第 1 章　探索现代数据科学领域

如果你正在阅读本书，那么很可能你已经听说过数据科学。可以说，它是科技和 STEM 领域中增长最快、讨论最多的职业之一，同时仍然保持着其相对的前沿性和神秘感。也就是说，许多人听说过数据科学家，但很少有人知道他们做什么、他们如何创造价值，或者如何从零开始进入这个领域。

本章将通过一个实际的描述验证数据科学的定义。随后将讨论大多数数据科学工作的内容，同时花一些时间描述不同类型数据科学之间的区别。接下来将深入探讨进入数据科学的各种途径，以及获得第一份工作为何如此具有挑战性。

在本章结束时，读者将对现代数据科学家有深入的理解，了解获得该工作的多种途径，在成为数据科学家的旅程中可以期待什么，预期的障碍，以及应该掌握的技能。

本章主要涉及下列主题。

（1）数据科学是什么。

（2）探索数据科学过程。

（3）分析数据科学的不同类型。

（4）审视数据科学的职业路径。

（5）解决经验瓶颈问题。

（6）理解预期的技能和能力。

（7）探索数据科学的演变。

1.1　数据科学是什么

首先，我们给出数据科学的定义。根据维基百科，“数据科学是一个跨学科的学术领域，它使用统计学、科学计算、科学方法、过程、算法和系统从嘈杂的、结构化的和非结构化的数据中提取或推断知识和洞察力”[1]。它包含多种技术、程序和工具来处理、分析和可视化数据，使企业和组织能够做出数据驱动的决策和预测。数据科学的主要目标是识别数据中的模式、关系和趋势，以支持决策制定和创建可操作的洞察结果。

《哈佛商业评论》将数据科学称为 21 世纪最“性感”的工作之一[2]，关于数据科学家获得六位数高薪的故事并不罕见。数据科学家通常被视为组织中的神谕，回答诸如“如果

向这群客户增加产品供应，我们能增加收入吗？”或“客户流失的常见原因是什么？”等复杂的商业问题。

在组织内部，对数据科学家技能的需求持续增长。美国劳工统计局预测，在 2022 年，数据科学家的工作岗位在未来 10 年内将增长约 36%[3]。对数据科学家需求的增长是由几个因素推动的，这些因素如图 1.1 所示。

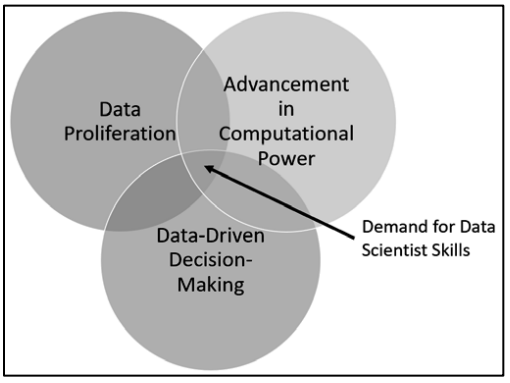

图 1.1　数据科学家需求增长的原因

原文	译文
Data Proliferation	数据激增
Advancement in Computational Power	计算能力的提高
Demand for Data Scientist Skills	对数据科学家技能的需求
Data-Driven Decision-Making	数据驱动的决策制定

首先是数据的激增。由数字设备、社交媒体和各种其他来源产生的数据呈指数级增长，这使得组织必须利用这些数据进行决策和创新。预计未来数据增长将持续下去，国际数据公司（IDC）预计到 2025 年，我们将每年产生 175ZB 的数据[4]。这是一个惊人的数据量。

组织机构希望利用数据可用性的激增为决策生成洞察结果。随着世界变得更加互联和复杂，基于证据的决策需求增长，导致对能够将数据转化为可操作洞察力的技能型数据科学家的需求增加。组织和企业越来越依赖于数据驱动的洞察结果在市场上获得竞争优势，优化运营，并改善客户体验。

最后，如果没有计算能力的提高以及工具和平台的进步，将数据转化为洞察力是无法实现的。计算能力的增强和高级算法的发展，特别是在机器学习（ML）和深度学习（DL）方面，使得高效处理和分析大量数据成为可能。此外，开源工具、库和平台的发展使数据科学更容易被更广泛的受众接受，从而促进了这一职业的发展。

因此，数据科学仍是一个不断发展的领域，预计它将与计算和技术创新（如生成式人工智能）并行增长。此外，随着公司继续以更大的兴趣拥抱数字时代，最大化数据的效用，并利用其背后的洞察力获得竞争优势，对数据科学家的需求也将进一步扩大。

然而，尽管数据科学通常被视为并描述为一项单一的功能，但你很快就会了解到它是一个多方面的学科，通常因团队、部门甚至公司而异。自然地，数据科学家的工作概况也是一个不断演变的描述，但我们将涵盖最常见的任务。

1.2　探索数据科学过程

数据科学工作通常是一个迭代过程，如果数据科学家遇到挑战，他们需要返回到早期步骤。对数据科学过程的分类方法有很多，但通常包括：

（1）数据收集。

（2）数据探索。

（3）数据建模。

（4）模型评估。

（5）模型部署和监控。

1.2.1　数据收集

数据收集和预处理涉及从各种来源（如数据库、API 和网络抓取）收集数据，然后清洗和转换数据，以准备进行分析。这一步涉及处理缺失、不一致或嘈杂的数据，并将其转换为结构化格式。

根据组织机构的不同，数据工程师团队会支持数据科学过程中的这一步骤。然而，数据科学家通常也需要管理这一过程。这要求他们对数据源有深入的了解，并能够编写结构化查询语言（SQL）进行查询，编写可以查询数据库的代码，或者使用如网络爬虫等自定义工具来收集所需的数据。

