连续结构.其主要的应用对象是微分方程等.计算机科学处理的主要是离散结构，如网络.而现实数据的特点介于两者之间：数据本身是离散的，而往往数据的背后有一个连续的模型，所以要发展针对数据的算法，就必须把计算数学和计算机科学研究的算法有效地结合起来

1.2对学科发展的影响

数据科学对学科发展提供了前所未有的机遇和挑战.要充分利用好这个机会,我们就必须建立起一套新的学科和教育体系.在高校的层面,要赋予数据科学应有的地位，建立起跨学科全方位的数据科学研究平台.同时要进一步完善和企业合作创新的机制，培养适应学术界和企业界需求的数据科学人才

数据科学也将对许多传统学科的发展带来极大的影响.首先是数学.数学的发展主要来自两个方面的推动力：一是来自数学内部,学科自身的完善带来的推动;二是来自外部,由其他学科、社会或工业发展的需要而带来的推动.就目前的现状而言，第一方面的推动力对数学的影响要远远超过第二方面的推动力.这样造成的结果是,一方面,数学作为一门学科,其重要性已经得到广泛的认可.而另一方面,数学家作为一个群体,其对社会和科学整体发展的影响是比较间接的在很多学校以及在整个科学界，数学家这个群体正显得越来越孤立，这是一件极为不幸的事情.它不仅大大影响了数学的发展,更是影响其他学科、技术乃至社会的发展.事实上,至少在理论研究方面，很多学科的瓶颈问题都是数学问题,这在近一百年前狄拉克就已经指出来了.所以在很多学科里，我们看见的都是非数学出身的科学家在进行数学方面的研究.对数学本身来说,这可能不是坏事情.但对数学家这个群体来说却是失去了他们本来应该发挥作用的空间

自然科学领域的基本出发点都是来自于物理学的基本原理.对数学家们来说,这是云个基本障碍,数据科学不一样它的基杰原理本身就来自于

效学.所以数据科学在数学和实际应用之间建立起了一个直接的桥梁.而这些实际应用正是智能化时代整个社会最为活跃的一部分.数学是这个学科的理论基础、尽管数学家们还没有充分发力,但对数学和数学家们来说这实在是一个千载难逢的机会

不仅如此，数据分析几乎涉及了现代数学的所有分支.甚至于像表示论这样的极其抽象的分支,在数据科学领域也有其发挥作用的余地，所以数据科学对数学的要求和推动是全面的，而不是仅仅局限在几个领域.数据应该成为数、图形和方程之外数学研究的基本对象之一

数据科学对计算机科学的发展也会带来很大的影响.图灵奖得主 JohrHopcroft曾经指出,在过去的几十年里,计算机科学的研究对象主要是计算机本身,包括硬件和软件.以后计算机科学的发展将主要围绕应用展开、而从计算机科学自身来看，这些应用领域提供的主要研究对象就是数据.虽然计算机科学一贯重视数据的研究，但数据在其中的地位将会得到更进步的加强

再看统计学.统计学一直就是一门研究数据的学科.所以它也是数据科学最核心的部分之一但在数据科学的框架之下,统计学的发展也会受到很大的冲击.这种冲击至少表现在两个方面、一是关于数据的模型将会跳出传统的统计模型的框架，更一般的数学概念，如拓扑、几何和随机场的概念将会在数据分析中扮演重要的角色.二是算法和分布式计算将成为研究的中心课题之一

1.2.1对传统学科的冲击

这里我们举两个例子，一个是社会学的例子，作为社会科学的一个分支,社会学一直是一门基于数据的学科.大到国家和社会层面的数据,小到家庭和个人的数据,这些是社会学研究的基本资料，从这个角度来看,社会学和数据之间的关系不是什么新的现象.但即便如此,数据科学的兴起仍然对社会学的研究有着巨大的影响.这至少表现在如下几个方面

