Humans Can Be Cranky and Data Is Naive: Using Subjective Evidence to Drive Automated Decisions at Airbnb

STEPHANIE CARTER
Airbnb

RICHARD DEAR
Airbnb

How can we build fairness into automated systems, and what evidence is needed to do so? Recently, Airbnb grappled with this question to brainstorm ways to re-envision the way hosts review guests who stay with them. Reviews are key to how Airbnb builds trust between strangers. In 2018 we started to think about new ways to leverage host reviews for decision making at scale, such as identifying exceptional guests for a potential loyalty program or notifying guests that need to be warned about poor behavior. The challenge is that the evidence available to use for automated decisions, star ratings and reviews left by hosts, are inherently subjective and sensitive to the cross-cultural contexts in which they were created. This case study explores how the collaboration between research and data science revealed that the underlying constraint for Airbnb to leverage subjective evidence is a fundamental difference between 'public' and 'private' feedback. The outcome of this integrated, cross-disciplinary approach was a proposed re-envisioned review flow that clearly separates public and private-to-Airbnb feedback with a single binary question. If implemented, it should allow Airbnb to collect additional evidence from hosts that can be utilized to make automatic decisions about whether guests need warnings or whether they have met an exceptional quality bar for a potential loyalty program.

SETTING

"Would you recommend this guest to other hosts? Describe your experience."

These are the first two questions of the review flow for hosts that has existed on Airbnb since January 2011. At that time, Airbnb had only been around 3 years and had around thirty thousand listings and under a hundred thousand guests who had stayed at Airbnbs. A review system was designed for the purpose of highlighting issues that occurred during stays and establishing a fair dynamic between hosts and guests when they review each other. Every time someone stays on Airbnb, the host of the place reviews the guest and the guest reviews the host. That information is aggregated and displayed for all future potential guests and hosts to see; together, it forms Airbnb's reputation system.

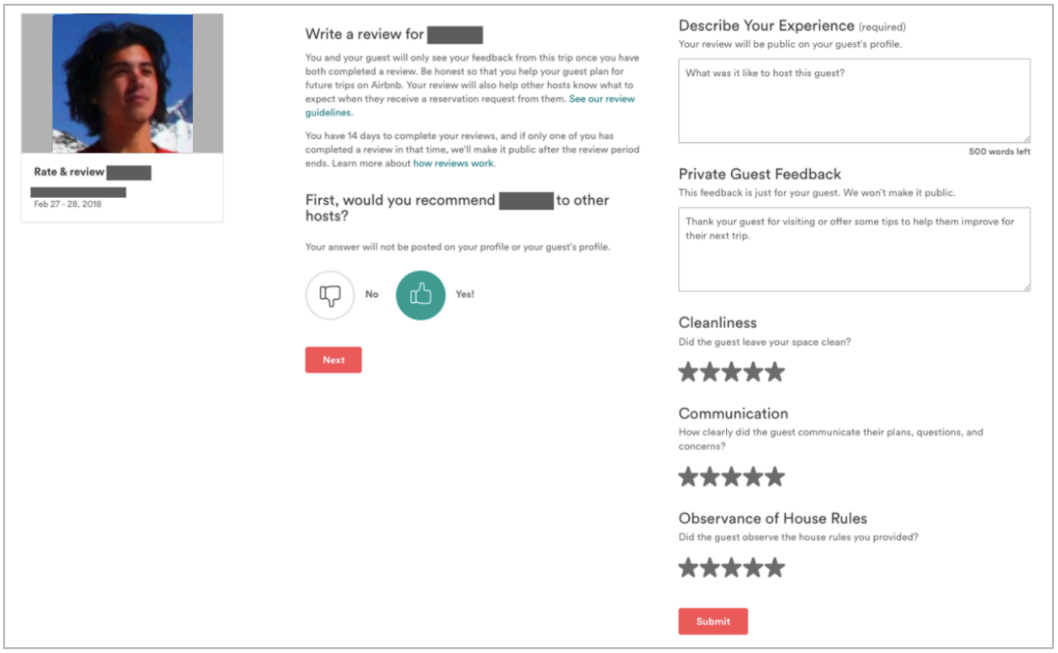

Figure 1. Existing review flow for a host to review a guest.

Nearly eight years later, Airbnb has grown into a community that generates millions of stays each year. People can also book hosted experiences around the world. Airbnb now has over 5 million listings worldwide, in more than 191 countries, across 81,000 cities. There are over 400 million guests who have stayed in Airbnb's apartments, villas, B&Bs, treehouses and many other types of inventory. There were 3 million people who stayed in Airbnbs the night of New Year's Eve (2017) alone. Each day, over a hundred thousand guests get reviewed on Airbnb.

At this scale and maturity, there are two very interesting business challenges related to the guest community.

1. How might we automatically identify which guests need to be warned about poor behavior? Airbnb has a high standard for the quality of its guest community. Yet, in 100 million stays, even a tiny (<0.1%) rate of poor guest behavior becomes a hit to our community and we take that very seriously. We want to do whatever we can to prevent guests and hosts from having less than perfect experiences. Manually investigating potential issues raised in reviews to identify guests that should receive a warning becomes a heavy operational cost.