1.2.2　数据探索

数据探索涉及进行探索性数据分析（EDA），以更好地理解数据、检测异常，并识别变量之间的关系。这一步骤的关键是寻找相关性并理解数据的分布。这涉及使用描述性统计和可视化技术来总结数据并获得洞察；因此，数据科学家应该能够使用汇总统计数据，编程实现描述

性可视化结果，或者利用如 Power BI 或 Tableau 等报告工具来创建强大的图表。

1.2.3　数据建模

利用在数据探索步骤中学到的知识，数据建模是数据科学家使用机器学习和统计技术构建预测性或描述性模型的步骤，这些技术能够识别数据中的模式和关系。这里，数据科学家应选择适当的算法，在历史数据上训练模型，并验证它们的性能。

1.2.4　模型评估

模型评估和优化涉及使用诸如准确度、RMSE（均方根误差）、精确度、召回率、AUC（曲线下面积）或 F1 分数等指标评估模型的性能。基于这些评估，数据科学家可能会细化模型或尝试替代算法以提高它们的性能。理解模型预测背后的根本原因是建立对其结果信任和确保其与领域知识一致的关键。因此，数据科学家必须确保模型解决了组织/业务目标。这里，数据科学家需要能够将他们的发现传达给可能的技术性和非技术性个体。

1.2.5　模型部署和监控

模型部署和监控涉及将模型实施到现实世界的应用中，监控它们的性能并维护它们，以确保持续准确性和相关性。例如，数据科学家可能与数据工程团队合作或使用如容器等工具实施模型。一旦部署，数据科学家还可能需要开发仪表板来监控模型的性能，并在性能超出预期范围时通知利益相关者。

可以看到，数据科学是一个包含许多与数据相关的任务的职业，特别是那些涉及以某种形式获取、准备和交付数据的任务。虽然数据建模是这份工作最吸引人的地方，但实际上其他所有任务大约占据了工作内容的 80%。这还不包括与数据无关的任务，如与利益相关者交流、收集需求、调试软件、查看电子邮件和研究。然而，这些任务并不一定是数据科学家独有的。

在了解了与工作相关的常见任务后，接下来探索数据科学的不同类型或风格。

1.3　分析数据科学的不同类型

现在我们已经定义了数据科学家角色的一些关键方面，很明显该角色通常涵盖了许多

不同的技能。数据科学家经常被要求执行各种与数据相关的任务，包括设计数据库表以收集数据、编程机器学习算法、理解统计学，以及创建引人注目的视觉效果以帮助向他人解释有趣的发现，但任何一个人要精通所有这些技能领域都是困难的。

因此，我们经常看到数据科学家在一两个领域特别熟练，而在其他领域则具备基本能力。他们的才能可以被认为是 T 形的，如图 1.2 所示。他们精通某些领域，如 T 的水平线，同时他们在少数领域具有深厚的知识和专长，如字母的垂直部分所示。

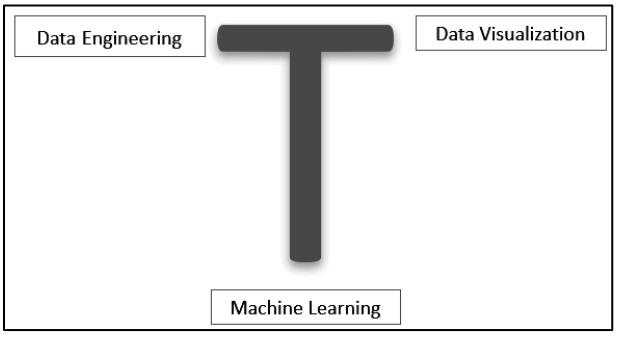

图 1.2　能力 T 形示例

原文	译文
Data Engineering	数据工程
Data Visualization	数据可视化
Machine Learning	机器学习

这个例子说明，有些人在数据工程和可视化方面能力足够，但在 ML 方面却出类拔萃。这些能力往往与一个人的独特经历或兴趣有关。也许他们曾经主修统计学，后来喜欢上了机器学习；也许他们曾经是一名商业智能（BI）工程师，在数据提取、转换和加载（ETL）方面拥有丰富的经验，从而能够更快地掌握数据工程概念。

无论出于何种原因，有些人对某些概念的掌握自然要比其他人好。在阅读本书时，请务必牢记这一点。虽然并不期望读者精通数据科学的方方面面，但我们希望你掌握相关的基础知识。不过，你几乎肯定会发现自己的“能力 T 形”，即三位一体的顶级技能组合，这将巩固你在数据科学领域的身份。

技能专长的组合数不胜数，下面回顾一下读者会遇到的一些最常见的组合。

（1）数据工程师。

（2）仪表板和可视化专家。

（3）机器学习专家。

（4）领域专家。

1.3.1　数据工程师

如前所述，数据工程是数据科学过程中的一个关键方面，涉及数据收集、存储、处理和管理。它专注于设计、开发和维护可扩展的数据基础设施，确保用于分析和建模的高质量数据的可用性。数据工程师以对数据管道 ETL 过程的监督而闻名。在一些数据科学家团队中，尤其是在较小的组织机构内，数据工程职责属于数据科学团队。因此，专门从事这一领域的数据科学家可以帮助团队项目进行数据收集和存储，了解机器学习过程的需求，如将数据结构化，以便它可以高效地输入深度学习算法中。

数据工程师有大量工具可供选择。我们不能指望任何一位数据工程师掌握所有这些技术，尤其是在相同的能力水平上。事实上，工程师的资历越深，他们在所选工具方面的能力就越强。此外，这并不是一份全面的清单。不过，你可以在数据工程师简历中看到以下内容：

