数据科学

鄂维南

数据科学的基本内容

大数据时代在科学领域里的表现是数据科学的兴起。常常听到有人问：多大才算是“大数据”？ “大数据”和“海量数据”有什么区别？其实根本没有必要为“大数据”这个名词的确切含义而纠结。“大数据”是一个热点名词。 它代表的是一种潮流、一个时代。它可以有多方面的含义。“海量数据”是一个技术名词。它强调数据量之大。 而数据科学则是一个学科、 一门新兴的学科。

为什么要强调数据科学？ 它和已有的信息科学、统计学、机器学习等学科有什么不一样？

作为一门学科， 数据科学所依赖的两个因素是数据的广泛性和多样性， 以及数据研究的共性。 现代社会的各行各业都充满了数据。 而且这些数据也是多种多样，不仅包括传统的结构型数据， 也包括象网页、文本、图像、视频、语音等非结构型数据。正如我们后面将要讨论到的， 数据分析本质上都是在解反问题， 而且是随机模型的反问题。所以对它们的研究有着很多的共性。比方说自然语言处理和生物大分子模型里都用到隐式马氏过程和动态规划方法。其最根本的原因是它们处理的都是一维的随机信号。 再如图像处理和统计学习中都用到的正则化方法， 也是处理反问题的数学模型中最常用的一种手段。所以用于图像处理的算法和用于压缩感知的算法有着许多共同之处。 这在新加坡国立大学沈佐伟教授的工作中就可以很明显地看出来。

除了新兴的学科如计算广告学之外， 数据科学主要包括两个方面： 用数据的方法来研究科学和用科学的方法来研究数据。 前者包括象生物信息学、 天体

信息学、数字地球等领域。 后者包括统计学、机器学习、数据挖掘、数据库等领域。这些学科都是数据科学的重要组成部分。但只有把它们有机地放在一起，才能形成整个数据科学的全貌。

用数据的方法来研究科学，最典型的例子是开普勒关于行星运动的三大定律。

开普勒的三大定律是根据他的前任，一位叫第谷的天文学家留给他的观察数据总结出来的。表9-1是一个典型的例子。这里列出的数据是行星绕太阳一周所需要的时间（以年为单位），和行星离太阳的平均距离（以地球与太阳的平均距离为单位）。从这组数据可以看出，行星绕太阳运行的周期的平方和行星离太阳的平均距离的立方成正比。这就是开普勒的第三定律。

行星	周期（年）	平均距离	周期^2/距离^3
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00

表9-1：太阳系八大行星绕太阳运动的数据

开普勒虽然总结出他的三大定律，但他并不理解其内涵。牛顿则不然。牛顿用他的第二定律和万有引力定律把行星运动归结成一个纯粹的数学问题，即一个常微分方程组。如果忽略行星之间的相互作用，那么这就成了一个两体问题。因此很容易求出这个常微分方程组的解，并由此推出开普勒的三大定律。

牛顿运用的是寻求基本原理的方法，它远比开普勒的方法深刻。牛顿不仅知其然，而且知其所以然。所以牛顿开创的寻求基本原理的方法成了科学研究的首选模式。这种方法在上个世纪初期达到了顶峰：在它的指导下，物理学家们发现了量子力学。原则上来讲，我们日常生活中所碰到的自然现象都可以从量子力学出发得到解决。量子力学提供了研究化学、材料科学、工程科学、生命科学等几乎所有自然和工程学科的基本原理。这应该说是很成功，但事情远非这么简单。正如狄拉克指出的那样，如果以量子力学的基本原理为出发点去解决这些问题，那么其中的数学问题太困难了。所以如果要想有进展，还是必须做妥协，也就是说要对基本原理作近似。

