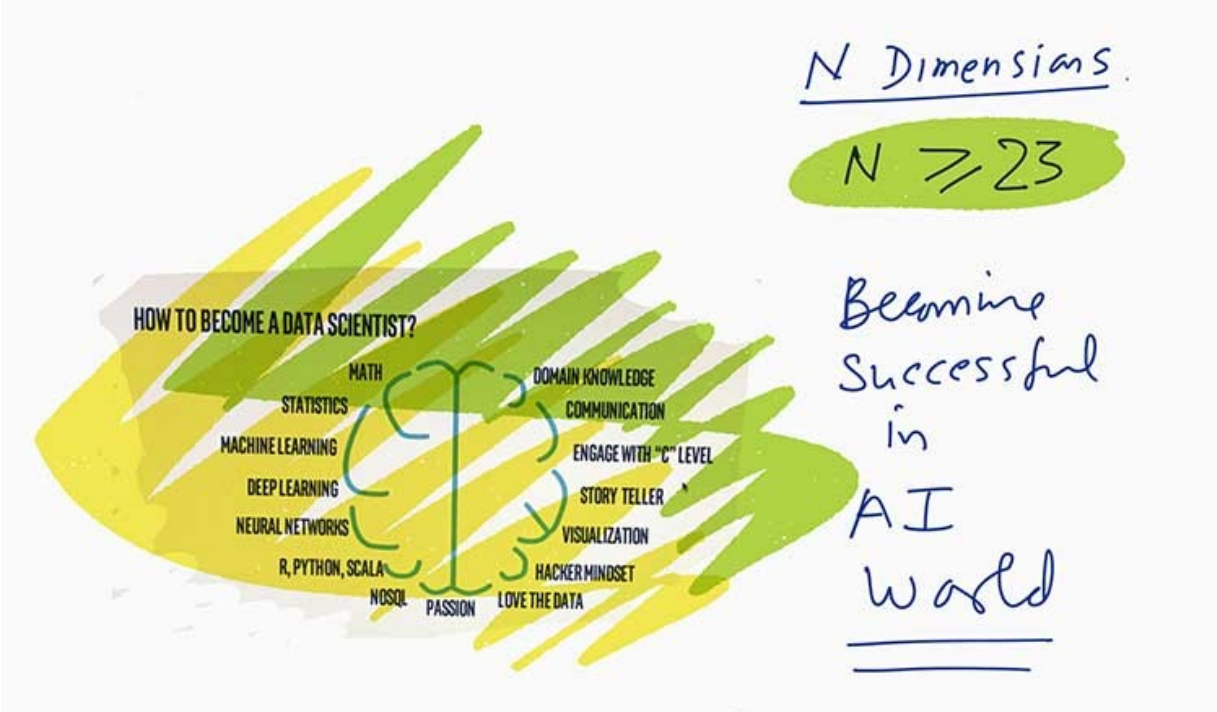
数据科学是一个信息的海洋—保持专注！

发布时间：2017 年 4 月 12 日 翻译

关于如何成为数据科学家的入门资料

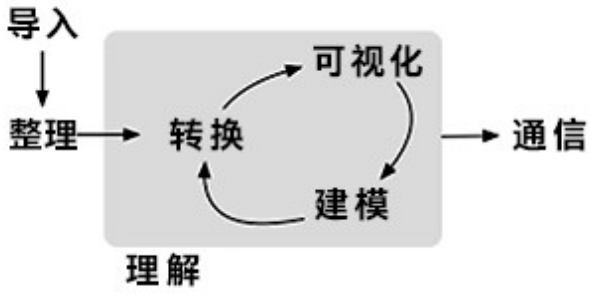
如何成为一名优秀的数据科学家？我应该学习 R* 还是 Python*？还是两者都学？我是否需要获得博士学位？我是否需要上很多数学课？我需要具备哪些软技能才能成功？项目管理经验如何？哪些技能可转移？从何处入手？



