Memorandum: “Big Data made in Germany”
Symposium 29. - 30. Juni 2017, Berlin

Paradigmenwandel in Wissenschaft, Wirtschaft und Gesellschaft: Datenfluten beeinflussen unser Leben und Arbeiten in immer stärkerem Maße. Prominente Beispiele für den rasanten Anstieg von Datenmengen im “Netzwerk der Dinge” sind Bereiche der sozialen Medien, der Industrie 4.0 oder alternative Mobilitätskonzepte. Methoden zur Analyse dieser Datenfluten bilden einen Wirtschaftszweig, dessen reale Bedeutung und Wert sich erst in der Zukunft in vollem Umfang abschätzen lässt. **Hier hat Europa im Vergleich zu Nordamerika** – mit seinen großen IT-Firmen – einen **Nachholbedarf**, den es in den nächsten Jahren zu kompensieren gilt. Deutschland kann durch gute Ausbildung und Infrastruktur eine Führungsposition einnehmen.

Mögliche Handlungsoptionen vorzuschlagen, war eine der wesentlichen Ziele des Symposiums “Big Data made in Germany”. Im folgenden werden diese wiedergegeben, sowie die wichtigsten Aussagen und Thesen.

Die **durch Big Data getriebene digitale Transformation der Gesellschaft beschleunigt sich** durch die Fortschritte bei der Anwendung künstlicher Intelligenz und des Maschinenlernens rapide.

Allein 90% des Datenaufkommens weltweit haben ihren Ursprung in den letzten zwei Jahren. Einerseits bedeutet dies, dass Anstrengungen unternommen werden müssen, die Prozesse nachvollziehbar und kontrollierbar zu machen. **Die rechtlichen Rahmenbedingungen für die Big Data Analytics müssen überhaupt erst geschaffen werden**, wie Thomas Wiegand, Direktor am Fraunhofer Heinrich Hertz Institut in Berlin und Moreen Heine, Expertin für Digital Government von der Universität Potsdam, unisono betonten. Die **Konformität der technischen Möglichkeiten beispielsweise im Bereich des Verkehrs mit dem Grundgesetz stellt eine enorme Herausforderung** dar, einschließlich der bekannten ethischen Grundfragen, worauf MDirig. Andreas Krüger vom Bundesministerium für Verkehr und digitale Infrastruktur nachdrücklich hinwies. Klaus Wiegerling vom Institut für digitale Ethik am Karlsruher Institut für Technologie warnte vor einer Entwicklung, die die **Ängste der Menschen vor der unsichtbaren Macht der Big Data** und nicht zuletzt dem Szenario des massenhaft drohenden Arbeitsplatzverlustes ignoriert. Viele Folgen der digitalen Transformation der Gesellschaft seien noch nicht überschaubar und müssten gesellschaftlich viel stärker diskutiert werden.

Die digitale Transformation stellt einen idealen **Nährboden für junge Unternehmen im Big Data Business** dar, denen sich eine unbeschränkte Zahl neuer Geschäftsmodelle eröffnet. Für die Startups sind die Verfügbarkeit von gut ausgebildetem, weiterhin lernfreudigem und international versiertem Personal, Risikokapital und ein florierendes Ökosystem von Kunden die entscheidenden Erfolgsfaktoren. Franz Färber, Executive Vice President bei SAP wies darauf hin, dass die Wertschöpfung im Big Data Business mit der Komplexität des Use-Cases anwächst und unterstrich damit die **Bedeutung der wissenschaftlichen Big Data**

Flagschiffprojekte: Large Hadron Collider (CERN) und Square Kilometre Array (SKA). Dort werden Methoden zur Sicherung der Authentizität von Daten, des kontrollierten Zugriffs und der exakten Analyse entwickelt, die in der nächsten Stufe der Big Data auch in wirtschaftlichen Prozessen relevant sein werden. Der Präsident der Deutschen Physikalischen Gesellschaft und frühere Direktor des CERN, Rolf Heuer, erläuterte, wie durch unabhängige, getrennte Analysen von zunächst nicht-öffentlichen Daten reproduzierbare Meßergebnisse abgesichert werden. Vielleicht liegt darin ein **Schlüssel zur öffentlichen Kontrolle der Big Data Analytics**.

