# 大数据、数据科学和物联网

◆吉姆•杜亚特/ 文

编者按

在第3届世界质量论坛暨第13届上海国际质量研讨会上，美国质量学会资深会员、大数据方面的专家小组成员吉姆•杜亚特发表了题为《大数据、数据科学和物联网》的演讲，剖析了管理层对质量管理和数据科学的一些误区，分享了不同类型数据专家、质量工作者提升数据应用效率的关键作用。本文根据现场翻译速记整理。

说到大数据，通常会提到速度、体量和多样性。速度包括流程速度和数据速度；体量包括生产体量和数据体量；多样性包括变量的多样性、数据类型的多样性，以及最重要的数据位置的多样性。我们必须知道数据在什么地方，然后设法采集数据，才能够进行分析。

## 大数据

为什么要说大数据？大数据可以让我们在进行有效分析时不依赖取样或样本质量，而使用来自连续或高速离散过程的实时数据流。这里面涉及到IT技术和OT技术，IT在数据领域更多的是指数据管理技术，而OT则倾向于数据运营分析。

另外就是查找DOE（试验设计）和控制图的最佳变量。我们要找到最佳变量，用于试验设计，更好地控制质量。其实，变量的数量非常多，如果说一个变量对分析或者数据的质量没有太大影响，就可以把这个变量省略掉，以便留下最好的、真正会带来重大影响的变量。

最后，基于大数据应用高级分析学，比如在企业信息数据方面详尽的收集及学习，可以帮助企业做出明智的决策。机器学习其实是可以归纳到高级分析学范畴的。2017年，我出版了一本书《Data Disruption》，把数据分为非活跃数据和活跃数据。所谓非活跃数据，指的是存放在数据仓库里的数据，而活跃数据指的是在事件发生时实时

产生的数据，比如从传感器或流程中获得的实时数据。

管理层相信什么？现在管理层相信的那些东西并不有利于数据分析。比如，管理层认为所有的运行数据应该归口于IT/OT部门来管理，而不是由质量工作者来管理。有时，他们还会霸道地认为，数据专家就是指IT/OT人员，但事实上，数据专家可以存在于各种岗位中。

有一种机构叫卓越分析中心（ACE）。那里集结了所有的数据专家，可以解答任何关于数据和数据分析方面的问题，比如业务流程、产品性能、数据来源、分析工作的目的、分析过程等。从事这方面工作的人，非常重要。但管理层认为，卓越分析中心可以独立于流程和产品之外；只需要它来处理复杂的问题，日常工作和简单分析并不需要它来做。这些看法都不对，显示出管理层对质量工作并不了解。

SME，在这里并不是指中小企业，而是精通某一领域的行业专家，比如在工程、质量、营销等方面的专业人士。管理层通常认为，管理并不需要他们的参与。事实上，质量工作者了解产品、流程、工艺等，他们不应该像IT人员那样单独地坐在某间办公室里。

管理层的思维方式，就好像隔着一道墙把一些问题扔出来给IT/OT人员，然后坐等答复。IT/OT人员对计算机、信息技术非常在行，但并不像质量工作者那样善于分析，所以很多时候把一些要求扔给他们，是勉为其难。所以一定要改变这种“隔着墙把问题抛出来”的方式，各个相关部门的行业专家或同事集中在一起讨论。

把想象力（Imagination）和分析（Analytics）结合在一起，我创造了一个新词叫Imagilytics。首先，要想象一下怎么使用数据、怎么帮助其他部门的人。如果其他部门有一个流程或者一个工艺，它必须保持稳定，你是不是能够帮助他们进行监测，从而实现稳定流程的目的。其实，质量工作者能做的事情越多，就越能向管理层证明其价值。

## 数据专家

数据专家通常有四类。

第一类数据专家是计算机技术专家，他们了解什么是云计算，通晓这些专用术语和技术，可以帮质量工作者收集数据，进行格式化或数据整理等，确保数据可以用于分析。

第二类数据专家是统计师和高级分析师，他们拥有统计训练或高级分析的行业背景，可以进行非常详细的统计工作，在开发预测分析和预测模型领域拥有专长，能够帮助企业建立统计分析模型。

第三类数据专家是擅长应用分析的业务分析师，俗称商业分析专家。他们熟悉模型或者其他分析手段，能够把数据导入第二类数据专家开发的模型或者软件中，应用于企业内部进行业务分析，帮企业做出更好的决策。

第四类数据专家的主要作用是确保沟通的顺畅。他们更像一位守门员，确保此前的数据工作满足公司需求，确保报告内容的准确，确保沟通渠道的畅通，并以管理者习惯的方式做出分析说明，帮助管理者做出明智的决策。

我有一篇文章叫《如何打造一支数据科学梦幻团队》，发表在福布斯杂志，里面详细阐明了这些细节。

通过四大分类可以知道，什么样的数据专家适合做什么样的工作。其实，这四类数据专家都需要相互沟通，虽然每个人都有自己的职责。对于质量工作者而言，你可以和四类专家交朋友或成为四类专家中的任意一种。你不一定要成为一个IT专业人员，如果有统计分析的专业背景，你可以成为第二类数据专家；如果你有相应背景或工作经历的话，也可以成为第三类数据专家。

