第 0 章 Chapter 0

发现、出发

最近一年里，知乎社区有不少朋友邀请我回答关于数据挖掘的问题，其中提问最多的是关于“如何改行做数据挖掘”。我想他们之所以邀请我回答这类问题，不是因为我做数据挖掘做得好，而是好奇我是如何改行做数据挖掘的？说来也巧，我本科是学电子的，研究生是学控制的，而我的职业理想是成为一个“先知”，但我并不知道如何才能实现这一职业理想。自公元 632 年人类最后一位先知默罕默德去世之后，将近 1400 年没人做先知了，既没有人可以指导我，也没有可以效仿的对象。2011 年到 2013 年发生了一系列事件，包括 IBM 的沃森在“危险边缘”节目中击败了人类选手、Google Brain 某些成果的展示、美国统计学家 Nate Silver 对于总统大选的预测等，这些事件都有一个共同点，那就是让“数据科学”从学术研究蜕变为实际的应用。这也让我意识到也许我可以做得更好——通过“数据科学”建造一个“先知”，虽然直到现在我还没有实现这个目标，不过我愿意把这一路积累的经验拿出来与大家分享，希望这些东西能够帮助各位读者实现自己的目标，或者找到自己的目标。现在，就让我们出发吧！

0.1 何谓数据科学

在家用计算机普及之前，数学、逻辑学、哲学及自然科学研究的目的都是为了追求完美的理论证明，或者是提供某种确定性的规则，用以解释某种自然现象，或者为某些技术提供理论依据。那个时候人类产生数据的能力和收集数据的能力还很有限，或许公司的经营账目和计算导弹发射弹道的演算纸就属于数据最集中的地方了。在那个年代，这些数

据分析和处理的工作大都是由人工完成的，最多也只会借助某些由机械或电子构成的计算装置罢了。在互联网兴起之后，人类将现实世界中的很多信息以数据的形式存储到网络空间中，比如生活中发生的一段故事，或者旅行中家人的照片，这些数据记录了人类的行为和社会的发展，甚至包括了自然环境的变化。当今，大量的、各种各样的数据快速产生，并存储在互联网中，而这些数据自然而然地构成了一个人造的环境，称为数据界（data nature）。通过对数据界中数据的研究，我们不仅可以了解数据本身的种类、状态、属性及变化形式和规律，还能从中洞悉人类的某些行为，了解人类的某些社会属性。并且这些研究方法还能扩展到其他依赖数据的学科中，比如气象科学、地震科学、金融学、基因科学，等等。在可以预见的未来，我相信，不仅在互联网行业中会有数据科学家的身影，在各行各业中，只要与计算机打交道，我们就不得不为已经产生和将要产生的数据做好充分的准备。所以，我认为在这个数字化的时代，不同的专业领域，都需要从大量的数据中寻找到一系列的理论和实践，这就是数据科学。

0.1.1 海量的数据与科学的方法

“如何才能成功？”无数成功学方面的书本和布道者都没法给出一个方程或流程图来向所有人解释这一过程。最多只能根据统计学（或者是臆想）列举出一些可能的必要条件，比如努力、机遇、贵人或仅仅只是运气好。我们能否对人类的行为做一个精确的建模？太难了，比如，不同的人对于成功的定义不同，有的人认为挣钱是成功，有的人认为出名是成功。再比如就算大家都认为成为企业家可以算作某种意义上的成功，但是企业的种类又各有不同，有的人是在电商领域成功的，有的人是在金融行业成功的，他们的成功经历也各不相同。

事实上，关于“成功”的变量我可以列举无数个，但即使穷尽了所有可能的变量，也还会遇到数据缺失的问题——一个人成功之前的数据又该如何准确地记录？这个世界有 60 亿人，如果每个人出生时就携带一个电子记录仪，那么就可以记录这个人生活中发生的所有事情。这有可能么？可能，不仅是可能的，而且我们现在就在做类似的事情，智能手机正源源不断地收集人类的数据并且存储到网络中，我们购物的数据、兴趣的数据、人口统计学的数据等都将用作描述我们每一个人的“数字化身”，这是存在于网络中的我们。并且随着智能硬件、物联网、工业 4.0 的推进，整个现实生活中的人类社会在网络中都会有一份“副本”。为了处理这些数据，并且从中找到对我们有价值的结果，需要更先进的技术与方法，其中将会涉及数据的收集、转换、存储、可视化、分析与解释等内容，这将会是一项非常有价值的课题。