2. How might we make fair and automated decisions of which guests would qualify as 'exceptional' for a guest loyalty program? Airbnb is exploring the idea of a guest loyalty program and identifying truly exceptional guests (according to the hosts they've stayed with) would be a valuable component of this program. In a community of over 500 million guests arrivals all time, we are at a scale where we need to be able to make these decisions in an automated fashion.

These two business challenges both rely on the evidence of guest behavior that hosts provide in the review flow after a guest stays with them.

However, the idea of using reviews as evidence for loyalty program rewards or to issue guests warnings presents unique challenges. Hosts and guests may perceive the same events differently, so making a decision using host reviews means letting algorithms make automatic judgements based on evidence that is inherently subjective.

Our mental model for how to differentiate guests was limited. Imagine the plight of a traveler if she lost her status at a hotel chain because of a clash in personality with the front desk receptionist, or wasn't able to stay at that hotel again because of a mismatch in communication style. At Airbnb, this challenge is multiplied by the cross-cultural and cross-language nature of the interactions between our hosts and guests, and our own biases as English-speaking Americans designing systems for global interactions.

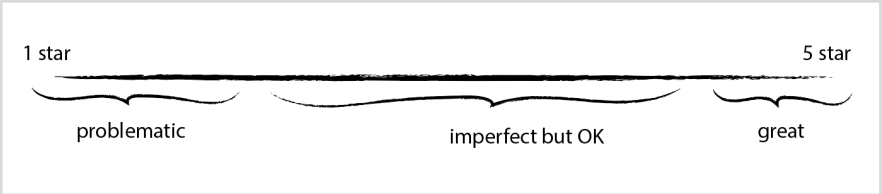

Figure 2. The original mental model for how we might be able to differentiate guests using ratings & reviews.

ACT I

To use subjective, human evidence of host reviews in potential automated decision-making systems, we knew we'd need holistic research across multiple disciplines (data science, research, etc.). Our first research question was a natural one: how well does the current review system work?

In technology companies, research and data science both typically have a few standard approaches for beginning investigation into a new problem space: in this case, it made sense for research to begin with interviewing hosts and guests as well as reviewing inbound feedback to understand the nuances of the experience of the current review system, and data science to follow these learnings with an opportunity analysis to estimate the scale of any potential user problems.

1. Research Methods: 1:1 interviews & review of in-bound feedback

Research began with one of the standard approaches to understanding a complex problem space: guiding open-ended discussions on the topic with groups of stakeholders (in this case, both guests and hosts) while also looking at pre-existing in-bound feedback about the review flow (in this case, reports submitted through a pop-up widget on the review flow).

We learned that hosts were uncertain about the 'right' way to review guests; they were applying different norms to how they shared feedback. For example, one host said, "If I feel disrespected, if rules weren't followed, I'll share everything publicly. It's my home." Yet

another host said, "Recently someone kind of conned me, I didn't review him because I didn't want him to review me, I felt blackmailed." Sometimes the uncertainty of the 'right' way to review became such a barrier that they did not leave any review at all which left Airbnb with an unreliable, incomplete picture of people's experiences.

We also heard from hosts that providing feedback could feel repetitive and there was a desire to reduce the cognitive overhead of free text reviews. As one host said, "It's too time consuming and confusing. I see the same things again and again. If I could just select those things it would save me time and effort." We saw an opportunity to introduce structured options to simplify the flow.

2. Data Science Method: Opportunity analysis

Opportunity analysis is about taking an intuitive notion of a problem and quantifying 'why should we work on this?'. In the interviews, we heard concerns about the uncertainty, inconsistency, and cognitive overhead of the review systems, so if we wanted to use these reviews as evidence for automated systems we needed to understand how often these problems were occurring. The first place to look was the aggregated review data.

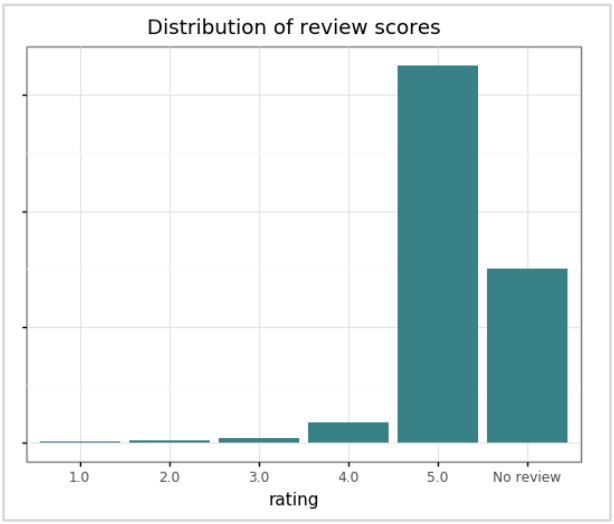

Figure 3. Ratings of guests left by hosts skew largely to 5 stars (1 to 5 stars, 5 is the best).

The vast majority of reviews were 5-star, and a substantial portion of guest stays also had no review at all. This was a challenge -- how many of those stays with no reviews were actually "less than ideal" stays, where the host felt uncertain so they left no review at all? To use statistical terminology, the 1:1 interviews made us pretty sure that these missing data were not missing at random, rather they were intentionally not completed. When comparing to 'guest reviews of stays' (the inverse type of review in the system), previous research showed that of the similar percentage of stays were left unreviewed. Follow-up research on the unreviewed stays indicated that indeed there was a portion who didn't review because they had a less than ideal experience but they didn't want to damage the hosts' reputation.

