21世纪天文学面临的大数据和研究范式转型

张彦霞，崔辰州，赵永恒
中国科学院国家天文台光学天文重点实验室，北京 100012

摘要
随着大型空间和地面观测技术的发展，天文数据的数据量、数据产出率和数据复杂性急剧增加，天文学步入了全新的数据密集型时代。阐述了天文数据的特征，分析了传统天文研究方法的局限性，提出了发展天文统计学和天文信息学的必要性和研究方向，介绍了天文大数据的典型应用——大型综合巡天望远镜的工作原理及应用，归纳了与天文统计学和天文信息学相关的国际组织，分析了与其相关的公共教育的现状及存在的问题，为天文大数据的健康发展提供参考方向。

关键词
天文大数据；天文信息学；天文统计学；巡天

中图分类号：P11 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016067

Big data and paradigm shift for astronomy in the 21st century

ZHANG Yanxia, CUI Chenzhou, ZHAO Yongheng
Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

Abstract
With the development of large space-based and ground-based observational technologies, the volume, output rate and complexity of astronomical data rapidly increase. The astronomy steps into a new data-intensive era. The characteristics of astronomical data were represented. The limitations of traditional astronomy were analyzed. The necessity and research direction of developing astrostatistics and astroinformatics were put forward. As a typical astronomical application, the operation and applications of LSST were introduced. The organizations related to astrostatistics and astroinformatics were summarized. The present situation and problems of public education about this respect were analyzed. All these were provided as references for the healthy development of big data in astronomy.

Key words
astronomical big data, astroinformatics, astrostatistics, survey

1 引言

从古至今，天文学一直都是以观测为基础的科学。伽利略望远镜指向了天空，开启了望远镜观天的时代。哈勃空间望远镜的出现，又让人们的视野一下子从地面扩展到了太空，观测窗口也由单纯的可见光波段转向多波段。多波段望远镜的出现，使得天文学开始发展成为全波段天文学。多目标多光纤望远镜的出现，使得天文学由定点观测变为巡天观测，标志着天文学巡天时代的开始。这一切的发展变化都离不开天文探测技术、望远镜技术、计算机技术、网络通信等技术的飞速发展。天文巡天项目已经成为天文数据获取的主要来源。20世纪40年代，美国帕拉玛巡天开始了最早的天空普查；20世纪90年代，英国和澳大利亚合作的2度视场星系红移巡天和美国Sloan数字巡天计划陆续开始；21世纪初，英国和澳大利亚合作的6度视场星系红移巡天和RAVE、中国郭守敬望远镜、欧洲空间局的盖亚项目已经开展或正在运行，这些项目都属于可见光波段。其他波段（如红外波段、射电波段、X射线波段、γ射线波段等）的巡天项目也在有序开展。21世纪，即将运行的大型综合巡天望远镜（large synoptic survey telescope，LSST）和平方公里阵列（square kilometer array，SKA）望远镜等大型巡天项目，将把天文学指引到全新的数据密集型时代，使得天文数据急剧增加，数据采用PB，甚至EB来计量，天文学已经步入大数据时代[1,2]，即将迎来大发现时代[3]。正是在这种背景下，天文统计学和天文信息学应运而生。

天文统计学是一门探讨如何从不完整的信息中获取科学可靠的结论，从而进一步进行天文学研究的设计、取样、分析、资料整理与推论的学科。它是天文学、天体物理学与统计学相结合形成的一门新型学科，应用统计学的理论和方法来解决天文学中面临的一切统计学问题。

天文信息学[4,5]是研究天文信息的获取、处理、存储、传输、分析、挖掘和解释等方面的学科。它是天文学、天体物理学、计算机科学、工程学和信息学相结合的一门新型学科，应用天文学、天体物理学、计算机科学、工程学和信息技术揭示大量复杂的天文数据所赋有的宇宙和天体的奥秘，主要是为了应对下一代望远镜产生的按指数增长的数据量、数据产出率和数据复杂性而面临的挑战和机遇。天文信息学即天文信息化，正在推动21世纪天文学由发现驱动和假设驱动到数据驱动和计算驱动的科学转型，数据密集型天文学研究方式已开启。

