# FB DS Interview

## How to prepare:

- Product design: Use the product design framework CIRCLES Method to explore possible personas and articulate the use cases.
- Execution: get things done and make critical decisions. AARM Method, Google HEART framework for User Experience
- leadership + drive : What is your biggest weakness?

### High-Level Talks

Alex Schultz — How to Get Users and Grow Youtube
Chamath Palihapitiya — how we put Facebook on the path to 1 billion users Youtube
Alistair Croll — Lean Analytics: Using data to build a better startup faster Youtube

FB Official Numbers: Link
Instagram stats: Link Link2

### Examples: StellarPeers

How would you measure the success of Facebook Stories? Link
How would you find the cause of a 15% drop in Facebook Groups usage? Link

### 经验总结帖:

如何准备Data science analytics interview, case study详解 by Alice007 Link
总结自己如何cracking the Data Challenge by Alice007 Link
product sense的经验+鸡汤 Link
非典型性面经（Facebook, LinkedIn, Pinterest） Link

### 地里相关面经

**1.**Link Link2 Link3 Link4

SQL – Composer
table composer，3 columns: userid l event l date,
event 包括 enter/post/cancel (enter就是开始在composer里面写内容，cancel就是开始编辑但是没有post而是终止了)
(1) what is the post success rate for each day in the last week?

```
SELECT
        date,
        ROUND(IFNULL(SUM(CASE WHEN event = 'post' THEN 1 ELSE 0 END)/
                                SUM(CASE WHEN event = 'enter' THEN 1 ELSE 0 END), 0), 2) AS
success_rate
FROM
        composer
WHERE
        DATEDIFF(CURDATE(), date) <= 7
GROUP BY 1;
```

(2) 在第一题的基础上，又给了一个table：user，4 columns: userid | date | country | dau_flag{0, 1}。其中dau_flag表示daily active or not
what is the average number of post per daily active user by country today?

```
SELECT
    country,
    ROUND(IFNULL(SUM(CASE WHEN C.event = 'post' THEN 1 ELSE 0 END)/
COUNT(DISTINCT U.userid), 0), 2) AS avg_posts
FROM
    user U
    LEFT JOIN
    composer C
    ON U.userid = C.userid AND U.date = C.date
WHERE
    date = CURDATE() AND U.dau_flag = 1
GROUP BY 1;
```

Product Sense
Given a tracking metric: Avg_ActiveUser_post -> The average active user post number per day:
We have a drop from 05/01/2018 to 06/01/2018 below：
05/01/2018 Avg_ActiveUser_post 3.0
06/01/2018 Avg_ActiveUser_post 2.5
Question: What is the reasons/factors you would think that cause this drop ?
Open. As long as it makes sense.

1.问的是上面的metric - average number of post per daily active user 突然从3下降到2.5，有哪些可能的原因，并且解释每个原因。

(1) ask for clarifications: problem with data collection? one-time event or progressively? seasonality? platform? region? special events?
(2) about the metric: numerator — any change in total # of posts from all DAU per day? privacy concern? a similar feature/product?
denominator — any change in # of DAU? new users are not less willing to reveal themselves online?

好像还问了个问题，是怎么样确定一个新的change是好是坏之类的，有哪些metric可以帮助measure。
What is the goal of this change? profit or engagement?
The goal of Facebook is to increase user engagement and retention, a new feature which helps to achieve this goal is good. To measure engagement/retention, use metrics like # of active users daily/monthly, fraction of users who used this new feature. avg # of posts per DAU, time spent

我问了是worldwide么，然后check了有没有突发情况或者seasonality，然后说大家either total number of posts下降了or dau增多了但是增加的dau发帖子不活跃，
follow up问怎么确定dau活跃程度，我说bucket by time on fb
time spent, avg # of posts/comments/likes

2. fb app现在想改版成ins app那种在页面最下方有一个发帖按钮的interface，问怎么设计a/b

testing

## 2.Link Link2

SQL — message
Table: date, timestamp, send_id, receive_id
Question: What's the fraction of users communicating to > 5 users in a day?

```
SELECT
        date,
        ROUND(SUM(CASE WHEN num_contacts >= 5 THEN 1 ELSE 0)/COUNT(userid), 2) AS
fraction
FROM
        (SELECT date, userid, COUNT(DISTINCT userid2) AS num_contacts
        FROM
        (SELECT date, send_id AS userid, receive_id AS userid2
         FROM message
        UNION ALL
        SELECT date, receive_id, send_id
        FROM message
        ) T1
        GROUP BY 1, 2
        ) T2
GROUP BY 1;
```

Product — Best friend?
# of interactions (likes, comments, post on timeline), # of photo tags, gifts, messenger data, demographical data

## 3.Link Link2

SQL — sms_message (fb to users)

| date | country | cell_numer | carrier | type |
|---|---|---|---|---|
| 2018-12-06 | US | xxxxxxxxxx | verizon | confirmation (ask user to confirm) |
| 2018-12-05 | UK | xxxxxxxxxx | t-mobile | notification |

confirmation (users confirmed their phone number)

| date | cell_number |
|---|---|

(User can only confirm during the same day FB sent the confirmation message)

1. yesterday how many confirmation texts by country.

```
SELECT
        country, COUNT(*)
FROM
        sms_message
WHERE
        type = 'confirmation' AND DATEDIFF(CURDATE(), date) = 1
GROUP BY 1
ORDER BY 2 DESC;
```

2. Number of users who received notification every single day during the last 7 days.

```
SELECT
        date, COUNT(DISTINCT cell_number) AS num_users
FROM
        sms_message
WHERE
        DATEDIFF(CURDATE(), date) <= 7  AND type = 'notification'
GROUP BY 1;


SELECT
        COUNT(*)
FROM
        sms_message
WHERE type DATEDIFF(CURDATE(), date) <= 7  AND type = 'notification'
GROUP BY cell_number
HAVING COUNT(DISTINCT date) = 7


3. 过去30天的 confirmation rate
SELECT
        T1.date, ROUND(IFNULL(num_confirmations/num_send, 0), 2) AS confirmation_rate
FROM
        (SELECT date, COUNT(*) AS num_confimations
         FROM confirmation
         WHERE DATEDIFF(CURDATE(), date) <= 30
         GROUP BY 1;
        ) T1
        JOIN
        (SELECT date, SUM(CASE WHEN type = 'confirmation' THEN 1 ELSE 0 END) AS
num_send
         FROM sms_message
        WHERE DATEDIFF(CURDATE(), date) <= 30
         GROUP BY 1
        ) T2
        ON T1.date = T2.date
GROUP BY 1;
```

如果有一个简化版的FB，只有简单的文字post和comment且comment 没有nested这个功能（只能一条一条）。 问如何判断一个post的comment包含conversation。有什么metrics可以监测。comments的话可以看有没有几个user back and forth的pattern

然后问了几道probability，FB ads 有lazy reviewer 和common reviewer, 一个reviewer是common的概率是0.8，是lazy概率是0.2。common给好评概率是0.6,差评0.4。lazy全给好评。问1.）一条ads 是好评的概率，2.) 100个ads里number of 好评的expectation。3.）有五个ads都是好评，是lazy的概率
(1)P(good) = 0.8 * 0.6 + 0.2 = 0.68  (2) E = 100 * 0.68 = 68 (3) P(lazy| 5 good) = 0.2/(0.6^5 * 0.8 + 0.2) = 0.76

**probability**:
two approaches:
a. 5% chance to be an ad per post.
b. every 20 post must have an ad in it.
1. compute each expected value and variance for number of ads in 100 posts.
expected = 5 for both, var_a = 100 * 0.05 * 0.95 = 4.75, var_b = 0

2. probability of getting more than 10 ads in 100 posts with approach a.