While the dynamic of guest versus host reviews are a bit different (hosts generally have more at stake), we expected that there was probably some similar behavior happening on the host side of things.

Furthermore, it was interesting that where reviews were left, the vast majority were 5-star. Research has shown that the extremely high ratings of hosts on Airbnb leads to loss of informative value for the guest; a host's reputation has a subsequent diminished effect on things like listing price of booking likelihood (Ert, Fleischer & Magen 2016). Yet, we knew that even though reputation systems are wildly inflated, they do matter in decision-making. Research has shown that reputation systems can significantly increase the trust between dissimilar users and there is an inverse relationship between risk aversion and trust in those with positive reputations. Having a high reputation is actually enough to counteract homophily. Specifically, research on 1 million requests-to-stay by guests on Airbnb data has shown a higher tolerance for individuals at farther social distances between guests and their selected hosts as the reputation of the host got better (Abrahao, Parigi, Gupta & Cook 2017).

At this point, we weren't sure exactly how to interpret the reality behind the inflation of the review ratings, but we would soon realize this was a hint of a deeper, fundamental problem of human psychology.

The combination of concerns from the qualitative research and the quantitative snapshot of the problem's scale made us concerned that the evidence about guest behavior collected in the current host review flow might be unreliable or incomplete. At an industry level, we knew review systems were imperfect but we wanted to see if we could go beyond face value and get a better signal. We suspected we might have to change the review flow -- but how?

3. Hybrid Method: Human judgements of analytically-sampled reviews

To move from identifying a problem to researching solutions, we realized we needed to answer a deeper, fundamental question: 'what is the difference between a problematic and a perfectly reviewed guest?' In trying to answer this, we landed on a new, hybrid methodology that was only possible with the combined skills of our disciplines. The hybrid method began with the realization that our review data was a unique data set that included both structured data (the rating), and unstructured data (the review text).

We agreed that when working with subjective evidence, in this case a review by a host, the 'ground truth' of what the evidence means has to be a human judgement. In this case, we can make that human judgement by closely reading the unstructured review text. Our goal was to compare our human judgements of 'problematic' and 'perfectly reviewed' guests from the unstructured text data with patterns in the structured data of the star rating.

Because the vast majority of Airbnb reviews of guests were 5-star, data science suggested we focus on two particular patterns of the structured data:

1. 'Perfectly reviewed guests' -- guests with at least 10 reviews, all of which had a perfect 5-star rating.
2. 'Potentially problematic guests' -- guests with at least 10 reviews, of which at least two reviews were only 1-star or 2-star.

Data science identified guests that fit these patterns, and then drew a sample of the *next* review received by each guest. For both 'perfectly reviewed' and 'potentially problematic' guests, we had examples of their next review being another 5-star, or a 1-or-2-star. We could then separately apply human judgement to four categories of reviews.

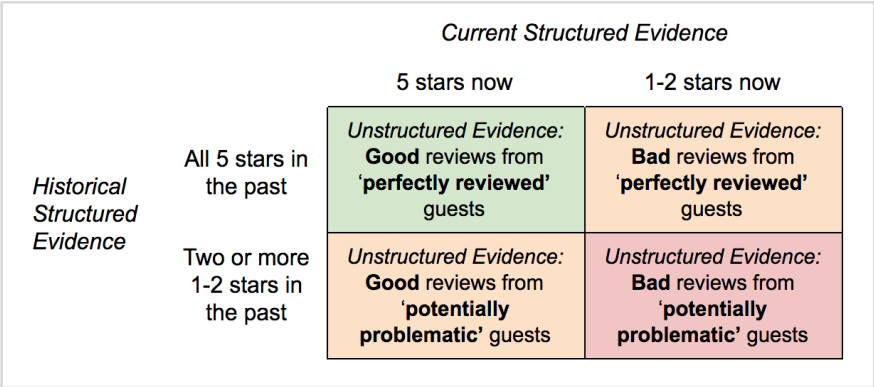

Figure 4. Four categories of reviews were sampled for analysis.

We printed out a few thousand samples, divided them into the four categories, and manually read through hosts' review text one by one, hand coding our observations. The results of this process of comparing how structured evidence related to human judgements of what a 'perfectly reviewed' or 'potentially problematic' guest is surprised us.

When we compared the one star reviews from the 'potentially problematic' guests with 1-star reviews from the 'perfectly reviewed' guests, we found there were clearly two types of one star reviews: 'actually problematic guests', whom it was clear should receive warnings if not be removed from Airbnb altogether; and 'potentially unlucky guests', who happened to run into a careless incident or what sounded like an overly-sensitive host. In a great deal of cases, these two types of reviews both received a 1-star rating.

> Actually Problematic: "Hosts beware of ___! ___ and her 3 friends stayed in my place. She is a very rude, entitled and ungrateful person… They all showered, left the heating on and did not say one word of complaint. Two days later she contacts me demanding a full refund despite the fact that she and all her friends used my place. If she was not happy she could have left and I would have refunded her."

> Potentially Unlucky: "I'm afraid I cannot recommend these guests to other hosts. They are polite girls but their lack of respect and care put us and our home at very real risk of fire. Somehow it appears a towel was left over a lamp that was on....the towel burned through and the lamp fortunately just melted as it was fire resistant. They informed me of some damage, paid to replace items and apologised. It's an experience I wouldn't want repeated."

