Data Application Lab

数据科学家求职培训白皮书

Data Scientist Training & Career Development Guideline

数据应用学院 编写

TABLE OF CONTENTS

前言

当我们第一次写这本“数据科学的职业指南”时，我们并没有预期到它将会获得如此大的关注，就如我们 2015 年秋第一次在美国加州洛杉矶 UCLA 做“大数据和数据科学职业发展公益讲座”一样，数以千计的人在几天内报名关注。大家的热情和兴趣，让我们相信，数据科学（Data Science）是一个正在快速发展、充满着令人兴奋的机会、但又略显模糊的领域。对于想要求职、转行进入数据领域的个人，往往缺少一个很好的全面建议。

在和越来越多的人交谈的过程中，我们发现只有很少的资源完成解释了如何打破进入数据科学的事业。网上的确有越来越多公共资源，MOOC 公开课、数据挖掘教材和文章，Python 语言的应用指南供大家选择，然而有限的时间内，除了快速学习能力以外，科学有针对性的学习指导方案，能帮助大家高效明确方向更显得重要。虽然市面上有许多面经和面试问题的资料，但是我们无法找到一份完整的指南。没有一份资料能够全面地涵盖数据科学面试过程的一切，包括如何获得面试和如何处理任何提供的职位面试机会。

我们希望能有这样一个指南，从行业招聘和被招聘的两边同时收集观点；我们和参与候选人筛选的招聘人员，决定招聘的管理人员，以及成功通过数据科学面试的候选人进行交流，以通过过去经历过的人的见解来揭示数据科学面试过程。所以我们通过与多位业内人士、学生的交流，在大家的帮助下著了这本书。

在 Data Application Lab（数据应用学院），我们通过我 Bootcamp 小班、研讨会吸引了数千名数据科学爱好者、学习者，并在一年内帮助了 300 多名学生如愿找到数据科学相关工作。我们建立了超过 500 人的导师和校友社区，这为我们提供了独特的优势，为数据科学面试过程分享重要的经验和见解。

想要收集所有讯息很困难，就像许多候选人在这个求职过程中很困难一样。数据科学领域的一些业界专家，包括首席数据科学家，必须经过至少六个月的准备和努力才能得到收获！大多数公司的数据科学面试流程旨在筛选出最有决心和最熟练的候选人。有时会设置一些非常难跨越、让人望而生退的障碍。虽然投资似乎是巨大的，但回报可以更大。

数据科学被称为“21 世纪最性感的工作”。数据科学家不仅高薪，他们还帮助解决很多会产生重大的社会影响的问题：从优化城市规划到“解决世界贫困”，甚至在发生流行病之前停止恐慌的扩散。数据科学家发现了 Banksy 的身份，他们掌握了 March Madness 中预测篮球赛的艺术。在数据科学方面的工作并不仅仅是获得良好的工资和良好的工作生活平衡， 更是关于运用智慧解决重要的大问题。

我们写这本指南，因为我们希望你从对数据科学好奇开始，转而明确自己的方向，最终开始积极地尝试寻找工作。我们希望通过分享前人的经验指点迷津，了解数据科学面试过程，以及需要做的准备。

希望这本指南能帮助你一举拿下数据科学面试。

数据科学家 2018 年求职新趋势

根据 LinkedIn 的 2018 年 11 日发表的文章，“LinkedIn Data Reveals the Most Promising Jobs and In-Demand Skills of 2018”，数据科学家的工作岗位的 year over year （YOY）增长是 45%,仍然有很大的市场增长。但是数据科学求职会有一些新的变化，在 2018 年开始的时候，我们来讨论一下数据科学求职的新趋势。

我们先来看看数据科学家的学科和学位的数据分析

数据科学家都拥有什么学科
计算机学位 20%
数学和统计 20%
经济学和社会学 19%
自然科学学位（物理、化学或生物等） 11%
工程学科 9%
数据科学专业 13%

Others

我们可以看到，数据科学的专业背景很丰富，本专业只占有13%，其他的都是转行的。2018 仍然会是这个趋势，数据科学为各种不同专业的学生提供新的就业方向。

数据科学家都拥有什么学位
Ph.D 27%
硕士 48%
本科 15%
MBA 2%
Others

数据科学家一直是博士生的比例很高，占到三分之一，但是2018 有向硕士倾斜的趋势。这里主要的原因有几个，一个是市场需求太大，没有足够的博士生的资源，另一个原因是培训项目越来越成熟和丰富，数据应用学院（Data Application Lab）就帮助很多硕士生增强了数据科学技能，找到相关的工作。

下面我们再来看看数据科学在技术和应用领域的新趋势。
根绝 Gartner 对 2018 技术革新的预测，人工智能将被广泛应用到决策分析，而数据科学提供了大数据和机器学习的有效结合，为 AI 在决策中的应用提供具体的解决方案。

2018 第二个趋势是智能 APP 被广泛开发和应用，智能 APP 不是去替代人，而是加强人的能力，这里面包括自动化的数据清理，数据准备，数据发现和智能分享。
第三，深度学习将更广泛的被应用。现在深度学习是处理图像，影视应用的利器，今年我们会看到更多的应用领域，像 GIS 地理信息系统，智能物体（Intelligence thing）。
第四，数据科学平台化自动化越来越越来越流行，IBM，GOOGLE，AWS 和 MS 都推出平台化产品，大量初创公司更是进入平台化，像 DataRobot，datascience.com 等等。

应对数据科学技术革新的变化，作为求职者，我们应该如何应对呢？

第一个就是要拥抱人工智能。数据科学家的已经掌握了机器学习技术，我们要更多学习其他 AI 的技术，例如自然语言处理技术（NLP），知识图谱。

第二，我们要拥抱平台，特别是应有云计算云分析的平台知识。云技术解决了大数据的存储和管理问题，下一步就是云分析了。Google，AWS，IBM 每一家的平台都有其特色，可以适当学习和了解。

第三，我们要更好的掌握主业领域的知识，对于 Heal care，Finance，e-commerce 里面的数据科学应用场景，要深入了解和学习。

Reference：

https://www.datascience.com/blog/data-scientist-skills?utm_medium=email&_hsenc=p2ANqtz-9ZjJGtu2y-vw6lBnsBithtwe3uUZtGr9Pc5eYDOejEOQx63Ev61zCXtnUuke-HvDIdY_RMpHuPnz-nOY4azat653h7aA&_hsmi=59545730&utm_content=59545730&utm_source=hs_email&hsCtaTracking=56136152-bdec-49b2-bdf0-e4336533bd19%7C1cabe095-0d8d-44bb-9857-ec0cb3fcef57

https://blog.linkedin.com/2018/january/11/linkedin-data-reveals-the-most-promising-jobs-and-in-demand-skills-2018

https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/

https://blog.dataiku.com/top-4-data-science-trends-to-watch-in-2018

在希望拿到数据科学面试之前，你应该知道这个术语是什么意思，以及你将要迎来什么样的工作。

“Data Scientist”这个术语是由美国前首席数据科学家 DJ Patil 与 Cloudera 的首席数据科学家 Jeff Hammerbacher 首先提出的。数

据科学是一门利用数据学习知识的学科，其目标是通过从数据中提取出有价值的部分来生产数据产品。它结合了诸多领域中的理论和技术，包括应用数学，统计，模式识别，机器学习，数据可视化，数据仓库，以及高性能计算。

该术语在第一次使用十年后，仍然有争议。实践者和学者们对数据科学的意义以及数据分析公司一直使用的不同之处还有很多争议。人们谈论大数据和使用机器学习来解决数据问题时，其实是正在冒险地进入一个有着全新定义的未知领域。

然而不同的公司基于自己的需求，有着对数据科学不同的定义。每个招聘经理都会对他们正在寻找的资质、技能和打算聘用的候选人有着不同的期待。

这种混乱模糊的定义使得数据科学面试过程对于很多候选人来说很难。实际上，数据科学可以根据你正在申请的角色和正在面试的公司而有着非常不同的定义。

数据科学中不同的职业定位

数据科学家倾向于用探索数据的方式来看待周围的世界，把大量散乱的数据变成结构化的可供分析的数据，还要找出丰富的数据源，整合其他可能不完整的数据源，并清理成结果数据集。新的竞争环境中，挑战不断地变化，新数据不断地流入，数据科学家需要帮助决策者穿梭于各种分析，从临时数据分析到持续的数据交互分析。当他们有所发现，便交流他们的发现，建议新的业务方向。他们很有创造力的展示视觉化的信息，也让找到的模式清晰而有说服力，把蕴含在数据中的规律建议给管理层，从而影响产品，流程和决策。

首先，我们来了解数据科学在不同行业到底可以胜任什么样的角色。

如果你是一名在互联网公司的数据科学家，那你可能会需要通过建模分析来去讨论这个网站的流量。这个模型可以被数据分析师拿去直接使用，得出结论。而作为数据科学家，你需要追根溯源地走到数据的深处去探一探根本，然后建立一个可用的模型。

如果你是一个产品公司的数据科学家，那你需要根据公司所经营的产品特点分析市场的统计计算，去看产品分布，销量曲线，从而来预测，如果改变了一些属性，产品的销量会如何改变。

如果你是一个咨询公司的数据科学家，那你需要根据咨询公司的规则建模，然后根据每一位客户的情况来做数据可视化，现在的你，需要用数据讲故事。

当然作为数据科学家，你还需要知道如何做机器学习，深度学习等等。

接着，我们通过一个具体数据科学项目的实例——“使用 Deep Learning 对 Yelp 上的图像进行分类”，来详细说明数据科学中不同的角色。

数百万张照片每天在 Yelp 上传，但可能难以获得每个餐厅所需的图像。有时，上传的照片是相同的类别，也许他们都是食物的照片或餐厅的外观招聘。对一个餐厅的整体评价需要不同种类的图像。你可以使用机器学习自动对图像进行分类。计算机可以在训练集的帮助下告诉你图像是餐厅外还是食物。

数据科学家创建了模型来帮助机器创建这些区别。他们可以通过手动标记的照片到图像标题中的关键字来思考他们需要的数据类型。这往往是一个更高级的角色，因为他们经常管理数据产品，并处理数据科学问题的所有方面，从算法选择到工程设计。

数据工程师创建系统以获取所有图像数据并存储它们，并实现数据科学家选定的一些算法。这往往是强大的技术人员的角色，但可能不太了解他们正在实施的算法背后的理论。

数据分析师查询并介绍带来的业务影响。有没有受到用户的欢迎？Yelp 由于最近的变化而产生多少流量？这些是数据分析师会问的问题。然后，他们传达他们发现的见解。这个角色往往会被更多的入门级人员和具有商业背景的人员填补，以便在技术基础上应用他们的见解。

此外，我们再通过具体的两个商业案例来进一步了解数据科学家所发挥的作用。

【Uber 的大数据应用】

数据分析在 Uber 产品设计和功能优化中发挥着重要作用。以 Uber 新司机产品组为例，产品开发流程主要有以下四个步骤。

第一步是识别问题，团队主要会使用户调研、数据挖掘等方式来了解用户的需求、发现产品机会。例如 Uber 团队在观察了新司机的上线行为之后发现，一般上线后 10 分钟以上还常常接不到单子的新司机，更有可能会弃平台而去。

第二步是定量分析。在完成第一阶段的调研后，产品团队往往会得到一个长长的需求列表。数据科学家会通过数据分析，对这些问题做优先级排序。具体怎么做呢？首先是确定相关性。例如以上提出的各种问题是原因变量 X，司机的 Retention 是结果变量 Y，数据科学家把 X 和 Y 进行相关性分析，并在数据之中寻找证据，从而判断哪些因素是真正相关的。确定了相关性之后，可以通过建立模型，做回归分析、决策树等等来选择出对 Y 影响最为显著的一些因素。

第三步是生成产品想法。产品团队一般会通过头脑风暴和行业对标两周方式来形成解决问题的方案。

第四步是看新的落地方案有没有效果。为了降低风险，通常会挑选一小批城市来进行测试。团队会尽量选择一些市场条件差异很大的城市，测试产品在不同条件下的表现。之后，通过用户的反馈，团队会决定是否能够将产品落地。

在了解了数据分析在产品开发过程中应用的一般流程后，我们再来通过一个具体的案例来加深对此的理解。

在针对新司机的留存方面，如何让一个刚刚加入平台的司机能够逐步发展成活跃稳定的平台用户，进而一步步成长为老司机？那么核心的问题就是要让司机接到足够多的单子，赚取丰厚的利润。Uber 团队通过数据挖掘发现，当一些司机身处非市中心地段（大多数人拥有私家车或是其他原因导致他们不太常用 Uber），司机上线后，等待很久也没有新订单出现，这种情况的出现会导致 Uber 失去很多新司机。为了让这些司机能够在“对的时间”（需求量大的时段）上线，Uber 团队设计了一个

Peak Hours 的功能。告诉司机什么时段需求较为旺盛，何时上线接单更容易。

在产品功能设计完成之后，那么就需要进行测试来看这项功能是否真正能够解决新司机留存这个问题了。然而，测试结果却显示，增加新功能的实验组和未上线该功能的对照组之间的司机反馈并没有显著不同。

为什么是如此意想不到的结论？数据科学家通过进一步分析发现，原因可能主要有以下两个方面：其一，只要生活在一个城市中的居民，对于城市各个时段的交通情况其实有一个常识性的了解。即使没有 Peak Hours 功能，他们也会清楚地知道 Peak Hours 是什么时段。其二，司机群体可以分为 Full-time 司机和 Part-time 司机。Full-time 司机一直在线，因此是否是 Peak Hours 对其上线行为没有影响。Part-time 司机往往白天有固定的工作，即使有 Peak Hours 提醒，他们也无法放下手头的工作去开 Uber。这就导致，无论对于哪种类型的司机，是否有 Peak Hours 提醒都无法很大程度上改变他们的上线行为。

虽然这个测试结果基本否决了产品团队在时段这个方向上的努力，但它也启示了数据科学家们，时间可能并非是一个很好的探索方向。因而在此之后，数据科学家们开始更多地关注与位置有关的探索。比如他们会提前一周告诉司机，何时在哪里有篮球赛或是演唱会等大型活动，可能可以接到很多单子。同时，他们也开始一些创新的尝试，比如给空载的司机付费，让他愿意把车开到另一个需求更旺盛的地方去搭载乘客，来实现运力的调配和优化。

通过以上案例和产品功能开发流程我们可以窥见，大数据分析在各个流程中的重要性不言而喻。从最初的需求痛点的发现和准确定位，到后期的产品测试与反馈分析，都离不开扎实过硬的数据分析功底。掌握数据分析的技能，从数据中提炼 insight，把握风云变幻的市场需求，必将是未来互联网等诸多行业的核心人才想要追求的。

【使用 Machine Learning 来预测 Airbnb 上的房屋价值】

Airbnb 利用机器学习的方法来降低开发成本，提高模型性能和工作速度。在这个案例中，将用一个具体的 LTV（Customer Lifetime Value）模型案例－预测房屋价值来说明。LTV 建模流程主要有 4 步：寻找相关特征；创造模型；选择模型；使用模型。

任何机器学习的第一步都是寻找相关特征。在以往的工作中，Airbnb 常常编写 Hive 查询语句来创造与寻找特征，但是这一过程往往很耗时，而且需要特定知识，共享性程度不高。为了高扩展性的应用，Airbnb 创造了 Zipline 特征仓库(feature repository)来高效地设计模型.这一仓库提供了各个层次的特征信息，例如，在房主、房客、市场等 。这使得数据科学家的工作更加得高效，令他们能灵活地调用或是创造特征。

在创造模型这一环节中，Airbnb 使用了 Python 里的 Scikit-learn 来进行机器学习。在创造模型之前，数据科学家们往往需要数据进行调整，以便更好地契合模型，这其中就包括了对缺失数据进行填补，对分类变量进行编码。一些开源工具例如 Scikit-learn，对创建数据流和如何转化特征，起到了关键性的作用。此外，利用数据流来开发原始模型，还能解决训练数据集和测试数据集之间的同步问题 。

在选择模型这一步骤中，Airbnb 主要运用的是 AutoML 框架。为了决定用哪个模型，往往需要衡量不同模型之间的利弊，例如复杂程度以及可解释性。 运用不同的 AutoML 框架可以节省选择模型的时间，加快工作流程。例如，数据科学家发现，XGBoost 模型相比其他一些基准模型在预测房屋价值方面表现更加，拥有更好的灵活性。

在使用模型的环节中，Airbnb 创建了一个名叫 ML Automator 的框架来自动翻译 Jupyter 上的笔记到机器学习的工作流程中。这一框架为那些熟悉用 Python 写模型，但不熟悉数据工程的数据科学家提供了方便。首先，这一框架要求使用者在 Jupyter 笔记

中指定特定模型，来告诉系统哪里定位训练数据集，如何计算参数等。接着，数据科学家要编写特定的转化函数来定义如何完成对数据的训练。当笔记本上的内容融入到框架后，训练模型就被存入 Python UDF 中，然后制造 airflow 工作流。随后，数据序列化，定期的模型更新，计算方法更新等日常批量处理工作都可以在这一工作流中进行，就大大为数据科学家节省了开发模型的成本。

数据科学家求职准备

求职阶段所需掌握的全部技能

数据科学家作为当今最火热的职能，其需要掌握的基本技能是融合了多个学科领域的，包括数学、统计、计算机相关技能、以及很强的沟通能力与从事相关行业的 business sense。下面会从几个基本领域给大家罗列一些数据科学家需要掌握的 skill set：

硬实力：

Math & Statistics

- Machine Learning Model: Linear/ Logistic Regression, Decision Tree, SVM, KNN, K-means, PCA, Neutral Network, …
- Statistical Modeling
- Experiment Design
- Beyesian Inference
- Supervised Learning: Classification, Regression
- Unsupervised Learning: Dimension Reduction, Clustering, Outlier Detection
- Optimization (Regularization, Gradient Descent)

Programming & Database

- Basic computer science knowledge
- Basic algorithm (Sorting, Searching …)

- Scripting Language such as Python
- Statistical computing language such as R
- Databases SQL and NoSQL
- MapReduce Concepts
- Hadoop and Hive/Pig
- Spark
- Other popular tools such as Excel, AWS etc.

