

开通VIP



MS Robotics方向诚心求比较JHU和Gatech

snowsmile1211 | 微信 | 设置 | 消息 | 提醒 | 签到领奖!

新手上路 | 退出



开始答题 积分: 156 | 用户组: 中级农民-加分请看右边栏-多参与|记录|反馈

请输入搜索内容

帖子

热搜: 美国找工作 定位评估 申请总结 绿卡移民

论坛 学位+学习 数据科学 Data Scientist 炼成记录-更新完毕 | 机器学习练成记录 ...

最近看过此主题的会员



yujing91



SLX



爱学习的

中国数据智能独角兽企业  
坐标杭州 | 个推诚聘  
数据/算法/分析/研发等岗位码农求职神器Triplebyte  
不用海投  
内推多家公司面试科技公司如何  
用数据分析驱动产品开发  
\$366 off coupon code: best深入浅出AB Test  
从入门到精通  
\$366 off coupon code: bestE轮2.5亿美元融资  
一起作业诚聘  
机器学习/数据/教育等职位生  
Dream  
招聘游:

返回列表

1

2

3

4

5

6

7

8

9

10

... 38

1

/ 38 页

下一页

查看: 253178 | 回复: 387

导读 VIP瞬间解锁



我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



发消息



分享帖子到朋友圈

Data Scientist 炼成记录-更新完毕 | 机器学习练成记录 - 已开新帖 [复制链接] | 试试Instant~

K姐 发表于 2013-11-20 10:47:27 | 只看该作者 | 只看大图 ▶

网课 应用中心 留学 加入我们 免米搜索 关于

本楼: 1楼 98% (7/86)  
全局: 顶 95% (7/86)1% (1) 踩  
4% (236) 踩机器学习练成记录请移步[这里](#)

数据科学: 简单说就是, 不要靠拍脑袋下结论, 要以数据为根据, 让事实说话。

能力范畴3个词: 统计, 编程, 表述

A PhD Data Scientist: Jack of All trades, master of one.

展开说: 统计 (能探索数据, 建模, 设计实验),  
编程 (能取数据, 洗数据, 至少能Prototype自己的data solution, 懂基本大数据工作原理(MapReduce)),  
表述 (化繁为简, 口头Present, 书面写报告和论文, 作图(静态和web)) -baidu 1point3acres

简历上 (+脑子里) 如果有这些: 你找工作基本没有问题:

Ttest, Regression, ANOVA, Logistic Regression, DOE, Machine Learning, Data Mining, MapReduce, SQL, R/Matlab, Python, Java

本文主要针对IT类行业做数据科学 It does not define a data engineer. Rather, it's a close call to a "full-stack data scientist". Master this list and you will not only be able to work for established firms, but startups too.

其他偏重传统行业应用的，应该对表述要求稍高，对其他要求稍低。  
 面试之前请务必花1周时间学习对方行业的基本内容，wikipedia即可，起码做到熟悉对方行业常用关键字。  
 如果目的就是有份还可以的工作，请照单子静下心来学习。  
 如果你希望做的很好，三个方面请突出至少一个方面。  
 要学过来，需要很多时间，如果希望不太费力就做data scientist, OK, dream on!

**请不要mark一份学习清单就.Equals(学习任务已经完成了)一样，一起来学起来吧~~~~~【墙裂建议贴出你的学习计划，大家一起监督讨论，几位版主有空也会来给建议，坚持下来的有积分奖励】**

. check 1point3acres for more.

我做了一套在线[课程](#)，跟大家分享自己的经验，希望对大家有帮助

### 1. Overview (2小时的分享)

第一场2000多人参加，以后考虑不定期重开

### 2. Analytics (8小时课程)

为期一个月的一轮课程，



### 3. Experimentation (8小时课程)

为期一个月的一轮课程，

以后根据需求会有新的offering，敬请关注

### 4. 其他课程还在开发中，

有兴趣的同学可以留言，说明你希望看见的内容和其他建议等

如果有不清楚的请google.

差不多一年前看市面工作还是很混杂的样子，今天又翻了翻，估计年底账目清算，很多公司很多新职位出来了，职位要求[解析在此](#)  
 感觉现在data scientist/researcher之类职位针对性更强，能更清楚看出来到底对方需要的是什么样的人：是啥都会一点的，还是会点统计的码农，还是Machine learning，还是优化、logistics 供应链，还是会点编程的统计师。  
 (data business person 一般不叫data scientist) 主要用SQL产生报表的BI analyst 也不在此列。

学习列表一来是准备面试用，二来本来平时就是要用的。我自己学完的mark as green

打算把我自己学的一些东西总结在这里欢迎补充。不定期汇总到首楼。

如果你想收藏本帖请点首楼下方的“收藏”-> 确定 -> 然后文章会出现在“快捷导航”-> 收藏里面  
 如果没有啥具体内容要补充的，请不必回帖了。想加分的可以加分，不加也无所谓。

请别问我某校的Data Science项目如何，你三围如何能否上某校。I have no idea.

基本上是must have:

## 统计Statistics 统计和机器学习

hypothesis testing, point/interval estimation

pvalue, power, (type 1/2 error)

clt, delta method, derive coef and var(coef) etc

t-test: assumptions, remedy. 适用问题范围 basics listed above 请看这个课 <http://onlinestatbook.com/2/index.html>

glm (lm, logistic regression, anova etc): assumptions, model selection and validation, diagnostics, remedy 适用问题范围

### times series

[Forecast with R](#)

[Time Series Analysis and Its Applications](#): With R Examples (Springer Texts in Statistics)

and its [Upitt](#) course

### bayesian

[Bayesian for hackers](#) (python)

[coursera](#) Graphical Model (VERY nicely explained)

Bayesian reasoning and machine learning [book](#) (quite difficult to read)

入门: A first course in Bayes 一下就看完了, 很不错

### longitudinal, mixed model

doe: all kinds of design, response [surface](#)

(?) survival

## Machine Learning

[Coursera Andrew Ng](#). From 1point3acres bbs

[stanford](#) Statistical Learning (Tibshirani & Hastie) -- 本书还出了一个本科版, 着重动手实践, 大量R, very easy to read. recommend starting from here.

Caltech 那个 learning from Data 我没能跟下来 Please, make sure you know your logistic regression inside and out!

## Deep Learning:

See my separate thread here: <http://www.1point3acres.com/bbs/...1&extra=#pid2601595>

Learn recommender system. From 1point3acres bbs

Learn some NLP

Make sure you KNOW how things work, not just how to call a certain package in a certain language!!!

## Experimental Design / Causal Inference

This is somewhat a niche area. But as a DS, you will most likely deal with some AB tests, if you are with a reputable internet company. It is not just using some tool to compute power for a chisquare test or t-test. Be sure you know the difference between observational study and designed experiment. Be sure you know when to use which. Students from biostat/epi background will have an edge here. If you are able to handle very complex expt design, then you are opening many doors -- think multi-sided Marketplace and interfering subjects (Uber/Lyft, Airbnb, eBay), Social network (Snap/FB/[linkedin](#)) problems, think about problems that can't cleanly randomize users (opt-in, marketing campaign, mobile app feature roll out).

## Optimization & MoreIntro to linear programming

<https://www.math.ucla.edu/~tom/LP.pdf> Good and easy read.

See stanford for additional courses on convex opt.

Prof. Ferguson also has some good reading material on **game theory** [https://www.math.ucla.edu/~tom/Game\\_Theory/](https://www.math.ucla.edu/~tom/Game_Theory/)

**Udacity Intro to AI** is a great course (also one of the very first MOOC in this world) that connects the many concepts together, including particle filters, Kalman filter, HMM etc.