It was not a hard-and-fast rule that all 'perfectly reviewed' guests with a 1-star review were only 'potentially unlucky', but it was certainly more common. If we were to make automated judgements about whether guests were 'problematic guests' based on the structured data of only one or two reviews, we would be very prone to unfair decisions based on incomplete data.

This discovery with respect to the 1-star reviews was mirrored with the *5-star* reviews. The vast majority of reviews were 5-star, regardless of whether the text of the review suggested the guest was 'consistently positively reviewed' or 'a truly exceptional, once-in-a-lifetime, would-invite-to-my-wedding personal connection.' The 'perfectly reviewed' guests were somewhat more likely than the 'potentially problematic' guests to receive what we considered an 'exceptional' review, but it again was not clear-cut: many of their reviews were also just fine and wouldn't put them in an exceptional category. Even some of the 'potentially problematic' guests also received what our human judgement considered were 'exceptional' reviews on occasion.

> Positively reviewed: "short but nice stay ... polite and nice guy."
>
> Exceptionally reviewed: "We are very lucky that we could meet ___ and ___. They are very interesting and friendly couple and it was so much fun to be around them. We are already missing our conversations. We were so impressed to find our place so clean and shiny after they left. You can not ask for better guests than ___ and ___. Our only regret is that their stay was far too short. Ps. ___, I really enjoy reading your book."

The fundamental problem had become clear: How can we trust this evidence of who is an 'exceptional' and a 'potentially problematic' guest in an automated system, if it takes so much of our nuanced human judgement to make these decisions? Since this problem was clear even among guests who had a long history of past evidence (10+ past reviews), we knew that for all our guests who had so far received only one or two reviews, there was simply not enough evidence to make a fair judgement.

Thanks to this hybrid research method, we now had a clearer mental model of the range of reviews in the system. Research has argued that the bonding power of the interactions have been diminished by the development of the online reputation systems in a sort of "disenchantment" created by technology (Parigi & State 2014). Despite this "disenchantment" amdist the inflated ratings, we still saw nuance and detail coming through in the review text. But to use the reviews we collected from hosts in automated systems, the reviews would have to provide evidence which could fairly distinguish between five different types of guest behaviors (below) -- and the current system barely even distinguished between two.

This led us to a fourth step in our methodology: designing prototypes of a new review flow that could provide more detailed and fair evidence, and putting these prototypes in front of hosts to gather feedback.

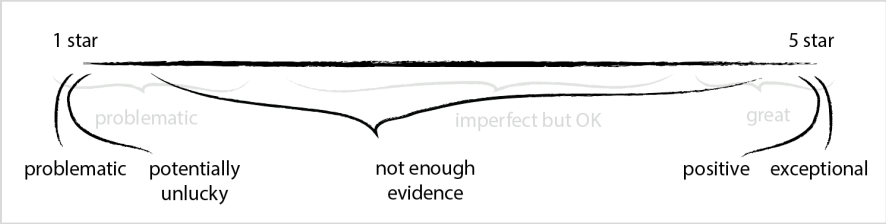

Figure 5. The updated mental model for how we might be able to differentiate guests using ratings & reviews.

4. Research Method: Prototype testing & participatory design

Applying our new understanding, we arrived at two key hypotheses for how we could prototype a better review flow:

1. Firstly, to help the star rating better reflect multiple types of guests (not just '5 stars' or 'not 5 stars'), we moved the point where we asked hosts for the star rating from up front, to the end of the review flow, after first asking hosts to relate the objective facts of their story. Our hypothesis was that this would lead to the star ratings being more spread-out, and thus better capture our more nuanced view of guest behavior.
2. Secondly, to further clarify the 'problematic' vs 'potentially unlucky' distinction, we added a question to ask how responsibly guests acted after any issues that arose. From the many samples we had reviewed before, we thought that any guest who responded responsibly after an issue had a high chance of being just 'unlucky', and not really a 'terrible' guest that shouldn't be on the platform. We also asked a question about how severe the issue was.

We designed two new interactive prototypes of the review flow that we believed would address the challenges of the existing review flow, and we invited hosts to share feedback.

Hosts were interviewed in pairs to stimulate discussion through disagreement and shared stories that would remind each other of their history of guest interactions. After discussing the key issues and reviewing the prototypes, we invited the hosts to share their ideas by drawing and explaining their own proposed review flow.

We realized were were way off track after our conversations with hosts. We thought we could create nuance and accuracy in the reviews through question wording, ordering and structured data capture, but there was a more fundamental issue at play: Hosts told us they were intentionally not sharing their true opinions of guests, because there's little incentive to review someone poorly. They inflate because of fear of retribution or a sense of guilt that they'd be individually responsible for any consequences to the guest (e.g. the guest won't get accepted in the future).

