Product Sense

CHAPTER 10

A magikarp, a one-legged man in an ass-kicking contest, and an ejector seat in a helicopter. These three are examples of things more useful than a data scientist with a weak product sense and business acumen. Because data scientists often work cross-functionally with product managers (PMs) and business stakeholders to help create product roadmaps and understand the root cause of various business problems, they are expected to have a strong product and business intuition. It's not just data scientists who can expect product-sense interview questions — these topics are also frequently covered during product analyst, data analyst, and business intelligence analyst interviews.

Between questions on the art of selecting product metrics, troubleshooting A/B test results, and weighing business trade-offs, the scope of product interview questions is massive. But fear not! In this chapter we cover both actionable strategies to approach the four most common types of product interview questions you'll face and long-term tips to develop your overall product and business sense. We also solve 18 real product-sense interview questions from companies like Amazon, Airbnb, and Facebook.

Four Most Common Types of Product Interview Questions

Before we dive into specific product management topics, it's important for you, the reader, to first get a glimpse at the four most common types of product-focused data science interview questions:

- ***Defining a product metric:*** What metrics would you define to measure the success of a new product launch? If a product manager (PM) thought it was a good idea to change an existing feature, what metrics would you analyze to validate their hypothesis?

- ***Diagnosing a metric change:*** How would you investigate the root cause behind a metric going up or down? What if other counter metrics changed at the same time — how would you handle the metric trade-offs?
- ***Brainstorming product features:*** At a high level, should a company launch a particular new product? Why or why not? For an existing product, what feature ideas do you have to improve a certain metric?
- ***Designing A/B tests:*** How would you set up an A/B test to measure the success of a new feature? What are some likely pitfalls you might run into while performing A/B tests, and how would you deal with them?

By keeping these frequently asked types of questions top of mind, we hope you'll better understand how the following high-level advice can be concretely applied to acing product questions.

Big-Picture Advice for Product Sense Interview Questions

Framework for Approaching Product Interview Questions

The tips below work for approaching product questions, as well as for the occasional business question:

- ***Ask clarifying questions:*** Make sure you understand the user flow for a product, who the end users are for the product, who the other stakeholders are that are involved with this problem, and what product and business goals we aim to achieve by solving the problem. Even if you've done your research into the company and product and know many details, frame your knowledge as a question so you don't inadvertently head down the wrong path. For example: "I know Robinhood's mission is to democratize finance for all. It seems this crypto wallet feature is meant to democratize access to crypto currencies, which can also help us better compete with Coinbase. Am I on the right track?"
- ***Establish problem boundaries:*** These are big problems; scope them down. Establish with your interviewer what you're purposely choosing to ignore to solve the problem within the time frame of the interview.
- ***Talk Out Loud:*** You've seen this tip now many times. These interviews are held to see your thought process — and until Elon Musk invents mind reading at Neuralink, you need to voice your thinking!
- ***Be conversational:*** Don't talk *at* the interviewer — talk *to* the interviewer. Engage them in conversation from time to time as a means of checking in. For instance, "I think a good metric for engagement on YouTube is average time spent watching videos, so I'll focus on that. How does that sound?"
- ***Keep goals forefront:*** It's easy to get lost in technical details. Never forget your answer stems from the company's mission and vision, which you hopefully articulated at the start of your conversation!
- ***Bring in outside experience tactfully:*** Because these problems are rooted in the real world, it's okay to flex your past domain experience. Just don't go overboard and come across as arrogant or cargo cult-y by saying, "This is the only way a problem should be solved, because that's how we solved it at Google."

feeling like there's too many tips to keep track of? Ultimately, if you can remember just one thing when solving product problems, it's this: pretend you've already been hired at the company as a data scientist. You're just having a meeting about the problem with another co-worker. When you adopt the mindset that you're already working for the company, behaviors like talking out loud with your "co-worker" or keeping the company mission top of mind should come naturally.

How to Develop Your Product Sense

Because data scientists help Product Managers (PMs) quantitatively understand the business and look for opportunities for product improvement within the data, they play a crucial role during the product roadmap creation process. As such, questions asking you to brainstorm new products and features are very common during product-focused data science interviews. The best way to improve your performance on this type of problem is to improve your general product sense.

However, don't let the term "product sense" faze you; this isn't an innate gift you're born with, but, rather, a skill that can be developed over time. By following the tips in this section to enhance your overall product sensibilities, you won't freeze up like a deer in the headlights in your next interview with Google, when you're asked to brainstorm features to help students better use Google Hangouts. Instead, you'll tackle the problem with the confidence of Sundar Pichai after yet another Alphabet quarterly earnings beat.

The Daily Habit You Need to Build Your Product Sense

An easy way to develop your product sense is through analyzing the products you naturally encounter in your daily life. When using a product, think about:

- Who was this product created for?
- What's the main problem it was designed to solve?
- What are the product's end-user benefits (this is bigger than simply what problem it solves!)?
- How do the visual design and marketing copy help convey the product's purpose and benefits?
- How does the product tie in with the company's mission and vision?

A great deal of good product sense is having empathy for a product's or service's users. That's why, when answering the above questions while analyzing a product or service, you must try to put yourself in a user's shoes.

Take Snapchat, for example. Sure, you can post photos to your story or send messages to people on Snapchat. But so can iMessage, WhatsApp, Instagram, and Messenger. At a deeper level, Snapchat is about being able to stay in touch with your closest friends in a casual, authentic way. That's why opening up the Snapchat app puts you directly on the camera, in order to make it frictionless to express yourself and live in the moment — two core elements of Snap's company mission. It's also why photos and messages disappear by default — this lowers the barrier to expressing yourself and pushes you to share whatever you captured rather than spending time editing a photo you know will soon be gone.

Contrast this with Instagram, which defaults to the feed to promote consumption rather than visual communication. On Snap's more polished rival, you are made to feel that your posts need to be perfect, lest they be judged by acquaintances and extended family. There's an associated permanence to the photos you post, which takes away some of the whimsicalness that Snap's optimized for. Similarly, Snap's default ephemeral messages set it apart from other messaging platforms like iMessage and Instagram.

While we could go on and on about the two apps, which have overlapping functionality but serve two very different user needs, we want to emphasize that the point of this exercise is to go beyond simply relegating the app to just a "dumb Gen Z" thing or "basically the same as Instagram." By thinking more critically about products in your everyday life, you can sharpen your product intuition, ace interviews, and eventually build successful products in the workplace.

Calibrate Your Intuition by Analyzing Reviews

The daily habit of analyzing products, who they're made for, and what benefits they offer is fine and dandy, but how do you know you're right? How do you know if your reasoning lines up with how others perceive the product and its benefits? One way to calibrate yourself and fine tune your intuition is by analyzing customer reviews.

By looking at the positive reviews and press for the product, you can see how user benefits are described and what is expected of the product. Reading these positive reviews helps you better articulate user benefits; it also helps you notice benefits you might have taken for granted. Similarly, by reading negative reviews, you can understand how the products you encounter are falling short in meeting user needs. By comparing the issues the negative reviews flagged against your own product evaluation, you can start to develop a more critical eye. Additionally, negative reviews are a great source of ideas for product brainstorming.

Reddit is a great place to see unfiltered conversations about products and services. For apps, also check out the App Store and Google Play Store. For enterprise products, check out G2 Crowd and Gartner Special Reports. For physical products, check out the reviews on Amazon. (While you're there, help us immensely by spending two minutes to rate and review our book — we greatly appreciate this!).

How to Build Your Business Sense

While it's exceedingly rare to be explicitly asked business questions like "How do you measure the health of the enterprise sales pipeline?" or "How do you model free cash flow from revenue?" having some general business knowledge is crucial. Why? Because cash rules everything around me (C.R.E.A.M.).

Wu-Tang reference aside, the honest truth is that, in the workplace, having the best technical skills possible won't matter if you solve the wrong business problems. By following the money and understanding how the products you work on help the business make more money, you give yourself a better chance of working on high-leverage technical projects. Stronger business sense helps even your product-sense get sharper, since ultimately knowing what products to build and what product metrics to improve upon stems from the company's business model and strategy.

The Daily Habit You Need to Build Your Business Sense

Much like the daily habit of understanding user incentives we recommended earlier in order to develop product sense, asking yourself the following questions when you encounter a new business can help develop your business sensee:

- ***Business Model:*** How does the business monetize? What product levers can be pulled to improve the business's ability to monetize?
- ***Metrics:*** Which key performance indicators (KPIs) would I measure if I were working on this business? What factors and variables influence those particular metrics?
- ***Landscape:*** How does the business fit into the broader ecosystem of the companies comprising that industry? What companies does the business compete with, and what companies does it partner with?

Another way to grow your business awareness is by reading some of the best business books out there. We'll admit, there's a lot of fluffy business books out there, most of which are just self-help manuals or written to pad the author's ego. And plenty should have been a blog post but got padded with filler to become a book. Then there's books like *How to Win Friends and Influence People* or *Think & Grow Rich*, which, while interesting, don't really help you foundationally understand business. As such, we curated a list of what business and product books helped us out the most in our career: acethedatascienceinterview.com/business-books.

How to Hack Your Domain Experience

We'll let you in on a psychological quirk interviewers have which you can hack: your interviewers likely live in a bubble. They're knee deep in the problem in workplaces designed to be hard to unplug — eating the company-provided free three meals a day, talking about the problem at dinner with their teammates, and thinking about work while commuting back home on the company shuttle. That's why it shouldn't be surprising that, come interview time with you, an outsider, they forget you don't have as much context as they do! Bubbles are real, y'all, and techies love to live in them!

While unfair, the data scientist's familiarity and time spent with the product and company can cloud their ability to accurately assess your product sense. Plus, you might be interviewing against internal candidates with better context on the problem. Or you might be competing against candidates who worked in similar businesses, so naturally get a leg up. That's why doing your homework is one of the best ways to level up.

Doing Your Homework: Uber Eats Example

As a real-world example, let's say that you had an upcoming interview with the Uber Eats team and wanted to prepare for any product or business case questions you might be asked. To prepare, you should learn how Uber as a whole makes its money and how much revenue comes from their transportation products versus their delivery business. You should dig deeper into Uber Eats in order to understand how it fits into Uber's overall strategy as a logistics and transportation company. In addition, learn about the common metrics used to measure two- and three-sided marketplaces. Finally, prior to the interview, you should attempt to uncover the key inputs for Uber's pricing and payout algorithms that determine how much it charges a customer and how much it pays the delivery driver and restaurant.

Most of the research described above can be done with information available free-of-charge on the web, by searching "company name business model." Google News reports what financial analysts say about the company. If the company is public, looking at its earnings reports provides another good way to hone your business sense, and this lets you directly see what key business metrics are being tracked. If the company you are analyzing isn't public, see if there are comparable public businesses, since they'll most likely use similar metrics to what the private company tracks internally.

Another great resource which candidates unfortunately tend to underutilize (to their great peril) is a company's engineering blog. This resource gives you an inside look into the business and the technical systems underlying it. For instance, Uber Eats (from the above example) often publishes blog posts with titles like "Optimizing Delivery Times on Uber Eats" and "Food Discovery with Uber Eats." You could discover and deduce a great deal about the business from posts like these. DoorDash's Engineering blog wouldn't be a bad place to look either!

Doing Your Homework Means Use the Damn Product

It's easy to be an armchair analyst, reading earnings reports and blogs about the product. But at some point, it's absolutely necessary to do your own due diligence by exploring the product on your own.

And even though you probably knew this, and maybe the recruiter told you to do this already, we are surprised by how many candidates we've coached that skip this crucial step.

So, let's be honest — how are you going to reason about a high-level business strategy if your fundamental understanding of the product is weak? In the Uber Eats example, you, like most candidates, have most likely ordered food before, but have you also tried to download the Uber Driver app? Have you looked up the UI that restaurants use to fulfill orders? Have you looked at Uber's marketing website targeted at signing up new eateries, which explicitly spells out the value props that Uber offers restaurants? Putting in this extra effort to understand the product from its multiple angles would easily put you in the top 1% of candidates.

This much company and product research is crazy, right?

Yes, dear reader, we realize this is a lot of work to do before each interview. But the more company and product knowledge you can sneak into your answers, the better you'll do. Additionally, a common interview question — especially at smaller companies — is "Did you have a chance to use our product?" followed up with "What did you think? Any ideas on how to improve it?" This amount of preparation will help you knock these questions out of the park.

Finally, even if there aren't product-sense questions directly concerning the information you researched, don't be dismayed. Your effort wasn't wasted! At the end of the interview, your research would enable you to ask the interviewer more intelligent questions than would have been possible otherwise. Moreover, it would demonstrate your passion and willingness to put forth extra effort. Instead of asking the interviewer run-of-the-mill questions like "What's your favorite part about working at Uber Eats?" you can instead comment, "I was reading about the food delivery time estimation algorithm on your blog and found X fascinating. I was curious why you used approach Y, and if you ever thought about trying out Z instead?" Suddenly, you're showing enthusiasm and insight into the business, offering a suggestion, and gaining the opportunity to learn something new and meaningful in the process.