再举另外一个例子，表9-2中形象地描述了一组人类基因组的SNP数据（Single Nucleotide Polymorphism data）。一组研究人员在全世界挑选出1064个志愿者，并把他们的SNP数据数字化，也就是把每个位置上可能出现的10种碱基对用数字来代表，对这组数据作主组分分析，就可以得到图9-1中的结果。其中横轴和纵轴代表的是第一和第二奇异值所对应的特征向量。这些向量一共有1064个分量，对应1064个志愿者。值得注意的是这组点的颜色所代表的意义。可以看出，人类进化的过程可以从这组数据中通过最常见的统计分析的方法，即主组分分析，而展示出来。

主组分分析是一种最简单的数据分析方法。它的做法是对数据的协方差矩阵作对角分解。

	SNP1	SNP2	……	SNPm
志愿者1	0	1	……	0
志愿者2	0	2	……	1
志愿者3				
.				
.				
.				
志愿者n	1	9	……	1

表9-2　SNP数据的示意图：n=1064，m=644258，0，1，… ，9分别代表碱基对是AA，AC，CC，… 。参见：Jun Z. Li et al, "Worldwide human relationships inferred from genome-wide patterns of variation", Science, 22, Feburary, 2008.

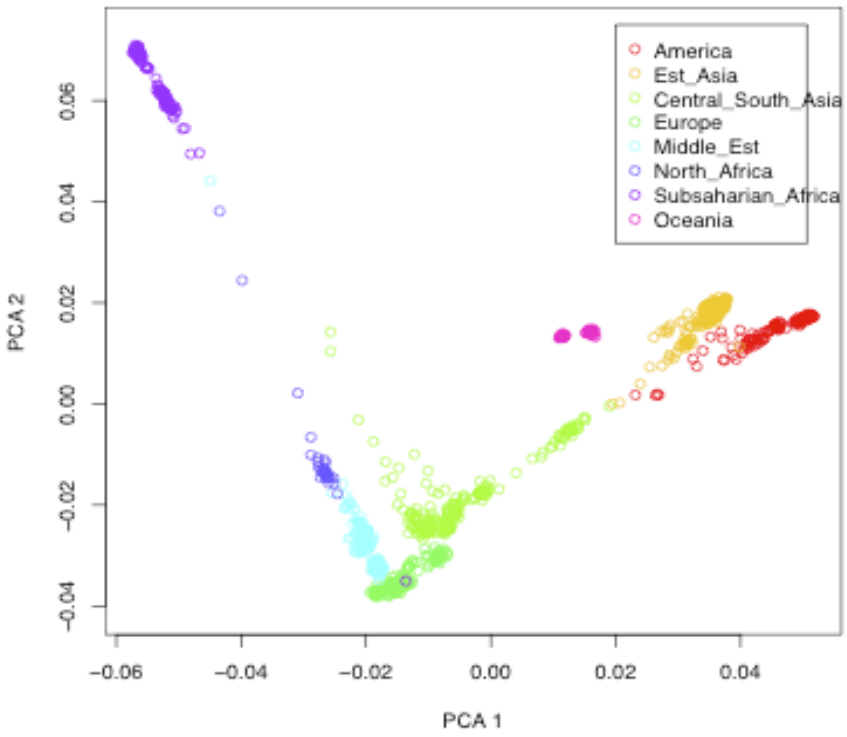

图9－1： 对SNP数据作主主份分析的结果告诉我们人类进化的过程。

这里横轴和纵轴分别表示最大奇异值和第二大奇异值所对应的特征向量。

此结果系姚远等根据 Jun Z. Li等人文章中的结果重新制作。

这样的问题，如果采用从基本原理出发的牛顿模式，则基本上是没法解决的。而基于数据的开普勒模式则是行之有效。尽管牛顿模式很深刻，但对复杂的问题，开普勒模式往往更有效。开普勒模式最成功的例子是生物信息学和人类基因组工程。正是因为它们的成功，材料基因组工程等类似的项目也被提上了议事日程。同样，天体信息学、计算社会学等等也成了热门学科。这些都是用数据的方法来研究科学问题的例子。图像处理是另外一个典型的例子。图像处理是否成功是由人的视觉系统决定的。所以要从根本上解决图像处理的问题，就需要从理解人的视觉系统着手，并了解不同质量的图像，对人的视觉系统产生什么样的影响。这