软实力：

Domain Knowledge

e.g. Natural Language Processing, Recommendation System, FinTech, Customer Behavior Analysis …

Communication

- Visualization skills
- Story telling skills
- Translate data-driven insights into decisions and actions
- Ability to engage with management or clients

KAGGLE 竞赛实战项目

Kaggle 竞赛是全球最大的数据建模和数据分析竞赛，也是全球范围内的数据科学家聚集与竞技的平台。数据科学家可以在该平台上参与由各大企业发布的竞赛项目，通过分析和建模

来解决企业急需解决的问题。数据应用学院的学员在完成基础知识的学习后，会在专业导师的带领下参加经由老师认真挑选的**具有较大商业价值，且当期正在进行的 Kaggle 竞赛项目**，真正通过实际的应用来巩固所学习到的知识。

截止到目前，数据应用学院辅导的学员在 Kaggle 竞赛中取得了很好的成绩，多次拿到单个项目的前 3%的名次，并在 2016 年 8 月取得了一枚宝贵的竞赛金牌！优秀的 Kaggle 竞赛过程将会成为学员的应聘时的一段宝贵经历。

HIGH-PROFILE 实战项目

Kaggle 竞赛项目毕竟和实际商业环境下的项目有所区别，仅仅通过参与 Kaggle 竞赛并不能帮助我们在开放命题条件下对整个商业模型和数据产品研发流程有所理解。 因此为了弥补 Kaggle 竞赛项目的不足，学院还为学员们提供了多个具有实际应用意义的数据应用项目，每个项目都经过了精心设计，可以作为学员面试时的 Demo 项目进行展示。

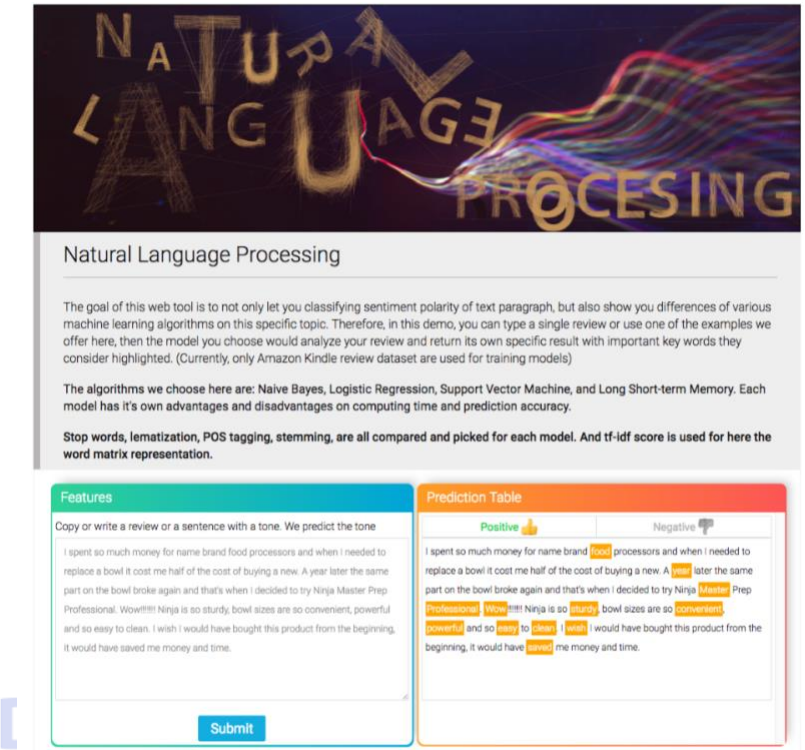
实战项目 I

NLP (NATURAL LANGUAGE PROCESSING) PROJECT （电商网站用户评价商业价值挖掘）

我们在电商平台购买商品时，通常会阅读其他购买人的评论来得知评价者对于商品的评价是好评还是差评。然而如何通过机器的自然语言识别自动识别一段文字的

情感评价，实现从数据到结果的自动化分析输入？如何借此挖掘电商网站近 20 年用户评价的价值？如何将 NLP 的潜在商业价值转化为现实的商业收益？在我们的 NLP 项目中，我们会通过结合不同的机器学习算法设计一项产品来帮助我们实现这一功能。我们将要设计的产品不仅仅可以实现对评论的情感评价，同时也会对其中的关键词进行高亮，并且通过简单的展示页面实现产品与用户操作上的交互。

- 抓取 Amazon Review Dataset 作为模型训练数据库
- NLP 处理流程的进阶学习与应用
- 多种机器学习算法的对比与评价(Logistic Regression, Support Vector Machine and Perceptron etc.)
- 网页用户交互界面设计与产品展示
- 产品实现对一段文字的情感评价与关键词提取



实战项目 II

FINTECH (FINANCIAL TECHNOLOGY) PROJECT （FINTECH 智能投资顾问）

通常情况下，Lending Club (美国 P2P 借款机构)中包含了成百上千的贷款项目，让投资人难以进行选择。在我们的 FinTech 项目中， 我们会使用过去所学的知识来设计一款智能投资顾问的数据产品，通过机器学习技术帮助投资人在

Lending Club 中鉴别项目的价值，以确定最优项目来进行投资。当新的贷款项目进入平台后，我们的产品会自动分析项目的各项指标，从而筛选出最佳的投资项目。我们还会设计简单的产品展示页面，实现产品与用户操作上的交互功能。

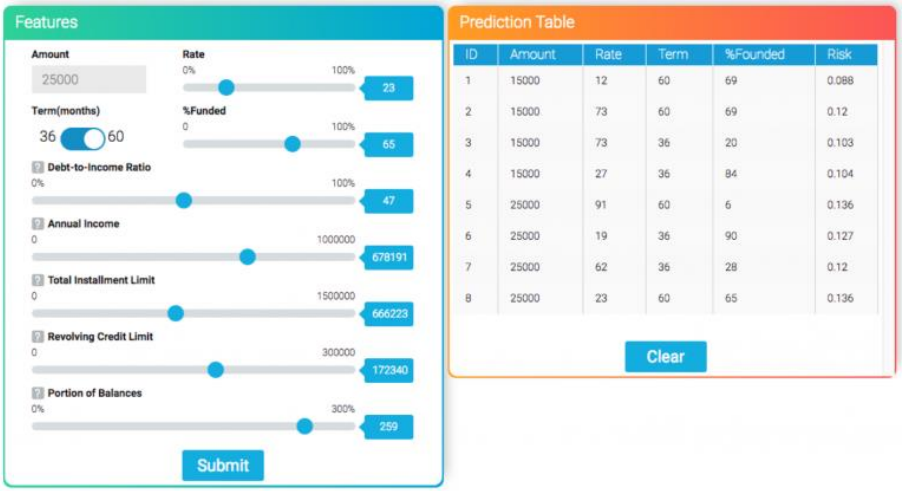
- 1,320,000+条大量数据的处理和 100+数据特征的筛选
- 通过 Gradient Boosted Regression Trees (GBRT) 算法构建机器学习模型
- 网页用户交互界面设计与产品展示
- 产品实现在 Lending Club 平台下对项目的评估与最佳投资项目的选择

FinTech-Intelligent Personal Investment Consultant

The centerpiece of model's entire operation is using algorithm to choose which notes to invest in. Often investors have to choose between hundreds or thousands of available loans at Lending Club. This model makes the process easier by using machine-learning to calculate which notes are more likely to perform better than others. The moment new loans are added to the platforms, the algorithm analyzes the variables of these loans and only invests in the best ones. The entire process, again, takes a split second.

The algorithm used is Gradient Boosted Regression Trees (GBRT). The advantages of GBRT are: Natural handling of data of mixed type (= heterogeneous features), Predictive power, Robustness to outliers in output space (via robust loss functions).

Five key features was picked according to the feature-importance scores: Debt-to-Income Ratio, Annual Income, Total Installment high credit/credit Limit, Revolving Credit Limit, Portion of Balances.

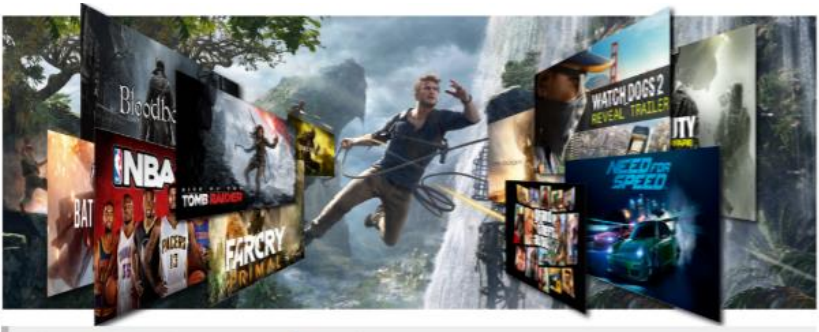


实战项目 III

GAME RECOMMENDATION SYSTEM PROJECT

推荐系统近几年发展十分火热，几乎所有的电子商务、社交网络、购物平台等都在不同程度上使用了推荐系统，在游戏平台中，推荐系统也是不可缺少的部分。在我们的游戏推荐系统项目中，我们基于 Steam 平台，对用户过去的游戏信息进行分析，根据游戏的受欢迎程度设计了推荐系统，为用户进行游戏推荐。用户同时也可以通过选择感兴趣的类别对结果进行过滤，对推荐结果进行进一步优化。学员将从产品定义、数据爬取、数据导入、数据分析、推荐系统平台设计、效果评估等方面，完成一系列完整的高水准产品研发过程。

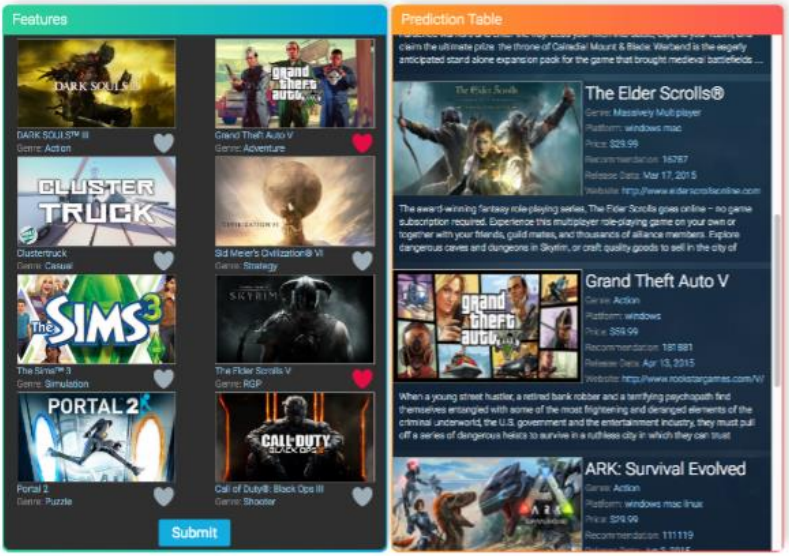
- Steam 游戏平台的数据抓取
- 300+的特征处理与筛选
- 基于 Popularity-Based Recommendations 算法实现对用户的游戏推荐功能
- 网页用户交互界面设计与产品展示



Steam Game Recommendation System

Game recommendation system changed the way the game vendors communicate with their gamers. Rather than providing a static experience in which users search for and potentially buy games, this game recommender system increase interaction to provide a richer experience. This recommender system identifies recommendations autonomously for individual users based on past their past playtimes, and on other users' behavior on the steam platform.

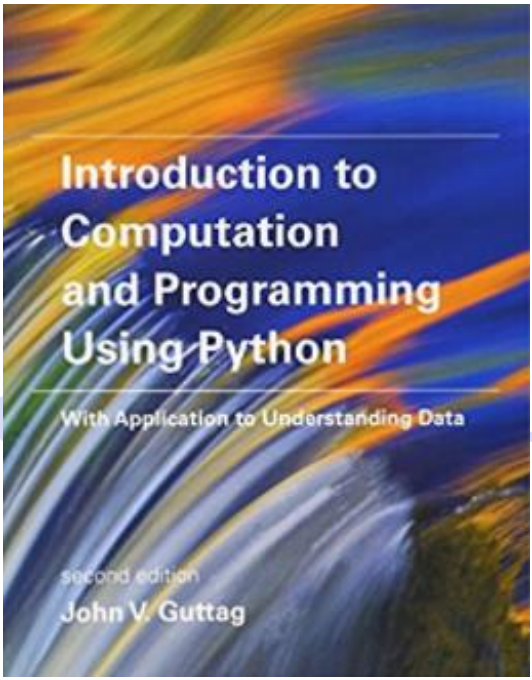
The approach used to recommend games is popularity based. Basically the most popular games would be recommended for each user since popularity is defined on the entire user pool. So everybody will see the same results.The user could filter their selection by choosing their own combination of categories: Action, Adventure, Casual, Strategy, Simulation, RPG, Shooter, and so on.



学习资料推荐

Python:

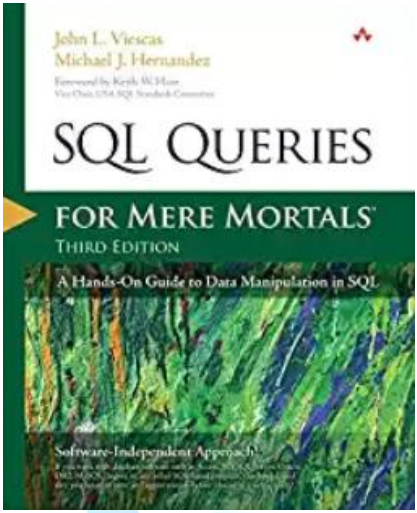
Introduction to Computation and Programming Using Python: With Application to Understanding Data (MIT Press) second edition Edition



R:
R for Data Science: Import, Tidy, Transform, Visualize, and Model Data 1st Edition

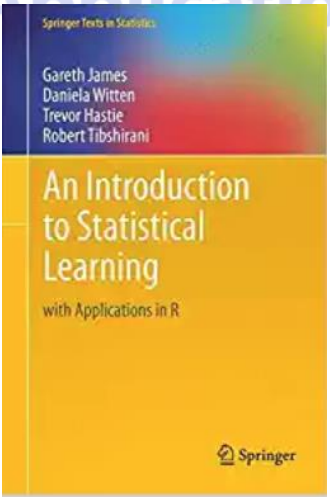
SQL:

SQL Queries for Mere Mortals: A Hands-On Guide to Data Manipulation in SQL (3rd Edition)



Machine Learning:

An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) 1st ed. 2013, Corr. 7th printing 2017 Edition



时间安排

1. 规划职业
2. 知识技能准备
    a. 各项技能网课 (2-4 月)
    b. 查漏补缺 (1 月)
    c. 制作项目(1-2 个月,精细制作)
3. 参加 Bootcamp, Kaggle 比赛
4. 面试
    a. 简历修改
    b. 不要海投
    c. 公司 research
    d. 面试
        i. HR 环节
        ii. 技术性面试
5. 根据面试反馈再一次查漏补缺, 修改简历
6. 反复, 耐心
7. 拿到 offer

常规: 2-12 个月获得满意 offer

如何获得数据科学面试？

数据科学面试过程的第一步不是面试，而是获取面试。争取面试机会这个过程本身可能需要几个月的努力！

九种获取数据科学面试的方式

我们发现传统的找工作面试途径在数据科学求职中一定程度上可以起作用。我们还发现了新的，主动的方法，特别是新兴的初创公司，非传统战术可以让候选人站在招聘竞争的前线。

分类一：传统的工作面试途径

1 | 数据科学工作公布平台和正常的公司网上申请

您可以向公司招聘网站上提交简历和求职信。然后，你可以等待和希望。我们不是说要避开这条路线，但不应该是你所依赖的唯一渠道。

使用 Indeed，Careerbuilder，Monster，和 LinkedIn 来搜索不同的数据科学信息。那么，有数据科学的特定招聘版块，如 Kaggle Jobs Board（2017 年被 Google 收购，全球最大的数据科学竞赛平台，聚集了众多数据科学人才）。

另外，对于还在学校的同学们，一定，一定，要利用好学校 Career Board，校友网络，LinkedIn 校友 Group 里的招聘内推信息。

2 | 与猎头合作

您可以联系招聘人员，帮助您与合适的雇主联系。有专门从事数据科学的猎头。他们手头有很多未对外公布的工作信息资源。通过 LinkedIn 搜索附近的数据科学猎头能快速帮助您找到最相关的人选。

3 | 去招聘会

数据科学的招聘会相对传统招聘会来说数量少很多，最好能多参与当地的活动(local meetup)。例如，Southern California Data Science Conference（https://www.ideassn.org/socal-2017/）为学生提供了大量的数据科学工作。

分类二：积极的工作面试路径

我们已经涵盖了传统的工作面试途径，这些选择是求职者的默认选择-“基础配置”。如今，拿到工作 offer， 有时候需要更多的主动争取和磨练。初创公司提供大量新的数据科学工作。他们的文化和招聘策略吸引了十年前也是创业公司的大型公司 。结果是在一个新的招聘环境，通常情况下，一个人必须积极主动地接触那些在建立自己的公司的决策者 。

4 | 参加或组织数据科学活动

您需要找到对数据科学界感兴趣的人才能找到隐藏的机会，并积极主动地融入社区。有几个事你可以做到这一点，从大型会议到小型人才聚会。

4.1 会议

Strata 会议（https://conferences.oreilly.com/strata）: Strata 会议是在不同城市举行的全球数据科学大会。演讲者来自学术界和私营企业; 这些主题围绕着数据科学趋势的发展。会议使你能够了解数据科学背后的技术，并有大量的 networking 活动。

KDD（Knowledge Discovery in Data Science）: KDD 是另一个大型数据科学会议。它也是一个组织，旨在引导数据科学背后的科学的讨论和教学。这些会议的成员和出席率为数据科学日益增长的趋势做出了巨大的贡献。

Data Science Association （https://www.ideassn.org）：2016 年 9 月，在美国洛杉矶主办了一次南加州第一届数据科学高峰论坛，并邀请了 25 位来自工业界、学术界、政府部门的数据科学界杰出演讲嘉宾。 本次大会是由数据应用学院(Data Application Lab)和数据科学协会(Data Science Association)主办的数据科学行业交流大会，是美国为数不多的以华人占主导的行业性数据应用大会。本次大会的到场的 25 位嘉宾包括来自微软，LinkedIn, IBM 等大型企业和学校，政府的高级数据科学家和顾问，从数据安全，云计算，数据营销，数据可视化等角度展示了数据科学在各行业的应用和发展，收到来自产，学，研三方的高度关注与积极参与，聚集了超过 800 的数据科学从业者和爱好者。2017 年 Southern California Data Science Conference 将在 10 月 21-22 日在洛杉矶举办。

