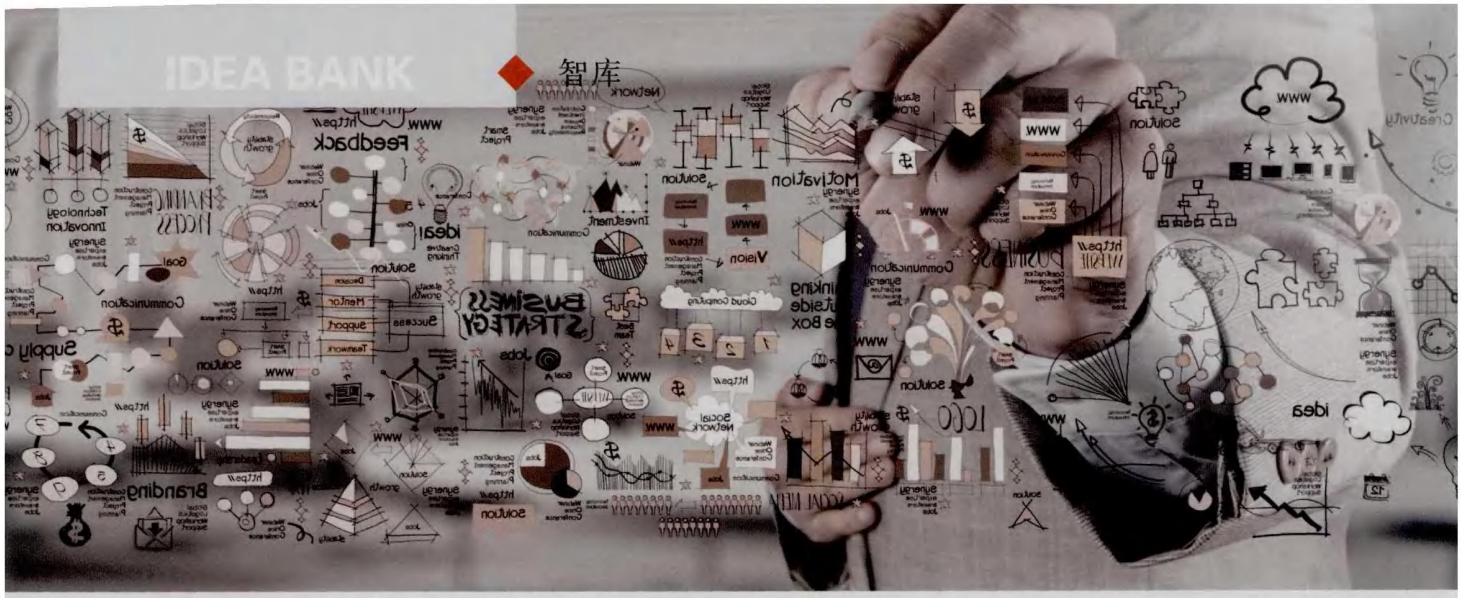


# 驾驭大数据（上）

文/ 徐端

在计算机走进千家万户后，人们开始进入信息时代。在智能手机、平板电脑几乎人手一部之后，各种智能设备带着形形色色的功能不断地产生大量数据，我们从信息时代逐渐走入大数据时代。大数据时代有着自己鲜明的特征，我们在考虑解决一些问题时，只有从习以为常的小数据时代的思维里跳出来，才能找到快速便捷的解决之道。

## 大数据闪亮登场

### 数据激增

2003年，刚进大学的小徐还没有自己的电脑，他省吃俭用花700元买了一个不知名品牌的MP3播放器，容量为128M。拿到MP3播放器后他非常欣喜，因为这个MP3播放器能存储大约50首普通压缩率的MP3歌曲，还能当软盘用。而他之前一直使用3.5寸的软盘来存储数据，一张软盘的容量仅为1.44M。之前他用来听歌的设备是一台索尼随身听，要听新歌只能花钱买磁带，每盘磁带大约30元，只能存储10首歌，而且没法自己挑选想要听的歌。

2013年，小徐已经参加工作多年，他平时使用智能手机听歌上网，使用平板电脑玩游戏、购物、看电影，家里的笔记本电脑已经用得越来越少了。可是，最近他想买一个移动硬盘来存储高清电影，他在网上浏览很久，最后花700元买了一个2T的移动硬盘。这个2T的移动硬盘大约能存储1000部高清电影，如果用来存储普通压缩率的MP3歌曲，大约能存储80万首。

不考虑货币购买力变化及产品功能等问题，只考虑数据容量，同样是700元，2013年购买到的容量是2003年的16000倍。可是，小徐还是觉得容量不够用，这10年里到底是哪里出了问题呢？

