisierung spielen ebenfalls in der Wissenschaft und damit auch in den Geowissenschaften eine immer bedeutendere Rolle (*Bell et al., 2009*). Der mit dem renommierten Turing Award ausgezeichnete Informatiker Jim Gray beschreibt diese Entwicklung in seinen vielzitierten Ausführungen zu „eScience – data intensive science" als viertes Forschungsparadigma, das die bisherigen Forschungskonzepte – die empirische Forschung, Theoriebildung und Simulation – ergänzt (*Hey et al., 2009*). Mit Data-Science-Ansätzen werden auf der Basis großer Datenmengen Muster und Relationen extrahiert und somit Informationen gewonnen, die zu neuen, möglicherweise unerwarteten, Erkenntnissen über komplexe Systeme führen. Jim Gray weist eindrücklich darauf hin, dass die wissenschaftliche Erkenntnisgewinnung aus den immensen Datenmengen, die durch immer mehr Sensorik und verbesserte Simulationsmodelle erzeugt werden, neue Konzepte und Strategien erfordert, die auf digitalen Technologien und Verfahren basieren. Als wesentlich für Data Science nennt er ein verbessertes Management von Daten, das Zusammenführen der Daten aus verschiedenen Quellen, die explorative Analyse der vielfältigen verfügbaren Daten, sowie die digitale Unterstützung des gesamten wissenschaftlichen Arbeitsprozesses: vom Finden geeigneter Daten, deren Verknüpfung/Integration und systematischen Analyse bis zur Publikation von Ergebnissen. Data Science ist somit eng verbunden mit der digitalen Transformation in den Wissenschaften.

Die Wissenschaft hat diese Herausforderungen aufgegriffen und in beispielhaften Lösungen umgesetzt. Vielfältige Initiativen zu Forschungsdateninfrastrukturen, wie die europäischen Projekte ESFRI, EUDAT und EPOS, oder Initiativen in der Seismologie (ORFEUS) oder Geodäsie (EUREF, IGS), wurden ins Leben gerufen, um Daten besser auffindbar, zugreifbar und wiederverwendbar zu machen. Die Verknüpfung von Daten aus unterschiedlichen Erfassungssystemen über die verschiedenen räumlichen und zeitlichen Skalen hinweg ist Thema mehrerer Forschungsvorhaben. Ein Beispiel ist das Projekt Digital Earth, an dem alle Helmholtz-Zentren des Forschungsbereichs „Erde und Umwelt" mitwirken. Verfahren aus dem maschinellen Lernen, Data Mining und Visual Analytics werden angewendet und erprobt, um die explorative Auswertung wissenschaftlicher Daten zu ermöglichen und damit Prozesse und Zusammenhänge, z. B. des Systems Erde, besser zu verstehen. Auch erfolgte die Entwicklung wissenschaftlicher Workflow-Systeme, wie Kepler oder Taverna, um den wissenschaftlichen Arbeitsprozess zu unterstützen, zu dokumentieren und zu reproduzieren. Die Helmholtz-Gemeinschaft trägt diesen Entwicklungen ebenfalls Rechnung durch Projekte wie Digital Earth oder den „Information and Data Science Inkubator", die die Konzepte, Methoden und Technologien von Data Science auf breiter Basis in die Wissenschaft tragen wollen.

Kontakt: D. Dransch
(doris.dransch@gfz-potsdam.de)

Data-Science-Entwicklungen am GFZ

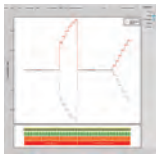
Am Deutschen GeoForschungsZentrum GFZ gibt es verschiedene Forschungs- und Entwicklungsaktivitäten, um das Potenzial von Data Science in Wert zu setzen und damit den geowissenschaftlichen Erkenntnisgewinn zu verbessern.