是社交网络的产生和网络科学的研究为社会学带来了一个新的研

仅供个人科研教学使用!第一章绪论

究层面,即介观层面.这不仅给社会学提供了新的研究方向，而且也给社会学的研究提供了新的实用价值，如信息传播、广告投放、热点分析等

二是新的数据来源和数据分析的方法将使社会学的研究进一步量化、去经验化

三是更多更加严密和系统的科学方法被引进到社会学的研究中，如数据采集的方法、北京大学中国社会调查中心开展的家庭访问调查就是一个很好的例子、他们不但注重调查中问答的结果，同时也记录了调查过程的数据，这样严密的科学方法一定会给社会学的研究带来极大的影响

以往,在人们眼里,社会学往往不是一门技术型的或实用型的学科.但随着社会学的进一步的量化，人们对社会学的看法将会发生很大的变化在不远的未来,社会学的研究将对产品营销、信息传播和舆情预警等实用领域产生深远的影响

我们要谈的第二个例子是语言学.跟社会学一样，语言学在历史上也是一个离实用技术比较远的学科.但近年来蓬勃发展起来的机器翻译、自然语言处理、语音识别和文本分析等技术给语言学的实际应用提供了一个绝好的机会，但值得注意的是，在所有这些领域,基于概率模型的处理方法的有效性远远超过了基于文法的处理方法的有效性.这对传统的语言学来说,不能不说是一个非常令人失望的结果

在麻省理工学院成立 150周年的一个纪念会上,当代语言学的奠基人乔姆斯基教授针对这一问题提出了他的看法、他认为概率模型的成功是有限的，而且其成功只是仅仅局限于逼近未被分析的数据这一方面，他的言下之意是说概率模型只是技术上的成功,不能算作是传统科学意义上的成功、因为它没有给传统的语言学问题如文法问题，带来新的认识.即便这个说法是正确的，它也只是站在了语言学的角度，而不是语言的基本社会功能的角度.而后者对社会的影响显然要高过前者、不过，需要强调的是，概率模型目前的成功并不代表语言学的研究都没有用处、只是说，语言学也必须正视概率模型的作用

1.2.2新学科的诞生：计算广告学

广告有着十分悠久的历史，但它一直都很难算得上是一门科学、尤其是在中国，最典型的广告，就是出现在媒体上，特别是在电视上的各种营销语言.近年来,由于百度等搜索引擎选择商业广告作为其主要赢利模式,一门新的学科，计算广告学，由此而诞生

计算广告学所处理的主要问题是怎样有针对性地投放广告、互联网上的广告有两个最基本的指标：点击率和转换率.点击率是广告被点击的概率.转换率是广告被点击以后引起商品成交的概率，由于后者更难估计,所以互联网上的广告往往以点击率作为主要指标.这就要求我们根据用户提供的信息、比方说其输入的关键词，预测不同广告的点击率.这是计算广告学的一个基本问题，解决这个问题的主要想法就是构造一个效用函数来估计用户对不同广告感兴趣的程度

目前像斯坦福大学和加州大学伯克利分校等知名学校都已开设了计算广告学这门课.美国国家基金委所属的几个数学研究所之一，地处北卡罗来纳州的统计与应用数学研究所也针对计算广告学举办了专题研讨会

1.3对科学研究的影响

大数据给科学和教育事业的发展提供了前所未有的机会,同时也提出了前所未有的挑战、它将对现有的科研和教学体制带来大幅度的变革，对科学与产业之间的关系、科学与社会之间的关系产生极大影响、总结一下大数据的影响将主要来自以下几个方面

首先,数据科学将成为科研体系中的重要部分,并逐渐取得与包括物理、化学、生命科学等学科在内的自然科学分庭抗礼的地位，未来的科研和教育体制应该由两条主线组成：一条是以基本原理为主线，现在的物理学、化学、机械工程等学科，以及生命科学、材料科学、天体物理、地球科

仅供个人科研教学使用!第一章绪论