样的理解当然很深刻，而且也许是我们最终所需要的。但从目前来看，它过于困难也过于复杂。解决很多实际问题时，并不需要它。而是一些更为简单的数学模型就足够了。

用数据的方法来研究科学问题，并不意味着就不需要模型。只是模型的出发点不一样，不是从基本原理的角度去找模型。就拿图像处理的例子来说，基于基本原理的模型需要描述人的视觉系统以及它与图像之间的关系。而通常的方法则可以是基于更为简单的数学模型，如函数逼近的模型。

怎样用科学的方法来研究数据？这包括以下几个方面的内容：数据的获取，存储，和数据的分析。下面我们将主要讨论数据的分析。

数据分析的中心问题

比较常见的数据有以下几类：

（1） 表格。这是最为经典的数据。

（2） 点集（point cloud）。很多数据都可以看成是某种空间的一堆点。

（3） 时间序列。文本，通话，DNA序列等都可以看成是时间序列。它们也是一个变量（通常可以看成是时间）的函数。

（4） 图像。可以看成是两个变量的函数。

（5） 视频。时间和空间坐标的函数。

（6） 网页，报纸等。虽然网页或报纸上的每篇文章都可以看成是时间序列，但整个网页或报纸又具有空间结构。

（7） 网络数据。

还可以考虑更高层次的数据，如图像集，时间序列集，表格序列等等。

数据分析的基本假设就是观察到的数据都是由背后的一个模型产生的。数据分析的基本问题就是找出这个模型。由于数据采集过程中不可避免地会引入噪声，通常这些模型都是随机模型。

数据类型	模型
点集	概率分布
时间序列	随机过程（如隐式马氏过程等）
图像	随机场（如吉布斯随机场）
网络	图模型，贝叶斯模型

表9－3. 常见的数学模型。

当然，在大部分情况下，我们并不感兴趣整个模型，而只是希望找到模型的一部分内容，如：

（1） 相关性。判断两组数据是不是相关的。

（2） 排序。比方说对网页作排序。

（3） 分类、聚类。把数据分成几类。

很多情况下，我们还需要对随机模型作近似。最常见的是把随机模型近似为确定型模型。所有的回归模型都采用了这样的近似。基于变分原理的图像处理模型也采用了同样的近似。另一类方法是对其分布作近似，例如假设概率密度是正态分布，或假设时间序列是马尔可夫链等等。

分析数据的第一步是赋予数据一定的数学结构。这种结构包括：

（1）度量结构。在数据集上引进度量，也就是距离，使之成为一个度量空间。文本处理中的余弦距离函数就是一个典型的例子。

（2）网络结构。有些数据本身就具有网络结构，如社交网络。有些数据本身没有网络结构，但可以附加上一个网络结构。比方说度量空间的点集，我们可以根据点与点之间的距离来决定是否把两个点连接起来，这样就得到一个网络结构。

（3）代数结构。比方说我们可以把数据看成是向量，或矩阵，或更高阶的张量。有些数据集具有隐含的对称性。这也可以用代数的方法表达出来。

在这基础上，我们可以问更进一步的问题。例如：

（1）拓扑结构。从不同的尺度去看数据集，得到的拓扑结构可能是不一样的。最著名的例子是3×3的自然图像数据集里面隐含着一个2维的克莱因瓶。(参见： Robert Ghrist, BARCODES: THE PERSISTENT TOPOLOGY OF DATA, BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY, Volume 45, Number 1, January 2008, Pages 61－75).

