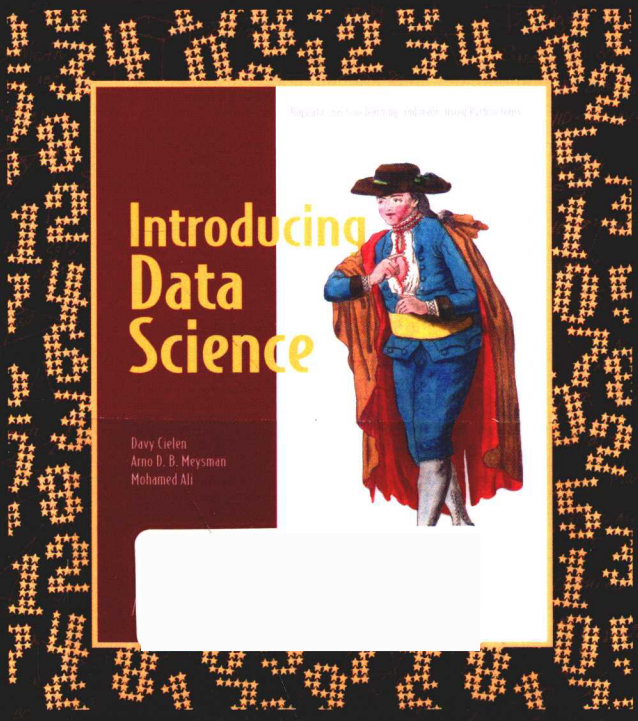
HZ BOOKS 华章IT
MANNING
数据科学与工程技术丛书

Python数据科学导论

戴维·西伦（Davy Cielen）
[法] 亚诺 D. B. 梅斯曼（Arno D. B. Meysman） 著
穆罕默德·阿里（Mohamed Ali）

王艳 刘义 于晨昕 王丽娜 陈南 译



INTRODUCING DATA SCIENCE
BIG DATA, MACHINE LEARNING, AND MORE,
USING PYTHON TOOLS

机械工业出版社
China Machine Press

数据科学与工程技术丛书

INTRODUCING DATA SCIENCE
BIG DATA, MACHINE LEARNING, AND MORE, USING PYTHON TOOLS

Python数据科学导论

戴维·西伦（Davy Cielen）
[法] 亚诺 D. B. 梅斯曼（Arno D. B. Meysman） 著
穆罕默德·阿里（Mohamed Ali）

王艳 刘义 于晨昕 王丽娜 陈南 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 数据科学导论 /（法）戴维·西伦（Davy Cielen）等著；王艳等译 . —北京：机械工业出版社，2017.8
（数据科学与工程技术丛书）
书名原文：Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools

ISBN 978-7-111-57826-0

I. P… II. ①戴… ②王… III. 软件工具－程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2017）第 202439 号

本书版权登记号：图字：01-2016-5128

本书涵盖的主题非常广泛，介绍了数据科学方方面面的知识，每一章都侧重于介绍数据科学的某一方面，为读者以后的深入学习打下基础。具体内容包括：第 1、2 章系统介绍大数据科学的背景知识及框架结构；第 3～5 章介绍机器学习相关知识；第 6～9 章介绍几个比较有趣的数据科学主题。

本书是学习数据科学知识的入门教材，在深入学习本书的实例前，需要掌握 SQL、Python 及 HTML5 的入门知识，并了解统计学和机器学习相关知识。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：朱秀英　　责任校对：李秋荣
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2017 年 8 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：14.75
书　　号：ISBN 978-7-111-57826-0　　定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

译 者 序

数据科学是一门新兴的学科，最早在20世纪60年代被提出，但当时并未受到学术界的广泛关注和认可。1996年在日本召开的“数据科学、分类和相关方法”会议，已经将数据科学作为会议主题词。目前，数据科学的研究在各个领域受到越来越多的关注。

数据科学的理论基础包括统计学、机器学习、数据可视化以及某一特定领域的知识。其研究内容包括数据科学基础理论、数据预测模型、数据计算以及数据管理。研究过程包括：数据的获取；对数据集的观测，并发现整体特征；进行数据分析，例如使用数据挖掘技术；发现数据规律，并完成数据预测。

