和自我介绍相关的：

1. 顺着话题接着问了分析过什么类型的数据，数据长什么样，有多复杂。
2. 问了你最有impact的project
3. 你觉得你喜欢linkedin的哪个产品

然后问我熟不熟悉linkedin的product最喜欢linkedin的什么product为什么喜欢。。。。。

Product Sense:

1. How to improve profile completion rate?
2. How to persuade the team to built a new feature？如何甄别一个用户的资料是否完整，如何提升用户资料的完整度。用什么matrix.完全的Product sense 题目。what product feature you want to improve, how do you persuade the team to build this feature
3. 识别经常出差的人

frequent traveller 问题, 问了怎么define frequent traveller, 找到之后有什么用，用什么feature 去找

在definition 上卡了一下，只考虑到了frequency of location change, 没考虑到distance of location change

问application, 说了可以personalized content, 打广告，面试官问还有什么，有点想不出来，随便说了个recommend connections

1. 工作申请数量下降分析原因

if the amonunt of job applications decrease significantly, as a data scientist, what would you do to find out why? 然后你提什么，面试官就会顺着你的思路问下去。这题大概讨论了20min。

1. 你想提高哪一个产品怎么说服PM
2. 500 Million users 你认为的5种user type

Estimate sales man的总数

如果在上述的user type中还有学生类别的话，你觉得他的比重如何

还问了linkedin 的different type of user 都有哪些。。linkedin 总共500m users，怎么样estimate each segment的market size

1. 如果你的service在过去10周的星期一的数据都很稳定，突然下降60%了，你如何找原因。这个问题就根据场景里面有什么可用的分类变量来解答就好了。我说了可能有seasonality或者location之类的原因
2. 那么如果在出现这个情况的前一周，我们上线了新的ML模型，如何判断是不是因为这个模型导致的。我能想到的就是AB testing，然后就讲述了下，可能是因为模型的评判标准没有跟客户实际需要或者我们的business goal一致导致的。
3. product: application rate 降低你怎么调查和解决
4. 有没有什么linkedin 的feature 你想improve? why, 怎么measure, 怎么说服大家。。
5. how would you classify job seekers among linkedin users. 然后也讨论了大概20min
6. 然后怎么建模设计job ranking 给individual user
7. opportunity sizing, 怎么样估计各个公司的sales professionals有多少然后给他们推销license
8. 假设Linkedin要识别出potential companies, 去让sales reps联系他们在linkedin上面做广告，都是围绕这个题展开的1. 你会用哪些features识别这些公司？2. 假设你有了data set，怎么证明你是对的？...
9. frequent traveller的问题，主要交流了怎样识别frequent traveller，怎么用这个信息

case interview 基本上是，穿插了很多细节问题。背景：

1.假设linkedin有一个算法能够找到经常旅行的人，你认为这个产品有什么用？

我说可以推广linkedin learning，　发温馨小广告等等。但后来有面试官提示我，可以卖给delta让他们去推广他们的产品。。。。

2.后退一步，如何设计这个算法？需要什么样的数据来找这样的人？

我举例：ｉｐ地址变化，按某种频率，　profile上的职业title，　industry等等

第3个更新：假设你已有这个metric,怎么用？怎么依据metric判断算法好坏？4.怎么找最优的FP, FN 的cut off rate. 5假设现在你已经有了这个系统，并且run了一年了，模型的优化是基于你提出的metric（user engagement相关的一个）。但现在我们想把这模型的结果卖给delta帮他们做广告，就意味着模型的目的变了（原来是为了提高用户engagement的，现在模型的准确性变得非常重要。）你认为该怎么改进模型？6.假设模型是个clustering算法，没有label data。怎么确定哪个算法好？7.你认为avg number of location changed per week 的distribution是什么样？描述看看？

16. 第三个讨论了LinkedIn Learning的product，从这里开始可能就跟interviewer的经历更相关，也跟我的经历相关。

面试人先问了我有没有online learning的experience，怎么看各个网站，我比较了coursera，edx，udemy，udacity

我认为coursera和edx和大学合作更多，udacity和公司合作更多，udemy是individual instructor，学知识去c和e，学习找工作技能去后两家

讲完以后我又顺便讲了以前做个的一个machine learning project，是预测上课的人是否会完成课程，怎么设计model，用什么feature

接着是讨论LinkedIn B2B Learning Product，如何提高engagement，总结了几点：

1）设计forum，或class channel，增加user interaction

2）educate customer，因为b2b用户和b2c用户不通，motivation不强，需要帮助用户找到合适的课程

3）survey learner，了解学习者的兴趣点，提高课程质量

17. 面试的内容主要是围绕Job Module，如何设计metrics，以及如何检验模型是否有效：

1. LinkedIn的job search给出的job ranking依靠某一个模型，如果想要考察这个模型的performance，你会看哪些metrics？

2.假设，这个模型，对于student这个category给出的job ranking 效果并不好，原因是什么；如果是high senior people，原因又会是什么。

3.如果你要设计这个模型，你会考虑哪些因素和metric呢？

4. Email Campaign：A/B testing，hypothesis testing，解释alpha，beta的意义，如何选择testing的sample size。

18.

