Additional Techniques for Brainstorming and Validating Metrics

These techniques can help you get an understanding of your users, which you can use to come up with ideas for metrics, validate your existing metrics, or even brainstorm ideas of what you might want to test in your experiments in the first place.

External Data

You might be able to use external data, that is, data collected by people outside of your company.

What External Data Is Available?

1. There's outside market share data, provided by companies such as Comscore and Hitwise, which includes things like how many users visit a site—for your site or for competitors or related sites.
2. There are also companies such as Nielsen, Forrester, and Pew Research that run and publish their own studies. For example, you can find studies surveying users on how many devices they use on a given day, or tracking the activity and recording detailed observations about the online usage of a panel of consenting users. Companies may have already done research using a variety of methods to answer questions that you may be interested in. Depending on what industry or type of site you work on, you might be able to find data from other people's experiments that is useful—and your company might already subscribe to these publications.
3. There are higher level aggregators of data, such as eMarketer, who provide summaries from all of these sources in easily consumable fashion.
4. There is also a whole treasure trove of published papers, where researchers have investigated all sorts of interesting questions in a rigorous fashion, either about how users behave in certain scenarios, how metrics are related, etc. Potentially relevant conferences include:
    a. CHI
    b. WWW
    c. KDD
    d. WSDM

What Can You Do With This External Data?

First, you can validate simple business metrics if your site or industry appears in one of these lists. For example, if you want to look at total visitors to your site, you can compare your number with the numbers provided by Comscore or Hitwise, or you could compare the fraction of shopping traffic in each "vertical" category to what you see on your site. However, the numbers you see will almost **never** exactly match your own data. Generally speaking, a better way to do the validation is to look at a time series of both your internally computed metric and the externally available one, and see if the trends and seasonal variability line up.

Second, you can provide supporting evidence for your business metrics, either direct measurable quantities (look, this is used by lots of sites) or to get ideas for which measurable metrics make good proxies for other harder-to-measure quantities.

Publicly available academic papers, such as the User Experience ones, often establish a general equivalence between different types of metrics. One example that Carrie worked on was a paper with Dan Russell, which compared user reported satisfaction with a search task to the duration of the task as measured on the

website. That gave a good general correlation for satisfaction with duration measured, though with some clear caveats. So this study helped validate a metric—duration—that could be computed at scale and then automatically converted to a metric that could not be compute at scale—user reported satisfaction.

Gathering Your Own In-Depth Data

Another way is to gather your own in-depth data, or hire an external firm to do it for you. There are three major methods commonly used to do this: User Experience Research (UER), focus groups, and surveys. These methods vary along two major axes: how many users you can study, and how in-depth you can go per-user, i.e., how qualitative to quantitative the study is. What you choose to do may depend on where you work—if you are at a large company, you may already have groups with data or can commission data. If you are developing your own app or website, you may choose something easy to do in an automated way such as surveying users on the site.

User Experience Research

With UER, you can go really deep with a few users, either in a lab or even in a field study. This is the most in-depth and intensive, and can be a combination of observing users doing tasks and asking them questions. The idea is to spot problems and draw insights from the observations and timely questions. This can be useful for generating hypotheses for problems that you might want to tackle fixing in experiments. You can also generate possible metrics to track those problems given the behavior you've observed. In another variation of a UER study, often called a **diary study**, you ask users to self-document their behavior rather than observing them in a lab or in the field. Diary studies can usually get more users, because you don't need a researcher to be spending time observing each participant, but they have more issues with self-reporting bias. UER studies are in-depth and qualitative, and you'll study maybe tens of users at most. They're great for generating ideas for metrics or for experiments, but you'll want to validate the results with methods that scale up to many more users.

Focus Groups

In focus grous, you recruit a bunch of users or potential users and bring them together for a group discussion. Once you bring the users together, you could show them screenshots or images, or you can walk them through a demo, and then you can ask them questions to elicit feedback. You can talk to more users than in a UER study, maybe hundreds of users, and you can often ask hypothetical questions, but even if you have a skilled facilitator, you run the risk of group-think and convergence on fewer opinions. Haven't you been in a room and there's been a few loud voices that dominate? That same dynamic can occur in focus groups.

Surveys

The final method is to run surveys or questionnaires, where you recruit a population and ask them to answer a bunch of questions. The number of questions can vary, as can the type of question. You can have multiple-choice answers, or open-ended questions where users give a free-form response. These can be done online, either on your site or other methods of reaching and targeting users, such as Google's Consumer Surveys. The main advantage of surveys is that you can typically reach thousands of users, if not more. The disadvantages are that the data is self-reported and users don't always give full or truthful answers, even if the survey is anonymous. When running a survey, you'll need to take care in how you recruit your population to ensure that it's representative, and you'll need to word the questions carefully since the wording may prime the participants to give specific answers. Surveys are useful as a way to get data for metrics that you cannot get from your system, but they're not really comparable to the metrics that you will

obtain from your own logs and data capture. For example, the populations may not be comparable since you may be reaching a biased population with your survey relative to the population using your website or app.

