

大数据背景下的应用统计专业硕士能力提升研究

成果总结

成果完成人：阮敬、吴启富、任韬、张玉春

一、研究背景和意义

大数据已经深入到国民经济生活的各个方面，大数据分析正日益成为人们社会经济生活进行预测和决策的重要手段，大数据分析能力也已成为现代社会的基本技能之一。目前，大数据方面的研究在美国、日本等发达国家已经上升为国家战略。然而在我国高等教育人才培养中，对大数据的关注还不够，传统数据收集和分析技术的知识结构已不能满足大数据时代的高要求，社会亟需具有大数据视野和分析能力的综合性人才。统计学专业作为研究数据的科学，已经在与传统计算机技术、数据库技术、网络技术等专业领域逐步融合，从学科发展趋势上看，需要对高层次统计专业人员进行培养，以适应大数据时代的来临。

我校是全国第一批获批应用统计专业硕士学位授予权的单位之一，同时也是全国应用统计专业硕士的发起单位之一。本项目负责人和主要项目组成员在大数据发展的前期阶段已经抓住先机，与北京大学、中国人民大学、中国科学院、中央财经大学联合成立了“大数据分析硕士联合培养协同创新体”，在国内率先培养大数据分析的高级人才。

因此，本成果的研究将对改革现有应用统计专业硕士人才培养方案，探索能够满足大数据分析需求的人才培养模式，设置合理的课程规划及其具体内容，使我校及统计学院在大数据时代抓住先机，迈入学科发展的前列等方面具有重大意义。

二、立项依据

1、研究现状

大数据除去概念观其本质有二：第一，如何架构大数据环境来管理、存储和提取数据；第二，如何根据“在线”和“实时”两种模式从数据中发掘出价值。

第一个本质其所关注和研究的是大数据硬、软件架构的问题，需要掌握服务器设备、存储设备以及必要组网设备（如交换机）、大规模集群系统、基于跨物理服务器的多点系统数据库系统、分布式文件系统、MapReduce 计算框架、实时流计算、内存计算、图计算框架等软硬件综合知识；该方面是计算机科学领域所关注的重点技术和内容。

第二个本质所关注的内容主要为大数据应用以及不同数据的关联，可以主要细分为两个重要方面：如何将传统单机数据运算与数据分析方法应用到分布式文

件系统的集群中以提升运算效率，需要进行算法应用的研究；如何在大数据架构基础上将已有分布式计算方法（统计分析方法、数据方法以及机器学习算法等）应用到解决实际问题当中。这两个方面是统计学关注的重点内容。

综合以上分析，大数据系统架构是计算机科学领域中较底层的分布式文件存储和管理系统（当前较为主流的开源体系为 Hadoop/MapReduce 框架，此外还有实时流计算、分布式内容计算、图计算框架等与之并存），大部分的商业软件如 SAS、Matlab 均可以将其直接整合，架构在其上（用户无需过多关注数据底层），面向用户提供数据管理和分析功能；同时一些开源软件如 R 和 Python、Perl、Ruby 等面向对象的开发性质脚本语言也可以运行于这些系统中，提供数据分析功能。

从美国 20 多所院校的著名大数据项目来看，他们的项目课程也是围绕上述两个本质性的问题来设计的。首先是以卡内基梅隆为代表的以大数据底层系统为主的项目，其所研究的是大数据系统架构的问题；其次是以北卡州立、麻省理工 MBA 等为主的大数据分析应用硕士，其所关注的重点内容为大数据分析在各领域中的应用问题。

2、本项目的实践基础

本成果负责人及主要成员是应用统计专业硕士大数据分析方向协同培养创新体的核心成员及创始人员，具备极佳的教学改革研究经验和资源。

本成果认为，培养大数据分析方向的应用统计专业硕士应当着力于“分析”能力的培养，这样才能在计算机领域中突出统计特色，在大数据时代才能占有一席之地。即：将大数据架构基础上的数据存储和分析环境作为解决问题的基本工具，培养学生搜集数据、整理数据、分析数据、呈现数据的基本能力，并将这些能力应用于解决实际工作的具体问题。在培养分析能力的基础上，需要开设一些必备的分布式运算以及系统开发等计算机技能选修课程，从而满足社会广泛需求。

