Forschungsgipfel 2015 (Berlin, den 20.05.2015)

Inputrede für das Themenforum „Open Access und Big Data“

Leitfrage: Wie führt die Öffnung von Datenbeständen zu Innovationen?

Mit der Entstehung von „Big Data“ beginnt das Zeitalter eines offenen Datenökosystems, dessen Management ein „Big Mind“ erfordert, um das Zusammenwirken der menschlichen Intelligenz und der künstlichen Intelligenz zu optimieren.

Unser Forum hat ein interessantes Thema „Open Access und Big Data“, bestehend aus 5 einfachen Wörtern. Darunter sind interessanterweise 4 englische Wörter mit 2 Adjektiven. Ich darf meine persönliche Auffassung zu Big und Open mit Ihnen teilen.

Big

„Big“ ist wissenschaftlich gesehen ein fuzzy Konzept und hängt von informationstechnischen Fortschritten ab. Noch nie würden unsere Erdoberfläche und das Verhalten von Lebewesen so intensiv und umfassend digitalisiert wie heute.

„Big“ im Volksmund ist eine Alltagserfahrung und deutet eine Massenbeteiligung an. Jeder darf seine emotionsgeladenen Erlebnisberichte, Eindrücke über „hier und jetzt“ und auch seine Ideen als User-Generated Content (UGC) im Internet mit seinen Mitmenschen teilen.

Also, Big verbindet die Wissenschaft mit der Gesellschaft.

Drei Typen von Datenlieferanten arbeiten konkurrierend, kollaborierend und komplementär zusammen:

- Amtliche Behörden mit flächendeckenden, strukturierten Basisdaten von hoher Qualität
- Unternehmen mit ergänzenden Basisdaten, semistrukturierter und kontrollierter Crowdsourcing nach wettbewerblichem Handlungsgrundsatz
- Private Personen mit eher unkontrollierter und unstrukturierter Crowd-Sourcing

Jede der drei Eigenschaften von Big Data – big, schnell und heterogen – ist wissenschaftlich herausfordernd:

Big ist Auslöser von Big Data. Man darf „big“ weder mit „gut“ noch mit „ganz“ gleichsetzen. Nichtsdestotrotz verbirgt sich in Big Data ein großes Nutzungspotential, aus umfangreicheren Stichproben mit höheren Wahrscheinlichkeiten nicht nur das bekannte Wissen zu bestätigen, sondern neues Wissen, insbesondere die Kausalitätsrelationen statt nur Korrelationen zwischen den Dateneinträgen zu entdecken.

Schnell - Big Data sind dynamische Datenströme, die nicht wie Flüsse aus allen Richtungen in einen großen Ozean münden, sondern sich netzwerkartig nach allen Richtungen hin entfalten. Zur Bewältigung der Datenströme sind smarte Systemarchitekturen für die inkrementelle und parallele Computing erforderlich.

Heterogen - Die Vielfalt der Big Data ist eine Voraussetzung für die Vitalität des Datenökosystems, sie erschwert zugleich die flächendeckende Qualitätssicherung sowie die Entwicklung von kosteneffizienten Data-Mining-Verfahren. Der Mehrwert in der Vielfalt liegt in erster Linie nicht in den einzelnen Dateneinträgen (Knoten), sondern in deren Beziehungen. Die naturwissenschaftlich messbaren Kenngrößen taugen offenbar nicht zur Bewertung der Mensch-Mensch-Beziehungen und der Mensch-Ding-Beziehungen. Es fehlt noch die Grundlage zur Modellierung semantischer Wechselwirkungen. Zwei Geotags lassen sich beliebig verlinken ohne dass sie eine Ähnlichkeit teilen. Die Kommunikationsintensität ist nicht unbedingt aussagefähig für die Beziehung zwischen zwei Menschen. Unsere Forschungsfrage lautet „wie kann man ein physisches Netzwerk mit einem sozialen Netzwerk zusammenführen?“