2 天文数据的分类和特点

在大型巡天时代到来之际，收集到的天体信息的数据量、质量和丰富性达到了前所未有的高度，三者紧密相关。巡天数据量急剧增长，可以获得全天的图像和星表；巡天数据的质量达到了定点观测的质量；星表不仅包含了天体的位置和星等信息，也包含它们的形状、轮廓、时间演化等丰富的信息。现代的大型巡天不仅仅使得数据体量增大，而且使得数据质量更好，富含的信息更丰富，远远超出天体认证的范畴。例如：阿塔卡玛大型毫米波/亚毫米波天线阵（Atacama large millimeter/submillimeter array，ALMA）、甚大天线阵（very large array，VLA）、SKA的日产数据量分别为250 TB、1.5 PB、0.5~10 PB；斯隆数字巡天（Sloan digital sky survey，SDSS）、全景式巡天望远镜和快速反应

系统(panoramic survey telescope and rapid response system, PanSTARRS)、LSST每晚产生的数据量分别为200 GB、700 GB、15 TB。可以看出,天文大数据已经摆在面前。数据科学家Kirk Borne针对高速变化和高度复杂的大数据的获取、清洁、管理、集成、存储、处理、索引、搜索、共享、传输、挖掘、分析和可视化等任务,提出了大数据具备“10V”特征[①],即体量大(volume)、复杂性高(velocity)、高速获取率(velocity)、真实性(veracity)、有效性(validity)、价值(value)、可变性(variability)、存储场所(venue)、相关词汇(vocabulary)、模糊性(vagueness)。这“10V”特征为解决大数据面临的挑战提供了参考。

天文数据以其获得方式的不同,主要分为观测数据、数值模拟数据;以其存在方式的不同,可以分为图像、星表、光谱、时序数据、网页数据;以其结构的不同,可以分为结构化数据、半结构化数据、非结构化数据。天文数据的特点包括:空间性、高维性、海量性、多模式、多尺度、多分辨率、时序性、缺值性、带误差、异构性、分布性、开放性等。天文数据的复杂性体现在:高光谱、非线性、非高斯噪声、非线性系统影响、形状和大小的变化、密度的变化、近似性、局部维数不同等。这些特点和复杂性都对天文数据的存储、传输、处理、分析和挖掘提出了严峻的挑战。

3 天文学研究的方法和方式转型

过去,分布的、异构的天文数据资源只有几百或上千个;而今,各个天文数据资源可以通过虚拟天文台接口统一获得;未来,随着大型巡天项目的发展,天文学将成为更加数据密集型的学科。天文数据的状态发生了翻天覆地的变化:由数据匮乏变为数据过剩,数据集扩展为数据流,静态数据演变为动态演化数据,任意时刻的数据转化为实时或准实时数据,数据集中存储到数据分布存储和计算,数据机构所有权转变为学科数据所有权。这些因素成为天文学发展和进步的动力和源泉,同时也为天文学的发展提供了机遇和挑战,促进天文学研究方式和方法的转型。

大型巡天时代,天文学面临着许多需要信息学和数据科学来解决的科学问题,例如:不同星表的概率交叉认证、距离估测、恒星和星系分类、图像中的宇宙线识别、超新星的寻找与分类、各种天体的形态分类、新类型天体或新类型子类天体的发现、分类器分类规则的提高、大型数据流的分类、天文事件的实时分类、大型数据集的聚类、大型数据集中的新奇、反常和异常的现象或天体的探测等[②]。

这一系列天文问题的解决,需要借助新的技术和手段。在做分类、聚类、相关分析、离群探测、时间序列模式发现等数据挖掘任务时,机器学习是关键技术。机器学习针对大数据的复杂性时,也面临着诸多挑战,如数据预处理、特征选择、降维、算法和模型的选择、数据不完备、不确定性估计、可扩展性、可视化。许多复杂数据存在超维结构(如聚类、相关性等),维数达到百维甚至万维,而且还在持续增长,需要选择合适的挖掘算法,并对挖掘结果进行解释。关于天文学中的数据挖掘和知识发现可参考参考文献[6-12]。

图灵奖得主、关系数据库的鼻祖Jim Gray提出了科学研究的4个“范式”:第一范式,实验科学;第二范式,理论科学;第三范式,计算科学;第四范式,数据密集型科学。在各行各业数据蜂拥阶段,数据密集型科学成为当前科学的主流。**图1**给出了假设驱动与数据驱动的科学的流程,目前

