8

Institutional Memory and Meta-Analysis

Individuals sometimes forgive, but bodies and societies never do
– Lord Chesterfield (1694–1773)

Why you care: *As your organization moves into the "Fly" maturity phase, institutional memory, which contains a history of all experiments and changes made, becomes increasingly important. It can be used to identify patterns that generalize across experiments, to foster a culture of experimentation, to improve future innovations, and more.*

What Is Institutional Memory?

After fully embracing controlled experiments as a default step in the innovation process, your company can effectively have a digital journal of all changes through experimentation, including descriptions, screen shots, and key results. Each of the hundreds or even thousands of experiments run in the past is a page in the journal, with precious and rich data on each change (launched or not). This digital journal is what we refer to as *Institutional Memory*. This section is about how to utilize the institutional memory through meta-analysis, and mining data from all these historical experiments.

It goes without saying that you need to capture and organize data as part of institutional memory. Having a centralized experimentation platform, where all changes are tested, certainly makes it easier. It is highly recommended that you capture meta information on each experiment, such as who the owners are; when the experiment started; how long it ran; descriptions and screen shots if the change was visual. You should also have results summarizing how much impact the experiment had on various metrics, including a definitive scorecard with triggered and overall impact (see Chapter 20). Lastly, you should capture the hypothesis the experiment is based on; what decision was made and why.

Why Is Institutional Memory Useful?

What can you get from mining data from all these experiments? This is what we refer to here as meta-analysis. We organize the use cases into these five categories:

1. **Experiment culture.** Having a summary view of past experiments can really highlight the importance of experimentation and help solidify the culture. Here are a few concrete examples of meta-analysis to do:
   - **How has experimentation been contributing to the growth of the broader organizational goals?** For example, if the company's goal is to improve sessions-per-user, how much session-per-user improvement over the past year is attributable to changes launched through experiments? This can be many inch-by-inch wins added together. Bing Ads shared a powerful plot that shows how their revenue gains between 2013 and 2015 were attributable to incremental improvements from hundreds of experiments (see Chapter 1).
   - **What are the experiments with big or surprising impact?** While numbers are great at helping organizations gain insights at-scale, people relate to concrete examples. We find it helpful to regularly share experiments that are big wins or that have surprising results (see Chapter 1). As we mentioned in Chapter 4, we can also share a regular report on experiments that have a big impact on the metrics people care about.
   - **How many experiments positively or negatively impacted metrics?** At well-optimized domains such as Bing and Google, by some measures success rate is only 10–20% (Manzi 2012) (Kohavi et al. 2012). Microsoft shared that a third of their experiments moved key metrics positively, a third moved negatively, and a third didn't have significant impact (Kohavi, Longbotham et al. 2009). LinkedIn observed similar statistics. It's always humbling to realize that without experimentation to offer an objective true assessment, we could end up shipping both positive and negative experiments, canceling impact from each other.
   - **What percentage of features launch through experiments? Which teams have run the most experiments?** What is the growth quarter over quarter or year over year? Which team is the most effective at moving your OEC? Which outages are associated with changes that were not

experimented with? When postmortems on outages must answer such questions, the culture changes because people realize that experiments indeed provide a safety net. For bigger companies where there are many teams involved in running many experiments, it helps to create the breakdown and encourages better accountability.

2. **Experiment best practices.** Not necessarily every experimenter follows the best practices. This is especially common when more and more people start to experiment. For example, does the experiment go through the internal beta ramp period that is recommended? Is the experiment powered enough to detect movement of key metrics? Once you have enough experiments, you can conduct meta-analysis and report summary statistics to show teams and leadership where they can improve. You can break down the statistics by teams to further raise accountability. These insights help you decide whether you should invest in the automation to address the biggest gaps. For instance, by examining experiment ramp schedules, LinkedIn realized many experiments spent too much time on early ramp phases, while others did not even go through the internal beta ramp phase (see Chapter 14). To address this, LinkedIn built an auto-ramp feature that helps experimenters follow best ramping practices (Xu, Duan and Huang 2018).
3. **Future innovations.** For someone new to your company or new to a team, having a catalog of what worked and what didn't in the past is highly valuable. This helps avoid repeating mistakes and inspires effective innovation. Changes that did not work in the past, perhaps because of macro environment changes may be worth trying again. As you conduct meta-analysis on many experiments, patterns emerge that can guide you to better ideas. For example, which type of experiments are most effective for moving key metrics? Which kind of UI patterns are more likely to engage users? GoodUI.org summarizes many UI patterns that win repeatedly (Linowski 2018).

   After running many experiments that optimize a particular page, such as the Search Engine Results Page (SERP), you could predict the impact that changes to spacing, bolding, line length, thumbnails, and so on has on the metrics. Therefore, when you add a new element to the SERP, you can narrow the space of experiments to run. Another example is looking at experiment heterogeneity across countries (see Chapter 3), you can uncover hidden insights on how countries react differently for features, which allows you to build a better user experience customized for these users.
4. **Metrics.** Metrics are inseparable from experimentation (see Chapter 7). You can look across your experiments and how various metrics are

performing to develop a deeper understanding of how to better leverage them. Here are some example use cases of meta-analysis for metrics:

- **Metric sensitivity.** While developing metrics, one key criterion is whether they can be meaningfully measured during experiments. A metric that no experiment can move statistically significantly is not a good metric (see Chapter 7). While variance is a key factor influencing sensitivity, how likely an exogenous change can impact a metric is also a consideration. For example, daily active users (DAU) is a metric that is hard to move in short-term experiments. Studying existing metrics by comparing their performance in past experiments allows you to identify potential long-term vs. short-term metrics (Azevedo et al. 2019). You can also construct a corpus of trusted experiments to evaluate new metrics and compare different definition options (Dmitriev and Wu 2016).
- **Related metrics**. You can use the movement of metrics in experiments to identify how they relate to each other. Note that this is different from metric-to-metric correlation. For example, a user who visits LinkedIn more often tends to also send a lot more messages. However, sessions and messages don't necessarily move together in experiments. One example of related metrics in experiments is early indicators, which are metrics that tend to show leading signals for other metrics that take time to show impact. This is especially useful if those slow-moving metrics are critical for decision making (see Chapter 7). By studying a lot of experiments, you can uncover these relationships. See Chen, Liu and Xu (2019) for how such insights are uncovered and utilized at LinkedIn.
- **Probabilistic priors for Bayesian approaches**. As the Bayesian view of evaluating experiments gains popularity, one key concern is whether you can construct reasonable priors. For more matured products, it is reasonable to assume metric movement in historical experiments can offer reasonable prior distribution. See Deng (2015). For product areas evolving rapidly, it is not clear whether empirical distributions from the past can reasonably represent the future.

5. **Empirical research.** The vast amount of experiment data also offers researchers empirical evidence to evaluate and study their theories through meta-analysis. For example, Azevedo et al. (2019) studied how a company can best utilize experimentation to improve innovation productivity. They proposed an optimal implementation and experimentation strategy based on thousands of experiments that ran on Microsoft's experimentation platform. Experiment randomization can also act as a great instrumental variable.

By looking at 700 experiments conducted on the “People You May Know” algorithm at LinkedIn between 2014 and 2016, Saint-Jacques et al. (2018) found causal evidence that it is not the strongest connections that help people land a job, but those that strike a compromise between strength and diversity. Lee and Shen (2018) looked at how to aggregate impact from many launched experiments. When a group of experiments is conducted, usually those with significant successful results are chosen to be launched into the product. They investigate the statistical selection bias in this process and propose a correction method based on studying the experiments run on Airbnb’s experimentation platform.