数据科学是当今科技界的热门话题。科学推动着全球许多趋势的发展，包括机器学习和人工智能。

在本文中，我们将通过一系列步骤介绍数据科学相关知识，以便对数据科学感兴趣的产品经理或业务经理能够走出迈向数据科学家的第一步或至少加深对其的了解。

第 1 步：定义问题陈述



我们都听到过这样的对话：“查看数据，告诉我你发现了什么。”当数据量比较小、数据大多为结构化数据，且数据信息有限时，这种方法可能有效。但当我们需要处理数 GB 或数 TB 的数据时，这种方法可能会带来无休止、艰巨的检测任务并得不到任何结果，因为没有问题可以着手。

科学虽然强大，但并不是魔法。任何科学领域的发明都可解决一个问题。同样，使用数据科学的第一步是定义一个问题陈述、一个需要验证的假设或一个需要回答的问题。它也可能侧重于要发现的趋势、要作出的预估和预测等。

让我们以用于监控健康和健身的移动应用 MyFitnessPal* 为例。我和我的几个朋友一年前下载了这款应用，然后几乎每天都使用一段时间。但在过去 6 个月中，我们大多数人已经完全不用了。如果我是 MyFitnessPal 的产品经理，那么我可能想要解决的一个问题是：如何提高客户参与度和保留率?

第 2 步：获取数据

如今的数据科学家需要访问多个来源的数据。这些数据可能是结构化数据，也可能是非结构化数据。我们经常获得的**原始数据**是非结构化数据和/或脏数据，需要对其进行清理和结构化处理，之后才能用于分析。大多数常见的数据源现在提供了在 R 或 Python 中导入原始数据的接口。

常见数据源包括：

- 数据库
- CSV 文件
- 社交媒体源，如 Twitter、Facebook 等（非结构化）
- JSON
- Web 擦除数据（非结构化）
- Web 分析
- 物联网驱动的传感器数据
- Hadoop*
- Spark*

- 客户访谈数据
- Excel* 分析
- 学术文献
- 政府研究文献和图书馆，如 www.data.gov
- 财务数据，例如来自 Yahoo Finance* 的财务数据

在数据科学领域，常见词汇包括：

- **观察或示例。** 这些可以被视为来自典型数据库的水平数据库记录。
- **变量、信号、特征。** 这些相当于数据库中的字段或列。变量可为定性或定量。

⇄ 观察或示例	⇄ 相当于数据库中的行。	⇄例如：Joe Allen 的客户记录。
	⇅ 变量、信号、特征	
	⇅ 相当于列	
	⇅ 例如：Joe 的身高。	

第 3 步：清理数据

几个术语用于指代数据清理，例如数据再加工、数据预处理、数据转换和数据整理。这些术语都是指准备用于数据分析的原始数据的过程。

数据科学分析中多达 70–80% 的工作涉及数据清理。
数据科学家对数据中的每个变量进行分析，以评估其是否值得作为模型中的一个特征。如果添加变量可提高模型的预测能力，则会被视为模型的预测指标。这样的一个变量被视为一个 **特征**，所有这些特性共同为模型创建一个 **特征向量**。这种分析称为 **特征工程**。

