大数据技术发展的十个前沿方向（下）

Ten Fronties for Big Data Technologies (Part C)

吴甘沙，男，现任英特尔中国研究院院长。2000年加入英特尔，先后在编程系统实验室与嵌入式软件实验室承担了技术与管理职位，期间参与或主持的研究项目有受控运行时、XScale微架构、众核架构、数据并行编程及高生产率嵌入设备驱动程序开发工具等。2011年晋升为首席工程师，共同领导了公司的大数据中长期技术规划，主持大数据方面的研究，工作重点为大数据内存分析与数据货币化。在英特尔工作期间，发表了10余篇学术论文，有23项美国专利（10余项成为国际专利），14项专利进入审核期。

doi: 10.11959/j.issn.2096-0271.2015044

10 前沿方向八：精益求精，提升精度

精度是衡量机器学习（人工智能）算法好坏的重要指标。传统上，精度概念包括正确率、识别率、召回率等。在大数据时代，精度被赋予了更多的含义。

10.1 传统机器学习的模型不适应大数据

首先是数据规模。传统的机器学习模型无法支持超大规模的训练数据集，当数据超过一定规模时，传统模型将面临计算量爆炸和精度提升停滞两大难题。关于计算量爆炸，通过分布式优化的方式来加速计算（假设计算资源的扩展能够赶上计算量的增加）或者通过采样、近似等方式减少计算量。关于后者，周志华老师在中国计算机学会大数据学术会议上介绍了一个经典的案例：很多机器学习算法采用AUC（area under ROC curve）作为评估函数，但它需要做两两配对（pairwise）的计算，计算量大且数据无法装入内存，而采用了Least Square Loss函数进行逼近后，极大地减少了计算复杂度。

当然，还有一种选择是采用相对“简单”的模型。无论模型是简单还是复杂，必须具备高可变（high variance）的特性，这样才能随着数据量增大获得精度提升。高可变的“简单”模型虽然计算复杂度不高，但多具有较大的特征空间，更多的数据能够提升其收益。相比而言，“高偏差（high bias）”的简单模型不具有大数据带来的边际效益。

其次是数据的复杂性和维度。高维和非结构化数据（图片、影音等）的爆发推动了复杂模型的广泛应用。无参、非线性、生成性的复杂模型，能够在数据量爆炸时得到可观的性能提升。复杂模型的典型代表就是深度学习，在实际应用中，上百亿参数、20多层的神经网络很常见。

通过对多种简单模型的组合（ensemble）来获得新的复杂模型，这种方式被证明是提升精度的有效方法。当然，选择简单模型组合时尽量要选择相互间相关性弱的模型。另一种混合模型的方式是参数模型和无参模型的组合使用，可以有效处理大规模的高维度数据，特别是在其不同维度的稀疏情况不一致时有奇效：参数模型用于小维度数据，无参模型用于较大维度的数据，两者组合就可以有效地处理大规模、高维度的数据。同样，线性模型和非线性模型也可以采用类似的方法进行组合应用。

随着数据规模和模型复杂度的同时增加，计算量急速增大，优化算法的重要性也日益凸显。在机器学习中，数据、模型以及优化算法都会对最终结果产生重要影响。传统上习惯用非常复杂的算法找到“最优”的答案，但在近年的商业实践中，“足够好”的算法正逐渐成为主流。有时候模型组合可能会导致计算复杂度过高，而缺乏实用性。一个显著的案例就是，Netflix因为数据大赛获奖算法复杂度过高，所以未能将其投入实际应用，而获奖算法恰恰是通过组合多种次优算法的方式得到的。

