6. KoMSO Challenge Workshop Big Data

Heidelberg, 3.-4. März 2015

1 Ausgangslage

Big Data ist ein ebenso weitgefasster wie allgegenwärtiger Begriff. Meistens werden damit Herausforderungen verstanden, die im Zusammenhang mit den immer größer werdenden Datenmengen im Internet, im Mobilfunk oder herausragenden Anwendungsfeldern wie z.B. Genomik/Proteomik, Wettervorhersage oder Finanzdaten entstehen. Die damit verbundenen ungelösten Probleme der Datenhaltung, -übertragung, -auswertung oder deren Integration in nachfolgende Prozesse sind fundmentaler Natur und stellen in allen Bereichen heutige Modelle und Vorgehensweisen in Frage. Hier sind mathematische Grundlagen- und Anwendungsforschung ebenso gefragt wie enge Kooperationen mit Informatikern und Vertretern der Informationstechnologien.

Unter dem Begriff *smart data* statt *big data* wird dabei verstanden, dass diese unüberschaubaren Datenmengen nicht einfach gespeichert und möglichst effizient über die weltweiten Netze übertragen werden, sondern dass sie mit Hilfe optimierter Auswerteverfahren annotiert und die wesentliche Informationen komprimiert und in geeigneter Form den jeweiligen Nutzern zugänglich gemacht werden.

Es ist allgemein akzeptiert, dass die spezialisierten Nutzer jeweils nur einen Bruchteil der in den Datenmengen vorhandenen Informationen wirklich benötigen und nutzen – die grundlegenden Konzepte für die dazu notwendige Datenanalyse ist mathematischer Natur. Wir sind der Überzeugung, dass die dringendsten Herausforderungen der Big Data Szenarien ohne Mathematik nur unzureichend, unbefriedigend und insbesondere suboptimal gelöst werden.

2 Challenge Workshop

Der Workshop war auf 50 Teilnehmer (je ca. 25 Teilnehmer aus Wirtschaft/Industrie und dem akademischen Bereich) beschränkt, um eine intensive Diskussion zwischen den Teilnehmern zu ermöglichen. Innerhalb kürzester Zeit war der Workshop ausgebucht, wir hätten ohne Weiteres eine Vielzahl von Teilnehmern gewinnen können.

Ein weiteres Indiz dafür, dass Wirtschaft und Industrie diesem Thema eine hohe Bedeutung beimessen, war die Qualität der vertretenen Firmen: fünf DAX-Unternehmen (BASF SE, Bayer AG, Merck KGaA, SAP SE, Siemens AG), weitere Großunternehmen

(Google Inc., Hoffmann-La Roche, Horiba Jobin Yvon GmbH, SAS Institute GmbH) sowie mehr als zehn weitere KMUs und nicht akademische Einrichtungen (u.a. Deutscher Wetterdienst, SKZ). Auf der akademischen Seite waren neben verschiedenen Universitäten das Max-Planck-Institut für dynamische Systeme und Selbstorganisation, Göttingen, Fraunhofer ITWM, Kaiserslautern, EMBL, Heidelberg vertreten.

Angeregt durch die überraschend offenen Vorträge insbesondere der Industrievertreter ergaben sich im Verlauf des Workshops verschiedene Diskussionskreise, die spezielle Fragestellungen im Umfeld von Big Data Anwendungen diskutierten. Die Ergebnisse der Diskussionen sind stichwortartig im Abschnitt „Schwerpunktthemen“ zusammengefasst.

Der Vorschlag einer mathematischen Big Data Initiative wurde einhellig unterstützt.

3 Schwerpunktthemen

Mathematische Beiträge zu den Big Data Herausforderungen werden insbesondere in den Bereichen Datenauswertung/Information Retrieval und Integration in nachgeschaltete Produktions-/Verfahrensprozesse erwartet. Das Spektrum der dabei potentiell zum Einsatz kommenden Verfahren reicht von der reinen Mathematik (topologische Datenanalyse, algebraische Strukturen und Invariantentheorie) und sich gerade etablierenden Bereichen wie hybride Verfahren und *machine learning* bis zu klassischen Verfahren der mathematischen Datenanalyse und Statistik und zu allgemeinen Verfahren des HPSC (*high-performance scientific computing*).

Die Interessen der beteiligten Firmen fokussierten insbesondere auf 1) Anwendungen im Pharmabereich mit dem Schwerpunkt hyperspektrale Datenanalyse und Grenzen herkömmlicher Statistikverfahren z.B. bei multi-kausalen Krankheitsbildern sowie den Bereich 2) Modellbildung und Forecasting (Google, DWD, Siemens) mit Anwendungen auf die Wettervorhersage und nachgeschaltete Dienstleistungen sowie *commodity pricing*.

Um im internationalen wissenschaftlichen Wettbewerb bestehen zu können, ist im Rahmen einer eventuellen mathematischen Big Data Initiative eine Fokussierung auf ausgewählte Themenbereiche notwendig. Diese Auswahl sollte sich an den Interessen der deutschen und europäischen Wirtschaft/Industrie orientieren. Die Kernthemen, die im Rahmen dieses Workshops insbesondere durch die Vertreter von Bayer, Hoffmann-La Roche und Siemens geprägt wurden (hyperspektrale Datenanalyse in der Pharmaforschung, Modellbildung und Forecasting), sollten im Rahmen einer nationalen Initiative durch ein bis zwei weitere Schwerpunktthemen ergänzt werden.

4 Thesen für eine mathematische Big Data Initiative

- Big Data (Speicherung, Übertragung, Auswertung, Anwendung) wird im nächsten Jahrzehnt eine zentrale Herausforderung für verschiedenste Industriefelder von der Informationstechnologie über den Pharma/Medizin Bereich bis zu technischen Anwendungen sein.
- In den Bereichen Datenauswertung und Integration in nachfolgende Prozesse wird die zukünftige Entwicklung weniger durch Hardware-Entwicklungen sondern überwiegend durch intelligente Konzepte zum Information Retrieval und dessen Integration in nachfolgende Prozesse bestimmt.
- Dies erfordert an erster Stelle mathematische Kompetenz zur Strukturierung der Daten, Algorithmen zu deren Auswertung und Konzepte zur Integration in nachfolgende Prozesse.
- In Deutschland ist in besonderem Maße mathematische Kapazität in den MSO-Kompetenzfeldern vorhanden.
- Eine fokussierte Initiative zur Unterstützung mathematischer Forschung im Big Data Umfeld in Kooperation mit Firmen könnte einen entscheidenden Beitrag leisten, um die damit zusammenhängenden Herausforderungen von Wirtschaft und Industrie der nächsten Jahrzehnte erfolgreich zu bewältigen.

Kontakt
Dr. Anja Milde
KoMSO Coordination Office
IWR – Interdisciplinary Center for Scientific Computing
Im Neuenheimer Feld 368
69120 Heidelberg
Germany

T: +49 (0)6221 – 54-8886
F: +49 (0)6221 – 54-8810
komso@iwr.uni-heidelberg.de