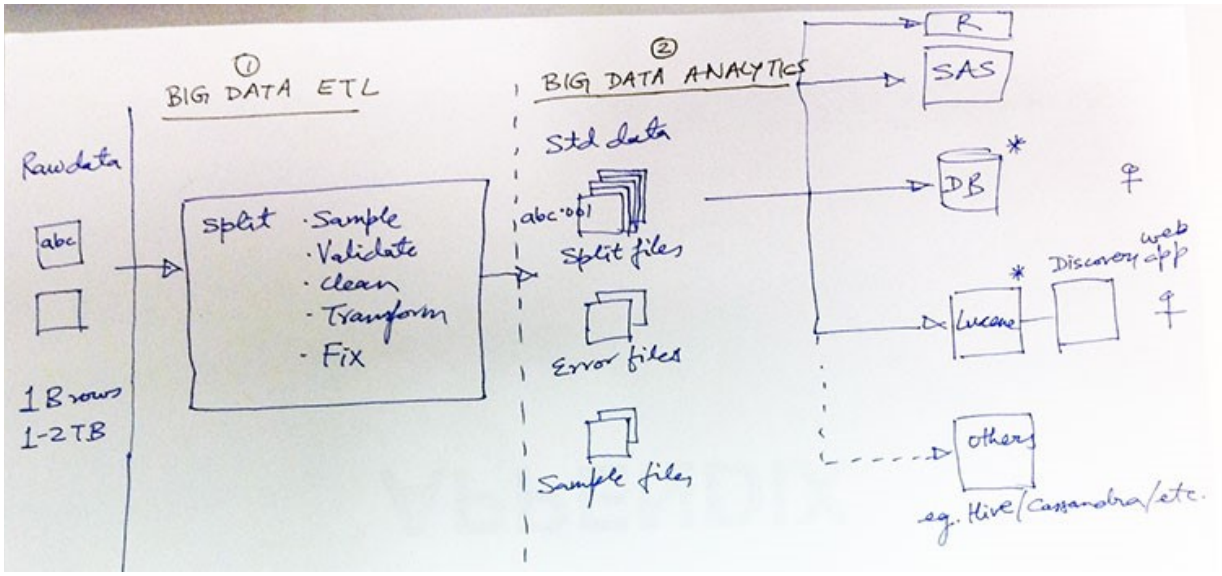
有时，变量可能需要进行清理或转换才能作为模型中的一个特征。为此我们编写脚本，也称为 **再加工脚本。** >脚本可执行一系列函数，如：

- 对变量进行重命名（有助于可读性和代码共享）
- 转换文本 (if "variable = "big" set variable = "HUGE")
- 删减数据
- 创建新变量或转置数据（例如根据生日计算年龄）
- 为现有数据补充额外数据（例如根据邮政编码获取城市和州）
- 将离散数值变量转换为连续范围（例如薪金到薪金范围；年龄到年龄范围）
- 日期和时间转换

- 将分类变量转换为多个二进制变量。例如，区域（可能的值为东、西、北、南）的分类变量可以转换为东、西、北、南这四个二进制变量，其中只有一个适用于观察。这种方法有助于在数据中创建更简单的连接。

有时，数据的数值在数量上有所不同，从而使信息难以可视化。我们可以使用 **功能缩放来解决此问题。**例如，考虑房屋的平方英尺和房间数量。如果我们使房屋的平方英尺数与卧室数量相似，我们的分析将变得更加轻松。

一系列脚本以迭代的方式应用于数据，直到我们获得足够有效的数据进行分析。为了连续提供用于分析的数据，需要利用新的原始数据重新运行一系列数据再加工脚本。**数据管道** 是指适用于原始数据的一系列处理步骤，用于确保其做好分析准备。



第 4 步：数据分析与模型选择

现在我们有了有效数据，可以进行分析了。我们的下一个目标是使用统计建模、可视化、以发现为导向的数据分析等功能熟悉数据。

对于简单问题，我们可以使用平均数、中等值、模式、最小值、最大值、平均值、范围、四分位数等进行 **简单的统计分析**。

监督式学习

我们还可以使用**监督式学习**，其数据集让我们能够访问一组特定特征变量（独立变量）的响应变量（从属变量）的实际值。例如，我们可以根据已离职员工的任职时间、资历和职位从实际数据中寻找趋势 (resigned=true)，然后利用这些趋势预测其他员工是否也会辞职。或者我们

可以使用历史数据来关联访客数量（独立变量或预测指标）与生成的收入（因变量或响应变量）之间的趋势。然后这种关联可用于根据访客数量预测网站的未来收入。

监督式学习的关键要求是实际数值的可用性以及一个需要回答的明确问题。例如：这名员工是否会离职？预计获得多少收入？数据科学家将其称为 "*为现有数据标记响应变量。*"

