数据科学人才的需求与培养

陈振冲，贺田田
香港理工大学电子计算学系，香港 999077

摘要
信息科技业已进入大数据时代。作为能够从大数据中挖掘知识的人才，数据科学家（data scientist）受到各行各业的青睐。首先从美国和中国主要的在线人才招聘平台收集数据，通过对比分析得出数据科学家与传统的数据分析师（data analyst）在工作性质、工作能力要求以及薪资待遇等方面的差别。其次，考察和总结了世界范围内优秀大学数据科学人才培养的概况，并与工业界的实际要求进行对比。根据以上两者之间的差异，就当前大学数据科学人才的培养提出了建议和对策。

关键词
大数据；数据科学；大学教育；人才培养

中图分类号：TP3　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016058

Data science: the demand and development of talents

Keith C C CHAN, HE Tiantian
Department of Computing, The Hong Kong Polytechnic University, Hong Kong 999077, China

Abstract
Information technology has entered the era of big data. As talents who can discover the knowledge in big data, data scientists are tremendously demanded. The differences between data scientists and data analysts in the job nature, entry requirement and even remuneration were presented. Through a careful survey of the current job markets in the US and China. Then, it was revealed the gap between the kind of talents that were required for the jobs and the kind of graduates that the universities were training out. After a gap analysis, the views to the kind of data science programs which we believe may best develop the talents for the current and future job market were presented.

Key words
big data, data science, university education, talent development

1 引言

信息科技已经进入大数据的时代。“大数据”的数据容量(volume)、增长速度(velocity)、多样性(variety)、多变性(variability)和精确性(veracity)相比以往都有了翻天覆地的变化。因此，传统的数据处理方法遇到了前所未有的挑战。大数据带来的巨大挑战，同时也是巨大的机遇。

数据资源是重要的现代战略资源，其重要性在本世纪有可能超过石油、煤炭、矿产，成为人类最重要的资源[1]。因此，如何存储、管理数据，发现大数据中有价值的信息，成为科学界、工业界炙手可热的问题。众所周知，“事”在人为，数据处理的策划、实施的主体是具备专业知识和技能的数据处理人才。大到一个国家，小到一间公司或一个集体，若想充分利用数据带来的价值，必须拥有具有专业知识和技术的数据人才。培养出足够的、合格的数据人才，对我国在未来掌握大数据的核心价值起着至关重要的作用。

在本文中，笔者首先阐述大数据时代催生出的新兴科学——数据科学，之于大数据处理的重要性；而后，通过总结工业界对于时下最热门的两个数据科学方面的职业（数据科学家和数据分析师）的要求，阐述工业界对于数据科学人才的一般要求；而后，再通过对比的方式得出大学教育培养数据科学人才与工业界要求的差异；最后对数据科学人才培养提出行之有效的建议。

2 大数据处理与数据科学

当今社会，伴随着计算机互联网技术的普及与发展，人类社会的诸多活动都会产生大量的数据。比如在科学研究方面，目前生物学的数据每年都以指数速率增长[2]，截至2013年，欧洲生物信息协会保存的基因和蛋白质的数据就达到20 PB[3]。此外，政府和企业的政务以及业务数据的总量也迅速增长。国内一项调查显示，接近30%的国内企业拥有500 TB以上级别的企业数据库用于存储企业数据[4]。截至2014年，全球各国政府和地区仅开放的数据集就已超过700 000个[5]。而人类使用互联网终端产生的数据，更是难以计数。面对如此巨大的数据总量，如何存储、处理、发现数据中有价值的信息，成为科学界、工业界炙手可热的话题。

为应对大数据带来的前所未有的挑战，一个以多学科多技术融合为特点的新兴基础学科[6]——数据科学，已经形成并迅速发展。从宏观角度而言，数据科学是一门利用数据学习知识的学科，其目标是通过在数据中提取的有价值的信息来生产数据产品。数据科学结合了诸多领域中的理论和技术，包括应用数学、统计、模式识别、机器学习、数据可视化、数据仓库以及高性能计算①。从其定义不难看出，数据科学与传统的数据分析学科有一定的交集，但不完全相同。可以看到，数据科学涉及的学科更多、更全面。

知识的获取是整个数据处理过程最重要的组成部分，也是获取海量数据之后的重要目的。目前，数据分析以获取知识的方法传统上称为data analysis，但近年来，data analytics这一术语也经常见到，特别是谈论大数据与社交媒体分析的时候。虽然这两个术语都可翻译为数据分析，但它们是否全无分别呢?

data analysis一般泛指传统的数据分析方法。针对不同的数据，data analysis运用统计学相关的方法，如T检验、F检验、卡方检验、逻辑回归等，获取数据中的相关知识。相比于data analysis，data

