What is Data Science?

数据科学是一个过程，而非结果 它是一种通过使用数据来理解不同的事物 和整个世界的过程 对于我来说，数据科学是当你有一个问题的解决方案或是模型 然后你可以用你的数据试着去证实这个模型或猜想 数据科学是一种 可以揭开数据背后的特征或趋势的一种艺术 你可以把数据翻译成一个故事 然后通过讲故事来揭示数据内在的涵义 通过这些特征 你可以为公司做出一些战略性的决定，或者设立一些重要的制度 数据科学是一个通过提取格式化与非格式化的的数据 来建立一个过程或系统的领域 数据科学就是对数据进行研究的科学 就像生物科学是研究生物的科学 物理学，是对物理相关反应进行研究的科学 数据是真实存在的，有它自己的特征 如果我们想通过它获得什么，我们就要研究它- 数据科学包含数据和一些特征 它的名字的由来是 在80-90年代，当一些教授在研究统计学课程时 他们认为把统计学课程叫做数据科学更加合适 但是什么是数据科学呢？ 我更愿意把数据科学看作一个人尝试去研究数据 去寻找他们正在研究的问题的答案 简而言之，数据科学更倾向于数据而非科学 如果你有数据，你有对数据的好奇心 你努力去研究这些数据 你熟练的操纵它，你观察它 你努力的去分析数据 试着从中找到一些特征来，这便是数据科学 数据之间是相互关联的，因为我们有许许多多可用的数据 我们过去常常为缺少数据担忧 现在我们的数据如洪水一般增加 过去，我们没有计算机程序，而现在有了 过去，软件是非常昂贵的 现在，他们开源并且免费 过去，我们不能储存许多数据 现在，只要付出微小的代价 我们就能拥有非常多的数据 所以，我们研究数据的工具 数据的多样性 储存和分析数据的能力 都很便宜和易得 数据无处不在，它就在这里 成为数据科学家，这是最好的时代

Fundamentals of Data Science

您询问的每个人都会对数据科学的内容略有不同的描述 ，但大多数人都认为数据科学具有重要的数据分析组成部分。 数据分析并不 新鲜。 最新消息是来自各种来源的大量数据：从 日志文件、电子邮件、社交媒体、销售数据、患者信息文件、运动表现 数据、传感器数据、安全摄像头等等。 与此同时， 可用的数据比以往任何时候都多，我们拥有进行有用的 分析和揭示新知识所需的计算能力。 数据科学可以帮助企业了解 其环境、分析现有问题并揭示以前隐藏的机会。 数据科学家使用数据分析通过调查 数据来增加组织的知识，探索如何最好地利用数据分析为企业提供价值。 那么，数据科学的过程是什么？ 许多组织都会使用数据科学专注 于特定问题，因此澄清组织 希望回答的问题至关重要。 这第一个也是最关键的步骤定义了数据科学项目的 进展方式。 优秀的数据科学家是好奇的人，他们提出问题以澄清业务 需求。 接下来的问题是：“我们需要什么数据 来解决问题，以及这些数据从哪里来？”。 数据科学家可以分析来自 许多来源的结构化和非结构化数据，根据 问题的性质，他们可以选择以不同的方式分析数据。 使用多个模型来 探索数据可以揭示模式和异常值；有时，这将确认组织 怀疑什么，但有时这将是全新的知识，导致组织采用 新的方法。 当数据揭示其洞察力时， 数据科学家的角色就变成了讲述故事的人，向 项目利益相关者传达结果。 数据科学家可以使用强大的数据可视化工具 帮助利益相关者了解结果的性质以及建议 采取的行动。 数据科学正在改变我们的工作方式； 它正在改变我们使用数据的方式，并改变组织理解 世界的方式。