第一问top 5 linkedin user type, 然后问我怎么区分marketing和sales professional，然后问我500 million linkedin users 怎么estimate多少sales professional。我说了我的逻辑，和一般consulting公司面试思路差不多。然后大概估算了一下，得到一个数字，他听完之后，问我是不是reasonable。我自己一想，estimate应该太高了，就回去重新考虑各种assumption。然后他问有什么别的思路来estimate。我猜测他也许可能想听modeling，然后就正经的开始brainstorm各种feature去定义一个sales person。第二大题是说，如果作为ds的你突然发现current week的number of job applications 突然下降了你怎么analyze？我先问了他很多问题，比如是不是specific to 哪个platform啊之类的。yes - 你要怎么继续dig，no - aggregate across all platform 怎么继续挖问题是啥？我说可能seasonality，他说那你怎么确定是time effect？如果不是seasonality，那还会是什么导致的？然后我又接着脑补，反正我脑补一样啥，基本接下来的问题就是if yes，你怎么弄，if no，what else。我其实不确定最后有没有回答到他的点上。第三题，他是这么说的，你最想改变linkedin product的啥？我看时间剩的不多，就随口一说那个recommendation email，jobs you might be interested in。没想到给自己挖了个坑。他follow up问，假如你写了个新的你觉得是厉害的algorithm，你怎么convince PM去真的launch这个algorithm。我就说了一系列down stream 的我认为的key metric，然后measure这个change的impact。他问那你怎么validate你的algorithm可以达到你说的improvement。我当时以为他要开始问我ab test了，没想到我刚开始说了开头，他也知道where I am going 就制止了我，说还不能launch，你就有个prototype你怎么prove。我就说好那我就自己做simulation。又是几个来回的follow up的问题。然后他问我如果可以identify那些不喜欢linkedin recommendation 的人，怎么去cluster他们。我就high level说了几个点。然后时间真的差不多了，他也就没问下去了。总体来说，我面完的时候我都觉得我肯定凉了。出乎意料的第二天，recruiter告诉我feedback positive，要去onsite了。

19.

how can you do if number of complains increased from 2000 on avg. to 3000 yesterday? how would you find where the problems are?

20.

1. What are the top 5 types of LinkedIn users?

2. LinkedIn has 500 million users, how to estimate number of sales professionals?

3. How to predict who will be signing up to a LinkedIn new product in an email compaign?

21.

Product case study。这部分由于面试我的面试官是负责metrics那部分的DS。所以围绕这个主题开始聊天。第一个是如何设计metrics来评判你的算法在job you may interested上的表现。第二个是如何根据不同的用户种类判断那些用户的profile的完成度是否高。如何建议他们完成哪些information。主要还是不断聊。不断根据情况和商业调整自己的想法。

22.

一个feature，是用来提高team efficiency 的，怎么知道这个feature 有没有用。

23.

如何重新设计message，怎么测试等等一系列，比较standard

24.

How to find a metrics to check if Linkedin is the primary source of some one posting on social network?

Initially I answer # of post on average per day. Eventually, I realized and answered the one interviewer prefer: # of hours between two posts. Then follow up: any cons? Answer: need to check if a post is valid or not. such as some one just post couple non-sense words every hour should not be counted. Follow up: how to detect? check the length of the post, check # of share, # of comment, # of likes, # of read for a given post. Conculsion: there are two sides of Ecology of a social network: 1) content creation, 2) content consumption. Follow up: what is the secondary metrics you want to use?

25.

Classic question: we saw different usage pattern of ios and android. More people use homepage tab in IOS, fewer people use homepage tab in android. Question: how to you analyze? answer: 1) check if there is population differences between ios and android users. How to check? demography, blahblahblah~~~. 2) Yes, there is population differences, how to eliminate (again propensity scoring, or segmentation based on demography), 3) No, there is no population differences. What is posible reason? (answer: App is poorly designed in android,etc, follow up: how to check thru data? see their website usages on desktop). 4) we now see in ios 80% use homepage, the other 20% use other different tabs, such connect, jobs,etc. But in android, the distribution of homepage tab and other tabs are more evenly. 4.1 - how to design a graph to show that? (barchart with side by side comparison, or pie chart). 4.2. - what is the reason you can think about? provided one hint: think about out of app behaviors. My answer - there maybe more out of app connecction thru other apps in android (such as connect, sign, jobs). So people can directly log in to that tab instead of go to homepage tab. True answer: andoird users tend to click to (Linkedin) notifications, which directly bring them to given tab. while IOS use then to not.

26.

下来主要就是考如何evaluate这个model。Interviewer有追问如何和PM communicate result, 怎么quantify impact

27.

怎么看待LinkedIn China，哪里有opportunity？

28.