伴随数据规模和数据维度的爆发，需要探索新的机器学习算法，目标是提升大数据时代下的机器学习应用的边际收益。

10.2 传统的机器学习忽略了数据的长尾特征

互联网的核心价值是解决信息不对称、交易成本苛刻和服务目标覆盖长尾末端的特殊要求。传统的机器学习方法，比如LDA做主题模型，只能发现高频主

题，因为具有指数分布的假设，而指数分布“割掉”了长尾，掩盖了低频的声音和丰富的隐含语义。从互联网经济大潮中成长起来的大数据分析，必须发现长尾人群的微弱声音。因此，大数据分析的首要问题就是优化和强化长尾效应。在新的商业需求下，基于指数假设的机器学习模型（如PCA、LDA、pLSA等）需要演进，比如腾讯公司的Peacock改进了LDA，以适应百万级别的主题。总体来说，改进的办法如下：基于更复杂的模型（如深度神经网络或Google Rephil这样的深度有向无环图模型）、模型的组合以及前文所述的混合模型，更好地探测数据中隐藏的信号。

10.3 精度是一个动态变量

引用海森堡的测不准理论，在大数据的世界里，观测行为会引起被观测现象的改变。很多大数据事实上也是测不准的，比如Google流感的预测，在2013年1月份，Google公司预计的流感情况远高于疾控中心实际测到的数目，所以《科学》和《自然》就对此发表了看法，《自然》认为大数据测不准，《科学》说这是大数据的傲慢。通常说舍恩伯格的大数据三大理念：要全集不要采样；拥抱混杂性、无需精确性；要相关性、不必有因果性。这些理念适合大方向思考，但是分析师不能把它们当作绝对真理。在流感预测案例中，即使是Google公司也把握不到全量的数据。纵然考虑了混杂性，Google公司融合了关键词和疾控中心的数据来调整模型，数据还是不能足够精确。

虽然无法完全解决“测不准”的问题，机器学习算法仍然希望能尽快抓住客观世界的快速变化。因此，在线学习和流式学习是大数据时代重要的办法。大数据的早期表现形式是批处理或离线的数据处理，同样，机器学习的主流方法也是离线训练、在线识别。当机器学习与大数据相遇，训练出来的模型所逼近的是过去的世界，而无法感知世界最新的变化。因此，在线学习或基于增量数据的学习变得非常重要，一边使用生产环境中的旧模型，一边纳入新的数据进行增量训练，快速更新模型并且部署到生产环境，不停顿地完成训练—验证—识别—再训练的闭环。

鉴于此，当前企业纷纷从数据仓库转为EDH（enterprise data hub）或DL（data lake）。因为传统数据仓库基于一个确定的问题定义进行数据的搜集和组织，并且把数据转变成相应的格式保存下来。一旦问题发生变化，再对数据结构做调整太困难、太昂贵。EDH是先把各种原始数据送进来，然后不断地提问题，相应地改变数据表示；不停地尝试更好的分析方法，相应地改变数据结构。

关于模型的选择，统计学大师George Box说：“所有模型都是错的，但是有些是有用的，关键是选择什么样的模型”。

必须指出，大数据不是简单的体量大，它的另一个主要思想是采用多源数据，在数据特征不多的前提下往往传统的简单模型也非常有效。比如常见的流感预测或者票房预测，简单的线性回归模型就能工作得很好。贝叶斯模型在很多场景被广泛地应用。《信号与噪声》的作者Nathan Silver多次成功地预测美国大选结果、奥斯卡获奖者，就是把贝叶斯模型用到了极致，证明了传统技术在大数据时代还有用武之地。

Isaiah Berlin有个比喻，有一种人是刺猬，一招鲜吃遍天，还有一种人是狐狸，一把钥匙开一把锁，以更开放的态度处理分析，选择最合适的模型。在更多的大数据场景中，还是需要根据问题选择合适的模型和方法。

模型的复杂度必须与问题匹配。这就是所谓的奥卡姆剃刀原理，当有多种模型能解释数据时，选择最简单的一个。如前所述，当数据量较小时，模型不能过于复杂，尤其是当模型的特征空间跟训练集规模相仿时，容易形成过拟合。另一方面，如果数据已经非常大，而模型过于简单，那么增加的数据量就无法带来效益的提升。