成果研究组成员对大数据分析硕士培养的相关问题研究具有相当丰富的前期调研成果和经验。我们创建了我校“STATISTICA 北京实训基地”和“JMP 实训基地”，与世界知名统计软件制造商 SAS、JMP、IBM SPSS、STATSOFT 和中国工商银行、中国移动、中国石油、京东、阿里巴巴、优酷、百度等业界翘楚有紧密的合作，为大数据分析硕士人才培养提供了实习实践实训条件；同时这些实训基地的合作方也跟我校有紧密的学术科研活动，能够为培养我校大数据分析人才的综合实践能力提供有力的实践指导。

此外，为吸取国外大数据分析硕士的教学培养模式，成果完成人及其成员于 2011~2014 年在美国中佛罗里达大学统计系、美国佛罗里达大学统计系、美国罗格斯大学统计系等均做过为期 1 年的访问学者，对本成果所研究的主要内容作了充分的准备工作，具有国际一流大学培养大数据分析人才的一手资料，并亲自参与了国外院校大数据人才培养的规划和实施工作。

此外，本成果完成人近年来还曾就大数据分析硕士人才需求、大数据分析硕士课题体系及教学内容、大数据分析硕士实践实训课程、大数据分析硕士教学效

果评估等具体内容亲自访问过耶鲁大学、北卡罗来纳大学教堂山分校、北卡州立大学、芝加哥大学、密歇根大学安娜堡分校、佛罗里达大学、英国伦敦政治经济学院、伦敦大学学院、日本岛根县立大学、京都大学、台湾辅仁大学、台湾淡江大学、香港科技大学、香港城市大学、澳门科技大学等几十所相关院校的统计院系以及美国 SAS 公司总部、美国 StatSoft 公司、美国蓝十字蓝盾保险公司、美国迪斯尼公司、美国环球影城公司、台湾“全国”意向调查公司、台湾史丹索特公司等诸多行业翘楚，就大数据分析硕士人才培养方面的具体问题及人才需求状况进行过翔实的调研，并搜集了一手数据和重要资料。

本成果在此基础上就我国大数据分析硕士人才培养的指导理论、课程系统设计、培养方案设计和人才培养过程中的目标、模式以及本土化人才向国际化实践创新人才转化途径等问题展开了深入研究。

三、研究内容

1、研究内容、目标和要解决的问题

第一，对现有大数据分析理论进行了梳理，研究了大数据理论与统计理论之间的关联，为应用统计专业大数据分析硕士人才培养提供理论基础。

大数据分析起源于互联网数据的高度聚集和流动。聚集数据往往以分布式形式存储在网络中的不同节点上，这需要对传统数据分析方法进行分解（MAP）和规约（REDUCE）上的改进才能快速得到分析结论；流动数据则大都以数据流的形式临时存储与内存当中，这同样需要对传统数据分析进行流分析的改进。研究这些新数据分析方法与传统统计分析理论之间的区别与联系，有利于培养学生的大数据分析思维，从而为应用统计专业大数据分析硕士的培养提供理论基础。

第二，通过大数据分析人才的需求分析，以应用角度来设置应用统计专业大数据分析硕士的课题体系与内容，培养学生的大数据分析能力。

大数据分析方向的硕士课程应当着力于“分析”能力的培养，这样才能在计算机领域中突出统计特色，在大数据时代才能占有一席之地。即：将大数据架构基础上的数据存储和分析环境作为解决问题的基本工具，培养学生搜集数据、整理数据、分析数据、呈现数据的基本能力，并将这些能力应用于解决实际工作的具体问题。在培养分析能力的基础上，需要开设一些必备的分布式运算以及系统开发等计算机技能选修课程，从而满足社会广泛需求。