Additional Reading

For more reading across all three (and more) methods, see:

- http://www.nngroup.com/articles/which-ux-research-methods/
- http://www.usability.gov/what-and-why/user-research.html

Retrospective Analysis

If you are running experiments, you must have logs or other data capture mechanisms to see what users do. Running analyses on this existing set of observational data **without** an experiment structure is called **retrospective analysis**, or **observational analysis**. These types of analyses are useful for generating ideas for A/B tests and for validating metrics. For example, if you observe something in a UER study, you could then go looking for that pattern in your logs to see if that observation bears out and is worth creating a metric about. You could look at the logs to see what the distribution of the latency of video loads might be and when the next user action is to see if that's an interesting area for exploration. If you observe that a few students in a UER study are getting stuck on a particular quiz, you could analyze all interactions with that quiz to see if that is borne out at scale. Oftentimes, the usefulness of these retrospective analyses is determined by how you frame the question and the analysis.

Long-term prospective experiments

One last way to come up with or validate metrics is to use long-term experiments. Specifically, what do you do when the metric you care about is only measurable in the long-term, such as whether users are getting jobs, long-term revenue, or other such metrics? While you could do the retrospective analyses as described above, that only gives you correlated results, not causal. Another option is to run some A/B tests and measure a change in that metric in the long-term—these will be long-running tests. Then, given the measured change in the long-term metric, you can look to see what metrics are measurable in the short-term that best predict the change in the long-term metric. For example, one thing Diane and Carrie care about at Google is long-term revenue, and they want to know whether showing users fewer ads but good ads increases their interaction with ads and increases long-term revenue, even though showing fewer ads decreases revenue in the short-term. They've run experiments over months and years to measure those long-term effects, and then build models to determine which short-term metrics best predict the long-term effects.

Human Evaluation

One other method frequently used, especially in search or other ranking-oriented systems is human evaluation. In human evaluation, you're paying people (raters) to complete a specific task. A canonical task for search would be to give the rater a query string and a result, and ask the rater how relevant that result is to that query. Typically, raters would receive (potentially long) instructions so that their rating of relevance is calibrated to some scale. Depending on how complicated the task is, you may need more or fewer instructions. Some tasks are relatively straightforward ("Do you prefer side A or side B", "Is this image pornographic?") and can get progressively more complicated ("Please label this image", "How relevant is this result for this query"), especially as you want more calibrated results.

Human evaluation, or "crowd-sourcing" can be helpful for getting a lot of labeled data. As the tasks get more complicated, the labels from different people are more likely to disagree unless you have clear guidelines. The quality of the data that you get from pay-systems such as Mechanical Turk can vary

depending on the amount that you pay and what incentives you give -- you should be prepared to be doing quality control on the data that you get from such a system and to throw some of it away as junk.

You can read more about human evaluation for search at Google here. Systems that allow you to create tasks and pay people to complete tasks include Mechanical Turk or MicroWorkers. There are numerous academic papers on human evaluation here and here.

Application to the Audacity example

In addition to the examples discussed in the videos, here are some possible ways to apply these techniques to the Audacity business.

Audacity might compare their class completion to published completion rates, e.g. for MOOCs at large or for competitors.

Audacity might also use focus groups to get a feel for what types of classes should be prioritized next, the typical reasons a student might not complete a lesson or class, or how often students are interrupted and why.

They might also use surveys to get a feel for how many students get a job and how instrumental the online classes are in getting that job. For example, the survey could ask whether the student's interview questions touched on the topics covered in class. Then Audacity could see if number of classes taken, time to complete classes, time between classes, or another metric is predictive of success in finding a job. Depending on the results, one of these might be a good proxy for the ultimate mission that Audacity cares about, although this metric would need to be revisited periodically.

Or if Audacity suspects that, for example, video load time correlates to other metrics they care about, they could verify this using retrospective analysis or additional prospective experiments.

Choosing a Technique

First, you need to consider your goal. Do you want to figure out how to measure a particular user experience? Or do you want to validate metrics? External data and retrospective analyses are best for validating metrics: the data is usually collected at scale, in other words, over a large enough population that there are likely fewer sampling biases or other measurement issues. Practically speaking, what you use may depend on what you have on hand or what you can get, and the bigger the decision, the more time you may want to invest in getting additional data. The key thing is that you want to compare different ways of measuring the same thing to see if you can draw the same conclusions. Just remember that you don't expect the absolute values to match, you are primarily looking to see if trends match.

On the other hand, if you're trying to come up with new metrics to measure user experiences, then it depends. If you have no idea about what metrics to gather in the first, place, more detailed, qualitative, brainstorming type of interactions, such as user experience research studies or focus groups work well. If you have no way of getting the data, because the interactions aren't on your site, something like a survey might work well. If it's an issue of time to measure, then something like retrospective analyses paired with long-term experiments might work.

All of these techniques have different trade-offs. You have to consider how many people you'll be able to collect data from. This affects the generalizability of the results and whether you'll have seen sufficient variability. Number of users often trades off against how much detail you'll be able to get. For example, logs usually have the scale but not the depth of information as you might get in something like a UER study.

Finally, remember that using multiple methods to triangulate towards a more accurate measurement is often a good idea. For example, Diane and Carrie wrote a paper with Dan Russell talking about how using query logs and retrospective analyses weren't enough to actually answer questions, and using the other methods provided a lot of value.