① https://en.wikipedia.org/wiki/Data_science

analytics更加与时俱进。针对各式各样的数据，data analytics首先针对数据建立数学模型，然后运用各类方法，如数理统计类方法、机器学习、数据挖掘等，对数据进行深层次的知识发掘。根据之前提到的数据科学的定义（数据科学是一门利用不同方法从数据中获取知识的科学），它更倾向于运用data analytics为主、data analysis为辅的方式发掘数据中的知识。由以上两种不同的知识发掘方法，催生出两大类不同的数据科学工作者，即数据科学家（data scientist）和数据分析师（data analyst）。在下文中，笔者将着重分析和对比工业界对以上两类数据科学工作者的要求和大学教育培养数据科学工作者之间的差异。

3 数据科学人才的要求与培养现状

根据前文所述可知，能否培养出能够应对和处理不同类型数据的专业数据科学人才是能否应对大数据带来的巨大挑战的关键。作为向工业界输送人才的重要力量，大学教育理应成为数据科学人才培养的重要基地。本节将着重分析工业界对于数据工作者的要求和大学数据科学人才培养的概况，并总结二者的差异。

3.1 数据科学工作者——工业界的要求

虽然目前工业界雇佣数据科学工作者时会用各式各样的工作头衔，如数据工程师、数据分析师、数据分析员等，但根据前文的阐述，数据科学工作者可以大概分为两类：数据分析师与数据科学家。这两类数据科学工作者可以看作为实现不同层次的数据处理而设定的工作岗位。为了能够明确区分工业界对于二者的要求，笔者主要以中美两国两家在线招聘平台，拉勾网[②]和CareerBuilder[③]当日投放的招聘广告为样本，统计了中美两国对于数据分析师和数据科学家的岗位要求，并对统计结果进行对比分析。使用以上两个在线招聘平台的数据作对比分析主要基于下列原因：其一，拉勾网和CareerBuilder分别是中国内地和美国较大的IT类在线招聘平台，在这两个平台投放的招聘广告在一定程度上能够代表两国工业界对于数据分析师和数据科学家的岗位要求；其二，两个在线招聘平台均提供了详细的关键字搜索、分类搜索以及过滤功能，笔者能够借助这些功能实现不同角度的对比分析。

从2016年5月21日的广告投放数据，笔者统计了中美两国对于数据科学家和数据分析师的学历及工作经验要求，见**表1**。从**表1**可以看到，虽然具体的数据有所差异，但中美两国公司对于数据科学家以及数据分析师的要求基本相似。数据科学家需要更高的学历，例如：要求硕士以上学历的招聘广告在拉勾网上达到27.7%，在CareerBuilder上则接近42%。在其他调查报告中，对于数据科学家的学历要求也给出了相似的结果。例如，在一项由RJMetrics发起的调查中[④]，在过去4年成功获得数据科学家工作职位的数据科学工作者中，接近45%的数据科学家至少拥有硕士学历，而拥有博士学历的数据科学家也接近20%。相比于数据科学家，数据分析师更需要相对丰富的工作经验，例如CareerBuilder和拉勾网要求数据分析师有3~5年工作经验的招聘比例分别达到24.65%和56.6%，而要求同样工作经验的数据科学家的比例只有12.9%和26%。由**表1**的统计分析不难发现，数据科学家对于数据科学的理论要求更高，而数据分析师更倾向于强调数据处理的熟练程度。

② www.lagou.com
③ http://www.careerbuilder.com/
④ https://rjmetrics.com/press/new-study-finds-52-of-data-scientists-have-earned-that-title-within-the-past-4-years/

表1 数据分析师与数据科学家的学历及工作经验要求

要求		数据科学家		数据分析师	
		中国招聘广告数/个（总数：62个）	美国招聘广告数/个（总数：742个）	中国招聘广告数/个（总数：900个）	美国招聘广告数/个（总数：2 199个）
最低学历	本科	34（54.8%）	277（37.33%）	526（58.44%）	1 279（58.16%）
	硕士及以上	18（27.7%）	310（41.78%）	181（20.11%）	280（12.73%）
	无要求	10	155	193	640
工作经验	0~3年	44（70.97%）	148（19.95%）	154（17%）	475（21.6%）
	3~5年	8（12.9%）	193（26%）	510（56.67%）	542（24.65%）
	5~10年	10（16.13%）	85（11.46%）	236（26.22%）	240（10.91%）