## 统计软件Statistical Computing: R/Matlab/Python. SAS(?)

R and Matlab 基本被业界认为是等同的。不过 Matlab is not free, Octave is free 但是不是那么好用。请考虑自学R。反正你会 Matlab 的话 pick up R 也就分分钟的事情。

如果其他语言一个都不会, 只会 SAS Base/Stat, 并且你也不想学其他的, 那也许数据科学不适合你。如果你非要用 SAS 不可, 请你至少写过 macro。SAS 的确在大数据的建模里面非常有用, 但是跟其他行业差距较大, 如果组里其他人都是 R/Py/Java 你跟他们交流起来会异常困难。另外软件很贵, 很多地方未必愿意买。

注意, 我说的是, 会 SAS 是好事, 但是不能仅仅只会 SAS。

Python: Data Analysis with Python (book), pandas

R: data.table, or plyr, lubridate, reshape2, build a R package, there are now lots of such courses on both udacity and coursera. Start from any.

know how to get data from any source (DB, web, xml, plain text, etc). From 1point3acres bbs

EDA (exploratory) - Descriptive stats udacity

Inference - udacity

Plot/explain. 1point3acres

read code from your favorite packages

-----

## 编程 : A compiled language, and a scripting language

### Python

我比较偏好Udacity一遍教一遍做quiz的方式, 光做题不讲(codecademy) 我自己好像学不清楚 Udacity CS101  
 Udacity CS 215 (Algorithm, 比Coursera Princeton and Stanford要简单, 快速过一遍不错)  
 Udacity (Peter Norvig) [CS212 Design of a Computer Program](#) 非常好, 强烈推荐

### Java 数据结构和算法

1. Udacity [java](#) (这门课我花了40小时学完) 适合连什么是函数什么是赋值都不知道的人。
2. Data structure 数据结构建议必学 python: [Problem Solving with Algorithms and Data Structures](#)  
 Java: Berkeley 61B <http://www.cs.berkeley.edu/~jrs/61b/>  
 教材是Head First Java & Data Structures and Algorithms in Java,  
 my progress bar: week 5, lab1, hw1.
3. Algorithm: Udacity Algo in Python 比较laid back, 如果不太希望费劲, 可以上这个课, 不过还是严肃点好。。。Java Coursera Algo I&II (Princeton), 如果对这个话题有兴趣,  
 不限语言 [Stanford Algo I&II](#)也很好, 两者不可相互代替。

很少会有人学的第一门语言是C#, 所以C#还真没有什么特别入门的书, 不推荐。如果没从前没学C, java, C++直接看C#的书简直无法理解  
 C++比较难, 对data scientist 来说应用也没有java广。当然如果你是大牛, plz当我没说。

### Design pattern: 地里同学推荐的:

<http://courses.caveofprogramming.com/ns-and-architecture>  
<https://www.youtube.com/playlist?list=PLF206E906175C7E07>

根据我组里面试别人, 和我在其他地方面试, 量化一下: 数科的编程到底需要什么水平?

我假定你有了上述其他的全部功底, 除非职位特别强调是统计师, 或者叫Data scientist, statistics/[analytics](#), 并且职位说明里面对代码完全一带而过, 你都可以假设, 是需要一些代码能力的。

具体水平是:

IT公司数科: Leetcode Medium要可做。所以, [刷题](#)吧。

传统公司: 不知道

如果你是码农出身, 或者做更偏向data engineer的, 要求会更高

涉及知识点包括并且不限于:

浮点溢出  
 边界情况考虑  
 改进MapReduce算法 (beyond brute force)  
 如果涉及大数据, 对时间复杂度要求会比较高 Binary search, and be prepared to talk about complexity  
 very basic DFS/BFS  
 reservoir sampling  
 string manipulation  
 if DP dynamic programming is ever asked, it will be very basic  
 basic data structures

Most likely you don't need leetcode hard

-- 其他我想起来了慢慢补

顺手学掉的小零碎:

**Regex** (a couple of hours) <http://deerchao.net/tutorials/regex/regex.htm>

**SQL** (a week) <http://www.w3schools.com/sql/> Coursera: Intro to DB

SQL 面试要考到什么程度?

如果JD是DS, 我自己没见过特别特别恶心的, 但是肯定需要懂

JOIN 比 subquery 快

COUNT DISTINCT

WHERE

GROUP BY

HAVING 什么的

advanced: 需要懂 windowing function

随便google一个准备sql面试的link 都有这些信息

SQL必须是面试过程中最没有悬念的一段了

大数据:

**MapReduce:** some knowledge    Udacity series: <http://blog.udacity.com/2013/11/sebastian-thrun-launching-our-data.html>)

Coursera: [intro to Data Science](#)

Coursera: Big data and web intelligence

learning by doing --- yes! wrote my very first reducer for real life projects!    MongoDB (udacity) (NOSQL)

Spark/Scala - try this book: [Advanced Analytics with Spark](#) (very doable and easy to follow, superb examples)

Scala推荐 Coursera: functional programming in scala - 超级好

Spark MOOC <http://www.1point3acres.com/bbs/thread-135600-2-1.html>

Book: Learning spark

**Basic Engineering:** <https://see.stanford.edu/Course> It also has great content on optimization, which is harder to find elsewhere.

## **If you want to be a DS for IT firms, then Maybe:**

[jquery/ajax](#) (start from codecademy very simple js and jquery intro, then find books) w3c school one is also really good.

web services    get basic idea of how browsers work (udacity - Website Performance Optimization)

udacity web development (build a blog) (40 hours)

SE. 1point3acres

Software Development Life Cycles (udacity, mostly videos, as a quick intro only), amazingly, this one filled lots of holes in my knowledge base. Highly recommend

Also a book is mentioned here, worth a quick flip through, unfortunately, no ebook that I found works. Martin Fowler, Kent Beck, John Brant, William Opdyke, Don Roberts-Refactoring\_ Improving the Design of Existing Code

-- this is helpful not only for working in IT, but helps overall coding style/efficiency as well. Wished I'd known earlier.

Linux

Many servers are in linux. at least familiarize yourself with the command line stuff. There's a not so good course on [edx](#).

Basic shell script or similar

jq, sed, awk-baidu 1point3acres

## **综合/分析/表述/软技能**

软技能难以表述,

技巧不是最重要, 想清楚再开口才是关键。突然发现我导师的lab页面竟然是用这些问题开头, 深感心有戚戚。

化繁为简, 高屋建瓴的表达能力: hide complex formula/engineering details, 尽量传达big picture

个人经验是, 习得这些能力最好的办法是: 去讲, 不要自顾自的讲话, 请随时关注听众是否听懂, 鼓励对方马上提问, 回答问题要选取符合对方背景的关键字, 而不是"自己熟悉"的关键字。不要用缩写, 小范围术语。多讲清楚intuition, 少堆积公式。

1. 教一门自己专业的入门课, e.g 统计学生, 去给其他专业的人讲入门统计, 例子: 请给完全不懂统计的人讲, 什么是pvalue, power, false positive, randomization, inference etc.

2. Consulting - 有些学校会有这种session, 别觉得浪费时间, 去给别人讲懂, 去看看别人用你的专业技术做什么问题, 他们的思路跟你哪里不同, 你如何理解他们, 如何让他们理解你。

3. 做presentation - 不要像专业学术会议上那样去讲, 要向给别人上101课那样讲。讲的目的, 不是展示你的专业多么复杂深奥, 不是为了impress others with your technical prowess, 而是让对方懂, 最终听取你的建议。

Data Journalism (course, starting early 2014) --- it was not as good as I expected. I do not recommend it.

作图, 静态的最好能会ggplot (a few hours), 动态的d3, 如果你会javascript, also great!, 推荐读

Nathan Yau: books visualize this & Data points, and his flowing data blog

for d3: Interactive Data Visualization for the Web . free online tutorial by author: <http://alignedleft.com/tutorials/d3/about> 真的没那么难

作图是否好看并不是关键所在, 选用合适的图标来帮助解释道理才比较重要

html (a few hours, w3c)

css (a few hours, w3c), or codecademy, or the d3 book mentioned above

javascript (codecademy as a start, a book to follow later). From 1point 3acres bbs

Rcharts/highcharts

Udacity现在也有一门新开的vis课了

**Prototype** your data products: . From 1point 3acres bbs

mean stack. <https://thinkster.io/angulartutorial/mean-stack-tutorial/>

起码把AngularJS学了, 这个不光做数科有用。

R open CPU. R Shiny (limited usage with free version). If you are not into Angular, try the flask+React stack, 上手的确很快

(关于flask, udacity有课, react自学即可, 可以参考udacity 关于components的课)

虽然我们不是要做前段开发, 但是看起来也得至少有个半吊子前段, 请学习这MM的经验, 超赞 <http://www.1point3acres.com/bbs/thread-104335-1-1.html>

**Design:** (optional but nice to know) 如果没有兴趣请至少看 ([组合起来好看的颜色](#)) 如果你有兴趣让图好看, 请花一个周末翻看这几本:

1. Before and After
2. Nondesigner's design book
3. Don't make me think
4. The Wall Street Journal Guide to Information Graphics

#### Research/publication:

sharelatex (invite enough users to get free versioning) /writelatex.com  
Go to conferences, see what people are working on. Read their papers.  
如果你想找某些类型的工作, 上linkedin找到组员, 泛读他们的paper

Domain Knowledge: google/wikipedia is your friend

### =====

### 整体思路:

Doing Data science (book)  
Data Science in Business

===== . 1point3acres

### other 一些我感觉不太费时间但是会有用的小东西

excel, power pivot etc  
科普类的书: (都很简单易读)

大数据到底是啥? ?? [《Big Data: A Revolution That Will Transform How We Live, Work, and Think》](#)  
和很近似的一本 [《Automate This: How Algorithms Took Over Our Markets, Our Jobs, and the World》](#)  
随便翻翻就好了

然后当然还有Nate Silver [《The Signal and the Noise: Why So Many Predictions Fail-but Some Don't》](#)

===== . check 1point3acres for more.