Metrics for Product & Case Interviews

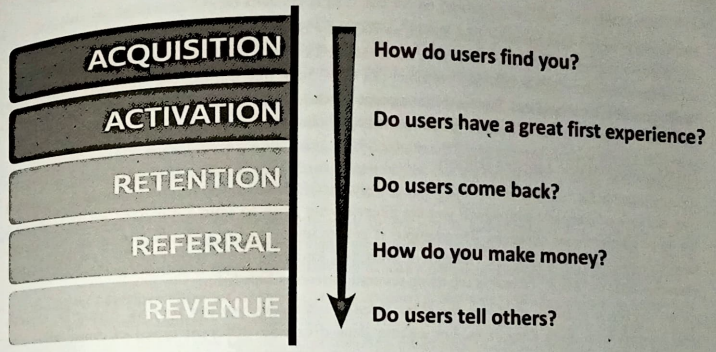
User Acquisition Funnel

Before we address the art of metric selection, it's key we cover the popular marketing concept of a customer acquisition funnel. There is no one funnel to rule them all — you'll see different frameworks depending on the product's use case (B2B vs. B2C), how the product is monetized (one-time purchase vs. recurring vs. freemium). One specific customer lifecycle we like is the pirate metrics framework created by the founder of startup accelerator 500 Startups, Dave McClure. It's called *pirate metrics* because the funnel steps form the acronym *AARRR*.

User Acquisition Metrics

User acquisition metrics try to capture how people are finding and trying out your product. Standard metrics used to track success in this area include new user counts, sign-up conversion rates, and customer acquisition costs (CAC). You'll often hear the word "top of funnel" when referring to this stage of the customer journey. For Pinterest, relevant user acquisition metrics might be the number of new users who download the app or the number of new customers who create an account per week. Related metrics Pinterest would track for this step include the sign-up conversion rate and customer acquisition costs (CAC).

AARRR Pirate Metrics



User Activation Metrics

User activation metrics refer to the point at which a user has successfully onboarded and reached the product "aha" moment — the spot where they experience the core value of the product. For DoorDash, it could be the number of people who make their first delivery order per week. On Instagram, it could be after a user views 10 unique posts in their newsfeed, or after they make their first post or story.

User Engagement

While not an explicit step in the AARRR pirate metrics framework, we thought it was important to bring up the term user engagement. After user activation, it's important to then look at how often and how well users interact with the product. The product interaction being measured could be time spent on platforms like Facebook or rides booked for Uber. A common way to incorporate frequency into the measure of user engagement is by looking at the unique number of users who took a core action within a fixed time period. For example, Facebook measures daily active users (DAU), weekly active users (WAU), and monthly active users (MAU).

User Retention Metrics

User retention refers to whether or not users keep coming back to a product or service over a prolonged period of time. The process of users joining and then leaving permanently is called churn. Maximizing retention and minimizing churn is a primary focus of most businesses, because acquiring a new customer is typically tougher and more costly than retaining an existing user or reactivating a prior user who has churned. Metrics to measure user retention include monthly retention and monthly churn.

User Referral

Referral, where a user shares the product with others, could technically happen at any time. However, in practice, users typically invite others only after having been activated, engaged, and retained on the product for a while. To quantify virality, growth marketers use the k-factor, which is just the

average number of referrals sent per user multiplied by the conversion rate of each referral. A k-factor over 1 would indicate exponential growth.

User Revenue

The popular funnel idea starts to break down, as you may be able to make money from the user at any time — not necessarily after they have been retained or sent referrals. One way to measure a user's revenue is the *lifetime value per customer* (LTV) — how much money a customer brings into the business before they churn. Sustainable companies seek to have a high LTV-to-CAC ratio, so that their marketing efforts pay off before the user eventually churns.

Do Your Homework: Metrics Edition

As mentioned in the "do your homework" section, you need to look up the primary metrics associated with the type of company with which you are interviewing prior to your interview, so as to have a leg up on metric-related questions. For example, if you're interviewing with an enterprise Software-as-a-Service (SaaS) company like Hubspot or Workday, knowing what terms like "ACV" (average contract value) or "MRR" (monthly recurring revenue) mean is helpful.

As we explained at the start of *Ace the Data Science Interview*, we are merely trying to refresh your memory with quick deep dives. For a better deep dive into different types of companies and their associated metrics, we recommend the book *Lean Analytics* by Alistair Croll and Benjamin Yoskovitz. The book devotes entire chapters to measuring such different business models as a freemium mobile app and a two-sided marketplace. (For a full list of books we recommend, visit acethedatascienceinterview.com/best-books-for-data-scientists)

What Makes a Metric Good or Bad?

Now that we understand the stages of the user journey, we are one step closer to addressing the most common product interview question you'll face: defining an appropriate metric for a new product or feature. An example interview question of this nature is "How would you measure the success of Facebook Dating?"

Product-focused tech companies ask this type of interview question because data scientists often help product managers determine the best analytics to measure the success of a new product launch or feature change. Note, these conversations don't necessarily happen just at launch time: often it helps to have the end result in mind before building a new product or feature. As such, often right after the product brainstorm phase, when ideas are being triaged to see if they should make it into the roadmap, conversations about what success would look like occur. Working to define metrics isn't just a skill for data scientists: data analysts and business intelligence engineers also play a role in building out the dashboards to visualize and monitor the health of the product post launch.

Before we jump into the framework for answering a product definition question, it's important we clarify what makes a metric good or bad in the first place. It's easier to start with examples of what makes a metric bad, since a good metric essentially manages to avoid the flaws of bad metrics.

Examples of Bad Product Metrics

Here's some common types of bad metrics, as well as an example of each for the Facebook dating question above:

- ***Vanity metrics:*** These metrics sound nice, but don't capture anything meaningful. For example, dating profiles viewed could be a proxy for engagement, but if this number is too high or low,

does it really tell us if the app is working well? Does it really impact the number of meaningful relationships formed from the product?

- ***Irrelevant Metrics:*** These metrics aren't tied to the business goal. For example, time spent using Facebook Dating. Does it really matter? Sure, it's an indicator that there is use, but it doesn't capture the true value a dating app offers. Time spent is better for media consumption-related products like YouTube and Netflix, not activity-driven dating apps.
- ***Impractical Metrics:*** An example of an impractical metric is the number of 3rd dates that occurred. While a good sign of a meaningful match being made, how would you even measure this? More advanced dating apps have a "did you meet?" user prompt, or they use NLP on the conversation to determine if they think you met, but this approach likely works only for a first date, after which the conversation usually moves off-app.
- ***Complicated Metrics:*** Is the metric easy to explain to stakeholders? If a complex metric changed, would it be easy to understand what actually happened, or would you have to break into its sub-component parts and definition to really understand which pieces were affected?
- ***Delayed Metrics:*** Number of marriages that occurred. Not only is this impractical to know (just like the 3rd date metric), it would take a very long time to figure out because people tend not to get married within a few months of connecting on an app. While downstream success metrics have their place, you want to track leading indicators so that data can be collected earlier, which enables you to make decisions and notice problems faster.

What Makes a Good Metric Good?

By reversing what makes metrics bad, you end up the following four qualities good metrics possess:

- ***Meaningful:*** Tied to business goals; isn't easily gamed; is actionable and can drive decisions
- ***Measurable:*** Simple to consistently and reliably track
- ***Understandable:*** Easy for stakeholders to understand; intuitive to know what is being measured based on the name
- ***Timely:*** Can be collected within a reasonable time frame

North Star & Guardrail Metrics

When asked an interview question about defining metrics, you need to go beyond the simple framework of only suggesting good metrics and avoiding bad ones. Besides mentioning the most important and relevant metric to the problem at hand (known as a *north star* metric), you should also mention guardrail metrics (also known as counter metrics). Guardrail metrics are business metrics that should not be degraded while optimizing the metric of interest. These are important to monitor because it's easy to boost a given metric at the expense of counter metrics or other KPIs.

For example, when Kevin was working on Facebook Groups, trying to reduce harmful content like hate speech and spam, his team's topline metric was to decrease the prevalence of harmful posts. One way to achieve this goal is to remove 99% of posts from the feed. You can't have harmful content on newsfeed if there isn't any content on the newsfeed to begin with! Obviously, such an approach doesn't make sense. That's why, for every A/B test, they'd also monitor guardrail metrics (posts made, posts viewed, number of likes and comments per post).

It is good to bring up counter metrics with the interviewer after they agree with your key success metrics. Proactively having this discussion shows that you realize building a product isn't just about

average number of referrals sent per user multiplied by the conversion rate of each referral. A k-factor over 1 would indicate exponential growth.

User Revenue

The popular funnel idea starts to break down, as you may be able to make money from the user at any time — not necessarily after they have been retained or sent referrals. One way to measure a user's revenue is the *lifetime value per customer* (LTV) — how much money a customer brings into the business before they churn. Sustainable companies seek to have a high LTV-to-CAC ratio, so that their marketing efforts pay off before the user eventually churns.

Do Your Homework: Metrics Edition

As mentioned in the "do your homework" section, you need to look up the primary metrics associated with the type of company with which you are interviewing prior to your interview, so as to have a leg up on metric-related questions. For example, if you're interviewing with an enterprise Software-as-a-Service (SaaS) company like Hubspot or Workday, knowing what terms like "ACV" (average contract value) or "MRR" (monthly recurring revenue) mean is helpful.

As we explained at the start of *Ace the Data Science Interview*, we are merely trying to refresh your memory with quick deep dives. For a better deep dive into different types of companies and their associated metrics, we recommend the book *Lean Analytics* by Alistair Croll and Benjamin Yoskovitz. The book devotes entire chapters to measuring such different business models as a freemium mobile app and a two-sided marketplace. (For a full list of books we recommend, visit acethedatascienceinterview.com/best-books-for-data-scientists)

What Makes a Metric Good or Bad?

Now that we understand the stages of the user journey, we are one step closer to addressing the most common product interview question you'll face: defining an appropriate metric for a new product or feature. An example interview question of this nature is "How would you measure the success of Facebook Dating?"

Product-focused tech companies ask this type of interview question because data scientists often help product managers determine the best analytics to measure the success of a new product launch or feature change. Note, these conversations don't necessarily happen just at launch time: often it helps to have the end result in mind before building a new product or feature. As such, often right after the product brainstorm phase, when ideas are being triaged to see if they should make it into the roadmap, conversations about what success would look like occur. Working to define metrics isn't just a skill for data scientists: data analysts and business intelligence engineers also play a role in building out the dashboards to visualize and monitor the health of the product post launch.

Before we jump into the framework for answering a product definition question, it's important we clarify what makes a metric good or bad in the first place. It's easier to start with examples of what makes a metric bad, since a good metric essentially manages to avoid the flaws of bad metrics.

Examples of Bad Product Metrics

Here's some common types of bad metrics, as well as an example of each for the Facebook dating question above:

- ***Vanity metrics:*** These metrics sound nice, but don't capture anything meaningful. For example, dating profiles viewed could be a proxy for engagement, but if this number is too high or low, does it really tell us if the app is working well? Does it really impact the number of meaningful relationships formed from the product?
- ***Irrelevant Metrics:*** These metrics aren't tied to the business goal. For example, time spent using Facebook Dating. Does it really matter? Sure, it's an indicator that there is use, but it doesn't capture the true value a dating app offers. Time spent is better for media consumption-related products like YouTube and Netflix, not activity-driven dating apps.
- ***Impractical Metrics:*** An example of an impractical metric is the number of 3rd dates that occurred. While a good sign of a meaningful match being made, how would you even measure this? More advanced dating apps have a "did you meet?" user prompt, or they use NLP on the conversation to determine if they think you met, but this approach likely works only for a first date, after which the conversation usually moves off-app.
- ***Complicated Metrics:*** Is the metric easy to explain to stakeholders? If a complex metric changed, would it be easy to understand what actually happened, or would you have to break into its sub-component parts and definition to really understand which pieces were affected?
- ***Delayed Metrics:*** Number of marriages that occurred. Not only is this impractical to know (just like the 3rd date metric), it would take a very long time to figure out because people tend not to get married within a few months of connecting on an app. While downstream success metrics have their place, you want to track leading indicators so that data can be collected earlier, which enables you to make decisions and notice problems faster.

What Makes a Good Metric Good?

By reversing what makes metrics bad, you end up the following four qualities good metrics possess:

- ***Meaningful:*** Tied to business goals; isn't easily gamed; is actionable and can drive decisions
- ***Measurable:*** Simple to consistently and reliably track
- ***Understandable:*** Easy for stakeholders to understand; intuitive to know what is being measured based on the name
- ***Timely:*** Can be collected within a reasonable time frame

North Star & Guardrail Metrics

When asked an interview question about defining metrics, you need to go beyond the simple framework of only suggesting good metrics and avoiding bad ones. Besides mentioning the most important and relevant metric to the problem at hand (known as a *north star* metric), you should also mention guardrail metrics (also known as counter metrics). Guardrail metrics are business metrics that should not be degraded while optimizing the metric of interest. These are important to monitor because it's easy to boost a given metric at the expense of counter metrics or other KPIs.

