10.2002
PP. 54-56

共进专栏：心中有数

（美国）林共进／文

处理大型资料的迷思

大型资料愈来愈普遍，此一现象并非偶然，而是必然。以下分别由宇宙发展、生物演化及人类活动等方面简单加以阐述大型资料形成的必然性。这部分没有严谨的学术证明，仅系个人观察的结果，希望能带给大家一些思考。

首先宇宙的发展本来就是愈来愈大，所以期间产生的资料当然也就愈来愈多，而这个"大"与"多"是远超乎我们想象的。根据当今比较得到认可的天文理论，现今的宇宙是在大约200亿年前诞生的。根据这个理论：宇宙一直在膨胀，这种膨胀是没有中心的，从任何一点看都能见到四周的星体远离我们而去。而且有趣的是，距离越远，退行（膨胀）速度越大，这就像一个正在充气的气球，表面上任何一点都会发现别的点正离它而去，而且距离越远，退离速度越大。宇宙的膨胀现象使我们想到，如果我们往回追溯，那么，宇宙会越来越小，就像涨大的气球放气一样，到最后就只剩下一个点了。因此科学家推论，宇宙是从点状宇宙——极小、极小的超微小宇宙，约为10的负34次方公分大小，发生大爆炸而开始膨胀的；此理论经爱因斯坦的一般相对论，及实际观测结果得到学界多数的认可。故此，目前宇宙空间尺度大约为200亿光年，大家知道，光是世界上跑得最快的东西，光速为每秒30万公里，它跑一年距离达9兆4600亿公里，可见1光年是很遥远的距离，而200亿光年即约10的23次方公里。宇宙由10的负34次方公分扩张到10的23次方公里，其间变化何等巨大！累积的讯息何等惊人！且还在不断扩大中。以我们所处的银河系为例，它的直径就有10万光年，而像太阳这样的恒星，银河系里差不多有2000亿颗，随着时间的推演，我们所知道的将愈来愈多。天文学家极力要在这庞大的资料中，找寻你我所在宇宙的过去与未来，天文学的发展已与处理超大型资料密不可分，愈研究就愈发现，从地球到苍穹（宇）、从亘古到永远（宙），实在太大了，人也实在显得太渺小了。

其次，从无生物到生物，以演化的观点来看，生物讯息也是愈来愈多。自古以来，人类对生命之探索一直很有兴趣。大家都知道母鸡生蛋，蛋孵小鸡，树木结出果子，生产种子，种子种植成树木。然而，最初的鸡或鸡蛋从何而来？最初的树木或种子从何而来？在探索生命过程中，140年前达尔文提出的"演化论"思想无疑的带来极大的震撼，引发人们对过去与未来的省思。根据达尔文的研究发现，生物是由低等演化至高等，构造由简单至复杂，所涵盖的资料讯息也是由少到多，以表一为例，体型较大、体制较复杂的高等生物（人、鸡），其DNA的总量和遗传基因总数要较低等生物（滤过性病毒、噬菌体）多得多。有些病毒的遗传基因数目较少，已被科学家发现且清楚排列出来，但像细菌这么简单的生物，遗传基因种类和数目，到目前为止还是弄不清楚，因此人类的遗传基因就更难确定。为什么难以确定？因为期间包含的资讯量太惊人了。以人类而言，人体每个细胞有46个染色体，共约65亿对

基(表一)，亦即平均每个DNA约有1.4亿对基(65亿/46)，其排列组合形成的密码变化极为惊人，因为DNA上的对基有4种，故理论上每个DNA至多可有4的1.4亿次方，亦即10的8千万次方种可能的组合方式，简直是超级天文数字，就算DNA上一小段的基因，其蕴含的资讯量亦极为惊人。照达尔文学派的理论，这是长期演化，由简至繁的结果；所以从生物演化的角度来看，生物自然界亦透露出大型资料走向的必然性，这也是生物科技在资讯技术发达的今日，能大行其道的重要原因。