我是这么想的:
approach a应该符合二项分布，p=0.05, q = 1-p = 0.95.
如果用二项分布解的话:
p(more than 10 ads) = 1 - p(less than or equal 10 ads)

p(less than or equal 10 ads) = p(ads = 0) + p(ads=1) + ... + p(ads=10)
等式右边的每一项可以用二项分布的概率密度函数解。。

但是这么解会比较耗时间。可以考虑用正态分布去估计（因为我们看到样本数目比较大np>=5且nq>=5）:
此处 mu = 5, var = 100*0.05*0.95 = 4.75, sigma = sqrt(4.75) = 2.18
**Z = (x - mu)/sigma = (10 - 5)/2.18 = 2.29**
如果没有Z表，可以这么估计:(我不太确定这么估计是否会让面试官满意，但是应该比没有估计好吧。。)
我们很熟悉的单尾Z<1.96的概率是0.975, 所以p(Z>1.96) = 0.025
所以p(Z>2.29) < p(Z>1.96) = 0.025
解答:
100个posts中有超过10个广告的概率不超过2.5%，具体数字根据查表得到为1.1%。
根据二项分布的解法，用Excel算了一下，结果是:1.1472%

3. expected number of seeing back-to-back ads in 100 posts with two approaches.

**product**:
why facebook require users to register accounts with phone number or email address confirmation? What pros and cons each have?
(1) recover password/account
(2) send notifications
(3) import contacts and find friends to improve engagement
(4) detect duplicated accounts
(5) privacy
(6) phone message may not be free, email requires internet connection

**product**: Link
instagram now have feature let users to switch accounts with one button
1. how to identify multiple accounts belonging to the same user?
(1) phone number/email address (2) user_name (3) following/follower intersection (4)

device id

2. total timespent flats and number of accounts increases. What are you hypothesis about why this happened? what data you need and how to testify your hypothesis? how do you determine if this feature is successful launch?
我的回答是可能有novelty effect，一开始推出这个feature大家觉得有趣就去多建了几个账号但是the way ppl interact with Instagram并没有变化所以avg time spent没有变。是否要 launch就要结合opportunity size以及你想的一些metrics来决定是否有practical impact，impact有多大来决定。

## 4.Link

SQL — ad4ad：
两个表
ad4ad: date, user_id, event(impression, click, create_ad), unit_id, cost, spend，记得还有一列，可能是ad_id，event是create_ad对应的行才有数值
users: user_id, country, age

ad4ad的背景，简单来说unit就是一个他未来会买到的广告的template，一个user可以看同一个unit很多次，也可以看到不同的unit，如果user create an ad了的话就不会再看到了。我补充一下这个ad4ad，fb想让一些通过fb宣传自己产品的隐性广告购买者也成为paid users。方法就是给他们提供一些广告的template，这些users就会在自己的newsfeed不断看到这些为他们设计的template，每次看到都算作一次impression，如果进一步点击了就算一次click，最后购买了的话就会create an ad。这样算一次成功的转化。

1. last 30 days, by country, total spend (问的是facebook的spend就是表里的cost） of the product

```
SELECT
        country, SUM(IFNULL(spend, 0)) AS total_spend
FROM
        users U
        LEFT JOIN
        ad4ad A
        ON A.user_id = U.user_id
WHERE DATEDIFF(CURDATE(), date) <= 30
GROUP BY country;
```

2. how many impressions before users create an ad given an unit?

```
SELECT
        t1.user_id, t1.unit_id, SUM(CASE WHEN event = 'impression' THEN 1 ELSE 0 END) AS num_impression
FROM
        (SELECT DISTINCT user_id, unit_id
         FROM ad4ad
         WHERE event = 'create_ad'
        ) t1
        LEFT JOIN
        ad4ad
```

```
        ON t1.user_id = ad4ad.userid AND t1.unit_id = ad4ad.unit_id
GROUP BY 1, 2;
```

3.avg number of impressions per user per item before user creates ad

```
SELECT
        unit_id, SUM(num_impression)/COUNT(DISTINCT user_id)
FROM
        (SELECT t1.user_id, t1.unit_id, SUM(CASE WHEN event = 'impression' THEN 1 ELSE 0
END) AS num_impression
         FROM
                (SELECT user_id, unit_id
                FROM ad4ad
                 WHERE event = 'create_ad'
                ) t1
                LEFT JOIN
                ad4ad
                ON t1.user_id = ad4ad.userid AND t1.unit_id = ad4ad.unit_id
        GROUP BY 1, 2;
        ) T2
GROUP BY 1;
```

product1：list few metrics related to ad4ad, why these metrics. If one metric goes down, what is the reason?
再说一下关于product的问题，如果一个metric go down了，大家不仅仅要break down the metric from user perspective(e.g. country, device, etc), 同时要想到这个metric是怎么算的，denominator和numerator是什么
metric下降可能是numerator变小了，可能是什么原因导致；然后denominator增加有可能是什么原因导致。还有如果别的创造广告的途径的收入下降了，对ad4ad好不好呢，可能会有什么影响呢等等，总之问的很细

avg就说不够robust to outliers，可以考虑用median,
metric 还可以提到revenue per targeted user = total ads revenue generated by ad4ad / total target ad4ad users

Product2:
1. metrics to measure the health:
        基于这个产品的feature，转换率还有profit是比较重要的。# or % of  users who used this feature per month; # of ads posted per month; overall, profit through this feature per month?
2. Which one is the most important to show to CEO? Profit
3. 你的角度都不错，也给了我一些启发。
我建议从profit的计算方法开始展开，也即从revenue和cost的两个角度出发。revenue低，可能是定价低，可以提高价格。但与此同时，或许购买人数会变小。这里我们需要做一个optimization。如果是cost的问题，有可能是购买率低，每次成功的转换都需要太多impression，说明我们推荐的不够好，不是target users。
4. 从Facebook角度来说推出这个feature有什么好坏?