此外，在中国范围内，还有数据库中国技术大会、中国数据挖掘会议、科学数据大会等。

4.2 Meetups

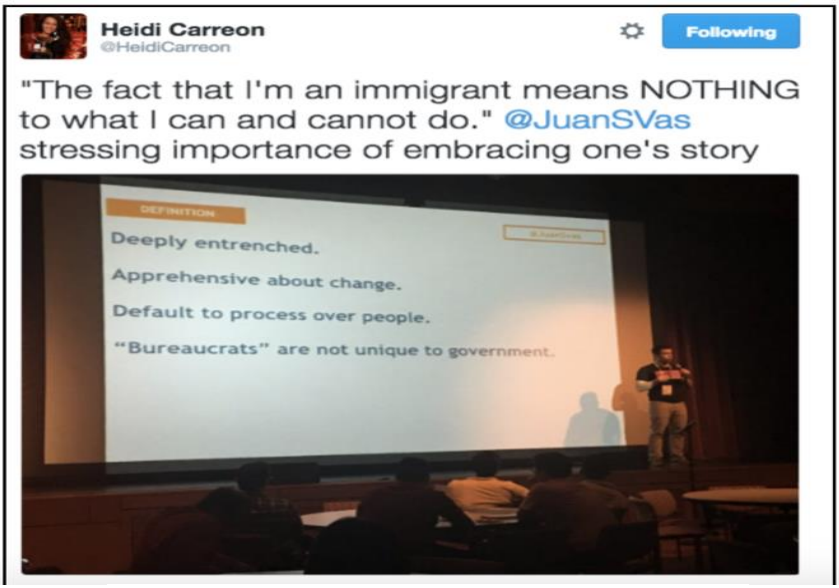
我们已经列出了数据科学界所在的主要会议，但是通常会有更小的聚会来连接当地的数据科学家。

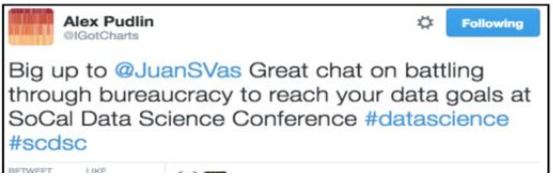
旧金山湾区往往拥有最多的数据聚会，尽管美国的每个主要城市通常都有一个。您可以通过 Meetup.com 查找您附近的数据科学会议。一些最大的数据科学会议，拥有 4000 多名成员包括：SF 数据挖掘，数据科学 DC，伦敦数据科学和 Bay Area R 用户组。

如果您找不到附近的活动，您也可以自己创建一个聚会。我们的数据应用学院的创始人 Jason 就是通过组织数据科学公益讲座，交流小组，成为数据科学 social connector，认识了很多志同道合的朋友，从而决定成立了 “数据科学人才项目孵化器”-Data Application Lab。

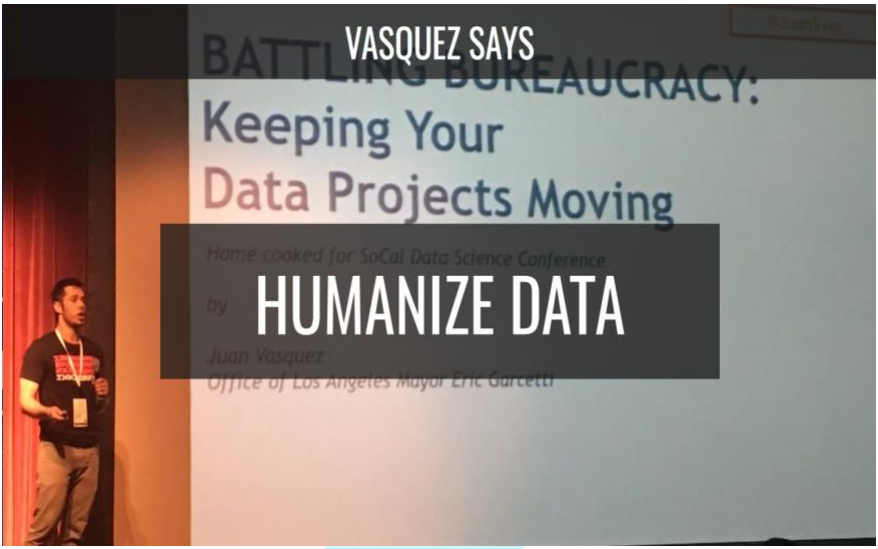
5 | 自由职业 Freelance 和构建项目案例

Juan Vasquez，Data Application Lab 的老朋友，是加州洛杉矶市长办公室的数据顾问和数据创新领导者。 Juan 本身是来美国求学的新一代移民， 坚信一个人他能做什么或不能做什么，并不会被他的移民身份所定义，他充满激情，有独特的 story-telling 魅力。在加入政府之前，他不认为自己是传统的数据科学家，而自称为“数据科学倡导者“， Juan 将其 story-telling 特长结合数据领域的应用发现，积极向公众传播，成为“数据领域的知识网红”，不久之后他被聘请加入了洛杉矶市长办公室成为数据创新青年领袖。





在南加数据科学大会上，Juan 和我们分享和讨论了，“如何运用数据科学提高行政系统的效率”。



美国的行政系统效率低下，饱受诟病多年，连最近的迪士尼动画片《疯狂动物城》都要调侃美国车管局，将其职员刻画为树懒，讽刺其效率低下。提升行政效率可谓是美国民众的迫切愿望。据 Juan 介绍，提升行政效率不是一件简单的事情，而需要提高的行政组织系统，并不仅有政府，很多具备规模的组织内部，例如，银行，医院，教育等部门，都需要提高他们的行政系统。

这些挑战让数据顾问 Juan Vasquez 的工作显得异常艰巨。在担任市长办公室数据顾问后，Juan 发现，洛杉矶郡的土地管理面临的一大障碍是，长久以来一直使用 Google sheet 来记录和管理土地使用情况，空洞而缺乏整体感，将这样的管理形式转化为在地图上呈现各个区域的土地使用情况，极大降低了使用该系

统的学习成本。数据科学技术在公共领域的小小改进和应用，确实能够让决策者在极短的时间内理解数据表达的意图，大大提升了交流沟通的效率。

从一个自由职业者到通过自己努力研究转型成为数据科学创新领袖，Juan 的经历告诉我们在这个持续发展的新领域，只要你有想法，有能力，不需要大公司的平台一样能探索数据科学的魅力。

如果您对数据问题感兴趣， 我们建议将你喜欢的，并把制作出的了不起的解决方案和作品集记录存储在一起（比如 Github），建立你的 Portfolio， 向大众讲述你的激情和故事。

6 | 参与开放资源和开放数据项目

世界上最有趣的项目不一定存在于公司的秘密数据库中。他们经常在 Github 的开源存储库中。这包括自然语言工具包（Natural Language Toolkit）项目，它可以帮助处理人类语言作为数据源以及构成 Python 数据科学和机器学习工具包的各种库。R 社区还在综合公共网站上托管了许多数据包。

许多领先的 CTO 将根据您对开源项目的贡献聘请，甚至可以通过该路线找到您。很容易判断一个人是否能够在一个团队中工作，并透过透明的开放源源创造出奇妙的东西。确保你可以充分利用资源发挥技术实力。

如果您不确定要分析哪些数据，以下是我们列出了您可以浏览的美国 19 个免费的开源公共数据集。

United States Census Data: https://www2.census.gov

如果你对社会议题感兴趣，关注经济民生的发展，美国人口普查部门（United States Census）的数据是一个可以优先考虑的全面数据集。

美国人口普查部门发布了包括国家，城市甚至邮政编码层面的人口统计数据。数据集对于创建地理数据可视化是非常棒的资

源，可以直接通过在人口普查网站上访问。或者可以通过 API 访问数据。使用该 API 的一种方便的方法是通过 chlorplethr。总的来说，这些数据非常干净而全面。

FBI Crime Data: https://ucr.fbi.gov/crime-in-the-u.s

看多了美国 FBI 相关的美剧，探索一下美国联邦调查局的犯罪数据集是不是听起来也很有趣？

FBI 上收集整理了 1995 年到 2016 年的数据，如果你有兴趣从时间唯独分析，可以尝试统计 20 年以上国家级犯罪率的变化，或者，你也可以在地理纬度上探索数据。

Centers for Disease Control and Prevention （CDC）：

https://wonder.cdc.gov

如果你对公共卫生和预防医学领域数据分析感兴趣，疾病控制中心（CDC）会是一个很不错的选择。

在 CDC Wonder 数据库里，存储了有关死亡原因的数据库，AIDS 艾滋病预防数据，空气污染和健康相关数据，空气污染颗粒数据等，数据分段包括：年龄，种族，年份等。

Medicare Hospital Quality:

https://data.medicare.gov/data/hospital-compare#

这些是医疗保险和医疗补助服务中心（Centers for Medicare & Medicaid Services）提供的官方数据集，可以用来比较美国超过 4000 所通过医疗保险认证医院的护理质量。

Bureau of Labor Statistics: https://www.bls.gov/data/

学经济的同学们，敲黑板，注意啦。

美国劳工统计局网站上可以找到美国的许多重要经济指标（如失业和通货膨胀）。大多数数据可以按时间和地理分段。

The Bureau of Economic Analysis: https://www.bea.gov/national/index.htm

美国商务部（US Department of Commerce）下面的经济分析局也有国家和地区的经济数据，如 GDP 和汇率。

IMF Economic Data: http://data.imf.org

和以上两个不同，如果感兴趣查看国际经济数据，比如汇率、利率、GDP，商品交易数据等，可以在 IMF 网站上找到。

Dow Jones Weekly Returns: http://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index

预测股价是数据分析和机器学习的主要应用。对金融投资市场感兴趣的朋友可以要研究的一个由美国加州欧文分校（UCI，统计大牛校）-- 提供的数据集，关于道琼斯指数的每周收益。

Boston Housing Data: http://www.kellogg.northwestern.edu/faculty/weber/emp/_session_3/boston.ht

Data Application Lab 也曾用加州数据，通过 GIS 技术，基于地理信息数据和回归模型，综合考量房屋特性，比较优势，地理位置等，对房产房价做了预测系统。

Enron Emails: https://www.cs.cmu.edu/~./enron/

Enron 崩溃后，发布了大约 50 万封邮件文本和元数据的电子邮件数据集。数据集现在是着名的，为文本相关分析提供了极好的测试基础。它具有现实世界数据的混乱。

Google N-Grams

如果您对真正大量的数据感兴趣，Google n-gram 数据集会在大量文本来源之间按年份计算单词和短语的频率。得到的文件是 2.2 TB。

Lending Club: https://www.lendingclub.com/info/download-data.action

Lending Club 是全球最大的 P2P（个人对个人）借贷公司，提供有关拒绝的贷款申请及其发行的贷款业务的数据。数据集本身

既适用于分类技术（将给定贷款违约）以及回归（对特定贷款支付多少）。

Data Application Lab 曾通过机器学习技术帮助投资人在 Lending Club 中鉴别项目的价值，以确定最优项目来进行投资，设计出了一款智能投资顾问的数据产品。当新的贷款项目进入平台后，我们的产品会自动分析项目的各项指标，从而筛选出最佳的投资项目。我们还会设计简单的产品展示页面，实现产品与用户操作上的交互功能。

Walmart https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

连传统零售行业沃尔玛在 Kaggle 比赛上也发布了包括 45 家门店和 98 个商品级的销售数据。这是一个很好的时间序列分析数据，并且呈现有趣的季节性因素影响，感兴趣的朋友可以一探究竟。

Airbnb https://www.airdna.co

Airbnb 发布用户会话数据作为内容的一部分，以创建分析和可视化。

Yelp https://www.yelp.com/dataset_challenge

Yelp 发布了一个学术数据集，其中包含了 30 所大学范围内的信息。

如果关注中国的开放数据，我们推荐基金会中心网 CFC：http://data.foundationcenter.org.cn/foundation.html

基金会中心网由国内 35 家知名基金会联合发起，于 2010 年 7 月 8 日正式成立。基金会中心网的使命是建立基金会行业信息披露平台，提供行业发展所需的能力建设服务，促进行业自律机制形成和公信力提升，培育良性、透明的公益文化。在过去的五年里，基金会中心网与国内外诸多知名基金会和社会组织建立了良好的合作关系，包括美国盖茨基金会、福特基金会、赠与亚洲、亚洲基金会、洛克菲勒基金会、德国宝马 BMW 基

金会、粮惠世界、墨卡托基金会、日本基金会、丰田基金会、美国基金会中心、欧洲基金会中心等等；同时也与国内外的一些知名大学建立了良好关系，包括哈佛大学、斯坦福大学、印第安纳大学、清华大学、北京大学、北京师范大学、浙江大学等等。

基金会中心网秉承使命，已经基本成为国内最具影响力的信息披露平台，在倡导慈善数据的应用方面发挥了一定作用，推出了基金会透明标准中基透明指数 FTI，有效地推动了基金会行业整体的透明度发展；建立良好的公共关系体系

Data Application Lab 于 2017 年暑假和基金会中心网、IBM 企业公民与企业事务部、DAL 数据应用学院发起，联合清华-青岛数据科学院、iJoin 社会创新咨询 x 粟米科技、可道、北京大学工学院研会、北京科技大学计通学院、贵州大学明德学院等社会 NGO 和高校共同举办 行了中国公益基金会数据骇客松。

大赛以 2017 年中国慈善周系列活动为载体，旨在利用数据科学技术，精确推荐捐助单位与基金会对接，量化公益事业的社会舆情关注，推进中国公益事业健康发展，帮助公益基金组织提高效率，吸引高校人才进入公益事业，吸引社会企业对公益基金会的关注，强化社会大众对公益事业的信任。

7 | 参加数据科学竞赛

如果开放源码项目的广泛范围不是您的项目类型，并且您的创造力在更为狭窄的情况下可以获得最佳成就，请考虑加入数据科学竞赛。

像 Kaggle，Datakind 和 Datadriven 这样的数据科学竞争平台让您能够处理真正的企业或社会问题。通过使用您的数据科学技能，您可以展示您的变革能力，并创造出最强大的面试资产。

我们 Data Application Lab 的金牌导师之一，Chris 导师，辅导 Data Application Lab 的同学取得 Kagge 比赛 “1 金牌 4 银牌 9 铜牌”的优异成绩。

8 | 和业内人士约咖啡，做信息采访（Informational Interviews）

是否想过，其实你的人际网络将为你提供新工作的最佳机会。你应该在你想要工作的领域寻求更多人的帮助，去了解他们有问题，是否可以帮忙解决。当然，因为大多数数据科学家都很忙，不一定有机会见你，所以 在接下来的章节里讲为你提供约见繁忙的数据科学家的小技巧。

你必须找到一种提供某种价值的方法，并能为他们面临的问题提供新的视角和解决思路。

这种类型的见面聊天可以转变成 Information Interview，我们所说的信息面试，不仅可以向在该领域的数据科学家寻求建议和信息，很多时候，欣赏你的数据科学家前辈会很乐于给你提供就业指导，甚至给你推荐可能的工作机会。 如果你这样有效的持续运作，你将不断扩大你的数据科学机会网络，也将更多地了解数据科学在行业中的运作。

9 | 数据骇客马拉松

根据 Learn by doing 的指导方针，数据骇客马拉松为你提供了一个独特的机会，你将在几天内和积极多才的团队合作，学习数据技能，解决数据问题。

这种数据骇客马拉松的一个例子是 Data Application Lab 在 2017 年 8 月举行的中美公益数据竞赛。 通过与他人合作提供真正的解决方案，你将在其他求职者中脱颖而出。许多雇主也期待 hackathons 帮助筛选人才，一些公司甚至赞助 hackathon 奖，希望找到他们的下一个数据科学家！

如何打造一份无法拒绝的简历

不合格简历：

Shuju Wang
Tel: 8181881818; Email: Jason.Geng@gmail.com (your English name and address)

EDUCATION BACKGROUND
The University of Texas at Dallas
May 2016
M.S., Information Technology Management
Nanjing University of Posts and Telecommunications
Jun 2014
B.S., Telecommunication Engineering

BUSINESS EXPERIENCE
XXX, Inc.
Jun 2015 - Sep 2015
Business analyst (you can adjust title, sequence of bullet points is also important)
- Generated daily and monthly sales summary report, produced payrolls, in order to develop quotation system and improve unloading efficiency (no result)
- Increased seasonal sales by 30% which covering over 150 different products and reached $850K/month (no method)
- Priced company's new products for North American market based on production costs, taxes, competitor's prices, etc.
- Created a spreadsheet to rate the importance of different requirements, in order to increase number of customers (should know what and how to highlight the importance part)

XXX, Inc.
Feb 2015 - May 2015
Industry operation analyst
- Developed drawing confidential system via analyzing usage frequency data which led to shorten the quoting and enquiry cycle
- Enhanced employee accountability through expending data collection range and method (should highlight your related skills and result)
- Optimized inventory space by getting rid of overstock and making profit from over 80% of excess inventory (express result in a simple and direct way)
- Developed and improved cost analysis system for contractor project and customized project through renewing classification methods

XXX, Co., Ltd.
Nov 2014 - Jan 2015
International market analyst
- Engaged in the company's Russian Market Development Project, successfully opened the gateway of exporting Chinese trucks into Russian Market. Russian Market becomes the company's largest overseas' market since then
- Processed orders for exporting automobile accessories to the company's subsidiary in the US

(if you don't have enough technical skills in this experience, why not mention sth else? Should also mention you people and management skills)

ACADEMIC PROJECT (you are a professional, do not mention too much about university)
Derivative Modeling and Risk Analysis (you are always the leader)
Jan 2016 - May 2016
- Classified income statement and balance sheet based on month from annual reports
- Used MATLAB for pricing the compound options by using Monte-Carlo Simulation
- Performed Monte-Carlo Simulation to improve accuracy of the result
- Analyzed data and computed financial ratios, profitability and risk ratios by Excel

(describe your project title and experience in a way the employer will understand)

