Web Performance Today

radware | strangeloop
Now part of Radware

# Big Data vs. Big Enough Data

*6 Mar 2013*

These days, there’s a lot of excitement around big data, and for good reason. It gives companies unprecedented power to harness customer information and increase their competitiveness in a time when the ability to compete globally has never been more important.

I have a lot of smart friends — some, like Eric Goldsmith, Cliff Crocker, and Buddy Brewer, who have been kind enough to come on as guests for my podcast — who are working with big data in meaningful and important ways.

But inevitably, where there’s excitement, there’s also hype. The tech community loves a new altar to worship at (and I’m putting up my hand here as well), and “big data” is  the official shrine of 2013.

I’ve also noticed a growing conviction that, given the choice between grabbing all the data or grabbing a sample of the data, we should always choose to work with all the data. The problem with this conviction is that it divides companies into two groups: the data haves and the data have-nots. If you don’t have access to billions and billions of data points, then there’s an understandable sense of frustration at being left behind.

In today’s post, I want to do two things:

- talk about when it’s okay — and possibly even better — to use big enough data, rather than big data, and
- as a caveat to the point above, explore the question of how big is big enough by using a recent example from our work here at Strangeloop.

## When is “big enough” good enough?

In TechCrunch a couple of months back, there was a really great interview [http://techcrunch.com/2012/11/25/the-big-data-fallacy-data-%E2%89%A0-information-%E2%89%A0-insights/] with Dr. Michael Wu, Principal Scientist of Analytics at Lithium. It may seem funny that I’m using it to argue for big enough data, because the thrust of the interview is that we actually need to look at even larger data sets, but I think the two arguments can peacefully co-exist.

Dr. Wu says:

> *While data does give you information, the fallacy of big data is that more data doesn't mean you will get "proportionately" more information. In fact, the more data you have, the less information you gain as a proportion of the data.*

In other words:

- Are massive data sets going to ensure that your insights are statistically relevant? Definitely yes.
- Are massive data sets going to deliver a proportionately massive number of amazing insights? Probably not.

In my opinion, there's one scenarios in which it's fine use "big enough" data:

### To generate a hypothesis to be further tested by bigger data.

I had a great chat with Eric Goldsmith a while back, where he gave the best explanation I've heard for how and why to mine big data. His mantra is "Mine the data for correlation and then experiment for causation." While Eric didn't specify the sizes of data sets he uses for mining and experimentation, I'm going to take the liberty of borrowing his mantra to offer this advice to anyone looking to make their data mining process more agile:

1. Start with a smaller (but still statistically significant) data set.
2. Identify trends.
3. Develop hypotheses.
4. Look to your larger data set to test your hypotheses.

These four points are all good and sound easy, but you still need some statistical significance. For example, variance can affect data set size, and so can the number of variables to analyze and correlate. "How big is big enough?" is the crucial question, which leads us to the second part of this post…

## So, how big is big enough? An experiment in finding the sweet spot.

Today I'm sharing just one example of how we answered this question here at Strangeloop. While it doesn't totally conform to the four points outlined above, it does demonstrate how we figured out how big was big enough in a specific scenario.

### Objective

This was an experiment conducted by Ken Jackson, one of our senior software engineers, to compare and analyze the effects of varying numbers of WebPagetest runs for a real customer's site (whose name I can't share, for obvious reasons) to show before-and-after acceleration results. **The goal was to test the assumption that 10 test runs is enough to deliver enough data for us to draw meaningful conclusions, given the amount of variance in the data we were collecting. It's important to note that we were looking at just one variable: load time.**

### Methodology

1. Using a WebPagetest private instance, gathered data for 100 runs, both treated and untreated, on the site's home page. This generated the baseline for comparing the other tests.

2. Fed that data into a statistical resampling exercise that simulated 10,000 treated vs untreated tests with 3 runs, 10 runs, and 30 runs.

3. To reduce variability as much as possible, used first-view only, a single browser (IE9), a single location (San Jose), and a single connectivity (cable).