答案是，大数据。

大数据时代已经悄然来临。不仅是小徐，几乎所有的个人、企业、政府都已经觉得原来购买的存储设备容量不够用。随着社交网络的逐渐成熟，移动带宽迅速提升，云计算、物联网应用更加丰富。更多的传感设备、移动终端接入网络，由此产生的数据及数据增长速度迅速攀升。

一项调查发现，九成企业的数据量在迅速上涨，其中16%企业的数据量每年增长一半甚至更多。调研机构IDC在2011年6月的报告显示，全球数据量在2011年已达到1.8ZB，在过去5年里增加了5倍。1.8ZB是什么样的概念呢？如果把所有这些数据都刻录存入普通DVD光盘里，光盘的高度将等同于从地球到月球的一个半来回也就是大约72万英里。相当于每位美国人每分钟写3条推特微博，而且还要不停地写2.6976万年，是不是很恐怖？这还不是最恐怖的，IDC预测全球数据量大约每两年翻一番，2015年全球数据量将达到近8ZB，到2020年，全球将达到

35ZB。

所谓大数据最直白的理解是海量数据，通常用来形容一个公司创造的大量非结构化和半结构化数据。

北京时间2012年3月29日，美国政府宣布“大数据研究和发展倡议”来推进从大量的、复杂的数据集合中获取知识和洞见的能力。该倡议涉及联邦政府的6个部门。这些部门承诺投资总共超过2亿美元来大力推动和改善与大数据相关的收集、组织和分析工具及技术。此外，这份倡议中还透露了多项正在进行中的联邦政府各部门的大数据计划。

其实，最早提出大数据时代已经到来的机构是全球知名咨询公司麦肯锡。麦肯锡在研究报告中指出，数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素；而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。

麦肯锡的报告发布后，大数据迅速成为计算机行业争相传诵的热门概念，也引起了金融界的高度关注。随着互联网技术的不断发展，数据本身是资产，这一点在业界已经形成共识。如果说云计算为数据资产提供了保管、访问的场所和渠道，那么如何盘活数据资产使其为国家治理、企业决策乃至个人生活服务，则是大数据的核心议题，也是云计算内在的灵魂和必然的升级方向。

事实上，全球互联网巨头都已意识到大数据时代数据的重要意义。包括EMC、惠普、IBM、微软在内的全球IT巨头纷纷通过收购大数据相关厂商来实现技术整合，这足以看出它们对大数据的重视。

**数据大小怎么算**

说起阿基米得，大家肯定不陌生。他是古希腊伟大的哲学家、数学家、物理学家，其留传于世的数学著作有10余部。传说他曾经与某位国王一起下棋。国王觉得只是单纯下棋太没意思，不够刺激，于是想赌点什么，阿基米得也同意了。阿基米得提议的赌法是：如果阿基米得下棋输了，就给国王当一辈子长工；如果国王输了，就得在下棋的64个格子里放上米粒。米粒的放法是：第一个格子1粒，第二个格子2粒，第三个格子4粒，第四个格子8粒……每往后一个格子，米粒就增加一倍。国王心想，这赌注太值得了，赢了可以让阿基米得当一辈子长工，输了也就输那么一点儿米粒，于是很爽快地答应了。国王害怕阿基米得反悔，专门找来了纸张和笔，和阿基米得正式地签下了对赌协议。

一盘下来，阿基米得胜出。国王愿赌服输，大手一挥，吩咐手下去准备米粒。手下的人赶紧拿来一个米袋，开始给阿基米得数米粒。很快，一袋子米空了。手下又拿来几袋子，这次空得更快。国王沉不住气了，他完全没想到这小小的棋盘计算出来的数字竟然这么大。阿基米得微笑着看着一切，似乎一切都在预料之中。国王找来一个精通数学的大臣，让他计算一下还差多少。大臣一听说这个赌法，脸都吓白了。

这个故事最后是怎么收场的，我们无从知晓，不过国王肯定是支付不了那么多米的。我们不妨粗略估计一下国王到底要给阿基米得多少米。棋盘一共有64个格子，所以阿基米得一共会获得1+2+4+8+16+…+263粒米粒，合计（264−1）粒。210=1024，为了便于估算，这里仅仅算作1000。264可以看作是（210）6*24=1.6×1019。假如我们把中国人口算作16亿人，阿基米得得到的米粒足够给每个中国人发1010粒。一粒米大约0.02克，所以每个人大约可以获得2万千克也就是4万斤米。够震惊吧？要知道，这里的数字还是往小了估算的。

