Are Tweets That Invoke Data More Likely to Be Shared on Twitter?

Justin Murphy

3/20/2018

By Justin Murphy. Code is available in the Github repository.

Does the invocation of data effect the likelihood that a message will be shared? Does data generate trust, helping messages to propagate? Or is data helpless to counteract distrust and political partisanship, because it is seen as only a higher form of deception? We tend to think of data as a way of supporting claims with evidence but we also have the popular expression that there exist "lies, damned lies, and then statistics..." These views imply different and mutually exclusive causal models, so which one is correct?

To better understand this question we have been experimenting with many approaches. Here we would like to share a peak at one way we are approaching the question.

I began by choosing two highly salient and contentious political topics: The Trump/Russia issue, and climate change. I began conducting searches for tweets that mention these keywords plus any one of the following ("graph," "chart," "map," "poll," "study", "data," and "figure") and any one of the following ("shows," "reveals," "indicates," "suggests," "proves"). An obvious drawback is that this certainly fails to capture many data-driven messages in the diverse styles in which they might be composed — that's why we're also experimenting with machine learning techniques that are stronger on this front, though with other drawbacks. The virtue of this method is we know exactly what we're getting, and there's no complicated machinery to question or critique. It appears to work very well on the specific goal of minimizing false positives (i.e. we do not capture all data-driven messages, but we are relatively unlikely to code as data-driven messages that are not in fact data-driven).

The graphs below give a sense of the basic patterns we're observing in this approach. None of the following visualisations control for confounding variables, which will of course be a major question as we develop our analyses. Note that because social media sharing is massively skewed count data (a very small number of messages get a very large number of shares), this complicates how we should interpret relationships between variables. For instance, we cannot use traditional standard errors and confidence intervals. Also, we might decide to remove the most retweeted tweets, but there would still be serious skew in this type of data, and choosing a cutoff could be arbitrary. These are issues we will have to think about more moving forward, but for the meantime we can share some basic figures. We removed reply tweets and tweets that are themselves retweets, so here we're focusing on the number of shares received by original status updates.

Figure 1 suggests that in the context of Trump/Russia conversations, data-invocation is associated with almost one extra retweet.
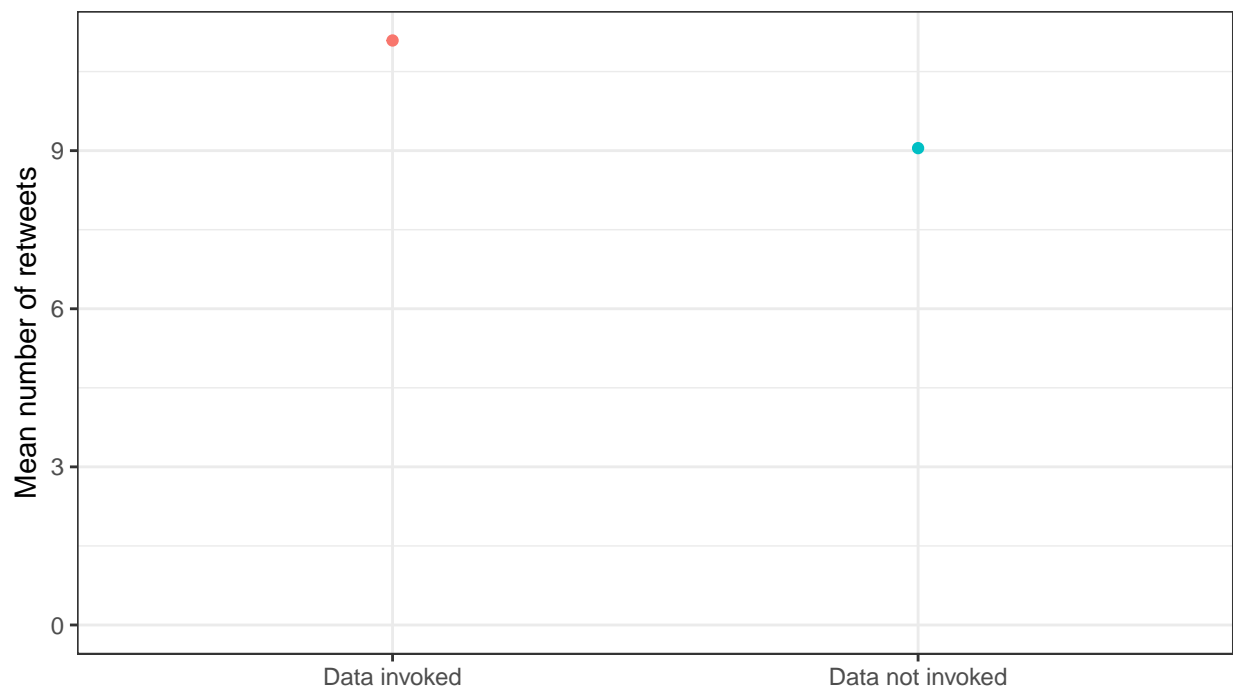
Figure 2 suggests that there is not really any association between data-invocation and the probability a message will be shared, in our sample of Trump/Russia messages.

Figures 3 and 4 show the same visualisations but for samples of tweets about climate change. Tweets that invoke data only receive a negligibly greater quantity of retweets, although tweets that invoke data are slightly more likely to be shared.

For the time being, I would not make any major conclusions from these data! Other than the basic facts of the in-sample bivariate relationships, we can't say too much more. I guess that's the beauty of research blogging–whereas fashionable online magazines and newspapers have to give you interesting and compelling results, research bloggers only update you on the incremental progress of a long and winding path. Our answer(s) to this question will become more clear as we account for confounding factors, collect larger samples, consider additional issue areas, and triangulate with different approaches.