【表一】 生物细胞内总基数

生物名称	核酸	基的组数(对)	遗传基因的数量(个)
人	双螺旋DNA	6.5×10^{9}	80,000～100,000
人的精子	双螺旋DNA	3.3×10^{9}	-
鸡	双螺旋DNA	2.3×10^{9}	-
果实蝇	双螺旋DNA	2.2×10^{8}	-
酵母菌	双螺旋DNA	3.7×10^{7}	-
大肠杆菌	环状双螺旋DNA	4.3×10^{6}	-
噬菌体T^2	线状双螺旋DNA	2.0×10^{5}	～100
噬菌体T^7	线状双螺旋DNA	38,000	～30
噬菌体φX174	环状双螺旋DNA	5,500	～8
滤过性病毒	线状双螺旋DNA	23,000	～10
噬菌体MS^2	线状双螺旋DNA	3,300	～3

说明：表内"～"符号是指大约数字。资料来源：知识解码，韦端、饶志坚，2000，晓园出版社。

最后，我们要从人类活动过程来看。根据人类科学家的计算，宇宙诞生于200亿年前，人类的祖先则在500万年前从它和黑猩猩的共同祖先分支，而有历史记载时间不超过5000年。若以24小时代表地球的年龄，那么人类历史活动长度还不到0.1秒，非常短暂，但人类却在这相对短暂时间内，快速发展，超越其他生物，一跃成为地球的主宰，这真是一项奇迹。就以刚过去的百年来看，汽车、飞机、电话的发明，使人类的起居起了革命性变化，太空梭、雷射、电脑、网际网路、核武、基因工程、复制羊都是远超过前人想像的产物，显示人类科技知识与成果，正以等比级数般的速度飞快累积，这期间各领域的资料数据库也不断扩大、扩大、再扩大，此与人类科技发达、活动频繁、分工细密与生活数据化有关，它不但是人类活动下的产物，也是提升人类生活品质的重要资讯来源。以台湾信用卡消费资料为例，20年前几乎只是少数有钱人使用的信用卡，由于所得提高，消费能力增强，再加上银行采用各种奖励措施积极抢攻信用卡市场，2001年底信用卡发卡数达4295万张，较2000年底增长32.5%，平均每人拥有1.9张信用卡，签帐笔数3.4亿笔，金额达7719亿元，占全体民间消费比重12.7%。在如此庞大消费市场诱因下，发行银行及百货商家如何从这一年3.4亿笔消费资料库中，分析消费者个人特征(如所得、职业、年龄、性别、教育程度)与消费特性(如消费金额、地点、日期、时间、商品)间之关联性，进而促销符合消费者需求之商品，提升服务品质，已是许多有心业者积极在做的事。当然，消费活动增加，消费资讯量也跟着增加，大型资料的形成也就理所当然了。

十七世纪犹太籍哲学家史宾诺莎强调理解是自由之道(他有句广为传颂的格言："不要哭，不要笑，要理解")。而各种大型资料库中背后隐藏的讯息是帮助我们理解的重要来源。对现在的条件了解得愈多，我们可以将这世界看得更透彻，就愈可以知道未来。人类对宇宙万物的认识虽仍有限，但知识增加的程度正以等比级数般的快速累积，如同开采一座巨大的知识金矿，从最早的用手，用血肉之躯去挖，挖1000年所得仍极有限，然后是用斧、锹、铲去挖，速度较快，挖100年的所得即超越前1000年收获，再演变至用挖土机、怪手开挖，挖10年就超越前100年的努力，今日则是以炸药、自动化机械全面大幅开挖，只要少许时间，即有重大成果。至于如何更有效率、更爆炸性的来开挖这些蕴藏超乎想像丰富的金矿，那就是接下来我们要讨论的主题了。

为了了解大量资料所内含的讯息，专家们提出了指标的观念。所谓"指标"(index)指的是将整体庞大的数据用少数(甚至少到一两个)指标来代表。常见的有平均数(mean)中位数(medium)、变异数(variance)、比率(ratio)等。比方说，两个国家比富裕用的是国民平均所得。比文化用的是认字率。当然用一个数字来涵盖所有资料很难去得到完全的面貌，只能尽力而为。而制订合理的指标是一门很深奥艰涩的学问。多年前我们在制定交通肇事率这种简单的指标便面对许多难题。所谓"交通肇事率"当然是两笔数字的比率。分子很明显的是该年度的交通事故的次数，但分母呢?如果把人口数当分母来除，结果中国大陆的交通肇事率最低，因为人一大堆但没有太多车(所以交通事故少)。如果把车辆数当分母，结果香港、日本的交通肇事率最低，因为车一大堆，但只有偶尔开，平常都搭公共运输(所以交通事故少)。有聪明人士便提出用总行驶的里程数来为分母，结果美国的交通肇事率最低，因为美国高速公路一口气可以开上百里上千里(所以分母可以很大)。如此单纯的指标都不容易，更不必说那些复杂的经济指标或社会指标了。