（2）函数结构。尤其对点集而言，寻找其中的函数结构是统计学的基本问题。这里的函数结构包括：线性函数，用于线性回归；分片常数，用于聚类或分类；分片多项式，如样条函数；其他函数如小波展开等。

数据分析的主要困难

我们碰到的数据通常有这样几个特点。一是数据量大。大家只要想一想，万维网上有多少网页，这些网页上有多少数据，就可以对现在碰到的数据量之大有

点感觉了。第二是维数高。前面提到的SNP数据是64万维的。第三是类型复杂，比方说这些数据可以是网页或报纸，也可以是图像，视频。第四是噪音大。

这里面最核心的困难是维数高。维数高给我们带来的是维数诅咒（curse of dimension）：模型的复杂度和计算量随着维数的增加而指数增长。例如非参数化的模型中参数的个数会随着维数的增加而指数增长。

怎样克服维数高带来的困难？通常有两类方法。一类方法就是将数学模型限制在一个极小的特殊类里面，如线性模型，如假设概率密度遵循正态分布，如假设观测到的时间序列是隐式马氏过程等。另一类方法是利用数据可能有的特殊结构，例如稀疏性，低维或低秩，光滑性等等。这些特性可以通过对模型作适当的正则化而实现。当然，降维方法也是主要方法之一。

总而言之，数据分析本质上是一个反问题。因此，处理反问题的许多想法，如正则化，在数据分析中扮演了很重要的角色。这也正是统计学与统计力学的不同之处。统计力学处理的是正问题，统计学处理的是反问题。

算法的重要性

跟模型相辅相成的是算法以及这些算法在计算机上的实现。特别是在数据量很大的情况下，算法的重要性就显得尤为突出。

从算法的角度来看，处理大数据主要有两条思路。

一是降低算法的复杂度，即计算量。通常我们要求算法的计算量是线性标度的，也就是说计算量跟数据量成线性关系。但很多关键的算法，尤其是优化方法，还达不到这个要求。对特别大的数据集，例如说万维网上的数据或社交网络数据，我们希望能有次线性标度的算法，也就是说计算量远小于数据量。这就要求我们

采用抽样的方法。但怎样对这样的数据进行抽样，比方说对社交网络进行抽样，仍还是一个未解决的问题。

第二条思路是云计算，或并行计算，它的基本想法是把一个大问题分解成很多小问题，然后分而治之。著名的MapReduce软件就是一个这样的例子。

下面举几个典型的算法方面的例子。这些例子来自于2006年IEEE国际数据挖掘会议所选举出来的数据挖掘领域中的10个最重要的算法。

（1）k-平均（k-means）方法。这是对数据作聚类的最简单有效的方法。

（2）支持向量机： 一种基于变分（或优化）模型的分类算法。

（3）期望最大化（EM）算法。这个算法的应用很广，典型的是基于极大似然方法（ maximum likelihood）的参数估计。

（4）谷歌的网页排序算法，PageRank。它的基本想法是：网页的排序应该是由网页在整个互联网中的重要性决定。从而把排序问题转换成一个矩阵的特征值问题。

（5）贝叶斯方法。这是概率模型中最一般的迭代法框架之一。它告诉我们怎样从一个先验的概率密度模型，结合已知的数据来得到一个后验的概率密度模型。

（6）k-最近邻域方法。用邻域的信息来作分类。跟支持向量机相比，这种方法侧重局部的信息。支持向量机则更侧重整体的趋势。

（7）AdaBoost。这个方法通过变换权重，重新运用数据的办法，把一个弱分类器变成一个强分类器。

其它的方法如决策树方法和用于市场分析的Apriori算法，以及用于推荐系统的合作过滤方法，等。

就现阶段而言，对算法的研究被分散在两个基本不相往来的领域里：计算数学和计算机科学。计算数学研究的算法基本上是针对像函数这样的连续结构。其主要的应用对象是微分方程等。计算机科学处理的主要是离散结构，如网络。而数据的特点介于两者之间。数据本身当然是离散的。但往往数据的背后有一个连续的模型。所以要发展针对数据的算法，就必须把计算数学和计算机科学研究的算法有效地结合起来。