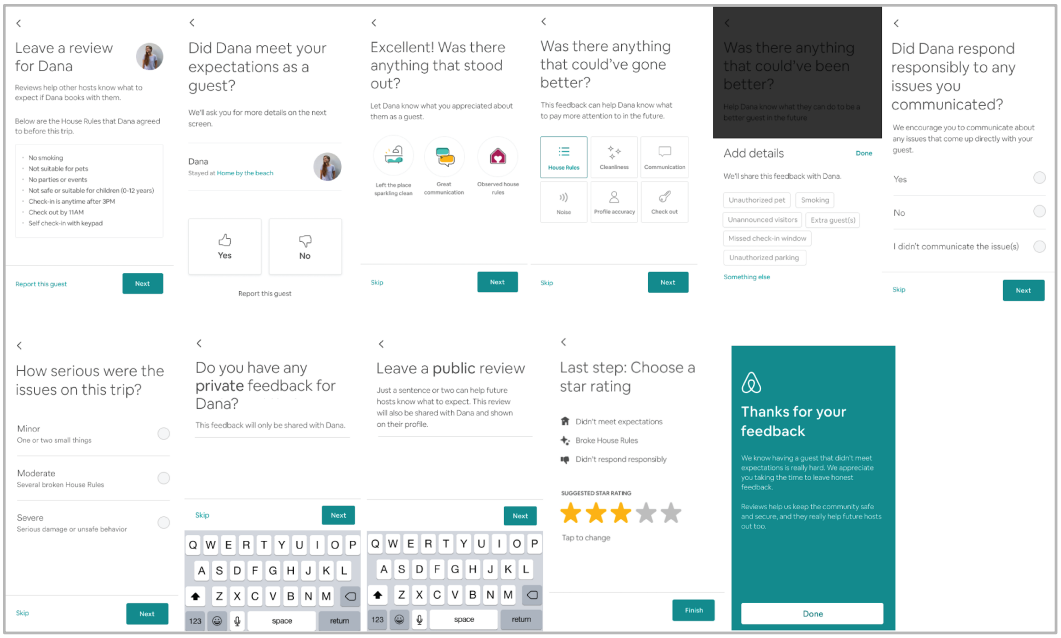

Figure 6. Screens from the interactive prototype of a mobile review flow designed for testing with hosts.

> "It's not the setup. It's about the guilt to say something nasty or unwillingness to grade or fear of retribution… The people I'd leave bad reviews for are the ones that might come back at you… and I don't know where they live."

> "I want to maybe note not to host them again but I don't want to ding her."

We began to form a new hypothesis: the underlying problem here is an intentional mismatch between hosts' public opinions and private opinions of guests. And if this is true, we can't solve it just by tweaking the order of review questions or asking them to go into more detail with structure content. We needed to design a system that would capture the intentional mismatch between public and private reviews.

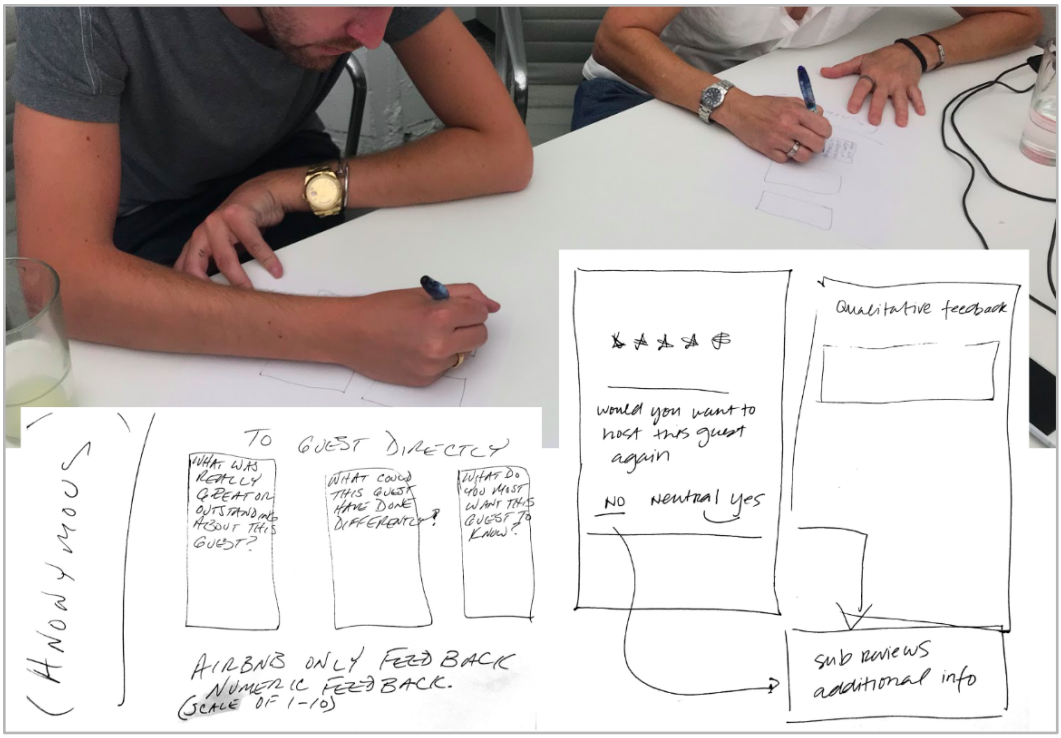

Figure 7. Pairs of Airbnb hosts sketch out their visions for an optimal guest-review flow.