Open

Die Offenheit ist ein doppelseitiges Schwert. Auf der einen Seite sind wir oft geplagt von der sog. Komplotttheorie und dem digitalen Schatten. Die digitalen Nachrichten und Fußstapfen der individuellen Personen werden leicht manipuliert und fortgepflanzt. Die Komplotttheorie und die digitalen Schatten mit ihrem Beigeschmack machen Big Data ethisch fraglich. Auf der anderen Seite ist die Offenheit auch ein Gegenmittel gegen den wissenschaftsbeschädigenden und menschenrechtsverletzten Datenmissbrauch. Wissenschaftler sind dafür verpflichtet, das offene Portal im Zusammenhang mit der Selbstheilungsfunktion zu gestalten und zwar nach dem Motto „Mehr Augen helfen zur gegenseitigen Entdeckung und Korrektur der Fehler in den Daten“. Die Wahrheit kommt früher oder später zum Vorschein, wie das Öl über dem Wasser.

Die Offenheit fördert auf der einen Seite die nachhaltige Datengenerierung. An der Stelle, wo einige Institutionen oder Freiwillige inaktiv geworden sind, springen neue Beitragende ein, somit wird der Datenstrom kontinuierlich erzeugt und gepflegt. Im Bereich Katastrophenmanagement spielt die “just-in-time”-User-Generated-Content wegen ihrer radikalen Offenheit, Flexibilität und ihrer ortsbezogenen Intelligenz eine unverzichtbare Rolle zur Linderung der Frustration „Informationen sind dort am wenigstens vorhanden, wo sie am dringlichsten benötigt werden“. Auf der anderen Seite leiden wir noch unter den sog. digitalen Spaltungen auf Makroebene (z.B. politisches System, Entwicklungsstand der IT-Infrastruktur) und Mikroebene (z.B. Alter, Gender, Computerkenntnisse). Die ungleichmäßige Verteilung der Datenlieferanten macht die Repräsentativität der Big Data fraglich. Auf der anderen Seite bietet aber Open Source Plattformen einen guten Ausgangspunkt, um die gegensätzlichen Meinungen für offene Bewertung durch die Partizipation der Gesellschaft unterzubringen. Jeder darf dazu stoßen und mitreden. Längerfristig werden die digitalen Spaltungen schrumpfen.

Also, Offenheit bringt mehr Vorteile als Nachteile mit sich. Hierfür sind zwei positive Beispiele:

1. Open Science

Mehr und mehr Forscher wählen die Option der kostenfreien On-line-Veröffentlichung, teilen ihre Testdaten / Stichproben mit weltweiten Kollegen. Der transparente Ideen- und Datenaustausch erlaubt die Wiederholung der Experimente, die konstruktive Ergänzung sowie die Fehlerkorrektur, erhöht dadurch die gesellschaftliche Sichtbarkeit der Forscher und führt letztendlich zu einer

verbesserten wissenschaftlichen Integrität und Berechenbarkeit. Exzellente Wissenschaftler sind zunehmend bereit, hochqualitative Dateneinträge in Open-Access-Portal wie Wikipedia zu veröffentlichen, ein Zeichen für die verbesserte Akzeptanz.

2. Open Geodaten
Nach der häufig zitierten These, dass mindestens 80% aller Daten einen Raumbezug aufweisen, bilden dann Big Geodata den Hauptbestandteil der Big Data und spielen daher eine zentrale Rolle zur Gestaltung des Big-Data-Ökosystems. Big Geodaten sind gemeinnützig und sind für das Allgemeinwohl zunehmend als Nahrungsmittel offen zugänglich, zumindest ab einer bestimmten Auflösung. Deutschland hat einen Wettbewerbsvorteil und ist bereits exzellent positioniert im internationalen Vergleich hinsichtlich der Konzipierung der raumzeitlichen Dateninfrastruktur und der Standards für die Geodateninteroperabilität.

Unser Zeitgeist im 21.Jh. ist nicht mehr die Einmischung in die inneren Angelegenheiten anderer Länder und die territoriale Ausdehnung (wie im 19.Jh.), oder die Unabhängigkeit (wie im 20.Jh.), sondern die Wechselwirkung:

- zwischen Ländern auf einer Makroebene
- zwischen Wissenschaft und Gesellschaft auf einer Mesoebene
- zwischen Natur- und Sozialwissenschaften auf einer Mikroebene

Das Wissen über diese Wechselwirkungen finden wir z.T. in Open Big Data.