（数据来源：2016年 5月 21日，拉勾网（中国）、CareerBuilder（美国））

除去比较学历和工作经验的要求，笔者还对数据分析师和数据科学家的工作职责要求进行了对比分析。根据拉勾网和CareerBuilder于2016年5月21日的招聘广告投放数据，**表2**列出了中美两国对于数据科学家和数据分析师工作职责要求的对比情况。

从**表2**可以看出，数据分析师被要求参与更多的应用性工作：如40%的招聘广告标明数据分析师需要撰写数据报告，而要求数据分析师利用工具软件进行数据挖掘的招聘广告超过了50%。相比之下，中美两国对于数据科学家的要求更强调数据科学理论，例如：美国至少80%的数据科学家工作要求应聘者具备建立数据模型的能力，而国内几乎所有的数据科学家岗位都要求应聘者具备数据建模的能力；此外，中美两国对于数据科学家的应聘者的算法设计能力、统计推理和数据挖掘理论以及决策支持方面的能力也有较高要求。而这些理论方面的岗位要求，在数据分析师的岗位要求中基本不会涉及。由**表2**的统计可以看出，就工作职责而言，数据科学家与数据分析师的区别也是显而易见，例如：数据科学家需掌握更全面的数据科学理论和应用知识，而数据分析师则更强调应用。由于工作性质、职责不尽相同，数据分析师与数据科学家的薪资待遇也不完全相同。笔者

表2 数据分析师与数据科学家的职责要求

工作职责	数据分析师		数据科学家	
	中国	美国	中国	美国
撰写数据报告	40%	40%	10%	20%
利用工具分析数据	50%	60%	0	20%
利用工具挖掘数据	40%	20%	0	0
数据库管理	10%	50%	0	0
实现分析和优化模型	30%	20%	30%	10%
建立数据模型	50%	40%	100%	80%
设计算法	30%	20%	80%	80%
统计推理分析和挖掘数据理论	0	10%	70%	60%
提供决策支持	20%	20%	50%	70%

（数据来源：2016年5月21日，拉勾网（中国）、CareerBuilder（美国））

通过调查CareerBuilder投放的数据科学家的岗位招聘广告发现，大多数招聘公司给出的年薪都在10万~20万美元，少数公司对于优秀的数据科学家可以给出更高的年薪。相比之下，数据分析师的年薪普遍低于10万美元，只有少数公司能够给经验丰富的数据分析师更高的劳动报酬。

根据**表1**和**表2**的统计，可以区分工业界对于数据分析师与数据科学家的要求。一名合格的数据分析师需要具备较强的实际应用能力，能够收集和管理数据，利用工具或软件分析数据，生成分析报告或撰写数据报告，能够实现不同的算法；而一名合格的数据科学家需要具备分析、研究、解决问题的能力，能够根据不同的数据建立数据模型，设计和实现数据分析、知识获取的算法，并且能够与商业或决策部门合作，利用从数据中获得的知识提供决策支持。只有具备以上相应能力的应聘者，才能成为符合工业界要求的数据科学人才。

3.2 供需失衡——数据科学人才的需求

在大数据的时代背景下，公司和企业都已认识到数据所能带来的巨大价值。但是数据科学人才的供应却明显不足。调查了近2 000家各类形式的商业团体的数据分析人才状况[15]，超过40%的受访公司承认自身缺乏具备深度数据分析能力的数据人才。据麦肯锡预测，到2018年，仅美国本土专业数据分析人才的缺口就将达到14万~19万人之多[14]。

同样通过对在线工作招聘网站数据的分析，可以在一定程度上了解工业界对数据科学人才的强烈需求。仅透过中国香港www.indeed.hk和美国CareerBuilder在线招聘网站的关键字data scientist和data analyst的查询，收集到2016年4—5月在以上两地投放的数据科学人才相关的招聘广告总计超过3 000条。为了解不同行业对于数据科学人才的需求，基于在线招聘平台2016年5月21日的广告投放数据，统计了不同行业投放数据科学相关职位的招聘广告的数量信息。

表3给出了拉勾网在2016年5月21日当日数据科学相关人才招聘广告中公司的分类统计及职位提供数。不难看出，和数据密切相关的产业，如移动互联网、金融、数据服务以及电子商务产业，都需要大量的数据科学人才。同时，从整体的人才需求而言，各行各业根据自身的特点，都有一定量的数据科学人才需求。可以说，在大数据的时代背景下，数据的巨大价值和利用专业数据人才管理数据、发掘知识的理念已经深入人心，不同的行业都希望结合自身特点，利用本行业特有的数据创造更大的商业价值。因为各行各业对于数据科学人才均有需求，这就对数据科学工作者的全面性提出了更高的要求：能够处理不同行业、不同类型的数据；能够利用不同方法发现数据中的知识和价值。

