Lecture 1: A/B Testing

Bryan S. Graham, UC - Berkeley & NBER

January 29, 2019

In 2008, then Senator Barack Obama, used a well-designed online presence to assemble an e-mail list of roughly thirteen million supporters, whose donations financed his campaign for the presidency. His fundraising operation was so effective that he became the first candidate to decline public campaign financing in the post-Watergate era.

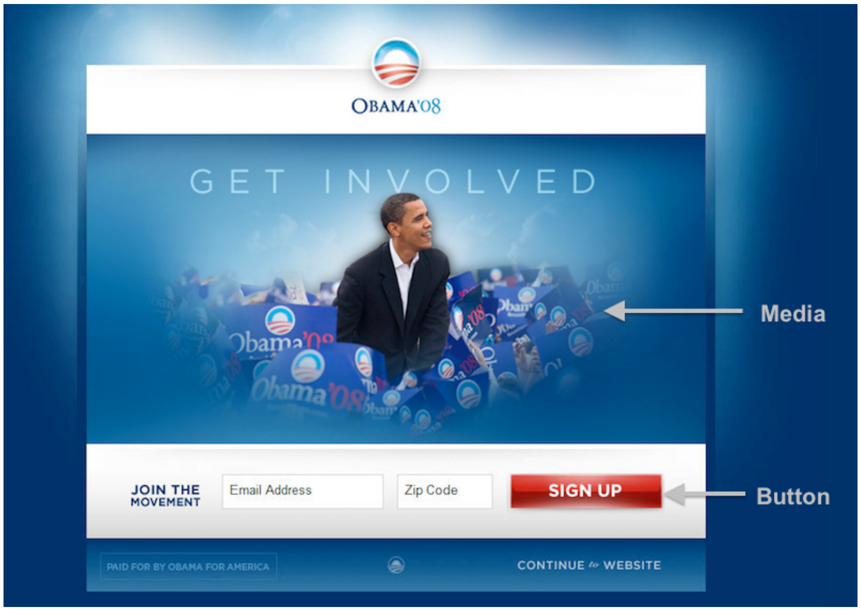
Not well-known at the time, Senator Obama's web-campaign was the product of repeated experimentation and redesign. According to Brian Christian, in an article for *Wired* magazine, in the early stages of Obama's 2008 campaign, visitors to his website were greeted by (see Figure 1)

> a luminous turquoise photo of Obama and a bright red "Sign Up" button. But too few people clicked the button **?**.

With the help of Dan Siroker, at that time a member of Google's web browser team, campaign staff ran an experiment. They tested three alternative phrasings for the sign up button: "Learn More", "Join Us Now" and "Sign Up Now". They also experimented with different photographs of the Senator (as well as video clips). These experiments revealed that the combination of "Learn More" and a particular black and white photograph of the Obama family could increase visitor sign up rates – leading to valuable e-mails and ultimately donations – by an extraordinary 40 percent (see Figure 2). **?** estimates that the home page redesign led to an $60 million in additional campaign donations and about 300,000 extra volunteers.

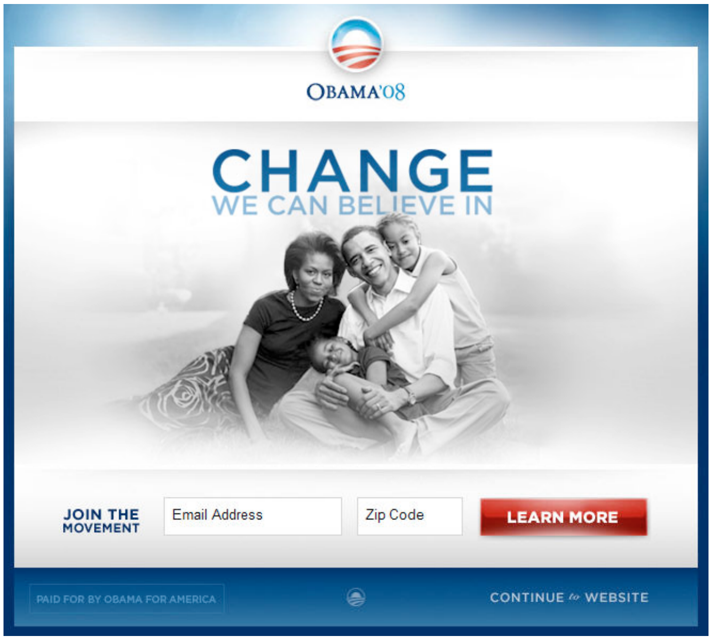
The type of testing done by the Obama campaign is called A/B testing. This type of testing is routinely used in internet-based industries. Consider a newspaper editor choosing a headline for the paper's lead story. In the past an editor would need to rely on past experience and gut instinct in making this decision. Now newspapers can randomly direct visitors to their website to versions of their home page with different headline phrasings. If one headline

Figure 1: Obama's original homepage, December 2007



Source: ?