Case study: Twitter data analytics <http://tweettracker.fulton.asu.edu/tda/>. check 1point3acres for more.

===== . 1point3acres

有人推荐的 MS data science 学习curriculum <http://datasciencemasters.org/>

=====

大家给我推荐的帮助整理思路, 用正确的方式做事的工具: It's more important than you think!!

<http://software-carpentry.org/lessons.html>

coursera reproducible research, 学转knitr, 不要copy paste anything

-baidu 1point3acres

Udacity Git Course (最好, 没有之一)

=====

最后, 没有什么比亲自干活和得到feedback更有用。

数据科学是一种 apprenticeship model, 找合适的人带着做事, 成长会很快。

数学, 统计, 数据科学

○ 评分

参与人数 **104** 大米 **+1247** 理由

[收起](#)

 hwh510444568 + 1 给你点个赞!

	huankuaidexiao + 3	给你点个赞!
	SuperGuy10 + 3	很有用的信息!
	sophie_1mu3fen + 3	很有用的信息!
	zhang.chi1 + 3	很有用的信息!
	Verdi + 3	很有用的信息!
	happydreamer + 3	很有用的信息!
	dazeze122222 + 2	给你点个赞!
	chenyx95 + 3	很有用的信息!
	hazelzhou + 3	很有用的信息!
	gaotianhang1022 + 3	很有用的信息!
	evanlys1993 + 3	很有用的信息!
	anna2016 + 3	很有用的信息!
	hohomuma + 3	很有用的信息!
	wbbssr + 3	很有用的信息!

[查看全部评分](#)

上一篇: [【求助】这个分数是不是废了](#)

下一篇: [Doing Data Science - by Rachel Schutt; Cathy O'Neil](#)

#### 相关帖子

藕丝 二轮店面 面筋	【求refer】Cornell Statistics PhD Wait List 求 refer
yahoo 实习包, 求米看面经	鸭虎二面
2/28/2019 Oath intern 电面	cmu metals被wl, 真心求refer!!
yahoo new grad apm 店面	凉凉的yahoo/藕丝 二轮店面
Waitlist Cornell Orie 求不去的朋友refer一个 谢谢	纽约雅虎面经
有需要refer UT MSITM 的吗	雅虎auth 实习面经
oath电面贵经	Yahoo/oath19summer-intern一面(视频)
结束失学 nyu cs	Oath/Yahoo backend实习
雅虎汤博乐店面	藕丝校园面经
yahoo 纽约面经	Oath 一轮店面

#### ○ 本帖被以下淘专辑推荐:

- [数据科学](#) | 主题: 42, 订阅: 108
- [统计找工相关](#) | 主题: 31, 订阅: 35
- [统计master](#) | 主题: 15, 订阅: 16
- [Stats/DS](#) | 主题: 4, 订阅: 5
- [JOB](#) | 主题: 22, 订阅: 0
- [数据科学Data Science/Analytics](#) | 主题: 17, 订阅: 59
- [BA](#) | 主题: 156, 订阅: 33
- [成为数据分析师](#) | 主题: 5, 订阅: 7
- [DS](#) | 主题: 12, 订阅: 3

 收藏 1457  评分  分享 37  淘帖 9

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.

如果觉得我的统计、数科经验有用, 欢迎来论坛分享你的学习心得, 跟大家一起刷公开课, 分享求职心得等。

为活跃论坛, 没有发过进程或总结帖的, 私信问问题恕不回复。





我的人缘 0

23主题

596帖子

9738积分

发消息



我的人缘 8

892主题

1万帖子

7万积分

一匹黑马



发消息

最后，申请相关问题我回答不了。

《数据科学求职面试40+真题讲解》

回复

微信

使用道具

举报

earlgrey 发表于 2016-3-10 12:04:23 | 只看该作者

来自 2楼

本楼： 0% (0)

0% (0)

全局：顶 97% (570)

2% (14) 踩

li3939108 发表于 2016-3-4 15:00

K姐你好，本人ECE（CE track） PhD，coding基础还可以，底层到应用层啥都知道一点，自己拿Ruby on Rails写 ...

Casella and Berger打基础很好，不过对于找工作准备面试不太够

All of Statistics 这本书也是master level统计课的教程，里面的topic更多一些，也更现代一点，当然不是所有topic都有用

linear regression, DOE, ML 的问题都有可能出现  
既然你还在学校，可以找找学校这方面的课，看看有没有时间去蹭课，他们用的什么教科书  
这些课我在学校都上过，复习用的都是以前的课本

ML: 可以搜 stanford cs 229， 那组notes写的还挺清楚

评分

参与人数 2 大米 +43 理由

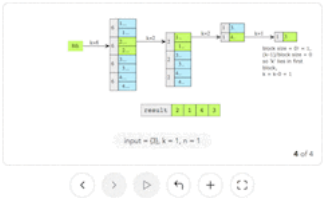
[收起](#)

li3939108 + 3 感谢分享!

K姐 + 40

[查看全部评分](#)

80道经典面试算法题Step by Step画图详解  
提供C++/Java/Python/JS等多种语言答案  
12道算法题免费试看。平均每题六毛钱。  
一亩三分地用户额外20%折扣。



回复

评分 举报

楼主 | K姐 发表于 2017-10-6 23:38:50 | 只看该作者

来自 3楼

本楼： 0% (0)

0% (0)

全局：顶 95% (4786)

4% (236) 踩

已经pivot away from DS -baidu 1point3acres

最近看的一些内容其实已经不是DS了，我也不敢说就已经"学到头"了，但是感觉深度上再去努力ROI就不大了，而个人对广度现在更有兴趣些

1。  
深度学习，多少学到能跟别人瞎聊的程度，下面需要hands on做点项目玩  
Ng那个课比较浅但是经验部分还可以在等fast ai的pytorch

另外有帖子总结DL。

2 infra、系统设计  
往码农的系统设计方向靠拢



工作里面还是有蛮多参与做数据系统的机会

广大数据科学的同学，请珍惜这种机会。很多时候码农年轻的时候不太能参与到设计的，数据类因为很多是新东西而是legacy code，反而很年轻的时候就可以参与到设计。

信息很多但是毕竟杂。最近开始看某书，以后会总结。

**3 代码和软工基础.** check 1point3acres for more.

多多少少能写但是始终没有顺溜到有什么需求就随手能实现，不惧的程度。

很希望还是能走过这个坎。

有点疑虑刷题并不能解决我的困惑。但是也没有找到更好的办法，也没有找到特别喜欢的项目想亲手做。

信息很多了轮不到我去收。

#### 4 软技能

信息很多了轮不到我去收。

一亩三分地管理员 + 所有产品线 PM

玩转Data Science

ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复

评分 举报

 楼主 | K姐 发表于 2017-10-18 21:52:09 | 只看该作者

 来自 4楼

本楼：  0% (0)  
全局： 顶 95% (4786)

0% (0)   
4% (236) 踩

## 已经pivot away from DS

强化学习和优化 (AI)

优化的TBD

强化学习：

Udacity那个intro AI 不错但是动手部分有点少

另外那个深度一点的课程不缴费完全做不了作业但是讲的还可以

Berkeley: CS188 Artificial Intelligence (edx) 录像在youtube都有

这个里面有很多其他好资料 <http://rll.berkeley.edu/deeprcourse/>

怕不见了先抄来这里

CS189 or equivalent is a prerequisite for the course. This course will assume some familiarity with reinforcement learning, numerical optimization and machine learning. Students who are not familiar with the concepts below are encouraged to brush up using the references provided right below this list. We'll review this material in class, but it will be rather cursory.

