数据科学中的“数据智慧”

郁 彬 / 文　　张心雨　吕 翔 / 译

本文转载自统计之都 http://cos.name/2015/05/the-data-wisdom-for-data-science/
转载得到了原文作者郁彬教授的授权

在大数据时代，学术界和工业界的大量研究都是关于如何以一种可扩展和高效率的方式对数据进行储存、交换和计算（通过统计方法和算法）。这些研究非常重要。然而，只有对数据智慧（data wisdom）也给予同等程度的重视，大数据（或者小数据）才能转化为真正有用的知识和可被采纳的信息。换言之，我们要充分认识到，只有拥有足够数量的数据，才有可能对复杂度较高的问题给出较可靠的答案。“数据智慧”对于我们从数据中提取有效信息和确保没有误用或夸大原始数据是至关重要的。

“数据智慧”一词是我对应用统计学核心部分的重新定义。这些核心部分在伟大的统计学家（或者说是数据科学家）图基（John W. Tukey）的文章[1]和伯克斯（Geogre Box）的文章[2]中都有详细介绍。

将统计学核心部分重新命名为“数据智慧”非常必要，因为它比“应用统计学”这个术语能起到更好的概括作用。对于这一点，最好让统计学领域之外的人也能了解到。因为这样一个有信息量的名称可以使人们意识到应用统计作为数据科学一部分的重要性。

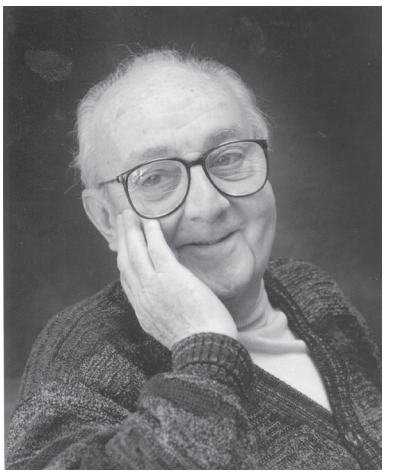
依据维基百科对“智慧”词条进行解释的第一句话，我想说：**“数据智慧”是将领域知识、数学和方法论与经验、理解、常识、洞察力以及良好的判断力相结合，思辨性地理解数据并依据数据做决策的一种能力。”**

“数据智慧”是数学、自然科学和人文主义三方面能力的融合，是科学和艺术的结合。

[1] 参见 http://projecteuclid.org/euclid.aoms/1177704711。

[2] 参见 http://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480949#.VR2_eWYhByU。

约翰·图基（1915-2000）



乔治·伯克斯（1919-2013）

如果没有实践经验者的指导，仅通过读书很难学习到“数据智慧”。学习它的最好方法就是和拥有它的人一起共事。当然，我们也可以通过问答的方式来帮助你形成和培养“数据智慧”能力。我这里有 10 个基本问题，我鼓励人们在开始从事数据分析项目或者在项目进行过程中要经常问问自己这些问题。这些问题是按照一定顺序排列的，但是在不断重复的数据分析过程中，这个顺序完全可以被打乱。

这些问题也许无法详尽、彻底地解释“数据智慧”，但是它们体现了“数据智慧”的一些特点。

1. 要回答的问题

数据科学问题最初往往来自统计学或者数据科学以外的学科。例如，神经科学中的一个问题：大脑是如何工作的？或银行业中的一个问题：该对哪组顾客推广新服务？要解决这些问题，统计学家必须要与这些领域的专家进行合作。这些专家会提供有助于解决问题的领域知识、早期的研究成果、更广阔的视角，甚至可能对该问题进行重新定义。而与这些专家（他们往往很忙）建立联系需要很强的人际交流技巧。

与领域专家的交流对于数据科学项目的成功是必不可少的。在数据来源充足的情况下，经常发生的事情是在收集数据前还没有精确定义要回答的问题。我们发现自己处在图基所说的“探索性数据分析（Exploratory Data Analysis, EDA）”的游戏中。我们寻找需要回答的问题，然后不断地重复统计调查过程（就像伯克斯的文章中所述）。由于误差的存在，我们谨慎地避免对数据中出现的模式进行过度拟合。例如，当同一份数据既被用于对问题进行建模又被用于对问题进行验证时，就会发生过度拟合。避免过度拟合的黄金准则就是将数据进行分割，在分割时考虑到数据潜在的结构（如相关性、聚类性、异质性），使分割后的每部分数据都能代表原始数据。其中一部分用来探索问题，而另一部分通过预测或者建模来回答问题。

2. 数据收集

什么样的数据与第 1 条中要回答的问题最相关？

实验设计（统计学的一个分支）和主动学习（机器学习的一个分支）中的方法有助于解决这个问题。即使在数据收集好了以后考虑这个问题也是很有必要的。因为对理想的数据收集机制的理解可以暴露出实际数据收集过程的缺陷，能够指导下一步分析的方向。