Unabhängig von der Größe des Unternehmens sind **Nachwuchskräfte aus Deutschland Mangelware**. Dies unterstrich auch Michael Franzkowiak von der jungen Softwareschmiede Contiamo aus Berlin. Die Universitäten sollten daraus Schlussfolgerungen ziehen. Neben einer Verstärkung der Ausbildungsqualität könnten sie mit eigenen Digital Labs den initialen Kontakt zwischen Startups und Investoren sowie Vertrauen und die Risikobereitschaft fördern. Es ist bemerkenswert, dass die Open-Source-Tool und Cloud-Computing-Entwicklung heute fast ausschließlich in den USA stattfinden. Diesen Punkt machte auch Volker Markl, Direktor am Berlin Big Data Center klar, der auf die strategische Besetzung von Kontrollpunkten hinwies. Es sind deshalb **Maßnahmen zu überlegen, wie man auch in Deutschland und Europa mehr Gestaltungseinfluss gewinnen könnte**. Die Wissenschaft ist in den strategischen Überlegungen bereits sehr weit vorangeschritten, allein es fehlt der "Vollzug". Joachim Wanbsganß, Vizepräsident der Astronomischen Gesellschaft, wies auf die einschlägigen Publikationen der Schwerpunktinitiative "Digitale Information" der deutschen Wissenschaftsorganisationen und des Rats für Informationsinfrastrukturen hin, die online verfügbar sind. Dort wird vor allem auf die **Notwendigkeit von Investitionen in Köpfe** als Kern einer de-institutionalisierten, breit angelegten Strategie hingewiesen. Bemerkenswert ist hier der Gleichklang mit den Akteuren im Big-Data-Business hinsichtlich der Forderung nach einer koordinierten und entschlossenen Stärkung der daten-basierten, multi-disziplinären Forschung an den Hochschulen nach dem Motto "Leistung aus Vielfalt".

Am ersten Tag des Symposiums präsentierten Referenten aus Wissenschaft und Wirtschaft ihre Einschätzungen der zu erwartenden Trends und Anforderungen im Bereich Big Data. Hierbei wurden Projekte und zukunftsweisende Konzepte des Transports, der Analyse, der Speicherung und der Datenverarbeitung vorgestellt.

Big Data spielt in der Grundlagenforschung (z.B. in Physik, Astronomie, Geowissenschaften, Life Sciences) **eine zunehmend wichtigere Rolle**. Insbesondere die naturwissenschaftlichen Experimente mit den größten Big-Data-Volumina und den komplexesten Anforderungen an die Datenanalyse sind herausragende **Entwicklungsmotoren für Innovationen und Ausbildung**. In der Forschung werden wegweisende Lösungsansätze entwickelt, die zeitversetzt auch in industrielle Anwendungen einfließen.

Mittelfristig werden zwei Flaggschiff-Projekte die weltweit größten Datenproduzenten sein: Der Large Hadron Collider (LHC) am CERN in Genf und das Square Kilometre Array (SKA), das aus Tausenden Radio-Teleskopen bestehen und in Südafrika und Australien errichtet wird. In beiden Großforschungsvorhaben übersteigt die Produktion an Rohdaten den