最后，把分析工作推向社会分工是获得更高精度的方法。如社会化分析平台Kaggle通过竞赛和众包的方式与数十万人的参与，往往能够获得最佳的模型。为了支持大规模的协作分析，学术界在基础设施上已经有所创新。比如DataHub加入了数据的版本控制和多语言支持，美国加州大学伯克利分校教授Joe Hellerstein最近提出，需要一个开放和厂商中立的元数据服务，从而提供跨组织边界协作分析的生产力。

11 前沿方向九：人机的角色变化

在机器学习/人工智能发展历程中，人机角色分工一直朝着使机器更加容易使用、更加广泛应用的目的发展前进，进而把稀缺的专业技能变为大众普及的基础服务。

11.1 机器所承担角色的提升

computer这个词最早出现在19世纪末的哈佛天文台，指一些负责精细计算的人，而现在这个概念已经被机器——计算机取代。人跟机器的关系一直在此消彼长，越来越多的人的职责和能力范围的任务开始由机器完成。传统意义上的数据分析流程，往往基于预先给定的假设和模型，由此出发采集数据样本、细化模型，再用测试样本进行验证，然后修正假设模型，重新开始循环。数据分析应用的成功与否，常常依赖于预定的假设模型，依赖于人。而在大数据场景下，数据是全集的或者接近全集的，预先的假设模型的作用被极大弱化甚至消失；相反，在全集数据中通过机械的数据挖掘穷举所有数据相关性，用相关性来取代主观假设。理想情况下，数据自己找到线索，相关性主动找到应用。

传统意义上的机器学习模型擅长做结构化的数据分析，做语义分析的案例很少。大数据分析能够超越抽象语义的限制。《魔球（Moneyball）》是数据分析与体育（棒球）跨界融合的典型案例，其宣扬的理念是可以用很低的成本找到最合适的队员，获得很高的胜率。然而真实故事里有“不足为外人道”之处，他们花了更多的钱来请高水平球探，对球员的心理能力（如抗压能力和意志力水平）做评估，而这些属性是客观数据无法描述的。这些“球探”代表了超越计算机的领域专业知识。现在机器语义分析能力增强，能够部分取代人的经验推理。

传统上的数据分析和可视化非常依赖分析专家的个人天赋和职业技能，只有少数数据科学家可以直观展示出大量数据中蕴含的各种规律。而领域专家和普通技术人员常常对错综复杂的数据束手无策。最新发展趋势是机器降低人的专业能力门槛的要求。比如MLBase[28]可以自动地找到最好的机器学习模型，VizDeck[29]则通过机器学习找到最好的可视化方式，Scorpion[30]通过可视化自动找出数据中的outliers，并且推知导致outliers的数据记录。

随着工具变得更为人性化，普通从业人员能够更好地从数据中提取价值。以数据可视化为例，出现了很多可视化的工具、库和框架，能够帮助用户专注于数据理解

本身，轻松对各类数据（包括文本、网络/图数据、时空数据和多维数据）进行表现手段丰富的可视化。帮助用户关注数据的内容而摆脱手段的羁绊。同时，数据可视化也从静态展示发展成动态交互过程。一次可视化从对单一视图的展示发展成对新问题的探索过程，形成新的领域——可视化分析，这归功于实时可视化技术的发展。在可视化的界面、交互组件的设计上越来越人性化，实时地、自然地实现多侧面、多分辨率和多焦点的交互。在一些商业领域的决策过程中通过对海量数据进行处理，实现了数据可视化、决策讨论、动作发生、再可视化的闭环式应用。

