Can I A/B Test That? Yes, with Few Exceptions

Ronny Kohavi
Aug 26, 2023

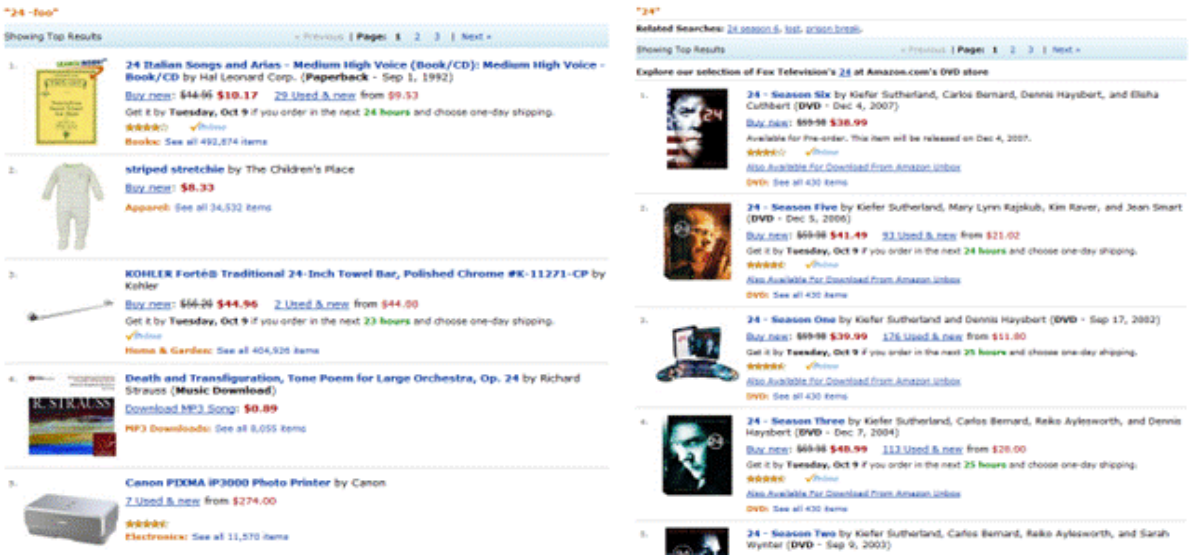
Introduction

There are common organizational debates about what *could* be tested in an online controlled experiment, or A/B Test. In our book (https://experimentguide.com) we have a chapter on Ethics, and there are clearly some examples that are out of bounds in some fields, like medicine. On a website, application, or service, it is rare for people to propose ideas that could cause material harm. While it is definitely important for the test to be legal, substantial user harm is unlikely. The usual concerns are around "industry standards" or strongly held beliefs, which are poor reasons not to experiment.

Amazon Example: Behavior Based Search

I want to share such an example from my Amazon days: Behavior Based Search (initially shared in http://bit.ly/expSurvey Section 2.4 back in 2009). Back in 2004, when I worked as director of data mining and personalization, we had the signature feature: "People who bought item X bought item Y," and we started generalizing it to "People who viewed item X bought item Y" and "People who viewed item X viewed item Y." A proposal was made to use the same algorithm for "People who searched for X bought item Y." We called it Behavior-Based Search (BBS) and gave it to an intern as a summer project.

The results were amazing. An extreme example was an underspecified search for 24. As shown in the figure below, search returned random stuff: CDs with 24 Italian Songs, clothing for 24-month old toddlers, a 24-inch towel bar, etc. The BBS algorithm gave top-notch results with the DVDs of the show "24," that is, things that people purchased after searching for "24" on Amazon.



This was hugely debated, with some good arguments:

1. The team (personalization) is encroaching on another team's charter. I was in charge of personalization, and this was search. I should sell the search team on the idea, let them put it in the backlog, but not build the feature and test it in the wrong team. (I tried selling it but given their backlog and the low priority this idea received, it was estimated they might get to it in about two years.)
2. We did not use the search terms directly. There is a whole science to search with bags of words, cosine similarity of vectors, and our team has very little understanding of the literature. We proposed a trivial algorithm that violated many of the basics.
3. We did not respect the user's keywords. A seemingly good example was raised at the time, where if you searched for "Sony HD DVD Player," our algorithm would return Toshiba HD DVD players; it's as if we ignore the explicit keyword "Sony." How could we be so dumb?
   For those old enough to remember, Sony was betting on Blu-Ray, and almost all their good DVDs were Blu-Ray. Our algorithm was very effective at noticing that people who typed this query realized that there's no good Sony HD DVD player, and most ended up buying a Toshiba, which our algorithm recommended.

We weren't going to get search support, so we implemented a UI hack: we would stick our BBS results at the top of search results when we were confident of the recommendations and do nothing otherwise.

