互联网 数据挖掘 数据分析 数据科学家 大数据 修改

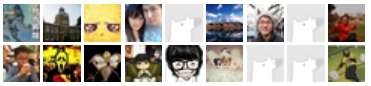
如何成为一名数据科学家？ 修改

我自己粗浅的理解为需要以下几个方面：
1. 业务知识
2. 数理统计和数据分析
3. 计算机相关知识
3.1 数据处理与收集（ETL？）
3.2 机器学习和数据挖掘
这几方面完全是自己的一个猜测，恳请大牛们不惜赐教！ 修改

4 条评论 分享 • 邀请回答 举报

关注

4799 人关注该问题



13 个回答 按票数排序

1102

Han Hsiao，一颗透明的心灵和会流泪的眼睛
罗素、冯聪、rock lee 等人赞同

-
版本更新，2014年5月14日更新一些内容。
-
如果展开讲，这个问题可以写一篇综述了。最近刚好有空，打算认真写写。

仅仅在几年前，数据科学家还不是一个正式确定的职业，然而一眨眼的工夫，这个职业就已经被誉为“今后十年IT行业最重要的人才”了。

一、数据科学家的起源
"数据科学"（DataScience）起初叫"datalogy "。最初在1966年由**Peter Naur**提出，用来代替"计算机科学"（丹麦人，2005年图灵奖得主，丹麦的计算机学会的正式名称就叫Danish Society of Datalogy，他是这个学会的第一任主席。Algol 60是许多后来的程序设计语言，包括今天那些必不可少的软件工程工具的原型。图灵奖被认为是“计算科学界的诺贝尔奖”。）

1996年，International Federation of Classification Societies (IFCS)国际会议召开。数据科学一词首次出现在会议（Data Science, classification, and related methods）标题里。

1998年，C.F. Jeff Wu做出题为“统计学=数据科学吗？ 的演讲，建议统计改名数据的科学统计数据的科学家。（吴教授于1987年获得COPSS奖，2000年在台湾被选为中研院院士，2004年作为第一位统计学者当选美国国家工程院院士，也是第一位华人统计学者获此殊荣。）

2002年，国际科学理事会：数据委员会科学和技术（CODATA）开始出版数据科学杂志。

2003年，美国哥伦比亚大学开始发布数据科学杂志，主要内容涵盖统计方法和定量研究中的应用。

2005年，美国国家科学委员会发表了"Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century"，其中给出数据科学家的定义：

> "the information and computer scientists, database and software and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection"

信息科学与计算机科学家，数据库和软件工程师，领域专家，策展人和标注专家，图书管理员，档案员等数字数据管理收集者都以可成为数据科学家。它们主要任务是："进行富有创造性的查询和分析。"

2012年，O'Reilly媒体的创始人 Tim O'Reilly 列出了世界上排名前7位的数据科学家。

- Larry Page，谷歌CEO。
- Jeff Hammerbacher，Cloudera的首席科学家和DJ Patil，Greylock风险投资公司企业家。
- Sebastian Thrun，斯坦福大学教授和Peter Norvig，谷歌数据科学家。
- Elizabeth Warren，Massachusetts州美国参议院候选人。
- Todd Park，人类健康服务部门首席技术官。
- Sandy Pentland，麻省理工学院教授。
- Hod Lipson and Michael Schmidt，康奈尔大学计算机科学家。

相关问题

互联网公司的数据科学家（Data Scientist）职责和日常工作内容什么？ 4 个回答

什么是好的数据科学家？ 5 个回答

如何成为一名数据科学家？ 13 个回答

搜索引擎工业界做搜索质量评估的一般流程是怎样的呢？ 0 个回答

分享问题

微博 站内私信

问题状态

最近活动于 11:21 • 查看问题日志
被浏览 **35865** 次，相关话题关注者 **926886** 人

具体有时间再补充，感兴趣的朋友可以Google Scholar 一下他们的文献。

关于数据科学家的更多讨论：
你能列出十个著名的女性数据科学家吗？Can you name 10 famous data scientist women?
谁是最富有的数据科学家？Who are the wealthiest data scientists?
请列出对大数据最具有影响力的20个人？Who Are The Top 20 Influencers in Big Data?

二、数据科学家的定义
数据科学(Data Science)是从数据中提取知识的研究，关键是科学。数据科学集成了多种领域的不同元素，包括信号处理，数学，概率模型技术和理论，机器学习，计算机编程，统计学，数据工程，模式识别和学习，可视化，不确定性建模，数据仓库，以及从数据中析取规律和产品的高性能计算。数据科学并不局限于大数据，但是数据量的扩大诚然使得数据科学的地位越发重要。

数据科学的从业者被称为数据科学家。数据科学家通过精深的专业知识在某些科学学科解决复杂的数据问题。不远的将来，数据科学家们需要精通一门、两门甚至多门学科，同时使用数学，统计学和计算机科学的生产要素展开工作。所以数据科学家就如同一个**team**。

曾经投资过Facebook，LinkedIn的格雷洛克风险投资公司把数据科学家描述成“能够管理和洞察数据的人”。在IBM的网站上，数据科学家的角色被形容成“一半分析师，一半艺术家”。他们代表了商业或数据分析这个角色的一个进化。

> for example – a data scientist will most likely explore and examine data from multiple disparate sources. The data scientist will sift through all incoming data with the goal of discovering a previously hidden insight, which in turn can provide a competitive advantage or address a pressing business problem. A data scientist does not simply collect and report on data, but also looks at it from many angles, determines what it means, then recommends ways to apply the data.