这个数字在那个年代是绝对的大数字。而这样的数字，在大数据时代，可以说是司空见惯的了。大家可能接触比较多的是各类电子文档和影音资料，比如一个10万字的txt文档大约200K，一首未经过压缩的APE格式歌曲大约30M，一张CD的容量大约为700M，一张普通DVD的容量大约为4.3G……关于KB、M、G这些表示文件大小的单位，我们一般比较熟悉。可是，你听说过T、P、E、Z、Y、D、N等单位吗？

这些单位我们不常遇到，但是在大数据里常常遇到。大数据又叫海量数据，光从名字看就知道数据的规模之大了。现在，个人、企业、政府手中的数据都处于井喷期，不断地大量爆发着。由于这些数据量是如此之大，已经不是以我们所熟知的多少G（1G=1000兆，即2的30次方字节）和T（即1000G）为单位来衡量，而是以P（1000T）、E（100万T）或Z（10亿T）为计量单位的。

那么，这些单位都是什么关系呢？它们之间如何换算呢？

在十进制的世界里，人们用以记录数字大小的数字符号有10个，分别是从0到9，数数的方式是0、1、2、3、4、5、6、7、8、9、10……而在计算机里，使用的是二进制，记录数字大小的符号只有0和1，数数的方式是0、1、10、11、100、101、110、111、1000……二进制数系统中，每个0或1就是一个位（bit），8bit为1Byte，称为1字节。字节是计算机文件大小的基本计算单位。一个英文字母占用一个字节，一个汉字占

用两个字节。

按照从小到大的顺序，单位分别为：bit（比特）、Byte（字节）、KB（千字节）、MB（兆字节）、GB、TB、PB、EB、ZB、YB、DB、NB。从KB到NB，人们习惯省略后面的“B”而直接用“多少K”或“多少N”这样的说法。

它们按照进率1024（2的十次方）来计算：

1Byte=8bit

1KB=1024Bytes

1MB=1024KB=1048576Bytes

1GB=1024MB=1048576KB=1073741824Bytes

1TB=1024GB=1048576MB=1073741824KB=1099511627776Bytes

1PB=1024TB=1048576GB=1125899906842624Bytes

1EB=1024PB=1048576TB=1152921504606846976Bytes

1ZB=1024EB=1180591620717411303424Bytes

1YB=1024ZB=1208925819614629174706176Bytes

1DB=1024YB=1237940039285380274899124224Bytes

1NB=1024DB=1267650600228229401496703205376Bytes

越到后面的单位看上去越像天文数字，我们似乎没有办法感觉到它们到底有多大。百度公司对此给出了更形象的描述：百度新首页导航每天就要从超过1.5PB的数据中进行挖掘，这些数据如果打印出来将超过5000亿张A4纸。这些纸全部摞起来超过4万千米高，接近地球同步卫星轨道，平铺可以铺满海南岛。而2020年新增的数字信息成长幅度将是2009年的近45倍。如今，只需两天就能创造出自文明诞生以来到2003年所产生的数据总量。

1.5PB的数据已经是这么大了，后面的EB、ZB、YB、DB、NB就真是大得不可想象了。再回头看看阿基米得的米粒，是不是也不算大了呢？

**大数据是什么**

2010年1月12日16时53分，加勒比岛国海地发生里氏7.0级大地震，首都太子港及全国大部分地区受灾情况严重。截止到地震发生后15天，世界卫生组织确认，此次海地地震已造成22.25万人死亡，19.6万人受伤。此次地震中遇难者有联合国驻海地维和部队人员，其中包括8名中国维和人员。地震发生后，国际社会纷纷伸出援手，表示将向海地提供人道主义援助。

地震发生后，海地人散落在全国各地，而当地的通信本身就很落后，从世界各地赶来的援助机构到达后，一直都搞不清楚到底该向哪里提供援助。他们只能以传统的方式，通过飞临灾区上空或赶赴灾区现场来查找需要援助的人群。就在这时候，一家独立的信息分析平台通过广播公布了手机短信紧急求助号码，结果收到数千条有关被困人员的信息。散居在美国各地的大量海地裔美国人翻译了这些信息，并把它们标注在“危机地图”上。这个数据分析平台的志愿者们通过互联网向海地的美国海岸警卫队发送即时消息，告诉他们搜寻地点，最终成功营救了不少当地居民。

这是大数据一次非常精彩的亮相。这家独立的信息分析平台是来自东非肯尼亚的一个开源数据分析平台——Ushahidi，它们一直收集和追踪有关暴乱、难民、强奸、死亡等事件的短信报告工作，并按照报告者提供的位置在地图上标明这些事件，并从中分析事件频发的位置，并进行预测和加强管制。和新闻报道和灾害应对小组相比，这个数据分析平台可以在更短

的时间内收集到更多的证据，这些证据的基础便是来源于对数据分析而进行准确的地理定位，通过在实时变化的地图信息来实施营救计划，在灾害面前，只有数据是最为冷静和理性的。