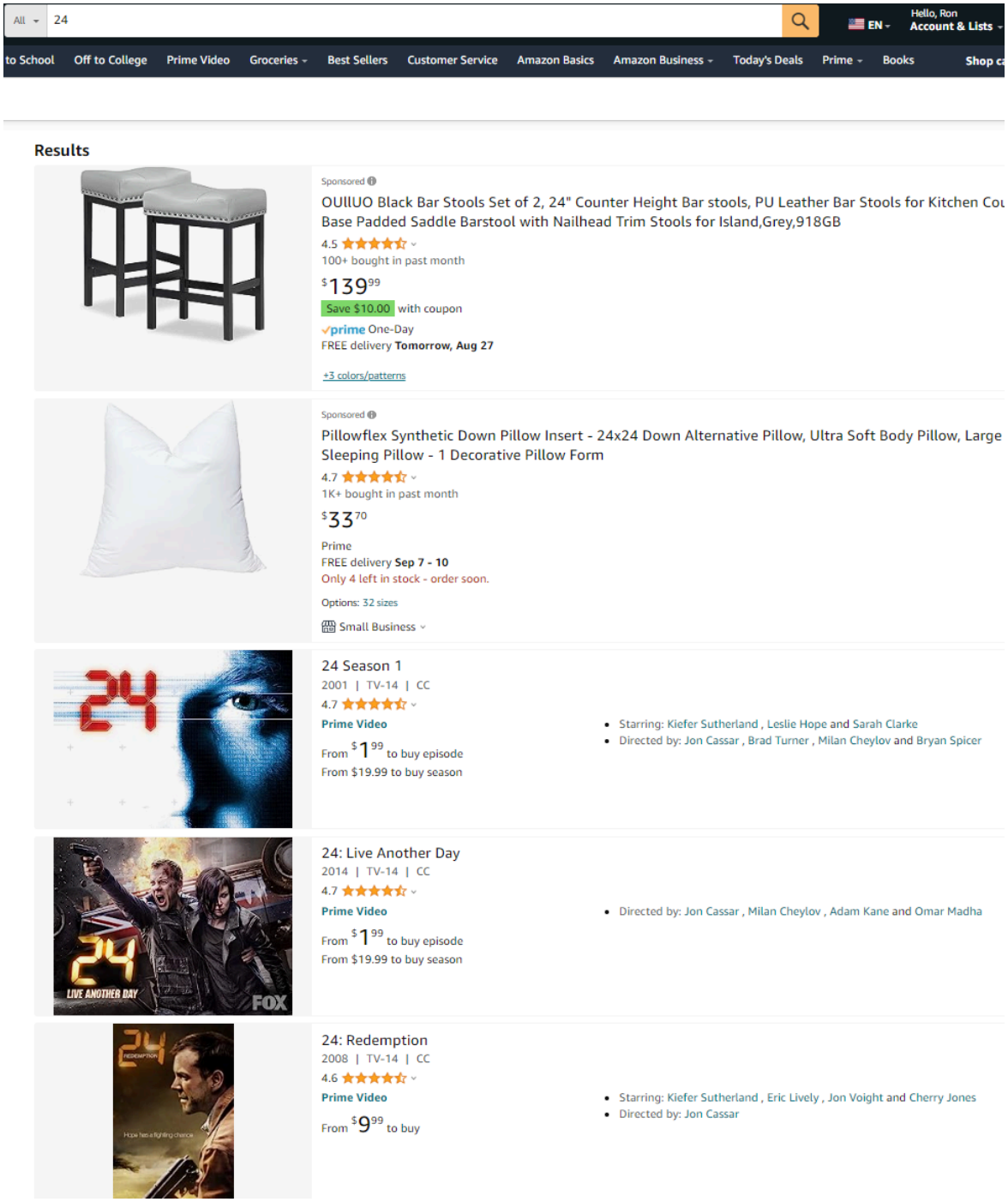
This escalated, and all the way to the top. Credit to Jeff Bezos, he was of the opinion that we should be able to run any reasonable test and learn from it. We ran the test with the UI hack.

The test was amazingly positive to key metrics, and after a few iterations, we got a 3% increase to Amazon's revenue. As you can imagine, this was hundreds of millions of dollars at the time. Very successful internship 😊