（1）编程语言：Python、SQL、Scala、R、C++。

（2）数据存储：关系数据库（例如，MySQL、PostgreSQL、Oracle）、NoSQL 数据库（例如，MongoDB、Cassandra、DynamoDB）、数据仓库（例如，Snowflake、Redshift、BigQuery）、分布式文件系统（例如，Hadoop 分布式文件系统（HDFS）、Apache Cassandra）。

（3）数据处理和分析：Apache Spark、Apache Flink、Apache Storm、Apache Beam、MapReduce、Hadoop、Hive、Apache Kafka、Amazon Kinesis。

（4）数据集成和 ETL：Apache NiFi、Talend、Apache Airflow、AWS Glue、Google Cloud Dataflow、dbt。

（5）数据版本控制和协作：Git、GitHub、GitLab、Bitbucket、Azure DevOps。

（6）数据可视化和商业智能：Tableau、Power BI、Looker、QlikView、Domo。

（7）云平台和基础设施：Microsoft Azure、Google Cloud Platform（GCP）、Amazon Web Services（AWS）。

（8）容器技术：Docker、Kubernetes。

1.3.2　仪表板和可视化专家

数据可视化是使用图表、图形和地图等视觉元素对数据和信息进行图形表示。它使利益相关者能够理解数据中的复杂模式、趋势和关系，从而做出更明智的决策。数据可视化有助于简化复杂数据并以易于消化的格式呈现，识别数据中的模式、趋势和相关性，支持

数据驱动的决策，并有效地向广大受众传达洞察和发现结果。将数据可视化与引人入胜的叙述结合起来，可以成为推动组织行动的强大动力。许多新闻机构聘请擅长数据可视化的优秀数据科学家，向他们的受众传达复杂信息。

仪表板和可视化专家在不同组织中有不同的称呼，但最常听到的一些名称包括商业智能（BI）工程师、数据分析师、数据可视化专家、数据叙述者，等等。他们通常是具有描述性统计、数据叙述和开发关键绩效指标（也称为 KPI）的强大背景的个人。仪表板和可视化专家使用的一些最常见的工具包括：

（1）编程语言：Python、SQL、R、JavaScript。

（2）数据存储：关系数据库（例如，MySQL、PostgreSQL、Oracle）、NoSQL 数据库（例如，MongoDB、Cassandra、DynamoDB）、数据仓库（例如，Snowflake、Redshift、BigQuery）。

（3）框架：Dask、Plotly、ggplot2、Shiny、Matplotlib、Seaborn、DB.js。

（4）数据可视化和商业智能：Tableau、Power BI、Looker、QlikView、Domo、Funnel、Excel。

（5）云平台和基础设施：Microsoft Azure、GCP、AWS。

1.3.3　机器学习专家

当大多数人想到数据科学家时，他们会想到设计和实现机器学习算法的人。机器学习专家和工程师利用计算机在没有显式编程的情况下从经验中学习并改进，通过开发算法和模型分析数据、识别模式，并根据这些模式进行预测或决策。他们在构建智能应用程序和系统中发挥着关键作用。机器学习专家对使用哪种学习算法以及如何调整其参数以实现最佳性能有很强的理解。

因此，他们有很强的研究倾向，以保持对最新定量问题解决方法的了解，并且特别擅长机器学习的开发、部署和维护任务。他们拥有强大的工具集，因为他们非常精通软件开发原则。虽然这不是固有的规则，但许多机器学习专家往往在统计学、运筹学、计算机科学和/或信息系统方面有很强的背景。机器学习专家使用的工具可能包括：

（1）编程语言：Python、SQL、R、Java、C++。

（2）框架：TensorFlow、Keras、scikit-learn、PyTorch、H2O、Hugging Face。

（3）数据存储：关系数据库（例如，MySQL、PostgreSQL、Oracle）、NoSQL 数据库（例如，MongoDB、Cassandra、DynamoDB）、数据仓库（例如，Snowflake、Redshift、BigQuery）、分布式文件系统（例如，HDFS、Apache Cassandra）。

（4）数据处理和分析：Apache Spark、Apache Flink、Apache Storm、Apache Beam、

MapReduce、Apache Kafka。

（5）数据集成和 ETL：Apache NiFi、Talend、Apache Airflow、AWS Glue、Google Cloud Dataflow。

（6）数据版本控制和协作：Git、GitHub、GitLab、Bitbucket。

（7）云平台和基础设施：Microsoft Azure、GCP、AWS。

（8）部署：Docker、Kubernetes、Flask。

1.3.4　领域专家

领域专家是具有特定行业或领域内丰富知识和专长的数据科学家。例如，那些在计算机视觉（CV）或自然语言（NL）问题上积累了丰富知识和专长的人。他们利用自己的领域知识开发定制的机器学习模型和数据分析技术，以满足其领域的独特挑战和要求。然而，也有一些非技术领域的专家，他们由于自己的专业背景，对特定行业或商业问题有着深入的了解。例如，具有数字营销背景的人可能在需要理解媒体组合建模或数据驱动归因的数据科学角色方面具有优势，而具有航空业经验的人则可能在路线优化模型方面具有优势。

由于领域专家往往拥有特定领域的专长，他们通常已经熟悉自己特定行业的工具。例如，数字营销专业人士必定对包括 Google Analytics、Adobe Analytics、HubSpot 等在内的众多 MarTech 平台有一定的经验。