我们说了那么多大数据，那么，到底什么是大数据？

维基百科上，所谓“大数据”指的是：“网络公司日常运营所生成和积累用户网络行为数据增长如此之快，以至于难以使用现有的数据库管理工具来驾驭，困难存在于数据的获取、存储、搜索、共享、分析和可视化等方面。”

“大数据”作为时下工厂行业最火热的词汇，随之而来的数据仓库、数据安全、数据分析、数据挖掘等围绕大数据的商业价值的利用逐渐成为行业人士争相追捧的利润焦点。

早在1980年，著名未来学家阿尔文 托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。不过，大约从2009年开始，“大数据”才成为互联网信息技术行业的流行词汇。美国互联网数据中心指出，互联网上的数据每年将增长50%，每两年便将翻一番，而目前世界上90%以上的数据是最近几年才产生的。此外，数据又并非单纯指人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，也产生了海量的数据信息。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现赢利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。中国物联网校企联盟认为，物联网的发展离不开大数据，依靠大数据可以提供足够有利的资源。

随着云时代的来临，大数据也吸引了越来越多的关注。大数据通常用来形容一个公司创造的大量非结构化和半结构化数据，这些数据在下载到关系型数据库用于分析时会花费过多的时间和金钱。大数据分析常和云计算联系到一起。

“大数据”这个术语最早期的应用可追溯到apacheorg的开源项目Nutch。当时，大数据用来描述为更新网络搜索索引需要同时进行批量处理或分析的大量数据集。现如今，大数据不再仅用来描述大量数据，还涵盖了处理数据的速度。

从某种程度上说，大数据是数据分析的前沿技术。简言之，从各种各样类型的数据中快速获得有价值信息的能力，就是大数据技术。明白这一点至关重要，也正是这一点促使该技术具备走向众多企业的潜力。

大数据可分成大数据技术、大数据工程、大数据科学和大数据应用等领域，目前人们谈论最多的是大数据技术和大数据应用，工程和科学问题尚未被重视。大数据工程指大数据的规划建设运营管理的系统工程；大数据科学关注大数据网络发展和运营过程中发现和验证大数据的规律及其与自然和社会活动之间的关系。

## 大数据的新思维

### 免费的才是最贵的

据说，在非常遥远的古代，人们都是不穿鞋子的。有一次，一个国王到外面考察民情，走了一天的路后脚疼得难受。因为路上的石子实在太多了，硌得脚很疼。国王心想：“我只是走了一天的路就这么难受，可怜我的子民们每天都要走这样的路啊。我得想个办法。”他边摸着自己的牛皮座椅边思考着，突然受到启发：“牛皮足够坚硬和平整，又不尖锐，还耐磨，如果把所有的公路铺满牛皮，人们走起来不就不会硌脚了吗？”于是，他下令把全国所有的公路都铺上牛皮。他认为这样一来，全国的百姓都可以不被石子硌脚了。这时，一个聪明的大臣看不下去了，心想：全国这么多大大小小的路，这得多少牛皮啊？于是向国王提醒道：臣民们只要把自己的脚包上牛皮就可以了，不需要那么多牛皮的。国王一下子醒悟过来，赶紧更改了命令。

据说，这就是皮鞋发明的缘由。同样是为了不硌脚，国王的方法成本大得不可思议，而大臣只换了一下思维角度就得出了更好的办法。这就是经济学的办法。经济学要讲成本计算，而人类行为的规律揭示：每个人为自己的脚负责，是最经济的办法。不仅脚，其他事务也是如此。再说，如果是国王用全国人的钱为全国道路铺上牛皮，有多少人会珍惜这条牛皮公路呢？因为反正是免费的，谁会在乎？但如果是自己买皮鞋，他们就不会随意糟蹋脚上的牛皮了。

这是很简单的道理，但在生活中人们常常不知道这一点。很多时候，我们陷入了“牛皮公路”的错误思维而不自觉，扮演着那个自以为得计的国王。

2013年，国外著名的社交网站Facebook预计将实现60亿美元的收益，而创造这么多收益的Facebook居然没有向用户收取一分钱。Facebook的所有服务对用户都完全免费，如果你是


Facebook的用户，你会不会觉得你使用Facebook的服务简直是在占这个网站的便宜呢？

如果你这么觉得，你就已经陷入“牛皮公路”的思维了。Facebook不是慈善机构，它的管理者不是国王，他们的网站不是供所有人免费使用的牛皮公路。事实上，正如2010年《时代》周刊评选出的100位最具影响力的人之一的思想家杰伦 拉尼尔所说：“Facebook的用户今年将为这家公司创造60亿美元的收入，却得不到一分钱的报酬。”