Database Analysis Feb 2015- Apr 2015

修改后简历：

Shuju "Data" Wang

2901 S King Dr. Apt. 1918, San Jose, CA 94401| Cell: (818) 188-1818 | Jason.Geng@gmail.com

EDUCATION

The University of Texas at Dallas, Dallas, TX — August 2014 – May 2016
M.S., Information Technology Management, GPA: 3.8
Candidate of PMI Certified Associate in Project Management (CAPM)
Relevant Courses: Exploratory Analysis, Prescriptive Analytics, Database Foundation, Statistical Modeling and Inference, Prediction and Machine Learning

Beijing University of Posts and Telecommunications, Beijing, China — September 2010 – June 2014
B.S., Telecommunication Engineering, GPA: 3.6
Relevant Courses: Mathematics, Linear Algebra, C++, Probability, Data Structure, Database, Image Processing

SKILLS

Software: MatLab, Python, Tableau, R Programming, MySQL, MS Office: Excel, Project, Visio and Access
Language: Mandarin

WORK EXPERIENCE

XXX, Inc., Dallas, TX
Business Analyst — June 2015 – September 2015

- Gathered various requirements, analyzed current and future state, and performed GAP analysis to develop quotation system, increasing seasonal sales by 30% with $850K/month
- Built analysis model to improve inbound arrangement and scheduling for 150 different products, increasing unloading efficiency by 40%
- Priced company's new product for North American market, increasing the product's profit by 27%
- Increased number of customers by 5% via creating a spreadsheet to rate the importance of different requirements

XXX, Inc., Dallas, TX
Industry Operation Analyst — February 2015 – May 2015

- Analyzed usage frequency data to develop drawing confidential system, reducing quotation & enquiry cycle and increasing Product Qualification Ratio by 20%
- Managed data collection range to enhance employee accountability, decreasing defective rate by 27%
- Optimized warehouse utilization and maintained low inventory holding cost, saving total cost by 13%
- Developed and improved cost analysis system for contractor project and customized project through renewing classification methods, increasing project probability by 30%

XXX Co., Ltd., Beijing, China
International Market Analyst — November 2014 – January 2015

- Improved the accuracy of obtaining customer requirements by 10% via creating product design, planning and control matrices
- Built a cost baseline and calculated out task cost via Three-point estimation for forecasting & controlling
- Figured out critical path via PERT table & MS project, facilitating project progress and saving resource cost by 8%
- Performed scenario & sensitive analysis for process improvement, analyzed potential risks, and assessed performance using visual tools

PROJECT EXPERIENCE

Derivative Modeling and Risk Analysis Project
Team Leader — January 2016 –May 2016

- Conducted a quantitative structuring analysis to customize compound options for various clients
- Programmed MATLAB code for pricing the compound options by using Monte-Carlo Simulation, CRR-binomial lattice and implicit finite difference method
- Performed Monte-Carlo Simulation by using variance reduction techniques, such as control variates, conditional Monte-Carlo and Antithetic variates, and trinomial tree to improve accuracy of the result
- Performed a risk-return analysis by stimulating the mean of profit, return to investment, and excess return; sharpe ratio, and 95% value at risk

Database Analysis Project
Team Leader — February 2015 – April 2015

- Conducted WBS to define the project scope and specification. Monitored the whole process to ensure the project advance on time each month by writing status reports, milestones and Gantt charts
- Wrote queries and tested for errors by using SQL Server, executing an advanced program result
- Draw entity relationship diagram and relational database schema vis Access, getting a complete logical relation

如何获得招聘官的青睐

通过我们之前的科普，大家应该有了很多接触数据科学的机会。然而距离成功的拿到最后的数据科学 offer 还任重而道远。这部分我们将结合包括 IBM 在内的北美各大名企的招聘官们的指导、着重介绍一些数据科学求职的小贴士。

数据科学求职的 Full Cycle 如下图所示：

- 数据科学求职的 Full Cycle -

我们将会针对每一个环节，给大家带来最详尽的介绍。

第一步：职业评估和规划

如果你正在认真的阅读的这本数据科学白皮书，那么你在这一环节上已经迈开了第一步。正确的认识自己的优势，并准确的匹配行业所需的技术，才能保证求职的顺畅和成功。

同样是数据科学领域，不同的角色将对不同的技能组具有非常不同的期望。比如虽然数据工程师（Data Engineer）可能不会有很多业务展示 business presentation 技能，但他们被期待有能力主导所有类型的编程挑战。相反，数据分析师（Data Analyst）将更多地倾向于他们的 SQL 技能，而不会面临非常挑战的技术问题，但往往他们需要具备最棒的表达和沟通能力。下表阐述了行业的需求和不同级别的难度，大家可以对照自己的特长，来选择最适合自己的职业发展道路。

职位 Position Title	**数学／统计知识 Math/Statistics（比如 P-value 分析，AB testing**	**数据库 Database Querying (SQL)**	**算法 Algorithms（eg Supervised learning, Entity Resolution）**	**软件工程 Software Engineering（Python, Java, Object Oriented）**	**大数据／系统工程 Big Data/Systems Engineering (Spark, HBase, Hadoop)**
Product Data Scientist 产品数据科学家	★★	★★	★★	★★★	★★★
Data Engineering 数据工程师	★	★★	★	★★★	★★★
Data Scientist 研发数据科	★★★	★★	★★★	★	★

学家					
Business Intelligence Data Scientist 商业智能数据科学家	★★	★★★	★★	★	★
Business／Data Analyst 商业数据分析师	★	★★★	★	★	★

- 不同数据科学岗位的技能要求侧重点 -

以下是具体职位角色的总览介绍：

产品数据科学家：具有数据工程技能的，能实现端对端（end-to-end）工程产品开发的数据科学家。产品数据科学家领导团队建立数据产品。他们调整算法，并对数据如何为最终用户提供服务有最终发言权。他们经常有工程能力来实现这些想法。相较于其他数据科学岗位，产品数据科学家对专业要求最高，需要具备很多难以获得的技能。

数据科学家：技术技能，商业技能和数学知识的独角兽组合。数据科学家了解如何创建和优化数据算法，以及如何解释其发现。他们可能需要比他们的数据工程师同行知道更少的编程，但是他们仍然需要了解处理大规模数据的方法。

商业智能数据科学家：商业智能数据科学家专注于从企业数据中挖掘价值。和数据分析师不一样的是，他们将足够了解统计方法和不同的机器学习算法 。他们能建立 Dashboard 并完成各种分析研究，以帮助各个团队做出更好的决策。

数据工程师：数据工程师常常不许具有对统计学和数学方面的高级知识，但是他们必须在每一个技术挑战的前提下，证明他们熟练掌握可以处理大量数据的算法。
数据分析师：一个入门级角色，通过查看数据和解释结果，很大程度依赖于制定可视化报告。这个角色通常需要很强的 SQL 和 Excel 能力，如果有数据可视化能力更佳。

第二步：职场沟通和人脉拓展

有效率、准确的职场沟通和人脉积累可以帮助求职者快速的打开局面，深入行业核心。

NETWORKING，让你的人脉关系网络为你内推工作

不论是在国内还是国外的求职过程中，强大的关系网都是一个求职者最有力的坚强后盾。在美国文化里的 networking 就有拓展人脉网的意思。通过 networking 来获得内推(internal referral)是美国找工作的特色之一。虽然印象中中国更是一个“关系型社会”，然而美国也是非常看重个人推荐，因为美国是信任感很强的国家， 招聘官也更倾向于招聘有熟识的同事，朋友提供背书的求职应聘者。

大多数人没有意识到建立和维护你的人脉网络对你的数据科学求职是至关重要。招聘公司很看重强力推荐，特别是内推。如果你办法在你申请的公司内找到内部员工给你提供帮助和内推，

那么可以确保你的简历能够被查看，甚至可以让你在面试过程中跳过一些步骤！

我们非正式地调查了 Data Application Lab 通过招聘过程中的一些校友。事实证明，公司内部人员的转推荐中能有 85％的机会受邀获得电话面试，而那些普通直接申请的，通常只有 10％的机会有可能拿到面试机会。对比数据就可以发现，用前一种“找内推”的方法，能给你的求职成功率提高不止一个数量级。

积累人脉的有效方法

不给方法的讲道理都是耍流氓。对于刚出国留学人生地不熟的，也不像国内有那么多多年积累的同学朋友，如何有意识的积累人脉，寻求潜在工作机会呢？

第一，长远看来，和共事过项目的同学、同事保持良好关系，主动给其他人提供帮助，加深彼此的了解，体现价值，在你真正需要找工作的时候，有一个数据领域强大的人脉网络积累，很轻易就有朋友能为你介绍一些职位。

第二，可以多参加行业相关的大会去做志愿者服务，也能逐步积累资源。

第三，也可以使用所谓的信息面试（information interview）技巧。信息访谈是一个很好的机会去准确地了解公司正在发生的事情以及他们的优先事项。这在实际的面试中是非常有益的。大多数人，甚至是完全陌生的人，如果你表现出对他们正在做的事情真正感兴趣，并且愿意提供志愿帮助，那么他们会非常慷慨地为你答疑解惑、提供一些帮助。一旦你预约了信息面试，请

确保你提前做好准备。你需要查看公司网站和其他资料、对公司和聊天对象进行了研究，并能够大致了解公司的运营情况和一天中可能遇到的问题。

第四，通过参与线下活动（比如 Meetups，聚会或者专门针对 LinkedIn，Angellist 和 FounderDating 等）来约见网络用户。你可以诚实地表达你的意图，特别是你对特定公司和数据科学的兴趣。

除此之外，你也可以请求约一个 coffee chat。在聊天过程中，你可以在发表有趣的观点想法来解决他们正在解决的问题，或了解他们公司的问题。如果你做得很好，能成功地给对方留下一个好的印象并证明你可以帮助公司，那么你正在喝咖啡的人可以成为一个强大的内部推荐人，并帮助你跳过通常的招聘流程进入第一轮面试。

别忘了你还可以通过你人脉网络来结实更多人。你可以你的人脉介绍对你有帮助的人认识。比如，你可以通过任何 LinkedIn 公司页面找到该公司的职员。如果有些人恰巧是你的第二联系人，你可以看看他们是如何连接到你的。你可以联系你们的共同联系人来介绍你们认识。

礼貌又专业的职场沟通

和人脉建立了最初的联系之后，维系良好的关系和实时更进也很重要。在和行业里的有效人脉沟通时，要特别注重礼节和专业性。从 Thank you letter 到 follow up email， 每一个环节都需要做到简明扼要和有礼貌。

我们归纳了一些经典职场场景，和这些场景下可以用来参考的沟通模板：

场景一：LinkedIn/FounderDating/Angellist 添加好友请求

Hi [name],

I come across your information on XXXXXX. Your profile seems to be very interesting that we both 【common ground】. As a data enthusiast, I am highly interested in your XXXXXXX. Therefore, I would love to add you to my professional network on XXXXXX and have more discussion with you on 【Topic】. I am looking forward to hearing back!

Cheers,

[your name]

附上你的 LinkedIn 链接，简历，个人网站，最近的项目作品

场景二：信息面试（information interview）请求

Hi [first name],

I was super interested in the problems 【Company Name】is facing in data science. I've been aspiring to break into the field, and being a passionate follower of the 【Company-related blogs】, I noticed that building trust with data/ 【Other Issues with data】is an important part of what drives 【Company】 .Based on my background in psychology and statistics, I might be able to help come up with some creative ideas on 【Company's Issue】.

I'd love to take you out to coffee and get a greater sense of what problems 【Company Name】 has. Perhaps I can help! When do you have some spare time for a quick chat?

Cheers,

[your name]
Optional info: [Why are you interested in the company], [something the company has done that you love], [how you can help]...

场景三：寻求推荐/介绍

[Greeting],
[Talk about last point of contact], [talk about interest in company and problems faced by a specific role], [ask to be introduced to hiring manager to help solve those problems]
e.g.
Hi [first name],
It was great seeing you at the potluck! I've been looking around, and I'm interested in the problems Uber is facing, specifically the ones faced by data scientists on the growth team. Would you mind introducing me to the hiring manager or somebody on the team so I could see if I could help?
Cheers,
[your name]

场景四：面试后 Follow up

[Greetings],
[Ask your interviewer how they prefer to be addressed during the interview], [Talk about problems you can help solve], [State that you're looking forward to next steps]
e.g.
Hi [ask how your interviewer prefers to be addressed],
It was a pleasure talking with you about Google's data science problems. I think I can help with some of the problems you've enumerated, and I look forward to the next steps in the process!

很多人可能会有一个传统的观点，关于什么是一个好的工作申请流程 。可是他们已经失去了一个更重要的一点：时代在改变，领域在进化，过去的方法也许不再百发百中管用。

我们一直强调，学术界和工业界有着根本的区别，首先第一点，就是从如何展现自己开始。

我们和招聘人员，学生和招聘经理进行了交谈，他们都同意 LinkedIn 是招聘的黄金标准。拥有最佳的 LinkedIn 简介可以让雇主优先考虑你，让猎头帮你找到更匹配适合的机会。

如果你没有保证你的 LinkedIn 头像是一个专业充满能量的职业照片，那么你已经输在起跑线上了。

虽然大多数工作申请中还是要求提交简历，然而简历不再是让你拿到面试机会的主要原因。招聘人员只会当简历放在面前，才会仔细查看简历，而一个很棒的 LinkedIn 可能会持续帮助猎头和招聘官主动联系你，给你带来可能的入职工作机会。

关于 LinkedIn 的重要建议

1） 别害羞，别低调。

和简历不一样，LinkedIn 是一个平台可以让你补充尽可能多的细节，特别是通过多媒体文件的方式将你的工作影响呈现出来。大多数招聘经理在面试之前，都希望看到你的 LinkedIn，看到更丰富立体的人物形象。

2）确保你的工作职位名称和行业标准、招聘关键词是一样 。比如，说你曾经是“数据科学家”或“数据分析师”，会比自己创造出来的职位名称，如“数据痴迷爱好者”，“数据挖掘机”等更方面搜

索。当然，这些称谓也十分有趣能体现个性色彩，你可以权衡利弊。

3）从另一个方面来说，添加你的个性化元素，包括兴趣和爱好，并确保它们在你的 LinkedIn 中显而易见，也能让你和其他应聘者相比显得与众不同。招聘经理，会评估候选人的技术技能和文化匹配度。能够展现自己拥有独一无二的思考视角和见解，也会增加你的求职信息，帮助你脱颖而出。

比如，可以通过描述一些志愿活动经历，展现出你对该领域的热爱和激情。很多求职者就会将自己参加数据科学大会志愿服务、参加数据骇客马拉松的经历、或者自己对某本书的深刻解读和看法分享出来，也是很不错的尝试。

虽然很多求职者在找工作过程中也不断探索，对各种机会保有开放态度，不会设置非常明确的行业和职位。 但是你需要知道你正在寻找什么相关的工作机会，并确保在你的 LinkedIn 上体现出来。最好能精心的设计你的个人资料，以便帮助你得到你想要的职位。比如，如果你不想找初级的岗位，，请避免列出“数据输入”等技能。如果你对行业有非常明确的偏好和热爱，热衷于解决某类问题， 请清晰的描绘出来，利用“吸引力法则”。

4）确保你知道你正在申请什么职位，并将与之匹配的行业关键字和技能关键字列在你的申请文件中。如果你对金融领域的数据科学工作感兴趣，请毫不犹豫地将行业术语放在你的简历和 LinkedIn 上。你也可以研究公司使用的技术，像 Yelp 和 AirBnB 这样的公司经常会在博客上展示他们相关的数据项目。如果你有一个你所研究的技能正好匹配你正在寻找的工作，请添加它

吧！如果你认为职位需要 Python 和 R 技能，请确保你的简历和 LinkedIn 标记了这些技能，并且把你锁熟练掌握的各种包，比如 Python（Numpy，Pandas，matplotlib，scikitlearn，sklearn），R（ggplot2，shiny，tidyr）也列上去，能更说明你的专业性。

5）在 LinkedIn 方面，Endorsement 功能也在这方面发挥了积极的作用，一定要积极主动的询问与你曾经共事的伙伴来认可你的技能，给出推荐。

越来越多的猎头和招聘官会更看重 LinkedIn 。招聘人员平均花 30 秒钟之前查看简历，然后就放在一边了。确保您已经完善简历和 LinkedIn 的格式，消灭语法错误，填写了正确的关键字，以及选用有力的动词，生动展现出你的工作成果和影响力。

第四步：简历和求职信写作

完善了你的 online image 之后，接下里对于拿到面试机会，更重要的就是简历 resume 和求职信 cover letter 的写作了。

不同于在学术界，列举一系列的令人印象深刻的论文和学术工作可以胜过其他一切；申请行业工作时，应当尽可能简洁，并列出你的最重要以及相关的工作成就和影响。

现实往往很残酷。也许你花了几周甚至一个月细心逐字逐句修改的简历，其实并不会被认真一字一字读。面对成百上千的申请者，更多情况下，招聘人员只是一扫而过简历，经验表明，招聘官平均在一份简历上只会花 **30 秒**。

关于简历的重要建议

1）保持简短，最好控制在一页。请记住，招聘官在考虑认真阅读之前，是通过快速扫视你的简历以获取他们感兴趣的信息。
2）确保你突出展示相关的技能点（可以**加粗**）。招聘人员和招聘经理会在进一步研究之前，看看您是否在技术技能上合适。
3）保证有趣的话题和确保关联性。每段工作经历描述中，最多用三句话（三个 bullet point）描述你的经历和成绩。 你想要清楚地标识您的经验与你申请的工作要求相关联。人生经历漫漫长，简历不是自传，不可能所有经历和细节都放在上面。

4）用数据量化，证明你的成果！不要只是简单的说“做过…”。需要将你工作的结果和影响写出来。比如，你需要清晰表述出“创建了一个 xx 模型，帮助 xx 提高了 xxx%的预测准确率”，而不只是简单的陈述“创建了 xx”。

除了简历之外，求职信（Cover letter）也属于经典的求职必备两件套。有趣的是，通过最近我们和招聘人员的交流，我们发现公司目前已经很少阅读 cover letter 了。当然，如果有明确说明需要 Cover letter，还是要准备一个。重点精力还是放在你的简历或你的 LinkedIn 上。
如果你想主动出击，你也可以尝试发送一个简短的个人总结，通过电子邮件发给招聘官。这样做有一个好处：一个公司可能有不同组都在招聘，即使这个招聘官觉得你不合适，也可以方便他转发邮件给其他组的同事，说不定就会有其他机会。我们过往 Data Science 班的同学就有类似经历，第一次 Amazon 亚马逊数据科学家面试之后所申请的组招满了，就被推荐到另外一个组，最终拿到了其他业务组的 offer。所以最好保持简短，不要超过几个段落， 集中在三条最能说明你特色金额个人价值的过往经验。