ACT II

We had a hypothesis: review ratings were not consistent with our human judgements of what happened during the stay, because hosts were intentionally not telling us their true opinions. We needed a methodology by which we could investigate this hypothesis further, and more fully understand this mismatch in what hosts were saying publicly with how they sometimes really felt. At this point we were basing our hypothesis largely on the self-stated views of a handful of hosts. Yet this was a topic with so much opportunity for bias; we needed a way to tease apart the public-private distinction.

This is when we realized we had a data source we hadn't yet used: in the past, Airbnb used to give hosts an additional option when leaving reviews to also give private open text feedback to Airbnb. Even though we no longer ask this question, we still had past data we could utilize. Our hope was that if hosts were indeed often intentionally leaving a public review that didn't represent their true opinion, they would at least sometimes give genuine feedback in the private review to Airbnb. Because we had records of both public and private parts of the reviews, we could look for these mismatches.

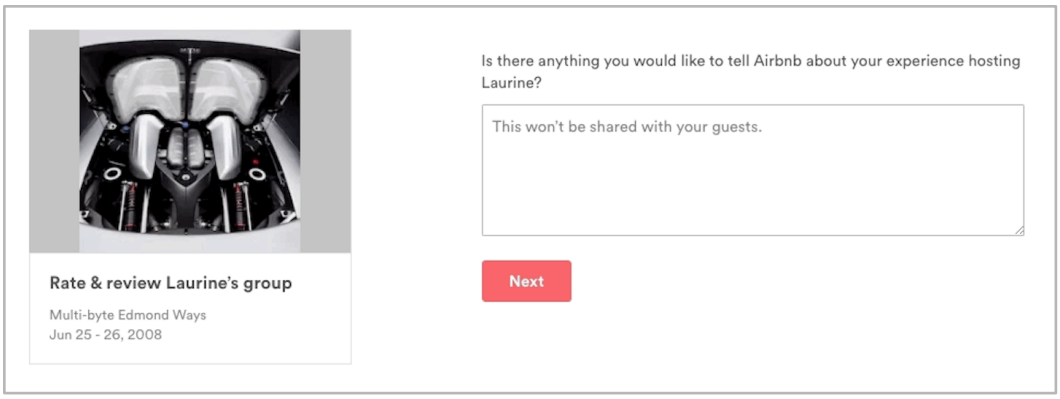

Figure 8. Final question in an old review flow that gave hosts the option to share private feedback with Airbnb.

5. Data Science Method: Sentiment classification

Our goal was to figure out if the mismatch between some hosts' public review and private feelings existed at scale in our review data. One data science technique was ideally suited to this challenge: 'sentiment classification', meaning to classify reviews as 'positive' or 'negative' based on the unstructured written text. If we could identify reviews as positive or negative based on the text alone, we could automatically identify cases where the public review was positive but the private review negative, and estimate how often this was occurring.

The trick to building any classification model, including a sentiment classifier for positive/negative text, was to have the right training data. In our case a unique opportunity presented itself, in that we had public review text that also matched to public review ratings. Our process was as follows:

1. Label public reviews that came with a perfect 5-star rating as 'positive', and public reviews that came with a 1-star or 2-star rating as 'negative'. In fact, we found we actually had to add a third category of 'unsure' to capture uninformative private reviews, such as when the host simply wrote 'No,' meaning, 'No I don't have any private feedback to give.'
2. Train a sentiment classifier on the public review text using the positive and negative labels. Because vast majority of the public reviews are positive, we downsampled the positive training data to create a balanced dataset. We used traditional sentiment classification methods where the public review texts are represented by the collection of meaningful words in them and their frequencies, thus the nuances of word order and grammar are ignored.
3. Confirm the accuracy of the classifier on a separate time period of public review data and ratings.
4. Use this classifier to classify the private review text (that lacks any rating label) as positive or negative. Then, count the percentage of the time that positive public text is written alongside negative private text. This is the 'public/private mismatch rate'.

The accuracy of the model when tested on the public reviews was 91%, with slightly higher accuracy at correctly identifying positive reviews (96%) than negative (83%). This result gave us high confidence that the model could correctly identify whether a review without a known rating label -- i.e. the private reviews -- were positive or negative.

The result we found was that there were enough reviews with 'positive sentiment' public review text that had 'negative sentiment' private review text that we knew the hosts were indeed not always willing to share their true feelings publicly.

Table 1. Stand-out examples of public-private mismatch

Public review	Private review to Airbnb
Easy communications and reliable guest. No problems	I was not there at the time of his visit but the cleaner reported the house to have been left "untidy" and with dirty dishes in the sink. They did pay a cleaning fee so perhaps they assumed that was acceptable.
[ ] is a decent responsible lady. The apartment was well-kept which is very important for us. I hope [ ] had a comfortable stay at ours.	Honestly I'd better not have such a guest next time… There was a plenty of dirty plates all over, shoes were worn inside the flat which is not common in Ukraine.
[ ] and her group were great guests. They were easy going and left the house nice and tidy.	They were a bigger party of people than expected... They failed to wash up properly leaving dried up food on plates, cups and cutlery. The oven was messy and I spend half hour cleaning it… Unfortunately, they didn't leave the house secure.
[ ] has a very friendly personality.	This woman is VERY HIGH maintenance! She has no boundaries or filters. One request after another, some of which were out of line. She would be more suited to stay in a hotel rather than someone's private home.
[ ] was very nice and clean.	[ ] asked to stay extra days and wanted to pay cash... He also completely ignored the checkout time and then needed to leave a lot of belongings at my place after he 'checked out' several hours later... He also is generally just a strange person who made me feel very uncomfortable and unsafe at times. I think he is probably nice but I would not host him ever again.
[ ]'s parents took care of my apartment. All went well. Thank you !	I won't rent to her again. Poor communication. Everything was very hard to deal with. First, I needed to give her the keys for the apartment. She was upset because no one was going to be there to receive them. She asked me to leave the keys at her house, which I did. She was still upset, no reason. After a few days, I asked her how things were going. She told me that Internet had been down for the last two days. I don't know why she wouldn't let me know right away. I called the Internet company and they told me that there was a problem in the whole building, not just my apartment. I shared the information with her, she was still upset with me. I ended up offering a very generous refund of 15% Still upset... I would define her as a crazy customer. Trouble maker.