指标的制定固然不容易，在一般的资料中，指标的计算倒不是太大的问题。但当资料量相当庞大时，指标的计算和统计的分析便面对不同的困难。拿最简单的平均值和中位数为例。我们都知道平均值便是将所有观测值加起来除以其总个数。而中位数便是所有观测值排序，从小排到大，那排在正中央的数便是中位数。这可能是最基本而简单的统计指标。当资料太庞大时，我们面对什么样的困难呢?首先是我们的记忆体不够大，一般的PC(personal Computer)可能有个40GB(1GB=1024MB)，而当我们的资料需要的memory是几百个GB甚至几个TB(1TB=1024GB)时，我们没有能力把所有资料同时叫出来求和，我们只能逐步(sequentially)的做。其次是计算时间，即便我们的memory够大，执行计算的时间可能相当大。以排序而言，目前最快速的排序方法也是OrderO(nlogn)，在资料相当庞大时，则可能需要几年的时间来运作，那么如何不排序又能求得中位数便成了问题。

资料的复杂性基本有两个项目，一个是资料本身的复杂性，一个是使用分析方法的复杂性。资料本身的复杂性我们以其所占的记忆体(memory)大小来区别。如表二所示，当资料量只有10^2项以下，我们称它为"细小型"(tiny)资料，这大约可以记载在一张笔记纸上。当资料在10^4项以下，我们称为小型(small)资料，这大约可以记载在几张纸上。当资料量在10^6项以下，我们称它为中型(medium)资料，纸张已不足以记载，我们得用电脑磁碟片(floppyDisk)来储存。当资料在10^8以下，我

们称它为大型(Large)资料，磁碟片已不够用，我们必须用电脑的主记忆体(HardDisk)来储存。当资料量在10^{10}以下，我们称它为巨大型(Huge)资料，需要数个HardDisk来储存，资料量在10^{12}以及10^{15}我们分别称它为庞大型(Massive)和超大型(supermassive)资料，这时必需用特殊的电脑设备来储存。

【表二】资料大小的分类

资料类别	资料大小(以Bytes计算)	储存方式
细小型	10^2	一张笔记纸
小型	10^4	几张笔记纸
中型	10^6	电脑磁碟片
大型	10^8	电脑的主记忆体
巨大型	10^{10}	数个电脑的主记忆体
庞大型	10^{12}	特殊磁带
超大型	10^{15}	特殊磁带

分析方法的复杂性则以其计算的复杂来区别，我们利用数学上的Order(bigO)来作判别。简单的说，从头到尾把资料走一回便可以得到答案的方法是O(n)。比方说求平均值便是O(n)的演算法。又好比传统clustering的方法必需先求得所有点之间的距离才可以分析。这是C(n，2)的演算法，也就是$O(n^2)$。当然Order越小，算起来越快。如表三显示，常用的分析方法可以简单列成下列几组。划一个scatterplot是$O(n^{1/2})$；计算平均值，变异数是O(n)；求FastFourierTransformation或快速排序(QuickSort)是O(nlogn)；解回归分析，对矩阵分解均是O(nc)；多数clustering方法是$O(n^2)$；多元分析找outlier则是$O(a^n)$。

【表三】分析方法的复杂性(AlgorithmicComplexity)

$O(n^{1/2})$	Plot a Scatterplot
O(n)	Calculate Means, Variances, Kernel Density Estimates
O(n log(n))	Calculate Fast Fourier Transforms
O(n c)	Calculate Singular Value Decomposition of an r x c Matrix; Solve a Multiple Linear Regression
$O(n^2)$	Solve most Clustering Algorithms
$O(a^n)$	Detect Multivariate Outliers