回归是一款用于监督式学习的常用工具。单因素回归使用一个变量；多因素回归使用多个变量。
线性回归假设因子与响应变量之间的未知关系是线性关系 Y = a + bx，其中 b 是 x 的 **系数。**

现有数据的一部分被用作 **训练数据**，以计算这一系数的数值。数据科学家通常使用 60%、80% 或 90%（偶尔）的数据进行训练。一旦为 **训练模型** 计算系数的数值，则将利用剩余数据（也称为**测试数据**）进行测试，以预测响应变量的数值。预测响应值与实际值之间的差异是称为 **测试误差指标的有效指标。**

我们在数据科学建模方面的探索是为了 **最大限度降低测试误差指标**，从而通过以下方式提高模型的预测能力：

- 选择有效的因子变量
- 编写有效的数据再加工脚本
- 选择合适的统计算法
- 选择所需的测试和训练数据量

非监督式学习

当我们尝试学习基础数据本身的结构时，就采用非监督式学习。它没有响应变量。数据集未标记，原有洞察不明确。我们什么都不清楚，所以不打算预测什么！
这种方法对探索性分析有效，可用于回答下列问题

- 分组。我们有多少类客户群？
- 异常检测。这种情况正常吗？

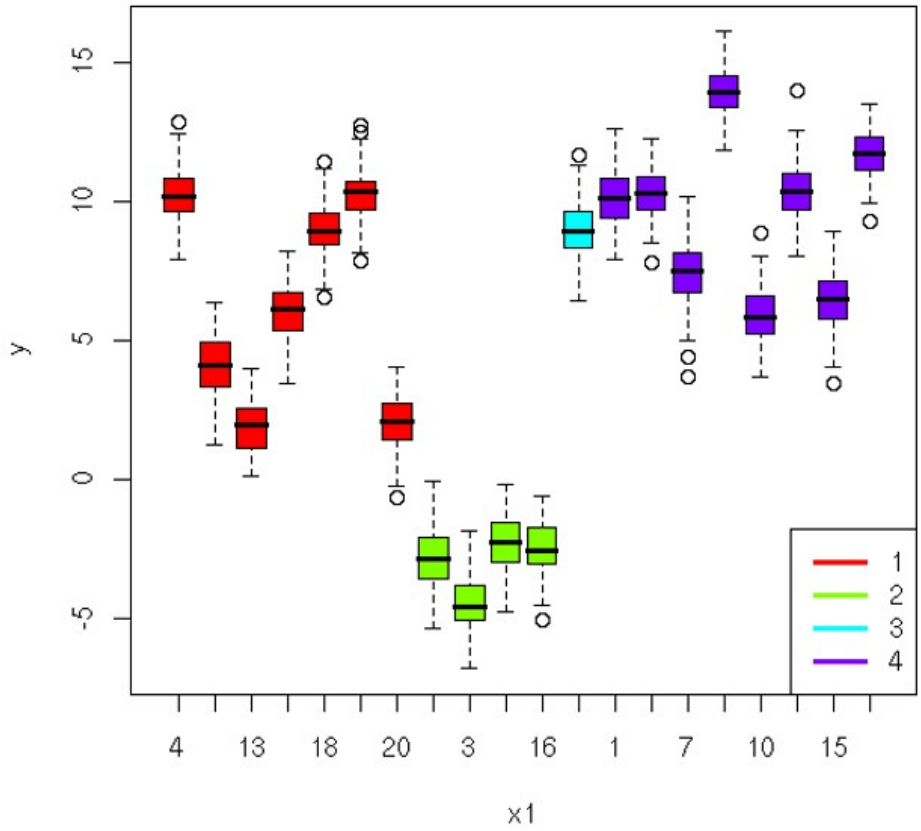
方差分析 (ANOVA)是一种常用技术，用于比较两个组或多个组的平均数。它被称为 ANOVA，因为“方差估计”是计算的主要中间统计数据。使用各种距离度量比较不同组的平均数，其中**欧氏距离**是一种常用方法。