- **Anjul Bhambhri**，**IBM**的大数据产品副总裁。

数据科学家是一个好奇的，不断质疑现有假设，能盯着数据就能指出趋势的人。这就好像在文艺复兴时期，一个非常想为组织带来挑战并从挑战中学习的人一样。

- **Jonathan Goldman**，**LinkedIn**数据科学家。

2006年的6月份进入商务社交网站LinkedIn，当时LinkedIn只有不到800万用户。高德曼在之后的研究中创造出新的模型，利用数据预测注册用户的人际网络。具体来讲，他以用户在LinkedIn的个人资料，来找到和这些信息最匹配的三个人，并以推荐的形式显示在用户的使用页面上——这也就是我们熟悉的"你可能认识的人（People you may know）"。这个小小的功能让LinkedIn增加了数百万的新的页面点击量(数据挖掘的应用典型之一推荐系统）。

- **John Rauser,** 亚马逊大数据科学家。

数据科学家是工程师和统计学家的结合体。从事这个职位要求极强的驾驭和管理海量数据的能力；同时也需要有像统计学家一样萃取、分析数据价值的本事，二者缺一不可。

- **Steven Hillion, EMC Greenplum**数据分析副总裁。

数据科学家是具有极强分析能力和对统计和数学有很深研究的数据工程师。他们能从商业信息等其他复杂且海量的数据库中洞察新趋势。

- **Monica Rogati, LinkedIn**资深数据科学家。

所有的科学家都是数据学家，因为他们整天都在和海量数据打交道。在我眼中，数据学家是一半黑客加一半分析师。他们通过数据建立看待事物的新维度。数据学家必须能够用一只眼睛发现新世界，用另一只眼睛质疑自己的发现。

- **Daniel Tunkelang**，**LinkedIn**首席数据科学家。

我是bitly 首席科学家Hilary Mason的忠实崇拜者。关于这个新概念的定义我也想引用她的说法：数据科学家是能够利用各种信息获取方式、统计学原理和机器的学习能力对其掌握的数据进行收集、去噪、分析并解读的角色。

- **Michael Rappa**，北卡罗莱纳州立大学教授。

尽管数据科学家这个名称最近才开始在硅谷出现，但这个新职业的产生却是基于人类上百年对数据分析的不断积累和衍生。和数据科学家最接近的职业应该是统计学家，只不过统计学家是一个成熟的定义且服务领域基本局限于政府和学界。数据科学家把统计学的精髓带到了更多的行业和领域。

- 林仕鼎，百度大数据首席架构师。

如果从广义的角度讲，从事数据处理、加工、分析等工作的数据科学家、数据架构师和数据工程师都可以笼统地称为数据科学家；而从狭义的角度讲，那些具有数据分析能力，精通各类算法，直接处理

数据的人员才可以称为数据科学家。

最后引用Thomas H. Davenport（埃森哲战略变革研究院主任） 和 D.J. Patil（美国科学促进会科学与技术政策研究员，为美国国防部服务）的话来总结数据科学家需要具备的能力：

- 数据科学家倾向于用探索数据的方式来看待周围的世界。（好奇心）
- 把大量散乱的数据变成结构化的可供分析的数据，还要找出丰富的数据源，整合其他可能不完整的数据源，并清理成结果数据集。（问题分体整理能力）
- 新的竞争环境中，挑战不断地变化，新数据不断地流入，数据科学家需要帮助决策者穿梭于各种分析，从临时数据分析到持续的数据交互分析。（快速学习能力）
- 数据科学家会遇到技术瓶颈，但他们能够找到新颖的解决方案。（问题转化能力）
- 当他们有所发现，便交流他们的发现，建议新的业务方向。（业务精通）
- 他们很有创造力的展示视觉化的信息，也让找到的模式清晰而有说服力。（表现沟通能力）
- 他们会把蕴含在数据中的规律建议给Boss，从而影响产品，流程和决策。（决策力）

三、数据科学家所需硬件技能

《数据之美 Beautiful Data》的作者Jeff Hammerbacher在书中提到，对于 Facebook 的数据科学家“我们发现传统的头衔如商业分析师、统计学家、工程师和研究科学家都不能确切地定义我们团队的角色。该角色的工作是变化多样的：

在任意给定的一天，团队的一个成员可以用 Python 实现一个多阶段的处理管道流、设计假设检验、用工具R在数据样本上执行回归测试、在 Hadoop 上为数据密集型产品或服务设计和实现算法，或者把我们分析的结果以清晰简洁的方式展示给企业的其他成员。为了掌握完成这多方面任务需要的技术，我们创造了数据科学家这个角色。”

(1) 计算机科学

一般来说，数据科学家大多要求具备编程、计算机科学相关的专业背景。简单来说，就是对处理大数据所必需的Hadoop、Mahout等大规模并行处理技术与机器学习相关的技能。

- 零基础学习 Hadoop 该如何下手？
- 想从事大数据、海量数据处理相关的工作，如何自学打基础？

(2) 数学、统计、数据挖掘等

除了数学、统计方面的素养之外，还需要具备使用SPSS、SAS等主流统计分析软件的技能。其中，面向统计分析的开源编程语言及其运行环境“R”最近备受瞩目。R的强项不仅在于其包含了丰富的统计分析库，而且具备将结果进行可视化的高品质图表生成功能，并可以通过简单的命令来运行。此外，它还具备称为CRAN（The Comprehensive R Archive Network）的包扩展机制，通过导入扩展包就可以使用标准状态下所不支持的函数和数据集。R语言虽然功能强大，但是学习曲线较为陡峭，个人建议从python入手，拥有丰富的statistical libraries，NumPy ，SciPy.org ，Python Data Analysis Library ，matplotlib: python plotting 。

- 如何系统地学习数据挖掘？
- 做数据分析不得不看的书有哪些？
- 怎么学习用R语言进行数据挖掘？

(3) 数据可视化（**Visualization**）

信息的质量很大程度上依赖于其表达方式。对数字罗列所组成的数据中所包含的意义进行分析，开发Web原型，使用外部API将图表、地图、Dashboard等其他服务统一起来，从而使分析结果可视化，这是对于数据科学家来说十分重要的技能之一。