这些只是数据科学内可以专攻的不同领域或风格。你不需要成为所有这些领域的专家，但需要展示在所有这些领域都有一定程度的能力和成长的意愿。通常在从事数据科学项目时，你会出于必要或热情而倾向于这些领域中的一个，其间获得的实践经验将是关键的，并增强求职者的候选资格，特别是当招聘经理寻找具有该技能集的人士时。

这些数据科学的不同风格很大程度上是一个人之前经验的结果，无论是在技术领域还是在其他领域。例如，一位软件工程师可能很适合转向机器学习或数据工程，而数据分析师可能更容易转向数据工程师或商业智能工程师。可以看到，所有数据科学的风格在技能、工具和任务上都有相当大的重叠。

根据之前的一些描述，你可能已经想象到自己在“等式”中的定位。下面花一点时间明确讨论一下通往数据科学职业的一些常见路径。

1.4　审视数据科学的职业路径

数据科学领域正在迅速发展，同时吸引了来自不同背景和学科的专业人士。这种动态

的格局催生了众多数据科学的职业道路，并为数据科学带来了独特的视角、技能和经验。本节将探索 3 种主要类型的数据科学家：传统型、领域专家型和非传统路径数据科学家。

1.4.1　传统型数据科学家

传统型数据科学家遵循了传统的教育路径进入数据科学领域。他们通常在计算机科学或数学方面拥有扎实的背景，且往往辅修另一个学科。其他常见的专业包括运筹学、统计学、物理学和工程学。这些人通常会在这些领域获得更高的学位，包括硕士学位甚至博士学位。他们严格的学术训练使他们对统计方法论、编程语言和高级算法有着深刻的理解。

传统型数据科学家对数据科学领域的基本数学和统计原理有着全面的了解。他们精通概率论、线性代数、微积分和优化技术，这些构成了许多机器学习算法和统计建模的基础。这一理论基础使他们能够理解各种方法的细微差别，并研究针对特定问题的最合适方法。

凭借计算机科学的背景，传统型数据科学家擅长使用数据科学中常用的编程语言，如 Python 和 R。他们的编程技能使他们能够操作数据、实现机器学习算法，并为特定问题开发定制解决方案。此外，他们还精通使用专门的库和框架，如 TensorFlow、PyTorch 和 scikit-learn，以加快数据科学项目的开发。

简而言之，传统型数据科学家的特点是拥有深厚的 STEM 学术背景、对统计原理有全面的理解，并精通编程和数据操作。如果你的背景是传统型的，我们建议你在求职面试中将自己定位为精通机器学习的人才。除此之外，还要进一步强调你所拥有的任何研究经验。

1.4.2　领域专家型数据科学家

领域专家型数据科学家是最初在特定行业（如市场营销、金融、医疗保健或供应链）开始其职业生涯的专业人士，后来才转向数据科学领域。凭借对自身领域的深厚理解，这些个体逐渐掌握了数据分析和编程技能，以补充他们的专业知识（例如，一名公司财务主管利用领域专业知识开发了一种机器学习算法，用于标记欺诈性交易）。领域专家拥有一种独特的能力，能够利用他们的领域知识从数据中挖掘出相关洞察，使组织能够做出推动增长和效率的数据驱动决策。

领域专家对其所处行业的复杂性和细微差别有着全面的理解，这使他们在数据驱动的项目中成为宝贵的资产。他们对行业特有的挑战、趋势和最佳实践的了解使他们能够识别关键的商业问题，并构建相关且有影响力的数据驱动解决方案。凭借丰富的领域知识和分析技能，领域专家型数据科学家擅长开发针对其行业的定制化解决方案。此外，他们能够敏锐地将商业问题转化为数据驱动的假设，并利用对行业独特的理解指导他们的分析。这

种针对性的方法使他们能够产生直接满足其行业需求和优先事项的洞察结果。

此外，领域专家精通其各自领域常用的分析工具和软件。这些专业工具可能包括行业特定的数据平台、可视化软件或机器学习框架，使他们能够高效地处理和分析其领域独有的数据。他们针对这些工具的专业知识使他们能够比缺乏行业特定知识的同行更快、更有效地提供洞察结果。

最后，领域专家型数据科学家的一个重要优势是，他们有能力将复杂的数据见解传达给行业内的非技术利益相关者。此外，他们还了解所在领域的背景和术语，因此能够以一种能与业务合作伙伴产生共鸣的方式来展示研究结果。这项技能对于推动数据驱动决策、确保其工作价值得到组织的认可和理解至关重要。

总之，如果你在面试领域拥有专业知识，我们建议你将自己定位为领域专家型数据科学家。强调对行业及其挑战的深刻理解，使你能够提供有针对性、有影响力的数据驱动型解决方案。此外，还要强调能够使用行业术语有效地传达复杂的见解。你的领域知识和数据科学技术将使你成为任何组织在其领域内的宝贵资产。

1.4.3　非传统路径数据科学家

非传统路径数据科学家是那些来自被认为非传统背景的个体，他们涉足数据科学领域。这些专业人士可能来自多样化的领域，这些领域较少关注定量任务，包括心理学、音乐甚至新闻学等领域。这种非传统背景可以为他们提供独特的视角和创造性的问题解决能力，用他们多样化的经验丰富数据科学领域。

非传统路径者拥有广泛的教育和职业背景，这使他们具备多样化的技能和知识。他们可能最初在不同领域追求职业生涯，后来才发现自己对数据科学的热爱。这种多样化的经验通常会导致更广泛的跨学科问题的解决方法，使他们能够建立联系和洞察，这些可能会被那些接受更传统训练的同行忽视。例如，非传统路径者可能会以与传统主义者或领域专家不同的方式处理机器学习和人工智能（AI）伦理问题（这是 AI 领域日益相关的话题）。通过解决人道主义问题，如灾难响应、公共卫生、食品安全和人权，他们也可能会视机器学习和 AI 为创造世界的工具，此外，AI 也可能引起土木工程师对智能城市的兴趣，或政治学专业人士在刑事司法系统中检测隐含偏见的兴趣。