0.1.2　数据科学并不是新概念

在过去的几年中，大数据、人工智能、数据挖掘等词汇被媒体炒得热火朝天，一方面我乐于见到我所从事的工作受到人们的关注，另一方面我也发现越来越多的人开始疑惑。就像本书开篇中所提到的那样，我每天都会收到来自不同工作领域的人（有时候是记者或化工专业的从业者，有时候是程序员或数学系的学生，有时是一些在实际工作中遇到困难的工程师）的提问，有的是希望能澄清一些概念，有的是问如何入门，有的是希望我针对他遇到的麻烦提一些建议。我很乐意帮助他们，顺便抱怨一下某些不负责任的媒体，是它们把大数据吹得天花乱坠，把各种神秘的力量都赋予数据科学，好像数据科学家就是新时代的先知一样，能够预测未来，改变人类的命运。而且媒体给公众传递的信息是这样的：大数据是上个月才出现的，Google 在上周才提出了深度学习方法，一举解决了人工智能难题。我担心在这样冒进的社会氛围下，这些被扭曲的报道掩盖了事实的真相，那些对这个领域感兴趣的人会被吓跑，这颗科学史上的新星会陨落（在我收到过的提问里，甚至有人问：大数据的浪潮是不是过去了，现在学还来得及么？）。如果要追溯数据科学的起源，可以从 1974 年在美国和瑞典同时出版的《计算机方法的简明调查》一书中看到，作者彼得·诺尔对数据科学下过这样的定义“数据科学是处理数据的科学，一旦数据与其所代表的事物的关系被建立起来，就能为其他领域与科学提供借鉴”。

在“大数据”出现以前，统计学家觉得他们所做的就是数据科学，他们会通过分析一些数据来为公司或政府提供一些决策上的帮助。比如，大型上市公司的财报，或者每一次美国大选之前所做的民意调查就属于此类范畴。当然，不能认为互联网时代的数据科学是新瓶装旧酒，经历了这么多年的沉淀和积累，加上广泛的需求，数据科学发展出了一套与之相适应的理论和方法。我也希望能帮助更多的人了解数据科学，促进数据科学的发展。

0.1.3　数据科学是一个系统工程

现代工业界喜欢谈生态和闭环，其实数据科学也要贯穿数据的整个生命周期。下面将数据的生命周期简单地划分为如下几个阶段。

- ❑ 数据采集
- ❑ 数据清洗
- ❑ 数据处理
- ❑ 数据查询与可视化

数据采集传统的手段主要来自于经营数据和网络爬虫采集的数据。现在还包含一些“数据化”的过程，2013 年一篇题为“ The Rise of Big Data”（大数据的崛起）的文章中提

到了“数据化”的概念，即数据化是一种流程，可以将生活中的方方面面转化为数据。各种手机上的传感器，智能穿戴等设备采集数据的过程都属于数据化。

数据清洗主要负责处理数据中的噪声或缺失数据。由于填写表单时的疏忽，或者是爬虫程序的故障，再或者是传感器失灵等原因，总是会产生一些我们意料之外的数据，这些数据可能不符合某些格式的要求，或者会缺失部分数据，需要通过数据清洗来剔除或修正这些数据。如果数据量巨大，这就需要我们有处理海量数据的能力。

数据处理可以使用统计学的方法或机器学习的方法从数据中发现我们想要的价值，通常所说的数据挖掘就是在这一步中进行的。之所以这里没有使用“数据挖掘”这个词，是因为有些时候，在某些项目中仅仅使用简单的统计方法就可以得出很有价值的结论，并没有使用数据挖掘的专门技法。而且，与普通人的直觉相反，数据挖掘结果的价值往往是通过与业务的紧密结合才能体现出来的，胡乱套用算法往往得不出任何有价值的东西。比如，通过历史房产中介的销售数据（包括房屋的价格、面积、层数、每层住户数等信息）来为新的楼盘定价、预测目标客户群体就是两个不同任务，前者通常只需要简单统计（实际上我们过去一直就在这么做）即可，而后者可能就要使用分类预测算法了。

数据查询与**数据可视化**这两项是为了将处理过后的数据呈现给需要的人。有的时候是需要索引巨量的数据，比如搜索引擎。有的时候是规律性的结果需要以图表的形式呈现，比如一些信息图（尽管目前大多数信息图都是人工统计的数据），或者在处理之前对大数据集进行探索。