工具对人的增强更多体现在基础设施层面，尤其是基础设施的社会化对大数据的普及起到了关键的作用。目前，Apache Hadoop的生态环境日趋复杂（由于各个不同组件往往用动物作为吉祥物，业内把Hadoop生态戏称为动物园），大数据基础环境部署的困难常常为人诟病。云计算把数据存储、计算甚至是机器学习的流水线做成了公共基础设施，而创业公司可以不懂分布式计算、容错、Hadoop或Spark，只要具有数据思维，就可以利用云上的基础设施完成数据应用的创新。一些很有创意的创业公司（如Decide.com、Prismatic）开始由几个算法工程师组建而成，而基础设施的事情Amazon公司替他们解决了。Spark的商业化领导者Databricks的愿景是让小数据的使用者很容易切换到大数据，Databricks Cloud正是其实现战略意图的核心。

11.2 人在机器学习过程中的作用

人本来是机器学习中最重要的角色。

机器学习首先需要高质量的数据标记，特别是对于监督学习，其学习的基础是高质量的标记数据。在机器学习这个领域，好的标记数据集能够极大地推动研究的整体进展，比如当前计算机视觉的研究受益于ImageNet。高质量的数据标记需要大量的人力，有时甚至是专业人士。以前的做法是外包至低成本国家，而现在众包（比如土耳其机器人）成为更通用的做法。有一些众包平台开始考虑游戏化机制，比如ESP Game让人们边玩游戏边完成对图像的标记。

无论是外包还是众包，仍然需要人来做。但是一些新的机器学习方法减少了对数据标记的依赖，从而弱化了人的初始作用。深度学习让无监督学习得到了更多的重视，因为它不需要标记数据；半监督学习在过去几年中有了长足的发展，它结合了少量的标记数据和较多的无标记数据；转移学习（transfer learning）采用了举一反三的思想，为另一个目的而标记的数据可以转而用于这个目的，从而弥补相关标记数据的不足。

特征工程（feature engineering）是机器学习中另一个需要大量专业人力的环节。往往模型工作好与坏的关键在于特征的选取，而人的经验非常重要，尤其是一些好的特征（golden feature）依赖于领域知识。所以，一支特征工程团队在项目的早期阶段有很好的效益，但是长期的边际效应越来越趋向于零。现在自动化特征抽取得到了长足的进步，非监督学习具有自动学习特征的能力。在信息维度异常丰富的数据中，可能具备几十亿、几百亿的特征，这是人力穷举无法完成的。深度学习很好地解决了这个问题，它的非监督学习能够逐层提取巨量的特征。有意思的是，这些特征不只是用于深度的神经元网络，还能够作为浅层学习的特征。

机器学习工具越来越易于使用，参与数据分析的人不再是传统意义上的专业

数据科学家、工程师。非专业人士、领域专家越来越能够成为数据分析的主宰者、数据价值的提取者。传统机器学习里面的很多“黑魔法”开始被标准化、具备高易用性的工具取代。而工具发展的趋势是机器学习全流水线。scikit-learn最早做了有益的尝试，通过简单的脚本在一个分析环境中完成端到端的所有工作。后来Spark等主流平台跟进，并且引入了一些新的非常有效的工具，如KeystoneML语言标准化对多种数据类型的处理，Volex对模型的迭代和生命周期进行管理。

11.3 人仍在闭环中（human in the loop）

虽然看到了机器角色的增强和人作用的弱化，但是相信在相当长的一段时间内，人仍将在整个分析闭环中起到重要的作用。

比如，在数据准备（清洗、治理）阶段，人的作用是不可或缺的。现在的很多工具都在如何引入人的干预上做创新，从而保证数据准备的目的是准确的，清洗的程度是合适的，数据表示是符合未来的分析的。