Erprobung und Weiterentwicklung von maschinellen Lernverfahren

Bei der Erforschung von Naturgefahren werden am GFZ Verfahren des maschinellen Lernens genutzt und erprobt, um das Verständnis von Prozessen, wie z. B. Erdbeben, Hochwasser oder Weltraumwetter, und ihre Auswirkungen zu verbessern. So werden maschinelle Lernverfahren eingesetzt, um Erdbebenereignisse zu klassifizieren, und es wird untersucht, ob mit Hilfe dieser Verfahren die Vorhersage des seismischen Verhaltens bestimmter Regionen verbessert werden kann. Im Bereich der Hochwasserforschung findet maschinelles Lernen Anwendung, um die komplexen Schädigungsprozesse bei Hochwasser zu analysieren und die damit verbundenen Prozesse besser zu verstehen, zu beschreiben und für Prognosen zu modellieren. Auch bei der Vorhersage des Weltraumwetters, das Auswirkungen auf Stromnetze und Satelliten hat, kommt maschinelles Lernen zum Einsatz. Mit Hilfe dieser Methoden wurden ein globales Vorhersagemodell der Plasmasphärendynamik erzeugt und ein Verfahren für die Vorhersage potenzieller geomagnetischer Stürme entwickelt (siehe Beitrag von Kreibich et al., S.10 in diesem Heft).

Maschinelles Lernen ist auch für die Auswertung von Bilddaten aus Satellitenmissionen eine wesentliche Methode. Ziel der Datenauswertung ist es, Zustand und Veränderung der Landoberfläche der Erde aus den Satellitenbildern noch präziser zu ermitteln. Die Zusammenhänge zwischen Eigenschaften der Satellitenbilder und Parametern der Landoberfläche sind oft nur sehr schwer mit Hilfe von deterministischen Modellen zu beschreiben. Maschinelle Lernverfahren hingegen können beliebige Daten verknüpfen, um daraus effizient Zusammenhänge, Rückschlüsse und Vorhersagen abzuleiten. Am GFZ werden auf der Basis geowissenschaftlicher Expertise verschiedene Verfahren des maschinellen Lernens für die Auswertung von Satellitenbilddaten genutzt und weiterentwickelt. Einsatzfelder sind beispielsweise die Atmosphärenkorrektur, die Analyse von Bodeneigenschaften und die Schätzung der Photosyntheseleistung von landwirtschaftlichen Nutzflächen (siehe Beitrag von Segl et al., S.18 in diesem Heft).

Entwicklung von Visual-Analytics-Konzepten und -Werkzeugen



Eine Voraussetzung für die Extraktion von Information aus Daten ist, dass Wissenschaftlerinnen und Wissenschaftler die Daten verstehen können. Mit zunehmender Größe und Komplexität der Daten wird dies eine stetig wachsende Herausforderung. Damit Wissenschaftlerinnen und Wissenschaftler sich „ein Bild“ von den Daten machen können, entwickeln Experten aus der Informatik und aus den Geowissenschaften am GFZ in enger Kooperation Visual-Analytics-Verfahren und -Werkzeuge, die Methoden aus Statistik, maschinellem Lernen und Data Mining mit interaktiver Visualisierung verknüpfen. Damit lassen sich umfangreiche und komplexe Datenbestände explorieren sowie Strukturen und Muster erkennen. Beispiele hierfür sind Verfahren für (1) die Analyse von Geoarchiven, wie z. B. Seesedimente, um Klima- und Landschaftsentwicklungen der Vergangenheit aufzuklären, (2) die Analyse von komplexen Simulationen, bei denen viele Größen miteinander interagieren, wie z. B. geochemische Simulationsmodelle in der Fluidsystemmodellierung, und (3) die Validierung von Simulationsmodellen, wie z. B. der glazial-isostatischen Anpassungen in der Erdsystemmodellierung (siehe Beitrag von Unger et al., S. 26 in diesem Heft).