第三，将大数据分析人才培养目标与行业应用相结合，培养真正对社会有用的复合应用型人才。

行业大数据应用往往建立在已有大数据存储系统中，如主流的 Hadoop、Storm 等大数据分析环境。如何将这些主流分析环境引入课程，在课堂中构建真实的数据收集、数据处理和数据分析环境，以案例教学的形式生动而真实的模拟实际解决问题的需求，让学生在校园中就能充分接触到行业应用的前沿，这是本成果的主要内容之一。

第四，将大数据分析硕士的师资力量整合，实现科研部门、教学部门与实际行业部门之间的联合培养协同创新。

大数据时代的数据之间具备一定的必然联系，培养大数据分析人才的师资力量在不同学校、不同学院以及实际行业企业或部门中都是有联系的。本成果依托大数据分析硕士联合培养创新体，将各校中涉及到的计算机学院、信息学院、统计学院等及其他数据应用领域的实际部门中的师资力量进行整合，统一规划培养课程和具体内容，共同培养适应社会发展潮流的大数据分析高端人才。

第五，在联合培养大数据分析硕士协同创新体中结合现有应用统计专业硕士研究生培养计划进行大数据课程的衔接与课程内容规划等具体问题，为大数据分析硕士培养提供制度保证和具体规划措施。

2、主要特色和创新

第一，本成果整合了国内著名高等院校、科研机构 and 行业实际部门的师资力量，在我院应用统计专业硕士培养基础上联合培养大数据分析人才，所依托的大数据分析联合培养协同创新体在培养模式上具有较强的创新；同时使得我院师资力量与国内外著名高校师资有紧密交流，有助于提升教学水平；

第二，本成果立足于我院现有应用统计专业硕士的培养，拓展了现有应用统计专业硕士的研究方向，并与国际主流应用统计专业硕士接轨，使我院统计专业高端人才培养走在全国前列并提升我校知名度；

第三，本成果的教学计划和课程规划建立在现有应用统计专业硕士培养的基础上，能够保证联合培养协同创新体所培养的学生所掌握的知识体系保持一致，有助于加强校际学生之间的交流，极大提升我院应用统计专业硕士研究生的科研水平和应用能力；

第四，本成果整合校内外相关实习、实践、实训资源，以点到面充分发挥统计专业类学生的技术优势，在大数据背景下能够拓展本专业的就业方向 and 途径。

四、本研究所取得的教学成果及其实践应用情况

本教学团队经过长达 5 年的研究，取得了如下成果，具体应用情况如下：

1、成果负责人就本成果的主要内容应用在与中国石油天然气集团公司的数据开发合作项目中，指导学生基于我国石油大数据进行了数据分析和预测集成系统的开发，并获得 2014 年全国应用统计专业硕士学位研究生案例大赛及教学成果一等奖奖励。

2、根据研究成果所编写的教材——《实用 SAS 统计分析教程》（中国统计出版社 2013 年第一版）得到了美国耶鲁大学、美国佛罗里达大学、美国中佛罗里达大学、北卡罗来纳大学、SAS 总部、香港科技大学、台湾辅仁大学、台湾淡江大学、澳门科技大学、清华大学、中国人民大学、中央财经大学等国内外著名院校专家的一致好评，并被首都经贸大学、中央财经大学、江西财经大学、安徽财经大学、南京财经大学、河北大学等全国十几所院校选为课程指定教材和重要参考书籍。

3、本成果所提培养大数据分析思维的嵌入式教学模式充分强调与实际应用的紧密结合，成果已获得美国 StatSoft 公司、SAS 公司 JMP 事业部、台湾史丹索特公司等业界翘楚认可并在其业务运营中进行大力推广，本团队成员均被受聘为其培训讲师讲授本成果内容。

4、成果完成人的教学团队按本成果所提理念及教学改进方法，培养并指导研究生获得全国应用统计专业硕士研究生案例大赛一等奖 1 次、全国大学生统计建模大赛三等奖 1 次和优秀奖 1 次，全国研究生数学建模大赛一等奖 1 次，挑战杯大赛一等奖 1 次，以及行业协会统计应用方面的优秀论文奖 1 次（具体材料见附件）。

5、成果完成人按照本成果规划实施的内容，近年来共培养研究生获得国家奖学金 2 人、北京市优秀毕业生 1 人。