第五步：模拟面试

希望你所有为数据科学面试所做的一切工作最终获得回报。我们将详细剖析准备面试的几个部分。

面试过程是如何

数据科学面试相比普通面试有特殊性，行为问题与一大堆技术问题混在一起。能够首先拿到面试机会，意味着你已经有了很大的进步，但是你还有待进一步的发展。

根据你申请的职位和招聘机构，数据科学面试将会有很大的不同。某些组织将非常严格，让您经历几个技术挑战。其他人会更多的看重公司文化适应，特别是如果你有强有力的内部人推荐，可能就直接让你进入最后一轮面试。

一般严格的面试包括以下步骤：

1-电话筛选

这通常由人力资源部门的人员完成，并作为过滤器来节省招聘经理的时间。有时候会出现基本的技术问题来筛选严重不合格的候选人，但大多数时候，这个电话筛选就是建立文化适应的开始，并确保候选人具有足够的沟通能力。

在这次电话会议中，你需要了解数据团队面临的问题和你应聘团队的组织结构。准备一些有深度，可以表现出对业务及其运营状况有深刻理解的问题，在面试中询问他们。

2-TAKE-HOME ASSIGNMENT

电话初筛后，公司经常为候选人发送准备好的小测，要求一段规定时间内完成。这是筛选技术薄弱的候选人的好方法，也可

以排除那些可能没有足够的时间投入到招聘过程中候选者。一些公司完全免除了这一点，但是那些用 Take-home assignment（Analytical Test）的公司往往把它作为一个测试，来节省他们招聘经理的时间。

一个测试的例子包括对给你提供的特定数据集进行深入分析。当测试设计良好时，这也是一次机会，让你可以了解更多关于你将来要完成工作的问题类型。在这里，你可能会在数据中找到有关见解，并以 storytelling 的故事呈现。另一个例子就是要求清理数据集中可能可能有的错误， 最后一个例子将涉及与业务相关的具体问题，例如根据职务说明的数据为申请人建立职位推荐系统。

只有那些通过测试的候选人才能与招聘经理进行面对面交谈。如果你拒绝完成测试，你会被迅速淘汰。

建议大家花时间做这些测试，并尝试看看它与公司正在经历的什么问题有关。大家可以使用测试来作为一种方式去了解你将要测试的技能，以及有关公司如何考虑你的角色。请确保你最大限度地发挥你的时间和贡献。这可以帮助你在招聘过程中展示您出与其他候选人的不同之处的机会。

3-与招聘经理打电话

您可能会收到另一个电话面试，将专注于数学、统计或编程问题。这将由招聘经理或技术人员完成。这可能是在公司邀请您进行现场面试之前的最终评估。电话通常将分为三个部分。有

时，这是在一个长时间的电话中完成的；另外一些情况会在三个短暂的电话中完成的，每次约三十分钟。

数学/统计面试电话

这个电话主要是对核心数学和统计概念进行评估，这将在某种程度上取决于您正在申请什么角色和公司。 互联网公司将倾向于关注您对 A / B 检验的了解，你对 P 值如何计算的理解以及统计意义的含义。能源公司可能会更加重视对回归和线性代数的测试。无论你正在和什么类型的面试官交流，你都需要勾画出解决问题的整个思维过程。

如果你被问及 A / B 检验，请详细描述 A / B 检验的过程，展示出你过往相关的经验，注意细节和陷阱。一方面要展示你的数学和对统计学理论功底，同时也要注意业务相关的细节，以及这个统计测试结果对公司有什么指导性的意义。

编程面试电话

这部分采访过程是相当典型的，也是最接近其他技术访谈。你将通过电话评估你的能力，通过呈现伪代码或编译就绪代码来解决编码挑战。如果你正在申请数据分析师职位，很可能会考察你如何用 SQL 查询数据。如果申请数据科学家、数据工程师等职位，你可能会被问及 Java、Python 相关的 编程和脚本语言的问题，以及任何一种你在简历提到熟练掌握的编程语言。

你的面试官也可能使用 HackerRank 或 Collabedit 等工具来评估你在线测试的状态。在这种情况下，你的招聘经理将和你线上

直接交流，考察你 white board coding 和解决问题的能力。 你也可以提前使用这些工具进行面试培训！
目前网上有很多非常好的编程面试辅导相关的资源，从 Cracking the Coding Interview 到 InterviewCake，请好好利用这些资源。

通常面试中还会问到很多关于数据结构的问题，包括 hashmaps 哈希表，tree 树型结构，堆栈和队列。好好准备这个面试，就像软件工程师如何准备编码面试一样，相信“熟能生巧”，尽早开始编程练习。准备记下纸上的代码，并通过电话进行解释，或者准备输入笔记本电脑上的代码，你一定会顺利通过。

和招聘经理面试

最后，你将和招聘经理直接交流。这一环节主要是评估你沟通的良好程度，以及是否适合团队合作。这可能是和技术电话面试分开的单独电面，或者它可以是包含所有三个部分的最后一块。在这个电话中，招聘经理试图了解更多关于你的性格，你的工作积极性，你的团队合作能力以及你的原始聪明才智水平。
大多数招聘经理心里都有一个合适的人选模型，你越和这个人选模型接近，你就越有可能被邀请到现场面试。
你之前的面试准备和前几轮跟面试官沟通中对公司的了解都能为该轮面试。你对招聘经理面临的业务问题以及他们正在寻找的那种人的了解越多，你就越有能力将自己呈现为完美契合。
定制您的沟通目标，并自信和清楚的表达，你会更容易进入下

一轮。另一方面，沟通和社交能力也是很重要的一方面， 因为工作场所将迫使你密切合作，共同度过大量时间，尝试通过“Airplane”测试，想象如果出差，主管经理评估他们是否愿意花几个小时与你在一起，确保你显示出你的社交能力让你的主管可以一起去。

4-现场面试 ONSITE INTERVIEW

最后，如果你已经通过早期的面试，你将会和你以后工作的上司面对面。他们将从技术和非技术角度对你进行评估。他们正在寻找确定你是否适合的人选，他们还可能通过让白板考题测试你的技术实力。

5-技术面试 TECHNICAL CHALLENGE

如果在现场面试中没有考到技术部分，请做好准备以一种或另一种形式挑战你的技术技能，特别是对于倾向于数据工程的角色。你可能会发现这与软件工程面试过程相似，你将被要求使用白板，并写下如何实现某些算法或解决某些问题。
这里软件工程相关的基础和知识，如时间复杂性/大 O 表示法，以及对数据算法背后的数学和统计数据的熟练掌握就可以发挥展现实力了。

6-和管理层面试

如果你通过招聘经理的层层考核，你经常会和一位高级管理人员进行最后面试。在创业公司中，这往往是共同创始人或 CEO 本身。

如果你这么做，恭喜！不要认为这是理所当然的，这是一个公司倾向于给你工作 offer 的迹象。通常情况下，只有通过技术面试的候选人才能到达这里，所以现在你需要强调你如何通过对业务本身的了解以及其面临的具体问题来提高对公司的影响力。在这一点上，你不是要尽可能地证明自己，最主要的是避免出现明显的错误。

行为面试

所有的面试中，都会包括行为问题(Behavioral Questions)。数据科学面试也是一样。面试官打算测试你的软技能，看看你是否符合公司文化。

对于行为面试而言，最基础也是最最最重要的就是一定要准备好你的故事。你会被非常详细地问到你过去的工作。做好把过往经历和工作描述的尽可能详细的准备，包括从你用过那些工具到为什么你做出不同的决定的所有细节。准备好一个连贯的故事来告诉面试官你曾经做过哪些惊艳的事情来提高你过往公司的业务结果。

面试例题讲解

Q1: 告诉我过去所做的一个数据科学项目？

考察意图：了解你从过去经验中获得的知识和贡献的深度。它测试你能够讲述你的工作的故事的能力，以及你是否可以配合它对你所使用的公司产生影响。
回答思路点拨：

- 尝试描述展示产品和工程的项目
  所谓经验，就是你为该项目提供了分析，并产生了洞察力，使之具有可操作性。例如，如果您通过主题提取技术在文本数据集中确定了关键主题，那么您应该解释这些主题如何促进公司在数据产品中的发展。
- 从业务目标角度详细介绍您的具体贡献和结果
  面试官想知道您在试图了解项目总体目标时所做的具体做法。
- 多次操练
  这是一个非常常见的问题，所以需要起码准备 2-3 个可以深入讲出所有细节的项目故事。

Q2: 你喜欢不喜欢你以前的职位？
考察意图：确定您面试的角色是否适合您，并确定您从前一职位移动的原因。
回答思路点拨：

- 了解角色
  使用人力资源联系人获取有关内幕信息。

- 分析角色及其挑战
  积极了解关于角色、团队、历史和关键即时业务目标的宝贵资料。
- 自圆其说
  最后强调为什么现在申请这个职位和为什么适合

第六步：高阶面试准备 - 案例面试和数据挑战

除了行为面试之外，数据科学面试对于技术也有一定的要求。作为国际生，只有展现出强大的技术能力才能证明自己的不可替代性。

技术面试的 Technical Questions 分为六大类：

- 概率与统计

统计和概率往往是数据科学面试中的“主食“。这些问题的目的是测试你的思维和你如何在有不确定性下进行推理。这是一个数

据科学家需要掌握的基本技能。你必须得对概率论和统计的基本知识有很清晰的了解，这里点出几个必须了解的知识点：组合事件概率，条件概率，期望值，置信区间，参数估计，p-value，t-test，A/B test 以及 假设检验。

面试例题：

Q1: During the lunchtime suppose you are sitting by the street, you find that there is 64% chance observing at least one Tesla in an hour. What is the probability that you observe at least one Tesla in half an hour?

考点：Probability Theory

解：The chance of no Tesla passing by in an hour is 1-0.64 = 36%. An hour can be viewed as two consecutive 30-min time period. Therefore, the chance of no Tesla passing by in one 30-min time period is squareroot(36%) = 0.6. So the probability that you observe at least one Tesla in half an hour is 1-60%=40%.

Q2: You are testing for a rare disease, with 1% of the population is infected. You have highly sensitive and specific test: 99% of sick patients test positive, and 99% of healthy patients test negative. Given that a patient tests positive, what is the probability that the patient is actually sick?

考点：Statistical inference; Bayes' Theorem($P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$);

Marginal Probability($P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$)

解：P(sick) = 1%

P(+|sick) = 99% => P(+|not sick) = 1-P(+|sick) = 1%

According to marginal probability, P(+) = P(+|sick) * P(sick) + P(+|not sick) * P(not sick)

= 1% * 99% + 1% * 99% = 0.0198

According to Bayes' Theorem, P(sick | +) = P(+|sick) * P(sick) / P(+) = 99% * 1% / 1.98% = 0.5

- 数据库处理

在这类面试问题中，主要考察了对于数据库的掌握，和能否有效的用正确的口令，抓取有用的数据并获得有用的信息。这一类型的问题，主要考察了 SQL 的句法。常见的考点有 GROUP BY, sub-queries, INNER JOIN vs JOIN, WHERE vs HAVING…

面试例题：

Q1: Create three tables: a) accounts, which contains accountid; b) dates, which contains dateid; c) facts, which contains three columns - dates, accountid and revenue. The facts table records the expense of an account every day if there is expense. If there is no expense then there won't be a record in the facts table. Given this scenario, write a SQL query that generates a list of all accounts on every day in the last 30 days that had no expense.

考点：SQL(CREATE TABLE, JOIN, subquery...)

解：

```sql
#create tables as required
CREATE TABLE dswhitepaper.accounts(accountid BIGINT);
CREATE TABLE dswhitepaper.dates (dateid DATE);
CREATE TABLE dswhitepaper.facts(
    accountid BIGINT,
    dateid DATE,
    revenue NUMERIC);

#join tables and retrieve no sale accounts
SELECT d.dateid, a.accountid
FROM accounts a
CROSS JOIN dates d
LEFT OUTER JOIN(
    SELECT DISTINCT f.accountid, d.dateid AS dateid
    FROM facts f CROSS JOIN dates d
    WHERE (d.dateid - f.dateid) <= 30
        AND (d.dateid - f.dateid) >= 0) active
ON a.accountid = active.accountid AND d.dateid = active.dateid
WHERE active.dateid is NULL;
```

Q2: What's the difference between HAVING and WHERE?

考点：SQL 概念

解：The difference between the two is in the relationship to the GROUP BY clause. WHERE comes before GROUP BY; SQL evaluates the WHERE clause before it groups records. HAVING comes after GROUP BY; SQL evaluates HAVING after it groups records.

- 实验设计

几乎每个数据科学公司都会在技术面试中问关于 A/B testing 的实验设计问题。这是因为 A/B testing 在现代科技公司中测试网站或者项目起着举足轻重的地位。但是一般的毕业生往往没有这方面的从业经验。所以公司也时常会考察一些基础的统计类的实验设计问题，来确保求职者在入职后也能很快适应并展现出超高的素质。

同时一些实验结果分析中的统计概念也会被考察。常见的考点包括了 p value, power，t test，ANOVA test 等等。

面试例题：

Q1: What is the definition of power? What are some factors that affect the power of the test? How does the power of the test relate to P value in a certain test?

考点：统计概念（power, p-value...)

解：The power of a statistical test is the capability to detect that an effect actual exists. In this context, it is the capability of detecting that the metrics of treatment is different from control, when it is truly the case. Probabilistically, the power of test is: Pr(reject null hypothesis | null hypothesis is false). It can also be expressed as 1-β, where β is Pr(fail to reject null hypothesis | null hypothesis is faIse). β is also usually termed as Type II error.

Qualitatively speaking, here is the list of factors that could affect the Power:

- Sample size (n). Other things being equal, the greater the sample size, the greater the power of the test.
- Significance level (α). The higher the significance level, the higher the power of the test.
  α is also known as Type I error.
- Variability or Standard Error

p-value is related to α. Once the α is set, a statistic e.g. t-statistics is computed. Each statistics has an associated probability value called p-value, or the likelihood of an observed statistic occurring due to

chance, given the sampling distribution. In other words, p-value measures how extreme the data are.
For example, with a very small p-value (<< 0.0001) we say there are 0.01% chance we observe this data is due by chance. Therefore we believe the data is indeed very extreme. By comparing p-value with Type I error we can decide if we are going to reject the null hypothesis.
If we set a larger α, we in fact allows a larger p-value to be considered significant, therefore increase the power of the test.

Q2: We want to add a new feature to our product. How to determine whether people like it or not?
考点：实验设计，A/B Testing
解：The general idea is to do A/B Testing:

- Define metrics to measure if the feature has met the business expectation
- Prepare two version of the products; one with new feature, and one without. Simultaneously serve both versions to customers.
- Split test population into a treatment group and a control group. Make sure such split did not introduce any bias. A common way to do this is Simple Random Sampling (SRS) 50% vs 50%.
- Prepare the following elements as of a standard hypothesis testing:
    - You need to define a null hypothesis for your hypothesis testing. The null hypothesis can be that the treatment metrics is not different from the control.

- Pre-define the Type I error you can tolerate in your weblab. Based on this, one can judge based on the p value if the result is significant.
- Assess the power of your statistical tests.

There are also two practical considerations when running a A/B Testing on a website:

- Running time: increase running time can get more samples and successfully fend off noise. However, if the metrics drop it can bring real loss.
- Dealing with outliers. There are two simple ways to deal with outliers. One way is to truncate at a certain percentile of metrics of the population, e.g. 99.9%. OR you can visualize the outliers and have a hand picked threshold.

- 产品度量与分析

对于一般常规学校教育中往往会忽略这一部分的训练。然而产品表现的度量与分析又偏偏在数据科学工作中有着重要的应用。面试官往往会问一些关于产品的问题。面试官会先描述一个商业情境，然后问一些开放性的问题。建议大家在准备这一类型的技术面试问题时，可以上网多搜集一些工业界常见的 metrics。在这基础上，面试者也有可能会夹杂着一些概念性的问题来综合提问。

一些常见易考的 metrics 有 Active user，conversion rate, impression, Click Through Rate(CTR), Bounce Rate, Revenue, Profit...更多的 metric 指标可以参考我们最后一章给大家整理的“商业常见产品 Metrics 一览”。

面试例题：

Q1: An Internet company recently enforced users to use a standalone Messenger App and deprecated the chat functionality on the company's mobile App. To track the performance of this move, what metrics will you use?

考点：Product Metrics

解：In this case, we analyze the metrics before and after launching the standalone messenger App. One consideration point when choosing the metrics is to look at the total metrics before and after the change, without delineating the impact on the Company's general purpose mobile App.

1. The number of daily active users and monthly active users are good indicators to see whether we lose users, or experience reduced activity by enforcing them to use a standalone Messenger App. We can compare the number of total users of both mobile App and messenger App after the change, with the number of users of the mobile App before the change. This metric should provide insight on whether the deprecation has significant impact on the App users.
2. The total time spent on the old integrated mobile App vs. the total time spent on the new App plus the total time spent on the standalone Messenger App.
3. The number of messages sent or received. Again, compare the metric between the standalone Messenger App and the old integrated mobile App.
4. The time taken for a user to send or resend a massage. Compare the metric between the standalone Messenger App and the old integrated mobile App.

Also if we need to maintain the user engagement of the mobile App after the change, a few more metrics should be considered:
1. Revenue generated by the general purpose of mobile App before and after the change
2. Click through rate or number of impression of the general purpose mobile App excluding the ones generated from the chat functionality.