Datenverkehr im weltweiten Internet. Eine Kernaufgabe besteht darin, schon während der Datennahme den Anteil der physikalisch interessanten Daten herauszufischen und die immensen Datenmengen auf ein vertretbares Maß zu reduzieren. Der Transport der Daten ist gegen Verfälschungen zu sichern. Transparente Standards für Datenintegrität und Datenverschlüsselung setzen **öffentlich kontrollierte Plattformen** voraus. Im Mittelpunkt stehen quantifizierbare und reproduzierbare Messungen und die für ihre Durchführung erforderlichen Methoden, sowohl in Bezug auf die Software als auch die Hardware. Intransparenz, Manipulation, Datenkorruption oder unkontrollierte Rückwirkungen auf die Messvorgänge sind dabei auszuschließen. Dies gehört elementar zur wissenschaftlichen Kultur und sollte zwingend auch bei der anwendungsbezogenen Nutzung von Big Data Analytics vorgeschrieben werden, bedarf aber der zuvor stattfindenden Entwicklung entsprechender Verfahren und der Einrichtung von Plattformen, über die die Einhaltung von Standards überwacht werden kann. *Möglicherweise kann die wissenschaftliche Methodik zum **Bürgerlichen Gesetzbuch für die Welt der Big Data Analytics** werden.*

Die **Datenanalyse**-Workflows **der beiden Flaggschiff-Projekte unterscheiden sich grundlegend** und erfordern die Entwicklung unterschiedlicher Datenverarbeitungsmodelle. **LHC** zeigt einen prinzipiellen Weg auf, wie eine Daten-Kompression erfolgen kann, die einen Rohdatenstrom von 1 Petabyte pro Sekunde auf ein Archivwachstum von 100 Megabyte pro Sekunde reduziert. Trotz dieser starken Reduktion werden die langfristig zu speichernden Daten in den Bereich von Exabytes pro Jahr vorstoßen, wobei die Analyse auf einzelnen Datensätzen basiert. Beim SKA werden pro Jahr bis zu 300 Petabytes an prozessierten Archivdaten erwartet, was deren kohärente Analyse zu einer großen Herausforderung macht. Das extrem breite Spektrum an wissenschaftlichen Fragen beim **SKA** erfordert ein ganz **neues Ökosystem von Datenanalysemethoden** mit hoher Flexibilität.

Während Transport und Management der Datenmengen beim LHC und SKA mit den heutigen Methoden prinzipiell handhabbar sind, stellt die **Entwicklung von neuartigen datengetriebenen Algorithmen** einen entscheidenden Erfolgsfaktor dar, um die wissenschaftlichen Ziele zu erreichen. Die zu erwartenden großen Archivdatenbestände machen die Entwicklung neuer Computing-Modelle notwendig. Es wurde die **Etablierung einer globalen und evolutionär anpassungsfähigen "Data Cloud"** vorgeschlagen, die aus einigen wenigen Datenzentren für die Archive besteht sowie aus einer Vielzahl anwendungsbezogener Rechenzentren, die über einen globalen Netzwerk-Bus mit sehr hoher Bandbreite Zugang zu den Archiven erhalten und diesen zusammen mit einer zugehörigen Software-Entwicklungsumgebung an ihre Klientel wissenschaftlicher Nutzer vermitteln. Zu den Aufgaben dieser Science-Data-Cloud würde es gehören, eigene Entwicklungsarbeiten anzustrengen, um auch unter Ressourcenbeschränkung eine steigende Performanz zu gewährleisten.

Big Data ist nicht nur eine Frage der Größe von Datenmengen. Es ist nicht davon auszugehen, dass die Probleme der Wissenschaft durch Unternehmen wie Amazon, Facebook oder Google gelöst werden. Allerdings kann die Wissenschaft von derartigen Unternehmen den effizienten Umgang mit Big Data lernen und davon profitieren. Darüber hinaus sollte das **Wissen um den**

Umgang mit großen Datenmengen und die Datenvolumina nicht ausschließlich in den Händen internationaler Konzerne liegen.

