Computer-Supported Research

Konzeptpapier zum CoSRe-Projekt

Clemens Beckstein, Christian Knüpfer, Birgitta König-Ries

14. April 2011

Motivation

Die Wissenschaften befinden sich am Beginn zu einem neuen Zeitalter: neue Messverfahren erzeugen riesige Mengen experimenteller Daten, Computer gestatten die Exploration dieser Daten, immer größere und komplexere Modelle erlauben präzise Vorhersagen, das Internet führt zu einer weltweiten Kollaboration, der Innovationszyklus wird immer schneller. Jim Gray bezeichnet dies als „vierten wissenschaftlichen Paradigma“[1]

Diese Datengetriebenheit stellt die Wissenschaften aber auch vor neue Herausforderungen. Zwar werden immer mehr Routineaufgaben von Computern übernommen. Für die Bewältigung der Informationsflut ist aber eine Computer-Unterstützung des gesamten wissenschaftlichen Zyklus notwendig: von der experimentellen Datenerhebung über die Datenanalyse und anschließende Modellbildung bis hin zur Simulation und einer Modellrevision auf Grund der Simulationsergebnisse (siehe Abbildung 1). Wissenschaftliches Arbeiten, das von einer solchen Art der Computer-Unterstützung getragen wird, bezeichnen wir als *Computer-Supported Research*.

Zwei Beobachtungen sind für die Unterstützung des gesamten wissenschaftlichen Zyklus sehr hilfreich: Zum einen ist es für die Verarbeitung von Daten unwichtig, ob sie aus Experimenten in der realen Welt oder aus Simulationen stammen. Zum anderen werden Modelle selbst heute mehr und mehr als First-Class-Objekte, d.h. als Daten behandelt: sie werden zwischen Programmen ausgetauscht, in Datenbanken gespeichert, über das Internet verschickt und durch Programme manipuliert.

Unterstützung des wissenschaftlichen Zyklus

Wie genau soll aber die Computer-Unterstützung erfolgen? Dafür können die einzelnen Schritte des wissenschaftlichen Zyklus betrachtet werden (Abbildung 1). Wie man der Abbildung 1 entnehmen kann, besteht der wissenschaftliche Zyklus aus zwei Zyklen der Modellrevision: die Modellrevision erfolgt einmal aufgrund von experimentellen Beobachtungen in der realen