为什么这么说呢？这又是一个大数据的案例了。很多人暗暗觉得，Facebook不是一个慈善机构，它应该有自己的赢利方式，只是自己不知道它是如何赢利的罢了。这是非常正确的思维方式，事实也确实如此。Facebook的价值正是数以亿计的用户在使用过程中不知不觉积累的大数据形成的。通过分析用户的喜好、身份资料、个人信息和浏览习惯，Facebook就能够猜测到每个用户的喜好，比如，你最容易被哪类广告吸引，每个网站页面都有一个“喜好”按钮，哪怕你从来不摁，你的信息也会被反馈给Facebook。

在大数据时代，数据就是金矿，而创造数据的用户便是产生金矿的原材料。Facebook的主要产品是社交网络，而造就一个良好的社交网络的最重要因素是它的内容。为Facebook提供内容的，正是一个个用户。用户提供的内容使网站变得美好，而他们的个人信息使得网站变得有价值。

这一切都解释了为什么像Facebook这么一家雇员少于5000人的公司，如今市值超过650亿美元。在思想家拉尼尔看来，这是一种巨大的不公平，也是大数据时代的一个巨大缺陷。像Facebook一样的公司，通过收集我们的各种行为数据获得巨大利润，而我们的行为本身却被视为是毫无价值的，似乎他们无须为我们的劳动付出任何报酬。这么看来，在大数据时代，表面上我们是在免费使用着某些公司的各种资源，而实际上是我们付出各种劳动，某些公司免费搜集着我们产生的数据，没有给我们任何报酬。这么一说，阿里巴巴创始人马云曾说“免费的才是最贵的”确有一定的道理。

那么，怎样才是合理的呢？让我们从小数据时代获得一些启示吧。比如，我们走在街头上，一个陌生人走过来请求我们帮助完成一项问卷调查。这种事情是常有的，当然，我们可以选择不合作。不过，很多时候我们都会帮忙完成。作为答谢，对方一般会准备一点儿小礼物，一支笔、一个小本子之类的。这算不上什么报酬，只能说是调查者对占用了被调查者的时间表示歉意的一种表达。那些如同Facebook一样的公司应该学会这种传统。首先，他们采集我们的数据，应该像在街头找我们做问卷调查一样征求我们的同意，而我们可以选择不同意。在我们表示同意他们收集数据后，他们应该认识到，他们应该礼节性地表示点什么。不然，这看似免费的服务才真正是最贵的。

### 一切皆可数据化

阿基米得曾经说：“给我一个支点，我就能撬动地球。”从某种意义上我们也可以说：“给我一组数据，我就能复制地球。”为什么这么说呢？数据到底能告诉我们多少信息呢？

在回答这个问题之前，我们不妨这么假设一下：现在我们正在野外的一块空地上挖掘，突然我们挖出了一个不明物体，这是一个规则的长方体。我们手上唯一的工具是尺子，现在我们量出了它的长、宽、高，也就能够在纸上画出这个长方体并算出它的体积。接着，我们发现这个长方体实际上是一个实心的大金块，那么根据黄金的密度我们可以算出它的质量，并根据当前黄金的价格给其估价；如果我们发现这块金块是贵重的文物，却不知道具体是什么时候的，我们可以把它带到实验室对它做C14鉴定，了解它具体制造于哪一年，进而推测是谁制

造的，这中间又发生了哪些故事……

从一开始我们只知道它是一个长方体到后来我们掌握了它的来龙去脉，这一步步里我们是如何增加对它的认识的？其实，我们只是逐步采集到了这么一些数据：

1.这是一个长方体；

2.这个长方体的长、宽、高的值；

3.我们已知的知识告诉我们：体积=长×宽×高，质量=体积×密度，黄金的密度=19.3克/厘米 ，由此得出物体质量；

4.由当前的金价，我们可以计算出这块金块值多少钱；

5.C14的半衰期为5700年，计算出这块金块的C14含量，就知道它制造的年代。

……

这一过程中，我们采集到的具体数据越来越多，最后得到的信息也越来越多。我们采集到的数据的多少，决定了我们准确描绘它的程度。对一块金块是如此，对这个地球同样是如此。当我们掌握的数据足够多，多到我们足以完美描绘出这个地球的任何一个特征，我们就能够将它数据化。同样，我们采集到一个人的数据足够多时，就能很好地用数据描绘这个人。