For example, when Kevin was working on Facebook Groups, trying to reduce harmful content like hate speech and spam, his team's topline metric was to decrease the prevalence of harmful posts. One way to achieve this goal is to remove 99% of posts from the feed. You can't have harmful content on newsfeed if there isn't any content on the newsfeed to begin with! Obviously, such an approach doesn't make sense. That's why, for every A/B test, they'd also monitor guardrail metrics (posts made, posts viewed, number of likes and comments per post).

It is good to bring up counter metrics with the interviewer after they agree with your key success metrics. Proactively having this discussion shows that you realize building a product isn't just about

optimizing a specific set of feature-related metrics, but more about holistically making sure the product is benefiting the business at-large.

3-Step Framework to Answer Product Metrics Definition Questions

Now that we understand what good metrics look like, and the stages of the user journey, we can cover the specific framework you can use to answer the question, "How would you measure the success of Facebook Dating?"

Step 1: Clarify the Product & Its Purpose

We need to start by defining the problem. This can be done by clarifying the business purpose behind the product, the product's goal, who this feature is for, and what the user flow looks like. For the Facebook Dating example, you might clarify:

- Why is Facebook interested in a dating product? Is it to boost engagement on the app? Is it to create another surface for ads, which can help increase ad revenue?
- What's the point of Facebook Dating — is it meant for casual dating, or people looking for marriage?
- Is the dating feature for all people, or a certain age demographic or orientation?
- How does the product work? Is this a stand-alone app from Facebook, or integrated to be within Facebook itself? Is there a Tinder-esque swipe-based UI? Are there any gimmicks involved, like women initiating the conversation, similar to Bumble?

In the situation with Facebook Dating, let's assume that you learn that it's an app, similar to Tinder, but with more elaborate profiles to foster more intentional swipes and matches. You learn it's being launched in New Zealand first, because it approximates the larger English-speaking populations like the U.S. and the U.K. where Facebook eventually wants to launch the product.

Step 2: Explain the Product & Business Goals

While you might expect that you can just ask the interviewer what the product and business goal are, often, the interviewer is looking to you to synthesize what the main purpose behind the product is, and how it ties into the business. You'll get bonus points for tying it back to the company mission. This is exactly why earlier we strongly recommended you to do company and product research — it makes a huge difference on this step.

In the case of Facebook Dating, it relates back to the product czars wanting to drive engagement to the app. If you log in to check your matches, it's easy to see one of the super-targeted ads and drive revenue for the company. Plus, it keeps the rest of the Facebook ecosystem attractive — if you check your notification about a new match, maybe you also check your notification about a group post or a friend's birthday at the same time. Plus, with in-built messaging, it keeps people within the Facebook, Instagram, Messenger, WhatsApp ecosystem.

Lastly, the Facebook company mission is to help you develop meaningful relationships. Keeping in touch with family, interacting with your community via groups, and messaging with your friends are all part of the mandate. So why not help you in the romantic relationship department as well?

Step 3: Define Success Metrics

It is crucial to determine and measure the main actions a user needs to take in order to drive the product and business goals. You can follow the user-acquisition funnel we talked about earlier to

serve as a framework on how to guide the analysis. As you are defining the metrics, remembe restate how they align with the product and business goals you had mentioned earlier.

Acquisition metrics: How many users sign up and how many users fill out their profile could great metrics to measure top-of-funnel. This is important, because people don't think of Facebo as a dating company, and it's important to know if Facebook is in the news for privacy and data-u concerns. This reputation can get in the way of users trusting it for their most intimate needs.

Activation metrics: These could be onboarding metrics if the profile is filled out and a certain numb of photos are uploaded. Why? Because rich detailed profiles are the point of a dating app seeking be more long-term-relationship focused. Other activation metrics might be if they view ten profile or if they manage to get a single match. You want them to hit a point of user delight.

Engagement metrics: Some example engagement metrics can be: how many matches were form per user, how many matches were made overall, swipes done per user, and real life meetings (th could be indicated by a phone number swap). This would reflect if they were able to foster actu meaningful connections. Retention metrics could be an off-shoot of these core engagement metric like what percentage of users swipe on the app after 28 days of signing up, or, of the users w managed to get a match last month, what percentage of them managed to get a match this month?

Revenue metrics: Revenue could be measured by the number of paid memberships or upgrad bought. If Facebook Dating is completely free but ad-supported, then revenue from ad impressio could be measured per user.

However, paid memberships or ad revenue isn't a priority on launch for a company like Faceboo which is primarily trying to nail product-market-fit first. Facebook tends to take a long-term approa to monetization. Their biggest worries are feature adoption and user retention; they know they c easily toss in ads to support the product when needed.

4-Step Framework for Diagnosing Metric Changes

The second most common product interview question that data scientists face is one of diagnosi metric changes. For example, you might be asked: "Instagram's average number of comments per po is declining — how would you troubleshoot this?"

Step 1: Scope Out the Metric Change

Before even jumping into a solution, you need to clarify and gather context.

Questions to ask:

- ***Metric Definition Nuance:*** What does the metric in question mean? Are we dealing wi proportional metrics? If so, by isolating which part of the ratio changed — the numerator or t denominator — you're able to offer a more targeted hypothesis for what is driving the met change.
- ***Importance:*** Is this metric actually consequential to the team? With thousands of metrics bei tracked at a company like Facebook, there's constant fluctuation in all parts of the business, b not all changes are equal or relevant.
- ***Time frame:*** Is this a singular occurrence or an ongoing issue? A sudden change or an ongoi trend? At what granularity (over a day, a week, a month) has the trend occurred?
- ***Magnitude:*** How big is the change, in both relative and absolute terms? Are we comparing t change to last week or on a year-over-year basis?

Say that the interviewer tells you that the number of posts has been increasing on the platform. The number of comments being made on the platform has also been increasing, but at a lesser rate, causing the average number of comments per post to be decreasing. This has been a slow decline over the last 6 months, and today's average comments per post is down 10% from this time last year. This trend concerns leadership because people engaging with posts in the comments section is an important part of Instagram, and helps the product be more interactive rather than consumption oriented.

Step 2: Hypothesize Contributing Factors

Now that you know what the metric change entails, it's time to start brainstorming possible issues that could have occurred to cause the metric to change. Generally, contributing factors fall into four different buckets:

- ***Accidental Changes:*** Is the metric drop even real? Were there any data generation processes or bugs with instrumentation and logging that caused a change that isn't actually reflective of user behavior?
- ***Natural Changes:*** Is the change simply due to seasonality? Could the day of the week, or the fact that it's a holiday, or a change in weather cause the product changes you are seeing?
- ***Internal Changes:*** Issues like new feature launches, bug fixes, intentional product changes, or new marketing campaigns going live can cause metrics to change.
- ***External Changes:*** Competitors launching new products, or more macro events such as a pandemic or a recession, can cause shifts in user behavior.

One good way to brainstorm factors, especially internal ones, is to walk up the product funnel. You can do this by starting with the feature at hand (local) and working your way upstream to broader issues that could affect the metric. For example, in the Instagram case, we can look at the following:

- Were there any UI changes to how users can make comments? Maybe the comment composer UI got less emphasis? Or maybe there is stronger comment moderation in place, leading to more automatically deleted comments?
- Were there any changes to how users can give feedback on a post (like, comment, or share)? Maybe the UI changed so that liking or sharing a post to your story got more emphasis than commenting on a post, leading to cannibalization?
- Were there any changes to the types of posts in the feed? Maybe there are more ads with comments turned off? Or maybe there are more reels, which are more consumption based, rather than things people comment on. Maybe the ranking model changed, favoring posts which are more likely to be liked and shared, rather than commented on?
- Were there any changes to the feed itself? Maybe comments and posts are both being cannibalized by stories, reels, or Instagram Explore?

As you walk backwards through the product funnel, listing a few of the likely culprits behind the metric change, the interviewer may stop you and ask you how you'd validate each hypothesis. Otherwise, stop yourself — it's easy to list off 20 potential reasons for why something happened, but you don't want to overdo it. The meat of these types of problems is explaining what metrics you'd look at to validate each factor, not how many you can come up with.

Step 3: Validate Each Factor

Validating hypotheses means slicing and dicing data into many different segments, hunting for insights that validate or disprove your hypothesis. For example, you could slice along user demographics

(i.e., age, gender, location, language, device type) to see if any of these factors correlate to a trend of fewer comments per post.

You should also look at upstream metrics. This is where the product funnel approach we mentioned earlier becomes applicable, where you continuously zoom out until you find something that can explain the metric change. For the Instagram example, metrics to look at include:

- Number of views on the comment composer or comments sections of a post, to understand if there is less emphasis on commenting.
- Ratio of likes to comments per post, and ratio of shares to comments per post, to understand if this issue is specific to comments or a more general issue with post engagement.
- Amount of engagement per post relative to trends in engagement per story and engagement per reel to see if the issue is Instagram feed losing engagement to other features within the app.

Generally, an interviewer isn't expecting you to ramble on about every dimension you'd cut or every metric you'd check. By prioritizing which hypotheses are most likely based on your product sense and product research, you can narrow down the space and focus the conversation on the most likely culprits.

Step 4: Classify Each Factor

After you've explained how you'd validate each potential factor, the interviewer will often share the results of your hypothetical data analysis. Based on this information, the next step would be to bucket each of the hypotheses into the following categories:

- ***Root cause:*** The root cause of the metric change
- ***Contributing factor:*** While not the root cause, still contributing to the root cause
- ***Correlated result:*** Factors that are symptoms of the root cause but not a contributing factor
- ***Unrelated factor:*** Factors that are unrelated to the metric change

For the Instagram example, you find that overall activity on Instagram's feed looks normal; people are still viewing posts, and liking and sharing posts at the same rate. This tells us the issue is localized to comments — not a more general engagement or Instagram feed issue.

From slicing comments per post by account age, you notice that posts made from newer accounts are experiencing a sharper decline in the average number of comments than posts made from older accounts. Running cohort analysis, based on when a poster joined Instagram, you are able to confirm that about 6 months ago, the average number of comments made per post by an old user starts to diverge from a new user. This might give you intuition that something changed for new users about 6 months ago. From auditing the product and checking in with the PMs of the Instagram onboarding team, you find out that there was an interstitial added by the Trust & Safety Team that makes new users aware they can turn off comments on a post. This feature has led to some percentage of posts simply having no comments on them at all. By removing posts with comments turned off from the analysis, you realize there was no decline in the number of comments per post. Boom! You found the root cause for the metric change!

Assessing Metric Trade-Offs

Lauded economist Thomas Sowell insisted, "There are no solutions, only trade-offs." As a data scientist, you'll be asked to weigh in on tough business decisions, where there is no clearly right answer. That's why, to assess your judgement, you might get asked a question like "We tested a new feature that shows more ads on LinkedIn. It increases revenue by 1% but hurts time spent by 3% — should we ship it?"

The steps to approach a metric trade-off problem are very similar to the steps in the frameworks for defining a metric and troubleshooting a metric change. First, to reason about a trade-off, you'd need to know more details about what the metrics in question actually mean. Thus, you follow the first step of the metric troubleshooting framework, where you clarify what the metric definitions are (except now you have two metrics to ask clarifying questions about instead of one).

The second major aspect of solving this type of problem is understanding the product and business goal. By applying your own research into the company, along with your intuition, and then asking smart clarifying questions, you can determine which metrics are more important than others. By doing so, you can reason on how to proceed with the trade-off. After collecting more information about the trade-off, next steps usually are either to

- revert the feature change since the trade-off is not acceptable,
- minimize the trade-off's impact by brainstorming new product interventions, or
- accept the metric trade-off since it's justified.

It can be hard to give a definitive recommendation on what to do, because, in reality, you'd be solving this collaboratively with other stakeholders. However, since this is an interview, the interviewer isn't looking for a specific correct answer, but more your thought process. As such, it's okay to explain in what scenario or what more information you'd need to make each of the three different recommendations.

For the LinkedIn example, you could mention that depending on how hard or easy it was to get a 1% revenue gain or a 3% engagement gain from other features, you'd know what to do. This is reasonable, since maybe 1% revenue gains are easy to achieve through more sophisticated means (via better targeted ads), whereas a 3% engagement drop is very hard to recover from and could unwind progress in other strategic areas for the business. You could also mention how maybe the decrease in time spent is mostly from users who are getting poor quality ads, and that there should be product features like "hide this ad" or "block this advertiser" that can minimize the engagement drop while still maintaining the revenue gain.

A/B Testing & Experimental Design

While the mathematical underpinnings of A/B testing were discussed in "Chapter 6: Statistics," in this chapter we dive into the nuances of real-world A/B testing, because it's brought up in many product data science interviews. Usually, at the end of a success metrics definition problem, you'll be asked, "How would you test this new feature?" Interviewers will ask you to walk through the full experimental design setup for a hypothetical test. Often, interviewers steer the conversation towards one of the many practical testing pitfalls you may face.