凭借其非传统背景，非传统路径者为数据科学带来了独特的视角，使他们能够从不同的角度解决问题。他们的创造力和创新思维可以促进新方法、模型或可视化技术的发展，挑战现状并推动数据科学的可能边界。这种跳出框架的思考方式是宝贵的，特别是在解决复杂或新颖的问题时。

此外，凭借其独特的背景，非传统路径者非常适合与来自不同学科的专业人士合作，利用他们独特的视角解决复杂问题。他们与跨学科团队有效合作的能力可以促进创新解决方案的发展，这些解决方案结合了多个领域的强项，并推动组织的成长和成功。为了促进与不同背景者的合作，他们通常需要有效地向不同受众传达复杂的想法和洞察。非传统路径者通常理解数据科学中叙述故事的重要性，使用数据可视化和叙述传达他们的发现成果。这项技能使他们能够在技术专家和非技术利益相关者之间架起桥梁，从而促进合作。

总而言之，如果你作为非传统路径者进入数据科学领域，我们建议你在求职面试中将自己定位为一个适应性强、能够带来独特视角以促进创造性问题解决的人。此外，还应强调沟通和协作的能力。随着数据科学领域的持续扩展，其专业人士的多样性只会增加。传统型、领域专家型和非传统路径者各自都带来了独特的优势和视角。当然，这些只是数据科学专业人士的一般分类，你可能兼具所有这些特征。强调个人优势将使你能够在数据科学面试中获得最佳定位。

尽管所有这些路径都有其优势，但没有一个是完全没有障碍的。数据科学中一个常见的误解是存在一条完美路径或全面性的路径，该路径没有任何瓶颈。虽然的确某些路径比其他路径有更多优势，但它们都有需要解决的空白。虽然其中一些空白是特定于风格或路径的，但它们都有一个共同点：获得第一份数据科学工作。

1.5　解决经验瓶颈问题

那么，你想成为一名数据科学家吗？欢迎来到《饥饿游戏：数据科学版》。

虽然这听起来可能有些夸张，但对数据科学家的需求不断增长已经将面试过程变成了一个候选人各种背景和专业知识的战场。

但不要害怕，就像《饥饿游戏》一样，胜算可能会偏向你这一边。

竞争不应该吓退你进入这个领域。你已经通过阅读这本书展示了你的兴趣和承诺，随着阅读的深入，你将学会如何准备数据科学面试，无论你的背景如何。此外，我们将分享策略以填补你经验中的空白，使你成为更有竞争力的候选人。记住，你有自己的优势和弱点。你可以通过专注于你的空白领域并理解独特的技能脱颖而出。

信不信由你，候选人在他们的经验中存在空白是非常常见的。在接下来的几节中，我们将回顾两个熟悉的经验空白来源：学术和工作经验空白。除了指出这些空白区域外，我们还将提供一些建议来帮助你填补它们。

1.5.1　学术经验

求职者经验中常见的空白之一就是他们的学术背景。雇主可能更青睐拥有数据科学、计算机科学或相关领域正式学位的求职者，这使得没有传统学术背景的求职者难以脱颖而出。你可能不是工程师或程序员出身，但是对数学或计算机有所了解，但尚未深入假设检验等细节内容。不必担心。弥补学术背景空白的第一步就是找出空白。反思自己的教育和经历，并向自己提出以下问题：

（1）在数据科学的哪些领域感到最不自信？

（2）需要更多接触哪些技术或概念？

（3）在面试或项目工作中，我在哪些主题或任务上遇到最大的困难？

（4）面试的工作通常需要哪些模型？

一旦识别出了自己的空白，即可制订一个行动计划有效地解决它们。以下是几种方法，可以帮助你填补学术经验的空白，并增强数据科学候选人资格。

1. 获取相关认证

从知名组织或平台（例如，DataCamp、Codecademy、Sololearn、Alison、Udemy、Udacity、Google 认证等）获取数据科学、机器学习、人工智能或相关领域的认证。这些认证可以帮助你获得信誉，展示专业知识，并证明你对学习的承诺。

2. 参加研讨会和训练营

参加研讨会、训练营或短期课程，以获得数据科学技术和工具的实践经验。例如，Meetup.com 和 LinkedIn 是寻找本地或虚拟数据科学小组的有用网站。这不仅有助于提高技能，还可以让你与该领域的其他专业人士建立联系。

3. 利用大规模开放在线课程（MOOC）

从顶尖大学或平台报名参加 MOOC，学习数据科学的概念和技术。常见的网站包括 Coursera 和 edX。这些课程可以帮助你在该主题上打下坚实的基础，并提升非传统学术背景。

4. 建立强大的作品集

创建一个展示你的数据科学项目、编码技能和解决问题能力的作品集。强调独特视角以及非传统背景如何为数据科学方法做出贡献。

5. 与数据科学专业人士建立联系

通过社交活动、在线论坛或 LinkedIn 等社交媒体平台与数据科学领域的专业人士建立

联系。这可以帮助您深入了解行业，了解工作机会，并建立可能导致指导或工作推荐的人际关系。

相关资源（如书籍、在线课程和教程）可以帮助您获得必要的知识。为完成这些活动制定一个现实的进度表，不要被大量在线课程所淹没。在制定学习计划时，设定可实现的目标并对自己保持耐心是很重要的。记住，数据科学是一个广阔的领域，要精通它需要时间。对此，可设定专门的时间去执行学习计划。此外，还可通过论坛、社交媒体和社交活动与数据科学社区互动，向他人学习并保持动力。