说很多人登录linkedin不成功是因为之前后缀是前公司或前学校的email失效了，她问我as a data scientist, how do i solve this problem? 九月那会儿没练过这种case study，所以不知道怎么approach，我就很一个接一个的提了一些方法，但是她总举一些不适用的case反驳我。但是其实说实话，到现在我也不太知道怎么给这个问题提出一个acceptable的solution。最后几分钟给我提问的时间，我问team现在的一些challenge是什么，她就讲了又一个问题，说希望能够减少linkein上的harassment，fake news之类的有negative impacts的信息以及传播，然后又一副期待眼看我希望我提出一些想法。。。。

29.

考查了一个problem diagnostic的问题，大致是说他们design了一个东西，希望用户们在遇到xx问题的时候可以通过用这个东西来解决问题，但是用户们都不用，都打客服，问我怎么分析和解决。

A/B Testing:

1. AB testing 两个问题。一个问题如果给你一个email campaign你如何设计AB testing，如何计算sample size跟时长。第二个是你同时有一个email campaign跟new msg function，如何做ab testing

我觉得这个就是先算你的sample size ,比如1万个Page View, 然后再算你平均每天的page view 比如是10万个，然后你assume每天想放多少个%of traffic in the A/B testing,这里一般建议放少一点，因为有safety/stability的各种考量，那么假设我放1%的traffic，那么需要run十天，最后还可以说说很多东西都有weekday vs weekend effect 所以这样run两周最好，可以完整的看到effect

我认为他想问你的事根据已经sampling的状况跟hypothesis + z test 算出你需要的sampling size 那你的sampling size 就决定你要testing的时间长短

2. Discuss one project. then want to test if a given treatment works for something (I forgot). what is key metrics? short term, long term? Follow up: what if your two group is not comparable? (answer: propensity scoring). Follow up: what if your propensity scoring is not working in both high end and low end score region?

第一题是统计, 主要考了binomial，然而后来问了一个我没有想到的问题：如果fb有1百万个用户，1000个广告，并且每个广告能够被每个用户看到的概率是uniform distribution，那么至少能看到一个广告的用户数目是多少，要求不能笔算，要心算估测出一个具体的数字。这个公式好写，就是用二项分布就ok，但是怎么算呢？我其实猜到了跟泰勒展开有一点关系，把小量近似成0来做，但是当时大脑一片空白，有点方..面试官感觉不是想刁难我，后来给我解释了怎么做。这里我想说，大家面试的时候遇到了问题千万别慌，诚恳的说自己这里有点不明白，大局为重，之后好好发挥还是有机会的。 产品题也比较新，问的是fb group功能如何衡量表现，以及需要哪些数据来选择是focus在小group还是大group(小group就是人数比较少的那种)，你认为哪种更值得被投资和继续发展

Stats and ML:

上来就是让我画各种distribution，然后问我各种关于distribution的问题，但是都不难。后来就是个比较standardized的ab testing的问题。比较特别的是他问了我同时更改两个features我应该怎么isolate the effect，然后我就说做四组实验，然后做chi-square test。瞎说的也不知道对不对

1. <https://stackoverflow.com/questions/1986859/unbiased-random-number-generator-using-a-biased-one>
2. + bias ／variance tradeoff
3. l1/l2 regularization

word2vec那两个model CBOW + Skip-gram 和解释negative sampling

bag VS boosting

markov property. independent from path，只和前面的一个状态量有关。

先过简历，列举classfication algorithm，　然后要具体讲一个logistic regression，要求记得数学公式，知道怎么解决，l1/l2 regulation，　各有什么利弊，我简历上写作过clustering，所以具体问了clustering的算法，选这个的理由，问了ｔｏｎｓ　ｏｆ　ｃｏｎｃｅｐｔｓ，　ｐｒｅｃｉｓｉｏｎ，　ｒｅｃａｌｌ，　ｒｏｃ什么的，设计一个怎么提取job description 里面的skill，答了ｎ－ｇｒａｍ＋ｃｒｆ，ｃｏｄｉｎｇ也很基本，ｉｓｌａｎｄ　ｃｏｕｎｔ，　ｄｆｓ小哥就说可以了。

上来先问了一堆Machine Learning的基本东西。因为楼主是做neural network，就问了一堆neural network的东西，问neural network和SVM（我简历提到了）的原理，楼主用了一堆数学的东西，经面试官提醒，再用一些intuitive的东西进行解释。还有问什么是overfit，如果avoid，还有performance（training/testing）如何measure（false-positve之类的东西）还算比较基础吧，大家认真准备就好。

2. 大概一小时吧先是各种基础知识，logistics regression的目标函数，L1，L2 regularizer，cross validation，SVM，然后问了我一下，怎么把SVM的output按照概率输出，这个当时我没答出来，但是我瞎说了一个，我说假设svm的residual按照高斯分布，然后我们训练的时候多训练一个高斯模型好了……并且当时我跟他讲，我这个是自己瞎想的……事后发现店面的面试官没挂我，而且其实我答得也不算特别特别的离谱，只不过把我的高斯模型换成logistic regression就可以了

3. 如果计算[linkedin](http://redirect.viglink.com/?key=a1aa544c3b328def412653f9fc432107&u=http://linkedin.com) member的average secondary connection