- 有哪些值得推荐的数据可视化工具？

(4) 跨界为王

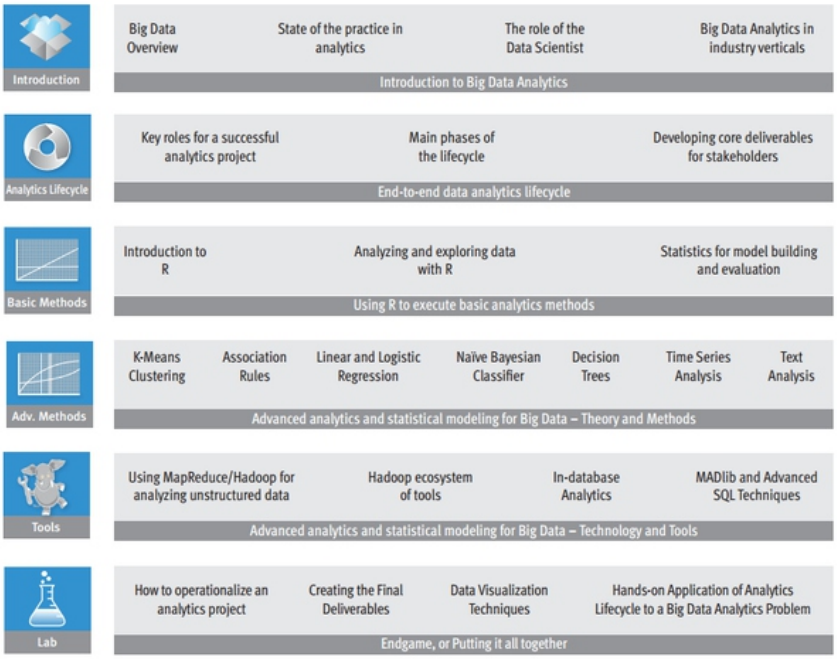
麦肯锡认为未来需要更多的“translators”，能够在IT技术，数据分析和商业决策之间架起一座桥梁的复合型人才是最被人需要的。”translators“可以驱动整个数据分析战略的设计和执行，同时连接的IT，数据分析和业务部门的团队。如果缺少“translators“，即使拥有高端的数据分析策略和工具方法也是于事无补的。

> The data strategists’combination of IT knowledge and experience making business decisions makes them well suited to define the data requirements for high-value business analytics. Data scientists combine deep analytics expertise with IT know-how to develop sophisticated models and algorithms. Analytic consultants combine practical business knowledge with analytics experience to zero in on high-impact opportunities for analytics.

天才的”translators“非常罕见。但是大家可以各敬其职（三个臭皮匠臭死诸葛亮），数据战略家可以使用IT知识和经验来制定商业决策，数据科学家可以结合对专业知识的深入理解使用IT技术开发复杂的模

型和算法，分析顾问可以结合实际的业务知识与分析经验聚焦下一个行业爆点。
推荐关注：facebook.com/data

四、数据科学家的培养
位于伊利诺伊州芝加哥郊外埃文斯顿市的美国名牌私立大学——西北大学（Northwestern University），就是其中之一。西北大学决定从2012年9月起在其工程学院下成立一个主攻大数据分析课程的分析学研究生院，并开始了招生工作。西北大学对于成立该研究生院是这样解释的：“虽然只要具备一些Hadoop和Cassandra的基本知识就很容易找到工作，但拥有深入知识的人才却是十分缺乏的。”
此外，该研究生院的课程计划以“传授和指导将业务引向成功的技能，培养能够领导项目团队的优秀分析师”为目标，授课内容在数学、统计学的基础上，融合了尖端计算机工程学和数据分析。课程预计将涵盖分析领域中主要的三种数据分析方法：预测分析、描述分析（商业智能和数据挖掘）和规范分析（优化和模拟），具体内容如下。
(1) 秋学期
* 数据挖掘相关的统计方法（多元Logistic回归分析、非线性回归分析、判别分析等）
* 定量方法（时间轴分析、概率模型、优化）
* 决策分析（多目的决策分析、决策树、影响图、敏感性分析）
* 树立竞争优势的分析（通过项目和成功案例学习基本的分析理念）
(2) 冬学期
* 数据库入门（数据模型、数据库设计）
* 预测分析（时间轴分析、主成分分析、非参数回归、统计流程控制）
* 数据管理（ETL（Extract、Transform、Load）、数据治理、管理责任、元数据）
* 优化与启发（整数计划法、非线性计划法、局部探索法、超启发（模拟退火、遗传算法））
(3) 春学期
* 大数据分析（非结构化数据概念的学习、MapReduce技术、大数据分析方法）
* 数据挖掘（聚类（k-means法、分割法）、关联性规则、因子分析、存活时间分析）
* 其他，以下任选两门（社交网络、文本分析、Web分析、财务分析、服务业中的分析、能源、健康医疗、供应链管理、综合营销沟通中的概率模型）
(4) 秋学期
* 风险分析与运营分析的计算机模拟
* 软件层面的分析学（组织层面的分析课题、IT与业务用户、变革管理、数据课题、结果的展现与传达方法）