Als Geodätin und Kartographin habe ich oft das Gefühl, dass die globalen Navigationssatelliten und Fernerkundungstechnologien uns nicht in erster Linie den Eindruck vermitteln, den Weltraum erobert zu haben, sondern sie erlauben uns, aus einem höheren Blickwinkel zu sehen, wie zerbrechlich unser blauer Planet ist und die Topographie, die wir auf der Erde miteinander teilen, viel wertvoller ist als die Ländergrenze, die uns trennt.

Wir leben in einer gemeinsamen aber unsicheren Welt. Die Unsicherheit betrifft sowohl die digitale Welt also auch die physische Welt. Unsere Haltung gegenüber der Unsicherheit muss aber geändert werden. Wir sollen lernen, die Unsicherheit als einen Immunitätsstärkenden Bestandteil unseres Lebens zu akzeptieren, bewusst und konstruktiv mit der Unsicherheit umzugehen. Wenn wir so denken, werden wir dazu neigen, alles sinnvoll zu regulieren statt zu überregulieren bis Innovationsideen in immer dicker werdenden Regelwerks ersticken.

Fazit im Zusammenhang mit dem Forschungsgipfel
Ein Wettbewerbsnachteil von Deutschland besteht m.E. in der Überregulation der Sicherheitsmaßnahmen, die zum Scheitern einiger großer Projekte geführt hat. Die Ängstlichkeit könnte eine Teilerklärung sein für die Tatsache, dass heutiges Deutschland in der inkrementellen Innovation stark, aber in der Durchbruchsinnovation eher schwach ist. Deutschland ist als Land der Ideen bekannt. Aber es nützt wenig, wenn die Ideen zu langsam oder nur anderswo umgesetzt werden. In der Wissenschaft wie auch in der Wirtschaft herrscht die Wettbewerbsregel: nicht mehr die größeren oder die stärkeren schlagen die kleineren oder die schwächeren, sondern die schnelleren schlagen die langsameren. Also, lieber Fehler zu überholen als von Fehlern überholt zu werden. Ich plädiere hier nicht für eine leichtsinnige Beschleunigung, sondern möchte unterstreichen, dass

Innovationen die Risikobereitschaft voraussetzen. Wenn wir einen Innovationsprozess überregulieren, um die Unsicherheit zu minimieren oder gar zu beseitigen, dann haben wir den Wettbewerb bereits halbwegs verloren.

Wenn ich für die deutschen Hochschulen sprechen darf, würde ich ihre Internationalität noch erhöhen. Die Exzellenzinitiative in den letzten 10 Jahren hat den Wettbewerbsgeist unter den deutschen Universitäten erheblich stimuliert. Die Gesetzänderung (91b Absatz 1 GG) hat eine nachhaltige Entwicklung offenbar begünstigt. Dennoch wird weder der Umfang noch die Qualität der deutschen Hochschulen in diversen Rankingsystems wahrhaft widergespiegelt. In einer Zeit, wo Aufmerksamkeit ein knappes Gut geworden ist, sind wir mehr denn je zur Selbstdarstellung verdammt. Wir dürfen nicht erwarten, dass andere uns die nötige Beachtung schenken, wenn wir nicht für uns selbst die Werbetrommel rühren. Die Outgoing- und Incoming-Mobilität soll als Teil des Campuslebens verinnerlicht werden, um jungen Talenten zu erlauben, so früh wie möglich ihre Fähigkeit zum Change-Management, ihre Anpassungs- und Kommunikationsfähigkeit in einem multikulturellen und multidisziplinären Lern- und Arbeitsumfeld zu entwickeln, und zu verstehen, dass zu einer Fragestellung mehrere Lösungswege existieren und dass Steine aus anderen Hügeln Jade polieren können.