广告太多也会影响其他用户的使用体验等等。

Link2
Product:
你会用什么metrics来衡量这个feature？如果有个metrics下降了10%， 你会怎么做，你可以看到fb任何数据，你会怎么来做（大概意思就是你怎么segment去看问题出现在哪里）

因为我觉得我product回答的都还可以的， 但是被通知要加试一轮product analysis，所以做出了一下总结，希望后面的人不要犯同样的错误！
附上总结，1.在回答product的问题的时候，不要emmmm。。。面试官会以为你被难住了，你不会而且没自信，其实我这只是我的个人习惯我只是想表示我在思考，其实完全可以说： please kindly allow a moment fororganizing my thoughts.
2. 在说metrics的时候不要想到哪个metrics就甩过去，因为你可能会被follow up很深。这次就是被问了很深，幸好自己比较了解我自己选择的几个metrics。Metrics哪里好哪里不好？要有准备会问到。还会被问到哪个metrics更好。如果当时脑子卡了一下，可以问一下这个公司这个产品当前的目标，对你进一步选择metrics 有帮助，
3. metrics下降了这种问题，不要只想着denominator，numerator， 也要顺着产品去想想问题出现在哪里
4. 最后的忠告是千万别犹豫！！！如果说 当你想选择table context里面的segment，千万别犹豫，说出来，是对的，没必要一定要自己想出来个什么来，脸皮不要薄。想到啥说啥，千万别怕错，你想说的很可能是对的！！

## 5.Link Link2 Link3 Link4 Link5

SQL — SPAM
-- Q1: how many posts were reported yesterday for each report Reason?
-- Table: user_actions
-- ds(date, String) l user_id l post_id l action ('view','like','reaction','comment','report','reshare') l extra (extra reason for the action, e.g. 'love','spam','nudity')

```
SELECT
        extra AS reason,
        COUNT(IDSTINCT postid)
FROM
        user_actions
WHERE action = ‘report’ AND DATEDIFF(CURDATE(), date) = 1
GROUP BY 1;
```

-- Q2: introduce a new table: reviewer_removals, please calculate what percent of daily content that users view on FB is actually spam?
--no need to consider if the removal happen at the same post date or not.
-- ds(date, String) l reviewer_id lpost_id

```
SELECT
        U.date,
        U.user_id,
        COUNT(reviewer_id)/COUNT(DISTINCT U.post_id)AS percentage
FROM
```

```
        user_actions U
        LEFT JOIN
        reviewer_removlas R
        ON U.post_id = R.post_id
GROUP BY 1, 2;
```

Q3: How to find the user who abuses this spam system?
第三问我是output了一个table：第一列是user id，第二列是这个user report了多少个post是spam，第三列是这个user report spam的post中到底是有多少是真的spam，所以就是left join了一下。面试官也说ok能够提供想要的结果。

```
SELECT
        user_id,
        COUNT(reviewer_id)/COUNT(DISTINCT U.post_id) AS fraction_spam
FROM
        user_actions U
        LEFT JOIN
        reviewer_removals R
        ON U.post_id = R.post_id
WHERE
        U.action = 'report'
        AND DATEDIFF(CURDATE(), date) <= 30
GROUP BY 1;
```

Product:

- How would you test if this filter works?
  - num  of spam reported every day
  - num of shares/comments/likes
  - time spent
  - # of active user per day, week, month
- If we experiment, how would you conduct it?
  - A/B testing
- How to select a sample group
  - Random to avoid bias
- How many people would you select for your sample group
  - Use formula for n (minimum sample size)
- After getting results from A/B testing, what to do next?
  - T-test on metrics to see if there's a difference
- What's a t-test? What's t-score? What's P-value? Explain p-value to someone who doesn't know stats.
- Let's say the filter worked but revenue went down, what would be your hypothesis?
  - Revenue comes from click of ads, # users * CTR * price/click, # users and price/click are the same, then CTR gets smaller, people spend more on good contents
  - Short term vs long term, in the long run it's good for the product
- Given revenue decrease, how would you make recommendations? (doesn't have to be yes or no answer)
  - Short term vs. long term: how much does revenue drops? User experience vs. revenue. Short term revenue drop vs. long term brand perception and long term revenue gain.

PRODUCT:
Q3: Facebook has decided to be proactive about SPAM, instead of merely reactive. We decide to address the SPAM problem through a Machine Learning solution predicting whether a given post is Indeed SPAM. We want to use the predictions in order to downrank/deprioritize suspected SPAM from news feed.

Q1. Facebook用machine learning 建了一个model来rank content以达到filter spam的目的，需要关注什么metrics来评价这个model.

Q2. 在用ab testing的时候发现用了新的spam model之后revenue下降了。面试官确定了首先这个model不会touch到ads，就是说ads不会被filter out。并且DAU/WAU/MAU和time spend没有变化，也就是说user方面没有变化。那么可能的原因是什么。
然后我问面试官revenue主要来自什么，面试官说是click ads。我说那么ads click的revenue主要可以break down成#user x CTR/CTP x price/click. 这个情况下只可能有变化的是CTR，也就是说因为用来新的model以后，这个平台的整体content质量更高了，那么user就更喜欢花更多时间去explore这些content，那么点击广告的时间就相对来说哦变少了，revenue也下降了。面试官说是这样的，采用新的model之后用户可能会花更多时间去看video之类的，那么用在ads上的时间就变少了。

#DAU x CTR/CTP x price/click x # ads per user

题目背景介绍很长，大意就是以前spam都是人工看，费时费钱，现在整出来一个ML算法来看。然后筛掉这些。结果发现其实筛掉的不一定对。所以改成了把挑出来的downrank到最下面去。问如何evaluate这个。

楼主一开始有点懵，不知道到底要evaluate算法，还是evaluate downrank这个事情，是和最开始的人工筛选比较还是和之前直接全筛掉比较。

考到的考点有:
1) 哪些metrics? --我选了DAU和avg time spent 衡量engagement。最后的最后烙印告诉我其实他想要的是衡量算法，其实最直观的应该是spam reported before vs after。
2) 样本是random挑选么？ -- 我回答应该从至少report过一次的人里挑，因为一般人可能不举报，说明不敏感。
3) ML这个模怎么建？吓了一跳考起来ML了，幸亏不是很细致，BS了一段
4) spam降了发现engagement也下来了怎么回事？--我依稀记得地里说过distribution，但是当时没细想，临时一问想不起来distribution如何make sense，憋了半天说了一个可能spam的东西有人喜欢看，这些人很生气。烙印又问还可能是什么呢？这下憋不出来了，冷场更多。。。最后烙印放弃换下一题。
5）一个升一个降如何衡量效果。--我说weighted avg来看

Link
**产品部分，**Facebook 做了一个算法来down rank SPAM post。

**1.你会用什么 metrics evaluate the performance of the algorithm?**
这里我分了product和user两个维度来答，首先product方面，可以用spam rate；用户方面，可以用DAU/MAU，以及engagement of posts( likes, comments, shares)。

最后站在公司角度长期来看会影响Revenue，（这里是受面经的影响，直接把后面第三题会问的revenue说了出来....其实这两者看上去是没有直接关系的）面试官这里follow up，让我解释为什么spam filter会影响公司Revenue，因，我有一点点慌，但是还好临时圆了场，说spam 太多会降低用户的使用体验，长期来看也许会流失用户，从而影响公司的revenue。面试官表示make sense。
这里只要能自圆其说应该就问题不大。

**2. how do you know this happens is due to the use of the new algorithm? how do you form your control/experiment group?**
讲一下A/B test的基本流程就好
分组最重要的一点就是Randomization