- 编程

如果统计和概率是数据科学面试中的“主食“，那编程问题就会是“配菜“。数据科学需要大规模处理数据，这需要编辑程序来自动处理所要求的大量的工作。像 Facebook, Twitter, LinkedIn 等大型互联网科技公司都要求数据科学求职者们具有超强的 coding 能力。要求对于基础的算法需要熟悉，但对具体编程语言种类没有固定要求。这类型的问题基本和面试 software engineer 的面试问题差不多。准备时，可以参考的网站包括 leetcode.com, hackerrank.com 等，书目的话可以看 Cracking the Coding Interview。

这类问题的难点在于有时候问题问的内容不是很清晰，所以回答前要弄清到底在问什么。如果有任何不清楚的，一定要仔细的询问面试官，请面试官澄清所有的要求、前提、条件等等。

面试例题：

Q1: Implement pow(x, n) where x is a double and n is an integer.
考点：Coding

解：This is a numerical computation problem. The naive approach is to compute the product of n copies of x, which takes O(n) times. But using the recursive equation pow(x, n) = pow(x^2,n/2), we can reduce the integer by half after each iteration until the n becomes 0 where x^0=1. In this way, we can have O(log(n)) computation time. Note that n is integer, we need to handle corner cases such as 1) if n is negative then pow(x, n) = pow(1/x, n), and 2) if n is odd, we have pow(x, n) = pow(x,n-1). Note that (n-1)/2 is an integer when n is odd.

```
def pow(x, n):

    if(n == 0):
        return 1
    if(n == 1):
        return x
    if(n < 0):
        return pow(1.0 / x , -n)
    if(n % 2 == 1):
        return pow(x, n-1) * x
    else:
        return pow(x * x, n / 2)
```


- 机器学习

Machine Learning 作为大热话题，已经成为了技术面试里的热门考点。除了考察你对基础机器学习概念的理解，还特别喜欢问关于 supervised learning 方面的问题。比如说 bias 和 variance 之间的 trade-off，什么叫过度拟合（overfitting）且如何避免等等。有些时候面试官也会让你运用机器学习、选择合适的机器学习算法来解决现实生活中的商业问题。注意，在面试时，可以和面试官讨论一下问题的情景和背景。比如说有哪些 input，需要

预测的是什么，training data 的规模是多少，最重要的特征有哪些等等。

面试例题：

Q1: Why is “Naive” Bayes naive?
考点：ML 概念（Naive Bayes）
解：Naive Bayes is considered “Naive” because it makes an assumption that is virtually impossible to see in real-life data: the conditional probability is calculated as the pure product of the individual probabilities of components. This implies the absolute independence of features — a condition probably never met in real life. As a Quora commenter put it whimsically, a Naive Bayes classifier that figured out that you liked pickles and ice cream would probably naively recommend you a pickle ice cream.
Q2: “People who bought this, also bought….” recommendations on Amazon are a result of which machine learning algorithm?
考点：ML 概念和应用
解：Recommender systems usually implement the collaborative filtering machine learning algorithm that considers user behaviour for recommending products to users. Collaborative filtering machine learning algorithms exploit the behaviour of users and products through ratings, reviews, transaction history, browsing history, selection and purchase information.

** 此章节的 Technical interview Questions 摘选自:

You, Jane. *Data science interviews exposed: your one stop source for Data Science job interviews*. S.I: CreateSpace, 2015. Print.

与数据科学家面试官的采访

［采访一］

采访嘉宾：Peter 老师，数据领域资深专家，现就职于 Uber，曾在多个领域的大中小企业任职并担任招聘面试官，拥有丰富的职场经验，在面试方面有独到的见解。

Q: 您在招聘的时候，最看重应聘者的哪些素质？

A: 不同类型的公司对于应聘者的要求各有侧重。对于大公司而言，需要应聘者就某一特定领域具有较深入的造诣，讲究“专”。而对于中小型公司，希望应聘者知识面比较广，对数据科学有 general sense，讲究“广”。

Q: 为了顺利通过面试，您有什么好建议？应聘者该如何有效准备面试？

A: 如果应聘大公司，应聘者应该对自己申请的 team 有一定的了解，知道这个组希望招什么样的人，这点很重要。比如，这个组是做 risk analysis，就需要应聘者对于风控模型有深入的了解。如果这个组专门做优化的，那应聘者对于优化模型的各种方法需要掌握到位。如果应聘中小公司，communication skill 和对抽象问题的理解能力很重要，可能给出的问题比较抽象，需要自己去 define。

Q: 您通常会在面试中问什么样的问题，想测试应聘者的哪些能力？

A: 我个人比较喜欢问一些“广”的问题。比如，在 Santa Monica，有多少车在路上开？这样的问题其实很能看出应聘者的思维方式，他的 assumption 是什么，approach 是什么。如果是比较具体的职位，例如是和统计相关的岗位，会问一些专业问题，像贝叶斯， a/b testing 等。

Q: Uber 和其他科技公司相比，在招聘环节上会有什么不同之处？

A: 一般公司招聘会有这样几个流程。首先第一轮是电话面试，往往是 hiring manager 或者 recruiter 了解一下大致情况。第二轮是 technical interview，考察对 SQL, Python, Programming, Algorithm 等一些基本功的掌握情况。第三轮是 behavioral interview。Uber 会在 technical interview 之后加入一轮 take-home data challenge , 关于这方面的准备，大家可以多练练 Kaggle 上的题。面试结束后，往往都有一个 cutoff score，根据这个分数，来决定是否发 offer。

Q: 对于非理工科背景的同学找数据领域的工作，在面试时有何建议？

A: 我们欢迎各个背景的同学。当然，data 领域不同职位的要求各异，希望同学们扬长避短，发挥各自的专业优势。对于非科班出身的同学，建议多找一些实践的机会，例如在学校找一些相关的课题，或者在 kaggle 上面练练手。另外，SQL, Python 这些 hard skills 也一定要牢牢掌握。

Q: 最后，有什么职业发展的心得送给同学？

A: 在职业发展的前半期，其实每个人都给自己播种下了一个点，一个 dot。大家趁年轻，一定要多去尝试，多去发现机会。到后面，你会发现播下的种子总会有发芽的那一刻，所经历的都是有用的，就像乔布斯说的“connect the dots”。拿我自己来举例，之前在找工作的事情有尝试咨询行业，练了 case interview，也尝试过金融领域，考了 CFA。这些虽然看似和我现在做的 data

并没有太大的关系，但是却锻炼了我思维方式，让我能从不同的角度看问题，对我的职业发展产生了积极的影响。

［采访二］

采访嘉宾：小梁老师，数据领域资深专家，现担任某 b2c 电商零售公司首席数据科学家， 曾在生物制药，咨询，时尚等多个领域拥有从业经历。

Q: 您在招聘的时候，最看重应聘者的哪些素质？

A: 主要看专业技术能力，商业嗅觉以及逻辑分析能力。

Q: 为了顺利通过面试，您有什么好建议？应聘者该如何有效准备面试？

A: 希望同学们能抓牢基础知识。例如，在电话面试环节中，我们会考察 SQL, Scripting Language(Python/Java/Ruby 任意)等一些基础的概念，所以需要扎实的功底。在 onsite 面试中，更看重的是应聘者的表达能力，如果遇到不懂的，也不要装懂，多问多交流。应聘者需要会用 white board 交流想法，通过写一些 pseudo script 来清楚表达自己的想法。

Q: 您通常会在面试中问什么样的问题，想测试应聘者的哪些能力？

A: 我个人会问一些和专业技术相关的问题，考察对 SQL, Scripting Language(Python/Java/Ruby 任意), advanced modeling 的掌握程度。还会问一些项目的经历，如何做项目管理，从最开始的数据准备到最后的成功展示是如何规划的，包括项目中所遇到的一些问题，以及如何解决的。

Q: 您的公司和其他科技公司相比，在招聘环节上会有什么不同之处？

A: 一些小的 startup 比较喜欢关注应聘者过去的项目经历，以及应聘者能给公司的成长带来什么影响。

Q: 对于非理工科背景的同学找数据领域的工作，在面试时有何建议？

A: 非理工科背景的同学需要尽量强调出本专业能对 data analysis 带来什么不同的视角，有益于更全面的分析问题。同时，对常用分析工具有一个基本了解其实已经足够了，用人公司更看重的是你的 passion 和发展潜力。

拿到 OFFER 的阶段

你的目标是获得尽可能多的工作 offer，以便你可以评估和谈判。虽然过程本身很困难，可能需要比你预期的更长的时间，一旦您开始获得 offer，你就可以获得更多工作机会。

首先必须强调，把控期望和保持希望的重要性。我们采访的几位数据科学家谈到几个月到五十年的时间，等待着从一所有名望的学校进入一个高级学位到安全的工作。他们中的很多人不得不采取入门级的位置来踏上门。

你可能听说过很多关于数据科学的伟大的事情，但你只会经历很多艰苦的工作和等待。

确保你衡量所呈现给你的是什么，选择你应得的一切，一旦你花了所有的辛勤工作赚钱。

接受工作 offer

如果您成功完成一个流程，您可能会有一个优惠或多个优惠。恭喜！

接受报价是对您有关公司的大量时间的承诺。一定要考虑到这一点。有几个因素可以用来确定报价是否适合您。

- 公司文化

这可能是确定报价是您应该接受的最重要因素之一。确保你询问你将要成为的一部分公司文化。寻找公司有真正享受彼此花时间的人的迹象

逃避一般的描述符和企图，这些描述符和公司很难界定自己的文化，或是将这个问题浪费掉。伟大的公司投入大量的时间和精力来确保他们拥有喜欢他们做的事情的真棒人物。这将在你的质询中脱颖而出。

您还应检查外部和客观来源，例如 Glassdoor 上的公司评论。接近目前的员工，以及您在 LinkedIn 上可以找到的以前的员工，以获得他们的故事。你会经常找到可以给你一个很好的预览工作在你的新工作的坦白的故事就像。

- 团队

公司文化是生活在其中的团队的延伸，但你应该很高兴每天上班，并与其他人一起工作。确保你正在和一个你可以学习的团队合作。你是你花费最多时间的五个人的组合，你将花费大量的时间与你的办公室团队。

- 位置

确保您在公司所在的地方很舒适，尤其是在距离很远的地方。你不能没有困难地移动，重要的是你放心你住在哪里。像天气和过境系统这样的事情在一定程度上是重要的，特别是如果你要和这些条件一起生活多年。

- 谈判你的工资

令人惊讶的是，有 18％的人从来没有谈判工资，尽管那些通常会把工资提高到 7％的人。当您第一次获得报价时，您处于独一无二的杠杆点，这是时候测试你的价值。

如果你有一些平均工资来谈判，总是更容易。如果您有具体的优惠提议，您将在谈判桌上更加坚强。

以下是可以开始研究的一些事实和数据。

Indeed.com 的数据分析师的平均工资为 65,000 美元，数据工程师的平均工资为 100,000 美元，数据科学家的平均工资为 11.5 万美元。区域不同，薪水最高的地区在高科技海湾地区。根据 O'Reilly Media 的数据，加利福尼亚州是所有地区的最高范围和中位数。在全球范围内，美国的数据科学工资的中位数和范围最高，而英国，新西兰，澳大利亚和加拿大则并不遥远。亚洲和非洲往往拥有最低的中位数。

最高收入行业往往是技术和社交网络公司，而支付最低的行业往往是教育和非营利性行业。

这个薪水也根据所使用的技能和工具而有所不同。 O'Reilly 对行业中数以百计的受访者进行了明确的调查。一项公开研究，结果表明导致不同平均工资的各种不同因素。就像一个例子，使用 Scala 语言的人是一种专门的编程类型，收入高于 10 万美元的中位数，而使用 SPSS（专有工具）的人的收入要少得多。

在您接受报价之前，请确保您了解您对公司，团队和金钱的承诺。

如果您已接受 offer，恭喜！你已经完成了这个整个过程的目标，找到了你想寻求的工作，并且这个工作承诺你很好的薪酬，并且能带来重大的社会影响。

数据科学资源总汇

数据求职 CHECK LIST

- ▢ Map out the role your skills fit
- ▢ Map out the industries and types of companies you want to work for
- ▢ Prepare your LinkedIn, CV, and email templates
- ▢ Research each company and role you want to aim for thoroughly
- ▢ Reach out proactively to individuals within companies with informational interviews
- ▢ Build strong networks and referrals
- ▢ Tackle the data science interview
- ▢ Keep up hope
- ▢ Negotiate your offer

商业常见产品 METRICS 一览

网站类产品常用指标

· 页面浏览量（Page View，PV）：在一定统计周期内（通常为 24 小时）所有访问者浏览的页面总数。该指标重复计算，即如果一个访问者浏览同一页面 3 次，那么 PV 就计算为 3 个。PV 之于网站，就像是收视率之于电视，从某种意义上已成为投资者衡量商业网站表现的最重要尺度之一。严格意义上来说，PV 只记录了页面被加载显示出的次数，并不能真正确保用户进

行了浏览，有些网站会利用这一特性“刷”PV，例如在页面中嵌入不可见的 iframe。还有的网站编辑为了完成 PV 指标，会将一篇长文（或组图）拆分成多页，从而制造出阅读量大的假象。

· 独立访问者（Unique Visitor，UV）：在一定统计周期内访问某站点的不同 IP 地址的人数。通常在同一天内，UV 只记录第一次进入网站的具有独立 IP 的访问者。如果某人访问网易首页，又点开了三条新闻，则记作 4 个 PV 和 1 个 UV。UV 反映了网站覆盖的绝对人数，但没有体现出访问者在网站上的全面活动。此外，由于校园网络、企业机关等一些部门通常有统一的对外 IP 出口，依靠 IP 来判断的 UV 也并不能做到完全准确，更优的做法是结合 Cookies。

· 访问数（Visit）：访问者从进入网站到离开网站之间的整个交互过程，视作一次 Visit。它可能包含一组页面浏览行为。通常界定同一访问者的两次不同 Visit 的判定方法是间隔时长，如 30 分钟。这意味着如果同一访问者连续的两次页面访问之间间隔为 15 分钟，则视作一次 Visit；如果间隔 41 分钟（因故暂时离开或阅读了一篇长文），则被切分为两次 Visit。

· 着陆页（Landing Page）：指访问者浏览网站时所到达的第一个页面，又称用户捕获页。针对着陆页的分析追踪可作为判定外部广告或其他营销推广活动效果的依据，因此着陆页应当是经过恰当优化的。

· 退出页（Exit Page）：指访问者浏览网站时所访问的最后一个页面。退出页数量大，并不等同于网站的黏性差，此时

应当参照退出数与页面浏览量的比值，即退出率。若某个页面本不该有较高的退出率（如在线购买流程的下单环节），则需要检查该页面，防止其成为整站的流量漏洞。

· 跳出率（Bounce Rate）：用于衡量整站或网页的黏性。跳出，指访问者仅仅浏览了一个网页就结束了访问（Visit）。整站跳出率 = 全站跳出数/全站页面浏览量，它反映了整站的导航效率；而针对单独页面计算的跳出率 = 该页面跳出数/该页面浏览量，它是对单个网页导航能力的评价。一般而言，跳出率越高代表网站的问题越大。

· 展现数（Impressions）：又称印象数，指广告在浏览器中被加载的次数。只要广告内容被加载出一次（如刷新了页面），展现数就加 1。服务器打点数（Hit）：打点指服务器收到一次请求。如访问者浏览了一个仅有 10 张图片的网页，则打点数记作 11，其中包括 1 次网页请求和 10 次加载图片的请求。

· 转化率（Conversion Rate）：转化，指达成了某种预设的目标，如引导用户完成下载、注册、新闻订阅、走完新手介绍流程等。转化率是计量这种转化成效的指标，可用于衡量网站内容对访问者的吸引程度和宣传效果等。例如，广告条的转化率 = 通过广告条点击进入着陆页的流量/广告条的展现数；注册的转化率 = 完成注册流程的用户数/到达注册页面的流量。

· 停留时间（Duration）：指一次访问的持续时长。通常较为简单的计算方法是用最后一次访问的时间减去访问第一张页面的时间（但这将无法统计最后一次访问的持续时长）。
· 初访者（New Visitor）：初次访问网站的访问者。通常用 Cookie 判断，并以一定时限为统计周期，通常为一个月。如果上月某人访问过网站，次月再次访问，则对于次月内的第一次访问行为而言，这个访问者仍视作该月内的一个新的初访者。
· 回访者（Return Visitor）：相对初访者而言，如果一个访问者在该月内重复访问，则视作回访者，也就是“回头客”。该指标衡量网站内容对访问者的吸引程度和网站实用性。统计周期内所有初访者数量 + 所有回访者数量 = 独立访问者数量。
· 访问来源（Referrer）：指一次访问或一个网页浏览的流量来源，又被称作“推荐来源”。访问来源可从不同维度进行划分。如按来源网站的性质，可划分为来自搜索引擎、网站推荐（如友情链接、广告条、软文植入）、无网站来源（用户直接进入网站，如从浏览器收藏夹点入、直接在地址栏输入域名）等；按来源网址的形式，可划分为来自域（如 fanbing.net）、网站（如 www.fanbing.net）或 URL（如 http://www.fanbing.net/about.html）；按照内外部，可划分为站外链接或站内来源。

· 其他属性：有的第三方统计工具可结合自身收集的其他数据，获取访问者进一步的信息，如地域分布、系统环境、性别比例、年龄分布、学历分布、职业分布等。

软件及移动应用类产品常用指标

· 新增用户数（New Users）：指首次打开应用的用户数量，通常通过设备识别符（如苹果系统的 UDID）来识别用户的唯一身份。由于传输统计数据需要联网，因此即便是首次打开应用，若未能联网，也统计不到。此外，卸载再安装通常不会算作新增用户，老用户的版本升级也不会计算在内。当然，如果下载了应用但并未安装，或安装之后没有启动过，也无法统计为新增用户。

· 活跃用户数（Active Users）：指统计周期内有过特定使用行为的用户数量。同一用户在一个统计周期内多次使用记作一个活跃用户。这里“使用行为”的定义因应用而异，有的团队将启动即视作活跃，有的则需要满足启动 + 执行某种操作（如浏览过至少一条新闻），还有的则索性将常驻后台的守护进程没有被杀死也统计进了活跃范畴中。因此如何计量活跃用户数，归根到底还是看团队真正追求的是什么。活跃用户数一般看“日活”（Daily Active Users，DAU）和“月活”（Monthly Active Users，MAU）。