- Reinforcement learning and MDPs
  - Definition of MDPs
  - Exact algorithms: policy and value iteration
  - Search algorithms. From 1point 3acres bbs
- Numerical Optimization
  - gradient descent, stochastic gradient descent
  - backpropagation algorithm
- Machine Learning
  - Classification and regression problems: what loss functions are used, how to fit linear and nonlinear models
  - Training/test error, overfitting.

For introductory material on RL and MDPs, see

- [CS188 EdX course](#), starting with *Markov Decision Processes I*
- [Sutton & Barto](#), Ch 3 and 4.
- For a concise intro to MDPs, see Ch 1-2 of [Andrew Ng's thesis](#)
- David Silver's course, [links below](#)

For introductory material on machine learning and neural networks, see

- [Andrej Karpathy's course](#)



我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马



发消息

- [Geoff Hinton on Coursera](#)
- [Andrew Ng on Coursera](#)
- [Yaser Abu-Mostafa's course](#)

#### Related MaterialsJohn's lecture series at MLSS

- [Lecture 1](#): intro, derivative free optimization
- [Lecture 2](#): score function gradient estimation and policy gradients
- [Lecture 3](#): actor critic methods
- [Lecture 4](#): trust region and natural gradient methods, open problems

#### Courses

- [Dave Silver's course on reinforcement learning](#) / [Lecture Videos](#)
- [Nando de Freitas' course on machine learning](#)
- [Andrei Karpathy's course on neural networks](#)

#### Relevant Textbooks

- [Deep Learning](#)
- [Sutton & Barto, Reinforcement Learning: An Introduction](#)
- [Szepesvari, Algorithms for Reinforcement Learning](#)
- [Bertsekas, Dynamic Programming and Optimal Control, Vols I and II](#)
- [Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming](#)
- [Powell, Approximate Dynamic Programming](#)

#### Misc Links

- [A collection of deep learning resources](#)

一亩三分地管理员 + 所有产品线 PM

玩转Data Science

ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复

评分 举报

 楼主 | [K姐](#) 发表于 2017-10-18 21:58:07 | 只看该作者

 来自 5楼

本楼: **【顶】** 0% (0)  
全局: 顶 95% (4786)

0% (0) **【踩】**  
4% (236) 踩

## 已经pivot away from DS 之 data infra/data system

下阶段看的一个话题已经不是data science /analytics范畴了，但是也不是随便一个SWE就知道的范畴，因为主要还是跟数据相关，仍然放在这个版了，这个贴跟之前发的都不一样，不建议DS的人盲跟。如果你属于业务导向，business导向的，完全不需要看这些。analyst也完全不需要看。=====

Why?

在DS、DA这个领域玩了好一阵子以后，感觉除了DL其他基本上可以做到融会贯通了，希望技术上有更宽的发展，因为感觉更深的深挖，好像能跳槽的地方就一只手数过来并且还不一定想去了。If I dig any deeper, I may be cornering myself into a narrow niche with few patrons. Also I have always been interested in seeing connections between things.

具体应用上，最大的感受是，常常方法并不一定需要多深，真正的hurdle 常常是一地鸡毛的事情，比如上游数据质量，速度，下游数据应用在业务上的场景，速度（？）什么的。

学学周围码农在干什么，有助于合作，有助于参与数据系统设计，有助于设计别人没有想到的系统和使用场景 -- 周围懂数据和业务使用的人，他们经常完全不懂系统和实现，而懂系统和实现的人，可能没有谁比我更懂数据的了。。。中间这个gap也影响到一些做事方式。To bridge the gap 应该是很有趣的。

先占个座，学习过程中会继续发心得和总结

一亩三分地管理员 + 所有产品线 PM

玩转Data Science

ML Engineer by profession.



我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马



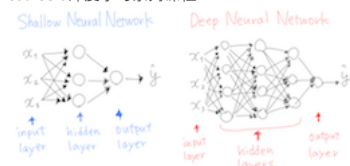
发消息

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

coursera深度学习系列课程



回复

评分 举报

楼主 | K姐 发表于 2017-10-18 22:49:11 | 只看该作者

来自 6楼

本楼: 【顶】 100% (2)

0% (0) 【踩】

全局: 顶 95% (4786)

4% (236) 踩

转眼这个帖子已经快四年老了，几点update

- DS unicorn已成真

当年写这个帖子的时候，有说法是DS三条腿：编程，统计和domain，很少有人能三条都强。venn diagram中间那一块被认为不存在，是unicorn。如今这个状态已经改变。已经有新的一批人才被培养出来。他们可能有MS/BS in CS 然后去读一个stat/OR phd，但是导师来自两个系，或者EE PhD出身，要码能码，要推公式能推公式。或者单独在quantitative领域出身（统计，数学，物理，econ，IEOR）但是业余达成不错的码水平。

而Domain这件事，无非是兴趣，坚持和积累的问题。跟前两条并不冲突，一般工作个几年就有了。

- 领域逐渐成熟和分化

DS仍然是一个新兴领域，但是很多大公司的数据文化开始形成。从前说DS没人知道你具体是analyst还是ML eng。但是现在，如果你是G家QA，大家能猜到你是统计大拿但是不见得会编程，深度不错广度未知；如果你是FB product analyst，大家会认为你平常一般写SQL和沟通较多，产品感觉不错，但是统计深度不见得深，基本不会码，基本不会机器学习；如果你在L家做ML，你的title更大的可能是software engineer而不是DS；其他很多家也有类似说法。

不同出身的人，在一个个新鲜的科技公司里面占据了早期位置，导致了不同公司完全不同的数据文化。有的公司虽然叫DS但是基本做的仍然是DA的活，有的公司明确区分DS和DA并在薪酬和发展空间上完全区别对待。-baidu 1point3acres

- 行业的一些个人理解

大体上，DS行业仍然会有不错的发展，但是我预计越来越会实现码农和数据的分化。DA算是比较清晰定义了。但是目前市面上号称做数据产品的人，整体滥竽充数的仍然是大多数，凭着年轻和热情随随便便设计一个产品，看起来都work但是其实数字错的离谱的例子，哪怕在很大的新兴公司内，也比皆是，甚至可能是主流。这方面，大厂可能比新厂领先上好几年甚至10几年。

这些缺陷多来自早期团队主力人员经验的缺乏，和数据产品的复杂性+难以甄别。跟一般软件产品不一样，看起来好像work，但是其实差了十万八千里的例子非常常见。

如果使用人都不知道需要去检查结果对不对，和要如何去检查结果对不对，自然缺乏对质量的重视和鉴定能力。他们经常也不懂得appreciate别人的经验，也无法理解这些差别会对业务造成多大的影响。

当然没人愿意承认现成系统里面存在这么严重的问题，因此在需要改进的时候还是逃避为主。

解决办法要么是一开始就花大价钱请好专家，（对于新公司，这个真不一定合适）

现在出现的新公司经常并不是技术导向，而是产品，设计，运营等导向，这种情况下，技术专家未必愿意早早进去，而且早期也的确并不需要over engineer。

但是至少需要保证能记得早期做的trade off，在找到product market fit后，需要定期revisit当初的tradeoff，这个时候的解决办法，只能是勇于换血。

大胆的一些预测：未来有做系统的，但是越来越多的系统会是个data intensive系统。目前鱼龙混杂，质量参差不齐的情况会慢慢得到改善，但是会经历有很大的浪费和痛苦。DA类似business于技术层面的API，仍然存在，暂时无法想象能用AI取代。DA和码农之间会存在一些精通数据系统的专家architect，负责设计符合业务的数据intensive产品，这些architect会精通数据系统自身的特性，并且善于在业务需求和系统可靠性之间找平衡。DA的薪酬会与另外两种人拉开距离，但是在business方向完全可以有很好的发展。

今年的TFX和鲲鹏已经展示出比较完备的机器学习系统。除了传统监督学习之外的很多如今需要各家单独开发的基础数据工具，未来几年会成为缺省自带状态。其中有些是我自己每换一家公司就教别人一次，零零星星有公司做，但是离拿来就可用还有相当大从差距，很多不小的公司仍然做成一坨的内容包括：

- 可靠的数据平台，从定义和产生log到存储，ETL，etc  
目前这部分还ad hoc到匪夷所思的地步



我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



发消息

- 自动的多种质量检测 (TFX里面号称有很多)  
目前不存在, 基本是靠人品和手工检查
- mobile rollout 平台  
FB和U有, G估计也有
- (AI controlled deployment): monitoring and anomaly detection, 自动修正, alert 反馈  
有几家有, 无general 系统
- real time/streaming analytics system  
有, 但是不是plug and play的, 需要大量domain knowledge 才能正确setup和使用
- end to end实验平台  
貌似没有深度结合产品的, optimizely只有前端。一直疑惑为什么MS EXP不单独spin off开公司
- 自动insight finding, 自学习的系统代替许多如今的手工操作  
没有自动系统, GA里的是个还不错的开端但是coverage 远不够
- intuitive 的visual analytics 涵盖无法自动提取的信息, 帮助人肉眼识别  
不知道, 不熟
- 足够flexible的ML 平台  
有点了
- 监督学习高效获取label的系统  
没有, 可以做
- 根据业务场景强化学习和优化的系统  
貌似难以generalize。。。这就是目前的holy grail了, general AI

一亩三分地现在已聚集一大批数据方面的talent, 相信大家的经历就可以较完备的map out 常见公司的数据文化, 帮助大家跳槽的时候能有清晰的选择。

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.