本书第1、2章介绍了数据科学的理论背景和框架，为本书其他章节的学习提供了基础。第3～5章介绍了将机器学习应用在不同的大数据集上的案例。第6～9章介绍了数据科学中一些有趣的主题，如NoSQL数据库、文本挖掘等。在阅读本书前，建议读者对SQL、Python、HTML5以及统计或机器学习有一些了解。本书作者Davy Cielen、Aron D. B. Meysman和Mohamed Ali具有丰富的大数据及数据科学经验，他们分别在比利时和英国联合创立了两家数据科学公司，专注于大数据处理及数据科学相关的研究，并为很多大公司提供数据科学领域的咨询工作。

本书由多位译者联合翻译，其中，王艳翻译了前言、第2、4章和附录D；王丽娜翻译了第1章和第5章；刘义翻译了第3、8章以及附录C；于晨昕翻译了第6、7章以及附录A和附录B；陈南翻译了第9章。感谢机械工业出版社华章公司的编辑在翻译过程中提出的诸多宝贵建议。由于译者的水平及经验有限，难免存在错误和纰漏，恳请广大读者指正。

译者

2017年3月

前　言

本书传递的知识永存我们心中。人类之所以为人类，人类之所以是现在的样子，数据科学技术功不可没。这本书不仅介绍计算机驱动的数据科学相关知识，还将教给读者洞察连接的能力，以及如何以事实为依据演绎出结论，如何从过去的经历中汲取经验。人类比地球上的任何其他生物更依赖于大脑。人类的生存依赖于人脑，人类在大自然中的位置完全取决于人脑的特性。古往今来，这一战略解决了人类所面临的所有问题，在不久的将来，人类也不太可能改变它。

当谈到原始计算时，人类的大脑只能引领我们走到目前的境地。现在，我们每天都接收到海量的数据，人脑分析已经无法跟上大数据时代信息所包含的潜在内容，我们已掌握的知识更难以满足人类的好奇心。因此，我们利用机器为我们做一部分工作，比如：模式识别，创建连接，以及为人类的众多问题探寻答案。

对知识永无止境的探索是人类的基因，依赖计算机为人类完成一些力所能及的工作是我们的使命。

致谢

非常感谢 Manning 出版社所有参与本书制作的人员，在你们的帮助下本书得以顺利出版。

感谢 Ravishankar Rajagopalan 对本书的书稿做了全面细致的技术校对，感谢 Jonathan Thoms 和 Michael Roberts 给了许多专业的建议。另外感谢众多的评审人员，他们在本书的制作过程中提供了许多极有价值的意见反馈，他们是：Alvin Raj, Arthur Zubarev, Bill Martschenko, Craig Smith, Filip Pravica, Hamideh Iraj, Heather Campbell, Hector Cuesta, Ian Stirk, Jeff Smith, Joel Kotarski, Jonathan Sharley, Jörn Dinkla, Marius Butuc, Matt R. Cole, Matthew Heck, Meredith Godar, Rob Agle, Scott Chaussee, Steve Rogers。

首先，我想感谢我的妻子 Filipa，她给了我灵感和动力，让我得以战胜所有的困难。感谢她在我的职业生涯和创作这本书的过程中，始终陪伴在我身边。感谢她担负起家庭的重担，当我不在的时候独自照顾我们的小女儿，让我有了充裕的时间去追求我的目标并实现抱负。谨以此书向我的妻子致敬，非常感谢她为我们的小家庭所做的无私奉献。

同时，我想感谢我的女儿 Eva 以及我未出生的儿子，他们给了我极大的欢乐并让我笑口常开。他们活泼有趣、充满爱心，是上帝送给我的最好的礼物，也是我所期望的最完美的小孩，和他们在一起总是充满了乐趣。

特别要感谢我的父母，谢谢他们对我长期以来的支持。他们无尽的爱和鼓励让我从容完成了这本书，实现了人生的一个阶段目标，并继续我人生新的旅程。

同时，真诚地感谢同我一起共事的小伙伴们，谢谢大家齐心协力，一起攻坚了一个又一个难题。特别要感谢 Mo 和 Arno，他们给了我最有力的支持和很好的建议。非常感激大家在本书的创作过程中付出的时间和精力，你们棒极了！没有你们，我可能都不会写这本书。

最后，真诚地感谢每一位支持我、理解我的朋友们。我常常忙得没有空闲时间，谢谢你们的关爱和一如既往的支持，让我能够专心创作并完成这本书。

Davy Cielen

非常感谢我的家庭和我的朋友们，他们在我完成本书的过程中，给了我一如既往的支持和鼓励。外面的新鲜事物很多，能在家完成这本书的创作真的很不容易，谢谢大家！特别要感谢我的父母，我的兄弟 Jago，还有我亲爱的女朋友 Delphine。不管我有什么疯狂的想法和离奇的举动，你们一直坚守在我身边，不离不弃。