**3. ads revenue下降了，原因是什么?**
**首先确认没有seasonality / geography 的influence**
然后确定了这个filter不会touch到ads，就是说ads不会被filter out。并且DAU/WAU/MAU和time spend没有变化，也就是说user方面没有变化。
之后 revenue breakdown = #user * price/view * average view （注意这里不是CTR)

if price too low： 面试官说这是ads组的事情，不需要我们来考虑

那就只能是average view下降，面试官让我继续解释，为什么spam filter 会导 ads view下降。这时候我由于受之前面经影响很深，思维固化了，直接说出因为spam filter，这个平台的整体content质量更高了，那么user就更喜欢花更多时间去explore这些content（比如video, photos, posts from friends)，那么看广告的时间就少了。

面试官表示ads view下降 和 time spend无关，这时候我又再次有点慌...后来沉默了一会儿，面试官掏出手机打开insta，给我演示了用户是如何看到广告的，突然我就开窍了，是因为ads是穿插在post里的，如果spam 变少了，user更容易看到他们想看的content，就会减少screen scoll length，相当于看到广告的可能性降低。面试官说make sense。

**4. 这套算法上线后，Spam Rate下降，ads revenue也下降，我们如何决策是否继续这套方案?**
这里由于算法已经上线，所以是一道和A/B test无关的题。

最后整个面试过程结束只花了30分钟，面试官就笑了，说还有十五分钟时间我们要怎么办呢==，我就只能闲扯问了几个fb相关的问题拖到了40分钟。
面试官说fb里一切都是以product为先，无论是engineer, desiner, 还是data scientist都是以产品为单位来分成组，所以每个岗位都需要懂产品，这一点我非常赞同。

## 6.Link
SQL — given table with: (user, group, time, displays, clicks) for a payment page.
1） # of clicks and displays in given day

2) click through rate,
3) click through rate for each group,

4) group 1 click rate: 10%, group 2: 15%, think about possible differences, group 3 click rate 150%, think about possible reason,
5) how to identify click but not displays

Stat
1） 1000 people, each time select 10 (w/o replacement) 问每个人on average多少次会被抽到，2) 1000 people, each time select 10, (w/ replacement), 问每个人on average多少次会被第一次抽到，3）画distribution of page shared (#users vs # page share). 标mean, median, p1, p99, 问根据经验mean是多少，median是多少，取出day 1所有page share=2的population, 问trends for day 2-30, 同样取出day 1 所有page share=5的population, 问同样的trend. 那个方差大，是什么分布?
Explanation: Link



product
问父母加入脸书对子女的影响，1）如何判断父母是否加入，2）直觉上有和影响？3）如何分析数据，4） 如何test
This way works to show correlation between two, but cannot show causal relationship. So you will need to further test this. Couple different ways, 1) design some test to randomize A/B. For example, randomly select from those their parents send them invitation. Just don't show them. (of cause create some customer experience problem), 2) design some treatment. For example, if you think the impact is positive, you send a notification to group A asking them to invite their parents. B group is control. If you think the impact is negative, for those parents already connect with them, Group A introduce a "enhanced privacy setting" to opt them out of their parents post, group B for control. cons: the impact is treatment+behavior

## 7.Link Link2

SQL — table: ad_account,date, spend, status (open,close,fraud)
1.求在active accout里是fraud的概率（active means spend>0）

```
SELECT
    SUM(CASE WHEN status = ‘fraud’ THEN 1 ELSE 0 END)/SUM(CASE WHEN spend > 0 THEN 1 ELSE 0 END) AS prob_fraud
FROM
    table;
```

2.求有多少account是今天被label成fraud

```
SELECT
    COUNT(DISTINCT ad_account)
FROM
    table
WHERE
    date = CURDATE() AND status = ‘fraud’;
```

3.如果给被report为fraud的账户申诉的机会，有什么financial benefit
(1) second chance, those who post the ads will like it, user satisfaction
(2) some user may abuse the report mechanism, they report every ad they see. This will definitely stop some potential advertisers

Link2
如果说要准备这类面试的话，我建议你把FB所有产品都点一遍，每一个button每一个feature都在自己脑子里面过一遍，考虑这样几个问题："这个feature到底有什么用？本质是什么?为什么FB非要加这个feature？它到底有什么独特之处？假设它没有launch，我在launch之前要考虑哪些因素，怎么去衡量成功？如何去做test？这样做是否有局限？怎么改进？"我自己就是这样把FB所有产品走过一遍的。
如果有一起准备的小伙伴的话，可以互相challenge。
面试时候的话，尽可能把这个当作一个交流，优先梳理框架，在白板上整理出来，而不是一上来直接和贯口一样把metrics都倒出来。

Group: Active Group%, Time Spent on Group/Total Time Spent on Facebook, New Member Growth Rate
Best Friend: About Similarity, %Post/Photo Tagged Together, Messanger records
NewsFeed: Favor towards certain post type? Load Speed? Text font? Internet Quality and Coverage? Device Type?

## 8.Link Link2 Link3
SQL — comment distribution
有content_id, content_type (comment/ post), target_id。如果是comment，target_id就是post的id，如果是post则target_id为NULL。求comment distribution。

```
SELECT
        num_comments, COUNT(post_id) AS num_posts
FROM
        (SELECT target_id AS post_id, COUNT(*) AS num_comments
         FROM table
         WHERE content_type = 'comment'
         GROUP BY 1
        ) T1
GROUP BY 1;
```

然后加问：如果现在content_type变成post, video, photo, article，要求计算每一个content type的comment distribution。

```
SELECT
        T1.content_type, T2.num_comments, COUNT(T2.post_id) AS num_posts
FROM
        (SELECT content_id, content_type
         FROM table
         WHERE content_type != 'comment'
        ) T1
        LEFT JOIN
```

```
        (SELECT target_id AS post_id, COUNT(*) AS num_comments
         FROM table
         WHERE content_type = 'comment'
         GROUP BY 1
        ) T2
        ON T1.content_id = T2.post_id
GROUP BY 1, 2;
```

SQL，考的是post type distribution那道题:You have one table named content_action which has 5 fields:Date,
User_id (content_creator_id),
Content_id (this is the primary key),
Content_type (with 4 types: status_update, photo, video, comment),
Target_id (it's the original content_id associated with the comment, if the content type is not comment, this will be null)

Question:
1.find the distribution of stories（photo+video） based on comment count?
2.what if content_type becomes (comment/ post), what is the distribution of comments?
3.Now what if content_type becomes {comment, post, video, photo, article}，what is the comment distribution for each content type?

Distribution
想跟大家详细讨论一下科技公司里面特别喜欢问的distribution问题。出题形式一般是，问一个特定的random variable符合什么样的distribution。比较常见的几种题型如下：
1. What's the distribution of the comments per post?
2. What's the distribution of comment length?
3. What's the distirubtion of page shared per person?
在onsite的时候，还有公司喜欢追问，if we take all people that share 2 page in day1, what can be their distribution of page share in day2?