（EMC的在线课程：Data Science and Big Data Analytics Training ，收费T_T，大家可以了解下学习路径）

(5)分享一些免费的课程
以下课程免费，讲师都是领域的专家，需要提前报名，请注意开班的时间。

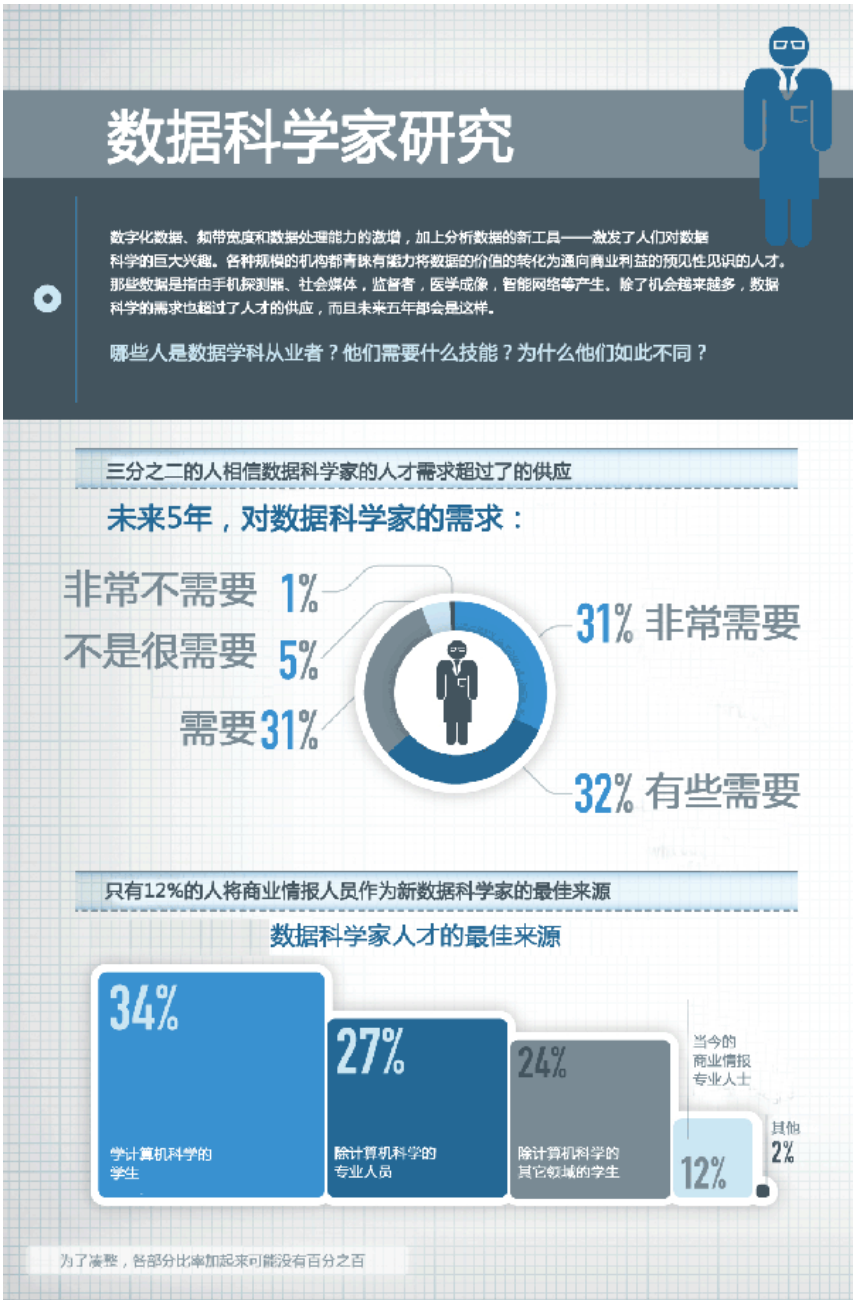
- Coursera.org ：统计学。
- Coursera.org ：机器学习。

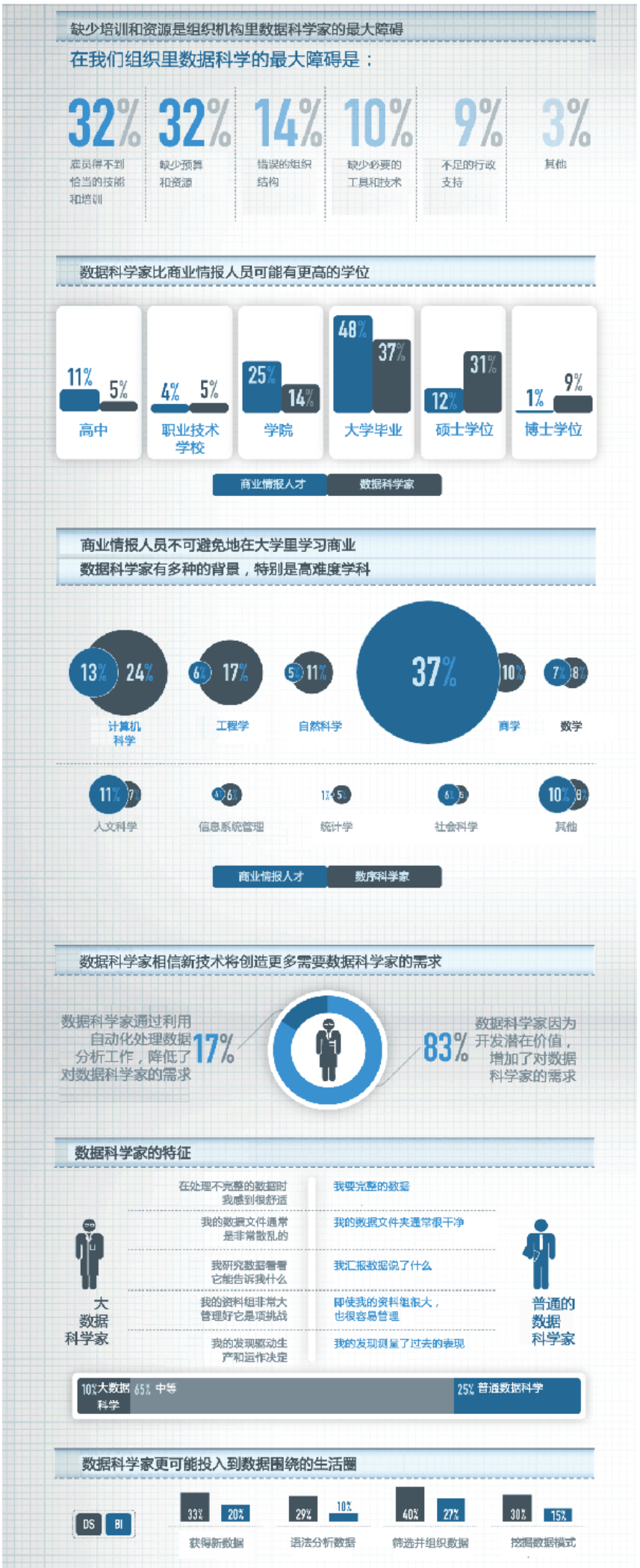
- Coursera.org ：数据分析的计算方法。
- Coursera.org ：大数据。
- Coursera.org ：数据科学导论。
- Coursera.org ：数据分析。