To see that apparently-positive reviews actually had negative issues revealed in private was a striking discovery. Especially given that such a small percentage of reviews have negative public ratings -- this assured us that our public evidence of the quality of the guests was definitely underestimated.

This application of a data science modeling technique had succeeded: we'd proven that the evidence from the user interviews was showing up at scale. Intentional public/private

mismatches existed, and with our quantitative estimate, we knew that these intentional mismatches should be the focus of our attention rather than the mere inaccuracy of the public review questions asked.

This also made us think back to our earlier observation that the vast majority of reviews being 5-star seemed too high. It was too high, because hosts weren't always publicly sharing what they really thought -- and thanks to modern computational methods, we had the evidence to both prove and quantify how often this was happening.

6. Research Method: Remote prototype testing interviews

The final step of research was a set of remote interviews with hosts to gauge their willingness to report honestly using a new private feedback method. We designed a prototype of a new review flow that simplified the number of questions asked (drawing out nuance publicly was superfluous) and included a final question that we had heard almost every host previously interviewed speak to in one way or another: "Would you host 'this guest' again?" Hosts repeatedly had said that they didn't want to publicly thumbs down some people but they really didn't want to personally host them again. We decided that was the perfect question to ask that would yield the most honest answer. The final screen in the new review flow asks hosts this key question and indicates it will not be shown publicly.

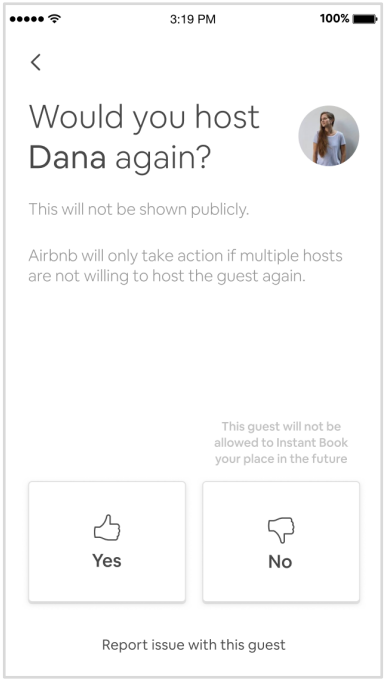

Figure 9. The final question of the new review flow asks for a private, binary rating; this should help unlock the public/private mismatch.

Hosts liked the addition of this question. They talked about the "gray area" where guests aren't terrible but also aren't great, and they would be more willing to share this honest

assessment using the private question of 'hosting again.' Interestingly, they took as a given that if they said they wouldn't host someone again, Airbnb would never allow that guest to book with them again, even when the version of the prototype they saw didn't explicitly promise this.

The callout on the review flow that "Airbnb will only take action if multiple hosts are not willing to host this guest again" is intended to protect guests who have unfortunate, isolated clashes with hosts, whether they are cross-cultural or personality-based. The hope is that these issues won't repeat themselves if they are indeed isolated, so the guests won't be unduly punished for single offenses.

We believe this structured "would you host again" question will yield useful data that we can act on at scale because it will begin to paint the picture of our updated mental model of guests (problematic - potentially unlucky - not enough evidence - positive - exceptional). In the past, when we had a private feedback question, it didn't direct hosts to focus on the key question of whether they'd host this guest again. It also couldn't reliably be acted on at scale because it was a free form text box. If this new review question gets implemented, future data should indicate whether we are able to use the answer to this question, aggregated over many reviews, to identify guests that need to be warned about poor behavior as well as identify exceptional guests for a potential loyalty program.

OUTCOME

The outcome of this integrated, cross-disciplinary approach was an understanding of the underlying constraints behind Airbnb being able to leverage subjective reviews to make automated decisions. The fundamental difference between 'public' and 'private' feedback is at the crux of this challenge. With this new knowledge, we were able to re-envision a review flow that separates and utilizes the public and private components differently. If implemented, we hope this will allow hosts to share feedback without guilt or fear of retribution while Airbnb can still collect reliable structured evidence from them. This can be fairly used to make automated decisions that will enable two key business goals: identifying guests that need to be warned about poor behavior or identifying exceptional guests for a potential loyalty program.

Our process of getting to the heart of the challenge was not direct; it included mishaps alongside major 'aha' moments, and each step revealed something unique that was essential to informing the solution.