又如，现在虽然可以使用机械的方法发现海量数据中的相关性，但在无数的相关性中发现真正的线索，就需要数据分析师的直觉。直觉就是在潜意识里自动完成的逻辑推理。怎么训练直觉？可以通过学习大量侦探小说和悬疑小说里面的推理过程。这样的推理过程不只是建立模型，还需要数据，则需要很多先验的知识。这些知识怎么来？可以通过广泛的阅读，跨界思想的碰撞，还需要获得上下文的知识，将其融入业务应用中。数据分析师深入业务部门，和业务人员融入到一起，这才能防止数据采集和分析脱钩、数据分析和业务应用脱节，这些过程不能用机器实现。美国加州大学伯克利分校教授郁彬认为数据科学是SDC3（statistics、domain knowledge、computing、collaboration、communication），这里说的正是D和后两个C。此外，communication还反映在分析结果的艺术化展现和精彩的故事讲述将使分析事半功倍。现在分析师所学的内容要从STEM到STEAM，STEM是科学（science）、技术（technology）、工程（engineering）、数学（mathematics），STEAM多出的“A”是艺术（art），这一点上机器短期内很难取代人。另外，艺术不只是这种优雅美观的可视化，还有一个很重要的就是讲故事，有了分析结果之后怎么用更具亲和力的方式表达出来。比如啤酒加尿布的故事，就符合了讲故事的3D：戏剧性（drama）、细节（detail）、参与这个对话的感觉（dialogue）。虽然这个案例是编纂的故事，但是它的易传播性和启发性使得更多人愿意去投入数据分析。前文所述《魔球》也是这样，对故事做适当的加工，用一个精彩的、抑扬起伏的故事讲述数据分析怎么改变棒球运动。这种源于生活、高于生活的拔高是机器望尘莫及的。

另外，人的大规模协作分析或人类计算（human computing），能够完成大量机器所不能完成的任务。

- 在数据库里有个所谓DB-hard的问题，即自然语言表述的不唯一性和歧义性给数据治理带来了挑战，美国加州大学伯克利分校AMBLab的CrowdDB通过众包解决了数据字段规范化的问题。
- 机器学习可以看作模型表示+评价函数+优化方法，而优化方法是寻找最佳模型的必要步骤。Kaggle将企业和科研中海量的数据分析问题与其20万注册数据分析师进行对接，通过悬赏和海选的方法完成了优化过程。
- Duolingo[31]是另一个有趣的案例。

如果Google翻译是集中化的、权威数据主导的分析过程，Duolingo则是社会化、民主化、普通人主导的大规模协作翻译过程，所获得的效果甚至优于Google翻译。在Duolingo平台上，用户学习目标语言过程中必须完成大量的翻译题目，而这些题目来自互联网，因此其学习的过程同时也是对互联网翻译的过程，其惊人的规模效应和积累效用从下例中可见一斑：100万用户通过80 h的学习就能把整个维基百科从英语翻译成为西班牙语。

总之，human-in-the-loop machine learning或active learning已经成为业界的一个热点问题。

12 前沿方向十：智能之争

人工智能在近年成为流行词汇，它代表着生物智能和机器智能的一种博弈。这个博弈的一边是生物智能，生物智能擅长的是模式匹配。人的认知过程就是不停地进行匹配、识别、联想，从记忆中提取数据。而机器智能则是通过计算完成，大量的计算是机器擅长的，比如在大的搜索空间寻找最优解（国际象棋战胜人类世界冠军）、海量信息的检索（沃森电脑在Jeopardy的知识问答中战胜人类）、从计算中总结隐藏的规律等。因此，人工智能也分成了几个派别。

第一个派别认为机器智能并不一定要学习人的生物构造，机器有机器的特点。他们经常引用的一个例子是，当莱特兄弟不试图模仿鸟类的翅膀，而是开始研究空气动力学的时候，人类才有了飞上蓝天的机会。所以机器智能并不一定要学习生物智能，它可以通过更擅长的计算、更完美的数学模型实现智能。这里有很多大师，如统计学大师Michael Jordan、老派的Peter Norvig、新派的邢波。Jordan认为统计是大数据的基础，炒作那些没有数学基础的“新方法”将使大数据进入“寒冬期”。