对学科发展的影响

回到本章的主题，数据科学对学科发展提供了前所未有的机遇和挑战。要充分利用好这个机会，我们就必须建立起一套新的科学和教育体系。在大学的层面，要赋予数据科学其应有的地位，建立起跨学科，全方位的数据科学研究平台；进一步完善和企业合作创新的机制；培养适应学术界和企业界需求的数据科学人才。

数据科学也将对许多传统学科的发展带来极大的影响。首先是对数学。数学的发展主要来自两个方面的推动力：一是来自数学内部，学科自身的完善带来的推动；二是来自外部，由其它学科，社会或工业发展的需要而带来的推动。就目前的现状而言，第一方面的推动力对数学的影响要远远超过第二方面的推动力。这样造成的结果是，一方面，数学作为一门学科，其重要性已经得到广泛的认可。而另一方面，数学家作为一个群体，其对社会和科学整体发展的影响却难以得到承认。在很多学校以及在整个科学界，数学家这个群体正显得越来越孤立。这就是为什么数学家们经常发现自己处在一个很尴尬的位置。这是一件极为不幸的事情。它不仅大大影响了数学的发展，更是影响其它学科、技术乃至社会的发展。

事实上，至少在理论研究方面，很多学科的瓶颈问题都是数学问题。这在近一百年前狄拉克就已经指出来了（参见前文）。所以在很多学科里，我们看见的都是非数学出生的科学家在进行数学方面的研究。

数学家们为什么不擅于帮助解决其它学科的问题呢？在自然科学领域，有一个基本的原因，那就是要解决自然科学的问题，首先要有基本原理，也就是通常所说的模型。我们把它们叫做数学模型。但实际上这些模型都是来自于物理学的基本原理。对数学家们来说，这是一个基本障碍。

数据科学不一样，如前所述，数据科学的基本原理本身就来自于数学。所以数据科学在数学和实际应用之间建立起了一个直接的桥梁。而这些实际应用正是来自于象信息服务等现代产业中最为活跃的一部分。这对数学来说，实在是一个千载难逢的机会。

不仅如此，数据的分析几乎涉及到了现代数学的所有分支。甚至于像表示论这样的极其抽象的分支，在数据的领域也有其发挥作用的余地。所以数据科学对数学的要求和推动是全面的，而不是仅仅局限在几个领域。数据应该成为数、图形和方程之外数学研究的基本对象之一。

数据科学对计算机科学的发展也会带来很大的影响。图灵奖得主John Hopcroft 曾经指出，在过去的几十年里，计算机科学的研究对象主要是计算机本身，包括硬件和软件。 以后计算机科学的发展将主要围绕着应用展开。而从计算机科学自身来看，这些应用领域提供的主要研究对象就是数据。虽然计算机科学一贯重视数据的研究，但数据在其中的地位将会得到更进一步的加强。

再看统计。统计一直就是一门研究数据的学科。所以它也是数据科学最核心的部分之一。但在数据科学的框架之下，统计的发展也会受到很大的冲击。这种

冲击至少表现在两个方面。一是关于数据的模型将会跳出传统的统计模型的框架。更一般的数学概念，如拓扑、几何和随机场的概念将会在数据分析中扮演重要的角色。二是算法和计算机上的实现将成为研究的中心课题之一，这在前面已经讨论过，这里不再重复。

应该说，在很长的一段时间里，统计这门学科没有受到足够的重视。普林斯顿大学还取消了统计系。近年来，学术界和应用领域都已经逐渐地认识到统计的重要性。许多学校都有计划要发展统计，但苦于难以吸引到高质量的统计人才。如果把视野拓宽一点，我们就会发现，发展数据科学则是更加有利的做法：因为它既更加适应未来的需要，又能尽快地把应用数学、计算数学和计算机科学等学科中的有生力量调动起来以开展工作。