3.3 数据科学人才的培养

作为向各个产业培养和输送人才的

表 3 不同产业公司投放的数据人才招聘广告数量

产业分类	移动&互联网	电商	金融	工业服务	教育	娱乐文化	游戏	信息安全
广告投放量/个	>500	389	>500	166	76	106	85	33
产业分类	O2O	电脑硬件	社交媒体	旅游	健康医疗	生活服务	广告	数据服务
广告投放量/个	258	37	41	28	36	52	51	445

（数据来源：2016年5月21日，拉勾网）

表 4 QS 世界排名前 50 大学中开设数据科学相关硕士培养计划的学校统计

大学	国家
哈佛大学	美国
斯坦福大学	美国
伦敦大学学院	英国
芝加哥大学	美国
新加坡国立大学	新加坡
约翰霍普金斯大学	美国
康奈尔大学	美国
爱丁堡大学	英国
哥伦比亚大学	美国
加州大学伯克利分校	美国
密歇根大学	美国
美国西北大学	美国
曼彻斯特大学	英国
布里斯托大学	英国
加州大学圣地亚哥分校	美国
华威大学	英国
伦敦帝国学院	英国

基地，大学理应承担起培养合格的数据科学人才的责任。为应对数据科学人才需求的挑战，国内外的大学均在一定程度上改变各自的教学计划或内容，希望能够培养更多的数据科学人才。为调查世界范围内优秀大学的数据科学人才培养情况，依据QS2015全球大学的排名情况，着重了解了QS排名前50的大学数据相关的教学培养计划。在本次调查和统计中，重点关注每所大学的全日制硕士教育，调查教学计划中是否开设数据科学相关的专业。此次调查和统计过程中并未考虑本科教育的原因是：其一，绝大多数学校在本科教育中并未将数据科学作为一门独立的教学学科，而仅开设一定量的数据科学相关的课程，如数据挖掘、算法设计等；其二，相比于本科教育，硕士培养的方向更加精细化，这也为培养专业的数据科学人才提供了前提条件；其三，根据前文的叙述，尽管工业界对于数据人才的要求不尽相同，但是硕士水平的人才的比例仍然占很大一部分。基于以上3点原因，着重考察优秀大学在硕士培养计划中是否考虑到数据科学人才的培养，这能够在一定程度上揭示当前大学教育对于数据科学人才培养的重视程度。

首先，将QS世界排名前50的大学中设有数据科学相关的硕士培养计划的大学做了整理，见**表4**。根据统计，在2015—2016年度，QS世界排名前50的大学中，仅有17所大学开设数据科学相关的硕士培养计划。也就是说，超过60%的大学在硕士阶段没有数据科学相关专业。作为替代，这些未开设数据科学相关专业的大学设有一定量的关于数据科学的课程供硕士学生选择。这个现状和目前学术界与工业界的“大数据热”形成了鲜明的对比。

通过观察这17所大学所在的国家，发现这17所学校仅仅来自3个国家，分别为美国10所、英国6所、新加坡1所。通过大学所在地的分布，可以看出，作为大学教育整体领先的欧美地区，对数据科学专业的重视程度也相对较高。因此英美两国的优秀大学中，均有一定比例的大学开设了数据科学相关的专业。同时，这也契合了前文中所叙述的问题，英美两国的大学对工业界大量的数据人才需求做出了及时的应对，比如开设专门的硕士培养计划，向社会输送专业的数据科学人才。

其次，详细调查了各个大学数据科学相关专业的硕士培养计划，包括培养计划的名称、开设的院系和培养计划中着重处理的数据类型。通过此项调查，可以了解到不同大学对于数据科学人才培养的侧重点。**表5**给出了该项调查的详细结果。在开设数据科学相关专业的17所大学中，硕士培养计划的名称、开设院系以及针对的数据类型不尽相同。8所大学的计算机院系开设了数据科学硕士培养计划（伦敦大

学学院、芝加哥大学、加州大学伯克利分校、曼彻斯特大学、布里斯托大学、加州大学圣地亚哥分校、华威大学以及伦敦帝国学院）。除芝加哥大学外，另外7个开设在计算机院系的硕士培养计划并不强调处理特定的数据处理类型。这一特点同时也呈现在由统计类、信息类以及数据科学类院系所开设的硕士培养计划中。而由商业、运筹学以及公共健康类的院系开设的数据科学硕士培养计划，则倾向于应对特定的数据类型，诸如公共健康数据、金融及商业数据。从这些统计数据可以推断，由计算机、统计、信息类院系开设的数据科学硕士培养计划将培养教学中更大的比重放在如何将数据科学理论应用到不同数据的处理和发掘方面，而商业类院系开设的硕士培养计划更倾向于利用数据对科学理论处理和发掘商业以及金融数据。