Prologue

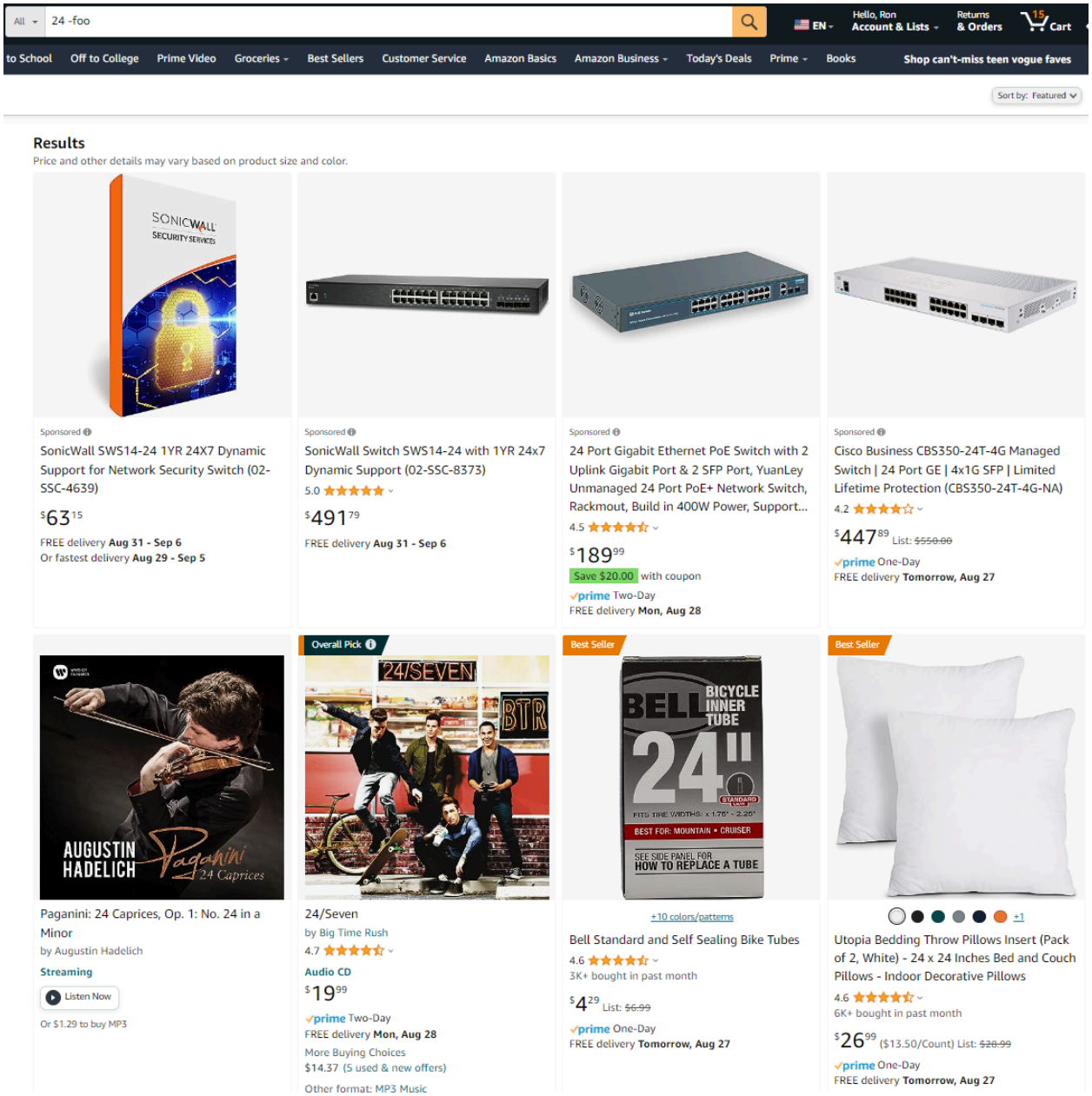
The feature was still not loved by the search team. The UI hack was live for several years before BBS became an input to search and integrated properly at the backend.

Search ads are still unable to handle 24 correctly. The figure below shows two irrelevant ads for 24" stools and 24x24 down pillow. It's very very very unlikely that anyone searching for a stool or pillow would only type 24. Paul Kotas should integrate BBS 😊



Search itself has improved since 2004, but it's still unable to handle "24" as a search query. By searching for "24 -foo" (24 excluding results with foo) we break BBS since nobody searches for this string.

Amazon now returns four ads with hardware network switches that have 24 in the name (usually indicative of 24 ports). After the ads, we have 24 Caprices No 24 by Paganini, 24/Seven CD, a 24" bicycle inner tube, and a 24x24 pillow. All these are pretty terrible results for the query.



Summary

When someone asks “Can I test that,” I would ask three questions

1. Is it legal? You shouldn’t test something that’s illegal.
2. Is it ethical? If there are doubts, bring the test to an Internal Review Board.
   We slowed Bing users, which some viewed as unethical, but I think the learning about the importance of performance (summarized in our book) outweighed the short-term harm. Look at the Belmont report for guidance.

3. If the test is positive, will you ship it to all users? The answer is usually, yes, but when it's not, and the reason is to test something extreme for the learning, bring it to an Internal Review Board to assess whether the learning benefits could outweigh the risk. The Facebook Contagion experiment falls here.

In most cases, the answer should be: if YOU think (integrating all available data) the test has the potential to win and you prioritized it highly with your resources, go ahead and test it.