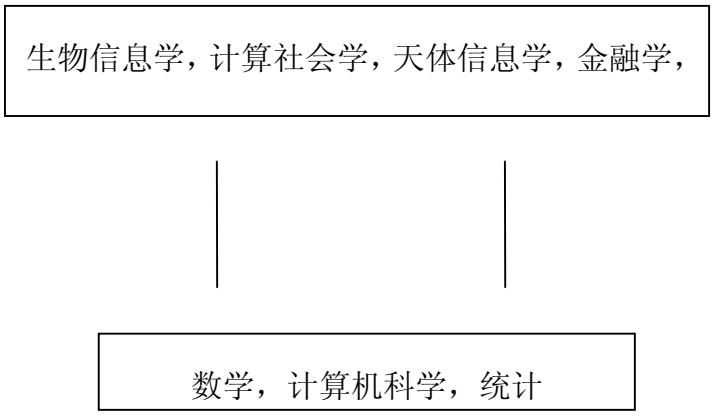

表9－4．数据科学的学科框架。

对传统学科的冲击

这里我们举两个例子。一个是社会学的例子。作为社会科学的一个分支，社会学一直是一门基于数据的学科。大到国家和社会层面的数据，小到家庭和个人

的数据，这些是社会学研究的基本资料。从这个角度来看，社会学和数据之间的关系不是什么新的现象。但即便如此，数据科学的兴起仍然对社会学的研究有着巨大的影响。这至少表现在如下几个分面。

一是社交网络的产生和网络科学的研究为社会学带来了一个新的研究层面，即介观层面。这不仅给社会学提供了新的研究方向，而且也给社会学的研究提供了新的实用价值，如信息传播，广告投放，热点分析等。

二是使社会学的研究进一步量化、去经验化。在过去很长的时间里，由于数据的稀缺，社会学在很大程度上是一门经验科学。大量数据资源的获取为社会学的更进一步量化提供了可靠的途径。

三是更多更加严密和系统的科学方法被引进到社会学的研究中，如数据采集的方法。北京大学中国社会调查中心所开展的家庭访问调查就是一个很好的例子。他们不但注重调查中问答的结果，同时也记录了调查过程的数据。这样严密的科学方法一定会给社会学的研究带来极大的影响。

在人们眼里，社会学往往不是一门技术型的或实用型的学科。但随着社会学的进一步的量化，人们对社会学的看法将会发生很大的变化。在不远的未来，社会学的研究将对产品推销，信息传播和舆情预警等实用领域产生深刻的影响。

我们要谈的第二个例子是语言学。跟社会学一样，语言学在历史上也是一个离实用技术比较远的学科。但近年来蓬勃发展起来的机器翻译，自然语言处理，语言识别，文本分析等技术给语言学的实际应用提供了一个绝好的机会。但值得注意的是，在所有这些领域，基于概率模型的处理方法的有效性远远超过了基于文法的处理方法的有效性。这对传统的语言学来说，不能不说是一个非常令人失望的结果。

在麻省理工学院成立一百五十周年的一个纪念会上，当代语言学的奠基人乔姆斯基教授针对这一问题提出了他的看法。他认为概率模型的成功是有限的，而且其成功只是仅仅局限于逼近未被分析的数据这一方面。他的言下之意是说概率模型只是技术上的成功，不能算作是传统科学意义上的成功，因为它没有给传统的语言学问题如文法问题，带来新的认识。应该说，这种看法是比较保守的。按照这种逻辑，生物信息学也只是工程上的成功，不是科学意义上的成功。按照前文的说法，自然语言的概率模型可以看成是一种开普勒模式的做法。而乔姆斯基只认可牛顿模式。科学发展的历史已经告诉我们，这两种模式都十分重要。而具体到语言学来说，承认并认真应对概率模型的成功才是真正可取的方法。

新学科的诞生：计算广告学

广告有着十分悠久的历史。但它一直都很难算得上是一门科学。尤其是在中国，由于管理上的漏洞，最典型的广告，就是在媒体上，特别是在电视上，由各种各样的明星说上几句不负责任的话。近年来，由于雅虎，谷歌等搜索引擎选择商业广告作为其主要赢利模式，一门新的学科，计算广告学，由此而诞生。