这种问题我翻找了以下，对于1,2,3，常见答案一般喜欢说log normal 或者poisson distribution。原因是因为认为 X=0的概率最大，且容易出现长尾效应，我觉得这种答案是有道理的，但是可能在面试中不是特别足够，所以专开一贴进行讨论。

Firstly,我觉得poisson distribution不是特别合适。传统意义上，poisson distribution一般用来衡量固定时间内，event出现k的概率，event一般为binary（1l0）。本质上poisson distribution 是一种bionomial的特殊形式，在n趋近于无穷，但成功次数的expection lambda相对固定的时候使用。相对应的推导可以看这里，https://medium.com/@andrew.chamberlain/deriving-the-poisson-distribution-from-the-binomial-distribution-840cc1668239 。在这种情况下，其实情况1是比较合适的，我们可以将每一个post可能有的comment为n，when a user leave a comment, the event = 1 other wise 0. 潜在中的话，n的确趋近于无穷。但是这种解释有一个致命的问题，就是每一个post可以产生的post的expectation一定不相同，大v转发的动辄过万，小透明的肯定没人理睬，所以poisson 分布里每次抽样的，得到的expectation都是lambda这个假设被严重违背了。个人觉得，如果是求daily active user/daily likes之类的distribution会更加合适。

排除掉了poisson我觉得log normal其实是一种比较合适的估计。这里我看了两篇论文：https://

www.sciencedirect.com/science/article/pii/S1877050914005006 和 https://epjdatascience.springeropen.com/articles/10.1140/epjds14，一个是关于distibution of the reweets per tweet， 一个是distribution of comment length. log normal 的论证基础是，有一个变量X服从normal distribution，而变量Y = exp（X)，因而Y服从log normal（即log（Y)服从normal distribution）。在tweet的论文中，作者提到了power of law，大概意思是如果一篇文章本来很受欢迎，那么更多人会转发它,造成exponential的效果，如果我们认为大家会转发文章的这件事情本身-r服从normal distribution, 由于叠加效应,最终这篇文章的转发量应该是(1+r)^k, r如果大于0，将呈现几何增长，（1+r）^k在数学上，可以用exp(rk)表示，所以最后的转发数应该服从log normal。我觉得这种说法还是比较有道理。对于distribution of commen length,第二篇论文提出了一种看法，即人们comment发长的意愿是服从normal distribution的，但是由于心理学上的一种效应，人体的感官对comment的长度感应不敏感，当人们的感受成proportional increase 时，实际长度成exponential增长，举个实际例子：大部分人觉得comment打一个字，和打10个字，感觉没有什么不同，都很短，但是实际上comment length却增长了10倍。Research还做了两个实验，想要证明这个观点。如果上述关系成立那么大家对长度的感官服从normal distirbution , L(length of comments)确实就应该服从log normal。

这两个说法我觉得都有一定的根据，欢迎大家讨论与补充，面试中也可以了解更多，发挥更好。

最后对于if we take all people that share 2 page in day1, what can be their distribution of page share in day2?我认为应该是符合normal distribution的，在day1所有share 两个page的population中，在接下来的时间内，the proportion of share less should be symmetric to sharing more.我暂时没有想出比较solid的stats理由，希望大家能对此提出自己的意见和看法。

## 9.Link Link2 Link3 Link4 Link5

SQL — marketplace user log
Session table
Date I sessionid I userid I action (enter/click/send/exit)
Time table （sessionid都是unique的）
Date I Sessionid I time_spent (s)

Q1: Average sessions/user per day within the last 30 days

```
SELECT
    date, COUNT(DISTINCT sessionid)/COUNT(DISTINCT userid) AS avg_num
FROM
    session
WHERE DATEDIFF(date, CURDATE()) <= 30
GROUP BY date;
```

Q2: Time distribution of each user, 先问我大概会是怎么样的一个分部 — exponential distribution ? Not clear

```
SELECT
    total_time, COUNT(DISTINCT userid)
FROM
    (SELECT
        userid, SUM(IFNULL(time_spent, 0)) AS total_time
    FROM
```

```
            session S
            JOIN
            time T
            ON S.sessionid = T.sessionid AND S.date = T.date
    GROUP BY 1) temp
GROUP BY 1
ORDER BY 1;
```

Product — Marketplace
Q1: launch a call-to-action button “Sell Your Product” on the top banner, what's the reason behind this launch?
我说答案可能有两个方向，
第一个是我们的确发现了用户有这个需求，我们通过观察他们的clickstream和timespent是有可能得出他们需要一个button去引导他们完成selling post，所以我们设计了这个button
第二个我说这算是e-commerce行业的一个常识，用户不会去做他看不到的事情，所以我们需要引导用户，这样有利于我们提高CTR和转化率

Q2: How do you evaluate the impact and make sure this button is actually working?
这里个人认为很明显期待你答A/B testing
然后他问了我key metrics怎么设计，control 和test group都是什么；这里注意key metric不能是CTR，因为control group的设计里根本没有这个button，所以CTR会有bias，
我选择的是conversion rate，# of posts of selling product/total session in the test period, 然后他问我为什么分母是这个，我说因为那个时间段里，根据第一题SQL，用户一个session里会做很多事，其中之一就是post，所以这么算我觉得比较合理，他说好的，make sense。

Q3: 现在ran了A/B testing，发现CR反而降低了，问我为什么
我先确定了有没有external factor，seasonality，data collection error； 他说没有
然后我就回到公式本身，我说那就可能是两种情况，# of post 降低了，或者session数量增加了，session增加的原因可能是最近我们做了什么promotion，或者临近毕业季，大家都想卖东西

Q4: 他在这打断我，说ok，就是# of post降低了，问可能的原因
我说那应该是用户的行为有所变化，但不一定是坏事，我们需要先看一下所有用户的time spend有没有增加，我们的 # of transaction有没有增加，买家端的用户变化，以及我们所处的时间段，因为毕业季可能卖家多，开学季买家多之类的；
其次，那就是比较糟糕的情况，就是我们确实流失了一些user engagement，可能他们确实churn to ebay etc。 他说确实是，可能是买家比较多，用户在实验之前po的东西都被买了，现在这个时间段并不是高峰

Q5: 现在我们想做一个recommendation algorithm （就是类似于淘宝的猜你喜欢），你会怎么去设计这个功能
1. item based
我们可以根据每一个用户的购买记录去给他推类似的商品，但是这个方向有一个弊端，就是marketplace不像传统电商平台，这里是个C2C的，所以可能他的购买记录不是很丰富，其次，用户可能不想一直购买类似的商品，可能会引起厌烦
2. user based
这个方向就可以很好的弥补第一个direction的弊端，因为每一个user都是facebook user，我们可以利用他们的network，去找到跟他们类似的user，甚至跟他们在一个group里的人，去给他们推他的朋友买过的东西，这里的similarity是可以用algorithm模拟出来的。