同时，谢谢我的教母，还有我的教父，他正在与癌症作斗争，但他们的积极乐观让生活充满了希望。

还要感谢我的朋友们，他们给我买啤酒。也谢谢我女朋友 Delphine 的父母，她的兄弟 Karel 和未过门的妻子 Tess，谢谢你们的热情款待和美味佳肴。

大家为了美好的生活而努力奋斗着。

最后并且是最重要的一点，我想谢谢本书的合著者也是我的铁哥们 Mo, 以及本书的另一位合著者 Davy，谢谢你们深刻的洞察和独特见解。为了成为一名企业家和数据科学家，我们每天共享跌宕起伏的人生，这是一段多么精彩的旅程，我相信我们的未来会更精彩。

Arno D. B. Meysman

首先最重要的一点是我要感谢我的未婚妻 Muhuba，谢谢她的爱、理解、关心和包容。最后，感谢 Davy 和 Arno，和他们一起度过了很多开心时光并让我们的创业梦想成真。他们坚持不懈的奉献是我完成本书至关重要的资源。

Mohamed Ali

关于本书

我只能带领你到门口，前面的路靠你自己去前行。

——Morpheus，《The Matrix》

欢迎阅读本书！你可能已经注意到本书涵盖的主题非常广泛，我们希望本书能为你介绍数据科学方方面面的知识，为你以后深入的学习打下基础。数据科学的领域非常宽广，即使用10倍于本书的厚度，也难以详尽介绍数据科学的所有知识。因此在内容的组织上，我们精心挑选出一些比较有趣的内容。本书的每一章都侧重于介绍数据科学的某一方面。真心希望学完本书后，每一位读者都能从中受益。

希望本书是一个切入口——一扇通往激动人心的数据科学世界的大门。

阅读导览

本书的第1、2章系统地介绍大数据科学的背景知识及框架，有助于读者理解本书后面章节的内容。

- 第1章重点介绍数据科学及大数据，以Hadoop应用实例作为结尾。
- 第2章介绍数据科学过程，包括数据科学项目的一般步骤。

第3～5章介绍将机器学习应用到不断增加的大数据集上的案例。

- 第3章介绍数据量不大的情况，此时数据可存储在一般计算机上。
- 第4章介绍大数据带来的一系列挑战，这里的大数据是就数据的规模而言的，即海量数据（large data）。数据虽然可以存储在硬盘上，但存储在RAM上将变得日益困难。如果没有计算集群技术，一切将变得非常困难。
- 第5章开始介绍大数据（big data）相关知识，使用多台计算机协同工作是不可避免的。

第6～9章介绍数据科学几个比较有趣的主题，它们或多或少相互独立。

- 第6章介绍NoSQL并阐述为什么它与关系数据库不同。
- 第7章介绍如何将数据科学应用到流数据中。这里所面临的主要问题不是数据规模，而是数据生成和旧数据过时的速度。
- 第8章介绍文本挖掘。不是所有的数据都以数字开始。当数据以文本格式存在

时，比如邮件、博客、网站等，文本挖掘和文本分析变得非常重要。

- 第 9 章通过介绍若干实用的 HTML5 工具来关注数据科学过程的最后一部分——数据可视化和原型应用程序的构建。

附录 A～附录 D 介绍前面章节出现过的 Elasticsearch、Neo4j、MySQL 数据库以及 Python 代码包 Anaconda（对数据科学来说尤其有用）的安装和设置。

本书读者

本书是学习数据科学知识的入门教材。对于资深的数据科学家而言，本书只触及数据科学的基础知识。对于其他读者而言，深入学习本书的实例前，应有以下预备知识：SQL、Python 和 HTML5 的入门知识，以及统计学或机器学习的相关知识。

代码下载

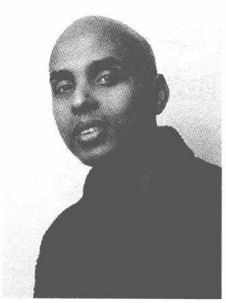
本书的实例采用 Python 语言脚本讲解。在过去的十几年中，Python 语言发展成一门广受推崇并广泛使用的数据科学语言。

本书包含的大多数案例代码可以从以下在线代码库中查阅：https://www.manning.com/books/introducing-data-science。