如果觉得我的统计、数科经验有用, 欢迎来论坛分享你的学习心得, 跟大家一起刷公开课, 分享求职心得等。

为活跃论坛, 没有发过进程或总结帖的, 私信问问题恕不回复。

最后, 申请相关问题我回答不了。

**免费看**六节系统设计面试题: Grokking the System Design Interview

System Design Interviews: A step by step guide

Designing a URL Shortening service like TinyURL [Preview](#)

Designing Pastebin

Designing Instagram [Preview](#)

Designing Dropbox

Designing Facebook Messenger

Designing Twitter


Designing Youtube or Netflix

Designing Typeahead Suggestion

Designing an API Rate Limiter ("New")

回复

评分 举报

 楼主 | [K姐](#) 发表于 2017-11-12 00:59:21 | 只看该作者

 来自 8垓

本楼: [【顶】](#) 100% (2) 0% (0) [【踩】](#)  
全局: 顶 95% (4786) 4% (236) 踩

最近接触的一些神书:

[Sutton Reinforcement learning](#)  
[Bengio/goodfellow的deep learning](#)  
[Design Data intensive applications](#)  
Linkedin的[statistical methods in recommender system](#)

学不完的焦虑, 只能是不去多想, 天天争取看一点。恩。这样。

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.



我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马



发消息

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

#### 去Fiverr找HR修改简历



回复

评分 举报

楼主 | K姐 发表于 2017-11-25 16:17:44 | 只看该作者

来自 10楼

本楼：【顶】 100% (1)  
全局： 顶 95% (4786)

0% (0) 【踩】  
4% (236) 踩

学不完的焦虑里面，整出了一丢丢眉目的样子。check 1point3acres for more.

#### 1. learn to rank / recsys / bayesian / opt

就是传统计算机专业会学，但是统计系似乎没有教太多的部分：

之前那些书都是神书，就是个人看书速度不太行，听video做quiz比较适合我

目前coursera上有一批这种内容，打算还是咬牙上了算了，不然永远是个不明不白，没有懂的很透彻。From 1point 3acres bbs

优化那部分不知道够不够，如果不够就留到下阶段去弄

还有个定价的课看起来也有趣，也许往经济学方向看看也会有意思。check 1point3acres for more.

最近的一个项目愣是找econ，统计，cs，ee，OR，ML的人都问了一圈，同一些基础的东西，在各个行当有不同的叫法和应用，有趣的要命。

#### 2. data infra/ data sys /data eng/ building data products

也是神书+coursera上一些课也许可以evaluate

更多是实践，下面一年中肯定会做这些工作

#### 3. deep learning

<http://www.1point3acres.com/bbs/thread-200846-1-1.html>

Andrew Ng的 intro DL 尤其是course 3, 4, 5 仍然是很值得上的好东西

下面就是打算手工弄下pytorch with fast.ai

肯定能在工作里面找到好应用的，到时候贴出来

一亩三分地管理员 + 所有产品线 PM

玩转Data Science

ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

#### 全球28万学生4.7/5分推荐：The Web Developer Bootcamp



**The Web Developer Bootcamp**  
300 lectures - 43.5 hours - All Levels  
The only course you need to learn web development: HTML, CSS, JS, Node, and More! | By Colt Steele

\$14.99

\$199.99

4.7

(93,262 ratings)

回复

评分 举报

楼主 | K姐 发表于 2018-2-19 15:17:26 | 只看该作者

来自 11楼

本楼：【顶】 100% (2)  
全局： 顶 95% (4786)

0% (0) 【踩】  
4% (236) 踩

Data engineering 基础的一些资料

Yandex: Big Data for Data Engineers Specialization <https://www.coursera.org/specializations/big-data-engineering>

Essentials: HDFS, MapReduce and Spark RDD

Analysis: Hive, Spark SQL, DataFrames and GraphFrames

Applications: Machine Learning at Scale

Applications: Real-Time Streaming

Services: Capstone Project



我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马



发消息



我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马



发消息

配合 [Designing Data Intensive Application](#) 一书不错

Big data 的书其实非常多，但是多到简直感觉看不过来，Hadoop的大象书其实当年也看过，一直也没有觉得理解的多清楚  
Spark的两本也都看过，也是感觉没有看的特别懂  
Kafka Definitive Guide还没来得及看  
感觉不懂这些系统的情况下试图设计ML系统就是个joke，会浪费无穷时间反复犯前人早就犯过错

如果还有空的话，

Intermediate: [Coursera Cloud computing \(5-10hr/week X 5 week X 5 courses, + 2 capstone projects\)](#)

Course 1 covers a ton of classic concepts that I see day in and day out without grokking

(C++)

. From 1point 3acres bbs

Hands on: [6.824: Distributed Systems](#) (full semester, 4 months)

Edx: <https://www.edx.org/course/architecting-distributed-cloud-microsoft-devops200-9x-1>

Edx: Each 5hrs/w X 5 weeks

<https://www.edx.org/course/reliable-distributed-algorithms-part-1-kthx-id2203-1x-0>

<https://www.edx.org/course/reliable-distributed-algorithms-part-2-kthx-id2203-2x>

CMU 15312 of MCDS

Book: [Andrew Tanenbaum : Operating Systems](#)

Book: [Andrew Tanenbaum : Network](#)

我应该不会把这些都看，先列在这里，能补多少补多少吧。

我也只是数据的使用者，并不是data infra设计者，但亲眼看见不懂系统的人做ML完全不成样子，希望自己不要瞎hack things together

一亩三分地管理员 + 所有产品线 PM

玩转Data Science

ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问题恕不回复。

最后，申请相关问题我回答不了。

Udemy评价最高的Mobile App课程：205个视频教你做app



**React Native - The Practical Guide**  
205 lectures • 16.5 hours • All Levels  
Use React Native and your React knowledge and take your web development skills to build native iOS and Android Apps | By Academichy | 10,000+ ratings

**\$14.99**

~~\$169.99~~

4.7

回复

评分 举报

Zzzed 发表于 2017-8-24 22:51:16 | 只看该作者

推荐

本楼: **【顶】** 100% (12)

0% (0) **【踩】**

全局: 顶 98% (49)

2% (1) 踩

现在网上MOOC关于Data Science的种类繁多，选择太多往往无从下手，对于转行的同学总想在最短的时间内获得最容易理解而且实用的知识，我最为其中之一深有体会，现在就结合我自己的经历说下自己在这方面的心得。

在我看来，Data Science/Analytics 大致需要掌握以下几方面的技能：

## 1. SQL, 数据库相关的技能

这个是所有从事数据分析的第一步：获取数据，而绝大部分的数据储存在数据库中，所以SQL的技能很关键，事实上也是以后也会占用你工作的大部分时间。

SQL不难，但是想要快速熟练的掌握光靠背几个 select, from, where, group by 是远远不够的，最好的联系方法是能一边写一边看得出的结果，从而搞清楚每条语句实际在背后对数据做了什么操作，逻辑是什么。

SQL也是数据分析面试时重点考察的方面，Google, Facebook, Uber, Slack等等这些大的科技公司都会去着重考察，不需要你会很fancy的命令语句，但是会让你利用简单的命令语句去实现很复杂的逻辑关系，这方面的资源比较入门级的有 SQLZOO 和 W3 School的SQL部分，这两个相对来说好快速上手，而且都是我前面说的可以让你一边写SQL一边看你query出来的结果，这样会让你对命令语句具体对数据本身做了什么。

进阶的资源有微软在edx上的一门MOOC: Querying with Transact-SQL, 这门课也适用于初学者，不过学习的时间要长一些，因为内容会讲的深一些（比如window function 和 table expression）