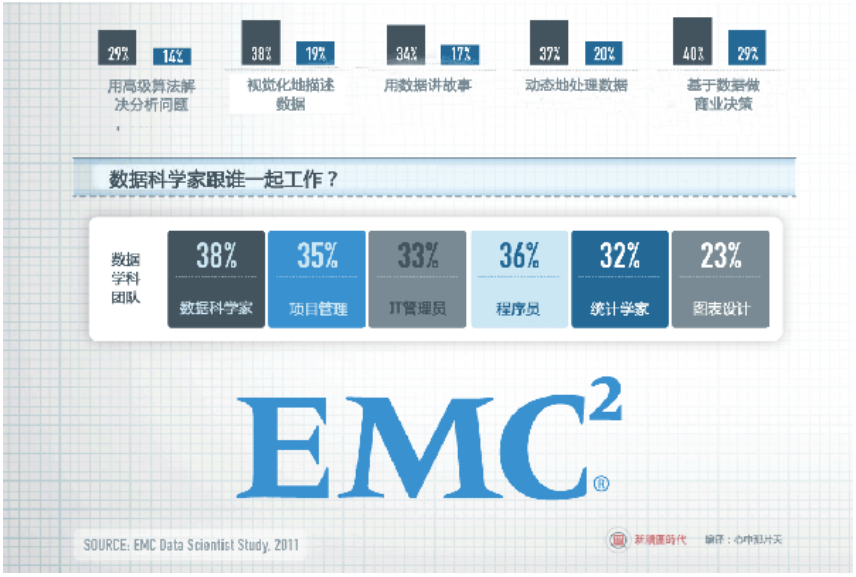
名校课程，需要一定的英语基础和计算机基础：

- Statistical Thinking and Data Analysis ：麻省理工学院的统计思维与数据分析课。概率抽样，回归，常见分布等。
- Data Mining | Sloan School of Management ：麻省理工学院的数据挖掘课程，数据挖掘的知识以及机器学习算法。
- Rice University Data Visualization ：莱斯大学的数据可视化，从统计学的角度分析信息可视化。
- Harvard University Introduction to Computing, Modeling, and Visualization ：哈佛大学，如何在数学计算与数据交互可视化之间架起桥梁。
- UC Berkeley Visualization ：加州大学伯克利分校数据可视化。
- Data Literacy Course -- IAP ：两个MIT的数据研究生，如何分析处理可视化数据。
- Columbia University Applied Data Science ：哥伦比亚大学，数据分析方法。需要一定的数据基础。
- SML: Systems ：加州大学伯克利分校，可扩展的机器学习方法。从硬件系统，并行化范式到MapReduce+Hadoop+BigTable，非常全面系统。

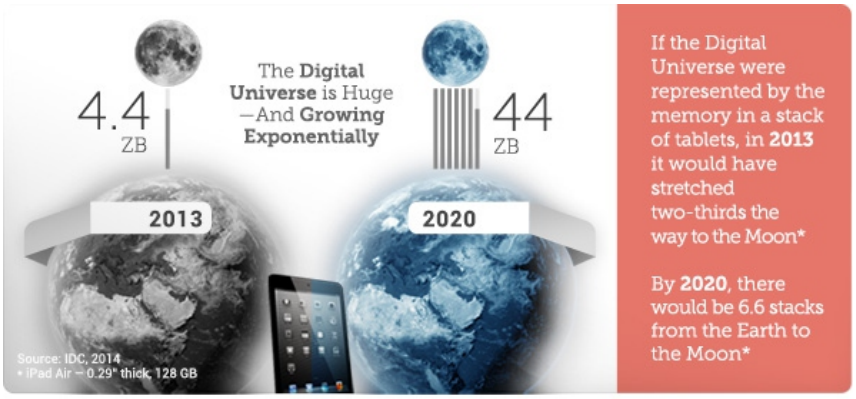
五、数据科学家的前景




缺少培训和资源是组织机构里数据科学家的最大障碍
在我们组织里数据科学的最大障碍是：
32% 雇员得不到恰当的技能和培训
32% 缺少预算和资源
14% 错误的组织结构
10% 缺少必要的工具和技术
9% 不足的行政支持
3% 其他
数据科学家比商业情报人员可能有更高的学位
高中 11% 5%
职业技术学校 4% 5%
学院 25% 14%
大学毕业 48% 37%
硕士学位 12% 31%
博士学位 1% 9%
商业情报人才
数据科学家
商业情报人员不可避免地在大学里学习商业
数据科学家有多种的背景，特别是高难度学科
计算机科学 13% 24%
工程学 6% 17%
自然科学 5% 11%
37%
商学 10%
数学 7% 8%
人文科学 11% 7%
信息系统管理 6%
统计学 5%
社会科学 6%
其他 10% 8%
商业情报人才
数据科学家
数据科学家相信新技术将创造更多需要数据科学家的需求
数据科学家通过利用自动化处理数据分析工作，降低了对数据科学家的需求 17%
83% 数据科学家因为开发潜在价值，增加了对数据科学家的需求
数据科学家的特征
在处理不完整的数据时我感到很舒适
我要完整的数据
我的数据文件通常是非常散乱的
我的数据文件夹通常很干净
我研究数据看看它能告诉我什么
我汇报数据说了什么
我的资料组非常大管理好它是项挑战
即使我的资料组很大，也很容易管理
我的发现驱动生产和运作决定
我的发现测量了过去的表现
大数据科学家
普通的数据科学家
10%大数据科学
65% 中等
25% 普通数据科学
数据科学家更可能投入到数据围绕的生活圈
DS
BI
获得新数据 33% 20%
语法分析数据 29% 10%
筛选并组织数据 40% 27%
挖掘数据模式 30% 15%

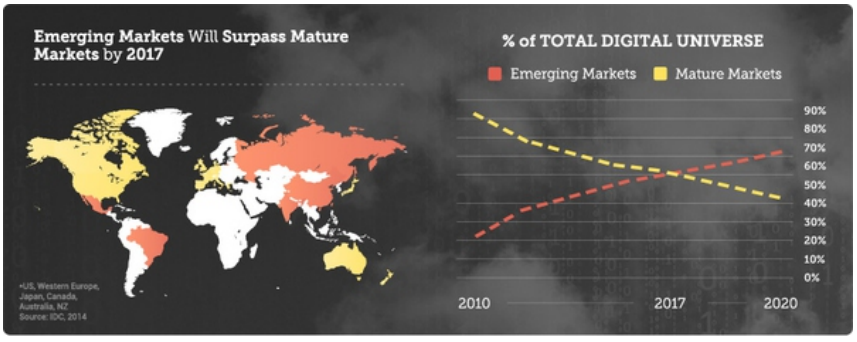


（EMC - Leading Cloud Computing, Big Data, and Trusted IT Solutions ，关于数据科学家的研究）



Like the physical universe, the digital universe is large – by 2020 containing nearly as many digital bits as there are stars in the universe. It is doubling in size every two years, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes. EMC预测，按照目前的情况数字宇宙以每两年一番的速度倍增，在2020年将到达44ZB（1ZB=1.1805916207174113e+21B）。EMC做出了5点比较大胆的预测。