[1] `http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf`

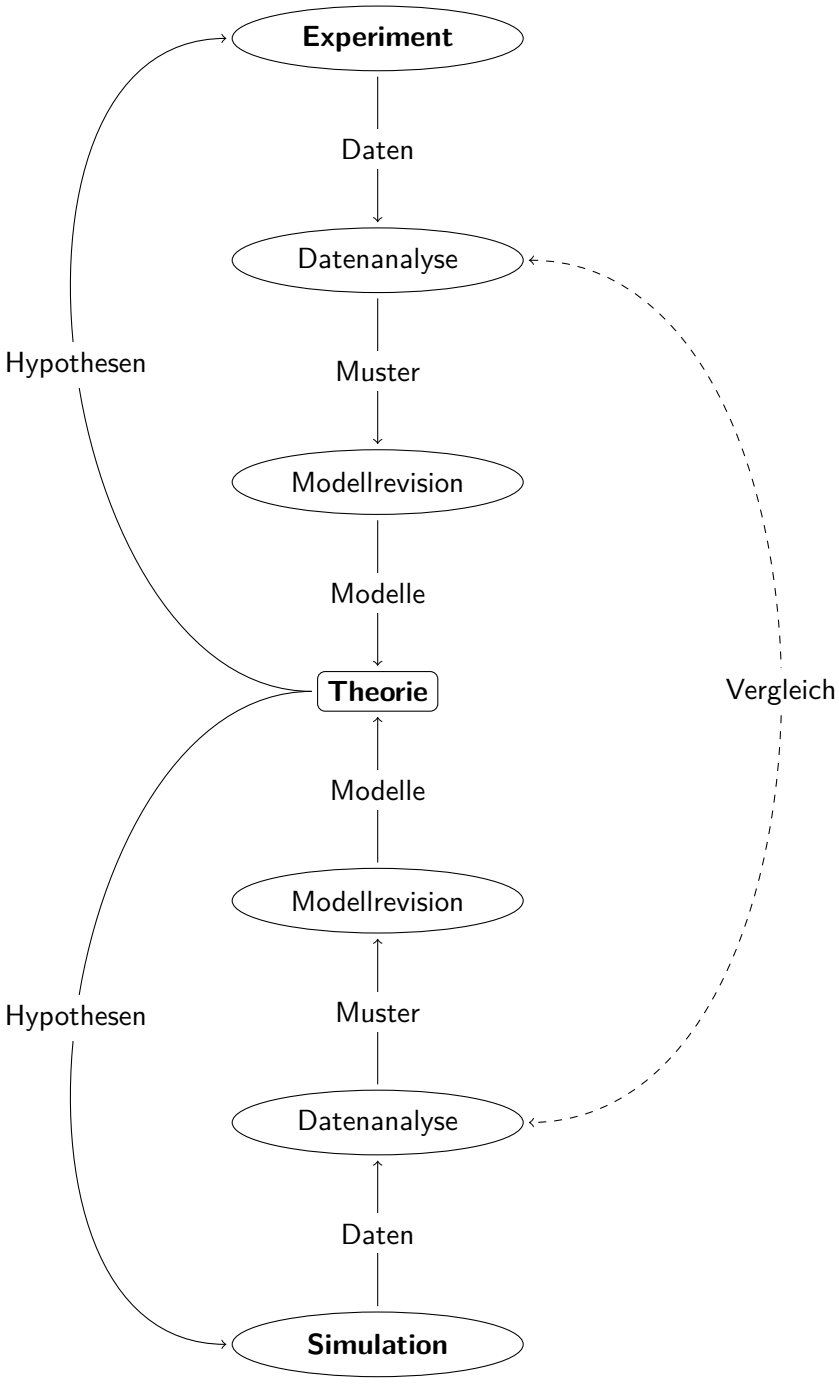

Abbildung 1: Wissenschaftlicher Zyklus

Welt. Zum anderen liefert die Überprüfung von Vorhersagen, wie sie in Simulationen gewonnen werden, Ansatzpunkte für die Modellrevision. Dabei müssen experimentelle Daten und Simulationsdaten miteinander verglichen werden.

Bei der experimentellen Datenerhebung kommen bereits heute Computer zum Einsatz: komplexe Abläufe werden automatisch gesteuert, Messergebnisse werden gefiltert und gespeichert. Da die Datenerhebung teuer ist und in der Regel nicht beliebig oft wiederholt werden kann, ist es angebracht, die Daten nicht nur für eine wissenschaftliche Untersuchung zu verwenden, sondern zu speichern und für andere Untersuchungen auch langfristig verfügbar zu machen. Neben der technischen Realisierung der Speicherung stellt die Kuration der Daten dabei eine besondere Herausforderung da: die Daten müssen von offensichtlichen Fehlern befreit werden. Damit soll eine hohe Datenqualität gewährleistet werden. Zusätzlich müssen ausreichend Meta-Informationen verfügbar gemacht werden, die eine spätere Verwendung der Daten durch Dritte gestatten. Erst durch diese zusätzlichen Informationen werden die Experimente nachvollziehbar und die Daten nachprüfbar. Dies sichert das wissenschaftliche Prinzip der Reproduzierbarkeit. Sowohl die Daten selbst als auch die Meta-Informationen müssen dabei in computerverstehbarer Form abgelegt werden, um die Suche in riesigen Datenbeständen unterstützen zu können.

Sind die richtigen Daten für eine wissenschaftliche Untersuchung aus eigenen Experimenten erhoben bzw. in Datenbanken gefunden wurden, geht es darum, diese zu analysieren. Dies erfordert oft den Einsatz massiver Rechentechnik, in vielen Fällen auch die Erstellung maßgeschneiderter Algorithmen und Entwicklung von Spezialhardware durch Experten. Hier wäre natürlich die Verfügbarkeit von generischen Methoden wünschenswert, z.B. als Web-Services. Bei der Auswahl und Adaption der Methoden ist Computer-Unterstützung notwendig, die den Wissenschaftler durch den Analyse-Prozess leitet.