然后他的反馈比较positive，他说是的，这个make sense

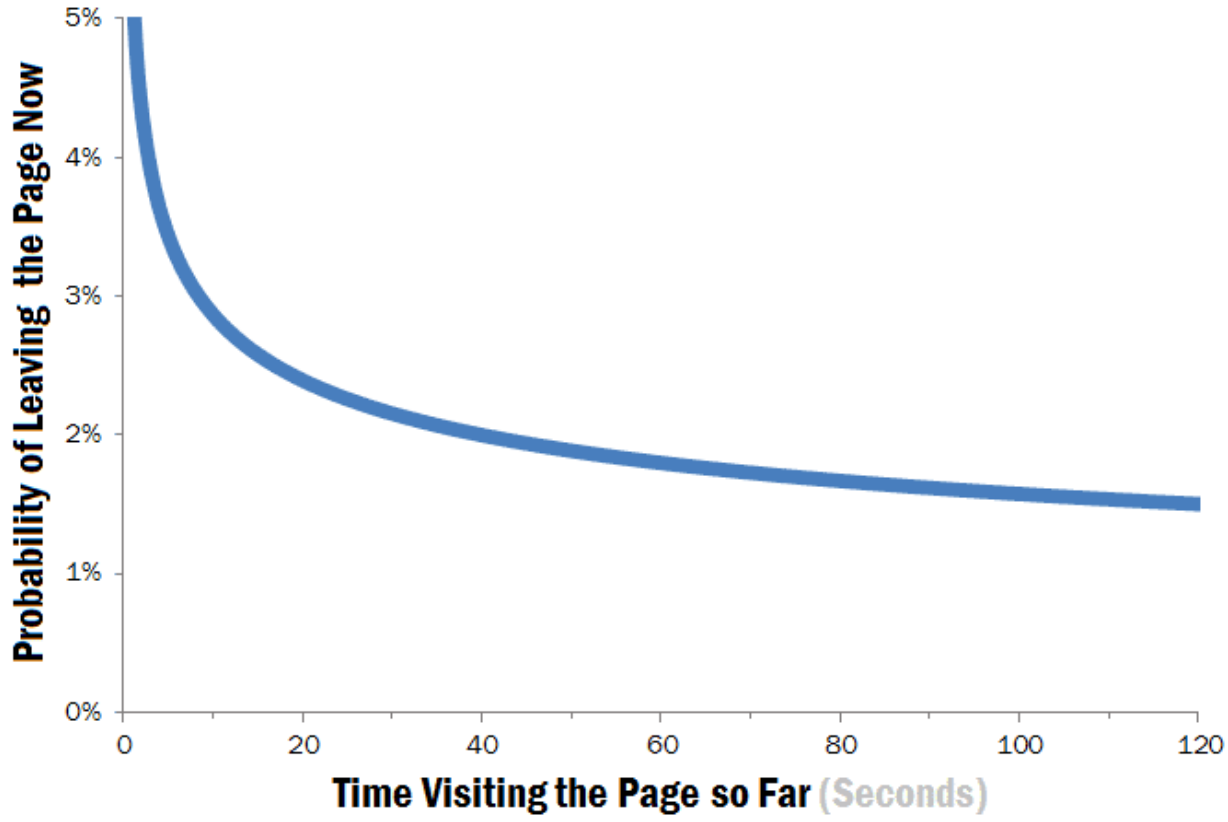
**Tips:**
因为我个人很喜欢facebook，所以准备了很多细节上的小问题
1. 不要小瞧面试开始前的chitchat，那也是你可以抓住机会differentiate自己的地方，一定要比较契合公司文化，建议大家去看看这个链接里的几个重要的value，试着结合自己的经验说说
https://www.facebook.com/careers/

2. SQL的格式要整齐，显得你比较严谨

3. 个人认为这个岗位还是更偏向于product，建议多去读读博客，看看关于key metrics的insights，当然地里的面经也很重要

4. 可以适当的对interviewer下下功夫，我在领英上发现我的interviwer之前创过业，所以我在自我介绍的时候着重说了下自己的创业经历。

Link
Product — Q5 Estimate the distribution of time spent on market place. Y-axis is # of people, x axis is time spent on marketplace. (mean, median, mode)
应该是exponential distribution，大部分人只是浏览了5sec到10sec甚至更短，如何让浏览时间变长是产品的努力方向。




It's clear from the chart that the first 10 seconds of the page visit are critical for users' decision to stay or leave. The probability of leaving is very high during these first few seconds because

users are extremely skeptical, having suffered countless poorly designed web pages in the past. People know that most web pages are useless, and they behave accordingly to avoid wasting more time than absolutely necessary on bad pages.
If the web page survives this first — extremely harsh — 10-second judgment, users will look around a bit. However, they're still highly likely to leave during the subsequent 20 seconds of their visit. Only after people have stayed on a page for about 30 seconds does the curve become relatively flat. People continue to leave every second, but at a much slower rate than during the first 30 seconds.
So, if you can convince users to stay on your page for half a minute, there's a fair chance that they'll stay much longer — often 2 minutes or more, which is an eternity on the web.
So, roughly speaking, there are two cases here:

- bad pages, which get the chop in a few seconds; and
- good pages, which might be allocated a few minutes.

Note: "good" vs. "bad" is a decision that each individual user makes within those first few seconds of arriving. The design implications are clear:

- To gain several minutes of user attention, you must clearly communicate your value proposition within 10 seconds.

Q6. What is the benefit of launching Call to action (e.g., learn more, buy)? What metrics used for measure the effect? Want to sell?
Goal is to encourage people to sell products
Metrics:
# of users who sell their products
# Conversion rate (# people clicked yes/# people viewed)
# Monthly/daily active users
# Engagement/Time spent on browsing marketplace

Q7. The message_sent/session drops 10% month by month. What is the reason?
Replied Seasonality effect (look at historical data), time span(two months data vs. six months data), maybe buyer will not use facebook messenger to leave a note to buyers, competitors, does they show similar trend?
In the end, the interviewer mentioned it is mainly because area effects, some regions drops, while some regions increases. Need further investigation on this problem.

Link
产品：FB开发一个新产品，Pet page, 比如说Ins上有很多人专门给自己的宠物建一个账号，如果在FB上launch这个功能，让人们给自己的宠物建pet page， 请问怎么measure这个产品
产品我就回答了[td]adoption  perspective:  adoption rate, growth rate such as WoW and the trend plot
DAU， MAU，etc

## 10.Link Link2

SQL 是说friend request的2 tables
一个request table - (sender_id, send_to_id, time)
一个accept table - (requester_id, accepter_id, time)

Q1. On date XXX, what's the acceptance rate/percentage?

SELECT

```
        COUNT(A.requester_id)/COUNT(R.sender_id) AS acceptance_rate
FROM
        request R,
        accept A
WHERE
        DATE(R.time) = 'XXX' AND DATE(A.date) = 'XXX';
```

To remove duplicates:

```
SELECT
        COUNT(T2.requester_id)/COUNT(T1.sender_id) AS acceptance_rate
FROM
        (SELECT DISTINCT sender_id, send_to_id
         FROM request
         WHERE DATE(time) = 'XXX') T1,
        (SELECT DISTINCT requester_id, accepter_id
         FROM accept
         WHERE DATE(time) = 'XXX') T2;
```

Q2. 谁的friend 最多

```
SELECT
        user_id, SUM(num_friends) AS num_friends
FROM
        (SELECT requester_id AS user_id, COUNT(DISTINCT accepter_id) AS num_friends
         FROM accept
         GROUP BY 1
         UNION ALL
         SELECT accepter_id, COUNT(DISTINCT requester_id)
         FROM accept
         GROUP BY 1
        ) T1
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1;
```