将资料复杂性(表二)和分析方法复杂性(表三)综合在一起我们便可以得到表四。对OrderO(n)而言，针对tiny、small、medium、large和huge分别是10^2、10^4、10^6、10^8以及10^{10}，而对$O(n^2)$则是10^4、10^8、10^{12}、10^{16}、10^{20}，对O(nlogn)则是2×10^2，4×10^4，6×10^6，8×10^8，10×10^{10}。现今假使我桌上有一部Pentium III的PC(目前比较快而普遍的机器)，此类机器每秒钟可执行10^7个运算，所以表四可以除以10^7而得到表五。表五所呈现的是每一个情况的计算时间。我们可以看出当资料量是tiny时，即使计算复杂度为$O(n^2)$，大约千分之一秒可以计算完毕。但同样的计算用在巨大型资料上则需要317000年，换句话说在有生之年是看不到计算结果的。当然工作站(work station)或更先进的高科技计算机可以大幅提高这速度，但要达到可以接受的速度(比方说10多分钟甚至小时)则还需要一段漫长的研究。

所以两个研究重点值得进一步开发：

(一)所有的分析方法(Methodology)与算则(Algorithm)必需标明其order让使用者知道该不该使用，这好比药品食物必需标示其成份让使用者知道该不该使用，其道理是一样的。

(二)多数较复杂的方法(O(nlogn)以上)并不适用于巨大型资料，必需研发其替代方案，即所谓的Single-Loop方法。

这两个研究方向在未来几年中将有进一步的结果可期。

【表四】综合资料大小与分析方法的复杂性

n	$n^{1/2}$	n	nlog(n)	$n^{3/2}$	n^2
tiny	10	10^2	2×10^2	10^3	10^4
small	10^2	10^4	4×10^4	10^6	10^8
medium	10^3	10^6	6×10^6	10^9	10^{12}
large	10^4	10^8	8×10^8	10^{12}	10^{16}
huge	10^5	10^{10}	10^{11}	10^{15}	10^{20}.

【表五】表四各种组合之运算时间(PentiumIII 10megaflop)

n	$n^{1/2}$	n	nlog(n)	$n^{3/2}$	n^2
tiny	10^{-6} seconds	10^{-5} seconds	2×10^{-5} seconds	0.0001 seconds	0.001 seconds
small	10^{-5} seconds	0.001 seconds	0.004 seconds	0.1 seconds	10 seconds
medium	0.0001 seconds	0.1 seconds	0.6 seconds	1.67 minutes	1.16 days
large	0.001 seconds	10 seconds	1.3 minutes	1.16 days	31.7 years
huge	0.01 seconds	16.7 minutes	2.78 hours	3.17 years	317,000 years

大型资料同时也带来另外两个冲击，其一，资料所以如此的大主要原因是因为它是个时间序列(time series)，而时间序列方法在DataMining几乎是空白。其二，资料量这么大因为它什么都收集，换言之，一般而言它常常是population(母体)而不是我们统计课本内所学的样本(sample)。如何分析母体数据对我们而言是个新挑战。如果我们有能力去计算，那么算出来的便是population mean，population median而不是估计值(estimate)。

所以传统的统计思维在这里需要修正，传统的统计思维从"资料收集"，"资料分析"到"统计推论"自成一个完整的体系。但对大型资料而言，资料为什么会如此大，因为它常常是所有的资料，所谓母体(populaton)，而不是我们一般经由抽样技术而得到的样本(sample)，所以没有所谓"资料收集"的问题。在"资料分析"和"统计推论"方面，传统的思维是假设一个可能的机率分布(Distribution)然后透过收集而来的样本来估计机率函数中的参数(parameter)，而种种估计的方法便成了许许多多研究的重点。一旦参数值敲定，从机率分布函数便可有效地统计推断。对大型资料而言，整个母体资料都有了，为什么还需要去假设分布函数呢?是否应该从手中大笔资料直接去决定分布函数，直接作统计推论呢?技术层面上还需进一步开发但这观念应有所改变。

过去我们一直以为"数大便是美"，数据不嫌多。但当科技已进步到大小通吃，资料量几乎可以无限上网时，我们面对完完全全不同的困难。这些困难是等待一一被突破，一一被解决。