关于作者

Davy Cielen，经验丰富的企业家、作家及大学教授。他与 Arno 和 Mo（本书另两位作者）一起在比利时及英国创建了两家数据科学公司，同时本书的三位作者在索马里兰合伙创建了第三家数据科学公司。这些数据公司关注大数据科学战略，偶尔也为其他大公司做数据咨询。Davy 是法国里尔 IESEG 管理学院的副教授，一直从事大数据科学领域的教学和科研工作。

Arno Meysman，企业家及数据科学家。作为数据科学家，Arno 兴趣爱好广泛，致力于医学分析、零售和游戏分析。他认为将数据的洞察分析与想象相结合，将在很大程度上帮助我们更好地改善这个世界。



Mohamed Ali，企业家及数据科学咨询顾问。他的兴趣集中在两个领域：数据科学及可持续发展的项目。

关于封面插图

本书封面的插图取自于 1805 年版的 Sylvain Maréchal 的地域服饰习俗的四卷纲要。这本书第 1 版于 1788 年在法国巴黎出版，当时正值法国大革命的前一年。书中每一幅插图均采用手工绘制。本书封面所用插图的标题是“ Homme Salamanque”，意为从萨拉曼卡来的男人，萨拉曼卡位于西班牙西部的一个省，与葡萄牙交界。该小城以天然美景、茂盛的森林、古老的橡树、崎岖蜿蜒的山脉和古老的历史而闻名遐迩。

“ Homme Salamanque”正是 Maréchal 收藏的丰富多彩的人物插图之一，这一收藏生动地展示了 200 年前世界不同城镇和地区的独特性和个性化。在那个时代，从人们的不同服饰和着装即可判断他们来自哪个地域，是否属于同一地域，彼此相隔有多远。该收藏栩栩如生地向我们展示了这一时期不同地域的分隔和距离感，也展现了这一历史时期与其他年代的显著区别。

从那以后着装规范逐步改变，区域服饰的多样性慢慢消失了。当今社会，人们很难再从服饰上区分出某个居民来自于哪一大陆。也许，文化多样性的消失带来的是更多样化的个人生活——肯定是更多变且快节奏的科技生活。

我们以 200 多年前丰富多彩的、多元地区的生活展示作为本书的封面，由 Maréchal 的图片回归现实生活，以此颂扬计算机业务的创造力、原创性，重新揭示多样性的价值。

目　录

第 1 章
大数据世界中的数据科学

本章涵盖：

- 数据科学和大数据的定义。
- 数据的不同类型。
- 洞察数据科学的过程。
- 数据科学和大数据领域的介绍。
- Hadoop 实例。

大数据是对使用传统数据管理技术（例如，RDBMS（Relational Database Management System，关系型数据库管理系统））无法处理的大规模、复杂数据集合的总称。关系型数据库管理系统一直被广泛使用，并被认为是唯一的解决方案，但对于处理大数据的需求，它显示出了缺陷。数据科学涉及了对大量数据的分析以及对数据包含的知识方法的提取。你可以认为，大数据和数据科学之间的关系，就如同原油厂和炼油厂之间的关系。数据科学和大数据都源自于统计和传统的数据管理，但现在又被认为是不同的学科。

大数据通常被认为具有三个 v 的特征：

- 量（volume）——有多少数据？
- 类（variety）——有多少种不同类型的数据？
- 速率（velocity）——在以什么样的速度生成新的数据？

这些特征往往又被辅助于第四个 v——准确性（veracity）：数据的精确性有多高？这 4 个属性是大数据不同于传统数据管理工具的数据。因此，大数据带来的挑战几乎是每一个方面的：数据采集、维护、存储、搜索、共享、转换以及可视化。此外，大数据需要特殊的技术进行提取。

数据科学是使用统计的方法处理当前产生的海量数据的一种延伸。他把源自计算机科学的方法添加到统计方法中。在 Laney 和 Kart 的一篇研究论文“Emerging Role of the Data Scientist and the Art of Data Science”中，作者通过对上百个数据科学家、统计学家和商业智能分析师工作描述的筛选，来找出这些职务的差别。数据科学家和统计学家最主要的差别是，数据科学家有处理大数据的能力，具有机器学习、计算以及算法搭建的经验。他们使用的工具也是有差异的，在数据科学家的工作描述中，高频率地提到使用 Hadoop、Pig、Spark、R、Python 和 Java 的能力。虽然本书重点关注 Python，但是不要过分担心这些工具，

本书会逐步介绍这些工具的大部分。对于数据科学来说，Python 是一个伟大的语言，因为它包含了许多可用的数据科学库，并被专业软件所支持。例如，广泛使用的 NoSQL 数据库有一套特定于 Python 的 API。正是由于这些特性以及 Python 的快速原型能力，使 Python 的影响力在数据科学领域逐步增长。