Product: FB打算有个feature, 给用户发text notification if their close friends update something. 这个feature有两步，一个问要不要sign up, 回答yes or no，说了yes就发。然后问如何决定要不要做这个feature，然后问了些metrics, 建议去看看AAREM这个framework，AAREM 是说一个 funnel process, Acquisition, Activation, Retention, Engagement, Monetization... 好像不同的人有不同的叫法，可以根据这个去查一下。

Link Link2

知道地里很多同学在准备product题目，其中实验设计经常出现，在这里想向大家请教两道题目，谢谢！

第一题：研究如果向用户推送close friends有update的notification，用户行为的变化

http://www.1point3acres.com/bbs/thread-273654-1-1.html

第二题：研究父母加入FB对用户行为的影响
http://www.1point3acres.com/bbs/thread-209706-1-1.html

## 我的问题是这两题应该用a/b testing还是cohort analysis?

a/b testing主要用于测试population里在同一时间一个变量对group A和group B的作用，group A和group B要comparable，要randomize。cohort analysis主要是分析population里面一小撮人，强调的是用户前后行为的对比。

a/b testing的问题：对于第一题，如果对愿意接受推送和不愿意的用户进行a/b testing是有bias的，愿意接收推送的人可能本来就很关心自己的close friends，所以有了推送之后，他们的活动会增加，这两个population就不是random的。对于第二题，如果直接比较有父母和没有父母的两组，很有可能是在有父母的里面，很多用户本来就不care父母，所以最后比较结果不显著。

Cohort的问题：没办法控制时间变量，可能有外在其他因素影响。

我的想法是，可否用两个cohort来进行前后对比。比如第一题，设置两个cohort，cohort A：接受的A，没接受的B。当cohort B在turn on前后没有显著差异的时候（确认没有外在条件影响），再分析cohort A在turn on前后是否有显著差异（但是感觉还是有点儿问题，比如如果外在事件只影响一个cohort呢）

Product — Close friend notification: 打算给用户的发text notification 告诉他们close friends 的Update, How to evaluate if we want to add this feature?
楼主回答说time spend on facebook, 看他没什么反应，然后又问这个notification是可以点击link到facebook上去的吗，他说是，于是我多嘴扯了一个CTR of the notification. 然后他追问你怎么evaluate CTR呢？然后我就马上跪了承认说这个metrics不好，没法draw the line，还是time spend on facebook吧.

Follow up: How to evaluate negative impact of this feature?
楼主问user是不是可以选择关闭这个服务，他说可以，楼主回答说看看percentage of user who close this feature

Follow up: If the engineer team roll out the feature on 1000 users, the time spend on facebook was 23 mins the week before roll out and 25 mins the week after. What would you say about this result?

楼主回答说先得rule out other factor may result the increase, do A/B test, 加个control to see if also see this increase trend.

Follow up: experiment 1000 user, 24 mins before, 26 mins after. Control arm 1000 user, 24 mins before, 24 mins after. What would you say about this result.
楼主脑子短路，在他提示variance是1.5mins，才反应过来这是在问hypothesis testing

不知道对不对啊求指正！

假设检验

uc: mean time spent on FB of the control group
ut: mean time spent on FB of the treatment group

构造了单边检验（这一点不太确定，感觉双边也可以？）

h0: uc = ut
ha: ut > uc

Assume alpha = 95%, 两组variance一样
用t检验
t =(ut-uc)/sqrt(var/1000+var/1000) = 2/sqrt(4.5/1000) = 30 >> 1.65
so reject h0

For the close friend notification question, I think we can break down the answer to three steps.

First, if the feature were successful, would it be good for Facebook? That is, would it move the key metric to the right direction? If we use engagement (time spent on Facebook, number of actions) as the metric, then the answer is yes. Sending text notification will potentially make people spend more time on Facebook and perform more actions (react to friends' posts).

Second, also the key to this question, is that we need to find a proxy for demand for this feature based on what users are doing today. The safest way for a data scientist to drive new features is to look at the data. If you find a demand for that feature (maybe users are going to friend's profile page to check for updates), then you can incentivize this behavior by simplifying the process. Sending text messages is a good way to do this.

Third, once you establish that the feature is good for Facebook's target metric and there's a demand for that today, we can run an A/B test to decide if we want to add it.

## **11.**Link Link2

SQL —
table 1: user1 | user2
123 456
456 123
123 789
789 123

table 2: sender | reciepient | action | date
123 456 create 2019-01-01
456 123 create 2019-01-01
123 789 create 2019-01-01

问每一对friends的interaction是多少？（一个create就是一个interaction）

```
SELECT t1.user1, t1.user2, count(t2.sender) as pair_total_interaction
FROM friend_pair t1
LEFT JOIN interaction t2
ON ((t1.user1 = t2.sender AND t1.user2 = t2.reciepient) OR (t1.user2 = t2.sender AND t1.user1
```

= t2.reciepient) )
group by 1,2

Product —
1. 一般说来一个user有很多friends是很好的，但是一个user有太多的friends也会有问题。你觉得会有什么问题?
(1) too many posts in the newsfeed, user may miss important updates from close friends who they care about most
(2) more spam or scams, higher chance that a user may add a friend which is a fake account sends them spams or scams

2. 如何来判断一个user的close friends
答：可以通过他们在一些特殊日子是不是会送礼物，照片里他们是不是会频繁出现被tag。
Best Friend: About Similarity, %Post/Photo Tagged Together, Messanger records

3. 如果有一个unfriend button， 如何向user推荐可以unfriend的人。
答： 可以给interaction（比如like，comment），互送gift， 照片里出现频率等factor加权重，然后算出一个score，根据score来排序。

Link
1.好友多是好事，但太多有时候也不好为什么
（好友太多，每天的post会太多，有些好友因为不熟并不关心他的动态，看不到想看的）
2. 基于你说的情况，我们称为crowded，那怎么定义这个存在crowded的situation。
（首先定义亲密好友，再比较亲密好友和非亲密好友的发帖比例）
3. 定义完亲密好友后，你打算用什么方法解决问题
推出unfriend的feature或者derank 非亲密好友的帖子
4.怎么定义亲密
共同好友，毕业学校，互动程度（互相点赞，互相发帖等等）然后weighted average出一个分数.
5.如果只有少部分人有互动怎么办
共同好友，text analtyics看post是否参加共同的活动，图像识别是否在同一张照片，是不是在同一个location
6. weighted average 的weight 怎么得到
我说可以用linear regression 或者其他machine learning mode得到
7.你要怎么implement这个model （当时楼主并没有意识到其实做supervised learning我们是没有y的，以为他问我怎么选模型，feauture engineering之类的）
楼主说了很多模型还有feature engineering以及evaluate模型的方法。。。
（被打断）8.用小白板指出没有y
我意识到问题，说那可以跟做survey得到y，或者用unsupervised model 做cluster
9.得到分数以后怎么做这个unfriend
我说根据分数升序排序推荐（分数越高越亲密）

## 12.Link

2 tabales: (1) date, userid, message_sends (2) date, userid, failed_message_sends. Write a query to obtain avg number of successful message-sends for users in 2 groups: (1)who did not have message_send failure on a given day. (2) users who faces message-send failure on that day)