2011年12月，英国电视4台播出了一部名为《黑镜》的迷你电视剧，全剧共两季，每季3集，每集都是一个独立的故事。虽然每集都有不同的演员上演不同的故事，但所有故事都是围绕我们当今的生活展开的。在《黑镜》第二季里，编剧查理 布鲁克为大家讲了3个故事，其中第一个故事是这样的：女主角是一个叫玛莎的女孩，她深爱的男友艾什因车祸意外去世。刚刚怀孕的玛莎痛不欲生，每天都沉浸在过去，怀念着有艾什的日子。艾什生前沉迷于各种社交网络，在网络上留下了不少东西，包括照片、视频、聊天记录、电子邮件等。而此时，一种新的电脑软件出现了，只要将艾什生前散落在网络上的各种内容全部整合在一起，经过一系列复杂的数据分析，这个软件就能够准确地掌握艾什的各种特征，包括形象、语言风格等。通过这些特征，这个软件可以再造出一个艾什出来。玛莎接受了这项服务。这样，玛莎可以像过去一样与虚拟的艾什进行网络聊天、手机通话等。

这当然不是死而复生，而是一个大数据时代的奇迹。如果顺着这个剧情设想，我们不难作出预测，在未来，现在不能数据化的东西都可以数据化，直到最后一切都可以数据化，包括一个人、一个世界。

这个故事到后来发展到玛莎订购了一个具有艾什特征的机器人，然后发现机器人毕竟只是机器人，没有艾什的灵魂，最终玛莎放弃了这个机器人。导演似乎是要告诉我们，科技到任何时候都无法代替一个真正的人。可是，灵魂到底是什么？不就是说机器人还不够像艾什吗？那也只是因为艾什留下的数据还不够大，如果艾什从出生到车祸死去前的所有行为特征都被采集到了，根据这个采集到的大数据定制出的艾什和真正的艾什又有何不同呢？

且不说这个剧里导演的考虑，这个剧给我们最直观的感受是：大数据分析可以强大到复制出一个人。在未来的世界里，一切都可以数据化，包括人。一切都保存在互联网的数据库中，当你有一天需要的时候，数据库服务商能够将这些数据调出来给你。

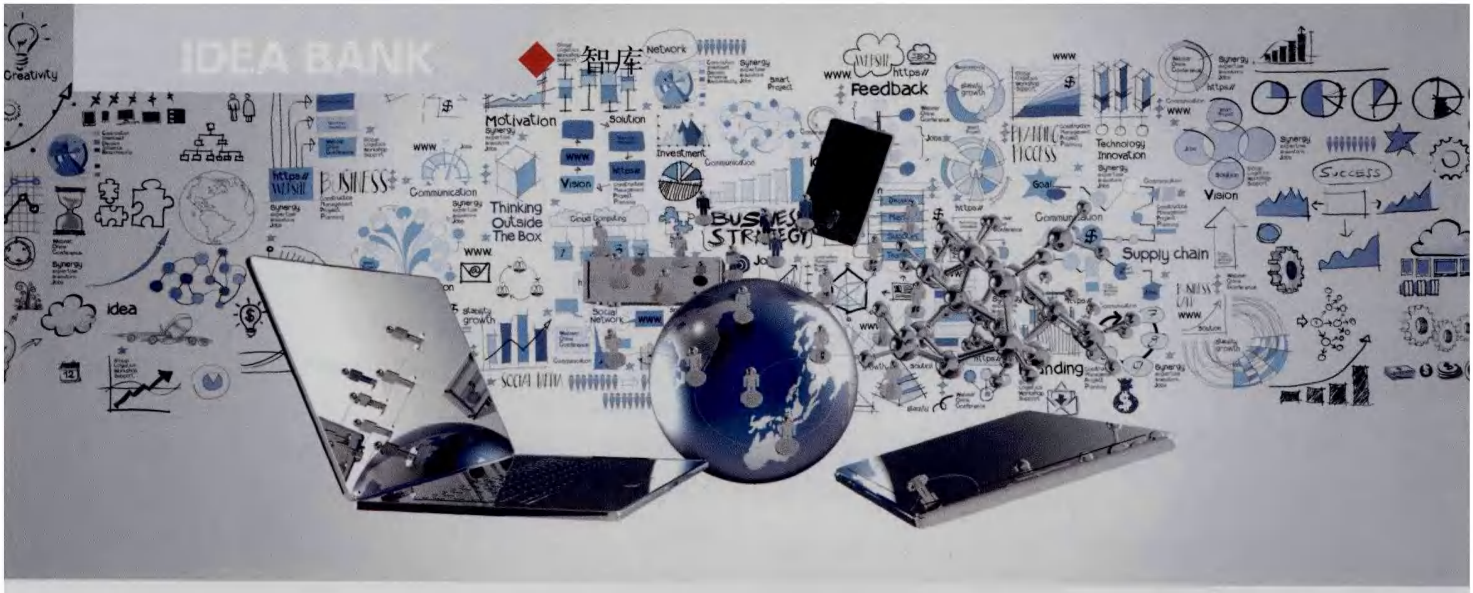
**一切都可以量化**

很多传统观念告诉我们，有些东西是可以量化的，而有些东西不能够量化。比如，一个书法家每天写了多少字是可以量化的，数数字数就知道了，而写字的优劣是没办法量化的，因为每个人欣赏眼光不一样；一个鱼缸里有多少鱼是可以量化的，数数就知道了，而整个地球的海洋里有多少鱼是没法量化的，实在没办法去估算……现在，我们需要转变这个观念。