Liqiu Meng

中文译文：
2015 年研究峰会专题论坛 **Open Access und Big Data** 发言内容。论坛的主题问题为“数据开放会带来创新吗？”

大数据（Big Data）概念的形成开启了一个开放式数据生态系统时代。我们需要一个强大的头脑 （Big Mind）来管理这样一个生态系统，以期优化人的智能和机器智能的综合效应。

我们这个专题论坛的题目很有意思，总共五个单词里竟有四个英文单词，其中包括两个关键词 “open”和“big” 。以下我试图说说自己对这两个形容词的理解，希望能够引起大家的一些共鸣。

先说“大”字。
“大” 从科学的角度看是个模糊概念。数据大不大和信息技术的发展有关。昨日的数据量用今日的计算能力衡量的话不算大，但在高度联网和安全的计算环境实现之前，今日的数据量和急增速度着实让我们招架不住。我们的地球表面和地球上生物的行为从未象今日这样广泛而密集地经历着反复不断的数字化。“大”从老百姓的角度看是一个司空见惯的口头语，隐含着大众参与的意思。每一位网民随时可以和其他网民分享他关于“此时此地”的情绪化的个人经历，印象和想法。如此说来，一个通俗易懂的“大”字很巧妙地把阳春白雪的科学和萝卜白菜的社会联系在了一起。

大数据的采集和供应者大致分三类：第一类是官方机构，它们有组织地采集高质量的，覆盖全球的格式化的基础数据，通过许可证的办法有偿或者无偿地发布数据；第二类是企业，它们补充一部分基础数据，也通过监控或者反馈系统收集格式化或者半格式

化的客户信息并且按照交易规则发布数据；第三类是个人志愿者，他们随意而无偿地提供半格式化或完全无格式的众包数据和其他网民共享。

大数据所面临的科学挑战和它的三个特征：大，快和多样化有关。

“大”是造成大数据的起因。“大”不等于“好”，也不等于“全”。但我们深信大数据的潜在用途。比如我们可以利用大样本来检验已有的知识，发现新的知识，特别是发现那些存在于大数据中的因果关系。和简单的相关关系不同，因果关系往往只能在时间信息足够的前提下才能得到挖掘。

“快”意味着大数据并非静态，而是动态的数据流。此外，这种数据流并不像河流那样从四面八方汇入大海，而是以网状的结构全方位地流出和流入，随时处在千变万化中。治理这样不断演进的数据流不能单纯地靠提高某一群电脑系统的计算和存储功能，而需要设计新的分布式计算机联网结构来促进增量计算和并行计算。

“多样化”是数据生态系统赖于生存的前提条件，但它却增加了数据质量管理以及开发高效数据挖掘算法的难度。大数据的价值主要不是体现在描述一个个数据节点的内容里，而是节点之间有待挖掘的关系中。但自然科学里常用的量化指标往往不适用于描述人与人，人与物之间的关系。比如：我们可以把任意两条消息链接在一起而并不要求它们之间有任何相似性，再比如：两个用户之间信息传输往来的频率并不能说明他们之间关系的密切程度。目前，我们还缺乏一个描述语义关系的理论基础。有待解答的一个科学问题是：我们如何将物理网络和社会网络结合起来？

现在让我来分析一下“开放”一词。

“开放” 是一把双刃剑。

开放性使我们常常遭遇阴谋论和数字阴影的困扰。任何消息或者关于某个人的数字踪迹都可以轻而易举地被篡改，歪曲和广泛传播。可谓真理不出门，谎言传千里。因此，我们不得不质疑大数据的伦理价值。另一方面，开放性又是最好的对抗那些损害科学和侵犯个人隐私的数据滥用行为的解药。科学工作者有责任设计出具有自我修复能力的开放门户，鼓励广大网民互相监督和更正数据中的缪误。真理早晚会显露，就像油会浮出水面一样。

开放性可以促进数据的可持续获取。当一些机构或者个人出于某种原因失去兴趣，停止提供数据时，新的机构或者其他个人就会取而代之，这样能够保证数据流的畅通和更新。在灾害管理方面，用户互相提供的志愿者数据由于其及时性，灵活性以及拥有关于现场的智能，能够有效缓解 “最需要的时刻和地方却最得不到数据”的尴尬。另一方面，我们却面临多重的数字鸿沟问题。国家体制不同和信息基础设施发展程度不同，造成宏观的数字鸿沟，而年龄，性别以及对计算机技术的熟悉程度不同则造成微观的数字鸿沟。这些鸿沟的存在意味着目前的大数据并不具备代表性，根据这些数据获取的知识里也就必然存在偏差。 但另一方面，开放门户毕竟为我们提供了一个良好的出发点，它能够包容不同的意见，接受广大网民的裁判，任何人在任何时候和任何地方，只要有上网的可能性，都可以加盟。久而久之，数字鸿沟会减小。