1.5.2　工作经验

对于候选人来说，另一个常见的空白与工作经验有关。进入数据科学领域可能具有挑战性，特别是当面临工作经验瓶颈时。雇主通常寻求具有先前经验的候选人，这为有志于成为数据科学家的人创造了一个两难境地：你需要经验才能得到工作，但你需要工作才能获得经验。本节将探讨工作背景空白的常见原因，并提供策略帮助你克服工作经验瓶颈。

你的工作背景可能与雇主所寻找的不完全一致，这包含几种原因，例如不同领域的职业转换。你可能是一名拥有有限经验或没有全职经验的应届毕业生，或者可能因个人原因（例如，照顾他人、健康、旅行）而存在就业空白，或者可能做过自由职业或合同工，这可能不被视为稳定或相关的工作经验。

理解工作背景空白背后的原因对于构建一个引人入胜的叙述和向潜在雇主展示你的价值至关重要。以下是几种方法，可以帮助填补工作经验空白并增强数据科学候选人资格。

1. 个人项目

开发并展示个人项目，证明你的技能、创造力和解决问题的能力。选择与职业兴趣或目标行业一致的项目，这将有助于构建作品集，并展示你对该领域的激情和承诺。

2. 实习、合作、奖学金和学徒

寻求实习、合作或学徒机会，以获得实践经验并与行业中的宝贵人脉建立联系。这些机会可以为你提供入门的途径，让你能够向经验丰富的专业人士学习，并建立可以带来未来工作前景的网络。针对于此，甚至存在一些在线实习机会。例如，Forage 提供了由摩根大通、沃尔玛、KPMG、Lyft、红牛、PWC、埃森哲、德勤、通用电气等顶级公司主办的虚拟体验。许多科技公司，如微软、亚马逊和谷歌，为应届毕业生和专业人士提供许多实习机会。一些组织还提供在线奖学金，如 Correlation One 和 Insight Fellows。

3. 自由职业和咨询工作

为企业和组织提供自由职业或咨询服务，即使是无偿的。这使你能够获得实践经验，提高技能，并建立成功的记录。此外，它还证明了你与客户合作和解决现实世界问题的能力。

4. 在线竞赛和黑客松

参加数据科学竞赛和黑客松，如 Kaggle 或 DrivenData 上举办的活动。这些活动使你能够解决具有挑战性的问题，与他人合作，并向潜在雇主展示你的技能。

5. 开源贡献

为与数据科学、机器学习或人工智能相关的开源项目做出贡献。这提高了你的技术技能，并证明了你与他人合作并向更广泛的数据科学社区做出贡献的能力。

通过采用这些策略，你可以克服工作经验的瓶颈，并将自己定位为数据科学就业市场的有力候选人。记住，坚持和适应性是成功的关键。保持对目标的专注，抓住学习和成长的机会，最终，你将突破工作经验的障碍，获得理想的数据科学工作。

现在，你已经对可能遇到的瓶颈问题以及解决这些问题的方法和资源有了正确的认识，接下来让我们更好地了解期望掌握的技能和能力。在回顾了硬技能和被低估的软技能之后，你将能够认识自己的能力差距，这不仅有助于你确定要利用哪些资源，还能帮助你以更有针对性和目标导向的方式阅读本书。我们鼓励读者通读全书，但你也可以直接阅读所关注的章节。

1.6　理解预期的技能和能力

事实是这样的：面试是数据科学工作申请流程中的关键组成部分，你可以借此机会向潜在雇主展示技能、知识和个性。面试过程至关重要，原因有几点：

（1）雇主可以评估你的技术技能、解决问题的能力和批判性思维。

（2）可以展示你的沟通技巧、团队合作精神和文化契合度。

（3）它让你有机会提问并收集有关公司和角色的信息，以确保与你的职业目标和价值观一致。

（4）为面试做准备对于在竞争激烈的就业市场中脱颖而出并获取理想的职位至关重要。

为数据科学面试做准备对成功至关重要。实际上，这是可以为职业生涯做的最有用的

活动之一。这不仅适用于希望在该领域获得第一份工作的数据科学新手，也适用于希望掌握新技能和技术的经验丰富的数据科学家。在本书的后续部分中，我们将通过回顾最常见的数据科学面试主题帮助你做好准备，包括技术和案例研究问题。此外，我们还将提供解决问题的技能、编码和数据操作技巧的问题。除了这些活动，你还应该针对公司、企业文化、产品和行业趋势进行准备。此外，还可以准备一些问题向面试官提问，以展示你的兴趣和参与度。

目前，你需要知道大多数数据科学面试包括两个主要领域：技术（硬）技能和非技术（软）技能。每个领域都有不同的目的，且需要不同的准备策略。技术部分评估你在数据科学、编程、统计学和机器学习方面的知识和技能。例如，它可能包括编码练习或算法问题、数据操作和清洗任务、统计分析或假设检验问题，以及机器学习模型选择和评估问题。与此同时，非技术部分则评估沟通技巧、解决问题的能力和团队合作能力。它可能涉及你过去的经验和成就、情境或解决问题的场景，个人的优势、劣势和工作方式，以及了解你的动机和职业抱负。