伴随着大数据的持续增长，对大数据利用的需要也越来越重要，每一个数据科学家在他们的职业生涯中都会遇到大数据项目。

1.1 数据科学和大数据的好处和用途

几乎所有的商业和非商业机构都在使用数据科学和大数据。使用大数据的实例非常多，本书中会提及一些用例，但是只涉及很少的一部分。

几乎所有行业的企业都在使用数据科学和大数据去分析他们的用户、流程、员工、实现和产品。很多企业还应用数据科学为用户提供更好的用户体验，以及交叉销售、向上销售，进而个性化他们的产品。例如 Google AdSense，从互联网收集用户数据，根据收集的信息，使互联网的用户得到更加匹配的商业信息。MaxPoint (http://maxpoint.com/us) 是另一个实时个性化广告的实例。人力资源专员使用人类分析方法和文本挖掘来筛选候选人，关注员工的情绪，研究同事之间的关系。在《 Moneyball: The Art of Winning an Unfair Game 》一书中介绍了人类分析理论。在该书（以及电影）中，我们看到美国棒球传统的队员位置安排是随机的，但是通过相关信息的分析，可以改变比赛结果。依靠统计数据，他们可以聘请合适的球员，并把他们安排到合理的位置对抗对手，让队员的能力、优势得到最大的发挥。金融机构利用数据科学预测股票市场，确定贷款的风险，并学习如何吸引新客户。在写这本书的时候，全球贸易至少 50% 是由机器自动执行，机器的算法是基于 quants 理论。从事交易算法的人通常被称为数据科学家，数据科学家帮助大数据和数据科学技术的发展。

政府部门也逐渐开始意识到数据的重要性。很多政府机构不仅依赖于内部数据科学家发现有价值的信息，也与公众共享他们的数据。用户可以使用这些数据来了解趋势或者构建数据驱动的应用程序。例如 Data.gov，美国政府的公开数据都在上边。数据科学家在政府部门会接触各种各样的项目，例如检测欺诈、其他犯罪活动或优化项目资金。一个著名的例子是爱德华·斯诺登泄露了美国国家安全局和英国政府通信总部内部文件，文件清楚地显示了他们如何使用数据科学和大数据监控数以百万的人。这些组织从广泛的应用中收集了 50 亿数据，比如谷歌地图、愤怒的小鸟、电子邮件和短信，以及其他数据源。然后应用数据科学技术来提取信息。

非政府组织（NGO）也轻车熟路地使用数据。他们用它来筹集资金和捍卫自己的事业。例如世界野生动物基金会（WWF），雇佣数据科学家帮助他们提高筹款的有效性。许多数据科学家利用自己的业余时间帮助非政府组织，因为非政府组织往往缺乏资源来收集数据和雇佣数据科学家。DataKind 是一个数据科学家组织，该组织致力于造福人类的公益事业。

大学也在应用数据科学做研究，同时也可以提高学生的学习经验。大规模网络公开课（MOOC）的崛起也产生了大量数据，这些数据帮助大学研究怎么使这种类型的学习方式补充传统课堂。MOOC 是一个无价的资产，如果你想成为一个数据科学家和大数据专家，一定要看这几个知名的网络公开课网站：Coursera、Udacity 和 edX。大数据和数据科学更新速度快，网络公开课让你从顶尖大学课程获得最新知识。如果你不了解它们，现在正是时间去学习。当你真正了解了数据科学，你会和我一样，爱上它们。

1.2　数据种类

数据科学和大数据中有许多不同类型的数据，每种数据都需要不同的工具和技术。主要分为以下数据类型：

- 结构化数据
- 非结构化数据
- 自然语言数据
- 计算机数据
- 图类数据
- 音频、视频和图像数据
- 流数据

让我们一起来探索这些有趣的数据类型吧。

1.2.1　结构化数据

结构化数据是数据依存于数据模型和记录驻留在一个固定的字段中。因此，结构化数据往往容易存储在数据库或 Excel 文件中（如图 1-1 所示）。SQL 或结构化查询语言是数据库中管理和查询数据的首选方法。你也可能遇到，结构化数据让你很难将其存储在一个传统的关系数据库中。家谱的分层数据就是一个这样的例子。