The Many Paths to Data Science

当我长大的时候，数据科学并不存在。 这不是我醒来后 说的，我想成为一名数据科学家，当我长大后。 不，它不存在。 我不知道我会从事数据科学工作。 当我长大的时候，没有那个叫做数据科学的领域。 我认为这是真正的新的。 数据科学直到 2009 年，2011 年才存在。 像 DJ 帕蒂尔或安德鲁·盖尔曼这样的人创造了这个词。 在此之前，有统计数据。 而 且我不想成为那些人。 我想做生意 然后我发现数据科学更有趣。 我研究了统计数据，这就是我开始的。 我经历了许多不同的阶段在我的生活，我想成为一名歌手， 然后成为一名医生。 然后我意识到我擅长数学。 所以我选择了一个专注于定量分析的领域。 从那时起，我确实认为我想要处理数据。 不一定是今天所知的数据科学。 我第一次接触数据科学， 当我是机械工程的第一年。 而战略咨询公司，他们利用数据科学来做出决策。 所以这是我第一次接触数据科学。 我有一个复杂的问题，我需要解决，当 时我们通常的技术无法解决这个问题。 我在最糟糕的时候毕业 了数学学位，就在经济危机之后，你实际上必须是有用的才能找到一份工作。 所以我去了，并获得了统计学学位。 然后，我做的工作很多，被称为数据科学家， 我突然成为了一个。 我的本科学位是商业专业，我主修政治 、哲学和经济学。 然后我在 纽约大学斯特恩商学院攻读了商业分析硕士学位。 当我离开本科毕业时，我加入的第一家公司，事实证 明，他们正在分析零售制造商的电子销售点数据。 我们正在做的是数据科学。 但是我们才真正开始使用这个词。 事实上，我会说四五年前是我们开始称之为分析和 数据科学的时候。 我在加拿大这里有几个实习选择。 其中一个选择是使用数据科学。 我曾经从事项目开发工作。 但我认为这是一个不错的选择。 然后我开始了我的数据科学实习。 我是一名土木工程师通过培训，所以所有的工程师都使用数据。 我会说， 在我的生活中，数据科学的传统应用始于交通研究。 我开始构建大型模型，试图预测街道交通情况，试图 确定拥堵和温室气体排放量或排气管排放量。 所以我认为这就是我的开始。 当 我在 多伦多大学的研究生时，我开始建造这些模型。 开始使用非常大的数据集， 从 50 万次旅行中查看 15 万个家庭的家庭样本。 而 且，我也是从 90 年代中期开始讲，当时这 应该是一个非常庞大的数据集，但在今天的情况下并非如此。 但我就是这样开始的 我继续使用它。 然后我搬到麦吉尔大学那里我是交通 工程教授 我建立了更大的数据模型，涉及数据和分析。 所以我会说，是的，交通研究把我带到了数据科学。

Advice for New Data Scientists

我对一个有抱负的数据科学家的建议是要好奇， 极具争论性和判断性。 好奇心是绝对必须的。 如果你不好奇，你不会知道如何处理数据。 判断，因为如果你没有 先入为主的概念有关的事情，你不会知道从哪里开始。 争论性的，因为如果你可以争辩，如果你可以为一个案件辩护， 至少你可以从某个地方开始，然后你从数据中学习，然后 你修改你的假设和假设，你的数据会帮助你学习。 你可能会从错误的点开始 你可能会说，我以为我相信这一点， 但现在有数据我知道这一点。 所以，这可以让你一个学习过程。 所以，好奇心能够采取一个立场， 强大的位置，然后向前推进。 数据科学家需要的另 一件事是分析平台的舒适性和灵活性：一些软件， 一些计算平台，但这是次要的。 最重要的是好奇心和采取立场的能力。 一旦你完成了这项工作，一旦你分析了， 那么你就得到了一些答案。 这是数据科学家最不需要的东西 ，这就是讲述故事的能力。 一旦你有了你的分析， 一旦你有了你的表格， 现在你应该能够从中讲述一个伟大的故事。 因为如果你不告诉一个伟大的故事从它， 你的发现将保持隐藏， 保持埋葬，没有人会知道。 但是你的突出地位几乎依赖于你讲述伟大故事的能力。 一个起点是了解您的竞争优势。 您是否希望成为任何领域或特定领域的数据科学家？ 因为，假设您希望成为数据科学家，并在 IT 公司或基于 Web 的公司或基于互联网的公司工作， 那么您需要一套不同的技能。 如果你想成为医疗行业的数据科学家， 那么你需要不同的技能。 因此，首先弄清楚你感兴趣 的是什么，你的竞争优势是什么。 您的竞争优势不一定是您的分析技能。 你的竞争优势是你对 生活的某些方面的理解，你超越了其他方面的理解。 也许是电影，也许是零售， 也许是健康，也许是电脑。 一旦你知道你的专业知识在哪里， 你就开始获得分析技能。 要学习哪些平台和那 些平台，这些工具将特定于您感兴趣的行业。 然后，一旦你对这些工具有了一定的熟练程度， 接下来就是将你的技能应用到真正的问题上， 然后告诉世界其他地方你可以用它来做什么。