Lösungen für Big Data

Big Data und die damit verbundenen Anforderungen an die Datenverarbeitung und Datenanalyse rücken aufgrund des Fortschritts in der Sensorentwicklung und der Computersimulation auch in den Fokus der Geowissenschaften. Big Data werden generell durch „5 Vs“ charakterisiert: Volume (Datenmenge), Variety (Vielfalt der Daten), Velocity (Geschwindigkeit der Erzeugung), Veracity (Zuverlässigkeit) und Value (Wert). Auch wenn die 5 Vs Herausforderungen in vielen Fachgebieten der Geowissenschaften mit sich bringen, sind vor allem diejenigen Disziplinen mit Big Data konfrontiert, die mittels kontinuierlicher Observationssysteme Daten erheben und verarbeiten. Zu nennen sind hier stellvertretend die Fachgebiete Geodäsie, Seismologie und Geomagnetismus. Sie bewegen sich bei der Erfassung, Verarbeitung und Analyse der Daten seit jeher an der Grenze des technisch und wissenschaftlich Machbaren. Am GFZ werden im Rahmen der geodätischen Satellitenmission GRACE-FO, der Globalen Navigationssatelliten-Systeme (GNSS), der Very Long Baseline Interferometry (VLBI), des globalen seismologischen Netzwerks und Datenzentrums GEOFON sowie der Erdmagnetfeld-bezogenen Satellitenmission Swarm und den damit verknüpften Datenzentren Lösungen für die Erfassung, Verarbeitung und Analyse von Big Data in den Geowissenschaften entwickelt (siehe Beitrag von Schuh et al., S. 32 in diesem Heft).

Die Geowissenschaften können von Konzepten und Technologien, die in der Informatik für die Verarbeitung und Analyse von Big Data entwickelt wurden, profitieren. Dies betrifft die Speicherung und Prozessierung von Daten wie auch die Extraktion relevanter Informationen und Muster aus großen Datenmengen. Dazu müssen jedoch die Informatik-Technologien und -Konzepte in die Geowissenschaften übertragen und entsprechend angepasst werden. Beispiele dieser am GFZ durchgeführten Adaptation sind die Entwicklung effizienter, skalierbarer geowissenschaftlicher Analysemethoden für Zeitreihen oder die Anpassung geowissenschaftlicher Methoden aus dem Bereich der Satellitenfernerkundung an existierende Big-Data-Technologien (siehe Beitrag von Sips et al., S. 40 in diesem Heft).

System Erde (2018) 8,1 | DOI: 10.2312/GFZ.syserde.08.01.1

Nutzung von Citizen-Science-Ansätzen

Neben den traditionellen Verfahren der Datenerhebung steht den

Geowissenschaften mit zunehmender Digitalisierung der Gesellschaft eine weitere Datenquelle zur Verfügung, nämlich Daten, die von Bürgerinnen und Bürgern erhoben werden. Die Beteiligung von interessierten Laien, die sich an wissenschaftlichen Beobachtungen beteiligen bzw. diese unterstützen, hat eine lange Tradition, z. B. bei der Vogelzählung oder Wetteraufzeichnung. Die wissenschaftliche Nutzung der so erhobenen Daten nimmt durch die neuen Möglichkeiten der digitalen Technologien weiter zu. Am GFZ werden Verfahren entwickelt und erprobt, wie geeignete Daten mittels Bürgerbeteiligung – Citizen Science – für die Quantifizierung von nächtlicher Lichtemission, die schnelle Schadensabschätzung bei Flutereignissen oder die dynamische Modellierung der Gefährdung von Gebäuden bezüglich Erdbeben erhoben und genutzt werden können (siehe Beitrag von Dransch et al., S. 46 in diesem Heft).

Etablierung eines Forschungsdatenmanagements nach dem FAIR-Prinzip

Data Science kann ihr Potenzial nur entfalten, wenn Forschungsdaten aufbereitet und mit Metadaten beschrieben sind und wenn es einen einfachen Zugriff für Mensch und Maschine auf die Daten gibt.

Das Prinzip von FAIR Data ist daher zur allgemein anerkannten Basis für das Forschungsdatenmanagement geworden. FAIR umfasst die Auffindbarkeit (Findable), Zugänglichkeit (Accessible), Interoperabilität (Interoperable) und Wiederverwendbarkeit (Reusable) von Daten. Das GFZ hat eine lange Tradition darin, Forschungsdaten öffentlich bereitzustellen. Beispielsweise wurde mit dem Suchwerkzeug RI@GFZ ein einfacher „Daten-Wegweiser" bereitgestellt. Die Vergabe von DOIs (Digital Object Identifier) erlaubt die eindeutige Identifizierung von geowissenschaftlichen Daten und Softwareprodukten. Die Beschreibung der Daten durch standardisierte Metadaten ermöglicht deren Wiederverwendbarkeit. Das Forschungsdatenmanagement am GFZ wird in enger Anbindung an Aktivitäten der Helmholtz-Gemeinschaft und internationaler Initiativen kontinuierlich weiterentwickelt (siehe Beitrag von Elger et al., S. 52 in diesem Heft).