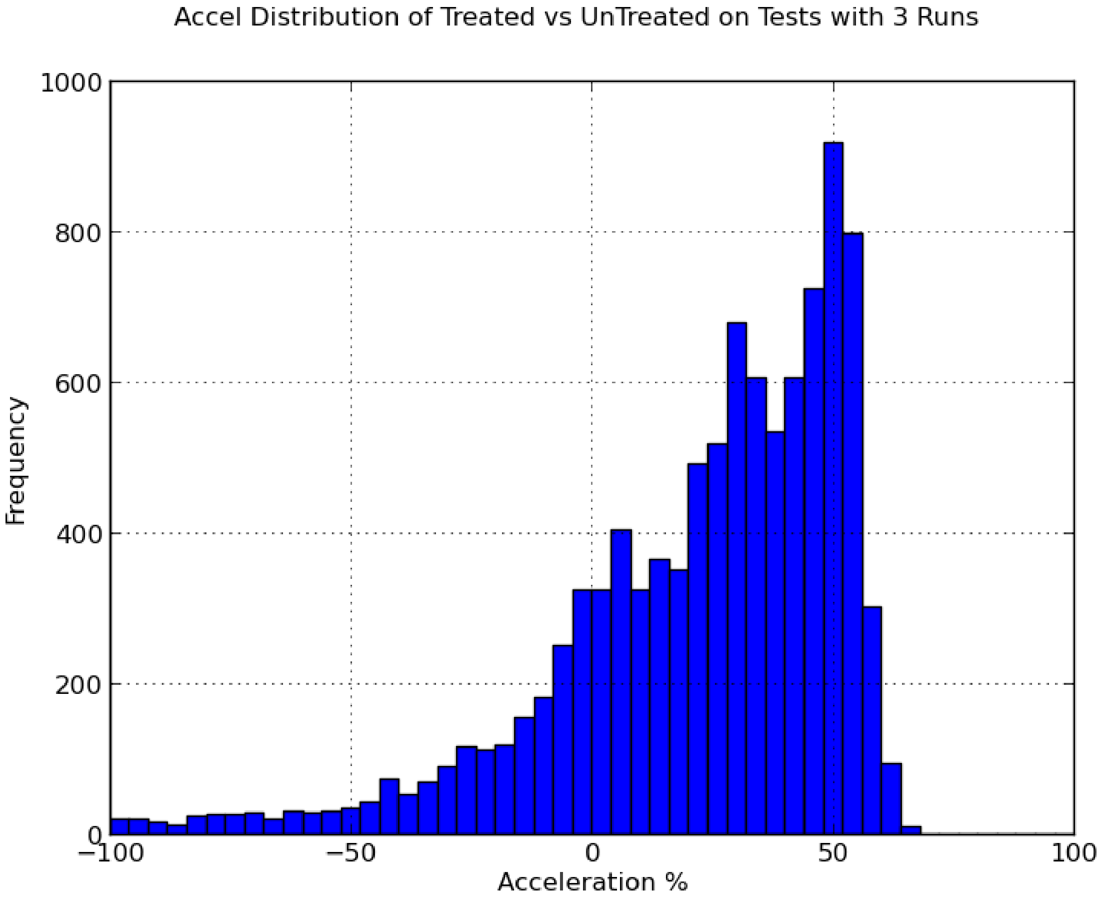
## Results

Overall, the acceleration from the median doc complete times for the 100 runs was 31%. So, as already stated, the goal was to find out which smaller set of runs, if any, yielded similar results.

The results are shown as a series of histograms below.