Experimental Design Setup Overview

When you're faced with a general A/B testing question like "How would you A/B test a 1-click job-apply feature on LinkedIn?" it can be challenging to not ramble aimlessly. To keep your answer focused, make sure to address the four main steps of setting up an experiment:

1. ***Pick a Metric to Test:*** While you'll be tracking a whole host of core and guardrail metrics, narrow each experiment down to a few key metrics that capture the essence of the goal of testing the feature change. Don't cheat by cherry-picking a metric post-hoc based on whatever ended up being statistically significant!

2. ***Define Thresholds:*** Decide on a particular statistical significance level (alpha) which is generally set to 0.05, as well as a power threshold (1 — beta), which is generally set to 0.8. The value for the power is dependent on the minimal detectable effect (MDE) and is usually set by consulting with stakeholders. That is, we can calculate the minimal detectable effect (MDE) at x% power (for example 0.8 MDE for 80% power) for any given sample size and sample variance.
3. ***Decide on Sample Size & Experiment Length:*** Based on the MDE and power, along with the metric variance, one can calculate the required sample size needed for the test. Using this required sample size, the length of the experiment can be determined based on the daily traffic of the feature in question. A good rule of thumb, though, is to run a test for at least two weeks, to account for day-of-week effects typical in most consumer products.
4. ***Assign Groups:*** When deciding the control group and treatment group, we want to randomize these groups sufficiently; otherwise, there will be confounding variables down the line. At large companies, this consideration is often abstracted away for you by the A/B testing infrastructure.

Real-World A/B Testing Considerations

Let's face it: in the messy real world, product and business constraints can get in the way of proper experimental design. As such, data science interviews often touch on the A/B testing pitfalls you're likely to encounter in practice, and how you'd guard against them. Unless you're a seasoned product data scientist with many battle scars from A/B tests gone bad, pay careful attention to the A/B testing nuances below.

When Not to A/B Test

Sure, A/B testing is essential to businesses like Facebook and Amazon. But right off the bat, we need to acknowledge that A/B testing isn't always the correct answer to every product and business problem that arises. As 16th century British playwright and data scientist William Shakespeare once said, "To A/B test, or not to A/B test, that is the question."

Some scenarios where A/B tests typically shouldn't be run:

- ***Lack of infrastructure:*** Having the data engineering infrastructure needed to reliably test isn't a trivial matter; keep this in mind when interviewing at smaller companies that might not have the dedicated resources to perform complex A/B tests.
- ***Lack of impact:*** Don't test things that don't matter. Try to size up the opportunity — if the test pans out, how much impact on the business would you expect? Is this benefit much larger than the time and engineering resources you'd have to spend on the test?
- ***Lack of traffic:*** Without enough people using a feature or performing a certain action, it can be difficult to make a statistically sound conclusion within a reasonable time frame.
- ***Lack of conviction:*** For high-traffic and high-value features like the Amazon "Buy Now" button, it might be worth trying out 60 variants of different copy, color, size, and iconography. But, at some level, conviction and intuition behind why a variant would be the winning option is crucial — especially if you are operating on a much smaller scale than Amazon. While a random monkey-throwing-darts approach might work for stock-market investing, it won't cut it for making great products!
- ***Lack of isolation:*** How would you A/B test a logo? It's not easy to have a "control" group, since these changes make the headlines and get rolled out to every user. In these cases, qualitative methods like customer interviews and focus groups can work better.

In cases where A/B testing is not useful, we can do the following:

- Conduct user experience research via focus groups and surveys to understand what options are better.
- Analyze user activity logs to get a better sense of what option is a better fit.
- Make the product change, but then run a retrospective analysis by looking at historical data to see if the metric that we are interested in responds as we expect.

Dealing with Non-Normality

As mentioned earlier, in the statistics chapter, A/B tests are just dressed up Z and t-tests. These statistical tests assume that the distribution of the random variable of interest (the metric we are testing for) is normally distributed. At larger tech companies, this assumption tends to hold, thanks to plentiful user data and the Central Limit Theorem. But what happens if this isn't the case for some reason?

Several methods exist to deal with non-normality of the test metric:

- ***Bootstrapping:*** The process of generating extra samples randomly for each variant, and then averaging results at the end, in order to invoke the Central Limit Theorem.
- ***Running alternative tests:*** The Wilcoxon rank-sum test is a popular alternative to the t-test and does not assume normality.
- ***Gathering more data:*** With budget and time constraints permitting, collecting more data can give you more confidence that what is being measured is more representative of the true effect on the population at hand.

Dealing with Multiple Tests Simultaneously

Recall the multiple testing problem we covered in the statistics chapter — if you run 100 A/B tests at the same time, by pure chance, you're bound to have some test succeed with a statistically significant result. In interviews, you'll be expected to have a high-level understanding of why this problem exists and how to deal with it.

One way to account for the multiple testing problem is to use Bonferroni Correction, which adjusts the significance level required based on the total number of tests running. Alternatively, you can control the false discovery rate (FDR) or the familywise error rate (FWER). Recall from the ML chapter false positives (FP) and true positives (TP). The FDR is equal to FP / (FP + TP) and, hence, is the rate of type I errors during multiple testing. In a similar vein, FWER is simply the probability of making one or more type I errors when performing multiple tests.

While statistical approaches are helpful, the reality of the situation is that experiments will unfortunately always interact to a degree. Often, the best you can do when faced with surprising or weird results, is to dig into the active experiments and see if anything could have impacted the primary experiment.

Dealing with Network Effects

Generally, it is assumed that during an A/B test, each user is selected to either the control or treatment group at random, that each user is independent, and there is no interference between the control and treatment group. However, this condition doesn't always hold in practice.

Consider a social network like Facebook — the actions of users are likely impacted by that of those around them (network effect). Therefore, it will likely be the case that the behaviors of those in the control group are being influenced by behaviors of those in the treatment group; hence, the true effect is not exactly as is stated.

As a concrete example, say Facebook is testing out a new kind of Facebook Live, which ends up being super entertaining. The test group might see a large increase in engagement metrics due to the new feature, but their increased engagement on the platform overall may drive them to also interact with their friends in the control group more. This spillover can lead to increased engagement for the control group, which causes the overall positive effect of the experiment to be understated because of the control group contamination.

One possible way to control for network effects is to create clusters of similar or connected people and divide these sub-networks into control and treatment groups for better isolation. Mathematically, this separation can be done through various graph partitioning methods (for example, normalized cuts) that place users into various groups based on social interactions.

Dealing with Novelty Effects

A/B tests may have an exaggerated initial effect, due to the novelty effect, where a new feature attracts social media attention and PR hype, causing inquisitive users to rush to check out the new changes. This flock of curious users leads to metrics like time spent and engagement to jump spuriously high. For example, when Facebook first launched emoji reactions like "haha" and "angry," post engagement massively spiked due to curious users trying the feature out. Eventually, though, reaction rates came back down and stabilized at a new, slightly higher, baseline rate.

In the opposite direction, there could be a primary effect, where users are fixed on what is familiar and have an aversion to changes in general. For example, when Facebook originally launched News Feed in 2006, it was reviled by its then 8-million student users. To get a sense of the backlash, consider this — a Facebook Group called "Students against Facebook News Feed" grew to 750k members in just two days. Ten percent of the entire user base was upset enough to join a group boycotting the news feed. We know how that experiment turned out!

To detect primary and novelty effects, you could analyze the test results for new users only — if there is a statistically significant change there, but not for the old users, novelty effects are likely at play. Similarly, to guard against the novelty effect, you can always just run the A/B test on new users. However, this opens up its own can of worms, if new users aren't similar to existing users. For example, at Facebook, back when Nick was on the New Person Experience team, most new people came from developing countries like India, Indonesia, and Nigeria. Usage patterns didn't approximate well to the average tenured user of Facebook, who resided in a more developed country.

Nuances of Taking Action on A/B Test Results

If you ran an A/B test, managed to avoid all the common experimental design pitfalls, got a positive result, and $p < 0.05$, you'd definitely launch it, right? Maybe.

Just because a test reaches statistical significance doesn't mean it's automatically shipped. At large tech companies that run experiments on hundreds of millions of users, it's easy for small differences to become detectable. So even with a small p-value, it's crucial to assess the effect size before shipping the change.

It's not just the direct effect size to consider — what happened to the counter metrics and guardrail metrics you'd set up? For example, if revenue went up, but retention went down, it's not obvious whether to ship it or not. As we mentioned earlier, it's best to confer with product and business stakeholders when making these metric trade-offs.

Another consideration to factor in is that any time you ship a change, there is human labor cost associated with it. There's always a cost of properly deploying and then supporting the feature change

— for example, increased customer support tickets from users who are confused by a new experience. And then there's the chance of hidden bugs in the new feature you're launching that didn't get caught in the test. You need to make sure the impact from the A/B test justifies launching the new variant.

Launch with A/B Test Holdouts

Suppose you ran a successful A/B test and stakeholders have given you the green light to roll out the new feature. Typically, even after launch, there remains a small group of users — usually just a few percent — who don't receive the new experience. This group is known as the A/B test holdout. Because A/B tests are typically run over a shorter time period, holdouts help you quantify the long-term lift from shipping features. This makes identifying potential novelty effects, where metrics tend back towards their original baseline over time, much easier.

For a practical example of how holdouts are used, when Nick was a part of Facebook's Growth Engineering division, his team would create a shared holdout group every quarter. By having the entire team's launched features not affect a tiny portion of users, the Head of Growth could easily measure the team's combined quarterly output, which was useful come performance review and promotion time.

Note that not every A/B test merits a holdout. For bug fixes or very sensitive changes, it can be better to launch the feature to the entire user base. For example, when Kevin was working on Facebook Groups, implementing a holdout for a new model that flagged child trafficking would mean that some small number of holdout users would still see such content. In these scenarios, after getting a signal that the new model was helpful without too much downside risk, we'd say, "F*ck it, ship it!" and launch the model to 100% of users without any holdouts.

Product Questions

10.1. Facebook: Imagine the social graphs for both Facebook and Twitter. How do they differ? What metrics would you use to measure how these social graphs differ?

10.2. Uber: Why does surge pricing exist? What metrics would you track to ensure that surge pricing was working effectively?

10.3. Airbnb: What factors might make A/B testing metrics on the Airbnb platform difficult?

10.4. Google: We currently pay the Mozilla foundation 9 figures per year for Google to be the default search engine on Firefox. The deal is being renegotiated, and Mozilla is now asking for twice the money. Should we take the deal? How would you estimate the upper bound on what Google should be willing to pay?

10.5. LinkedIn: Assume you were working on LinkedIn's Feed. What metrics would you use to track engagement? What product ideas do you have to improve these engagement metrics?

10.6. Lyft: Your team is trying to figure out whether a new rider app with extra UI features would increase the number of rides taken. For an A/B test, how would you split users and ensure that your tests have balanced groups?

10.7. Amazon: If you were to plot the average revenue per seller on the Amazon marketplace, what would the shape of the distribution look like?

10.8. Facebook: Besides posts Facebook is legally obligated to remove, what other types of posts should Facebook take down? What features would you use to identify these posts? What are the trade-offs that need to be considered when removing these posts?

10.9. Amazon: The Amazon books team finds that books with more complete author profiles sell more. A team implements a feature which scrapes Wikipedia and Goodreads to automatically fill in more information about authors, hoping to see an improvement in sales. However, sales don't change — why might this be?

10.10. Snapchat: Let's say Snapchat saw an overall 5% decrease in daily active users, a trend that had been consistent over the week. How would you go about determining the root cause of this?

10.11. Pinterest: Say you ship a new search ranking algorithm on Pinterest. What metrics would you use to measure the impact of this change?

10.12. Netflix: Say a given category, such as sci-fi TV shows, has less total watch time, compared to other similar categories. What metrics would you look into to determine if the problem is that people aren't interested in that category of content (demand problem), or if the category has interest but the content is bad (supply problem)?

10.13. Apple: Say you have data on millions of Apple customers and their purchases made at physical Apple retail stores. How could customer segmentation analysis increase a store's sales performance? What techniques would you use to segment brick & mortar customers into different groups?

10.14. Facebook: If 70% of Facebook users on iOS also use Instagram, but only 50% of Facebook users on Android also use Instagram, how would you go about identifying the underlying reasons for this discrepancy in usage?

10.15. Capital One: How would you assess the stickiness of the Capital One Quicksilver credit card?

10.16. Google: Say you worked on YouTube Premium, which is an ad-free version of YouTube bundled with YouTube Music — a music streaming service. You're launching the product in a few new countries — how would you determine pricing for each country?

10.17. Twitter: Should Twitter add Facebook-style emoji reactions (love, haha, sad, angry, etc.) to tweets?

10.18. Slack: What metrics would you use to measure user engagement at Slack? How would you be able to tell early whether or not user engagement was declining?