总体来说，开放性的利多于弊。以下两个积极的例子也可以说明开放性的好处：

第一个例子是开放科学
越来越多的研究者选择网上免费发表论文，和全球的同行共享实验数据和样本。这样透明的想法和数据交流有助于重复进行实验，获得建设性的补充意见和更正错误，也有助于提高研究者的能见度，最终会提升科学研究的严谨性。一些优秀科学家也越来越愿意在 Wikipedia 这样的门户提供经得起推敲的共享词条。这说明开放平台正在被广大的研究者接纳和认可。

第二个例子是开放地理基础数据
按照那个被反复引用的说法，80%以上的数据有相应的空间地理位置，大地理数据就是大数据中最重要的组成部分，它对于大数据生态环境的建立起着举足轻重的作用。大地理数据是公益数据，从某个分辨率开始，他们正日益成为大众共享的营养汤，服务于各种各样的应用领域。在规范数据基础设施和制订互操作标准方面德国目前在国际上处于领先地位，并拥有未来的竞争优势。

二十一世纪的时代精神不是干涉内政或扩张领土，也不是独立自主，而是相互依存，而且是多层面的相互依存，比如国家之间的相互依存，科学和社会之间的互相依存，自然科学和社会科学之间的互相依存。反映这些依存关系的知识可以部分地从大数据里挖掘出来。

作为大地测量和地图学工作者，我常常感觉到，全球导航卫星和遥感技术给我们留下的印象主要不是征服了太空，而是让我们从一个足够的高度注意到我们共同居住着的这颗蓝色行星有多么脆弱。我们在地球上共享的地形景观比那些分割我们的国界线重要得多。

不错，我们共同生活在一个并不安全的世界。这种不安全既针对物理的世界，也针对数字的世界。然而，我们对待不安全的态度应当有所改变。我们要学会接受不安全，把它视为增强我们免疫力的组成部分，学会有意识地和建设性地同不安全相处。如果我们能够这么思考，我们就会有意义地，而不是过度地规范创新环境以致于使创新思想窒息在越来越厚的条条框框里。

和研究峰会相关的结束语
我也想借此机会直言不讳地评论一下我个人观察和感受到的德国的一个竞争劣势。对安全措施的过度规范已经导致了一些大项目的流产。过度规范所对应的创新文化只能产生增量式创新，而不是突破式创新。

在国际上德国虽然至今仍被誉为一个产生优秀思想的国度。然而，如果不能及时实现或者只在别处实现，思想再多又有何用呢？ 和企业界一样，当今科学界的竞争原则不再是大的打败小的，或者强的打败弱的，而是快的打败慢的。也就是说，宁愿跟上错误，也不能被错误跟上。我在此并不是毫无原则地主张提速，而是想强调，创新需要冒险精神。如果我们一味地为了使风险最小化或者干脆把风险降低到零而过度规范，那么我们还没有开始竞争就已经输了一半。

最后请允许我说说德国高校，我希望高校进一步提高国际化程度和创业精神的培养。最近十年的联邦精英计划大大刺激了高校的竞争精神。联邦政府和州政府共同向高校提供永久性资助的新法规也保证了未来高校的可持续发展空间。但目前全球高校的各

种排名榜上还远远没有体现出德国高校的实际规模和创新水平，我们并没有从全球吸引到足够的优秀人才。主要原因是我们对宣传抱有偏见甚至轻视的态度，我们还缺乏为人才服务的意识，把知识转移狭义地理解成了把知识传授给学生或者转换成实用产品的过程，没有意识到知识转移也是让知识流向社会，特别是在国际社会上用一种知识去碰撞另一种知识的过程。在一个信息碎片化导致关注度不断下降的大数据时代，酒香最怕巷子深。高校管理者不能指望旁人当吹鼓手，不能守株待兔，而应当主动走出去广结联盟，共享资源。请进来和走出去应当成为年轻学者的必修课，只有这样，他们才有可能尽早地在多学科多文化的生活环境里掌握应变能力和自我表达能力，见证一个问题的多种解决方式，明白“它山之石可以攻玉”的道理。

孟立秋