上面列举的几个阶段，每一个都面临着巨大的挑战，虽然工业界有一些解决方案，但离成熟还远得很。并且在面对不同的公司、不同的开发人员、不同的业务需求时，要将这几个阶段有机地整合起来更是难上加难。在其中起到核心作用的人就称为“数据科学家”。

0.2 如何成为数据科学家

读者应该知道这个问题很难回答，失败的原因总是相似的，成功的经历却各有不同。从来没有人靠复制他人的经历就能获得同样的成就，就像“人不能两次踏入同一条河流”的哲学观点一样，没有人可以复制别人的经历，更何谈成就。因此在回答这个问题时，我只假设一些概念上的前提条件：良好的计算机科学基础，较高的英文读写水平，极强的自学能力，还有一些个人品质比如耐心、毅力、乐于分享，等等。不过最重要的还是“兴趣”，我相信能花上几十块钱购买这本书的读者一定是有兴趣的，因为这本书是给那些对数据科学有一些了解，希望学习具体方法的人准备的。所以，即使上面所说的前提条件你一个都不具备，只

要有兴趣，那么让我们从现在就开始吧。

我需要数学或计算机科学的学位吗

最好有！如果你恰好是在校大学生，又碰巧学习数学或计算机相关专业（在这个程序员匮乏的年代，所有必修 C 语言的专业都称为“计算机相关专业”），希望你能学习好学校的课程，下面是一份技能清单，如果其中有一些技能没有在你的课程安排里，那么最好是通过选修或自学的方式进行补充。

- ❑ 一门编程语言
- ❑ 算法、数据库、操作系统
- ❑ 概率与统计、线性代数
- ❑ 英语

对于已经错过了花季、雨季的社会人来讲，如果你并非从事计算机程序开发的相关工作，上述几项技能对你来说可能要求太高了。不过，你还是需要多付出一些努力来补上这些知识，当然是在读过本书之后。得益于互联网的发达，很多教学资源都能够从网上获取，这里也向各位读者推荐一些好的网站。

- ❑ 编程学习：

https://www.codecademy.com/

https://www.codeschool.com/

这是国外的两家编程学习网站，拥有交互式解释器、美观的讲义，有一些课程还有手把手的视频教程，可能读英文对你来说有点慢，不过这是一个好的开始。

- ❑ 算法学习：

http://www.brpreiss.com/books/opus7/

这是由布鲁诺·R·普莱斯所著的一系列算法图书的在线版，包括 C++ 版、Java 版、C# 版、Python 版、Ruby 版、Lua 版、Perl 版、PHP 版、Objective-C 版等，你能想到的常用编程语言都有对应的版本，它们中的一部分有过正式引进的中文版，或者有爱好者翻译的版本，当然推荐阅读原版。

另外，本书会带领读者复习一下概率与统计和线性代数的基本概念，以及介绍一些 SQL 方面的知识。最后，不要忘记本书的目的是通过数据科学实战学习 Python 编程。希望读者在读过这本书之后，能有充分的知识来支持后续的学习。

0.3 为什么是 Python

通过书名，各位读者就应该知道这是一本讲解 Python 编程的书了。数据科学只是个引子，我希望能通过相关的例子和练习激发出读者的兴趣，帮助读者除掉编程这条拦路虎。在很多非计算机相关专业的人的概念里，编程是要归为玄学分类的，通过一堆意义不明的符号就能驱动计算机完成各种各样的任务，是不是有点像魔法师口中所念的咒语。但事实上，计算机只能做两件事情，执行计算并记录结果，只不过它的这两项能力远远超过人类大脑的能力（读者可能看过一些文章，其中有些研究声称尝试估算过人类大脑的计算能力，发现人脑的计算能力仍然比现今最先进的计算机还要快很多倍。但是人类大脑中有些模块，比如视觉、语言，是人类经过亿万年的演化，大自然进行极致优化所产生的结果。这里对于计算和存储能力的比较仅是指数学计算和文字存储方面）。以我正在使用的笔记本来说，其拥有主频为 2.5GHz 的双核处理器，总计约等于 50 亿次 / 秒的计算速度。而 512GB 的硬盘则可以存储 10 万本书（按每本书 5MB 计算，实际上 5MB 大小的书应该算是鸿篇巨著了。假如按 UTF-8[⊖]编码，每个中文占 3 ~ 4 个字节（byte），而 5MB 约有 500 万个字节，这至少是一本百万字的书）。如果想要使用计算机这种能力强大的工具，就需要掌握一门编程语言，用来和计算机进行沟通。虽然我也想为各位读者科普一下众多的编程语言，不过这毕竟是一本教授 Python 编程的书，所以这里只通过以下几个方面来阐述一下用 Python 作为数据科学工具的理由。