## 3 RUNS

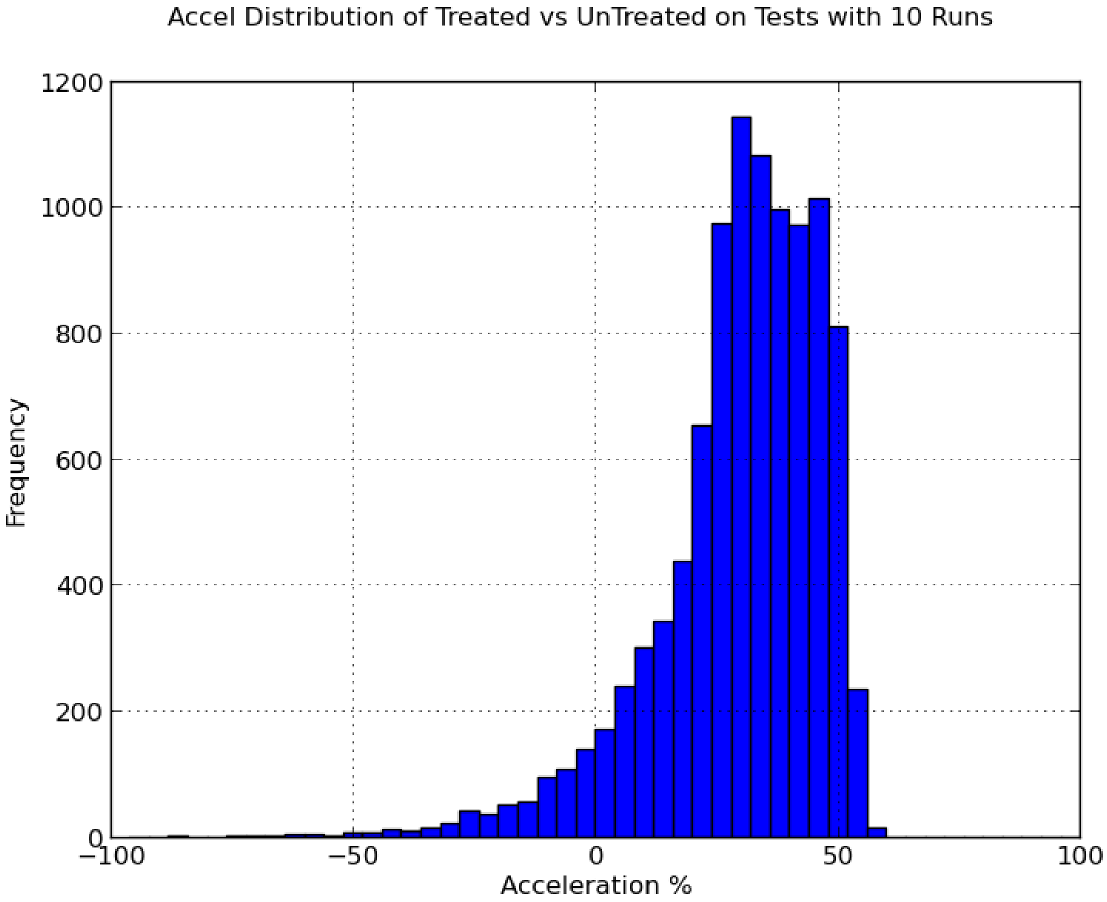
The first histogram below shows the results for tests with 3 runs. The height of each bar indicates the number of tests that gave a particular acceleration value. For example, the peak in this histogram shows about 900 treated vs untreated tests that resulted in a 50% acceleration. That seems a little high compared to the 31% obtained from running 100 tests, so already the 3 run results seem suspect.



The most striking thing about this graph is the number of bars that are less than 0% acceleration. There are cases where a 3-run test of treated vs untreated will show a negative acceleration even though we can be confident that the acceleration on this page close to 31%. In rare cases, a 3-run test can even show an acceleration of -100%.