① https://www.mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs

② http://www.lanl.gov/conferences/salishan/salishan2010/pdfs/Kirk%20Borne.pdf

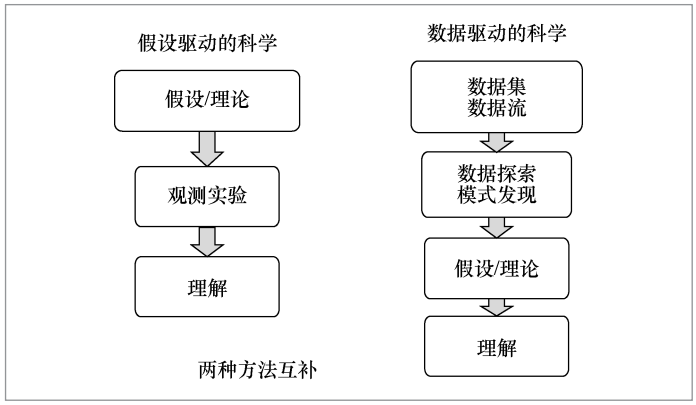


图1 假设驱动的科学与数据驱动的科学

数据驱动的科学占据主导地位，实际应用中这两种研究方式和方法互为补充。图2给出了小数据时代与大数据时代从数据到知识的过程。在小数据时代，科学家从望远镜获得数据后，需要自己亲自动手完成从原始数据到科学数据的转换，科学家需要花费大量的时间和精力进行编程、分析、预处理数据，最后再应用专业知识推理和导出模型。从数据到知识的整个过程，除数据由望远镜采集外，其他的工作都由科学家自己完成，在数据样本不大的时候，这样的科学模式是可行的。而在大数据时代，望远镜采集数据后，数据需要存储、传输、预处理，最后形成星表以供科学家使用。由于各个项目的望远镜系统、地理位置和环境、大气条件等情况不同，导致系统误差不同，相应的数据处理系统也应根据项目的需要而制定，这样一个项目需要具备一个数据预处理系统。这就需要大量人员的时间和精力投入数据预处理中。从望远镜观测得到的原始数据，需要经过降噪、仪器定标、图像处理、图像压缩、图像转化等一系列操作，才可以得到供科学家直接使用的星表，这样科学家可以专注于自己的科学，从而做起科学来更加容易、便捷。可以看出，在大数据时代，个人独立做科研的时代已经渐行渐远，合作科研方式正式开启。

③ http://www.ivoa.net/

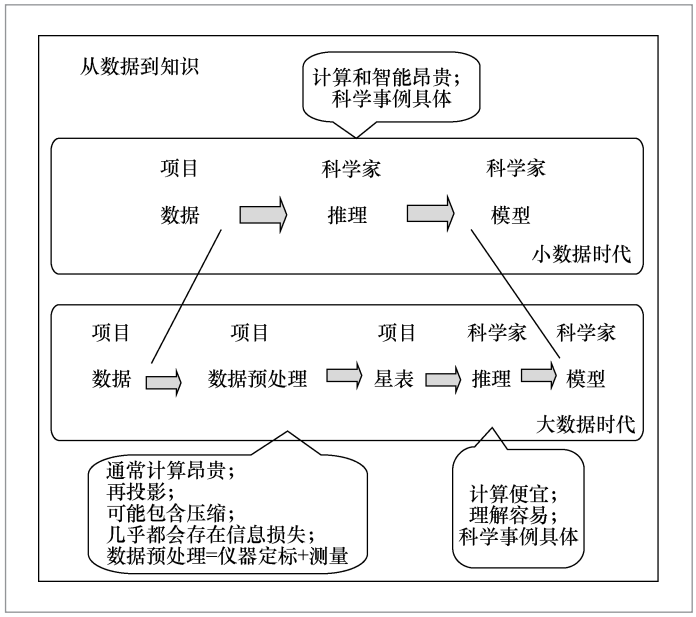


图2 从数据到知识的过程

4 虚拟组织的兴起

身处大数据时代，科学家必须意识到合作的重要性，而且必须是多学科多领域的合作，从而推动科研项目的发展，提高科学产出。科学社区针对大数据带来的挑战和机遇采取了应对措施，如：面向具体领域，不是以单位为基础，而是以分布的人员、数据、计算资源等构成的新型科学组织为基础。天文领域的国际虚拟天文台联盟③为天文大数据提供了一个完整的研究环境。自从2000年虚拟天文台发起至今，虚拟天文台取得了骄人的成绩，为广大天文学家所接受和使用。其成功在于多方面原因，如：所有数据以电子形式收集；精通计算机技术和数据技术人员的参与；制定统一的标准格式；收集的大数据都是有资金资助、机构授权的数据集；建立了数据共享的文化；动机来源于指数增长的数据；联合机构的支持/资助；数据没有商业价值和隐私。虚拟天文台优于其他领域的虚拟组织，有诸多优点，如：互操作和标准等的进步；拥有