## 2. 统计的基本原理

大部分传统的机器学习的算法来自于统计学，而且统计学的知识也被大量的用在了数据探索阶段（Explanatory Data Analysis）和工作中各种



我的人缘 0

5 15 100  
主题 帖子 积分

发消息



各样的Statistical Testing上面

这方面就是传统的统计知识，尽量选一些名牌大学的通俗易懂的基础统计课即可。

### 3. Data Science/Machine Learning Modeling

这块的课程最多，但也最难选，因为很多课程要么太注重理论，需要有很好的数学基础才能理解，要么就是相对来说太过简单，下面是我觉得蛮好的课程，兼顾了理论深度，理解难度和实践程度。

Udemy: Python for Data Science and Machine Learning Bootcamp

这门课的特点是讲解很清晰，信息量很大，讲师既讲了Python编程也讲了些ML的算法知识，不过相对来说不是很深。

Edx: Analytic Edge

如果你是个对理论数学化的东西不大感兴趣，只注重怎么把ML的算法应用在实际中，那这门课是很好的入门课。

这门来自MIT的神课介绍每个算法时都是通过一个相应的现实中真实应用的案例来讲的，而且讲的通俗易懂，全部课程的语言为，也很容易上手

Udacity: Intro to Machine Learning

这门课是Google X 实验室的创始人 Sebastian Thrun（同时也是Udacity的创始人）讲授的，全面的涵盖了主流ML的算法，中间每讲一个新的算法，都会穿插了很多小练习帮助你巩固新学到的知识，而且Sebastian作为业界大牛，对ML的讲解也很清晰直白易懂。

Stanford Online: Statistical Learning

如果你想探究ML算法背后的数学理论基础，那这门来自斯坦福的神课就是你的不二选择，虽然课程的数学理论涉及较多，但是只要跟着两位教授(两位大牛，其中一位发明了大名鼎鼎的LASSO Regression)一步步来，还是比较容易懂的，本课程也采用了较易上手的R作为编程语言以上这些都是Data Science/Analytics 的入门课，欢迎各位大牛继续补充！

当然还有很有名的coursea上的JHU的data science系列，我在这里就不多描述了。

希望能帮到你！

回复

评分 举报

 楼主 | [K姐](#) 发表于 2017-11-25 17:00:15 | 只看该作者

推荐

本楼： [【顶】](#) 100% (4)  
全局： 顶 95% (4786)

0% (0) [【踩】](#)  
4% (236) 踩

这个帖子已经四年啦，我还朝着成为full stack 数科人才的方向，努力前进。幸运的是一直的努力有回报、一路前行的路上也一直快乐为主旋律。

过去遇到点问题经常需要一个人纠结半天。如今即使很复杂的问题，也可以在熟人里面飞速找到很有启发的讨论。

过去遇到难题需要自己吭哧吭哧生啃下来。如今想学的东西，找几个人稍微讨论一下就学会了。不敢说轻松的飞起，但是学习效率那的确已经今非昔比。

过去只看得见离散的技术问题。如今对行业的发展，挑战和机遇，多少也形成了自己的看法。虽然对技术仍然充满热情，但是深刻体会到技术服务于企业在于产生价值，也能在价值和技术兴趣之间找到不错的平衡。

经历那么多的困难吃了那么多苦，却仍然对这个行业兴趣不减反增。四年前开贴的时候曾经感叹，牛人之所以牛，是因为聪明并且有genuine passion，受兴趣推动的持续努力，是任何外界驱动力比不了的。我聪明不足用兴趣算是弥补上了。

上班拿着还不错的工资，做着自己很喜欢的事情，一直持续在学新东西，从来都可以主动加班，加的开开心心。

越来越多的认识不同type的人，这把年纪了能有那么多忘年交的新朋友，能不顾你我的讨论我们共同的爱好。

能够学以致用，把自己感兴趣的技术应用于解决跟大家生活息息相关的各种场景。

这些我都要感恩。

感谢一路一起走来的小伙伴们~~~~这些年了面基过了好多，也有不少有了各种各样错综复杂的social connections，有正在，曾经，甚至再次一起共事的同事了：)

希望本帖能帮助有数科梦想的人 ^\_^

一亩三分地管理员 + 所有产品线 PM

玩转Data Science

ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。



我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马



发消息





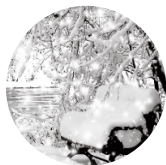
我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



发消息



我的人缘 0

29 140 499  
主题 帖子 积分



发消息

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复

评分 举报

楼主 | [K姐](#) 发表于 2018-4-20 13:26:30 | 只看该作者

推荐

本楼： [【顶】](#) 100% (3) 0% (0) [【踩】](#)  
全局： 顶 95% (4786) 4% (236) 踩

最近在公司花了些时间仔细阅读了数据平台（不是数据科学平台）的文档  
觉得受益匪浅

几种好用的学习资料列一下

1 youtube 看见不懂的名词，google完仍然读了跟没读一样的，其实youtube经常有几分钟讲清楚的

2 lightbend 这些书还是很实在，很良心的 <https://info.lightbend.com/rs/558-NCX-702/images/COLL-ebook-Fast-Data-for-Streaming-Apps.pdf>  
短小可读。虽然只谈streaming data 并不是最全面的，但是毕竟这个算是这种平台里面比较高级的配置了

3 就是咬牙坚持读，反复问人，多问几个不同的人。之前一直也觉得没有特别清楚，最近忍着集中读了一段时间后，似乎有点明白了。  
暂时还不确定上手能力有多重要：个人还是挺挣扎的。试图跟coursera上面一个课，作业debug慢的要死，进展缓慢  
但是conceptually似乎开始形成知识体系了

开心的记录个人的小小突破

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。

最后，申请相关问题我回答不了。

回复

评分 举报

七夜雪 发表于 2014-9-13 16:36:02 | 只看该作者

推荐

本楼： [【顶】](#) 100% (3) 0% (0) [【踩】](#)  
全局： 顶 100% (14) 0% (0) 踩

. 1point3acres

本帖最后由 七夜雪 于 2014-9-13 16:38 编辑

背景：学过C, javascript，不过是本科前两年，基本忘光。MATLAB用的多，少量STATA的经历（不知道这软件自由度这么低学校为啥这么喜欢）。EE PHD+ECON MASTER DOUBLE MAJOR, some experience in econometrics

目标：一年之内学完Python, Java, R, HTML5, Javascript, CSS, Machine Learning, MapReduce, SQL

大方向：Python和Java预计花时间最多，现在开始学习熟练。R不打算花巨多时间，准备有个大概的了解。HTML5系列明年再处理，到时候准备租个网站边学边弄。Machine learning准备用Python来实现（主要看这个帖子：<http://blog.renren.com/share/231...>  
[ose\\_time=1410188191](#)），会花一定时间。剩下的暂时计划不到。

短期计划（3个月）：Python已学习1个月，上了google的课程+小K贴出的算法和数据结构（graph没学完），自己写code把数据结构都实现了一遍（CS同学说主要就学数据结构）。学了recursive programming后花了一天写了个解数独的程序。graph学完后转战<Python for data analysis>，同时开始上手JAVA(用Core Java的书),并用JAVA实现基础数据结构。零碎的时间看看以前C的课件，主要熟悉pointer，然后看一些介绍概念的公开课。

Note: 其实最开始的时候很抵触coding，不过现在无奈coding就像是会开车一样，不知不觉变成了一个必须的技能。data science需要，就连EE的硬件也希望你会，于是我也找不到继续逃避的理由了。SIGH

○ 评分

参与人数 1 大米 +60 理由

收起

anonym + 60 坚持的不错，再接再厉！

[查看全部评分](#)

可叹停机德，堪怜咏絮才。

回复

评分 举报



我的人缘 0

5 39 94  
主题 帖子 积分

发消息



我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



发消息

wenbo5565 发表于 2017-7-24 13:33:36 | 只看该作者

推荐

本楼: 【顶】 100% (2) 0% (0) 【踩】  
全局: 顶 100% (4) 0% (0) 踩



求教学习的问题。之前读了一个统计的master后再美国工作了1年多, 主要在清理数据和用tableau做visualization. 马上准备读第二个master,想加深对machine learning的理解和能力。毕业以后想从事和predictive modeling/machine learning有关的工作。现在的水平是能用python和sklearn等package参加一些kaggle的比赛 最好成绩能达到15%-20%的水平 但是仍然感觉大多数的algorithm是blackbox. 请教该学习或者怎么进步? (之前的统计主要是frequentist的角度, 今后准备多修一些Bayesian和CS方向的课。请问没有data structure和algorithm的经验适合上那种作业都是数学推导的design and analysis of algo的课么?还是要先修data structure在考虑上design and analysis of algo) 多谢。