计算广告学所处理的主要问题是怎样有针对性地投放广告。互联网上的广告有两个最基本的指标：点击率和转换率。点击率是广告被点击的概率。转换率是广告被点击以后引起商品成交的概率。由于后者更难估计，所以互联网上的广告往往以点击率作为主要指标。这就要求我们根据用户提供的信息，比方说其所输入的关键词，预测不同广告的点击率。这是计算广告学的一个基本问题。解决这个问题的主要想法就是构造一个utility函数来估计用户对不同广告感兴趣的程度。

目前像斯坦福大学，加州大学伯克利分校等重要学校都已开设了计算广告学这门课。美国国家基金委所属的几个数学研究所之一，地处北卡州的统计与应用数学研究所也针对计算广告学举办了专题研讨会。

“科学能从谷歌那儿学到什么？”

这是2008年美国“连线”杂志（Wired Magzine）主编安德森在他的一篇评论文章（The end of theory: The data deluge makes the scientific method obsolete, Wired Magazine, 06.23.08）结尾时的问话。的确，谷歌不仅仅是信息产业界成功的典范，同时还是数据科学领域的先锋和开拓者。谷歌的成长史是一部创新和开拓的历史。

谷歌的起步是关于网页搜索排序的新概念和算法。谷歌之前，已经有了其它的搜索引擎，最著名的是雅虎。但所有这些引擎都没有解决好对搜索结果作排序的问题。佩奇和布林的想法是把网络的结构利用起来。事实上每个网页都是互联网上的一个节点。他们不是孤立的，不同的网页之间通过超链接联系在一起。如果一个网页有很多超链接指向它，就说明它具有权威性，应该排在前面。怎样给网页的权威性一个定量的刻划呢？ 设想一个醉汉在互联网上作随机游动，他访问得最多的网页就最具有权威性。这样就可以把网页排序的问题描述成为一个由互联网结构而派生出来的马氏链的不变测度的问题，也就是一个转移矩阵的特征值问题。这就是佩奇关于网页排序的基本想法。通过这种想法，佩奇和布林大大提高了互联网搜索结果的质量。

谷歌也是第一个将云计算由概念变为现实的企业。不言而喻，谷歌从一开始就需要处理大量的网页。它最初开发云计算的目的是建立一个能把大量的廉价服

务器集合在一起，以完成大型计算和存储的功能。这个平台必须是可扩展的，并行的，并且允许其中一些服务器出现故障。为了达到这一目的，谷歌开发了一系列的新技术和新的数据存储模式，其中包括谷歌文件系统（Google File System），MapReduce等。这些新概念和新技术已成为大数据处理的标准方法。与此同时，谷歌也建立起了面向未来的数据中心和云计算平台。这些基础设施使得谷歌在信息服务产业高居着一个得天独厚的位置。

谷歌之所以能做到这些，最根本的一点是它高瞻远瞩的眼光和胸怀。谷歌创始人佩奇和布林认识到，谷歌的根本利益在于互联网能否成为普通大众生活中必不可少的工具。做好了这一点，谷歌的商业利益就自然而然地来了。为做到这一点，谷歌坚持了由雅虎开创的互联网免费的原则。这个原则对互联网的普及起到了最为关键的作用。

事实上，谷歌的商业模式也是可圈可点的。它的赢利是靠互联网广告，而不是靠对用户的收费。在谷歌之前，Overture公司就已经在开展互联网广告业务，但谷歌把互联网广告推到了更高的层次。谷歌开发的Adwords系统，是计算广告学最早的实践典范。

互联网是一个极大的资源，一个由全世界的亿万网民共同构建的资源。而谷歌这样的公司，通过构建一系列新的概念和技术平台，十分有效地把这些资源变成了他们自己的资源。而在此同时，又给全世界的网民提供了十分有益的服务。谷歌的例子，是创新和产业发展密切结合、相互推动最成功的例子。