A day in the Life of a Data Scientist

Old problems, new problems, Data Science solutions

组织可以利用现在可用于组织的几乎无限量的数据

越来越多的方法。

但是，所有组织出于相同的原因最终都使用数据科学来发现

针对现有问题的最佳解决方案。

让我们看一下数据科学的三个示例，它们为

老问题。

在运输过程中，Uber收集实时用户数据以发现有多少可用的驱动程序，

是否需要更多，以及是否应允许大笔费用吸引更多的驾驶员。

Uber使用数据在正确的时间将正确数量的驱动程序放置在正确的位置，

费用是骑手愿意支付的。

在与运输相关的数据科学方面的另一项努力中，多伦多运输委员会

在解决交通流量的旧问题，重组这些流量方面取得了长足的进步

在城市及其周围。

使用数据科学工具和分析，他们可以：

收集数据以更好地了解电车的运行，并确定需要干预的地方

分析的客户投诉数据使用探查数据可以更好地了解流量

主要路线上的表现创建了一个团队以更好地利用大型

规划，运营和评估数据

通过专注于高峰时间通关并确定最拥挤的路线，每月损失的小时数

因交通拥堵导致的通勤时间从4.75小时下降。 在2010年到3个小时。 在2014年中。

在面对我们环境中的问题时，数据科学也可以发挥积极作用。

淡水湖泊满足人类和生态的各种需求，例如提供饮用水

水和生产食物。

但是世界各地的湖泊受到有害蓝藻细菌发病率上升的威胁

绽放。

解决这一长期存在的难题的项目和研究很多。

在美国，从缅因州到南卡罗来纳州的研究中心的科学家团队

正在开发和部署高科技工具来探索整个湖泊中的蓝细菌

东海岸。

该小组正在使用机器人船，浮标和配备摄像头的无人机来测量物理，

检测到蓝细菌的湖泊中的化学和生物学数据，收集了大量

与湖泊和有害水华发生有关的大量数据。

该项目还在建立新的算法模型来评估发现。

收集的信息将有助于更好地预测何时何地蓝藻细菌

盛开，使采取积极措施保护休闲湖中的公共健康

在那些提供饮用水的地方。

这样的跨学科培训为下一代科学家应对社会问题做好了准备

适当的现代化数据科学工具带来的问题。

它需要收集大量数据，清理和准备数据，然后对其进行分析以

获得为当今企业开发更好的解决方案所需的洞察力。

您如何获得有效的更好解决方案？

您必须：找出问题并建立清晰的

了解它。

收集数据进行分析。

确定要使用的正确工具。

制定数据策略。

案例研究也有助于定制潜在的解决方案。

这些条件存在并提取可用数据后，您可以开发一台机器

学习模型。

组织需要花费一些时间来使用数据优化数据策略的最佳实践

科学，但好处是值得的。

Data Science Topics and Algorithms

我真的很喜欢回归, 我想说回归可能是

第一个能够帮助我真正的了解数据的概念 所以我非常喜欢回归

我很喜欢数据可视化 我认为它是

将信息传达给人们的关键因素 这些人通常不太了解数据科学

人工神经网络

我真的对神经网络充满热情 因为我们有可从大自然学到很多知识

所以当我们试图模拟我们的大脑 我认为我们可以

通过算法中的这种生物行为开发一些应用

我十分喜欢基于R的数据可视化 最近邻 (算法) 这是最简单的

但是通常能够得到最好的结果 而不像一些过度的夸张算法

这种算法很可能过度拟合 来使其实现好的拟合

所以结构化数据更像是你熟悉的

微软Excel格式的表格式数据 你会得到行和列 这称为

结构化数据 非结构化数据基本上主要是来自

网络的数据 它不是表格 它不是以行和列的形式呈现

它是文本 有时它是视频和音频 所以你必须运用

更复杂的算法来提取数据 事实上很多时候我们花费大量时间和精力