要知道，凡事皆可量化。只要我们能够找到观察问题的方式，并从一个新的角度去衡量它，不管从这个新的角度衡量它到底精准度如何，只要它能让我们知道得比以前更多，那么它就是一种可行的量化方法。实际上，对那些看似不可量化的东西，人们总能找到相对简单的量化方法。

1938年诺贝尔物理学奖得主、著名的物理学家恩里科 费米在使用各种高明技巧方面很有天分，在量化工作方面也是如此。很多人都知道他的一些有关量化的有趣故事。

1945年7月16日，美国新墨西哥州洛斯阿拉莫斯附近的特里尼蒂沙漠进行了第一枚原子弹爆炸的试验。在其他科学家对量化爆炸当量的仪器进行最后校正时，作为基地观测爆炸情况的原子弹科学家之一的费米正在把一张纸撕成碎片。当第一波冲击波冲过营帐时，他把碎纸屑慢慢撒向空中，观察它们在冲击波的冲击下能飘多远，最远的碎片承受的就是波的压力峰值。费米知道一条简单规则，那就是碎纸片在风力作用下的漂移和他想要量化的数据有关。据此，费米得出结论：爆炸当量




至少有10000吨。这应该是一条新闻，因为其他观测者还没有算出这个下限。人们都在估计这次爆炸的当量，有说5000吨的，有说2000吨的，但都是非常感性的猜测，没有一个很好的估算办法去衡量，也没有其他的原子弹爆炸的参数去对比，因为这是原子弹的第一次爆炸。在人们根据仪器的读数作了大量分析后，最终的计算结果为18600吨，这证实了费米的猜测。

在整个职业生涯中，费米深谙快速估算的价值，并以教授学生们估算一些奇妙的数值而著称。学生们首次接触这些问题时，对所要量化的东西简直一无所知，最著名的例子就是“费米问题”。费米问他的学生该怎样估计芝加哥的钢琴调音师的人数，他们都是学科学和工程学的，开始时一般都会说他们对这个数据的相关知识知之甚少。

当然，也有一些解法是比较简单的，如通过查看广告一个个统计钢琴调音师的数量，或者通过发证机构来检查某种执照的数量等。但是，费米教给学生的是量化“无形之物”的方法，他希望学生们通过提问题并量化其数值，从而能真正了解并领悟到一些东西。

费米首先问学生们关于钢琴和钢琴调音师的其他问题，这些问题虽然也是不确定的，但相对容易一些，包括芝加哥当前人口数量（1930—1950年，略超过300万）、每家平均几口人（2或3人）、家庭平均拥有的需要定期调音的钢琴数量（10家里最多1家，但30家至少有1家）、每部钢琴需要调音的频率（也许平均一年1次）、一个调音师平均每天能调多少部钢琴（4～5部，包括交通时间）、一年工作多少天（约250天）等。此时，根据这些数据就可以计算结果。

芝加哥的家庭数量=芝加哥人口÷平均每个家庭的人口数

芝加哥拥有钢琴的家庭数量=芝加哥的家庭数量×有钢琴的家庭的百分比

芝加哥每年需要调音的次数=芝加哥拥有钢琴的家庭数量×每年需要调音的次数

一个调音师每年的调音次数=调音师每天调音的钢琴数×年工作天数

芝加哥的调音师数量=芝加哥每年需要调音的次数÷一个调音师每年调音次数

根据选择的不同特定值，所得结果应该是20～200，一般在50左右。后来费米可能从电话号码簿或行业协会弄到了真实值，当他把猜测值和真实值作比较时，发现他总是比学生们猜测的更接近真实值。或许20～200这个范围看起来很大，但考虑到这是学生们最初从“我们怎么猜得到”的态度开始一步步改进而得来的，就已经很不错了。这种解决费米问题的方法，被称为“费米分解”。这一方法不仅有助于估计不确定的数值，而且也给评估者提供了查看不确定性的来源。是每家平均拥有的钢琴数量不确定，还是钢琴每年需要调音的平均次数不确定，又或者是调音师每天调音的钢琴数量或者其他什么因素？弄清楚不确定性的来源，可以帮助我们量化相关事物，以便最大限度地减少不确定性。