- In 2013, while about 40% of the information in the digital universe required some type of data protection, less than 20% of the digital universe actually had these protections.
- Data from embedded systems, the signals from which are a major component of the Internet of Things, will grow from 2% of the digital universe in 2013 to 10% in 2020.
- In 2013, less than 20% of the data in the digital universe is “touched” by the cloud, either stored, perhaps temporarily, or processed in some way. By 2020, that percentage will double to 40%.
- Most of the digital universe is transient – unsaved Netflix or Hulu movie streams, or Xbox One gamer interactions, temporary routing information in networks, sensor signals discarded when no alarms go off, etc. – and it is getting more so. This is a good thing, because the world’s amount of available storage capacity (i.e., unused bytes) across all media types is growing slower than the digital universe. In 2013, the available storage capacity could hold just 33% of the digital universe. By 2020, it will be able to store less than 15%.
- In 2014, the digital universe will equal 1.7 megabytes a minute for every person on Earth.



Between 2013 and 2020 the division of the digital universe between mature and emerging markets (e.g., China) will switch – from 60% accounted for by mature markets to 60% of the data in the digital universe coming from emerging markets.
EMC预测在2017年左右新兴的市场将超越成熟市场，东亚国家是最具潜力的引爆点。（大家是不是有点小激动，前景一片光明）

六、结束语

推荐网站：

Data Science Central （数据科学中心，大牛云集，资源丰富，讨论者热情，各种课程）

祝每一个**DMer**都挖掘到金矿和快乐：）

参考文献：
[1].Data Scientists: The Definition of Sexy
[2].《大数据的冲击》. 城田真琴. 野村综合研究所创新开发部高级研究员、*IT*分析师，日本政府"智能云计算研究会"智囊团成员
*[3].*麦肯锡. *Big data: The next frontier for innovation, competition, and productivity*
[4].EMC. Executive Summary: Data Growth, Business Opportunities, and the IT Imperatives
[5].EMC Greenplum's Steven Hillion on What Is a Data Scientist?
[6].LinkedIn's Monica Rogati On "What Is A Data Scientist?"
[7].IBM - What is a Data Scientist?
*[8].*Data Science and Prediction
[9].The key word in "Data Science" is not Data, it is Science
[10].Data Science: How do I become a data scientist?
[11].A Practical Intro to Data Science
*[12].*解码数据科学家

编辑于 2014-05-14 39 条评论 感谢 分享 收藏 • 没有帮助 • 举报

1125

谢科，**Twitter Data Science Team**

Terence Lee、吴志强、Yuqi Qian 等人赞同

"Data Science = statistics who uses python and lives in San Francisco"

恰好我马上启程到Twitter的data science team，而且恰巧懂一点点统计和住在旧金山，所以冲动地没有邀请就厚脸回答了:D

我认为有几个大方面

1）学好python。
现在几乎所以公司的数据都可以api给你，而python的数据处理能力强大且方便。加之在machine learning的很多算法上，python也独俏一方。另外，它的简明方便迅速迭代开发，15分钟写完个算法就可以看效果了。

除此之外，py还有点酷酷的感觉。任何程序拿matlab和c++都是可以写的，不过我真没认识过哪个d愿意自己把自己扔那个不酷的框框里:D

对不规则输入的处理也给python一个巨大的优势。通常来说，在我现在日常的工作里，所有的数据都是以纯文本但是非格式的形式存储的（raw text, unstructured data)。问题在于，这些文本不可以直接当作各种算法的输入，你需要

1. 分词，分句
2. 提取特征
3. 整理缺失数据
4. 除掉异类（outlier）

在这些时候，python可谓是神器。这里做的1-4都可以直接在scikit-learn里面找到对应的工具，而且，即使是要自己写一个定制的算法处理某些特殊需求，也就是一百行代码的事情。

简而言之，对于数据科学面临的挑战，python可以让你短平快地解决手中的问题，而不是担心太多实现细节。

2）学好统计学习

略拗口。统计学习的概念就是“统计机器学习方法”。

统计和计算机科学前几十年互相平行着，互相造出了对方造出的一系列工具，算法。但是直到最近人们开始注意到，计算机科学家所谓的机器学习其实就是统计里面的prediction而已。因此这两个学科又开始重新融合。

为什么统计学习很重要？

因为，纯粹的机器学习讲究算法预测能力和实现，但是统计一直就强调“可解释性”。比如说，针对今天微博股票发行就上升20%，你把你的两个预测股票上涨还是下跌的model套在新浪的例子上，然后给你的上司看。

Model-1有99%的预测能力，也就是99%的情况下它预测对，但是Model-2有95%，不过它有例外的一个附加属性——可以告诉你为什么这个股票上涨或者下跌。

试问，你的上司会先哪个？问问你自己会选哪个？

显然是后者。因为前者虽然有很强的预测力（机器学习），但是没有解释能力（统计解释）。

而作为一个数据科学家，80%的时间你是需要跟客户，团队或者上司解释为什么A可行B不可行。如果你告诉他们，“我现在的神经网络就是能有那么好的预测力可是我根本就没法解释上来”，那么，没有人会愿意相信你。

具体一些，怎么样学习统计学习？

- 先学好基本的概率学。如果大学里的还给老师了（跟我一样），那么可以从MIT的概率论教材【1】入手。从第1章到第9章看完并做完所有的习题。（p.s.面试Twitter的时候被问到一个拿球后验概率的问题，从这本书上抓来的）。
- 了解基本的统计检验及它们的假设，什么时候可以用到它们。
- 快速了解统计学习有哪些术语，用来做什么目的，读这本【5】。
- 学习基本的统计思想。有frequentist的统计，也有bayesian的统计。前者的代表作有【2】，后者看【3】。前者是统计学习的圣书，偏frequentist，后者是pattern recognition的圣书，几乎从纯bayesian的角度来讲。注意，【2】有免费版，作者把它全放在了网上。而且有一个简易版，如果感觉力不从心直接看【2】，那么可以先从它的简易版开始看。简易版【4】是作者在coursera上开课用的大众教材，简单不少（不过仍然有很多闪光点，通俗易懂）。对于【3】，一开始很难直接啃下来，但是啃下来会受益匪浅。