第二个派别认为必须要了解人脑是怎么工作的。通过各种各样的脑计划绘制出人脑的机理地图，了解人们思维（mind）的工作方式，然后把计算架构往上演进。这个派别有很多生物学家，还有一些老派的科学家，如侯世达（《集异璧》作者）、彭罗斯（数学家，《皇帝的新脑》作者），还有一些民间代表，如雷·库兹韦尔。

第三个派别——计算智能（computational intelligence）方兴未艾。计算智能是上述两个派别之间的折中，他们认为可以用生物认识作为约束和启发，但还是以计算理论为基础来实现智能，比如人工神经网络、演化计算、模糊逻辑、人工免疫系统和群体智能等。人工免疫系统[32]其实就是模仿人体内的分布式免疫系统，即不同位置的淋巴结能够识别不同细菌病毒的特征，从而进行分布式的杀灭。现在主流的神经网络、深度学习科学家都属于这一类。这里不得不提Palm Computer的创始人Jeff Hawkins，他虽然不是科班出身，但赞助和支持了很多有益的工作，他提出的HTM（hierarchical temporal memory）模型[33]得到了美国DARPA-Cortical Processor项目的支持。

下面简略介绍一些现今国际上正在热烈讨论争论的问题。

第一，深度学习是否有可能包打天下？乐观者认为深度学习能够把所有的问题都解决了。以卷积神经网络（CNN）和递归神经网络（RNN）为代表的深度学习技术陆续在计算机视觉、语音识别和自然语言处理方面取得了突破。ImageNet取消object classification的比赛，标志着视觉方面的飞跃；百度公司最新宣布基于

LSTM（long short term memory）和CTC（connectionist temporal classification）的汉语语音识别在安静环境下达到了97%的识别率；而在自然语言方面，深度学习开始把问答和自然语言对话系统作为下一个突破点。

除了上述认知计算领域的进展，深度学习也开始解决人类不能胜任的非认知问题，如百度公司用其提升搜索质量、广告推荐的质量，取得了一定的效果。下一个有望受益于深度学习的是医疗健康领域，从医学影像分析到药物的研发，都可望获得突破。可以说，深度学习虽然一定程度上受到了过多的炒作，但其广泛的应用价值已经确保人工智能的另一个冬天不会到来。

但是质疑者说深度学习没有一个理论基础，缺乏机器学习算法的可解释性，是一些莫名其妙的手段的堆砌。包括Google公司自己发现深度学习可能存在一些内在缺陷[34]，比如两张图片人眼看起来是完全一模一样的，其中有一些细微的像素差别，但是深度学习只能认出一张，不能认出另外一张。为此，现在深度学习的大师们正在试图发展出一些理论，尝试从计算理论、生物隐喻上解释。比如Google公司的Geoffrey Hinton，提出了胶囊理论（capsules theory），模仿人类大脑中的皮质柱，如果将人的大脑皮质想象成一个有6层细胞厚度的皮层，它是由一个个圆柱体构成的。他希望用这个隐喻来改进深度学习每一层完全非结构化的问题，把每一层的神经元进行分组、功能化。另外，学术界开始探讨如何解决深度学习的知识表示问题。

第二，智能的未来是否一定就是类脑计算？目前这一领域的进展主要在两个方面：一方面是通过脑计划绘制大脑的数字机理地图，通过对思维的研究、对记忆的研究进一步了解人脑工作机制；另一方面是人工神经网络和Sparse Coding等“大脑启发计算（brain inspired computing）”技术的不断改进。比如反馈，人脑在从输入到结果的过程中，前向连接是后向连接（从处理到输入）的十分之一，也就意味着回路是前向连接的10倍之多。现在的人工神经网络还是前向多、回路少（即使有回路，如反向传播算法，也只发生在训练阶段）。另外，要增加时间因素。现在的很多人工神经网络没有时间因素，但是人是不断地在学习，其所见所想是有时间因素的，因此需要在线学习能力的提升。