回复

评分 举报

楼主 | K姐 发表于 2016-12-24 00:12:03 | 只看该作者

推荐

本楼: 【顶】 100% (2) 0% (0) 【踩】  
全局: 顶 95% (4786) 4% (236) 踩

添加一些关于OR的内容:

我自己也是OR盲, 这部分仅供参考

<https://www.quora.com/Are-there-good-online-courses-for-Operations-Research>

Nikos Makrymanolakis, M.Sc., ph.d (cand.) in the area

Some very good and relevant courses about OR subjects in coursera:

- \* [Discrete Optimization](#) by Professor Pascal Van Hentenryck
- \* [Algorithms, Part I](#) by Kevin Wayne and Robert Sedgewick
- \* [Algorithms, Part II](#) by Kevin Wayne and Robert Sedgewick
- \* [Algorithms on Graphs and Trees](#) by Alexander S. Kulikov and Michael Levin
- \* [Algorithms: Design and Analysis](#) by Tim Roughgarden

Most of the algorithms covered in the above section, are OR used algorithms. The discrete optimization course is excellent, focus entirely on optimization (you will love the professor).

Feng Mai, Assistant Professor at Stevens Institute of Technology

Operations Research is a broad field. For optimization I would recommend

Prof. Stephen Boyd's convex optimization (available on YouTube) and  
Prof. Pascal Van Hentenryck's discrete optimization (coursera).

<https://orc.mit.edu/academics/course-offerings>

Somewhat older list

<http://www.orcomplete.com/internet/enesbilgin/open-courses-on-operations-research>

The list from stanford

<https://see.stanford.edu/Course>

评分

参与人数 1 大米 +105 理由

收起



anonym + 105

查看全部评分

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.

如果觉得我的统计、数科经验有用, 欢迎来论坛分享你的学习心得, 跟大家一起刷公开课, 分享求职心得等。

为活跃论坛, 没有发过进程或总结帖的, 私信问问题恕不回复。



我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



发消息



我的人缘 0

43 1341 6196  
主题 帖子 积分



发消息



我的人缘 0

1 137 201  
主题 帖子 积分

发消息



我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



最后，申请相关问题我回答不了。

回复

评分 举报

楼主 | K姐 发表于 2015-6-29 11:22:09 | 只看该作者

推荐

本楼：**【顶】** 50% (1) 50% (1) **【踩】**  
全局： 顶 95% (4786) 4% (236) 踩

\(•̀•́)/ scala 代码可以开始干正事了!!!!

回复

评分 举报

nibuxing 发表于 2013-11-20 10:49:57 | 只看该作者

20楼

被K姐抢先了= =我一直也想等我会的再多一点写一篇自己养成的养成记录，不过谢谢K姐的分享，继续努力！希望在不久的将来能有更大的收获。

○ 点评

K姐

一起来补充内容吧？ 发表于 2013-11-20 10:50

回复

评分 举报

pureds 发表于 2013-11-24 13:15:42 | 只看该作者

21楼

本楼：**【顶】** 100% (1) 0% (0) **【踩】**  
全局： 顶 95% (23) 4% (1) 踩

多谢分享，确实感到DS/BA之类职位在市场上的定位更加清晰起来，虽说是交叉学科，但短时间内把统计，IT，和商业都熟练或精通是太难了，需要从自身的专业背景出发，做好定位，一步一步来积累。

回复

评分 举报

楼主 | K姐 发表于 2013-11-27 03:03:33 | 只看该作者

22楼

本楼：**【顶】** 0% (0) 100% (1) **【踩】**  
全局： 顶 95% (4786) 4% (236) 踩

bayesian methods for hackers  
<https://github.com/CamDavidsonPi... Methods-for-Hackers>

发消息



我的人缘 8

892 主题 | 1万 帖子 | 7万 积分

一匹黑马



发消息

回复

评分 举报

楼主 | K姐 发表于 2013-12-12 09:13:19 | 只看该作者

23楼

本楼: 【顶】 50% (1)  
全局: 顶 95% (4786)

50% (1) 【踩】  
4% (236) 踩

## Introduction to Hadoop and MapReduce

<https://www.udacity.com/course/ud617>

这课上完了。-baidu 1point3acres

讲解比较简单, 作业做起来略麻烦, 帮助上手笔记在这里:

Install VM as said (use winRAR in windows, 7zip will fail)

next steps:

read this:

<http://forums.udacity.com/questi...at-to-do-next#ud617>

especially this [http://www.youtube.com/watch?v=c\\_cJKZ4vzhA&t=57](http://www.youtube.com/watch?v=c_cJKZ4vzhA&t=57)

watch this:

<https://www.udacity.com/course/v...8873795/m-309382595>

the difference between file system on linux and on hdfs!!!!

even there's a local file directory called data, still need to create one on hdfs:

hadoop fs -mkdir data

hadoop fs -ls

Found 1 item

drwxr-xr-x - training supergroup 0 2013-12-11 17:16 data

then there's a HDFS folder called data.

now put the actual data into HDFS:

hadoop fs -put purchase.txt data (1st purchase.txt is the file on your local Filesystem, 2nd data is HDFS folder)

then you can check you do have this:

hadoop fs -ls data

Found 1 items

-rw-r--r-- 1 training supergroup 211312924 2013-12-11 17:17 data/purchases.txt

Run:

hs ../code/mapper.py ../code/reducer\_f2.py data/purchases.txt outdata2

packageJobJar: [../code/mapper.py, ../code/reducer\_f2.py, /tmp/hadoop-training/hadoop-unjar8573115774818496995/] []

/tmp/streamjob8981780528938293292.jar tmpDir=null

13/12/11 17:33:16 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.

13/12/11 17:33:17 WARN snappy.LoadSnappy: Snappy native library is available

13/12/11 17:33:17 INFO snappy.LoadSnappy: Snappy native library loaded

13/12/11 17:33:17 INFO mapred.FileInputFormat: Total input paths to process : 1

13/12/11 17:33:17 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/mapred/local]

13/12/11 17:33:17 INFO streaming.StreamJob: Running job: job\_201312111650\_0004

13/12/11 17:33:17 INFO streaming.StreamJob: To kill this job, run:

13/12/11 17:33:17 INFO streaming.StreamJob: UNDEF/bin/hadoop job -Dmapred.job.tracker=0.0.0.0:8021 -kill job\_201312111650\_0004

13/12/11 17:33:17 INFO streaming.StreamJob: Tracking URL: [http://0.0.0.0:50030/jobdetails.jsp?jobid=job\\_201312111650\\_0004](http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201312111650_0004)

13/12/11 17:33:18 INFO streaming.StreamJob: map 0% reduce 0%

13/12/11 17:33:30 INFO streaming.StreamJob: map 12% reduce 0%

13/12/11 17:33:33 INFO streaming.StreamJob: map 19% reduce 0%

13/12/11 17:33:36 INFO streaming.StreamJob: map 26% reduce 0%

13/12/11 17:33:40 INFO streaming.StreamJob: map 32% reduce 0%

13/12/11 17:33:43 INFO streaming.StreamJob: map 40% reduce 0%

13/12/11 17:33:46 INFO streaming.StreamJob: map 47% reduce 0%

13/12/11 17:33:49 INFO streaming.StreamJob: map 50% reduce 0%

```

13/12/11 17:34:01 INFO streaming.StreamJob: map 75% reduce 0%
13/12/11 17:34:02 INFO streaming.StreamJob: map 81% reduce 17% -baidu 1point3acres
13/12/11 17:34:05 INFO streaming.StreamJob: map 88% reduce 17%
13/12/11 17:34:08 INFO streaming.StreamJob: map 95% reduce 25%
13/12/11 17:34:11 INFO streaming.StreamJob: map 100% reduce 25%
13/12/11 17:34:17 INFO streaming.StreamJob: map 100% reduce 69%
13/12/11 17:34:20 INFO streaming.StreamJob: map 100% reduce 75%

```

```

hadoop fs -cat outdata1/part-00000
Baby    57491808.44
Books   57450757.91
. check 1point3acres for more.
get data out from HDFS to local
ls
code data
[training@localhost udacity_training]$ mkdir outdata-baidu 1point3acres
[training@localhost udacity_training]$ cd outdata/
[training@localhost outdata]$ hadoop fs -get outdata2a/part-00000
[training@localhost outdata]$ ls
part-00000