从技术上说，费米分解法不完全是量化，因为它不是建立在一种新的观测方式基础上的，但它确实是一种让你更加了解问题的评估方式。在大数据时代，数据在以我们无法想象的速度增长着，有些问题是无法实现非常精确的计算的，而费米分解就为我们提供了很好的思路。我们要避免陷入不确定性及无法分析的泥潭，为了避免被显而易见的不确定性压倒，应该从知道的事情开始提问。正如后面看到的，评测我们目前了解的事物的数量，是量化那些似乎根本不可量化的事物的重要步骤。

**大数据≠大价值**

电视连续剧《薛平贵与王宝钏》中有一段剧情，说的是王宝钏的二姐王银钏刻薄嫉妒、嫌贫爱富。她不但对母亲疼爱宝钏感到愤愤不平，还非常看不起沦为乞丐的薛平贵。王银钏曾对薛平贵百般羞辱，极尽嘲讽之能事，一心想让王宝钏和薛平贵棒打鸳鸯两处飞。后来，薛平贵飞黄腾达登上高位后，赐她金碗要她沿街乞讨以示惩罚。讨到金钱或食物算她好运，讨不到东西就活该她倒霉。这还不够，薛平贵还在惩罚里加了限制条件：那只金碗只许用不许卖，并派官兵在她后面监督。王银钏拿着金碗怎么也讨不到饭，因为别人要么认为她是神经病，要么觉得事有蹊跷不敢随意施舍。

这个故事到这里并没有结束，但我们只讲到这里。这里有一个疑问：薛平贵为什么要以这种方式惩罚王银钏呢？这其实是一种暗讽。薛平贵就像那只金碗一样，非常贵重，王银钏曾经离薛平贵那么近，却一点儿也不识货，就像拿着金碗讨饭一样。一个金碗，在识货的人手里才能体现出它的价值。像王银钏那样拿着的金碗，既不能卖也不能换东西，失去了它应有的价值。其实，大数据也是这样的。为什么这么说呢？

大数据并不等于大价值，就像金碗并不一定等于大价值一样。一个企业掌握着庞大的数据，如果没有对其进行数据分析，这些大数据就是一个沉重的负担。因为光是采集和储存这些数据都要耗费很多人力资源和时间成本，而采集到的数据没有给企业带来红利，只有支出没有收入。

从麦肯锡的调查来看，大数据确实给很多行业带来了价值，比如为美国的医疗行业带来了每年3000亿美元的价值，而其他的各行各业也一样可以从大数据中受惠。

大数据带来大价值，但是大数据不等于大价值。就像一座未开发的金矿不等于黄金万两一样。金矿只有通过开发成为金砖后才能产生价值，而数据只有通过技术和分析工具显现在大家面前，使得数据变成信息，然后信息分离出有用的信息，才能产生价值。大数据也是一样，无非就是数据的量不同。

大数据就像一座庞大的冰山，大量的数据都隐藏在海面之下，显现出来的只有一点点。如何将这些大量的数据挖掘出价值，这是和IT技术进步相关的。现在，计算机的硬件和软件计算能力都越来越强大，使得我们从大量数据中提取有用信息的速度也越来越快，很多以前我们无法计算的问题现在能够得到解决。

例如，富士通帮日本的医疗机构做数据挖掘，其中一个项目是将很多电子病历、抑郁症患者的DNA信息、抑郁症患者的重点发病地都结合起来。富士通和日本大学医院政府做实验，根据病例、气象、DNA、地域数据，分析抑郁症患者自杀的概率，建立数据模型，进行验证。这在过去是不可能做到的，但现在有IT技术后，可以把假设通过技术很快地运算并加以验证，这样，以前没有体现出价值的数据便体现出了价值。

另一方面，过去某些大数据可能也是可以进行分析的，但是因为数据量太大或者计算过于复杂，得到结果的速度实在太慢，等待结果出来时，数据的时效性可能已经过了。比如我们要预测第二天的天气，以前的计算机可能需要三四天才能够计算出来，而等计算出来，预测本身已经失去了意义。而现在，同样的计算可能只需要几个小时。这样，预测本身的价值就体现出来了。

大数据不等于大价值，但大数据分析做好后，大数据就会带来大价值。随着大数据技术的发展，一些现在将大数据视为负担的企业将越来越多地感受到大数据分析带来的甜头。（未完待续）BR


**作者简介**

徐端：湖北孝感人。关注互联网十五年，见证了中国互联网从出现到蓬勃发展的全过程。兴趣广泛、涉猎颇广，勤于思考、善于研究，然经常停于设想、怠于行动，人称“一只脚在IT界、一只脚在文化界的树懒。”