第三，是否需要发明专为类脑计算的计算架构。人工智能研究的先驱Hans Moravec曾经提出Moravec Paradox：成年人才能做的高阶任务（如推理和规划），现有的计算架构绰绰有余；而一两岁孩童就运用娴熟的低阶任务（如感知和协调运动），需要的计算能力远远超过了冯诺依曼架构的能力。举一个未必确切的比喻：天河2号1 800万瓦，5亿亿次（浮点）计算每秒，而人脑据估计是10亿亿次操作每秒，只耗电20 W（每天只需100多毫克的葡萄糖）。因此，针对特定的负载，人们希望能够实现低功耗的具有识别、联想、推理能力的新计算架构。新的架构也有不同路线：一类是传统人工神经网络的加速器，如中国科学院计算技术研究所的电脑、大电脑、普电脑，Yann LeCun的NeuFlow；另一类是更接近生物神经网络的处理器，被称为神经拟态（neuromorphic）架构，如IBM公司的TrueNorth、高通公司的Zeroth。前者的识别精度高，但没有在线学习能力；后者目前精度低，但能够在线学习，也许未来有不错的前景。当然，目前来说，所有这些架构都面临可编程性差的问题，因此，在较近的一段时间内，FPGA、GPU和众核可能是

更实用的计算架构。

所有这些问题是当前在智能之争上面讨论的问题。

13 结束语

目前来看，协作、开放的计算机科学（collaborative open computer science）已成为当今世界的主流。大数据在所有热门技术中具有最开放的技术生态，开源框架（如Theano、PyLearn2和Caffe）极大地加速了深度学习的普及，未来像GitXiv这样集合GitHub（开放源代码）、arXiv（公开研究方法）以及学术论坛的平台，将极大地促进计算机科学的发展。

英特尔公司一直在推动开放、协作的创新，资助、跟踪大学的研究，注重在10个前沿方向推动技术的发展。英特尔公司在全球范围内与大学有多个联合研究项目，在美国有9个研究中心，世界范围内有多家（包括在中国与清华大学、东南大学和中国科学技术大学联合建立的移动网络和计算英特尔协作创新中心，专注于5G网络和计算研究）。其中，一些大数据研究中心取得了很好的成绩。例如，卡内基梅隆的云计算中心，Spark是该中心早期自主的项目（研究主体在美国伯克利），还贡献了GraphLab、Petuum。在MIT的大数据中心的领导者之一就是新科图灵奖得主Michael Stonebraker，MIT中心的很多工作围绕新一代的DBMS，如内存数据库H-Store、流数据库S-Store、科学计算数据库SciDB、原位计算可视化、支持协作分析的DataHub等。美国斯坦福的大数据中心主要做可视化，由Pat Hanrahan教授领导，他是Tableau的创始人之一。还有，以色列的计算智能中心，对深度学习有很多贡献。这些中心的很多工作已经开源。

英特尔公司希望能够通过这些协作研究，了解大数据发展的前沿。同时，也能够使英特尔的架构更好地跟随大数据算法和系统的发展。

参考文献

[28] Pan X H, Sparks E R, Wibisono A. MLbase: Distributed Machine Learning Made Easy. Dept. Computer Science, UC Berkeley, 2013

[29] Key A, Howe B, Perry D, *et al*. Vizdeck: self-organizing dashboards for visual analytics. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, Scottsdale, USA, 2012

[30] Wu E, Madden S. Scorpion: explaining away outliers in aggregate queries. Proceedings of the VLDB Endowment, 2013, 6(8)

[31] Luis V A. Duolingo: learn a language for free while helping to translate the web. Proceedings of the 2013 International Conference on Intelligent User Interface, Santa Monica, USA, 2013

[32] Hofmeyr S A, Forrest S A. Architecture for an artificial immune system. Evolutionary Computation, 2000, 8(4): 443~473

[33] Hawkins J, George D. Hierarchical Temporal Memory Mdash; Concepts, Theory and Terminology. Numenta Inc, 2006

[34] Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. Proceedings of International Conference on Learning Representations, Banff, Canada, 2014 □