每一类数据专家都有自己的专业领域，但领域会有重合。第一、二类数据专家的交叉领域包括：验证数据策略；识别数据；收集数据。第三、四类专家的交叉领域包括：报告验证；报告解读。而对于所有数据专家而言，最关键的一项保证便是获取可使用的数据。我经常听到有人抱怨说“我们有数据，但没有办法使用数据”，如果出现这种情况，那么你需要与第一类数据专家合作，他们能帮你对数据进行预处理。

工程师主要进行数据的创建和收集，他们需要了解生产过程的详细信息，比如机器、材料、操作参数等。IT/OT人员，也就是第一类数据专家主要进行数据的存取、整理、格式化等，确保这些数据能够在其他语言环境下正确使用。他们需要通晓技术，比如机器人、感应器、软件、人工智能、数据结构等。第二、三、四类数据专家或数据使用者主要应用工具、先进方法进对高质量的数据进行解读、分析，他们需要了解产品的详细信息、规格、功能或性

能，以及客户期望。这三类专家也需要通力合作，确保企业能够根据最终数据做出好的决断。

这些数据专家需要具备什么样的素质，又如何和他们打交道呢？首先，他们需要有很好的沟通技巧，要不断学习专业知识、掌握基础技能。其次，虽然有些数据专家对于数据会很保守或过分保护，但作为管理者要尊重这一特点，要学会和数据专家交朋友。

第二类数据专家属于分析型的，非常重要。他们通常需要接受高级分析方面的培训（至少学习术语），尤其是需要有想象力，也就是我前面说到的“Imagilytics”，才能帮助别人创造更多价值。第三类数据专家通常需要将统计分析应用于企业的业务之中，同样需要“Imagilytics”，一般是擅长使用数据的质量从业人员。与第四类数据专家合作互动，往往可以获得“真相”，帮助你做出明智的决定。

## 数据的质量及分析应用

美国的著名作家马克•吐温说：“数据其实就是垃圾。在收集它之前，你最好知道准备用它干什么。”

大数据需要进行适当的整理、存储、格式化和访问。其实，数据库就像一个垃圾场，不管干净也好、混乱也罢，垃圾场就在那里。比如，excel数据表就像是在堆肥，如果放的时间太久就会发臭。很多数据的时效性很短，无法长时间使用，堆积在那里过一段时间后就必然变质。因此我们需要有组织的垃圾填埋场或回收中心，将数据垃圾过滤，进行可回收利用，使之成为具有价值的数据仓库或数据湖。如果没有办法拿到处理过的数据，就没办法进行后续分析。

我们都有过类似经历：用一张表格来定义“谁是我们的客户”“他们拿到了什么产品”“产品的基本规格”“流程有哪些”，然后还要确认“我们需要什么样的数据”“数据的来源是什么”“能否获得这些数据”等。这样一张表格，能够帮助我们更好地和别人沟通。譬如你可以跟IT人员说“我没有拿到想要的数据”，这张表格可以很清楚地向他展示你需要什么样的数据以及是否能使用这些数据。

再看一下机器学习。机器学习能够生成一些由机器产生的原始数据，而且可以存储预测模型，将数据进行导入和分析，了解或预测结果。也就是说，可以在流程的边缘端控制芯片上进行数据的处理。

处理数据时有一些非常有用的新工具，比如预测性模型、文本分析、聚类分析、网络图、视觉系统等。网络图要比帕累托图更管用，帕累托图通常只被用来分析频率，而网络图除了分析频率之外还可以分析时间和成本。

预测性模型会用到静态数据，通过静态数据创造出动态数据，然后在分析边缘进行使用和部署。大家知道汽车用的引擎里有芯片控制汽油流量，它是怎么做到的呢？通过模型来控制引擎中的汽油流动。这一过程需要有数据，对模型中使用的程序进行最优化，然后嵌入到芯片中。创建阶段，在数据仓库中访问“静态数据”，端到端使用所有可用变量，然后建模分析；采用阶段，使用保留集和测试数据完善模型，从实时数据中确定关键变量，最终确定最佳预测模型；部署阶段，将模型置于实时环境中进行验证和测试，将模型“刻录”到控制芯片中进行部署。

再来看预测模型的另外一个应用。比如有一种平底锅，它的涂层非常容易剥落。碰到这个问题，很多厂商仍习惯于从最显而易见的涂层工艺出发。如果那样的话，首先要研究涂层的工艺流程，通过鱼骨图查找涂层剥落的原因，然后做实验设计（DOE），确认关键变量并提出更改建议。这种方法确实可以在一定程度上降低涂层剥落发生的概率，但不能彻底解决问题，因为解决方案只是从涂层的角度出发。如果查看整个生产流程的话，可以发现在制作平底锅之前，需要将铝板压制加工成合适的厚度，其中有一个非常关键的指标叫颗粒大小，颗粒大小严重影响涂层是否会脱落。可见，影响平底锅涂层质量的关键变量因素，并不在涂层的工艺流程中，而在更早的铝板压制过程中。因此，我们需要端到端分析整个过程，首先检查生产过程的所有部分，创建端到端的流程图，然后列出过程和数据位置中的所有变量，使用机器学习创建预测模型，检查对涂层剥落影响最大的变量，再根据分析建议进行更改，将最重要的变量用于新的实验设计中。

我们今天着重讲了三个方面：谁来做事情；去哪里获取数据；拿到数据后到底做什么。大数据中有很多东西，但最关键的一点是，让管理层知道谁来做事情。只要具备这种环境，你就可以成为一位数据专家。