Product Solutions

Solution #10.1

A bad answer glazes past the nuances of the social graph and jumps straight into defining metrics. For clarity — not just for you, dear reader, but also for the hypothetical interviewer asking this question, we'll structure our answer into three steps.

Step 1: Explaining User Behavior on Each Platform

Before explaining how the social graphs of Facebook and Twitter differ, it's crucial to first consider how each platform's average user interacts with their respective platform. Facebook is mostly about friendships, and so two users on the same social graph are mutual friends. Twitter, on the other hand, is more focused on followership, where one user typically follows another (who is usually an influential figure) without getting a followback. Thus, Twitter likely has a small number of people with very large followerships, whereas, on Facebook, that pattern appears less often.

Step 2: Describing the Social Graph and Its Differences Between Facebook and Twitter

Modeled as a graph, let's say each user is represented as a node, and the edge linking two nodes denotes a relationship (typically, friendship on Facebook and fellowship on Twitter) between the two users whose nodes the edge connects. Most nodes on Twitter would have low degrees, but a small number of nodes (those of influential people) would have very high degrees, resulting in a "hub-and-spoke" social graph for that platform.

Step 3: Defining Metrics to Measure the Social Graph

One way to quantify the difference between the two platforms' social graphs is by looking at the distributions of friendships/followerships represented by the social graphs of each platform's typical users. Because a typical node's degrees — that is, the number of connections it has to other nodes — should capture the difference in these platforms' social graphs, one concrete metric would thus be the average degree among all nodes on each platform. Alternatively, to obtain an even more detailed understanding of the differences between the two social networks, we could construct box-and-whisker plots of the platforms' degrees among all their respective nodes.

Still another way of looking at the two graphs would be to check the distribution of degrees across all nodes in each platform's network. In all likelihood, the Twitter distribution would show a greater amount of right skewness than that of Facebook. Metrics that quantify a distribution's skewness or kurtosis could thus be used to describe the difference between the two platforms' degree distributions and, hence, social graphs.

Solution #10.2:

For any metrics definition question, it's important to first explain the business goal of the product or feature and then explain the related stakeholders, before ultimately landing on good metrics to measure the success of that feature.

Step 1: Explain Uber's Motivation for Surge Pricing

You don't have to be an econ major to realize that surge pricing is about fixing imbalances between supply and demand. In the case of Uber, such an imbalance could result from either a lack of drivers or an excess number of potential riders. Therefore, surge pricing's goal would be to increase supply by enticing more drivers to use the app through increased pay, and reduce demand by raising prices for riders.

Step 2: Consider Stakeholders Related to Surge Pricing

A nuanced answer would consider the various stakeholders involved in surge pricing, beyond just the immediately obvious drivers and riders. For example, a good candidate would mention associated business functions within Uber that could be affected by the surge pricing algorithm not working effectively.

Step 3: Define Metrics & Counter Metrics for Surge Pricing

Surge-specific metrics are the duration of the surge, the surge pricing multiplier, and the number of riders and drivers in the affected area. We should also track the following metrics during surge periods: number of rides taken, number of rides cancelled, total revenue made for both Uber and Uber's drivers, total profit made by both Uber and Uber's drivers, and the average ride wait time. These are all standard metrics, but critical to monitor to ensure the business is healthy during surge periods.

In addition, topline metrics like user's lifetime value (LTV), driver retention, rides taken, daily active riders, and drivers should also be tracked, so that we can be sure surge pricing isn't having adverse impacts on the business overall.

As with any good metrics definition question, a discussion on counter metrics is important. Even if the surge pricing is bringing in extra money, one counter metric to implement would be net promoter score (NPS). Surge pricing can annoy users, for whom frequently fluctuating sky-high prices can be a source of frustration. And then there's the potential for mistakes, or users in a less-than-sober state accidentally making a purchase (we authors can neither confirm nor deny that $158 Uber from San Francisco to Palo Alto after a fun night out). It can even be a PR risk. Like clockwork, every New Year's, there's a news story about someone getting drunk and taking an $800 Uber by accident. Between bad PR, frustrated users, and potentially increased support tickets, some metric to make sure it's a quality program is key.

Solution #10.3:

A good answer would demonstrate not just thorough A/B testing knowledge, but some understanding of the Airbnb product. Also, to demonstrate your problem-solving attitude, it can be a wise move to not just mention the difficulties, but briefly also mention techniques to deal with these A/B testing problems. Finally, to earn bonus points on this problem, remember to relate your own A/B testing war stories if you think it would also be relevant to Airbnb. By coloring your answer with your own experience, you're demonstrating your time spent in the trenches, which can help separate you out from the more green candidates.

Issue #1: Complexity of User Flow

Testing Airbnb platform metrics would be difficult because of the complexity of the company's booking process, which starts with a user search and often requires user-host communication before a booking can be finalized. Alternatively, a user can sometimes book a rental without having to contact the property host at all. Also, adding to the complexity of conducting A/B tests on Airbnb platform metrics, booking flows frequently depend on factors outside of Airbnb's control, such as host responsiveness to messages left by prospective renters. Therefore, since a booking can be instantaneous or a long, drawn-out process, timeboxing an experiment could be difficult. Because of these issues, any data generated by tests on the booking process would most likely be quite noisy.

To mitigate these issues, we want to make sure we are looking at the correct non-intermediary metrics. For example, there could be a few steps in between searching and booking, but the searching to booking conversion rate should be the main metric. Additionally, we want to employ best practices on managing the data generation and collection process (logging and other downstream event collection).

Issue #2: User Bucketing Due to Multiple People & Devices in Booking Flow

A second source of complexity arises because planning a vacation often involves multiple people. Since even a single person could employ multiple devices during the booking process, it could involve multiple and discontinuous uses of the Airbnb platform from different IP addresses and, to further complicate things, occur over an extended time period. Ideally, there would be a clear one-to-one user-to-device mapping, and, also ideally, the booking process would occur nearly instantaneously. In that case, testing would be easy since the variables being tested for a specific user (e.g., demographics, booking details, time needed to book, and so on) could be clearly identified and then measured. However, different members of a user group (e.g., a family) could be involved

in different parts of the booking process, or a single user could employ different devices during the process and might not be logged in to any of these devices during some parts of the process. Then, the correct user profiles of each would need to be determined during each contact in order to correctly identify them and, consequently, the correct A/B test group(s) to which they belonged.

There are a number of ways to address these various issues. For example, doing extra checks with internal and potentially external datasets (which various vendors can provide) could help address the device-mapping issue. Note that data cleaning can have nuances — for example, in the cases of multiple devices, it may be the case that the user deletes their cookies, which is a useful signal in itself that may be easily imputed or predicted. In such cases, where there is missing user information, the best approach is to see if you can use other variables to try and predict the missing information.

Issue #3: Long Time Horizon for Measuring Success

Lastly, successful consumption of a use of Airbnb's services — a happy stay at an Airbnb listing — happens over a much longer time horizon than, for instance, use of social media (where consumer enjoyment of the service is instantaneous). This delay makes it difficult to accurately measure the influence of various features of Airbnb's service through calculation of various success metrics, which can be done only much later. Plus, these are low-frequency events — it's hard to know if there was a statistically significant increase in bookings if the majority of users don't even make a booking in a given month.

As an example, consider measuring longer-term metrics such as user retention or customer lifetime value. Since A/B tests cannot be run for many months, we need to find a shorter-term proxy for such longer-term metrics. A machine learning approach works well here — for example, using various features to predict retention or customer lifetime value, and choosing any of the important features correlated with the target metric that can be measured on a shorter-term basis. Then, on the A/B test, we want to simply see the expected moves on these important features.

Solution #10.4:

At first, it may seem that Google's search market share goes unchallenged. You might think, "If we don't make a deal with Firefox, users would still default to Google on their own, because what are they going to do, *Ask Jeeves*?" However, to answer this question well, it's important to first fully explain the business motivations for why Google wants to remain Firefox's default browser. Based on these business considerations, we can then specify the metrics Google should use to price the deal.

Step 1: Explaining Google's Immediate Motivation to Be Firefox's Default

Google's advertising revenue, most of which comes from search ads, makes up more than 80% of Alphabet's revenue. Clearly there's a lot of money at stake when it comes to search. However, Google's motivations for closing a deal go beyond the immediate profit that's at risk.

For example, Google likely has a business goal of beating competitor search engines and making sure they stay beat. That's because some percentage of users probably don't care about what search engine they use and simply stick with the default one. Plus, Firefox defaulting to a competitor like Bing just might be the spark that Microsoft needs to invest more heavily in challenging Google's monopoly on search. In the case that Firefox defaulted to DuckDuckGo, there would be additional brand ramifications to deal with too. Because Firefox positions itself as a more privacy-aware browser, and DuckDuckGo's brand is built on protecting a searcher's privacy, a vote of confidence by Firefox in DuckDuckGo could bring Google privacy concerns to the forefront, fueling a negative news cycle and ultimately hurting Google's brand reputation.

Google's not just worried about beating direct competitors — vertical search engines such as Amazon, Expedia, and Yelp compete heavily with Google on specialized searches. Google boosts its own ancillary services, like Google Shopping and Google Maps, in search results. That means if Firefox were to default to a competitor, Google Shopping couldn't benefit from its top-of-page position in Google, which would cause them to cede market share to Amazon.

Similarly, if Firefox switched from Google to Bing, searching for a nearby restaurant wouldn't result in a Google Maps listing, but instead, might take a user to a Yelp page. Clearly, a whole ecosystem of products and business lines would be adversely affected if Google isn't able to be the default search engine on Firefox.

Step 2: Explaining Google's Secondary Benefits from Being Firefox's Default

There's also network effects to consider as well. Google's search relevance algorithm takes into account how users interact with their search results. This means the more people who use Google Search, the better the product becomes, creating a positive feedback loop of user engagement and product improvement. Furthermore, these network effects extend to products downstream of Google Search, like Google Reviews and Google Maps. More people searching on Google means more people reviewing businesses on Google, which helps Google Maps compete better against Yelp and Apple Maps in the lucrative local search market.

In considering the whole ecosystem of products that Google Search leads to, along with the multitude of business Google competes with, it's clear Google's motivations to be the default search engine on Firefox go far beyond the immediate search ads revenue at stake.

Step 3: Metrics Used to Inform Google's Willingness to Pay

A lazy way of setting the price is by assuming what Google paid previously to Firefox (~$450 million in 2020) was fair, and adjusting it slightly according to the change in Firefox's install base. A similar approach could be to price this deal relative to the deal Google has with Apple, where it pays ~$10 billion a year to be Safari's default search engine, and scaling the price to Firefox's market share. Depending on the interviewer, these answers can be seen either as clever or missing the point of the problem, so tread carefully!

For an answer based on first principles, you could look at the amount of search ad revenue Google gains from all Firefox users. If this segment completely stopped searching on Google, how much revenue would be lost? This could be one way to price the deal, but it assumes the worst-case scenario: that everyone wouldn't use Google anymore if it wasn't the default. To make the estimate more realistic, you could look at other browsers where Google isn't the default and see what percentage end up using Google. This way, you can get a more reasonable estimate of how much revenue you'd stand to lose, and then determine the price based on this number.

Instead of considering direct search ad revenue, we could also do a similar analysis, except base our bid on the total revenue generated from Firefox users. This number would account for all the downstream ways Google makes revenue from a Firefox user, and thus better account for the second order effects of weakened market share amongst Firefox users.

Solution #10.5:

Before jumping into defining metrics and brainstorming product improvements (like forcing all LinkedIn users to follow linkedin.com/in/nipun-singh/ and linkedin.com/in/kevin-huo/ for career advice), it's best to start by explaining the goal of LinkedIn's Feed.

Step #1: Explain Why LinkedIn's Feed Exists

LinkedIn's mission is to connect the world's professionals to enable them to be more productive and successful in their careers. The newsfeed helps fulfill this mission by helping users keep tabs on their professional network, stay up-to-date with industry news, connect with new people through engaging content, and more.

From a business perspective, newsfeeds (not just at LinkedIn, but other social media companies as well) tend to be very engaging products. For LinkedIn, the feed ensures people keep checking the product often, which is crucial, since without the feed, the product doesn't hold much utility for users not actively job hunting or networking. By helping to keep users on the platform, LinkedIn is able to make more money through displaying ads and sponsored jobs.

Step #2: How LinkedIn Can Measure Feed's Engagement

For a product as expansive and critical as its feed, LinkedIn needs to track a whole host of metrics. A few top-level engagement metrics include daily active users, weekly active users, and monthly active users on Feed. You could also track L7 and L28 (how many days in a week or month do users check the feed). To add a notion of duration to these visits, you can also track average user session time on Feed, and average daily, weekly, and monthly time spent on the feed.

Having users log on frequently and not engage with anything doesn't help LinkedIn's product goal with feed. To measure the depth and quality of engagement, we can track important user actions taken on feed. You could track the number of posts seen, posts liked, posts commented on, and posts shared per month. To make it easier to report on, you could combine all the activity into a single score. However, not all post engagement is equal. By weighting the value of a post *view* vs. post *like* vs. post *comment* differently, you can more accurately capture post engagement activity using just one metric. You could also make a similar metric to measure content creation, which incorporates the number of posts made, comments made, and posts shared.