（1）简单易上手

Python 被誉为可执行的“伪代码”，其语法风格接近人类的语言，即使是第一次看代码的人也能很容易理解程序所要实现的功能，读者可以试着阅读下面这段代码[⊜]：

```
for i in range(0, 10):
    print(i)
```

上面的代码中 range 代表一段区间，0 代表下界，10 代表上界，通常 Python 程序的上下界是左闭右开的一个区间。for 的含义表示“这其中的每一个数”，print 就不言自明了，代表打印结果到屏幕上。

除了优雅的语法之外，Python 还属于解释性语言，我们可以不经过编译、链接等步骤直接获得程序执行的结果。而且 Python 还拥有交互式解释器，可以让我们随时随地测试我

⊖ UTF-8（8-bit Unicode Transformation Format）是一种针对 Unicode 的可变长度字元编码，可以编码世界上大部分语系的字符，也是使用最为普遍的一种编码方式。除了 UTF-8 之外还有专门的中文编码——GBK，日文编码——Shift_JIS 有在使用。

⊜ Python 代码的缩进应该总是 4 个空格，这是保证程序正确性、可读性的前提。

们的代码，如图 0-1 所示。

```
[jilu:src:% python
Python 2.7.9 (v2.7.9:648dcafa7e5f, Dec 10 2014, 10:10:46)
[GCC 4.2.1 (Apple Inc. build 5666) (dot 3)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
[>>> print("学习Python")
学习Python
[>>> 1 + 1 * 9
10
[>>> 3 > 2
True
>>>
```

图 0-1　初次使用 Python

（2）资源丰富、应用广泛

已经有很多书讲解了 Python 相关的技巧，比如《编程导论》是麻省理工学院（MIT）计算机科学导论的课程；《 Python 编程实战》是一本 Python 编程技巧进阶的好书，介绍了在 Python 中如何实践设计模式；《机器学习实战》主要讲解了机器学习的常见算法，其中使用 Python 编写了全部的代码；《 Python 高手之路》对如何使用 Python 构建大型系统提出了很多有益的见解。而且使用 Python 的知名项目也很多，比如 OpenStack 开源云计算平台就是由 Python 编写的，还有世界上最大的视频网站 YouTube 也是使用 Python 开发的，等等。当然 Python 在大数据应用上也有其独特的优势，科学计算库 NumPy 和 SciPy、绘图模块 Pylab、统计库 Pandas、机器学习库 Scikit-learn 都是为 Python 所设计的，现在流行的 Hadoop 和 Spark 也都提供了 Python 接口。可以说在“大数据”“数据科学”领域，如果某一个产品不支持 Python，那么其前景将会是难以想象的。

（3）跨平台、免费

Python 官方提供了多平台的解释器，包括 Windows、Mac OS X、Linux 甚至更多的其他平台，你所写的 Python 代码，可以在不经修改的情况下移植，比如在 Windows 上开发，在 Linux 服务器上运行，不会有任何问题。而且 Python 是免费且开源的，不仅标准库可以随意阅读其源码，连官方解释器的 C 语言实现也可以获得其源码。Python 社区是鼓励分享的，读者不仅可以从中学到很多编程的技巧，甚至还可以做出一些贡献。

0.4 一个简单的例子

下面是一段用 Python 编写的有趣的代码，这里所用的模块并不会在本书中进行讲解，仅仅是向购买本书的你表示我的感激。

代码清单如下：

```
# ! /usr/bin/python
# -*- coding: utf-8 -*-
import sys

from colorama import init
init(strip=not sys.stdout.isatty())
from termcolor import cprint
from pyfiglet import figlet_format

cprint(figlet_format('welcome', font='starwars'),
       'yellow', 'on_blue', attrs=['bold'])
```

其输出的结果如图 0-2 所示。

图 0-2 打印艺术学的 Python 程序

这段代码非常酷，它会将一个英文单词转换成字符拼接的文字，如果你还看不懂该程序，也没关系，在学完第 1 章之后你就能明白这段代码的含义了，祝你阅读愉快。