将非结构数据结构化

然后对其进行分析 所以如果你有一些

适合表格或是行和列的内容 那么这就是结构化数据

但是如果你看到的是一个博客 或者如果你想从网页中获取信息

你将会有大量的网页

这些是非结构化数据 就需要更多精力才能从中获取信息

让我用最简单的术语来解释回归

如果你乘坐过计程车或出租车 你就会明白回归

其原理是这样的 你乘坐在出租车里的那一刻

你会看到有一个固定的金额 $2.50 你可以让车辆行驶

或者你也可以下车 这就是你在踏入出租车的那一刻你欠驾驶员的金额

如果你已经踏入出租车 你就必须支付这笔固定的款项

然后当它开始行驶每米或每百米 费用会以一定数额增加

所以在距离和你支付的超过起步价的那部分费用

的两者之间有一个关系

如果出租车没有行驶 你遇到交通堵塞 每增加一分钟

你需要支付更多 所以费用随时间增加而增加

行驶的距离增加也会使费用增加 当这些发生的时候

你已经支付了一个固定的起步价 这就是回归

回归给以告诉我们起步价是多少 以及时间和你所付费用的之间关系

以及行驶距离和费用间的关系

因为当不知道这些关系的时候

如果知道行驶的距离和费用 回归可以使我们

计算出起步价$2.50 它也将算出

费用和距离的关系以及费用和时间的关系

这就是回归

Cloud for Data Science

云是数据科学家的天赐之物。

主要是因为您能够获取数据，

获取您的信息并将其放入云端，

将其放在中央存储系统中。

它可以让您绕过

身体上的限制

您所使用的计算机和系统

使用，它允许您部署

分析和存储能力

不一定需要

是您的机器还是公司的机器。

云使您不仅可以存储大量

加利福尼亚或内华达州某处的服务器上的数据，

但它也允许您部署

非常先进的计算算法和

做事的能力

高性能计算

使用非您的机器。

随便想想

一些信息，您无法存储，

所以您将其发送到存储空间

我们称之为云

以及您需要使用的算法

你没有他们。

但是在云端

您有可用的算法。

您要做的是将这些算法部署在

非常大的数据集

即使您拥有自己的系统，也可以做到这一点，

你自己的机器，你自己的计算环境

不允许您这样做。

云是美丽的。

云的另一件事是

美丽的是，它允许

多个实体可以工作

同时具有相同的数据。

您可以使用相同的数据

你的同事在说

德国和印度的另一支球队

还有加纳的另一支队伍

他们正在集体工作，

他们之所以能够这样做，是因为这些信息，

以及算法和工具，

以及答案和结果，

无论需要什么，都可以在中心位置找到，

我们称之为云。 云是美丽的。

使用云使您可以立即访问打开

诸如Apache Spark的源技术

无需在本地安装和配置它们。

使用云还可以为您提供

使用最新的工具和

图书馆无后顾之忧

维护它们并确保它们是最新的。

可从以下位置访问云

在每个时区的任何地方。

您可以使用基于云的技术

从您的笔记本电脑，平板电脑，

甚至从手机上

比以往任何时候都更容易实现协作。

多个合作者或团队

可以同时访问数据

共同制定解决方案。

一些大型科技公司提供云平台，

让您熟悉

预构建环境中基于云的技术。

IBM提供IBM Cloud，

亚马逊提供亚马逊网络服务或AWS，

Google提供了Google Cloud平台。

IBM还提供了技能网络实验室或SN实验室

给在以下任何一个地方注册的学习者

IBM Developer Skills Network上的学习门户，

您可以在其中使用工具

如Jupyter Notebooks和Spark

集群，因此您可以创建

您自己的数据科学项目并开发解决方案。

通过实践和熟悉，

您会发现云如何显着地

提高数据科学家的生产力。