Because LinkedIn Feed is so closely tied to a significant source of revenue — ads — you should also separately track engagement on ads in Feed. Metrics like ad impressions and ad clicks in Feed could be bundled under the umbrella of feed engagement.

Lastly, for all of these metrics, we want to make sure we are measuring genuine engagement — not the results of automated spam or scraping bots, which inflate activity metrics.

Step #3: Ways to Improve Engagement on LinkedIn's Feed

As mentioned in the intro section on doing your homework before the interview, hopefully you've used the product a bit and had a chance to look at competitors. That can make tackling product brainstorm questions easier since you've scoped out the competition and can "borrow" ideas, much like Facebook and Instagram love to borrow from Snapchat.

To improve these metrics, LinkedIn would want to incentivize people to stay engaged. Example features that boost engagement include personalizing the News Feed to the user and encouraging people to post more to keep the news feed fresh. Specific ways to achieve these goals would include creating up-to-date ranking models that accurately rank how likely a LinkedIn user is to consume and engage with newsfeed content or adding some new, highly requested reaction type. Enabling new post types, like LinkedIn Live streams or LinkedIn Stories can also aid in keeping the content inventory engaging.

Another way to improve metrics would be to build a model using features that you believe would affect the metric. Generally, this will be a combination of user data (demographics) and event data

(browsing behaviors and session events). Decision trees or random forests can be useful here, as they tend to have high accuracy and also can easily display feature importance. After assessing model outputs, you can determine factors contributing to the target metric and decide on an action plan accordingly.

Each of the features suggested above can be A/B tested against core engagement metrics to determine if they would drive increased engagement on the newsfeed. Since other metrics would likely also be affected (for example, with a large increase in time spent and news accessed, more bad content would also be consumed), it is paramount that such A/B testing be evaluated holistically and potential trade-offs kept in mind. Note that, overall, this process implies that improving metrics improves what you intend to measure (i.e., that metrics drive behavior) rather than the other way around, which can have some consequences. As we painfully learned firsthand running A/B tests at Facebook, sometimes the metrics measured don't accurately reflect true user behavior and sentiment.

Solution #10.6:

A good answer would walk through the basics of user assignment and address basic pitfalls the Lyft rider app may face in A/B testing. A great answer would also mention advanced issues that can occur, like network effects, and how doing geo-based randomization can help keep groups balanced. For bonus points, we also explain ways to quantitatively prove that our groups are, in fact, balanced.

Step 1: A/B Testing Basics for Lyft's Rider App

Since we want to quantify whether and to what extent instituting a change in how Lyft operates (in this case, adopting new UI features) would improve a metric of interest (number of rides taken), the most feasible strategy is to conduct an A/B test, in which users are divided into two groups — one exposed to the change, and one not exposed to it.

However, we can't just arbitrarily split Lyft riders into two groups with one having the new UI and the other having the old version; splitting the data haphazardly in that way could, and most likely would, cause the demographics of one group to differ greatly from that of the other. This would introduce a source of variability in the outcome variable not related to the UI change and would most likely skew the distribution of the dependent variable being measured (i.e., number of rides taken). Therefore, we would have to choose A users and B users so as ideally to balance the groups with respect to such user characteristics as demographics, locations, etc. Employing stratified random sampling would provide the best means of ensuring homogeneity of groups.

Step 2: Accounting for Network Effects with Geo-Assignment

We also need to take into account marketplace dynamics. Consider any location...say New York City. If we give half of the riders the new features and keep half of the riders on the old features, then if the new features *do* help people book more rides, there will be more competition for drivers on the new features (and vice versa if the features are detrimental to conducting more rides). In either direction, the resulting effect is exaggerated due to these marketplace dynamics. Therefore, our best option is to test by using comparable markets (comparable meaning the metrics in aggregate should be similar across both markets).

Step 3: Account for Geo-Assignment Flaws

Every method has its drawbacks — geo-based user assignment included. Assuming that two comparable markets are independent may not be accurate in many cases. Additionally, even if user demographics are relatively balanced between the two markets, there is no guarantee that the users

will always stay comparable and that any metric changes tracked by these pools of users will be comparable forever. Lastly, external events may happen that cause the two markets to diverge in some aspect of comparable distributions (regulatory or political change, certain competitors launch marketing campaigns in particular areas, etc.).

To make sure there weren't any geo-based assignment issues, it's best practice to check on a few baseline metrics that aren't supposed to change by market, and validate that they stayed the same so you can have more confidence in the test.

Solution #10.7:

In statistics class, it's beaten over our heads how many phenomena follow a normal distribution. But in the business world, a great many distributions actually follow the Pareto principle, where 80% of outcomes come from 20% of the causes. Just how 80% of the world's income is earned by the top 20% of people, or how many tech companies have found that 80% of crashes come from 20% of bugs, we'd expect the 80/20 rule to be valid here too.

We'd expect many small sellers doing small amounts of revenue, with a long tail of a few power sellers with enormous amounts of revenue. Hence, we'd expect the distribution to be right skewed.

Solution #10.8:

This open-ended discussion on what makes for "bad" content is one that can test your product, business, and even PR savvy. There's no right answer — people debate this question everywhere, from Facebook's headquarters, courtrooms, and even the Senate floor. As long as you're able to brainstorm content removal features well and convey the many nuances of taking down posts, you'll be golden.

Step 1: Brainstorm What Posts Should Be Taken Down

Besides what Facebook is legally obligated to take down (exploitive photos of minors, copyright and IP violations, etc.) other types of content Facebook could potentially take down:

- ***Explicit Content:*** Nudity, sexually suggestive imagery, self-harm, excessive violence
- ***Hate Speech:*** Death threats, posts that incite violence, bullying, doxxing
- ***Misinformation:*** Conspiracy theories, fake news about vaccines or elections
- ***Content from Bad Actors:*** Everything from a terrorist organization or criminal organization
- ***Regulated Goods:*** Posts that promote selling or trading firearms, narcotics, and human organs
- ***Scams:*** Ponzi schemes, fake fundraisers or charities, posts from people who stole someone's identity and are trying to now solicit money

We should also mention that there are other types of posts which could possibly be taken down; in some cases, you could avoid this by tacking on a warning below the post, with a link to verified resources that fact-check or debunk the post. This way, you can reduce harm on the platform while still allowing for freedom of expression.

Step 2: Propose Features to Find Bad Content

In classifying content, features to be considered would include the type of content, the entity posting it (i.e., who posted it), and the context (when/where the post occurred). Here are examples of features demonstrating each of these aspects:

- ***Content:*** Contains inappropriate language (curse words), nudity (in photos), hate speech, or sensitive keywords (e.g., "vaccine," "election fraud").
- ***Entity:*** Posted by a suspected fake account or bot; an entity with a history of posts taken down in the past; an entity with an unverified phone number or email address; an entity connected to other bad actors, etc. Since it is likely that people rarely just commit one act of harm, they may masquerade as various accounts with similar behavior. Thus, it is important to keep track of all the detailed user information (IP, device ID, etc.) to try and triangulate such users.
- ***Context:*** How much spam the group or feed it was posted in has, the amount of "bad actors" within the group or feed posted, etc. Often rings form online where multiple people organize and engage in harmful activity.

We should also work with product operations and manual review teams to understand what types of bad posts they are seeing and the heuristics they use to find these bad posts, as this human intuition can help us generate new features.

Step 3: Explain Trade-Offs of Taking Down Content

As with any kind of classification problem under uncertainty, there's false positives and false negatives. Classifying a post as harmful and taking it down, when in fact it was benign, can confuse and anger users. They might be perplexed as to why their post was removed, which could lead them to post less of that type of content in the future. They can also feel like their voice doesn't matter in the purported community Facebook is building, and might deactivate or cancel their Facebook account in protest of censorship. Sometimes, these news stories even work their way to Capitol Hill, with congressmen and senators calling for regulation or breaking up the purported monopoly because someone they like posted something innocuous and it got taken down (which is perceived as damning evidence of Facebook bias and censorship).

In the case of false negatives, letting harmful content remain on the platform can have many ill effects. People can be confused or harmed (no, drinking hand sanitizer won't clean the COVID out of your system!). People who don't even see the original misinformation can be affected. Almost always, when something harmful goes viral, there are negative news stories, like "*5 million people saw fake news on Facebook claiming that Drake is NOT the best rapper of the 2010s*" which can cause a PR shitshow for the company. Long term, it can even make the company seem complicit with the misinformation that spreads.

As such, for different bad content types, there can be different sensitivities used, depending on the downside risk of having a false positive or a false negative. Additionally, we will want to tweak the algorithm for sensitive accounts. For example, if a political figure with a large following posted something questionable but allowed, but by accident the algorithm flags it and it gets taken down, Facebook gets tarred and feathered in the press for censorship. For sensitive accounts like news agencies, political figures, or governmental agencies, there likely should be a human in the loop to improve accuracy.

Solution #10.9:

Two words, one equation: Correlation ! = Causation

Just because more complete author profiles correlate with increased sales doesn't mean it's the cause. Maybe books with more complete author profiles had a more highly reputed publisher fill it out for them, which means they likely also have better designed book covers, and that's why they have better sales.

will always stay comparable and that any metric changes tracked by these pools of users will be comparable forever. Lastly, external events may happen that cause the two markets to diverge in some aspect of comparable distributions (regulatory or political change, certain competitors launch marketing campaigns in particular areas, etc.).

To make sure there weren't any geo-based assignment issues, it's best practice to check on a few baseline metrics that aren't supposed to change by market, and validate that they stayed the same so you can have more confidence in the test.

Solution #10.7:

In statistics class, it's beaten over our heads how many phenomena follow a normal distribution. But in the business world, a great many distributions actually follow the Pareto principle, where 80% of outcomes come from 20% of the causes. Just how 80% of the world's income is earned by the top 20% of people, or how many tech companies have found that 80% of crashes come from 20% of bugs, we'd expect the 80/20 rule to be valid here too.

We'd expect many small sellers doing small amounts of revenue, with a long tail of a few power sellers with enormous amounts of revenue. Hence, we'd expect the distribution to be right skewed.

Solution #10.8:

This open-ended discussion on what makes for "bad" content is one that can test your product, business, and even PR savvy. There's no right answer — people debate this question everywhere, from Facebook's headquarters, courtrooms, and even the Senate floor. As long as you're able to brainstorm content removal features well and convey the many nuances of taking down posts, you'll be golden.

Step 1: Brainstorm What Posts Should Be Taken Down

Besides what Facebook is legally obligated to take down (exploitive photos of minors, copyright and IP violations, etc.) other types of content Facebook could potentially take down:

- ***Explicit Content:*** Nudity, sexually suggestive imagery, self-harm, excessive violence
- ***Hate Speech:*** Death threats, posts that incite violence, bullying, doxxing
- ***Misinformation:*** Conspiracy theories, fake news about vaccines or elections
- ***Content from Bad Actors:*** Everything from a terrorist organization or criminal organization
- ***Regulated Goods:*** Posts that promote selling or trading firearms, narcotics, and human organs
- ***Scams:*** Ponzi schemes, fake fundraisers or charities, posts from people who stole someone's identity and are trying to now solicit money

We should also mention that there are other types of posts which could possibly be taken down; in some cases, you could avoid this by tacking on a warning below the post, with a link to verified resources that fact-check or debunk the post. This way, you can reduce harm on the platform while still allowing for freedom of expression.

Step 2: Propose Features to Find Bad Content

In classifying content, features to be considered would include the type of content, the entity posting it (i.e., who posted it), and the context (when/where the post occurred). Here are examples of features demonstrating each of these aspects:

- ***Content:*** Contains inappropriate language (curse words), nudity (in photos), hate speech, or sensitive keywords (e.g., "vaccine," "election fraud").
- ***Entity:*** Posted by a suspected fake account or bot; an entity with a history of posts taken down in the past; an entity with an unverified phone number or email address; an entity connected to other bad actors, etc. Since it is likely that people rarely just commit one act of harm, they may masquerade as various accounts with similar behavior. Thus, it is important to keep track of all the detailed user information (IP, device ID, etc.) to try and triangulate such users.
- ***Context:*** How much spam the group or feed it was posted in has, the amount of "bad actors" within the group or feed posted, etc. Often rings form online where multiple people organize and engage in harmful activity.

We should also work with product operations and manual review teams to understand what types of bad posts they are seeing and the heuristics they use to find these bad posts, as this human intuition can help us generate new features.

Step 3: Explain Trade-Offs of Taking Down Content

As with any kind of classification problem under uncertainty, there's false positives and false negatives. Classifying a post as harmful and taking it down, when in fact it was benign, can confuse and anger users. They might be perplexed as to why their post was removed, which could lead them to post less of that type of content in the future. They can also feel like their voice doesn't matter in the purported community Facebook is building, and might deactivate or cancel their Facebook account in protest of censorship. Sometimes, these news stories even work their way to Capitol Hill, with congressmen and senators calling for regulation or breaking up the purported monopoly because someone they like posted something innocuous and it got taken down (which is perceived as damning evidence of Facebook bias and censorship).