· 升级用户数（Updated Users）：指由已装的老版本升级到新版本的用户数量。时常有人问，像 QQ 这样保有量已经

很大的应用，为什么每天还能在应用市场上创造如此巨大的下载量？其中很重要的因素之一，就是将用户从老版本升级到新版本的下载行为统计了进去。

· 留存率（Retention Rate）：指用户在某段时间内开始使用应用后，经过一段时间，仍然继续使用，这部分用户占当时新增用户的比率，也就是“有多少人最后留下来了”。留存率用于衡量应用的质量和营销效果的好坏。通常新增用户如果因为真实需求而来（如从应用市场主动搜索并下载获得），则留存率较高；而因为博眼球的营销推广（尤其是有奖活动）进来的用户，留存率较低。并且，不同种类应用的留存率也有各自的基准，如游戏的首月留存率通常比社交类高，而工具类的首月留存率又比游戏高。留存率通常看次日留存率、3 日留存率、7 日留存率、15 日留存率和 30 日留存率。

· 总用户数（Total Users）：指历史上所有新增用户数之和。该数字由单纯地相加获得，存在一定水分，无法体现已经流失或极不活跃的用户情况。

· 单次使用时长（Duration）：指用户从一次启动到退出应用所耗费的时间长短，用于衡量应用的黏性。应用在后台运行并不会计入其中。不同类别的应用，单次使用时长可以千差万别。工具类产品解决问题目标明确，用户完成任务之后就会立即退出，比如看一下天气、优化一下内存占用等，用几秒就可以关闭。而视频播放类应用则能持续更久，通常可达到几十分钟。

· 平均单次使用时长（Average Duration）：计算方法是某日总使用时长/该日启动数，可用于更准确地评估用户的使用状态。因为一款应用在不同时段的使用时长可能存在差别，用户早上挤地铁时的一瞥与晚间睡觉前的沉浸使用，其单次使用时长本身是不具备可比性的，只有平均之后才能用于横向比较。

· 使用间隔（Interval）：指连续两次使用之间的时间间隔。如果一款定位于提供每日新闻资讯的应用的使用间隔过长，则说明对用户的黏性不够强，并未培养成每日使用的习惯，只是在偶尔想起来时看一眼。这就需要在产品上下功夫，或采取一些运营手段弥补，如定时推送当日的头条新闻。

· 转化率（Conversion Rate）：指应用内特定行为目标的转化情况，如让用户点击某个按钮、播放一段视频、邀请一批好友等。

· K 因子（K-Factor）：衡量产品的病毒传播能力，计算方法为每个用户平均发出的邀请数量/收到邀请转化成新增用户的比率。如果 K 因子大于 1，表明产品具有自我传播能力，会随着用户的使用而持续扩散。

· 每用户平均收益（Average Revenue Per User，ARPU）：简单的理解就是“能从每个用户那里收多少钱”，是衡量产品盈利能力的指标，也可用来检测不同市场渠道获取的用户质量。ARPU 的通常计算方法是产品在一定时限内的收入/活跃

用户数。结合单用户的获取成本，可以推断出产品是否能形成自我造血的持续发展能力。

· 每付费用户平均收益（Average Revenue Per Paid User，ARPPU）：与 ARPU 将收入平摊到所有用户头上不同，ARPPU 只计算从所有付费用户处获取的平均收益，据此更准确地把握付费用户的支付能力、消费习惯，并有针对性地对这部分付费用户重点运营和服务。

· 月付费率（Monthly Payment Ratio，MPR）：指一个月的统计区间内付费用户占活跃用户的比例。

· 生命周期价值（Life Time Value，LTV）：用户从第一次使用产品，到最后一次使用之间，累计贡献的付费总量。

**此章节选自“Growth Hacking 增长黑客”

数据科学专业词汇表

http://www.datascienceglossary.org

algorithm

A series of repeatable steps for carrying out a certain type of task with data. As with data structures, people studying computer science learn about different algorithms and their suitability for various tasks. Specific data structures often play a role in how certain algorithms get implemented.

artificial intelligence

Also, *AI*. The ability to have machines act with apparent intelligence, although varying definitions of “intelligence” lead to a range of meanings for the artificial variety. In AI’s early days in the 1960s, researchers sought general principles of intelligence to implement, often using symbolic logic to automate reasoning. As the cost of

computing resources dropped, the focus moved more toward statistical analysis of large amounts of data to drive decision making that gives the appearance of intelligence.

Bayes' Theorem

Also, *Bayes' Rule*. An equation for calculating the probability that something is true if something potentially related to it is true. If P(A) means "the probability that A is true" and P(A|B) means "the probability that A is true if B is true," then Bayes' Theorem tells us that P(A|B) = (P(B|A)P(A)) / P(B). This is useful for working with false positives—for example, if *x*% of people have a disease, the test for it is correct *y*% of the time, and you test positive, Bayes' Theorem helps calculate the odds that you actually have the disease. The theorem also makes it easier to update a probability based on new data, which makes it valuable in the many applications where data continues to accumulate. Named for eighteenth-century English statistician and Presbyterian minister Thomas Bayes.

bias

In machine learning, "bias is a learner's tendency to consistently learn the same wrong thing. Variance is the tendency to learn random things irrespective of the real signal.... It's easy to avoid overfitting (variance) by falling into the opposite error of underfitting (bias). Simultaneously avoiding both requires learning a perfect classifier, and short of knowing it in advance there is no single technique that will always do best (no free lunch)."

Big Data

As this has become a popular marketing buzz phrase, definitions have proliferated, but in general, it refers to the ability to work with collections of data that had been impractical before because of their volume, velocity, and variety ("the three Vs"). A key driver of this new ability has been easier distribution of storage and processing across networks of inexpensive commodity hardware using technology such

as Hadoop instead of requiring larger, more powerful individual computers. The work done with these large amounts of data often draws on data science skills.

binomial distribution

A distribution of outcomes of independent events with two mutually exclusive possible outcomes, a fixed number of trials, and a constant probability of success. This is a discrete probability distribution, as opposed to continuous—for example, instead of graphing it with a line, you would use a histogram, because the potential outcomes are a discrete set of values. As the number of trials represented by a binomial distribution goes up, if the probability of success remains constant, the histogram bars will get thinner, and it will look more and more like a graph of normal distribution.

chi-square test

Chi (pronounced like "pie" but beginning with a "k") is a Greek letter, and chi-square is "a statistical method used to test whether the classification of data can be ascribed to chance or to some underlying law." The chi-square test "is an analysis technique used to estimate whether two variables in a cross tabulation are correlated."A chi-square distribution varies from normal distribution based on the "degrees of freedom" used to calculate it.

classification

The identification of which of two or more categories an item falls under; a classic machine learning task. Deciding whether an email message is spam or not classifies it among two categories, and analysis of data about movies might lead to classification of them among several genres.

clustering

Any unsupervised algorithm for dividing up data instances into groups—not a predetermined set of groups, which would make this

classification, but groups identified by the execution of the algorithm because of similarities that it found among the instances. The center of each cluster is known by the excellent name "centroid."

coefficient

"A number or algebraic symbol prefixed as a multiplier to a variable or unknown quantity (Ex.: *x* in *x(y + z)*, *6* in *6ab*" When graphing an equation such as $y = 3x + 4$, the coefficient of x determines the line's slope. Discussions of statistics often mention specific coefficients for specific tasks such as the correlation coefficient, Cramer's coefficient, and the Gini coefficient.

confidence interval

A range specified around an estimate to indicate margin of error, combined with a probability that a value will fall in that range. The field of statistics offers specific mathematical formulas to calculate confidence intervals.

correlation

"The degree of relative correspondence, as between two sets of data." If sales go up when the advertising budget goes up, they correlate. The *correlation coefficient* is a measure of how closely the two data sets correlate. A correlation coefficient of 1 is a perfect correlation, .9 is a strong correlation, and .2 is a weak correlation. This value can also be negative, as when the incidence of a disease goes down when vaccinations go up. A correlation coefficient of -1 is a perfect negative correlation. Always remember, though, that correlation does not imply causation.

covariance

"A measure of the relationship between two variables whose values are observed at the same time; specifically, the average value of the two variables diminished by the product of their average values." "Whereas variance measures how a single variable deviates from its mean,

covariance measures how two variables vary in tandem from their means."

cross-validation

When using data with an algorithm, "the name given to a set of techniques that divide up data into training sets and test sets. The training set is given to the algorithm, along with the correct answers... and becomes the set used to make predictions. The algorithm is then asked to make predictions for each item in the test set. The answers it gives are compared to the correct answers, and an overall score for how well the algorithm did is calculated."

data engineer

A specialist in data wrangling. "Data engineers are the ones that take the messy data... and build the infrastructure for real, tangible analysis. They run ETL software, marry data sets, enrich and clean all that data that companies have been storing for years."

data mining

Generally, the use of computers to analyze large data sets to look for patterns that let people make business decisions. While this sounds like much of what data science is about, popular use of the term is much older, dating back at least to the 1990s.

data science

"The ability to extract knowledge and insights from large and complex data sets." Data science work often requires knowledge of both statistics and software engineering.

decision trees

"A decision tree uses a tree structure to represent a number of possible decision paths and an outcome for each path. If you have ever played the game Twenty Questions, then it turns out you are familiar with decision trees."

deep learning

Typically, a multi-level algorithm that gradually identifies things at higher levels of abstraction. For example, the first level may identify certain lines, then the next level identifies combinations of lines as shapes, and then the next level identifies combinations of shapes as specific objects. As you might guess from this example, deep learning is popular for image classification.

dimension reduction

Also, *dimensionality reduction*. "We can use a technique called *principal component analysis* to extract one or more dimensions that capture as much of the variation in the data as possible... Dimensionality reduction is mostly useful when your data set has a large number of dimensions and you want to find a small subset that captures most of the variation." Linear algebra can be involved; "broadly speaking, linear algebra is about translating something residing in an *m*-dimensional space into a corresponding shape in an *n*-dimensional space."

feature

The machine learning expression for a piece of measurable information about something. If you store the age, annual income, and weight of a set of people, you're storing three features about them. In other areas of the IT world, people

JavaScript

A scripting language (no relation to Java) originally designed in the mid-1990s for embedding logic in web pages, but which later evolved into a more general-purpose development language. JavaScript continues to be very popular for embedding logic in web pages, with many libraries available to enhance the operation and visual presentation of these pages.

k-means clustering

"A data mining algorithm to cluster, classify, or group your *N* objects based on their attributes or features into *K* number of groups (so-called clusters)."

k-nearest neighbors

Also, *kNN*. A machine learning algorithm that classifies things based on their similarity to nearby neighbors. You tune the algorithm's execution by picking how many neighbors to examine (*k*) as well as some notion of "distance" to indicate how near the neighbors are. For example, in a social network, a friend of your friend could be considered twice the distance away from you as your friend. "Similarity" would be comparison of feature values in the neighbors being compared.

latent variable

"In statistics, latent variables (from Latin: present participle of *lateo* ('lie hidden'),as opposed to observable variables), are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured). Mathematical models that aim to explain observed variables in terms of latent variables are called latent variable models."

linear algebra

A branch of mathematics dealing with vector spaces and operations on them such as addition and multiplication. "Linear algebra is designed to represent systems of linear equations. Linear equations are designed to represent linear relationships, where one entity is written to be a sum of multiples of other entities. In the shorthand of linear algebra, a linear relationship is represented as a linear operator—a matrix."

linear regression

A technique to look for a linear relationship (that is, one where the relationship between two varying amounts, such as price and sales, can be expressed with an equation that you can represent as a straight line

on a graph) by starting with a set of data points that don't necessarily line up nicely. This is done by computing the "least squares" line: the one that has, on an x-y graph, the smallest possible sum of squared distances to the actual data point *y* values. Statistical software packages and even typical spreadsheet packages offer automated ways to calculate this. People who get excited about machine learning often apply it to problems that would have been much simpler by using linear regression in an Excel spreadsheet.

logarithm

If $y = 10^x$, then $log(y) = x$. Working with the log of one or more of a model's variables, instead of their original values, can make it easier to model relationships with linear functions instead of non-linear ones. Linear functions are typically easier to use in data analysis. (The $log(y) = x$ example shown is for log base 10. Natural logarithms, or log base *e*—where *e* is a specific irrational number a little over 2.7—are a bit more complicated but also very useful for related tasks.)

logistic regression

A model similar to linear regression but where the potential results are a specific set of categories instead of being continuous.

machine learning

The use of data-driven algorithms that perform better as they have more data to work with, "learning" (that is, refining their models) from this additional data. This often involves cross-validation with training and test data sets. "The fundamental goal of machine learning is to generalize beyond the examples in the training set." Studying the practical application of machine learning usually means researching which machine learning algorithms are best for which situations.

naive Bayes classifier

"A collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms that all share a

common principle, that every feature being classified is independent of the value of any other feature. So for example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A Naive Bayes classifier considers each of these 'features' (red, round, 3" in diameter) to contribute independently to the probability that the fruit is an apple, regardless of any correlations between features. Features, however, aren't always independent which is often seen as a shortcoming of the Naive Bayes algorithm and this is why it's labeled 'naive'." This naiveté makes it much easier to develop implementations of these algorithms that scale way up.

neural network

Also, *neural net* or *artificial neural network* to distinguish it from the brain, upon which this algorithm is modeled. "A robust function that takes an arbitrary set of inputs and fits it to an arbitrary set of outputs that are binary... In practice, Neural Networks are used in deep learning research to match images to features and much more. What makes Neural Networks special is their use of a hidden layer of weighted functions called neurons, with which you can effectively build a network that maps a lot of other functions. Without a hidden layer of functions, Neural Networks would be just a set of simple weighted functions."

normal distribution

Also, *Gaussian distribution*. (Carl Friedrich Gauss was an early nineteenth-century German mathematician.) A probability distribution which, when graphed, is a symmetrical bell curve with the mean value at the center. The standard deviation value affects the height and width of the graph.

null hypothesis

If your proposed model for a data set says that the value of *x* is affecting the value of *y*, then the null hypothesis—the model you're comparing your proposed model with to check whether *x* really is

affecting *y*—says that the observations are all based on chance and that there is no effect. "The smaller the P-value computed from the sample data, the stronger the evidence is against the null hypothesis."

outlier

"Extreme values that might be errors in measurement and recording, or might be accurate reports of rare events."

overfitting

A model of training data that, by taking too many of the data's quirks and outliers into account, is overly complicated and will not be as useful as it could be to find patterns in test data.

P value

Also, *p-value*. The probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. It's a measure of *how surprised you should be* if there is no actual difference between the groups, but you got data suggesting there is. A bigger difference, or one backed up by more data, suggests more surprise and a smaller *p* value...The *p* value is a measure of surprise, not a measure of the size of the effect. A lower *p* value means that your results are more statistically significant.

Pandas

A Python library for data manipulation popular with data scientists.

Poisson distribution

A distribution of independent events, usually over a period of time or space, used to help predict the probability of an event. Like the binomial distribution, this is a discrete distribution. Named for early 19th century French mathematician Siméon Denis Poisson.

predictive analytics

The analysis of data to predict future events, typically to aid in business planning. This incorporates predictive modeling and other techniques. Machine learning might be considered a set of algorithms to help implement predictive analytics. The more business-oriented spin of "predictive analytics" makes it a popular buzz phrase in marketing literature.

predictive modeling

The development of statistical models to predict future events.

principal component analysis

"This algorithm simply looks at the direction with the most variance and then determines that as the first principal component. This is very similar to how regression works in that it determines the best direction to map data to."

prior distribution

"In Bayesian inference, we assume that the unknown quantity to be estimated has many plausible values modeled by what's called a prior distribution. Bayesian inference is then using data (that is considered as unchanging) to build a tighter posterior distribution for the unknown quantity."

probability distribution

"A probability distribution for a discrete random variable is a listing of all possible distinct outcomes and their probabilities of occurring. Because all possible outcomes are listed, the sum of the probabilities must add to 1.0."

quantile, quartile

When you divide a set of sorted values into groups that each have the same number of values (for example, if you divide the values into two groups at the median), each group is known as a quantile. If there are four groups, we call them quartiles, which is a common way to divide

values for discussion and analysis purposes; if there are five, we call them quintiles, and so forth.

random forest

An algorithm used for regression or classification that uses a collection of tree data structures. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree 'votes' for that class. The forest chooses the classification having the most votes (over all the trees in the forest).The term "random forest" is actually trademarked by its authors.

regression

"...the more general problem of fitting any kind of model to any kind of data. This use of the term 'regression' is a historical accident; it is only indirectly related to the original meaning of the word."

reinforcement learning

A class of machine learning algorithms in which the process is not given specific goals to meet but, as it makes decisions, is instead given indications of whether it's doing well or not. For example, an algorithm for learning to play a video game knows that if its score just went up, it must have done something right.

Root Mean Squared Error

Also, *RMSE*. The square root of the Mean Squared Error. This is more popular than Mean Squared Error because taking the square root of a figure built from the squares of the observation value errors gives a number that's easier to understand in the units used to measure the original observations.

scalar

Designating or of a quantity that has magnitude but no direction in space, as volume or temperature — ***n.*** a scalar quantity: distinguished from vector

serial correlation

"As prices vary from day to day, you might expect to see patterns. If the price is high on Monday, you might expect it to be high for a few more days; and if it's low, you might expect it to stay low. A pattern like this is called serial correlation, because each value is correlated with the next one in the series. To compute serial correlation, we can shift the time series by an interval called a lag, and then compute the correlation of the shifted series with the original... 'Autocorrelation' is another name for serial correlation, used more often when the lag is not 1."

standardized score

Also, *standard score*, *normal score*, *z-score*. "Transforms a raw score into units of standard deviation above or below the mean. This translates the scores so they can be evaluated in reference to the standard normal distribution. Translating two different test sets to use standardized scores makes them easier to compare.

strata, stratified sampling

"Divide the population units into homogeneous groups (strata) and draw a simple random sample from each group."

supervised learning

A type of machine learning algorithm in which a system is taught to classify input into specific, known classes. The classic example is sorting email into spam versus ham.

t-distribution

Also, *student's t distribution*. A variation on normal distribution that accounts for the fact that you're only using a sampling of all the possible values instead of all of them. Invented by Guiness Brewery statistician William Gossett (publishing under the pseudonym "student") in the early 20th century for his quality assurance work there.

time series data

Strictly speaking, a time series is a sequence of measurements of some quantity taken at different times, often but not necessarily at equally spaced intervals.So, time series data will have measurements of observations (for example, air pressure or stock prices) accompanied by date-time stamps.

unsupervised learning

A class of machine learning algorithms designed to identify groupings of data without knowing in advance what the groups will be.

variance

"How much a list of numbers varies from the mean (average) value. It is frequently used in statistics to measure how large the differences are in a set of numbers. It is calculated by averaging the squared difference of every number from the mean." Any statistical package will offer an automated way to calculate this.

vector

Webster's first mathematical definition is a mathematical expression denoting a combination of magnitude and direction," which you may remember from geometry class, but their third definition is closer to how data scientists use the term: "an ordered set of real numbers, each denoting a distance on a coordinate axis. These numbers may represent a series of details about a single person, movie, product, or whatever entity is being modeled. This mathematical representation of the set of values makes it easier to take advantage of software libraries that apply advanced mathematical operations to the data.