掌握数据科学面试是一项关键技能，可以成就或破坏你的职业生涯。虽然我们不会赢得所有面试，但为这些面试做准备感觉就像是在准备一场马拉松。当需要准备多个面试时，这种感觉尤为强烈。进入数据科学领域的关键是在预期的技能和能力上打下坚实的基础。如果你在面试过程中表现出色，则可以给潜在雇主留下持久的印象，并增加收到工作邀请的机会。此外，充分了解面试的结构可以让你为技术和非技术部分做好准备，通过突出优势和技能，你将顺利地踏上数据科学领域的成功之路。

让我们更深入地了解数据科学家应具备的硬技能和软技能。审视之后，你将对本书学习的技能有一个更清晰的概念。

1.6.1　硬技能（技术技能）

要在数据科学角色中表现出色，你必须具备各种硬技术技能的坚实基础。这些技能使你能够有效地操作、分析和解释数据，并开发和部署机器学习模型。本节将讨论在数据科学职位中取得成功所需的基本硬技术技能。

1. 编程语言

精通编程语言对于数据操作、分析和可视化至关重要。数据科学中最流行的语言包括：

- Python：一种多功能的高级编程语言，拥有广泛的数据科学库和工具，如 NumPy、Pandas、Matplotlib 和 scikit-learn（本书稍后将介绍一些关键的 Python 技能）。

- R：一种专门为统计计算和图形学设计的编程语言，提供了广泛的数据操作、可视化和建模的软件包。

2. 数据操作和清洗

数据科学家经常处理原始的、混乱的或不完整的数据。因此，必须熟练于数据预处理、清洗、转换和组织数据，以便为分析或建模准备好数据。通常需要熟练掌握 SQL，以便从数据库中提取数据并进行清洗和准备。

3. 数据可视化

数据可视化以图形格式表示数据，有效地传达洞察和趋势。基本的数据可视化技能包括使用 Matplotlib、ggplot2 或 Tableau 等工具创建清晰且信息丰富的可视化内容，并根据数据和目标受众选择合适的可视化类型。通过视觉叙事有效地传达洞察和发现结果。

4. 统计学

扎实的统计学基础对于制订数据驱动的决策和解释结果至关重要。数据科学中的关键统计概念和技术包括描述性统计学，它使用均值、中位数、众数、方差和标准差等度量来总结和描述数据。此外，候选人必须了解推断性统计学，它使用样本数据通过假设检验和置信区间等技术对总体或关系得出结论。此外，概率论用于理解事件及其关系的可能性，包括条件概率、独立性和贝叶斯定理等概念。

5. 机器学习

机器学习涉及训练算法，进而从数据中学习并进行预测或决策。基本的机器学习技能包括：

- 监督学习（SL）：基于输入特征构建模型以预测目标变量。在数据科学面试之前，你应该了解的一些 SL 技术，包括线性回归、逻辑回归和决策树。
- 无监督学习（UL）：在没有标记目标的情况下发现数据中的模式或结构。在数据科学面试之前，重要的是要理解聚类、降维和异常检测等技术。
- 模型评估：使用准确度、精确度、召回率、F1 分数和曲线下面积（AUC）等指标评估模型性能。

6. 云计算平台

AWS、Azure 或 Google Cloud 等服务为数据存储、处理和机器学习提供可扩展资源。越来越多的组织采用这些平台，它们可能会要求你知道如何使用它们进行数据科学活动，尽管大多数服务都提供证书以证明你在使用其服务方面的熟练程度。

为了在快速发展的数据科学领域保持竞争力，持续完善和更新技能至关重要。持续学习、参加研讨会，并参加在线课程或训练营，以保持你的技术敏锐性和相关性。

1.6.2　软技能（沟通技能）

虽然硬技术技能构成了数据科学家专业知识的基础，但软技能在确保角色成功方面同样重要。软技能是非技术性的人际能力，帮助你驾驭职业关系，与团队成员合作，并有效地传达洞察结果。本节将讨论在数据科学职位中所需的基本软技能。

1. 好奇心和持续学习

成功的数据科学家必须具备好奇心和持续学习的能力。培养好奇心和持续学习包括了解行业趋势、新工具和技术。此外，还要征求同行、导师和主管的反馈意见，以确定需要改进的地方。最后，还应参与职业发展活动，如参加会议、研讨会或在线课程。

2. 沟通

有效的沟通对于数据科学家至关重要，因为它使你能够清晰、简洁地解释复杂的概念和洞察结果，以适应你的听众。同样重要的是，你需要向技术和非技术利益相关者展示你的发现和建议。

3. 团队合作和协作

数据科学家经常在多学科团队中工作，与工程师、分析师、产品经理和其他利益相关者合作。基本的团队合作和协作技能包括积极倾听和理解他人的观点、需求和想法。另外，适应性也是协作的关键，并且要根据团队动态、项目要求或目标的变化调整方法和优先级。

4. 问题解决

数据科学家必须通过将复杂、现实世界的问题分解为较小的组成部分，分析可用数据，并制定适当的解决方案以解决这些问题。关键的问题解决技能包括分析思维，即识别数据中的模式、趋势和关系，并理解问题的基本结构。

5. 时间管理和组织

有效的时间管理和组织对于管理多项任务、满足截止日期和确定工作优先级至关重要。要在这些领域表现出色，可以考虑为短期和长期项目设定明确的目标。此外，还可创建一个结构化的时间表，为不同的任务和优先事项分配时间。最后，你应该定期评估进度，根据需要调整计划，并从过去的经验中学习。

这些硬技能和软技能造就了一个全面的数据科学家，他不仅具备使用数学和计算技术来解决业务问题的能力，还擅长有效管理多个项目、交付成果、利益相关者期望和紧迫的截止日期。虽然数据科学家通常不是组织中面向客户最多的角色，但最优秀的数据科学家在拥有强大的人际技能以协作和沟通问题、需求、警告、模型如何运作以及如何解释结果时，会脱颖而出。毕竟，只有沟通得当，你的工作才会出色。