虽然各个大学硕士培养计划的名称、开设院系以及针对的数据类型不尽相同，但作为数据科学相关的硕士培养计划，课程的设置应该或多或少具有一定的相似性。为验证以上推断，调查了17所开设数据科学相关硕士培养计划的大学的详细的课程设置情况，并进行了横向的对比分析。通过该对比分析，可以在一定程度上了解到目前大学教育对于合格的数据科学人才的一般要求。**图1**列举了8个多所大学开设的热门课程。从**图1**中可以看出大学教育对于数据科学人才培养的几点考虑，具体如下。

- 是否精通统计学相关的知识在很大程度上决定了一个数据科学工作者是否合格。众所周知，统计学、统计推理等学科在数据挖掘过程中扮演着重要的角色，诸多知识发掘方法都源于统计学中的模型。
- 坚实的数据分析方面的知识也是数据科学人才培养的重要组成部分，从**图1**中看到，8个大学硕士培养计划中开设了数据

表 5　各大学开设数据科学的院系以及针对的数据类型

大学名称	硕士培养计划	开设院系	数据类型
哈佛大学	计算生物和数量遗传学	公共卫生学院	卫生信息学
斯坦福大学	统计学：数据科学	统计系	所有
伦敦大学学院	数据科学硕士	计算机科学系	所有
芝加哥大学	计算分析和公共政策硕士	计算机系	公共政策
新加坡国立大学	商业分析	商学院	金融和商业
约翰霍普金斯大学	信息系统	商学院	金融和商业
康奈尔大学	数据分析	运筹学与信息工程学院	金融和商业
爱丁堡大学	数据科学	信息学院	所有
哥伦比亚大学	数据科学	数据科学研究所	所有
加州大学伯克利分校	数据科学和系统	电子工程和计算机科学系	所有
密歇根大学	商业分析	商学院	金融和商业
美国西北大学	分析学	工程学院	所有
曼彻斯特大学	数据与知识管理	计算机系	所有
布里斯托大学	高级计算	计算机系	所有
加州大学圣地亚哥分校	数据科学与工程硕士	计算机科学与工程系	所有
华威大学	数据分析	计算机科学系	所有
伦敦帝国学院	计算学：机器学习	计算机系	所有

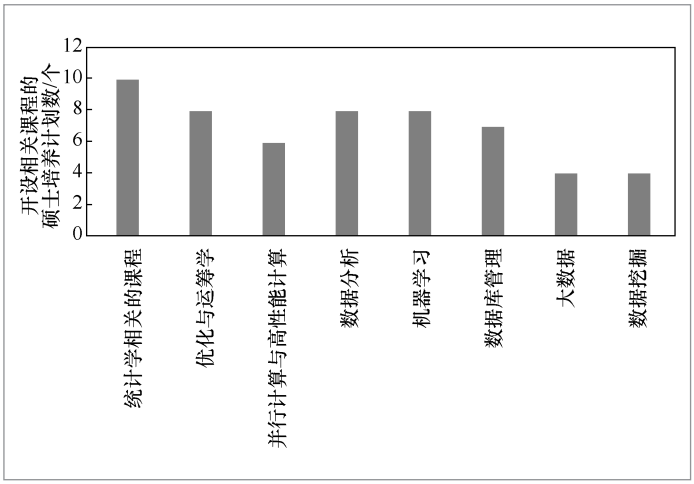

图 1 数据科学相关硕士培养计划下相似的课程设置

分析类的课程。

● 并行和高性能计算也是合格的数据科学人才必备的技能之一。伴随着大数据时代的到来，可收集的数据总量与日俱增，传统方法的效率不足以应对庞大的数据总量。因此，传统方法的并行以及高性能计算的应用能够大大加速数据处理和知识发掘的过程。

● 除统计学相关的课程外，其他知识发掘的方法，如机器学习、数据挖掘也是数据科学人才培养的重点。

从以上4个特点不难看出，目前大学培养数据科学人才基本按照数据科学的定义和范畴进行，但更着重培养学生掌握在一定数据类型中发掘知识的原理和方法，同时要求学生掌握数据存储、处理方面的理论。

表 6 中国内地与中国香港及美国较优秀的 10 所大学

美国	中国内地与中国香港
麻省理工学院	香港大学
哈佛大学	香港科技大学
斯坦福大学	香港中文大学
加州理工学院	香港城市大学
芝加哥大学	香港理工大学
普林斯顿大学	北京大学
耶鲁大学	清华大学
约翰霍普金斯大学	复旦大学
康奈尔大学	中国科技大学
宾夕法尼亚大学	上海交通大学