- **Facebook**

几乎大多数面试过程是积极的， 49%的受访者表示他们有一个愉快的面试经历，只有 23%的受访者认为他们没有。大多数候选者是通过当前员工或招聘官推荐的。在 1-5 的比例中，整个 Google 的面试过程的难度被评定为比平均水平高一点：3.4 分。其中 5 分代表最高难度。

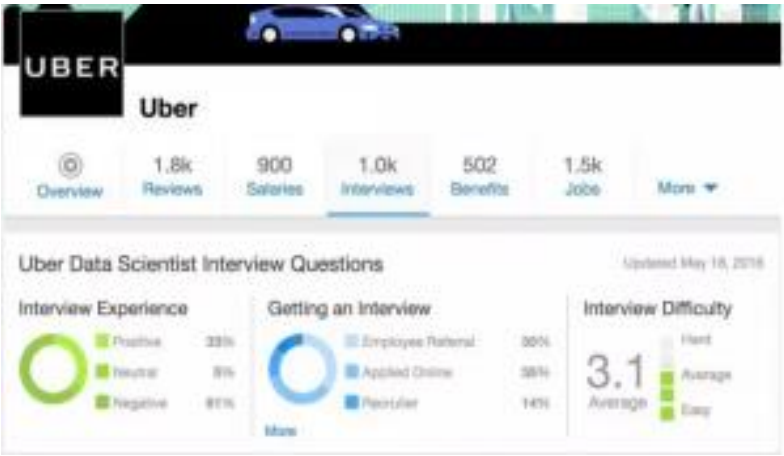
面试过程包括一轮电话面试， 一轮在家里完成的数据知识笔试，一轮屏幕共享的 SQL 的笔试和一轮公司现场面试。公司现场面试要求和团队中的每个人都进行一次 1 对 1 的面试。在整个面试过程的前期，问题多集中于 SQL, 之后多集中于机器学习和建

立 Facebook 的广告模型。开放性的场景问题多数是关于产品管理和数据科学，包括问题如怎么设计一个具体的 Facebook 的属性。根据描述，面试等待过程平均长达 3 个月，所以不用惊讶你需要花一段时间等待下一轮的结果。

Facebook 数据科学团队正在开展的工作：研究团队分享了他们目前正在做的工作：包括关于推动新的周期的研究，和人们如何在社交网络上互动的深度分析。

- **Uber**

Uber 数据科学面试有些负面反馈，有 61％的评论者说他们没有一个开心的经验。大量参加面试的人都是通过网上申请，与通过内部推荐获得面试的人数相同（35％）。面试过程的难度被评定为 3.1。



面试过程是标准化的：一轮电话面试，一轮 2 个小时的限制的笔试（分为 SQL 分析和一些操作简单数据库并回答的开放式问题），然后一轮现场面试—混合了技术问题和行为问题。技术问题围绕 Uber 的产品：你会被要求处理泊松分布，时间序列分析，以及用算法让驾驶员接受预订等问题。 Uber 的数据科学团队专注于优化一个乘客与驾驶员之间快速和省时的互动。他们的面试也是根据这个工作需求设计的。Uber 的数据科学团队正在做什么：这篇文章是 Emi Wang,一位 Uber 当前的员工介绍的 Uber 数据科学团队的日常工作。他们的工作在编写代码，做业

务分析和为新项目创建模型中交替，包括通过 Geosurge 调整供应和需求；Geosurge 是 Uber 动态定价的内部系统。

- **LinkedIn**

Linkedin 的面试经历积极反馈是消极反馈的两倍。很多候选者是通过网申获得面试邀请，所以可以试一试自己的运气！整个 Linkedin 的面试过程的难度被评定为比平均水平低一点：2.8 分。一位领英的招聘官说，面试过程包括一轮与招聘官的电话面试，一轮与团队领导的电话面试，然后一个现场面试。大多数候选者会收到一个长达 3 到 4 个小时的在家里完成的数据科学作业。Linkedin 的面试问题大多围绕你对 Linkedin 产品的兴趣，例如

如何推测员工的工资，或者如何从事一些已经建立好的属性（你可能认识的人）。Linkedin 团队非常看重 Python 和机器学习，尽管这部分知识会出现在面试的后期。前期还是多通过 SQL 和数据挖掘的问题来淘汰部分不合格的候选者。Linkedin 数据科学团队正在开展的工作：前 Linkedin 产品数据科学主管

Daniel Tunkelang 大概描述了每个人在团队里面的角色，以及团队在 2012 年的主要工作：包括更新社交圈，使其与用户更加相关，并更能代表用户的职位头衔。

- **Twitter**

Twitter 的面试经历反馈有 45%为中立，27%为积极，27%为消极。大多数申请者是通过网申拿到面试机会。Twitter 的面试难度被评定为 3.5，所以准备好要接受挑战。

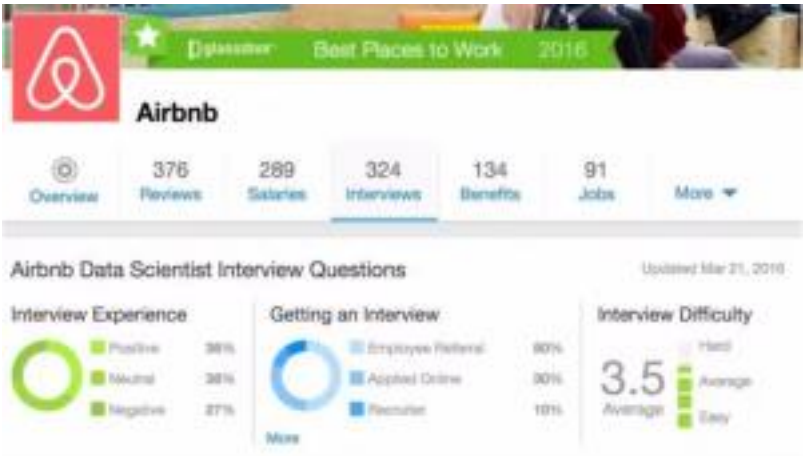
受访者认为面试结果回复得非常快，但是这个面试过程还是非常长的。首先是一轮网上的编程笔试，然后是两轮电话面试，一个关于编程，一个关于统计推理。之后公司现场面试包括两个 Skype 面试，一个关于数据科学，一个关于编程。

编程问题都是非常经典的软件工程的面试问题，但是 Twitter 的数据科学面试多数集中在开放性题目和与 Twitter 现在的工作相关的问题。候选者被测试到 A/B 测试的相关知识，同时他们用 collabedit.com 平台来做远程的编程笔试。一个候选者说他收到了很多白板问题 来测试他的机器学习理论和算法设计等知识。

Twitter 数据科学团队正在开展的工作:这篇文章是一个在 Twitter 工作了两年的数据科学家分享的个人经历。文章记录了部分工作：包括研究为什么某些国家有更高的多个 Twitter 帐户的比率和可能涉及的因素，以及有多少用户有资格获得不同的通知类型。

- **Airbnb**

36％的 Airbnb 面试经历是积极，27％是负面的。大多数面试邀请都是来自员工推荐：这一点看出 Airbnb 对他们内部推荐系统有强烈的权衡。面试的难度评分为 3.5 分。



面试过程的有少数公开的详细描述的数据，其中最有参考价值的是由 Airbnb 的数据分析主管公开的。他描述了先通过电话面试过滤部分申请者，之后是基本的数据笔试，之后是内部数据破解，然后是四次面试。这四次面试主要侧重于文化适应和与业务伙伴沟通的能力。

Glassdoor 上的评论证实了整个面试过程是恰当的。在家里完成的数据笔试主要侧重点在于 A/B 测试和结果的显著性。之后的公司内部的数据测试主要侧重于统计建模。尽管测试非常基础，但是给的时间非常少，所以你必须非常熟悉 Python 和 R 以便于能在极短的时间内最好地完成测试。Airbnb 数据科学团队和别的数据团队不同的一点是，他们非常关心候选者对 Airbnb 产品的想法以及过往的使用经历，所以务必准备好一些关于 Airbnb 应用程序使用体验的问题和你对产品的想法。
Airbnb 数据科学团队正在开展的工作：部分资料描述了数据团队如何在 Airbnb 整个团队中实现数据驱动型文化。

- **Yelp**

大多数人在通过网上申请得到 Yelp 的面试邀请。面试过程的难度被评为略高于平均水平，为 3.3 分。

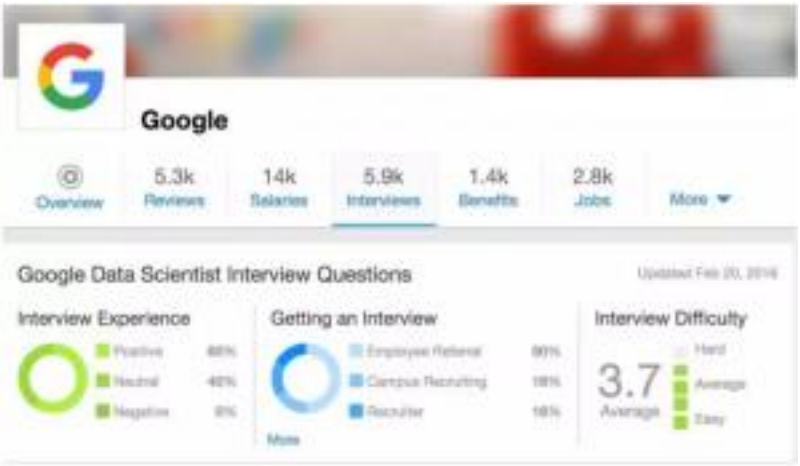
整个面试的过程如下：一轮限时的网上数据笔试，一轮电话面试，最后是一轮现场面试 – 分别与 4 个人面对面的面试。Yelp 的企业文化相当开放，员工自豪地分享他们使用的不同的工具，这一点和谷歌相似。Yelp 数据科学面试问题是相当标准化的。Yelp 数据科学团队正在开展的工作：其中一个是团队如何用深度学习对餐厅图像进行分类，以区别它们是食物的图像还是餐厅的内外部装饰的图像 。

- **Google**

大多数 Google 面试经历都是积极的，60％的受访者反馈了积极的体验。 员工推荐是获得面试的最佳方式，50％的受访者声称这是他们的获得面试的途径。 面试过程同时也被评为是所有面试中最难的：难度等级为 3.7。

最开始是一轮集中于技术的电话面试，然后是一轮高强度的现场面试—分别与谷歌的几个当前员工进行长达一个小时的面试。

电话面试混合了基本的计算机科学问题和统计问题，重点集中于 R/SQL 。现场面试问题侧重于如何切割数据。
Google 的数据科学团队正在开展工作：这篇“非官方”的 Google 数据科学博客分享了团队正在开展的大量项目，包括如何成为 Google 数据科学团队一员的初级课程。



- **J.P. Morgan**

摩根大通的候选人主要来自校园招聘，网上申请和招聘官招聘。面试难度低于平均水平：为 2.7 分 。
招聘过程开始于一轮 30 分钟电话面试，其次是与招聘经理和一个团队内的高层管理者的视频面试，最后是与几个人面对面的面试。 摩根大通的面试主要侧重于金融知识以及机器学习。 他们还强调与业务团队的沟通，要求候选人向非技术团队成员解释什么是线性回归。

摩根大通的数据科学团队正在开展工作：摩根大通使用 Hadoop 获取大量的客户和交易数据，并将其与社交媒体提及的内容结合起来，从而获得对他们所服务的客户的完整认识。

数据科学推荐读物

为了帮助大家更好的学习了解数据科学，我们找到了以下的关于机器学习和数据科学的免费电子书供大家参考学习。这些书目涵盖了关于统计基础的和关于机器学习的书籍。希望这些经典课题与最新热门课题的融合能够让你领悟到对于机器学习和大数据不一样的新的体验和兴趣。

1. Think Stats: Probability and Statistics for Programmers.

《统计思维：程序员数学之概率统计》

By Allen B. Downey

Think Stats 是一本写给码畜们的关于概率与统计学的初阶介绍类书籍。

这本书主要是介绍一些可以用来处理实际数据和讨论相关问题的基本方法。这本书讨论了一个基于美国国家卫生研究院（National Institutes of Health）数据的实际案例，来开展相关话题和知识点的讨论。这本书鼓励读者们去做一些基于真实数据集的 project。

2. Probabilistic Programming & Bayesian Methods for Hackers
《贝叶斯方法:概率编程与贝叶斯推断》
By Cam Davidson-Pilon

这本书相比于数学更注重与对贝叶斯方法论（Bayesian Method）和概率性编程的理解。贝叶斯方法论是对数学分析自然而然的估计与推论，然而贝叶斯方法论的推理非常繁杂难懂。一般情况下，关于贝叶斯推论的关键内容主要建立在概率论的两三个章节上，之后才会是真正讲解什么是贝叶斯推论。然而，按照这种讲解构架，由于贝叶斯的一些数学部分实在是很难被掌握，通常的书里只会介绍几个简单的，人为编造的案例。这些不符合真实世界的例子会让读者们有一种对于贝叶斯推论有一种 “so what” 的情绪。读者们无法认知到贝叶斯推论的重要性和实用性。事实上，这种想法只是其他作者最开始接触贝叶斯的初始理解而已。

3. Understanding Machine Learning: From Theory to Algorithms
《深入理解机器学习:从原理到算法》
By Shai Shalev-Shwartz & Shai Ben-David

机器学习是近几年来计算机领域里蹿红最快也确实有很多广泛应用的“小鲜肉”。 这本书的编写要义在于给读者一个原则性的对机器学习的介绍以及其联系到的算法案例。这本书介绍了如何通过实用且基本的机器学习和数学推导，来将原理转换为实际算法的理论解释。除了对于最基本东西的解释论述，这本书还包括了之前那些书目中没有提到的重要的课题。课题包括：计算机学习的计算复杂度，稳定性和凸性(convexity)的概念，随机梯度下降、神经网络和有结构的输出式学习的重要算法范例，以及 PAC-Bayes 和 compression-based bounds 等新兴概念。

4. The Elements of Statistical Learning

《统计学习要点》

By Trevor Hastie, Robert Tibshirani & Jerome Friedman

本书在大家都认知的一个基础框架上论述了在统计学领域上的一些重要的理论。尽管这本书的最主要主题是要讨论统计学知识，但它的重心却没有落在数学理论上。这本书为读者们提供了很多彩色插图和案例说明来阐明知识论点。这本书不仅仅对于统计学家来说很有价值，它对致力于科学工业进行数据挖掘的有志之士也有很大的阅读价值。这本书的知识网非常的广，从监督式学习（预测）到非监督式学习都有一定的设计。同时

书中还提到了神经网络，支持向量机，分级树和分级助推（这是相关话题在所有书籍中第一次被综合讨论）之类的其他话题。

5. An Introduction to Statistical Learning with Applications in R
《统计学习导论:基于 R 应用》

By Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

这本书对于统计学习基础方法的介绍。这本书是写给非数学专业的大三大四本科生，研究生和博士生的。这本书包括了很大量的 R 语言的实例。这些实例都附有如何将统计方法使用进真实世界情形设置的详细解释。这些资源对于想要成为优秀的数据科学家的人来说是很有实际练习价值的。

6. Foundations of Data Science

By Avrim Blum, John Hopcroft, and Ravindran Kannan

计算机科学的传统研究领域目前还处在一个很重要的地位，大家会使用计算机处理一些实际应用问题，然而越来越多的研究人员将目光转向了使用计算机从大量的数据库中提取和理解有用的信息。由于上述的原因，我们这本书呢就涵盖了未来四十年都会非常有用的原理学说。这些书中提到的原理就像我们四十年前会提到自动化相关理论和算法等原理一样，它们在这过去的四十年里对学生们在数据科学上的研究起到了很大的作用。

7. A Programmer’s Guide to Data Mining: The Ancient Art of the Numerati

《写给程序员的数据挖掘实践指南》

By Ron Zacharski

这本书的编写理念是引导读者通过实践编程来学习相关知识的。作者更建议读者去真正地一边通过 Python 去做书中给的练习和例子来学习书中的内容，而不是被动地去单纯阅读这本书。作者希望读者们可以积极地参与进这个编程的实战中，去尝试数据挖掘的技术。这本书是通过讲知识点分解成不同步骤来教学的。不同的步骤也就落在不同的小模块里面。当你真正学完这本书的时候，你就可以掌握一个对于数据挖掘技术的基本理解了。

8. Mining of Massive Datasets

《大数据:互联网大规模数据挖掘与分布式处理》

By Jure Leskovec, Anand Rajaraman and Jeff Ullman

你可能曾经为没有在斯坦福上过学而感到遗憾，这本书的存在可以在一定程度上弥补你心灵的缺失。它主要是基于斯坦福大学 CS246(大数据库挖掘)和 CS345A(数据挖掘)两门课来编写的。这本书的设置遵循了课程本身的设置理念，它是为没有相关经验的本科计算机学生准备的。如果读者想要去了解更深层的东西，大多数章节都附有可以让你继续阅读相关课题的参考书目。

9. Deep Learning

《深度学习》
By Ian Goodfellow, Joshua Bengio and Aaron Courville
这是一个帮助读者了解笼统的机器学习领域和深度学习课题的书。这本书的电子版已经完成，并且在网上长期免费阅读。

10. Machine Learning Yearning
《机器学习的渴望》 By Andrew Ng
人工智能(AI), 机器学习(machine learning) 和 深度学习(deep learning) 正在转变着第二产业。但是建立一个机器学习系统需要你做出以下一些决策：应该收集更多的训练数据吗？应该使用端对端的深度学习吗？如何处理与测试集不匹配的训练集？这本书对以上问题作出了回答。