1. *Research*: Our first finding revealed the inconsistency in hosts' approaches to leaving reviews and their desire to reduce the cognitive load of lots of free text.
2. *Data Science*: An opportunity analysis showed that the majority of reviews had 5 stars which seemed even higher than we expected. Almost a third of stays had no review; we had no way of knowing if they weren't reviewed because the guest was problematic or some other reason.
3. *Hybrid*: We used structured data (the rating), and unstructured data (the review text) to try and classify what is a 'problematic' from a 'perfectly reviewed' guest. We realized that human judgement was required to distinguish 'problematic' from 'potentially unlucky' and 'positively reviewed' from 'exceptionally reviewed.'

4. *Research*: We designed a review system that focused on drawing out the nuance of a rating with structured questions but, in doing so, discovered we were missing the big picture; hosts were actually intentionally not sharing their true opinions of guests because of guilt or fear of retribution. No level of nuanced questioning would get them to be more honest so long as the review was shared publicly. The current review system wasn't capturing their private opinions, which sometimes included not wanting to host a specific guest again.
5. *Data Science*: After realizing the public/private mismatch was the crux of the challenge, we dove into old data from when we used to ask hosts for private open-ended feedback on guests. We used semantic classification to identify if reviews were positive or negative and we were able to identify that there was indeed a mismatch in a decent percentage of the apparently-positive reviews, which was striking, given how few reviews actually get negative public ratings.
6. *Research*: Finally, we landed on a new, simple review flow that asks an additional key question privately of hosts: "Would you host this guest again?" If implemented and then aggregated over many reviews, the new data should allow us to both identify guests that need to be warned about poor behavior as well as identify exceptional guests for a potential loyalty program.

DISCUSSION

Reaching this outcome was only possible through a deep synthesis of research and data science. First, we applied human intuition when reading review ratings which revealed that our judgements differed from the recorded data. We saw that the aggregated stats about this data has major consequences at scale. Second, in-depth user tests of a new approach led to a hypothesis that we had misidentified the underlying problem. Data science techniques reaffirmed that our new understanding of the problem was correct. Third, we confirmed that a new approach could solve this problem by scaling an in-person UI test to a statistically significant sample. In reviewing our process, we recognize two generalizable principles that could inform future collaborations where a synthesis of qualitative and quantitative methods is paramount.

Principle 1: Shared artifacts

The first principle is to look for sources of information that can be accessed by multiple disciplines each in unique ways. In our case, we found shared artifacts in the samples of reviews; the review text was both a record of nuanced and layered human expression, and a natural language dataset that could be analyzed at scale for semantic patterns and linked to quantitative ratings. It was applying our different expertise to explore this complex and unique dataset together that generated some of our most important insights.

Principle 2: See-sawing

The second principle we arrived upon is a process of iterating back and forth between our disciplines — we've started calling this see-sawing. Multiple times, one of our approaches

seemed to reach a roadblock, but by switching focus to the other discipline a new way forward opened up. For example, our first quantitative analysis of ratings suggested this evidence was too abstract to be useful (almost all 5-star), but 1:1 interviews revealed the nuance that led to a new design to capture more accurate review data. However, when this design was tested with hosts, it seemed like we hit a different wall - hosts did not *want* to give us genuine feedback, and for good reasons. Yet, here a more quantitative way of thinking helped us move forward in seeing the public/private mismatch as a pattern that could be teased out, modeled, and classified. In retrospect, it's not always that the specific skills of the other discipline were necessary to move forward, but rather that the contrasting way of thinking helped shine a light on a new direction.

The conditions under which these principles are most useful are for problems that present puzzles of human psychology and emotion that vary greatly between individuals. In such cases, a purely quantitative approach is unlikely to reach useful conclusions through analysis, yet a purely qualitative approach will be limited in knowing how well individual stories and emotions can be generalized to be wider population. Such cases confront both disciplines with questions that they are ill-equipped to answer. In combination, however, research and data science display a remarkable capacity to combat each other's shortcomings, and reveal new insights about the complex patterns of human experience.

Stephanie Carter leads the Guest Experience Research team at Airbnb. Prior to Airbnb, she worked as a researcher for Facebook and for a med device design incubator; as a Resident Naturalist the jungles of Peru; and as an Urban Planner in post-Katrina New Orleans. She's driven by research's ability to understand and improve people's interactions with their built, natural, and digital environments. She studied Environmental Sciences & Art at Northwestern University and received her MFA in Design at Stanford. *stephanie.carter@gmail.com*

Richard Dear is a data scientist at Airbnb focusing on guest retention and loyalty. Prior to Airbnb he designed mobile games at two startups in China. He studied physics and philosophy at the Australian National University, though he remembers stressing about giving a TEDx talk on cross-disciplinary research more clearly than his actual courses. *richard.dear@airbnb.com*

REFERENCES CITED

Parigi Paolo. and Bogdan State
2014 Disenchanting the World: The Impact of Technology on Relationships. In: Aiello L.M., McFarland D. (eds) Social Informatics. SocInfo 2014. Lecture Notes in Computer Science, vol 8851. Springer, Cham.

Ert, Eyal, Aliza Fleischer, and Nathan Magen
2016 Trust and reputation in the sharing economy: The role of personal photos in Airbnb. Tourism Management 55 (2016), 62–73.

Abrahao, Bruno, Paolo Parigi, Alok Gupta, and Karen Cook
2017 Reputation offsets trust judgments based on social biases among Airbnb users. PNAS 114 (37): 9848-9853.