3.4 数据科学人才的培养——中美之对比

为了解中美两国大学培养数据科学人才的概况，还调查了中国内地、中国香港和美国较优秀的10所大学的数据科学相关硕士培养计划，见**表6**。经统计，在美国排名前10的大学中，共有5所大学设有专门的数据科学相关的硕士培养计划，分别为哈佛大学、斯坦福大学、芝加哥大学、约翰·霍普金斯大学和康奈尔大学。而中国内地和中国香港，仅有香港中文大学开设了全日制数据科学相关的硕士培养计划。

不过，以上略显悬殊的对比并不能说明中国的优秀大学不足够重视数据科学人才的培养，而是因为中国内地的硕士教育制度与美国和中国香港有一定的差异。在美国和中国香港地区，大学中设有专门的授课式硕士培养计划，而中国内地的大学多数采用授课和研究混合式的硕士培养计划。基于这个原因，中国内地很多大学并未直接给出明确的数据科学硕士培养计划，但是很多大学都设有专门的数据科学相关的研究院，通过这些研究院培养数据科学方面的人才。**表7**列出了3所大学开设的数据科学相关的研究院或研究小组，分别是清华大学的数据科学研究院、复旦大学的数据科学重点实验室和南京大学的机器学习与数据挖掘研究所。在这些研究院中，均设有数据科学相关的研究方向。同时，还可以通过其他几个实例来了解中国优秀大学对于数据科学人才的培养：如清华大学于2014年开设大数据硕士培养计划并于2014年9月开始招生；复旦大学也在2015年正式开设数据科学专业研究生培养计划[7]。因以上列举的培养计划主要针对专业硕士（非全日制），所

以它们并未计入第3.3节中的统计和对比。不过这也足够说明，数据科学人才的培养在我国同样得到了相当程度的重视。

3.5 人才培养与市场需求的对比

本节对大学数据科学人才的培养与市场需求进行了对比。二者之间的具体差异已经在**表**8中做了总结。从**表**8可以看出，当前大学教育对于数据科学人才的培养目标与市场的要求存在一定差异。

首先，很多大学仅开设一定量的数据科学相关的课程，而工业界更需要能够全面系统掌握数据科学理论的人才；其次，很多大学侧重数据科学相关的理论，如数学或统计建模、算法设计等，而工业界更倾向于聘用可以将数据科学理论应用到特定行业（如金融、商业、公共信息等）的人才；第三，由于学校教育的时间限制，学生对于不同工具、软件的掌握不可能做到面面俱到，而不同行业、不同的公司，要求使用的数据处理工具往往不尽相同；最后，就是巨大的人才缺口，如前文所述，在大数据的时代背景下，数据科学人才的需求总量远大于大学培养的人才总量。以上4点是目前数据科学人才培养与市场需求之间存在的突出矛盾。

4 数据科学人才培养的改进

根据数据科学人才培养与市场需求之间存在的突出矛盾，笔者给出以下关于数据科学人才培养的建议。

首先，各个大学应大力支持数据科学这一新兴基础科学的研究，发展和完善数据科学理论体系，为数据科学人才培养提供必要的理论和知识基础。第二，鉴于大学教育在本科教育更重视基础能力的培养，我国的优秀大学可以借鉴世界范围内优秀高等学府的经验，在硕士教育阶段开设专门的数据科学硕士培养计划，在本科教育阶段适当开设数据科学相关的基础课程，以培养不同层次的数据科学人才。依据目前大学培养数据科学人才的概况和工业界对于数据科学人才的需求，给出如下数据科学硕士培养计划以供参考。该培养计划根据数据科学的定义，将数据科学硕士培养分为4部分：相关基础学科学习、知识发掘方法的学习、数据科学理论在大数据背景下的应用以及数据科学在不同行业中的应用。前3个阶段可以看作数据科学理论体系的培养，最后一个阶段强调实际应用。接下来笔者将对这4个阶段分别进行详细的介绍。

表 7 中国内地优秀大学数据科学相关研究院举例

大学	数据科学相关研究机构
清华大学	数据科学研究院
复旦大学	上海市数据科学重点实验室
南京大学	机器学习与数据挖掘研究所

表 8 数据科学人才的培养与市场需求的差异

数据科学人才的培养	数据科学人才的市场要求
仅开设数据科学相关的课程	专门的数据科学方面的人才
数学或统计学建模，分析解决问题的能力	数学、统计学与特定学科（金融、商业、公共信息等）的综合运用
学习有限的数据处理工具	通过不同工具，理论完成数据分析，知识探索
少数学校开设专门的数据科学人才培养计划	大量的专业数据人才