方差分析用于将观察值分为类似的组，也称为 **集群**。可以根据相应的预测指标将观察值 **分类** 到这些集群中。
http://www.statsdirect.com/help/content/analysis_of_variance/anova.htm
(http://www.statsdirect.com/help/content/analysis_of_variance/anova.htm)

两个常见的聚类应用为：

- **分层聚类。** 一种自下而上的方法。我们从单独的观察值开始，并将它们与最接近的观察值进行合并。然后，我们计算这些分组观察值的平均数，并将这些组与最接近的平均数进行合并。我们将重复进行这一工作，直到形成更大的组。距离度量是提前定义的。这种方法很复杂，不适用于高维数据集。

18/5000 分层方差分析



- **K 均值聚类。**使用分区方法。
  - 根据我们的直觉，我们假设数据具有**固定数量的集群。**
  - 我们还假设了 **每个集群的起始中心。**
  - 然后将每个观察值分配给具有最接近观察值的平均值的群集
  - 重复这一步骤，直到所有观察值都已分配给集群。
  - 现在，我们根据分配给集群的所有观察值的平均值重新计算集群的平均值。
  - 将观察值重新分类到这些新集群中，并重复步骤 c、d 和 e，直到它们达到稳定状态。

如果未达到稳定状态，我们可能需要对开始时假设的集群数量（即 K）进行改进，或使用另一个距离度量。

第 5 步：可视化和有效通信

最终的集群可实现可视化，以便使用 Tableau* 或图形库等工具进行简单通信。

数据科学实践者的提示

在我了解数据科学的过程中，我遇到了在 Facebook、eBay、LinkedIn、Uber 和一些咨询企业等公司工作的实践者，这些公司正在有效地利用数据。这是我获得的一些重要建议：

- **了解您的数据。** 充分了解数据及其背后的假设非常重要。否则数据可能无效，这可能导致错误的答案、解决错误的问题，或两者兼有。
- **了解领域和问题。**数据科学家必须深入了解业务领域和要解决的问题，以便从数据中提取适当的洞察。
- **道德规范。** 不要为假设而牺牲数据质量。**问题往往不是无知，而是我们先入为主的观念！**
- 较大的数据集总能够提供更出色的洞察，这是一个谬论。虽然较大的数据量在统计上有显著意义，但大数据集也会带来更大的噪声。经常可以看到，较大数据集的 R 平方值小于较小数据集的 R 平方值。
- 虽然数据科学本身不是一个产品，但它可以为解决复杂问题的出色产品提供支持。有效沟通的产品经理和数据科学家可成为强大的合作伙伴：
    - 产品经理首先介绍要解决的业务问题、要解答的问题以及要发现和/或定义的限制。
    - 数据科学家在机器学习和数学方面拥有丰富的专业知识，专注于业务问题的理论方面。现代数据集用于执行数据分析、转换、模型选择和验证，以建立应用于 **业务问题的理论基础。**
    - 软件工程师致力于实施理论和解决方案。他或她需要对机器学习（Hadoop 集群、数据存储硬件和编写生产代码等）的机制有很强的理解。
- 学习编程语言。Python 最容易学习; R 被视为最强大的语言。

常用的数据科学工具

R

R 是一款许多数据科学家喜欢的工具，在学术界拥有特殊的地位。这款工具从数学家和统计学家的角度处理数据科学问题。R 是一种丰富的开源语言，约有 9,000 个额外软件包。用于在

R 中编程的工具称为 R Studio*。尽管 R 在企业中的采用率一直在稳步增长并且其普及在一定程度上归功于其丰富、强大的基于正则表达式的算法，但 R 学习起来比较复杂。

Python

Python 正在逐渐成为数据科学界使用最广泛的语言。像 R 一样，它也是一种开源语言，主要由软件工程师使用，这些工程师将数据科学视为一款使用数据解决面向客户的真正业务问题的工具。Python 学习起来比 R 简单，因为 Python 强调可读性和生产力。Python 也更加灵活和简单。