全球的天文数据网格和助力天文社区的科研；提供有用的网页服务；为社区提供了培训、科普教育资源。虚拟天文台还有不尽完美之处，如缺乏数据探索和挖掘工具，这正是科学产出之所在。不过他们已经意识到这方面的缺陷，正在加大力度发展。

多学科领域的交叉（如科学、计算机、信息技术等）和更广泛的社区参与（贡献者和用户），提供了交叉科学方法共享的机制和有用的网页服务，大大方便了不同学科的互通互联。对大多数的数据科学（如计算、机器学习、统计学等）而言，都面临着共同的挑战，一些应对方法也可以共享。如何将信息架构的发展、经验和解决方案从一个领域挪用到另一个领域？目前已有一些较为成熟的项目值得借鉴和学习。数据驱动的科学发现中心（CD3）是美国加州理工学院新成立的研究中心[④]，服务于全校的研究项目（如天文、物理、生物、地球物理等），是新的Caltech-JPL联合数据科学和技术中心的一部分。该中心的目标是帮助员工推进和执行数据密集项目，共享交叉学科的方法、思想、项目实施等成果。面向对象的数据技术（object oriented data technology，OODT）是要构建分布式数据密集系统的Apache开源框架[⑤]，用于获取和共享分布的资源，1998年获美国国家航空航天局（National Aeronautics and Space Administration，NASA）资助。它应用于行星科学、射电天文、地球科学、医疗卫生、气候检测、癌症诊断等，是NASA的第一个天文开源项目，同时也是基于Apache软件基金会的顶级项目，曾获得2003年度NASA软件比赛的第二名。现在数据集的高信息量足以驱动有价值的数据挖掘。数据融合可以发现那些仅靠单数据集无法发现的知识。数据的复杂性需要人工智能帮助人们理解和认识，未来虚拟现实将成为终极前沿。所有的信息和工具都可以通过网页联系在一起，网页已成为人们彼此联系、工作、学习的主要方式，数据的获取、处理方法、论文发表、教育等都可以通过网络空间来实现。

5 大数据时代的项目典范——LSST

LSST是位于智利的一架直径8.4 m的天文望远镜[⑥]，摄像机达到了32亿像素，可以拍摄出6个波段的图像，预计2020年投入使用。以聚光能力和视野宽度来说，LSST的规模是目前现役的和正在建造的任何一个巡天望远镜的10倍以上。每周可以对整个南半球天空巡查两次，每晚数据量将达15 TB。LSST预计在第一个月的运行时间内将观察到比以前所有望远镜加起来还要多得多的宇宙空间。LSST主要用来研究暗宇宙、宇宙的瞬间、太阳系的细节、银河系图像。

LSST项目的产品将分为3级：第一级是每晚探测的约1 000万个时间事件流和太阳系内约600万个天体的轨道星表，这些时域事件要在观测到的1 min内传送到相应的发布网络中；第二级是每年产生的约370亿个天体的星表（包括200亿个星系和170亿颗恒星）、约7万亿单历元探测事件、约30万亿的约定要观测的源，这些数据都可以通过在线获得，另外包括深的叠加图像；第三级是在数据中心可以将用户定制的处理和分析的服务以及计算资源提供给用户。就第三级而言，使天文学界可以基于LSST的软件、服务或计算资源创建新产品，即：基于已有的软件服务定制适合的测量和推理代码，让用户可以在LSST数据中心运行自己的代码，从而平衡投入产出比。在不久的将来，对于大型巡天项目，释放的数据会作为其主要产品，在巡天结束之际，软件以及由这些软件处理产生的特殊或暂源星表，同样也会作为其主要产品。释放的数据会作为