Die Datenanalyse sucht im Allgemeinen nach bestimmten Regelmäßigkeiten und liefert Muster. Auf der Grundlage der Muster wird ein neues Modell gebildet. Dieser hochinnovative Vorgang erfordert viel Spezialwissen und erfolgt meist auf der Basis bereits bestehender Modelle, die erweitert, verfeinert oder verallgemeinert werden. Somit ist Modellbildung im Prinzip meist eine (datenbasierte) Modellrevision. Dieser Prozess erfolgt momentan meist „per Hand“, was aber durch die Größe und Komplexität der Modelle sowie durch die schiere Menge an einfließenden Informationen immer schwieriger wird. Computer-Unterstützung ist bei der Auswahl geeigneter Ausgangsmodelle, bei deren Modifikation und Integration mit den betrachteten Mustern, bei der Organisation und Visualisierung einfließenden Wissens sowie bei der Konsistenzprüfung mit Grundannahmen und vorgegebenen Randbedingungen vorstellbar. Modelle können, wie oben erwähnt, als Daten betrachtet werden. Sie können somit in Modelldatenbanken abgelegt werden und müssen einem analogen Kurationsprozess wie die experimentellen Daten unterzogen werden: Ihre Qualität wird nach bestimmten Kriterien überprüft und sie werden mit Meta-Informationen versehen.

Die Modelle werden in Simulationen verwendet, um bestimmte Hypothesen zu prüfen. Die bei der Simulation anfallenden Daten gehen in einen neuen Modellrevisionszyklus ein. Zusätzlich kann ein Vergleich mit den experimentellen Daten vorgenommen werden. Dieser Vergleich erfolgt in der Regel nicht auf den Rohdaten, sondern auf einer qualitativen Ebene, wobei wieder eine Computer-Unterstützung bei der Auswahl der richtigen Vergleichsmethode und deren Durchführung erwünscht ist. Auch bei der Generierung von Experimenten und Simulationen auf der Grundlage von Hypothesen ist Computer-Unterstützung vorstellbar. Insbesondere bei Simulationen könnte eine automatische Exploration des Modellverhaltens erfolgen.

Kuration und Semantik für Computer-Supported Research

Die Kuration von Daten und Modellen, ihre Beschreibung in geeigneten Austauschformaten und die Speicherung in öffentlich zugänglichen Datenbanken ermöglicht die Wiederverwendung und fördert die weltweite Kollaboration zwischen Wissenschaftlern. Die oben beschriebene Computer-Unterstützung setzt insbesondere eine Formalisierung der Meta-Informationen für Daten und Modelle voraus. Durch geeignete Annotationen ist es möglich, die Bedeutung relevanter Aspekte von Daten und Modellen zu referenzieren.

Neben dieser formalen Semantik von Daten und Modellen benötigt Computer-Unterstützung im gesamten wissenschaftlichen Zyklus jedoch noch mehr Informationen in einer computerverstehbaren Form, Dafür müssen viele weitere Aspekte des einfließenden Wissens formal repräsentiert werden. Zu diesem Wissen zählen z.B. Grundannahmen, Randbedingungen und Hypothesen. Auch die Beschreibungen von Experimenten und Simulationen müssen geeignet formalisiert werden, um zum einen Bezüge zwischen Daten und experimentellem Setup herstellen zu können und zum anderen die Automatisierung von Experimenten und Simulationen zu ermöglichen. Eine besondere Herausforderung stellt die formale sprachliche Beschreibung der Dynamik von Modellen und realen Systemen dar, auf deren Grundlage qualitative Vergleiche möglich werden.

Ausblick

Langfristig führt die hier vorgestellte Vision einer computergestützten Forschung zu einer völlig neuen Art von Wissenschaft: der Computer wird zum Assistenten im gesamten wissenschaftlichen Zyklus. Dadurch wird es Wissenschaftler wieder möglich werden, ihren stark spezialisierten Fachbereich zu verlassen und Wissen und Daten aus anderen Bereichen mit einzubeziehen. Diese umfassende Integration wird zu völlig neuen Erkenntnisse führen und ein viel tieferes Verständnis realweltlicher Prozesse eröffnen.

Der Horizont der Forschung wird durch Computer-Unterstützung in zwei Richtungen erweitert werden: Zum einen wird es damit möglich, eine sehr viel größere Menge von Daten, Modellen und existierendem Wissen zu analysieren und dadurch neue Erkenntnisse zu gewinnen. Und zum anderen wird es mit einer geeigneten Computer-Unterstützung auch gelingen, hochkomplexe Zusammenhänge und Wirkbeziehungen aufzudecken, die einem menschlichen Auge allein prinzipiell verschlossen sind. Computer-Supported Research trägt somit das Potenzial nicht nur für eine neue Quantität sondern auch für eine neue Qualität in der Forschung.