SQL

SQL 是用于与数据库交互的基本语言，是所有工具所必需的语言。

其他工具

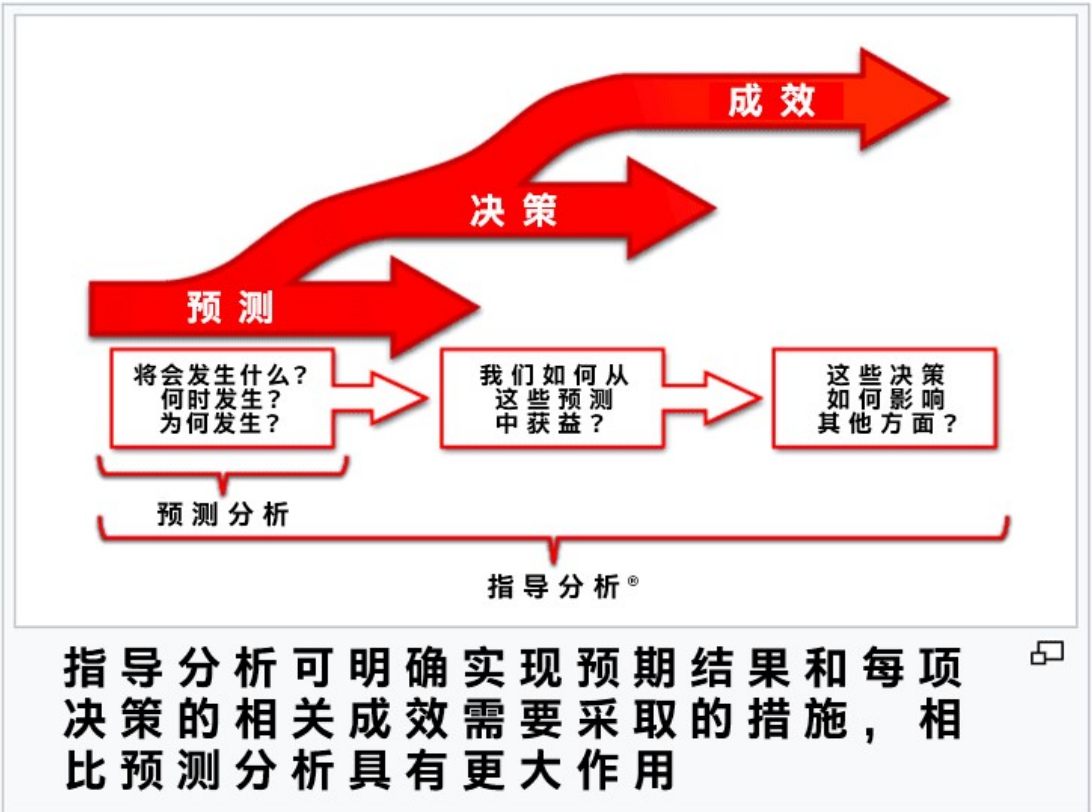
- Apache Spark 提供 **Scala***
- **MATLAB*** 是一个学术界长期使用的数学环境。它提供一个名为 Octave* 的开源版本
- Java* 用于 Hadoop 环境

需要哪些软技能？

下文将介绍您需要具备的重要软技能，您可能已经具备其中许多技能。

- **沟通。** 数据科学家无法独自坐在办公室里对 Python 程序进行编码。数据科学需要您与团队进行合作。您需要与高管、产品负责人、产品经理、开发人员、大数据工程师和 NoSQL *专家等各种人员进行沟通并建立融洽的关系。您的目标是了解他们正在试图构建什么，以及数据科学和机器学习能够起到怎样的帮助作用。
- **指导。** 作为一名数据科学家，您需要拥有出色的指导技能。您并不是公司的独立贡献者；您是首席执行官的最好伙伴，可帮助其塑造公司，即基于数据科学塑造产品和领域。例如，根据您的数据科学结果，您向管理团队提供透视分析结果，建议公司在巴西推出深绿色的鞋子; 如果这款产品投放在美国硅谷，则必将失败。您的发现可以为公司