注意，以上的书搜一下几乎全可以在网上搜到别人传的pdf。有条件的同学可以买一下纸制版来读，体验更好并且可以支持一下作者。所有的书我都买了纸制版，但是我知道在国内要买本书有多不方便（以及原版书多贵）。

读完以上的书是个长期过程。但是大概读了一遍之后，我个人觉得是非常值得的。如果你只是知道怎么用一些软件包，那么你一定成不了一个合格的data scientist。因为只要问题稍加变化，你就不知道怎么解决了。

如果你感觉自己是一个二吊子数据科学家（我也是）那么问一下下面几个问题，如果有2个答不上来，那么你就跟我一样，真的还是二吊子而已，继续学习吧。

- 为什么在神经网络里面feature需要standardize而不是直接扔进去
- 对Random Forest需要做Cross-Validatation来避免overfitting吗？
- 用naive-bayesian来做bagging，是不是一个不好的选择？为什么？
- 在用ensembe方法的时候，特别是Gradient Boosting Tree的时候，我需要把树的结构变得更复杂（high variance, low bias)还是更简单（low variance, high bias)呢？为什么？

如果你刚开始入门，没有关系，回答不出来这些问题很正常。如果你是一个二吊子，体会一下，为什么你跟一流的data scientist还有些差距——因为你不了解每个算法是怎么工作，当你想要把你的问题用那个算法解决的时候，面对无数的细节，你就无从下手了。

说个题外话，我很欣赏一个叫Jiro的寿司店，它的店长在（东京？）一个最不起眼的地铁站开了一家全世界最贵的餐馆，预订要提前3个月。怎么做到的？70年如一日练习如何做寿司。70年！除了丧娶之外的假期，店长每天必到，8个小时工作以外继续练习寿司做法。

其实学数据科学也一样，沉下心来，练习匠艺。

3）学习数据处理

这一步不必独立于2）来进行。显然，你在读这些书的时候会开始碰到各种算法，而且这里的书里也会提到各种数据。但是这个年代最不值钱的就是数据了（拜托，为什么还要用80年代的“加州房价数据”？），值钱的是数据分析过后提供给决策的价值。那么与其纠结在这么悲剧的80年代数据集上，为什么不自己搜集一些呢？

- 开始写一个小程序，用API爬下Twitter上随机的tweets（或者weibo吧。。。）
- 对这些tweets的text进行分词，处理噪音（比如广告）
- 用一些现成的label作为label，比如tweet里会有这条tweet被转发了几次
- 尝试写一个算法，来预测tweet会被转发几次
- 在未见的数据集上进行测试

如上的过程不是一日之功，尤其刚刚开始入门的时候。慢慢来，耐心大于进度。

4）变成全能工程师（full stack engineer）
在公司环境下，作为一个新入职的新手，你不可能有优待让你在需要写一个数据可视化的时候，找到一个同事来给你做。需要写把数据存到数据库的时候，找另一个同事来给你做。

况且即使你有这个条件，这样频繁切换上下文会浪费更多时间。比如你让同事早上给你塞一下数据到数据库，但是下午他才给你做好。或者你需要很长时间给他解释，逻辑是什么，存的方式是什么。

最好的变法，是把你自己武装成一个全能工作师。你不需要成为各方面的专家，但是你一定需要各方面都了解一点，查一下文档可以上手就用。

- 会使用NoSQL。尤其是MongoDB
- 学会基本的visualization，会用基础的html和javascript，知道d3【6】这个可视化库，以及highchart【7】
- 学习基本的算法和算法分析，知道如何分析算法复杂度。平均复杂度，最坏复杂度。每次写完一个程序，自己预计需要的时间（用算法分析来预测）。推荐普林斯顿的算法课【8】（注意，可以从算法1开始，它有两个版本）
- 写一个基础的服务器，用flask【9】的基本模板写一个可以让你做可视化分析的backbone。
- 学习使用一个顺手的IDE，VIM， pycharm都可以。

4）读，读，读！
除了闭门造车，你还需要知道其它数据科学家在做些啥。涌现的各种新的技术，新的想法和新的人，你都需要跟他们交流，扩大知识面，以便更好应对新的工作挑战。

通常，非常厉害的数据科学家都会把自己的blog放到网上供大家参观膜拜。我推荐一些我常看的。另外，学术圈里也有很多厉害的数据科学家，不必怕看论文，看了几篇之后，你就会觉得：哈！我也能想到这个！

读blog的一个好处是，如果你跟他们交流甚欢，甚至于你可以从他们那里要一个实习来做！

betaworks首席数据科学家，Gilad Lotan的博客，我从他这里要的intern :D Gilad Lotan
Ed Chi，六年本科硕士博士毕业的神人，google data science edchi.blogspot.com/
Hilary Mason，bitly首席科学家，纽约地区人尽皆知的数据科学家：hilarymason.com

在它们这里看够了之后，你会发现还有很多值得看的blog（他们会在文章里面引用其它文章的内容），这样滚雪球似的，你可以有够多的东西早上上班的路上读了：）

5）要不要上个研究生课程？
先说我上的网络课程：
Coursera.org
coursera.org/course/mac...
前者就不说了，人人都知道。后者我则更喜欢，因为教得更广阔，上课的教授也是世界一流的机器学习学者，而且经常会有一些很妙的点出来，促进思考。

对于是不是非要去上个研究生（尤其要不要到美国上），我觉得不是特别有必要。如果你收到了几个著名大学数据科学方向的录取，那开开心心地来，你会学到不少东西。但是如果没有的话，也不必纠结。我曾有幸上过或者旁听过美国这里一些顶级名校的课程，我感觉它的作用仍然是把你领进门，以