Grundlegende Technologiesprünge werden im Hardwarebereich in absehbarer Zeit nicht erwartet. Somit kommt der Softwareentwicklung eine entscheidende Bedeutung zu. Algorithmen werden dabei unerlässlich sein, allerdings ist ihre Funktionsweise nicht wie in einer Black-Box zu verbergen: Stattdessen sollten Algorithmen transparent sein, kontrollierbar und wohl verstanden. Die **großen nationalen Forschungsinstitutionen** (Fraunhofer, Helmholtz, Max-Planck) **und die Hochschulen verfügen über eine Vielzahl von hervorragenden Kompetenzen** in den Bereichen Daten-Management und Software-Entwicklung, wie auf der Tagung eindrucksvoll in den Vorträgen demonstriert wurde. Auch gibt es viele erfolgreiche **Kooperationen mit der Industrie**. So zum Beispiel analysiert SAP mit dem Max-Planck-Institut für Radioastronomie (Bonn) Pulsar-Daten, um die Analyse-Plattform HANA weiter zu entwickeln. Allerdings gilt es, die **interdisziplinäre Zusammenarbeit** über Fachgrenzen hinweg auszubauen und **strukturell besser** zu **koordinieren**. Die Helmholtz-Gemeinschaft hat mit dem umfangreichen Vorhaben “Large Scale Data Management and Analysis (LSDMA)” gezeigt, dass eine interdisziplinäre Zusammenarbeit zwischen Großforschungseinrichtung und Hochschulen erfolgreich gestaltet werden kann.

In den Universitäten ist die Forschung in Disziplinen wie Physik, Informatik, Statistik und Mathematik wohl etabliert. Die Absolventen sind sowohl für die Wissenschaft als auch für die Industrie qualifiziert. Über die DFG oder das BMBF können Universitäten finanzielle Unterstützung für Forschungsvorhaben erhalten. Allerdings liegt der Fokus eher auf der hardwarenahen Förderung. Die **Entwicklung von Software an der Schnittstelle zu den Use-Cases datenintensiver Fachgebiete wie Physik und Astronomie wird aktuell nur unzureichend gefördert.**

Die Ministerien (sowohl auf Bundes- als auch auf Landesebene) bieten eine Vielzahl von Förderprogrammen zur “Digitalisierung” an - selbst Experten haben kaum einen Überblick. Die Welt der Digitalisierung dreht sich immer schneller. Diese Entwicklung ist in den **Förderprogrammen** und den Forschungsinstitutionen durch eine **angemessene Dynamisierung** zu berücksichtigen: Die bisherige Projektförderung ist vom Ansatz her statisch und muss agiler werden. Zwischen der ersten Idee für ein Programm, über die Veröffentlichung der Ausschreibung, bis hin zur Bewilligung von Anträgen vergehen in der Regel mehr als zwei Jahre. In der Zwischenzeit wird die Ursprungsidee immer öfter von der Realität überholt sein: An einem Antrag, der das komplexe Antragsverfahren erfolgreich durchlaufen hat, wird bislang kaum eine nachträgliche Adjustierung der Ausrichtung vorgenommen.

Das **Ansehen von “wissenschaftlicher Softwareentwicklung” ist zu stärken**. Die bestehende Lücke zwischen der Informatik und den “Basis-Wissenschaften” sollte geschlossen werden. Die Hochschulen können durch Anreize in die Lage versetzt werden, ihre Studienangebote entsprechend zu erweitern und Querschnittsexperten als wissenschaftliche Daten-Ingenieure und Daten-Wissenschaftler auszubilden. Universitäts-Rechenzentren sollten Entwicklungsabteilungen für die Durchführung von datenwissenschaftlichen Projekten erhalten,

in denen multidisziplinäre Teams regelmäßig zusammenfinden und sich austauschen. Diese Datenwissenschafts-Zentren würden als Inkubatoren fungieren und durch Beratungsangebote sowie durch die Bereitstellung von Infrastruktur auch Universitätsausgründungen im Bereich der Datenwissenschaften stimulieren. Der Austausch von Ideen über Fachgrenzen hinweg und die Konstellation von sich gegenseitig ergänzenden Kompetenzen stellt einen idealen Nährboden für junge Gründer dar.