节省数百万美元。



- **出色的故事叙述者。** 好的数据科学家是一个很好的故事叙述者。在开展数据科学项目的过程中，您将拥有大量的数据、理论和结果。有时您会感到自己已经迷失在数据的海洋之中。如果发生这种情况，请退一步思考：我们想实现什么目标？例如，如果您的受众是首席执行官和首席运营官，他们可能需要根据您的演示在几分钟内作出决定。他们不关心您的 ROC 曲线，也不会浏览多达 4 TB 的数据和 3,000 行的 Python 代码。您的目标是根据可靠的预测算法和准确的结果为他们提供直接建议。我们建议您创建 4 到 5 张幻灯片，在幻灯片中利用可靠的数据和研究清楚地讲述这一故事。**可视化。** 良好的数据科学家需要使用可视化来传达结果和建议。您不能让别人读一份长达 200 页的报告。您需要使用图片、图像、图表和图形来呈现。

- **思维模式。**一个好的数据科学家具有“黑客”思维（这里的“黑客”是褒义）并且不断寻找数据集中的模式。
- **喜欢数据。**您需要使用数据并从中挖掘价值。当然，您可以使用很多工具来更全面地了解数据，但只需大致浏览即可获取很多信息。

我能成为什么？

现在是时候决定了。我应该成为什么类型的数据科学家？

- **了解管道。** 您需要从某个地方入手。您可以在数据科学项目中担任 Python 开发人员。您可以收集日志、传感器、CSV 文件等来源的输入数据。您可以编写脚本来使用和导入传入数据。数据既可以是静态数据，也可以是动态数据。您可能会决定成为一名使用 Hadoop 或 Hive* 等技术的大数据工程师。或者，您也可能会决定成为一名机器学习算法专家，即掌握相关技能并了解哪种算法最适用于哪类问题的人。您可能是一个数学天才，自由使用开箱即用的机器学习算法并根据自己的需求进行修改。您可能会成为一名数据持久性专家。您可以使用 SQL 或 NoSQL 技术来持久存储/提供数据。或者，您可能成为数据可视化专家，使用 Tableau 等工具构建仪表盘和数据案例。因此再检查一下上面的管道：**从导入到可视化。**创建机会清单。如 "D3 expert, Python script expert, Spark master" 等。
- **Check out the job market.**看看各种工作门户，了解当前需求。有多少工作可供选择？什么工作的需求最高？薪资结构是什么？大致浏览一下海湾地区有前途的数据科学工作 ►

- **了解自己。** 您已经了解了可以获得的管道和工作类型。现在该思考一下自己和自己的技能了。您最喜欢什么，您有什么经验？您是否喜欢项目管理？数据库呢？想想自己以前的成功案例。你喜欢编写复杂的脚本来关联和操作数据吗？您是否从事可视化工作，擅长创建引人注目的演示？创建一个“喜欢从事工作的清单”。 例如“喜欢编码、喜欢脚本、喜欢 Python”。
- **创建匹配。** 根据您的喜欢从事工作清单创建机会清单并继续开展计划。数据科学是一个信息的海洋。保持专注！！

◦ **硬件**

- 英特尔® AI DevCloud
- 英特尔® FPGA
- 英特尔® Movidius™ 视觉处理器
- 英特尔® 至强® 可扩展处理器

◦ **高校**

- 成为学生大使
- 教授的课堂资源
- 大学俱乐部赞助

◦ **连接**

- 开发人员之网项目
- 论坛
- 新闻简报

◦ **相关内容**

- 英特尔® AI Builders
- 人工智能研究项目
- AI4Good
- 英特尔® AI News
- 英特尔® AI DevCon

获取新闻简报

关注我们：