## 10 RUNS

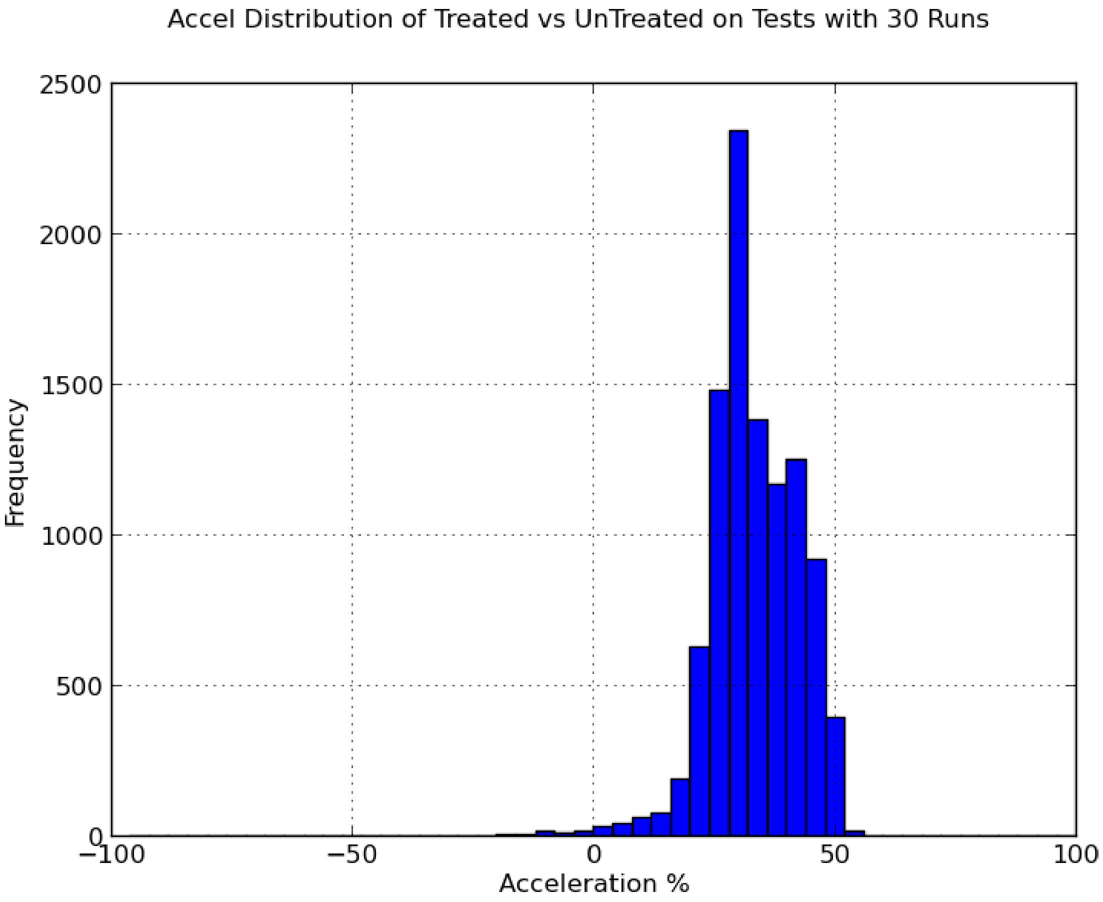
The next histogram shows the results from simulating 10,000 treated vs untreated tests with 10 runs. The peak is closer to 30% and overall the shape of the curve is much narrower.



However, there are still cases where the acceleration measured from a 10-run test would be negative – by as much as -50%. The width of the curve is still quite wide, so we would expect 10-run tests to often vary between 20 to 40% acceleration.

## 30 RUNS

The histogram for 10,000 simulated 30-run tests shows a stronger peak at 30% acceleration and the overall curve is narrower still.



Conclusions

The results indicate that up to 30 runs are often needed to reliably demonstrate acceleration value — even on a site with 30% acceleration and fairly low variability. As Ken pointed out when he wrote this up in our internal blog, even more runs would be needed on sites that have higher variability and/or lower acceleration.

Why? The more variance there is in your data, the more data you need. Many big data projects are about capturing hundreds of variables and comparing across many of them. But if there's not a lot variance — if there's a small number of variables or correlations between elements — then you can get away with using smaller data sets.

## Takeaway

It's impossible to cover all the finer points of the science of data collection and analysis in a single blog post. Our findings are specific to this unique scenario and shouldn't be extrapolated to other scenarios.

Instead, I'm hoping this post will serve as an example of a real-world situation that made us ask "How big is big enough?" and then made us look for a way to answer that question. In this case, we learned that the answer is "More than we think", due to the amount of variance in the data.

So while we had to make our data bigger in order to make it statistically viable, we didn’t have to blow it up to capital-B-capital-D Big Data proportions.

Getting back to my point at the top of this post: **big data doesn’t mean wasting compute cycles and testing forever, and it’s not about collecting a lot of the same measurement well past statistical viability.** Depending on the complexity of whatever you’re testing and the rate of variance in your results, you may be able to find a point at which you’ve controlled enough variables and have enough measurements that you don’t need to keep testing.

*By: Joshua Bixby*