（1）基础学科的学习

基础学科是数据科学人才培养的前提。在硕士培养的初始阶段，学校应该开设基础科目以夯实学生的理论基础。依据数据科学的发展现状，数据科学的基础学科至少应包括高等微积分、数理统计、矩阵论等数学方面的学科。开设这一类课程的原因有二：一是数据科学与数学类的学科

联系紧密，众多的知识发掘方法都需要学生以数学为基础去理解和学习；二是选择学习数据科学专业的学生可能有着不同的本科教育背景，学生可以根据自身情况酌情选择所修的课程。例如，出自数学、统计学相关专业的学生可能在以上提到的几个科目比较擅长，因此他们可以选择少修或者跳过种类基础学科的学习。而来自数学基础相对薄弱的专业的学生，在进修数据科学专业的初始阶段，应着重学习数学方面的基础理论，为今后的课程打好基础。

（2）系统地学习知识发掘的方法

知识的获取是整个数据处理过程中的关键，是处理数据的重要目的。在学生掌握相关基础学科理论的前提下，学校可以开设不同的课程，让学生系统地学习知识获取的方法。依据前文中提到的开设数据科学学科的大学的培养计划，笔者认为这一部分的课程至少应包括以下科目：统计推理、机器学习、数据挖掘和数据分析。通过学习统计推理、机器学习相关的课程，学生可以掌握一系列知识获取的概率模型，如贝叶斯模型、线性回归模型、逻辑回归模型等。通过学习数据挖掘、数据分析相关的课程，学生能够进一步将基础理论和实际的知识获取方法（算法）联系起来，如利用统计推理中的残差分析在数据中发现知识的算法[8,9]以及一些经典的数据挖掘算法，如决策树、*k*-means、*k*-NN等。知识获取方法这一部分是数据科学人才培养的关键，各个大学可以根据自身实际情况，尽量开设全面系统的课程，让学生从多个不同的角度深刻全面地理解数据科学理论中知识获取的方法。在培养学生掌握知识获取方法的同时，各个大学也应开设一定量的学习计算机程序语言的科目，以提高来自不同专业背景的学生掌握流行的计算机程序语言，如Java、R、C++、C#等。

（3）掌握高效的数据处理方法

在大数据的时代背景下，数据科学工作者面对的数据的容量、复杂度都今非昔比。海量数据带来的最直接挑战就是传统的方法难以处理如此巨大的数据集。因此，现代的数据处理方法在获取知识的过程中起到至关重要的作用。学校应在硕士阶段开设专门的课程以培养学生利用数据科学理论处理大数据的能力。根据世界范围内优秀大学的教学经验，我国的大学可以酌情开设针对大数据的高性能计算、并行计算、分布式计算、数据仓库、数据库管理等课程以及Spark、Hadoop等大数据处理平台的课程。通过学习这些课程，学生可以掌握如何高效地处理大数据，在大数据中获取有价值的知识，进而成为具备大数据处理能力的数据科学人才。

（4）数据科学在不同领域中的应用

在夯实数据科学理论的基础上，学校也应重视培养学生在不同类型的数据中获取知识的能力。毕竟行业、领域不同，数据不尽相同。为达到以上目的，学校可以尝试与工业界合作，以实习的方式让学生在工作中接触不同类型的数据，利用所学的知识尝试做数据科学方面的工作。当然，学校也可根据当前的市场需求，利用已经开放的数据资源，开设数据科学在热门行业、领域中的应用课程。通过这一类应用性课程的学习，学生能够根据自身的兴趣和未来的就业取向在数据科学理论的应用上有的放矢。因为数据量日益庞大，近来生物学方面的研究愈发依赖计算机技术，因此，学校可酌情开设计算生物学相关的课程，让感兴趣的学生学习。通过学习这些课程，学生可以了解和掌握数据科学在生物学领域中的应用，如在基因表达数据中的聚类分析[10]、在PPI网络中发现蛋白质化合物的方法[11]等。再比如，在搜索优化、定向推荐以及定向广告投放等时下流行的技术

中，一部分知识的获取是基于文档分类以及特征抽取的方法完成的。为培养有志在这个方面发展的学生，学校可以开设数据科学在自然语言处理方面的应用课程。通过学习这一类课程，学生可以掌握一系列自然语言处理和特征抽取的基本模型，并进一步研究复杂的、可并行的模型使特征抽取效率更高，准确率更高[12,13]。数据科学在不同行业、不同领域中的应用实例还有很多，在此笔者不一一列举。总之，通过接触和学习如何处理不同来源的数据，学生的实际应用能力可以得到大大加强。