数据科学的教育体系

在数据科学领域里工作的人才需要具备两方面的素质：一是概念性的，主要是对模型的理解和运用；二是实践性的，主要是处理实际数据的能力。培养这样的人才，需要数学、统计和计算机科学等学科之间的密切合作，同时也需要和产业界或其他拥有数据的部门之间的合作。目前还没有任何一所高校具有这样的平台。

数据科学的教育体系应该包括如下几方面的内容：

（1） 数学的基础知识。除了微积分、线性代数和概率论这三大基础中的基础以外，还需要随机过程、函数逼近论、图论、拓扑学、几何、变分法、群论等方面的基础知识。目前，可能还不是所有人都能看到这些内容跟数据的直接关系。但随着数据科学的不断深入发展，他们的作用会越来越明显。这些内容也不需要一门一门地教。数学系应该开出一些新的“高等数学”课程来覆盖这些方面的内容。

（2） 计算机科学的基本知识，如计算机语言、数据库、数据结构、可视化技术等。

（3） 算法方面的基本知识，包括数值代数、函数逼近、优化、蒙特卡洛方法、网络算法、计算几何等等。

（4） 数据的模型，如回归、分类、聚类、参数估计等。

（5） 专业课程，如图像处理、时间序列分析、视频处理、自然语言处理、文本处理、语言识别、图像识别、推荐系统等等。

（6） 其它专业课如生物信息学、天体信息学、金融数据分析等等。

这里（1）-（4）属于基础课，（5）-（6）属于专业课。专业课的设置还可以跟企业界合作，以满足不断变化着的实际需求。与企业界的合作也更有利于向企业界输送合适的人才。

本章结语

大数据给科学和教育事业的发展提供了前所未有的机会，同时也提出了前所未有的挑战。它将对现有的科研和教学体制带来大幅度的变革，对科学与产业之间的关系、科学与社会之间的关系带来大幅度的变革。总结一下，大数据的影响将主要来自以下几个方面。

首先是数据科学将成为科研体系中的重要部分，并逐渐达到与包括物理、化学、生命科学等学科在内的自然科学分庭抗礼的地位。未来的科研和教育体制应该由两条主线组成：一条是以基本原理为主线。现在的物理学、化学、机械工程等学科，以及生命科学、材料科学、天体物理、地球科学等学科的大部分都是沿着这样一条主线展开的。另一条是以数据为主线。它包括统计学、数据挖掘和机器学习、生物信息学、天体信息学、以及许多社会科学的学科。它还包括一些新兴的学科，如计算广告学。数据科学的兴起，将极大地推动许多社会科学学科朝着量化的方向发展，使他们逐步由经验性的模式转变成科学性的模式。

其次是科学研究和市场、和产业的联系将变得更加密切，从发现基本原理到产业化的周期将会被大大地缩短。这可以从谷歌的例子看出来。谷歌的发展，从搜索引擎的一个概念和算法上的突破到进入市场、变成产业，只经过了短短几年的时间。这样的例子在数据科学和信息产业并不陌生。但在传统的自然科学领域，从基本原理的突破，到技术、到产业，往往要经过一个漫长的过程。

再次，数据的主要来源之一是社会，如互联网、社交网络、公共交通、智慧城市等等。所以数据科学的研究与我们的日常生活、与社会有着密切的联系。比方说，谷歌和百度的网络搜索算法就对我们的日常生活产生了很大的影响。所以人们日常生活中的需要以及社会的需要将成为数据科学的主要问题来源之一。

科学研究最重要的一环是提出前瞻性的问题。提不出问题，就只能跟在别人后面，走一条从文献到文献的路子。对我国的科技界来讲，在很多学科，由于来自实际应用领域的限制，提出前瞻性问题的确是件很困难的事情。但数据科学则不然。由于我国人口众多这一特殊情况，和我们特殊的文化、文字、历史背景和社会发展的需要，我们在数据科学领域的很多问题自然就是前瞻性的。关键是我们能否用前瞻性的方法去面对这些问题。如果做好了这一点，我们在数据科学领域就自然而然地走到了世界的前沿。