Figure 2: Obama's redesigned homepage



Source: ?.

phrasing generates more article reads (or “click-throughs”), then that choice can be made permanent.

At large internet firms virtually every product design feature is subject to A/B testing. From the number of search results displayed for per page by Google, to the precise layout of Amazon’s check-out page. This form of product development is a key engine of profits in the web-based industries today. Although such testing also raised important ethical issues (see **?** for reporting on a controversial Facebook experiment).

? calls A/B testing a “revolution”. In one sense he is correct. The nature of web-based platforms makes scientific testing of the minutiae of web-page design easy (at least in theory). Consequently the scale of A/B testing is internet industries is breathtaking. **?** reports that in 2011 Google ran over 7,000 A/B tests on its search algorithm alone.

Unsurprisingly, however, the “revolution” label is largely a misnomer. **?** discusses an agricultural trial in the late 18th century conducted by Arthur Young. While James Lind, the Scottish Physician, is often credited with having conducted the first clinical trial, again in the 18th century, showing that scurvy could be prevented by consuming citrus fruits. Both these experiments were, essentially, A/B tests. **?** discusses work done by the famous Chicago economist, Milton Friedman, on evaluating alternative projectile designs for the U.S. Navy during World War II. Friedman even uses the A/B terminology, describing the testing problem as follows (as quoted by **?**):

> The Navy has two alternative designs (say A and B) for a projectile. It wants to determine which is superior. To do so it undertakes a series of paired firings. On each round it assigns the value 1 or 0 to A accordingly as its performance is superior or inferior to that of B and conversely 0 or 1 to B. The Navy asks the statistician how to conduct the test and how to analyze the results.

In this course we will learn how to design and conduct simple A/B experiments. The modern statistical theory of these experiments dates to the work of Jerzy Neyman’s 1923 dissertation and Fisher’s 1925 book *Statistical Methods for Research Workers* (**??**). Both these works were motivated by problems arising in the analysis of agricultural experiments. Both Fisher and Neyman played foundational roles in the establishment of the modern discipline of statistics. This is especially true for Neyman, who founded the Statistics Department at the University of California - Berkeley, the importance of which for the shape of 20th century statistics can not be overstated.

Although the theory of using experiments to support decision-making is well-established, many consequential decisions are made, more or less, blind. For example, school boards in the United States routinely enter into costly contracts with textbook publishers, armed with no real knowledge about the effects of these choices for student learning outcomes. Governments routinely pass laws with only limited evidence on their likely effects. This type of decision-making can be costly and, in extreme cases harmful (imagine if bloodletting had been exposed to rigorous evaluation earlier).

Evidence, which can be used to guide decisions, can come in many forms. It can come from reflecting on past experience, deliberation with experts, focus groups, informal experimentation (i.e., trial and error) and so on. These forms of evidence are valuable, and each of us likely use them to make decisions about our own lives on a regular basis. Here, however, we will talk about generating and interpreting formal *statistical evidence* in support of, or against, specific policies. The generation and use of statistical evidence should be of interest to any large organization that routinely makes expensive decisions.

Questions to which statistical evidence can be brought to bear upon include:

- How can UC Berkeley design its webpage to maximize solicitations of application materials?
- What types of bean varieties grow best in a lowland humid tropical environment?
- What hillside farming practices improve soil conservation?
- Are there cost-effective policies for encouraging large land owners to maintain forest-cover on their land?
- What types of educational inputs (textbooks, computers, smaller classes) are most effective for improving student outcomes?
- What types of educational pedagogies, methods of teaching training, are most effective?
- What policing activities lead to the greatest reductions in crime and citizen perceptions of safety?
- What shift structures best balance employee satisfaction and factory output?
- How do loan terms affect consumer demand and default rates?

Useful statistical evidence requires much more than data and computation. Issues of research design and measurement are paramount, and will be emphasized in this course.

Further resources

1. World Bank Impact Evaluation Blog

2. Poverty Action Lab Reseach Resources

© Bryan S. Graham 2019