数据科学目前还处于起步和发展的阶段，理论体系还需要完善。在将来的一段时间内，数据科学的理论、相关的知识获取方法以及应用还会进一步丰富，大学教育也应该根据不同时期数据科学的发展情况，调整培养计划，适应市场需求。

5 结束语

在大数据的时代背景下，各行各业均意识到了数据所能带来的巨大价值，因此纷纷向数据科学人才抛出橄榄枝，希望能借数据科学工作者的手，发掘数据中的潜在价值。在本文中，首先探讨了能够应对大数据处理的学科——数据科学；依据收集到的实例，考察了工业界对于不同类型的数据科学人才（数据科学家和数据分析师）的需求和岗位要求的异同；根据当前国内外优秀大学开设的数据科学相关的学科培养计划，总结了国内外优秀大学在硕士学历水平上培养数据科学人才的概况；根据大学教育培养数据科学人才的概况与工业界对于数据科学人才的具体需求，总结出大学教育培养的数据科学人才与工业界实际需求之间存在的突出矛盾；最后，根据供需之间的矛盾给出关于大学培养数据科学人才的4点改进建议，即重视基础学科的学习，系统地掌握知识发掘方法，掌握高效的数据处理方法以及精通数据科学在不同领域中的应用。人类也已进入大数据时代，能否培养出合格的数据专业人才关系到能否掌握数据的核心价值。作为为社会各界输送人才的基地，大学教育对于数据科学人才的培养至关重要。

参考文献:

[1] ZHU Y Y, XIONG Y. Protection and utilization of data resources[M]. Shanghai: Shanghai Scientific and Technical Publishers, 2008: 133-137.

[2] HOWE D, COSTANZO M, FEY P, et al. The future of biocuration[J]. Nature, 2008, 455(7209): 47-50.

[3] MARX V. The big challenges of big data[J]. Nature, 2013, 498(7453): 255-260.

[4] China Academyof Informationand Communications Technology. Survey on the development of big data in China[R]. 2015.

[5] FAN L, HONG X, CHAO H, et al. Challenge and countermeasure of governing government big data[J]. Big Data Research, 2016, 2(3): 27-38.

[6] PAN Z, CHENG X, YUAN X. Developing trend forecasting of big data in 2016 from CCF TFBD: interpretation and proposals[J]. Big Data Research, 2016, 2(1): 105-113.

[7] ZHU Y Y, XIONG Y. Training data scientists in the era of big data[J]. Big Data Research, 2016, 2(3): 106-112.

[8] CHAN K C C,WONG A K C, CHIU D K Y. Learning sequential patterns for probabilistic inductive prediction[J]. IEEE Transactions on Systems Man and Cybernetics, 1994, 24(10): 1532-1547.

[9] CHING J Y, WONG A K C, CHAN K C C. Class-dependent discretization for inductive learning from continuous and mixed-mode data[J]. IEEE Transactions

on Pattern Analysis and Machine Intelligence, 1995, 17(7): 641-651.

[10] AU W H, CHAN K C, WONG A K, et al. Attribute clustering for grouping, selection, and classification of gene expression data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, 2(2): 83-101.

[11] HU A L, CHAN K C C. Utilizing both topological and attribute information for protein complex identification in ppi networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 10(3): 780-792.

[12] LIU Z, ZHANG Y, CHANG E Y, et al. Plda+: parallel latent dirichlet allocation with data placement and pipeline processing[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 389-396.

[13] LIU Z, HUANG W, ZHENG Y, et al. Automatic keyphrase extraction via topic decomposition[C]//Conference on Empirical Methods in Natural Language Processing, October 9-11, 2010, Massachusetts, USA. [S.l.:s.n.], 2010: 366-376.

[14] MANYIKA J, CHUI M, BROWN B, et al. Big data: the next frontier for innovation, competition, and productivity[J]. McKinsey Global Institute, 2011.

[15] RANSBOTHAM S, KIRON D, PRENTICE P K. The Talent Dividend[J]. MIT Sloan Management Review, 2015, 56(4): 1.

作者简介

陈振冲(1959-),男,博士,香港理工大学学务长,电子计算学系教授。分别于1984年、1985年和1989年在加拿大滑铁卢大学计算机科学与统计学系获学士、系统设计工程方向硕士及博士学位,毕业后供职于IBM加拿大实验室,并于1994年加入香港理工大学电子计算学系担任教职工作至今。目前主要研究方向为大数据分析、生物信息学、计算生物学、数据挖掘、机器学习、模糊逻辑系统、遗传算法、人工智能以及软件工程。

贺田田(1985-),男,香港理工大学电子计算学系博士生,主要研究方向为数据挖掘、图聚类分析、生物信息学和遗传算法。

收稿日期: 2016-07-30