In the case of false negatives, letting harmful content remain on the platform can have many ill effects. People can be confused or harmed (no, drinking hand sanitizer won't clean the COVID out of your system!). People who don't even see the original misinformation can be affected. Almost always, when something harmful goes viral, there are negative news stories, like "*5 million people saw fake news on Facebook claiming that Drake is NOT the best rapper of the 2010s*" which can cause a PR shitshow for the company. Long term, it can even make the company seem complicit with the misinformation that spreads.

As such, for different bad content types, there can be different sensitivities used, depending on the downside risk of having a false positive or a false negative. Additionally, we will want to tweak the algorithm for sensitive accounts. For example, if a political figure with a large following posted something questionable but allowed, but by accident the algorithm flags it and it gets taken down, Facebook gets tarred and feathered in the press for censorship. For sensitive accounts like news agencies, political figures, or governmental agencies, there likely should be a human in the loop to improve accuracy.

Solution #10.9:

Two words, one equation: Correlation ! = Causation

Just because more complete author profiles correlate with increased sales doesn't mean it's the cause. Maybe books with more complete author profiles had a more highly reputed publisher fill it out for them, which means they likely also have better designed book covers, and that's why they have better sales.

Solution #10.10:

This is your classic metrics troubleshooting problem. Please raise your right hand, and repeat after me: *I do solemnly swear to stick to the 4-step framework for diagnosing metric changes.*

Step 1: Clarify the Scope of the Metric Change

Always start by asking clarifying questions about the metric in question. What is meant by "active users"? Is this a decrease in daily log-ins, or is it a decrease in usage of a specific feature? Did the metric suddenly drop by 5% one day, or was there a gradual decrease in active users over the week?

Also, is 5% a lot? What are we comparing this drop against — the daily active users this time last week, last month, or last year? How much does this metric typically fluctuate — maybe a 5% drop for a week isn't large enough or sustained long enough to be a cause for alarm. Or maybe there is seasonality at play — a 5% drop wouldn't be too surprising if this was the week after Christmas and New Year's.

For the sake of our solution, we'll assume that the interviewer tells us this was a decrease in the number of times logged-in users have opened the app, and that there is no seasonality at play. We'll assume the interviewer says the 5% drop is relative to last week, and that on a weekly basis, the daily active user count tends to be consistent, which is why stakeholders are so worried.

Step 2: Hypothesize Contributing Factors

Here are some potential reasons behind the 5% drop:

- Logging Issues: Data pipelines responsible for logging daily active users broke somehow, which makes it seem like a genuine drop, but it actually isn't.
- Upstream Issues: There could be a problem upstream of daily logins, like a bug in keeping users logged in or a decrease in push notifications being sent, which is having an effect on app opens.
- Product Changes: Did we change something inside the product, like how snap streaks work, or extend the expiration date of snaps, which is causing users to check the app less frequently?
- External Events: Has some large market experienced a hardship, like a natural disaster or an internet shutdown, which is causing users to not use the app as much?

Step 3: Validate Each Factor

For each of these factors, we can validate our hypothesis by looking at various metrics and talking to teammates:

- Logging Issues: We can check in with the SRE (site reliability engineering) team, data engineering team, and metrics and logging teams, to make sure pipelines are healthy and the drop is genuine.
- Upstream Issues: How is push notification volume looking? How are login and password recovery numbers looking — anything out of the ordinary? Did uninstalls spike in the last week?
- Product Changes: How many snaps were sent, and what was the average open rate? What about the number of messages sent between users and the number of stories posted and viewed — are these down by 5% too, or much more? If it's a more drastic drop, maybe something within the app broke that is causing users to not check the app as much. You can also directly check for product quality issues by seeing if bug reports or app crashes spiked within the last week.
- External event: Can you segment by market, language, and OS to check to see if this problem is local to any one subgroup, and then research into that area for changes that occurred in the last week?

Step 4: Classify Each Factor

Imagine after going through the above factors, your interviewer tells you that logging is working fine, and that this drop shows up across all markets, all ages, all genders, and all types of users (both new users and tenured users). However, you learn that the drop is 7% for iOS users but only 1% for Android users. Now we're getting somewhere!

You could bucket the iOS users by the app release number they are using, what carrier they are using, and what model of iPhone they are using. Say your interviewer tells you that people running the latest version of the app have a 20% lower chance of being a daily active user compared to the baseline iOS user. From this, you look at other upstream metrics for users on the latest app and compare them to the general iOS population. The interviewer then tells you that user logins and password resets both spiked the day the new app became available.

Now we've found our likely culprit — there must be a bug when upgrading to the newest app release that's accidentally logging out users, who are then forced to log back in or reset their password. Some subset of people probably can't recover their password, and thus drop off from being a DAP.

Solution #10.11:

We first clarify what the new search-ranking algorithm change is, then connect how this algorithm change relates to Pinterest's product and business goals. Once you have done this, you can suggest concrete metrics to measure the impact of the change.

Step 1: Clarify the Product Change

It's important to clarify the scope of the change. Questions to ask include:

- Did this algorithm change have any high-level goals in mind (e.g., prioritizing trending Pins, improving discoverability of niche Pins, increasing the personalization of search results)?
- Did this change involve any UX or UI changes perceptible to the end user, or was it solely a change on the backend?
- Do search results return just as fast for users as before?

Step 2: Explain Why Search Relevance Is Important for Pinterest

Pinterest is a visual discovery engine built for helping users find inspiration for their lives, from fashion to home decor to recipes and more. With billions of Pins on Pinterest, the ability to search for and find the content that sparks a user's imagination is key. At a higher level, to keep Pinterest's product competitive against similar content discovery platforms like Instagram Explore or Houzz, the search, content recommendation, and user personalization algorithms need to be top-notch.

When the discovery engine is working well, users will stay engaged with Pinterest and keep coming back for more. This user engagement and retention is key for Pinterest's ad-supported business model. Plus, with shoppable product pins — Pinterest's push to diversify away from ad revenue and into e-commerce — the ability for users to search for products and find exactly what they were seeking is even more critical to the business than before.

Step 3: Propose Metrics to Quantify a Search Algorithm Change

We could measure the direct amount of engagement the search functionality received. For example, time spent searching or the median number of searches made per user session could be used. While growth in these metrics could be a sign that users like the new search experience, it isn't definitive

proof of a successful algorithm change. For instance, if our search results weren't relevant, a user might perform multiple searches to find what they were looking for. Here, an increase in time spent and searches made actually indicates user frustration.

Thus, a more complete way to capture the impact of the change would be to also look at the downstream effects of a relevant search algorithm. For example, we could measure how often a search leads to a user pinning a search result to their board — a sign that the user found what they were looking for. To go one step further, we can quantify the direct monetary benefits of improved search. Here you could measure the revenue generated from purchases of buyable pins that came up as a search result.

For bonus points, you could also mention evaluation measures for information retrieval systems. If you take a binary approach to the results — for example, is this pin relevant or not — you can use a metric like "precision@10" to understand the percentage of relevant pins amongst the first 10 results. However, this won't take into account each pin's position within search results. If you want to account for the actual degree of relevance for each pin, a metric like normalized discounted cumulative gain (nDCG) measures how close the results are to the best possible result. The technical details of this measure are beyond the scope of this text.

Solution #10.12:

This product interview problem is a hybrid between a root-cause analysis question and a defining success metrics question. We'll first start by connecting the supply vs. demand question to Netflix's bigger business model and product goal. Then, we'll define some metrics we'd want to analyze in order to troubleshoot the root cause of the content supply or demand problem.

Step 1: Why Netflix Cares About Content Supply & Demand

Netflix's mission is to entertain the world. Netflix Studios not only has the power to produce original content, it can directly influence what entertainment hundreds of millions consume.

By deeply understanding what its users want to watch, thanks to the myriad of analytics Netflix collects, Netflix can greenlight new shows that delight niche audiences. Let's face it: you'd never see a show like *Indian Matchmaking* or *Orange is the New Black* on cable TV. As long as Netflix creates high-quality tailored content, its users will continue to engage and retain on the service rather than churn out due to stale inventory, or switch to competitors like Disney+ or Amazon Prime Video.

However, creating original shows is very expensive. You can't just *blindly* do things, like Sandra Bullock in *Bird Box*. Netflix needs to prioritize what types of shows to produce so that every dollar spent brings back many more in the form of increased customer retention and NPS. As such, it's crucial to vet how much demand there is for a show category before investing resources into producing Netflix originals or licensing more media from other studios.

Step 2: Content Supply vs. Demand Metrics to Investigate

Knowing that there is less total viewer watch time devoted to sci-fi TV shows compared to other similar categories doesn't tell you much. Sure, total watch time for sci-fi might be less compared to other categories, but what about on a per-show basis? Maybe there's just fewer available sci-fi TV shows, so even if people like the show, there just isn't enough inventory to support high total watch times compared to a category like comedy or drama, which has many more shows to watch. Knowing that even though the total watch time is low, the watch time per show is high would indicate a supply problem, not a lack-of-interest problem. This would signal that Netflix should create or buy more sci-fi TV shows — *not* drop the few sci-fi shows they do have.

Another way to get a clue on whether it's a user interest or show quality problem is by looking at metrics related to sci-fi TV show recommendations. Are people browsing for titles in this category but just not hitting play? This could indicate there is demand, but nothing catches a user's eye. One step down the funnel, what's the conversion rate between watching the first episode of a sci-fi series and finishing the first season? How does it compare to other categories? Maybe sci-fi TV shows just aren't bingeable. Or maybe the show is low quality (ok, maybe *terrible* quality), and people can't stand to finish a season, let alone a whole series (*Sense8*). To more directly measure supply quality, you can also look at user ratings on sci-fi TV shows. Do they tend to be much lower than in other categories?

When looking at all these metrics, we should segment by user attributes, like viewer country or language. Netflix is a global platform, and it's not fair to expect users to act as a monolith. Maybe sci-fi watchtime, relative to other categories, is okay in English-speaking markets, but more effort needs to be put into closed captioning and voice dubbing to serve other countries.

Bonus Points: Zooming Out

A good answer should also consider that it might not even be an actual supply or demand issue. Maybe, there's an issue at a step higher in the funnel, like sci-fi shows that, for some reason, don't tend to make it into Netflix show recommendations. Or maybe more mainstream shows get most of the advertising budget. In both cases, there is demand for sci-fi, and there is content available to meet their needs, but discovery is broken so people aren't aware of any sci-fi show besides *Stranger Things*.

We can also look outside of Netflix for an answer. Through consumer surveys or audience insights data from a company like Nielsen, we can benchmark engagement in sci-fi shows against broader interest in the category. This way, we could tell if there is generally less demand for sci-fi content, both on and off Netflix, or if Netflix in particular is underperforming in this segment relative to competitors and cable.

Solution #10.13:

Before we jump into the technical details of performing customer segmentation, it's important to flex your business muscles and explain what customer segmentation analysis is and brainstorm a few concrete ways the analysis results could boost store sales.

Step 1: Explain How Apple Benefits from Customer Segmentation

Each person that walks into an Apple store has individual needs, desires, and preferences. In an ideal world for Apple, they would be able to hyper-target their product offerings and store design to cater to a single person's needs, one at a time. Obviously, this isn't feasible. On the other hand, treating all Apple customers the same misses the variety of customer needs. As such, by grouping similar customers and creating customer segments, Apple can customize its in-store sales strategy to large groups of customers at once.

However, our approach does have a caveat: by relying on historical data for customer segmentation, we aren't able to analyze non-Apple customers since they wouldn't show up in prior store sales data. Therefore, it's important to let the interviewer know that a customer segmentation analysis should be complemented with some competitor research or market-level analysis.

Step 2: Brainstorm Ways Customer Segmentation Can Boost Sales

There's many different ways to segment users, and then use those insights to boost sales. For example, one axis you could segment users on is their tech savviness. This information could impact the different kinds of sales scripts Apple uses. For example, a store salesperson convincing a software

developer to buy a MacBook would use a very different pitch than when explaining the benefits of the product to a nontechnical person. It could also impact the store staffing — maybe each store should have a technical expert to field the toughest questions.

Another dimension on which you could segment customers is by the main type of product they bought. Say, for example, our analysis found three main types of people routinely walk into an Apple store: iPhone purchasers, MacBook purchasers, and Apple Genius Bar customers who pay for an issue to be fixed. By separating customers into three groups, we learn that iPhone purchasers are five times as likely to buy AirPods than MacBook purchasers or customers coming in to get their device fixed by the Apple Geniuses.

This insight from customer segmentation could mean that it's best to place the AirPods next to the iPhones, to increase the chance of a cross-sell. Another implication of this insight would be to train salespeople to upsell AirPods to customers who are about to buy iPhones, but not waste their time on the upsell for MacBook shoppers. Another idea is creating a new discounted iPhone and AirPods bundle to entice customers into buying both products at once.