Ausbildung im Bereich Data Science

Wissenschaftlerinnen und Wissenschaftler müssen zusätzliche Expertisen und Kompetenzen erwerben, um die neuen Data-Science-Konzepte, -Methoden und -Technologien für ihre Arbeit nutzen zu können. Das GFZ hat daher für die Ausbildung des wissenschaftlichen Nachwuchses mit Partnern des Geo.X-Forschungsverbunds die Geo.X Young Academy mit dem Schwerpunkt Geo Data Science ins Leben gerufen (siehe hierzu auch das Interview ab Seite 60 in diesem Heft). Auch beteiligt sich das GFZ an der „Helmholtz Einstein International Berlin Research School in Data Science" (HEIBRiDS), einem Verbund von sechs Helmholtz-Zentren und dem Einstein Center for Digitial Future. Geplant ist ergänzend die Einrichtung der „Helmholtz Information and Data Science Academy" (HIDA), die für alle in der Wissenschaft Tätigen Fortbildung im Bereich Data Science anbieten soll.

Auswirkung von Data Science auf die Geowissenschaften

Data Science verändert das wissenschaftliche Arbeiten, neue Methoden und digitale Technologien modifizieren die Forschung in mehrfacher Weise. Zum einen werden die traditionellen wissenschaftlichen Konzepte ergänzt. Zusätzlich zur Wissensgenerierung durch die Bildung und Überprüfung von Hypothesen und die Entwicklung von Simulationsmodellen auf Basis vorhandenen Wissens und physikalischer Gesetze, können explorative Verfahren wie Data Mining und maschinelles Lernen genutzt werden, um auf der Grundlage vieler existierender Daten Muster und Bezüge aus Daten zu extrahieren und Schlüsse für gesellschaftlich relevante Fragestellungen daraus zu ziehen. Auch der wissenschaftliche Arbeitsprozess erfährt eine Veränderung. Der Einsatz digitaler Verfahren bei der Generierung, Verwaltung, Publikation und Bereitstellung von Daten und Erkenntnissen erfordert eine Modifizierung bisheriger Arbeitsweisen. Beispielsweise müssen Daten aufbereitet und mit Metadaten beschrieben werden, damit sie als Quelle sowohl von Personen als auch von Algorithmen für Data Science genutzt werden können. Ein weiteres Beispiel ist die Dokumentation wissenschaftlicher Verfahren und Arbeitsprozesse mit Hilfe wissenschaftlicher Workflow-Systeme, welche die logische, aber auch technische Reproduzierbarkeit wissenschaftlicher Analysen ermöglichen.

Die zunehmende Integration digitaler Verfahren in den wissenschaftlichen Arbeitsprozess erfordert von den Wissenschaftlerinnen und Wissenschaftlern zusätzliche Expertise und zusätzlichen Zeitaufwand. Dieses muss zukünftig in den wissenschaftlichen Bewertungssystemen Berücksichtigung finden. Aktivitäten, die den Einsatz digitaler Verfahren befördern, sollten als wissenschaftliche Leistung anerkannt und entsprechend bewertet werden. Nur durch einen breiten Einsatz digitaler Verfahren kann Data Science ihr gesamtes Potenzial entfalten.

Data Science hat einen mehrdimensionalen Transformationsprozess in den Geowissenschaften angestoßen. Diesen zu gestalten und zu beleben wird in den Geowissenschaften eine Herausforderung für die nächsten Jahre sein. Das GFZ ist mit vielfältigen Aktivitäten daran beteiligt, wie die Beiträge in dieser Ausgabe des GFZ-Journals „System Erde" zeigen.

Literatur

Bell, G., Hey, T., Szalay, A. (2009): Beyond the Data Deluge. - Science, 323, 5919, pp. 1297—1298. DOI: https://doi.org/10.1126/science.1170411

Hey, T., Tansley, S., Tolle, K. (Eds.) (2009): Jim Gray on eScience: a transformed scientific method. – In: Hey, T., Tansley, S., Tolle, K. (Eds.), The fourth paradigm : data-intensive scientific discovery, Redmond [u. a.] : Microsoft Research, pp. xvii—xxxi, verfügbar unter https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/ [letzter Zugriff: 13.08.2018]