④ http://cd3.caltech.edu

⑤ http://oodt.apache.org

⑥ https://www.lsst.org/

所有星表的一部分，更频繁地广泛使用，而且会保留很长时间。LSST软硬件的总体设计思路就是要让这一切变为现实。

目前为止，还没有能力实现最优化地从数据到模型的推导，这主要是因为计算密集、I/O密集和数据量超大，此时的星表就覆盖面积而言已经是全天星表，而且精于多学科（如统计学、应用数学、软件工程等）的专家还未出现。不断增长的计算能力可以将天文学家从计算能力不足的困境中解脱出来。到2020年，大部分天文学家将成长为拥有多学科知识的生力军。另外，参与大型巡天项目和望远镜项目的人们，应该创建与这些项目相关的必要的软件和数据处理系统，这样天文学家才可以真正地投入大数据的洪流中一展身手。

传统天文学是数据饥渴的科学，研究的方式和方法自然受限于这种数据缺乏的状态。天文学过去是受硬件束缚的学科。那时天文学家对天文数据处理软件系统一直不太重视，甚至不需要复杂的编程。由于数据少而质量差，亦或没有足够的计算能力来正确处理数据，天文学家一度是幸运的，只是选择一些源观测，一个晚上最多观测上百来颗天体，一些简单的算法足以应付少量数据的处理，根本的挑战在于硬件。如今，天文学成为软件束缚的学科。各种情况发生了质的变化，收集的数据空前巨大。现在的望远镜和照相机技术飞速发展，达到了空前的水平，可获得高质量的数据。此外，计算机的能力也足以胜任比较复杂的计算，而且价格也越来越便宜。滞后的内存容量迫使人们追求更加复杂的算法，开发更好的软件处理系统。为了最好地发挥昂贵望远镜的效能，现在的大型项目从设计之初就开始软件与硬件同步发展。巡天计划的出现，使得数据从广度和深度上变得更加丰富起来，LSST正是这种转换模式的典范。在这种状态下，研究的成功将依赖于从现有的数据中挖掘出知识的能力。建造更大的仪器已经不再是最划算的事情。发展与之匹配的软件项目势在必行。这必将重新点燃类似虚拟天文台概念的驱动，不过这个驱动将侧重于算法、工具和已经成长起来的软件框架的再利用和合作。AstroPy和LSST项目已经开始推动类似的项目前行。

6 与天文统计和天文信息相关的国际组织和公共教育

天文大数据的核心价值在于如何从浩瀚的数据海洋中发现隐藏于其中的瑰宝，即发现人们感兴趣的稀有天体或现象，亦或未知天体或现象，从而推动天文学理论的发展。而推动这一过程实现的桥梁需要借助天文统计学和天文信息学。这两门新生学科，尤其天文信息学是近几年才蓬勃发展起来的，是天文学发展到一定阶段的必然产物。天文学研究的方式和方法是应时代而发展变化的，小型数据研究大部分是假设驱动，单个科研工作者足以胜任；中型和大型数据研究亦或是假设驱动，更准确地说是数据驱动，合作研究成为主流。正是在这种形势下，各种与天文统计和天文信息相关的国际组织成立。它们的共同目标就是促进大数据时代天文学的快速发展，培养面向大数据的新一代人才。

6.1 与天文统计和天文信息相关的国际组织

国际天文统计学会（International Astrostatistics Association，IAA）⑦于2012年正式成立，总部设在意大利米兰的布雷拉天文台，是全球第一个致力于天文统计学和天文信息学的科学学会。目前，会

⑦ http://iaa.mi.oa-brera.inaf.it/adm_program/modules/announcements/announcements.php

员数已超过550人，来自56个不同的国家和地区。学会的第一终极目标是加强天文学家/天体物理学家与统计学家的合作；第二个目标是让天文学家更好地理解新的统计方法，从而提高对天文大数据的分析和解释。学会不时地为会员提供详细的关于天文统计的文章、即将举行的会议、工作组和教育资源等信息。

国际天文学会天文信息和天文统计委员会[8]起源于2012-2015年度的国际天文学会天文统计与天文信息工作组，于2015年正式成立，主旨是促进现代的计算和统计方法在天文研究领域的应用。同年，国际天文学会时域天文学工作组正式成立，旨在协调全球变源的研究，包括变星和双星、地外行星、吸积天体等以及在射电波段、可见光波段和高能波段在内的多时域的巡天。并且美国天文学会也成立了时域天文工作组，鼓励和促进与时域天文相关的活动和合作，如科学运用时序数据、与其他波段信息的整合，进一步提高科学产出；组织和资助与时域天文相关的会议、工作组、小型会议。