及给你一个能跟世界上最聪明的人一个交流机会（我指那些教授）。除此之外，修行都是回家在寝室进行的。然而现在世界上最好的课程都摆在你的面前，为什么还要舍近求远呢。

总结一下吧
我很幸运地跟一些最好的数据科学家交流共事过，从他们的经历看和做事风格来看，真正的共性是

他们都很聪明——你也可以
他们都很喜欢自己做的东西——如果你不喜欢应该也不会看这个问题
他们都很能静下心来学东西——如果足够努力你也可以

【1】Introduction to Probability and Statistics
【2】Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009. 免费版
【3】Bishop, Christopher M. *Pattern recognition and machine learning*. Vol. 1. New York: springer, 2006.
【4】Introduction to Statistical Learning 免费版
【5】Wasserman, Larry. *All of statistics: a concise course in statistical inference*. Springer, 2004.
【6】d3js.org/
【7】highcharts.com/
【8】Coursera.org
【9】flask.pocoo.org/

发布于 2014-04-18 67 条评论 感谢 分享 收藏 • 没有帮助 • 举报

陈然，**CS@CMU**

40

哈哈、Entropy、袁昊 等人赞同

推荐一亩三分地W大的系列文章：
现在很火的数据科学data science到底是什么？你对做Data Scientist感兴趣吗？
数据科学家data scientist需要的三大核心技能：Data Hacking、Problem Solving and Communication

想成为数据科学家Data Scientist，需要申请读什么专业？
美国哪些公司招聘Data Scientist？看重数据科学家什么方面的背景？
数据科学家Data Scientist的职业发展前景如何？
推荐《说说Data Science的入门级工作》和《数据科学就业前景观察分析》，欢迎来分享你的经验和看法！
数据科学家Data Scientist能挣多少钱？
做数据科学家Data Scientist，硕士Master和博士PhD学位有啥区别？

发布于 2014-04-19 1 条评论 感谢 分享 收藏 • 没有帮助 • 举报

Albert Xiao，在路上

20

文渊、JOKER、hijiangtao 等人赞同

Quora回答：quora.com/Career-Advice...

编辑于 2014-01-19　4 条评论　感谢　分享　收藏 • 没有帮助 • 举报

李梦佳，好奇心很强的数学女博士

4

9kids、xiao ma、周庭锐 等人赞同

Road Map to Data Scientist

最好看大图。

哈哈手机党对不住了，请谨慎点击。

编辑于 2014-04-26　3 条评论　感谢　分享　收藏 • 没有帮助 • 举报

匿名用户

5

Galileo ZH、琳Lin、bsdelf 等人赞同

简单的说，在国内，学会扯淡；在国外，学过统计，并学会扯淡。

顺便顶一下肥濛的回答。

同学同事中有几位自我定义为数据科学家的朋友，为了不友尽，匿之。

发布于 2014-04-20　1 条评论　感谢　分享　收藏 • 没有帮助 • 举报

车明洲，二十一世纪的学术在民间

4 泓林、于忠鹏、duan raymond 等人赞同

我今年大一，现在在新加坡国立读一个叫“商业分析”的专业，这个专业专门为公司培养数据分析员。今年第一年招生，课程是计算机学院的一帮计算机科学和信息系统（Information System)的教授专门设计的，其中包括编程和IS（计算机学院）、统计（科学院）、商学院，三个faculty的课程。我学了一年下来最大的感受是，编程编不过学CS的，统计理解的没学统计的好，商学院的课做presentation被虐；但是他们对于别的领域不知道的知识我能略知一二。

以我现在的理解，其实所谓“数据科学家”，应该更加强调商业上的应用。与其把这个作为一种新的学科、新的知识，不如把它想成一种新的“职业”；这个”商业分析“专业也许没有对科学的发展、新知识的发现起到多少作用，但是能帮助学生在学习过程中对知识进行整合，从而培养出一种现在公司能用的“员工”。就像我们同学一直自嘲，我们不过是一帮”高级技工“罢了，只不过这种技工，现在市场需求不小，待遇也不差。

发布于 2014-04-24 1 条评论 感谢 分享 收藏 • 没有帮助 • 举报

---

1 **wise**，从事数据挖掘

燧石 赞同

[偷笑]全文都说学哪些工具，却没提及到熟悉业务、理解业务需求，多学点市场营销的东西，把算法和业务结合，这些都没提及到，就想做科学家？

发布于 2014-05-11 添加评论 感谢 分享 收藏 • 没有帮助 • 举报

0 王建飞

数据科学家必学课程：

用户3770845573_新浪博客

补充一点：大数据时代，数据科学家必须了解大数据技术，另外 还有一些相关领域：

物联网，云计算， 我个人认为其关系可以总结如下：

物联网=sensor+大数据+云计算

编辑于 2014-04-28 添加评论 感谢 分享 收藏 • 没有帮助 • 举报

---

0 匿名用户

Coursera和Udacity都有一门data science的课程path

发布于 2014-04-18 添加评论 感谢 分享 收藏 • 没有帮助 • 举报

---

0 李泽超

未来式的计算机世界，一定是云计算和大数据的世界，成为数据科学家，我认为既要懂数理统计方面的知识，同时还要深入了解业务，是该行业和领域的业务专家，能通过数据模型的建立，对未来发展的趋势进行预测

发布于 2014-05-17 1 条评论 感谢 分享 收藏 • 没有帮助 • 举报

0 **hijiangtao**，在学会如何玩互联网之前，我仍一直被互联...

> Mike Driscoll’s three sexy skills of data geeks:
> • Statistics– traditional analysis
> • Data Munging– parsing, scraping, and formatting data
> • Visualization– graphs, tools, etc.

编辑于 2014-07-12 添加评论 感谢 分享 收藏 • 没有帮助 • 举报

11 肥濛

yyy www、郑伟烁、张子谋 等人赞同

数据科学家，就是一帮学统计的互相往自己脸上贴金。

发布于 2014-04-18 10 条评论 感谢 分享 收藏 • 没有帮助 • 举报

---

5 个回答被折叠 （为什么？）

卢大虾，长生不老真的那么重要吗？ 修改话题经验

写回答...

匿名 发布回答