```

For quick tests, make some sample data (I just copied 20 lines from ~/udacity\_training/data/purchases.txt). Save it as sampleData.txt in your code directory.

```
head -40 purchase.txt > sample.txt
```

Then in a terminal, in the code directory, you can run

```
./mapper.py <sampleData.txt >mappedData.txt
```

-baidu 1point3acres

and then

```
./reducer.py <mappedData.txt
```

for me it's more like this:

```
python ./mapper_f3a.py <../data/sample.txt >../data/mappedData.txt
```

```
python ./reducer_f3a.py <../data/mappedData.txt
```

一亩三分地管理员 + 所有产品线 PM

玩转Data Science

ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问题恕不回复。

最后，申请相关问题我回答不了。

回复

评分 举报

 楼主 | [K姐](#) 发表于 2013-12-12 13:23:14 | 只看该作者

24楼

高分悬赏 **udacity Hadoop with python 求debug**

本楼: **【顶】** 0% (0)

0% (0) **【踩】**

全局: 顶 95% (4786)

4% (236) 踩

Final 作业的最后一步。check 1point3acres for more.

chained MR

症状如下:

如果我手工把第二步数据下载下来用这个办法跑: 可以得到正确结果:

```
python mapper_f23.py <../outdata2/part-00000 >toreducer2; python reducer_f23.py <toreducer2
```

这个是udacity自己论坛里面教的debug方法，屡试不爽



我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



发消息



我的人缘 8

892 1万 7万  
主题 帖子 积分

一匹黑马



发消息

同样文件，无论是直接读上次MR的结果，还是我下载然后重新上传，都有如下错误：

```
hs mapper_f23.py reducer_f23.py data/part-00000 outdataF2outc
packageJobJar: [mapper_f23.py, reducer_f23.py, /tmp/hadoop-training/hadoop-unjar3723990990451201080/] []
/tmp/streamjob2507367586069302810.jar tmpDir=null
13/12/12 00:20:04 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
13/12/12 00:20:05 WARN snappy.LoadSnappy: Snappy native library is available
13/12/12 00:20:05 INFO snappy.LoadSnappy: Snappy native library loaded
13/12/12 00:20:05 INFO mapred.FileInputFormat: Total input paths to process : 1
13/12/12 00:20:05 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/mapred/local]
13/12/12 00:20:05 INFO streaming.StreamJob: Running job: job_201312111650_0036
13/12/12 00:20:05 INFO streaming.StreamJob: To kill this job, run:
13/12/12 00:20:05 INFO streaming.StreamJob: UNDEF/bin/hadoop job -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201312111650_0036
13/12/12 00:20:05 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?jobid=job\_201312111650\_0036
13/12/12 00:20:06 INFO streaming.StreamJob: map 0% reduce 0%
13/12/12 00:20:11 INFO streaming.StreamJob: map 100% reduce 0%
13/12/12 00:20:36 INFO streaming.StreamJob: map 100% reduce 100%
13/12/12 00:20:36 INFO streaming.StreamJob: To kill this job, run:
13/12/12 00:20:36 INFO streaming.StreamJob: UNDEF/bin/hadoop job -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201312111650_0036
13/12/12 00:20:36 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?jobid=job\_201312111650\_0036
13/12/12 00:20:36 ERROR streaming.StreamJob: Job not successful. Error: NA
13/12/12 00:20:36 INFO streaming.StreamJob: killJob...
Streaming Command Failed!
```

回复

评分 举报

 楼主 | [K姐](#) 发表于 2013-12-12 13:28:44 | 只看该作者

25楼

本楼： [【顶】](#) 0% (0)  
全局： 顶 95% (4786)

0% (0) [【踩】](#)  
4% (236) 踩

两步MR. From 1point3acres bbs

mapper1. just output the filename and 1  
reducer1: simple word count reducer

--> output of step 1 MR is (filename, count)

mapper2:read output of step1, output (\_dummy\_, filename, count)  
reducer2:get all the keys called \_dummy\_, get the max of count, output. check 1point3acres for more.

核心也就这么一点点：

```
if thisCount > maxcnt:
    maxcnt = thisCount
    maxfile = thisKey
```

然后就输出了。

我无法想象为什么local run works but hs fail 了  
看样子也是跑出来结果了，但是结果无法stream?

这一步到底是试图把什么东西写到什么地方去呢？

```
13/12/12 00:20:36 ERROR streaming.StreamJob: Job not successful. Error: NA
13/12/12 00:20:36 INFO streaming.StreamJob: killJob...
Streaming Command Failed!
```

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.

如果觉得我的统计、数科经验有用，欢迎来论坛分享你的学习心得，跟大家一起刷公开课，分享求职心得等。

为活跃论坛，没有发过进程或总结帖的，私信问问题恕不回复。



我的人缘 8

892	1万	7万
主题	帖子	积分

一匹黑马



发消息



我的人缘 8

892	1万	7万
主题	帖子	积分

一匹黑马



发消息




我的人缘 0

43	1341	6196
主题	帖子	积分

最后, 申请相关问题我回答不了。

回复

评分 举报

 楼主 | K姐 发表于 2013-12-12 13:40:27 | 只看该作者

26楼

本楼:	<b>【顶】</b>	0% (0)	0% (0)	<b>【踩】</b>
全局:	顶	95% (4786)	4% (236)	踩

holy cow.....

错误是这句话前面不小心加了一个空行  
#!/usr/bin/env python

结果就找不到python

error msg not helpful at all

<http://stackoverflow.com/question/...example-not-working>

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.

如果觉得我的统计、数科经验有用, 欢迎来论坛分享你的学习心得, 跟大家一起刷公开课, 分享求职心得等。

为活跃论坛, 没有发过进程或总结帖的, 私信问问题恕不回复。

最后, 申请相关问题我回答不了。

回复

评分 举报

 楼主 | K姐 发表于 2013-12-28 04:29:45 | 只看该作者

27楼

本楼:	<b>【顶】</b>	0% (0)	0% (0)	<b>【踩】</b>
全局:	顶	95% (4786)	4% (236)	踩

上完了Udacity 很入门的java, 基础真的是讲的非常好的  
可惜内容稍浅。如果有后续课就好了。

暂时下面打算跟Berkeley 61B, 自己看Head first Java

<http://www.cs.berkeley.edu/~jrs/61b/><http://www.youtube.com/watch?v=Q...2A1049C&index=1>

Joyce好心share给我课程资料已经几年了, 现在才学起, 真是惭愧。

一亩三分地管理员 + 所有产品线 PM  
玩转Data Science  
ML Engineer by profession.


如果觉得我的统计、数科经验有用, 欢迎来论坛分享你的学习心得, 跟大家一起刷公开课, 分享求职心得等。

为活跃论坛, 没有发过进程或总结帖的, 私信问问题恕不回复。

最后, 申请相关问题我回答不了。

回复

评分 举报

 nibuxing 发表于 2013-12-28 04:58:30 | 只看该作者

28楼

小K 发表于 2013-12-28 04:29

上完了Udacity 很入门的java, 基础真的是讲的非常好的  
可惜内容稍浅。如果有后续课就好了。

我也正在跟CS61B, 上这门课的话我觉得Head first暂时不用看啦, 一开始教授会重新讲一点java基础, 还是挺好的。





snowsmile1

211

快速发帖

发消息

回复

评分 举报

下一 页 »

返回列表

1

2

3

4

5

6

7

8

9

10

... 38

1 / 38 页

下一页

上传

高级模式

发表回复

本版积分规则

提醒：发帖可以选择内容隐藏，部分板块支持匿名发帖。请认真读完以下全部说明：

■隐藏内容方法 - 不要多加空格：[hide=200]你想要隐藏的内容比如面经[/hide]

■意思是：用户积分低于200则看不到被隐藏的内容

■可以自行设置积分值，不建议太高（200以上太多人看不到），也不建议太低（那就没必要隐藏了）

■建议只隐藏关键内容，比如具体的面试题目、涉及隐私的信息，大部分内容没必要隐藏。

■微信/QQ/电子邮件等，为防止将来被骚扰甚至入肉，以论坛私信方式发给对方最安全。

■匿名发帖的板块和方法：<http://www.1point3acres.com/bbs/thread-405991-1-1.html>

☐ 回帖后跳转到最后一页