Step 3: Explain How to Perform Customer Segmentation

To perform customer segmentation, we could use K-means clustering. We could visualize the data or do hyperparameter tuning to find the appropriate number of clusters to segment the users into. Besides running K-means on the transactions data, we could also try to connect online sales data to in-store customers. While this analysis is primarily for understanding brick and mortar shoppers, cross-referencing in-person customers with their potential online purchases on Apple.com could help give a more complete picture of the customer.

Solution #10.14.

The first step would be to gather the basic data on both the iOS and Android users for both Facebook and Instagram. You could analyze user demographics such as age, gender, race, and location. You could also analyze user activity, looking at metrics such as time spent overall, and time spent on various activities (feed, in-app messaging, etc.) for both groups of users on both apps.

We can visualize the user activity metrics by each cut of user demographics to get a top-level understanding of where any differences may lie. For example, iOS users may, on average, spend much more time on the Facebook ecosystem than Android users do, and this "top-of-funnel" reason may lead them to use Instagram more also. Alternatively, iOS and Android users may, in general, be from different age groups, and this could be affecting their respective levels of Instagram usage, as Instagram isn't as widely used by older people.

Another set of factors to consider would be the actual Instagram's device and resource requirements, relative to Facebook's requirements. Maybe iOS devices have a much easier time downloading the Instagram app, since the app size is smaller for iOS than Android. Maybe the Instagram app only works on devices that have updated their OS within the last two years, and Apple devices tend to run the latest OS much more than Android devices.

Maybe the problem lies with the actual app experience — do Facebook and Instagram perform the same way on both platforms? What do app store ratings, number of bug reports, feed scroll latency, and percentage of sessions with app crash for both apps on both platforms look like? Maybe Facebook works equally well on both platforms, but Instagram has under-invested in its Android app optimization.

Finally, for a difference so big, across so many users, I'd make sure to talk with user experience researchers, folks from the product strategy teams, and the Android and iOS leads for both Facebook

and Instagram. While cutting the data may reveal the underlying reason, our gut intuition is that for such a large difference across so many users, there's likely a bigger structural or strategic cause for the disparity that a purely SQL-based analysis may not uncover. At the very least, by talking to other domain experts, we can add some more color to our analysis.

Solution #10.15.

Just like with every metrics question, a good answer should start out with a brief discussion of the business goal — in this case, into the goal of Capital One's credit card. It should also mention a few of the stakeholders involved with this business goal before determining the best metrics for measuring retention.

Step 1: Explaining Capital One's Motivation Behind Credit Card Retention

A credit card's "stickiness" is the frequency and duration of its use by its holders. The bank issuing a card is motivated to encourage a card holder to use the card frequently and over the long term — in other words, to increase its "stickiness," as this earns it a greater profit in interest (if a holder carries a balance) and directly on transactions. A card holder would prefer a non-cash option having low monthly repayments, a low interest rate, and, perhaps, rewards for using the card. The goal of a bank's reward system (cash back or other perks associated with card use) is to increase card usage and its customers' reliance on the card over the long term. Thus, if Capital One gave no incentives to cardholders to encourage them to use their card to pay for purchases, then users would most likely use their cards less frequently or, possibly, not use them at all — hence, a decrease in stickiness.

Step 2: Brainstorming "Stickiness" Metrics

For Capital One to assess the stickiness of its Quicksilver card, some potential metrics it might use are as follows:

Daily active users to monthly active users: Although this ratio can be over any interval that you deem appropriate, the goal of this metric is to see what percent of active users during a longer interval (in this case, a month) are active over a shorter interval (daily). A ratio of, for instance, 0.7 would suggest that 70% of cardholders who spend with the card on a monthly basis also do so on a daily basis. The higher this metric is, the stickier the product.

Month-over-month retention: If you were to create cohorts of people based on when they signed up for the card, you could determine what percentage of the cohort churns from any given time interval. In this case, a month-to-month time interval seems reasonable, as credit cards are billed monthly. Seeing what percentage of cardholders remain after X months enables you to see trends in duration of use before a customer leaves or the average duration of time a customer remains with the card before leaving. However, a card holder can also be inactive without actually closing the account (a "silent churn"). Many cardholders typically use various cards for one particular purpose or activity (e.g., one card for travel, one for dining, and so on). Therefore, attempting to track such behavior as well is advisable.

Transaction volume churn: Tracking the total amount spent by cardholders is critical, since this correlates to the amount of revenue Capital One could make. By looking at all users who spent money last month, and their total transaction volume, and comparing it to the total transaction volume of those same users this month, you can see if adoption is growing or shrinking.

Solution #10.16:

One good question to ask the interviewer before answering is "What's the goal of pricing?" For a new service, it could be okay to run the feature without profit if you believe you can aggressively gain market share and turn on monetization later. Assuming we don't want to run a free service, we can price YouTube Premium using cost-plus pricing, value-based pricing, or competitor-based pricing.

Cost-Plus Pricing

For a cost-plus pricing approach, we'd look at how much it costs to provide YouTube Premium in that country and then add a margin on top of that number to enable the business to earn a profit. We'd account for product localization costs, marketing costs to advertise to the new geography, and bandwidth costs to serve content. While not technically a cost, because there's no ads in this feature, you could also account for the lost ad revenue per user that YouTube would've earned.

Finally, because content licensing can be a bulk part of music-streaming service costs, and oftentimes media is licensed on a per country or per region basis, it's important to understand how expensive offering our music library would be. By adding up these costs and then tacking on a premium to this number, we've got one way to price the product.

Value-Based Pricing

A value-based pricing strategy would price the service relative to the amount of value a consumer perceives from using the service. The most direct way to gauge the perceived value of YouTube Premium would be to ask users themselves through consumer surveys and focus groups.

An alternative could be to see what an optimal price was in other countries where this product launched, and assume that price hits the optimal value offered. Then, adjust the price to the local market based on the market's average per capita income.

Competitor-Based Pricing

You could price the service based on competitor video and music streaming services like Netflix, Hulu, Spotify, and Apple Music. In each new country where YouTube Premium is being launched, by seeing how much more valuable (or less valuable!) our service is, we can adjust the pricing against the competitors. All else being equal, the larger the number of alternative options for a country's residents, the more likely that YouTube Premium's pricing would need to be discounted, especially given the relatively low switching costs available to consumers.

Wrapping It Up

Likely, an appropriate price for YouTube Premium would be a blend of all three pricing strategies. To triangulate on an exact answer, you'd want to consult with stakeholders like the sales, marketing, and finance teams. You could also always try A/B testing the prices, or running discounts or tiered memberships to get more signal into what an optimal price point may be.

Solution #10.17:

For questions dealing with whether a company should launch a certain feature or not, it's best to not prematurely discuss the proposed idea's merits. Instead, clarify with the interviewer what the feature actually is, what the company's hypothesis is behind proposing such a feature, and how it would impact key business metrics. Great answers would elaborate on the potential pitfalls of shipping the feature and, lastly, end with a final recommendation.

Step 1: Clarify Twitter's Product Hypothesis

Currently, users on Twitter can only like, comment, or retweet a tweet. Emoji-style reactions are likely being considered as a way to improve the engagement rate of Tweets. The hypothesis is that by reducing the friction for expressing more complex emotions like "this tweet was funny" or "this tweet made me mad," users will be more likely to engage with tweets.

This feature also addresses a common user issue: hitting "like" on sad news feels weird. By providing a more nuanced alternative to liking a tweet, there could be more engagement on tweets overall.

Step 2: Explain Twitter's Business Goal

If emoji-style reactions lead to more tweet engagement, there would be many positive downstream effects to Twitter's business. Firstly, users would be more engaged with their feed and interacting with more pieces of content, leading to longer and likely more meaningful sessions.

It's also not just about the amount of engagement given — it's about the positive impact on the receiving end too. People whose tweets get more engagement will get more notifications of people interacting with their post, driving them to check Twitter more often.

When tweeters notice their posts getting more engagement, that dopamine hit will surely incentivize them to post more on Twitter. This will improve tweet creation metrics, which would boost the amount of interesting content found on the timeline for all users. This positive flywheel of more engagement leading to more content leading to further engagement would improve both time spent on Twitter and user retention rates. More time spent and more user retention means more ads seen, which is crucial for Twitter's ad-supported business model.

Step 3: Identify Data to Support Product Hypothesis

We want to find some data that can guide us on whether emoji-style reactions are desired by users. The best way to do this would be to analyze current user activity and see if there is some unmet latent demand. If we have evidence that users are taking multiple steps to express the common reactions like "love" and "haha," then simplifying the process for expressing these feelings via emoji reactions would be a logical next step.

In terms of specific user data on Twitter that we can use as a proxy for demand for the reactions feature, we could look at current Twitter threads and analyze how often sentiments corresponding to the proposed reactions are expressed. Take the "haha" reaction, for example. If we see many short comments laughing at a particular thread or reply, using keywords like "lol" and "haha" and "lmao," we'd have a strong indicator that users want to express that they found the tweet funny. In that case, offering them an easier way to express that emotion would be beneficial.

We could also work with user research teams to conduct surveys on groups of users to confirm the data: do these people desire an expanded set of reactions? If so, for which reactions, and why? The user research and the data should collaboratively point to the same direction (that people want the ability to express more reactions).

Step 4: Counter-Argument for Shipping Reactions

Counter-arguments for emoji-style reactions are that it increases the complexity of Twitter. Right now, having a single reaction makes things intuitive. Another issue is that more complex emotions are expressed in comments. If we made reacting easier, we'd likely have fewer people replying to a tweet with messages like "this is so funny" or "i hate this." Plus, from an aesthetic and brand viewpoint, the Facebook-ification of the Twitter product may not be desirable.

There are also several challenges from the product and engineering side. For example, which reactions should be used? It may be redundant or confusing to have "love, "enjoy," and "like." reactions. Also, do we have the engineering resources to dedicate to build and test this nontrivial feature? It's not a given that the potential engagement boost warrants the time and money needed to launch the product. Has opportunity sizing been done to show the potential ROI on this project?

Lastly, all engagement isn't equal. Before running this kind of experiment, you'd want to align with stakeholders how important replies are versus reactions. By anticipating the likely trade-off that will need to be made if this feature is successful (overall reactions would be up, but number of comments would be down) and discussing this issue with stakeholders, you'd mitigate launch blockers and have an easier path to shipping it after the A/B test results come back in.

Step 4: Make Final Recommendation

If we have alignment that the goal is to increase engagement, have found good reason that the feature demand is legitimate, have intuition on why implementing the feature as described would drive engagement, and think the business benefits outweigh the development cost and time required to A/B test the feature, then and only then does it makes sense for Twitter to test reactions.

As mentioned in the counter-argument section, after we have A/B test results, we should align with stakeholders about the likely metric trade-off that will occur — increased reactions but decreased comments — before making the final launch decision.

Solution #10.18:

Step 1: Stakeholders for Slack Engagement

Slack's user engagement is important because it aims to be the go-to work and productivity tool. Since Slack serves as a place where people collaborate together, user engagement is critical to monitor and measure. A user in this context is simply any person who has an account on the platform. User engagement affects the business directly since Slack operates as a subscription-based model, where users pay per month for features on the platform. With more consistent user engagement, there is likely to be longer-term retention and new customers over time.

Step 2: Defining User Engagement Metrics for Slack

To measure user engagement for Slack, we could look at DAUs (daily active users), WAUs (weekly active users), and MAUs (monthly active users). An active Slack user is defined simply as anyone who signs into Slack on a given day. Since a product like Slack is meant to be used daily, tracking DAUs is crucial, and the ratios DAUs/WAUs and DAUs/MAUs are typically used, since a higher value of one of these ratios would mean the product was more sticky. Note that we would need to be aware of weekends and holidays (seasonality).

Since the product is a collaboration software, another core metric for engagement would be the number of messages sent. Again, as with the number of active users, we would want to look at messages sent per day, per week, and per month. Other auxiliary metrics we could track include the following: creating organizations, applying for membership in organizations created by another user, etc. However, the two primary metrics would be number of active users (i.e., those who sign in) and number of messages sent.

Additionally, measuring these trends at the cohort level would be important to ensure consistency over time. For example, DAUs could be dropping slightly, but the behavior among various cohorts

could be very different. In this case, identifying the cause(s) of why specific cohorts are seeing drops of greater significance than others would be important.

Step 3: Defining Leading Indicators for Engagement Decline

To receive an early warning of declining user engagement, we could look at trends over time in numbers of both DAUs and daily messages sent, as declines in either or both of these could be leading indicators of an overall decline in users. For example, a user who eventually leaves might initially be a DAU, but then slowly become a WAU , and, finally, become a MAU only. Such a lack of engagement would show up in DAUs and in numbers of messages sent, and, hence, tracking these would be important, as would calculating and tracking the ratios DAUs/WAUs and DAUs/MAUs. Again, looking at these metrics on a cohort level is also important.