然而，世界并不是由结构化的数据组成的，它是由人类和机器强加给它的。更多的时候，数据是非结构化的。

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

图 1-1　结构化数据的一个典型例子是 Excel 表格

1.2.2　非结构化数据

非结构化数据内容是上下文特定的或多样的，使其不容易适应数据模型。典型示例是大家通常使用的电子邮件（如图 1-2 所示）。虽然电子邮件包含结构化元素，如发件人、标题和正文文本。但是从一些人的邮件中找到谁写了一封电子邮件投诉一个特定的员工，仍然是一个挑战，因为现在有太多的方式指向一个人。成千上万的不同语言和方言进一步增加查找的难度。

一个人手工发出的电子邮件，如图 1-2 所示，就是一个完美的自然语言数据的示例。

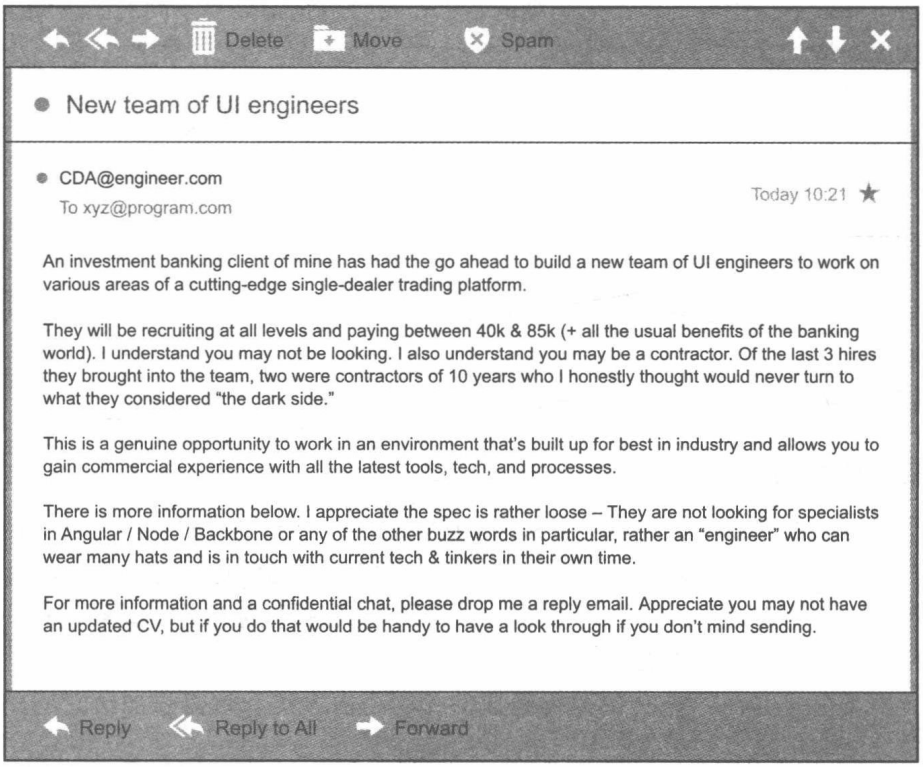


图 1-2　电子邮件是非结构化数据和自然语言数据的一个例子

1.2.3　自然语言数据

自然语言是一种特殊类型的非结构化数据；处理起来更加富有挑战性，因为它需要具备数据科学技术和语言学知识。

自然语言处理社区在实体识别、主题识别、总结、文本完成和情感分析方面取得了成功，但是一个领域的训练模型不能很好地推广到其他领域。即使是最先进的技术也不能够解读每一块文本的意思。这是人与自然语言的斗争。由于语言的含糊不清的性质，语义本身也是值得怀疑的。两个人听到的是同样的谈话，他们会得到相同的语义吗？当某人沮丧或快乐时说出相同的单词，意义可能会有所不同。

1.2.4　计算机数据

计算机数据是没有人工干预的，由计算机、过程、应用程序或其他机器自动创建的信息。计算机数据正在成为一个主要的数据资源，也是今后的发展趋势。Wikibon 曾预测，*互联网产业*（industrial Internet）（互联网产业一词由 Frost & Sullivan 定义，是指复杂的物理机械及网络化传感器和软件的集成）的市场价值在 2020 年达到约 5400 亿美元。国际数据公司（International Data Corporation，IDC）预估在 2020 年，事物的连接会是人类连接的 26 倍。这个网络就是通常所说的*物联网*（the internet of things）。

由于计算机数据的大容量和高速度，计算机数据的分析依赖于高度可伸缩的工具。例如 Web 服务器日志、呼叫详细记录、网络事件日志和遥测（如图 1-3 所示）都是计算机数据。