电气与电子工程师协会挖掘复杂天文数据的特别工作组[9]成立于2014年，隶属于IEEE计算智能学会的数据挖掘技术委员会，主要目标是解决现代天文学面临的问题，即如何将全天的无穷尽的数据流高速地转化为知识。广义而言，尽管急切需要机器学习与数据挖掘、计算智能方法，但是这些方面还有待进一步研究和开发。天文学家也开始广泛涉猎这些领域，保证工作组的出现和努力完全是为了更好地支持和推进天文科学的研究。计算机领域专家的有意义的贡献是与天文学家紧密合作分不开的。因此，组员将努力参与天文数据分析项目，而且最初组员的选择也反映了这种思想。

美国天文学会的天文信息和天文统计工作组[10]是以“方法、人才、会议”为宗旨，有3个主要战略目标：开发、组织和维护方法资源，如软件工具、文章、书籍、报告、研讨会以及其他的教育资源；提高人力资源，如建立演讲团、构建职业规划、建立存档的论坛、保证定期的新闻发布；组织专题会议。在2015年美国天文学会上，该工作组专门讨论了在天文领域计算和统计需求日益增加的情况下的教育和职业规划问题。

大型综合巡天望远镜的信息和统计科学团组[11]旗下有40余名数据科学家，主要致力于开发适合大型天文巡天数据的工具。该团组成员包括天文学家、统计学家、计算机科学家、机器学习专家，他们有一个共同的目标，就是要解决LSST面临的挑战，从而更好地实现LSST的科学目标。

美国统计学会的天文统计兴趣组[12]成立于2014年，搭起了统计学家与天文学家合作的桥梁。

天文统计和天文信息门户网站[13]是一个全新的服务于天文学家、统计学家和计算机科学家的交叉学科社区的网站。该网站由国际天文学会的天文统计和天文信息工作组主席Eric Feigelson和国际天文统计学会主席Joseph Hilbe维护和编辑。该网站的目标是致力于促进天文领域的高级算法研究和加强这些方法在更广阔的天文领域的应用。该网站提供了天文统计和天文信息领域的最新的文章摘要、一些相关课题的论坛、研究、专家文章、会议、各种各样的网络资源，如在线的课程、书籍、工作和博客等。该网站服务于国际天文统计学会、美国天文学会的天文信息和天文统计工作组、国际天文学会的天文统计和天文信息工作组、大型综合巡天望远镜的信息和统计科学团组、美国统计学会的天文统计兴趣组5个组织的公共教育。

6.2 与天文统计和天文信息相关的教育

科学家必须具备这种意识：将已有的

⑧ https://asaip.psu.edu/organizations/iau-commission-on-astroinformatics-and-astrostatistics

⑨ https://asaip.psu.edu/organizations/ieee-astrominer-task-force

⑩ https://asaip.psu.edu/organizations/aas-working-group-in-astroinformatics-and-astrostatistics

⑪ https://issc.science.lsst.org/

⑫ http://community.amstat.org/astrostats/home

⑬ https://asaip.psu.edu/

合适的分析技巧和方法应用到实际数据中，从而推导出最好的结论。然而，这种意识只有在天文统计和天文信息领域获得了足够多培训的时候才可以具备。目前的课程还远远不够，况且这些课程还没有对天文学系的学生开放或要求。实际科研中，科学家在软件编程方面比较薄弱的时候，还需要亲自动手写代码或软件。在“天文领域的软件使用用户：一份非正式调查”的报告中指出：所有使用软件的天文学家中90%的用户自己写代码，其中仅有8%的用户接受过扎实培训[14]。因此，美国天文学会的天文信息和天文统计工作组在美国西雅图2015年第225届美国天文年会上组织了一个关于天文研究中的天文信息和天文统计：迈向更好的课程的研讨会。会中提出了若干关于教育的议题，主要结论如下。

[14] https://asaip.psu.edu/

- 教师的培训是很重要的。非教学时间教师的培训会有帮助。在线有很多课程，应该给出一些指导性的建议，如哪些课程值得学习。
- 教师进修课程所需的经费是需要筹备的。这样的培训组织也是需要的。
- 天文学界要鼓励学生和职工从事数据科学方面的培训。
- 一些导师可能不支持学生参加数据科学方面的培训。如果导师不支持，学生就难于参加培训。
- 课程的变更常常是不可能的，或者说需要很长时间来推行。增加一门新课程意味着要减少一门其他课程。这样的修改需要经过缜密的思考和具备强大的动力。可以改变一种策略，课程变为必修课和可选课。如果可选课设计组织得好，有趣并且广告做得好，那样选择的学生人数自然会上升。
- 并不单单是天文学在这方面捉襟见肘，其他学科也面临同样的问题。可以与其他领域的同行交流，与领导沟通，尽可能实现大学课程的调整。
- 人才流失经常会发生而且是不可避免的。能够跟上学术领域有才华的科学家的步伐是比较具有挑战性的。在天文大数据时代，最好能与工业领域并驾齐驱，同时应该意识到数据科学是现代天文研究的重要组成部分。