(1)who did not have message_send failure on a given day

```
SELECT
        SUM(IFNULL(message_sends, 0))/COUNT(DISTINCT userid) AS avg_num
FROM
        table1
WHERE
        userid NOT IN (
        SELECT DISTINCT userid
        FROM table2
        )
        AND date = 'XXX';
```

(2) users who faces message-send failure on that day

```
SELECT
        IFNULL(SUM(T1.message_sends - IFNULL(T2.failed_message_sends, 0))/
COUNT(DISTINCT userid), 0) AS avg_num2
FROM
        table1 T1
        JOIN
        table2 T2
        ON T1.userid = T2.userid AND T1.date = T2.date
WHERE
        T1.date = 'XXX'
```

## 13.Link Link2

SQL spotify听歌问题，table1：timeluseridlsongidl table2: userid1luserid2

问(1)今天听的最频繁的歌曲是什么？(2)寻找一个list有userid和friendid：两个朋友有多于两首共同听过的歌曲

(1)

```
SELECT
        song_id, COUNT(user_id) AS num
FROM
        table1
WHERE DATE(time) = CURDATE()
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1;
```

(2)

```
SELECT
        T1.userid1, T1.userid2, COUNT(DISTINCT T2.songid) AS num_song
FROM
        table2 T1
```

```
    JOIN table1 T2 ON T1.userid1 = T2.userid
    JOIN table1 T3 ON T1.userid2 = T3.userid AND T2.songid = T3.songid
GROUP BY 1, 2
HAVING num_song >= 2;
```

Product sense: 如果看到total ads revenue下降，你该怎么办？如何figure out？怎么给Senior management汇报
Ads revenue = # active users * CTR * price/click
Long term VS short term

## 14.Link

SQL部分 — Ad_ROI
Table1 adv_info:  advertiser_idl ad_idl spend: (The Advertiser pay for this ad)
Table2 ad_info:  ad_idl user_idl price: (The user spend through this ad (Assume all prices in this column >0))

Q1: The fraction of advertiser has at least 1 conversion?
Q2. What metrics would you show to advertisers (其实就是在问ROI), 用SQL实现

Q1:

```
SELECT
    COUNT(DISTINCT advertiser_id)/(SELECT COUNT(DISTINCT advertiser_id) FROM adv_info)
FROM
    adv_info
    JOIN
    ad_info
    ON ad_info.ad_id = adv_info.ad_id;
```

Q2: two cases
Case 1: advertiser每个ad的ROI

```
SELECT
    T1.advertiser_id, T1.ad_id, SUM(IFNULL(T2.price, 0))/T1.spend AS ad_ROI
FROM
    adv_info T1
    LEFT JOIN
    ad_info T2
    ON T1.ad_id = T2.ad_id
GROUP BY 1, 2;
```

Case 2: 每个advertiser的平均ROI

```
SELECT
    T1.advertiser_id, ROUND(IFNULL(total_revenue/total_spend, 0), 2) AS adv_ROI
FROM
    (SELECT advertiser_id, SUM(IFNULL(spend, 0)) AS total_spend
    FROM adv_info
    GROUP BY 1) T1
    LEFT JOIN
```

```
(SELECT advertiser_id, SUM(IFNULL(T2.price, 0)) AS total_revenue
FROM
        adv_info
        LEFT JOIN
        ad_info
        ON adv_info.ad_id = ad_info.ad_id
GROUP BY 1) T2
ON T1.advertiser_id = T2.advertiser_id;
```

## 15.Link Link2

SQL — survey_log 列名：user_id, question_id, question_order, event = {imp, answered, skipped}, timestamp,
第一问是找conversion rate最高的question，我在写完answer rate以后有个follow up问题，说如果imp太少的问题怎么办，我答的是加个threshold，舍弃那些出现少于多少次的问题（这里我随便写的10，他没有说什么，不知道对不对）。
第二问是在用户已经回答了某一问题的情况下，如何安排下一问题使conversion rate最高，我这里就按地里讨论的一样说在已经回答了这一问题的用户中，选他们回答的其余问题里回答率最高的一个

求了个每道题的回答率，注意不能只求回答次数，要除以总的看见此题的次数
追问，即使按照回答率对题目进行排序，如果新来的用户已经skip掉了回答率最高和次高的题，如何动态调整题目顺序，获得此用户尽可能多的回答?
我在这题上挣扎了很久，思考了题目内容分类，题目之间的相似度分类，等等等等，最后发现其实是条件概率，要看用户之间的相似度，即已有数据中跳过了这两道题的用户回答率最高的是哪道……

(1) 找conversion rate最高的question

```
SELECT
        question_id,
        SUM(CASE WHEN event = 'answered' THEN 1 ELSE 0 END)/SUM(CASE WHEN event = 'imp' THEN 1 ELSE 0 END) AS answer_rate
FROM
        survey_log
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1;
```

(2)第二题，大概思路为找出在回答了现问题的人中，回答率最高的是那一个问题。因为，问题要知道两个变量，1，what's current question 2 what are questions the user  have answered? 下面subquery s找出了所有回答了current question 的人， subquery u找出了current user回答过的所有问题，因此，我们需要选出所有s 中回答率最高的问题，同时排除u中所有的问题。

```
#set the variables
set current_question = ?;
set current_user = ?;
#find the question_id that has the highest answer rate in the group that had answer current question
select question_id, sum(case when event = "answered" then 1 else 0 end)/sum(case when event = "imp"
```

```
then 1 else 0 end) as answer_rate
from
survey_log as L
join (select distinct user_id from survey_log where question_id = $current_question and event = "answer")
as s
on L.user_id = s.user_id
left join (select distinct question_id from survey_log where user_id = $current_user) as u
on L.question_id = u.question_id
where question_id != $current_question and u.question_id is null
order by answer_rate desc
limit 1;
```

Product

a. 先问how to measure health of facebook,我说可以分两个部分，一是website/sytem本身health，是不是function normally，另一部分是user engagement，比如新用户注册率，现有用户的interactions包括comment/share blablabla，然后DAU/MAU之类的，

b. 然后他直接打断说ok，现在就focus在comment上，如果要求你做一个dashboard specifically about comment，有哪些metrics你可以present。我先莫名想岔到之前面经的sql题了，跟他说可以show distribution of comments/users，然后他表现得不太满意，说ok，其他的呢，我就说可以present average comments per post/user，然后他继续问这个average comment per user是怎么计算的，denominator和numerator是啥，我回答numerator等于总评论数，denominator是distinct users who have left at least 1 comment per day。

c. 然后他说ok，现在把denominator换一换，改成DAU，我们发现这个总comment/DAU的比率跟一年前同一天相比，增加了50%,有哪些可能原因。我回答首先user mix可能有变化，可能今年新增的growth都是愿意发更多comment的用户，然后（随口乱说）content内容的变化吸引更多user leave comment，第三可能有新feature刺激用户留言更多之类的。然后他follow up了一句什么我忘了，好像是如何判断是哪种？我就先说我们可以segmentation by users' language/platform/device type/browser/location看看有没有哪个subgroups特别突出，有的话可以dive deep。