1.7　探索数据科学的演变

数据科学领域不断演变，无论是使用的工具还是工作的类型。这种演变是由技术进步、数据可用性的增加以及行业对数据驱动洞察的日益增长的需求所驱动的。因此，对于那些有兴趣进入该领域的人士来说，不仅要学习数据科学的基本知识，而且还要持续关注新的发展和技术，这是至关重要的。

1.7.1　新模型

数据科学领域演变的最显著方式之一是开发新的机器学习和人工智能算法和技术。随着人工智能的不断成熟，数据科学家能够构建更准确和强大的预测模型，这些模型可用于解决广泛的复杂问题。这包括实施从工业界和学术界等其他领域借鉴的方法，如流程改进、运筹学、博弈论、网络/图形分析和深度学习技术。

不言而喻，诸如在 ChatGPT 中使用的大型语言模型（LLM）之类的发展预计将对数据科学家的工作方式产生深远影响。例如，集成开发环境（IDE）中的 LLM 有潜力加快代码编写速度。这类似于开源软件（OSS）包的发展，后者已经提高了程序员的生产力。

1.7.2　新环境

数据科学领域演变的另一个方面是日益增长的云技术平台。虚拟化和无服务器技术使数据科学家能够访问强大的计算资源和可扩展的数据存储，使得处理大型数据集变得更加容易且成本效益更高。因此，云计算通过提供前所未有的机会，以及改变组织处理数据分析和机器学习的方式，彻底改变了数据科学领域。随着这些进步，数据科学家已经克服了传统的限制，如硬件限制、可扩展性挑战和资源分配问题。现在，数据科学家可以在单个物理服务器上创建多个虚拟机（VM），从而实现计算资源的有效利用。

例如，无服务器技术简化了模型部署和软件应用管理，因为它消除了基础设施配置的需求，并根据需求自动扩展资源。云计算平台，如亚马逊网络服务（AWS）、微软 Azure 和谷歌云平台（GCP）在基础设施即服务（IaaS）、平台即服务（PaaS）和软件即服务（SaaS）领域占据主导地位，使高性能计算、存储和专业工具的访问民主化，并赋予数据科学家巨大的计算能力。它们提供了强大的框架，如谷歌云 AI 平台和 Azure 机器学习，可以在不投资昂贵硬件的情况下对大型数据集进行复杂模型的训练。此外，基于云的数据湖，如 AWS 简单存储服务（S3）或 Azure 数据湖存储（ADLS），为大规模数据处理和分析提供了可扩展且成本效益高的存储解决方案。

总体而言，虚拟化、无服务器技术和云计算极大地扩展了数据科学的能力范围，实现了更有效、可扩展的数据分析，促进了创新，并加速了整个行业的人工智能驱动解决方案的发展。

1.7.3　新计算

计算能力的提高也将继续推动该领域的发展。随着数据集在规模和复杂性上的增长，以及人工智能算法的日益成熟，数据科学家需要更强大的计算资源来处理和分析数据。这导致了专门为数据科学设计的专用硬件和软件工具的发展，如 GPU，以及像 Hadoop 和 Spark 这样的分布式计算框架。此外，许多数据科学家现在转向像 AWS 和 Google Cloud 这样的基于云的计算平台，并按需访问可扩展的计算资源。

数据科学领域的技术进步迅速，数据科学家必须跟上计算能力的最新发展，并具备利用这些资源所需的技能和知识。

1.7.4　新应用

除了这些技术进步之外，数据科学领域还在其应用行业和应用领域中不断演变。数据科学现在被应用于医疗保健、金融、交通和物流等广泛领域。因此，数据科学家必须适应新的行业和领域，并能够应用他们的技能和技术解决新的和独特的问题。

鉴于数据科学领域的快速变化，对希望进入该领域的人士来说，跟上新的发展和技术至关重要。这需要对持续学习和职业发展做出承诺，并对新的想法和方法持开放态度。通过跟上该领域的最新进展，数据科学家可以确保他们保持竞争力，并能够为他们的组织和客户创造价值。

1.8 本章小结

在本章中，读者已经了解了现代数据科学的现状、角色所包含的内容、预期候选人所需具备的技能和能力，以及成为数据科学家的最常见的路径。此外，本章还讨论了数据科学的多样化功能，以及它如何培养出具有不同专业和背景的数据科学家的多元化劳动力。考虑到这一点，你可以确定求职路径可能是什么样子，或者希望填补哪些知识空白。

第 2 章将开始数据科学职位搜索之旅，从心理（和情感上）为你铺平前方的道路。我们将讨论一些被低估的技巧、如何识别合适的工作机会、如何发现工作机会、如何准备引人注目的申请，以及如何在不断演变的技术、项目作品集和简历的海洋中保持领先。

1.9 参考文献

[1] *Data science* from *Wikipedia*: https://en.wikipedia.org/wiki/Data_science.

[2] *Is Data Scientist Still the Sexiest Job of the 21st Century?* by *Thomas H. Davenport* and *DJ Patil*, from *Harvard Business Review*: https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century.

[3] *Data Scientists* from *U.S. Bureau of Labor Statistics*: https://www.bls.gov/ooh/ math/data-scientists.htm#tab-1.

[4] *The Digitization of the World* by *David Reinsel*, *John Gantz*, and *John Rydning*, from *International Data Corporation*: https://www.seagate.com/files/www-content/ our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.