针对天文领域出现的统计和信息问题，一些天文学家根据自己的科研经历总结和编写了天文统计和天文信息方面的书籍。**表1**列出了2012年以来相关的书籍。目前，天文信息方面的书籍还比较匮乏，还待这方面的有识之士去编撰。

关于天文统计和天文信息方面的大量的非正式培训不时地在各地开展。各种各样的会议和暑期班也相继举办，会议包括：现代天文学中的统计挑战、天文信息会议、21世纪宇宙学统计挑战会议、天文数据分析软件和系统会议等。暑期班包括：天文数据分析暑期班、天文学家的统计暑期班、天文大数据工具等。琳琅满目的关于统计和信息的资源可以在线获得。例如：通过YouTube可以找到16 900个关于贝叶斯计算的视频、7 710个关于非线性回归的视频、11 300个关于计算天体物理的视频。通过注册可以接受正式的网络培训课程（如Coursera、statistics.com等），尽管这些课程是面向其他领域的。数据海洋列出了直接面向教育的资源。研究者可以根据自己的需求选取适合自己的课程或学习资源。

7 结束语

天文学中仍然存在着许多未解问题，如什么是暗物质和暗能量？宇宙在诞生之初的10^{-35} s是否经历了膨胀阶段？星系是如何形成和演化的？到底有多少颗太阳系外行星？它们是否存在智慧生命？这些问

表1 2012年以来与天文统计和天文信息相关的书籍

书名	作者	出版年
Data-rich astronomy: mining synoptic sky surveys	CAVUOTI S	2015年
Bayesian methods for the physical sciences: learning from examples in astronomy and physics	ANDREON S, WEAVER B	2015年
Statistical methods for astronomical data analysis	CHATTOPADHYAY A K, CHATTOPADHYAY T	2014年
Statistics for astrophysics: methods and application of the regression	BURNET D F, VALLSGABAUD D	2014年
Astronomy and big data: a data clustering approach to identifying uncertain galaxy morphology	EDWARDS K J, GABER M M	2014年
Statistics, data mining, and machine learning in astronomy: a practical guide for the analysis of survey data	IVEZIC Z, CONNOLLY A J, VANDERPLAS J T, GRAY A	2014年
Astrostatistical challenges for the new astronomy	HILBE J M	2013年
Advanced statistical methods for astrophysical probes of cosmology	MARCH M C	2013年
Statistical challenges in modern astronomy V	SCHAFER C M, FREEMAN P E, HENDRY M A, WANDELT B D, JASCHE J	2012年
Modern statistical methods for astronomy with R applications	FEIGELSON E D, BABU G J	2012年
Practical statistics for astronomers	WALL J V, JENKINS C R	2012年
Advances in machine learning and data mining for astronomy	WAY M J, SCARGLE J D, ALI K M, SRIVASTAVA A N	2012年
Astrostatistics and data mining	SARRO L M, EYER L, O' MULLANE W, RIDDER J D	2012年

题的解答必将受益于大数据分析。面对天文大数据提出的挑战，知识发现工具有待改进和提高，如可用性、可扩展性、互动的数据挖掘和可视化。在机器学习/人工智能方面，实现协作的人机发现；在超维数据的可视化方面，加强人们的理解和认知力，实现可视化的数据探索和发现；在社区的参与和事业规划方面需要克服智能和方法的惰性，实行奖励和鼓励机制；需要探讨新型的发表和合作方式，超出发表论文的范畴，开发较好的合作工具；培养和造就下一代的大数据科学家。计算机技术和数据的指数级增长引发科学的转型，无论从定性还是定量上，数据驱动的科学有别于传统科学。任何一门数据密集型科学都面临着许多共同的挑战，它们的解决方案促进新的科学方法产生。大数据科学应用的核心是人的素质，专业天文学家和公众在数据科学方面的教育和素养必须大大提高。大数据时代标志着天文学家独自搞科研的时代结束，分享、合作、共赢成为大数据时代的主旋律，随着天文统计学和天文信息学的蓬勃发展和深入应用，天文学发现的黄金新时代即将来临。

参考文献:

[1] FEIGELSON E, BABU B. Big data in astronomy[J]. Significance, 2012, 9(4): 22-25.

[2] ZHANG Y, ZHAO Y. Astronomy in the big data era[J]. Data Science Journal, 2015, 14(11):1-9.

[3] TYSON J, BORNE K. Future sky surveys new discovery frontiers[M]//Advances in machine learning and data mining for astronomy. Boca Raton: CRC Press, 2012: 161-181.

[4] BORNE K D. Astroinformatics: a 21st century approach to astronomy[J]. Kirkborne Net, 2009, 42: 578.

[5] BORNE K D. Astroinformatics: data-oriented astronomy research and

education[J]. Earth Science Informatics, 2010, 3(1): 5-17.

[6] DJORGOVSKI S, DONALEK C, MAHABAL A, et al. Some pattern recognition challenges in data-intensive astronomy[C]// The 18th International Conference on Pattern Recognition, August 20-24, 2006, Hong Kong, China. [S.l.:s.n.], 2006: 856-863.

[7] 张彦霞, 赵永恒. 数据挖掘技术在天文学中的应用[J]. 科研信息化技术与应用, 2011, 2(3): 13-27.
ZHANG Y X, ZHAO Y H. The application of data mining technologies in astronomy[J]. E-science Technology & Application, 2011, 2(3): 13-27.

[8] 张彦霞, 赵永恒, 崔辰州. 天文学中的数据挖掘和知识发现[J]. 天文学进展, 2002, 20(4): 312-323.
ZHANG Y X, ZHAO Y H, CUI C Z. Data mining and knowledge discovery in database of astronomy[J]. Progress in Astronomy, 2002, 20(4): 312-323.

[9] BALL N M, BRUNNER R J. Data mingand machine learningin astronomy[J]. International Journal of Modern Physics D, 2010, 19(7): 1049-1106.

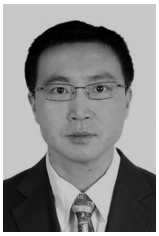
[10] DAS K, BHADURI K. Parallel and distributed data mining for astronomy applications[M]// Advances in machine learning and data mining for astronomy. Boca Raton: CRC Press, 2012: 595-615.

[11] BORNE K. Virtual observatories data mining and astroinformatics[J]. Planets Stars and Stellar Systems, 2013(2): 404-443.

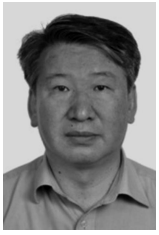
[12] BORNE K. Scientific data mining in astronomy[M]// Next generation of data mining. Boca Raton: CRC Press, 2009: 91-114.

作者简介

张彦霞(1974-), 女, 博士, 中国科学院国家天文台研究员、硕士生导师, 北京天文学会副秘书长, 国际天文统计学会会士, 国际天文学会天文信息与天文统计委员会委员。曾任国际天文学会天文统计与天文信息工作组委员会委员。主要从事天文大数据、天文信息学、天文统计学、多波段天文学等研究工作。在国内外学术刊物上发表论文100余篇。



崔辰州(1976-), 男, 博士, 中国科学院国家天文台研究员、硕士生导师, 国家天文台信息与计算中心主任, 北京市科技新星计划入选者。率先在中国科学院国家天文台建立起集科学数据库、高性能计算、信息化环境研发三位一体的科研信息化研究与服务体系。多年来, 一直以科技资源整合与共享为研究主线, 推动虚拟天文台的发展, 已发表论文70多篇。



赵永恒(1964-), 男, 博士, 中国科学院国家天文台研究员、博士生导师, 中国科学院“百人计划”和国家“万人计划”入选者。现任中国天文学会常务理事, 曾任国家重大科学工程LAMOST项目总经理、北京天文学会理事长、国际天文联合会第五委员会科学组织委员会委员、世界数据中心天文学科中心主任。主要研究方向为活动天体的理论研究、高能天体的观测分析、天文数据分析、天文信息技术等, 在国内外学术刊物上发表论文300余篇。

收稿日期: 2016-08-03
基金项目: 国家重点基础研究发展计划基金资助项目(No. 2014CB845700)
Foundation Item: National Key Basic Research Program of China (No.2014CB845700)