Figure 1: Figure 1
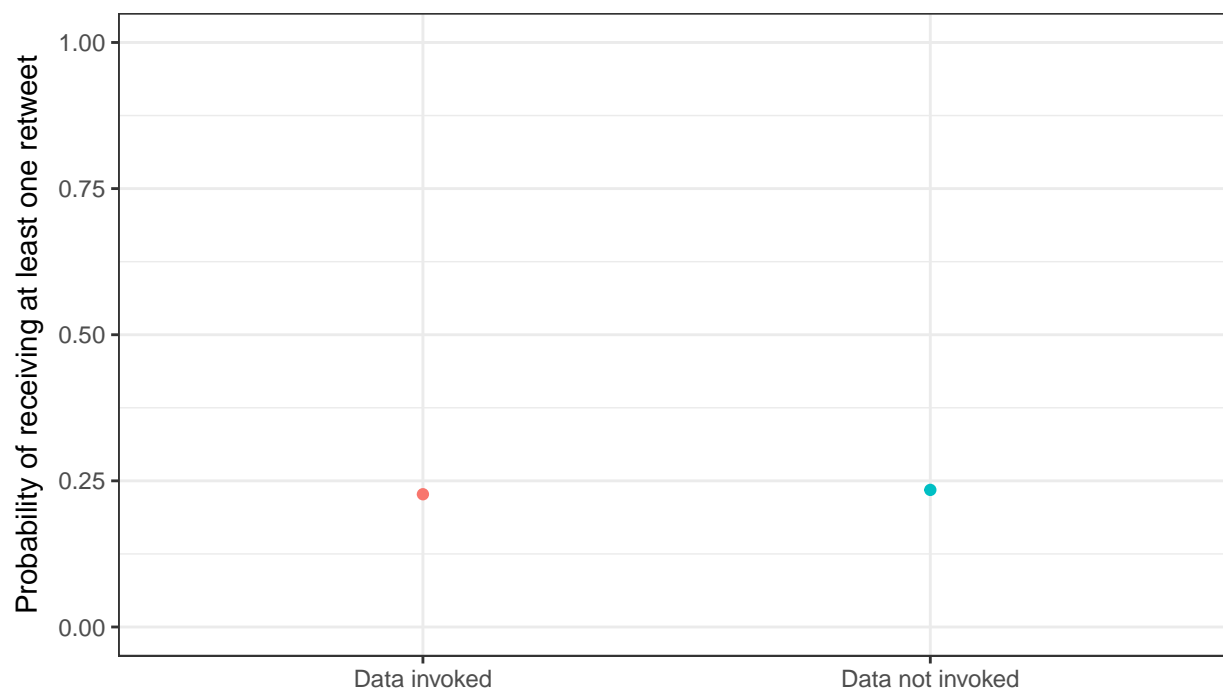
Does Invoking Data Increase the Probability a Message Will Be Shared at

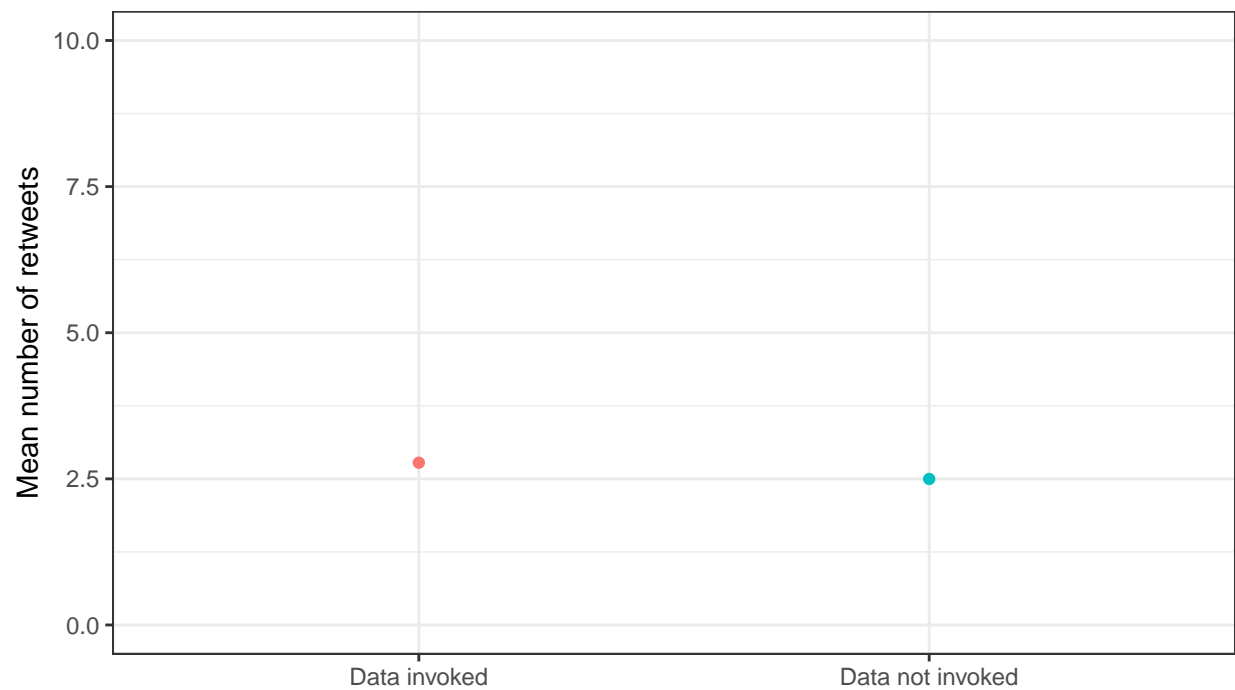The Case of Trump/Russia Discussions on Twitter (N = 11,593 messages)
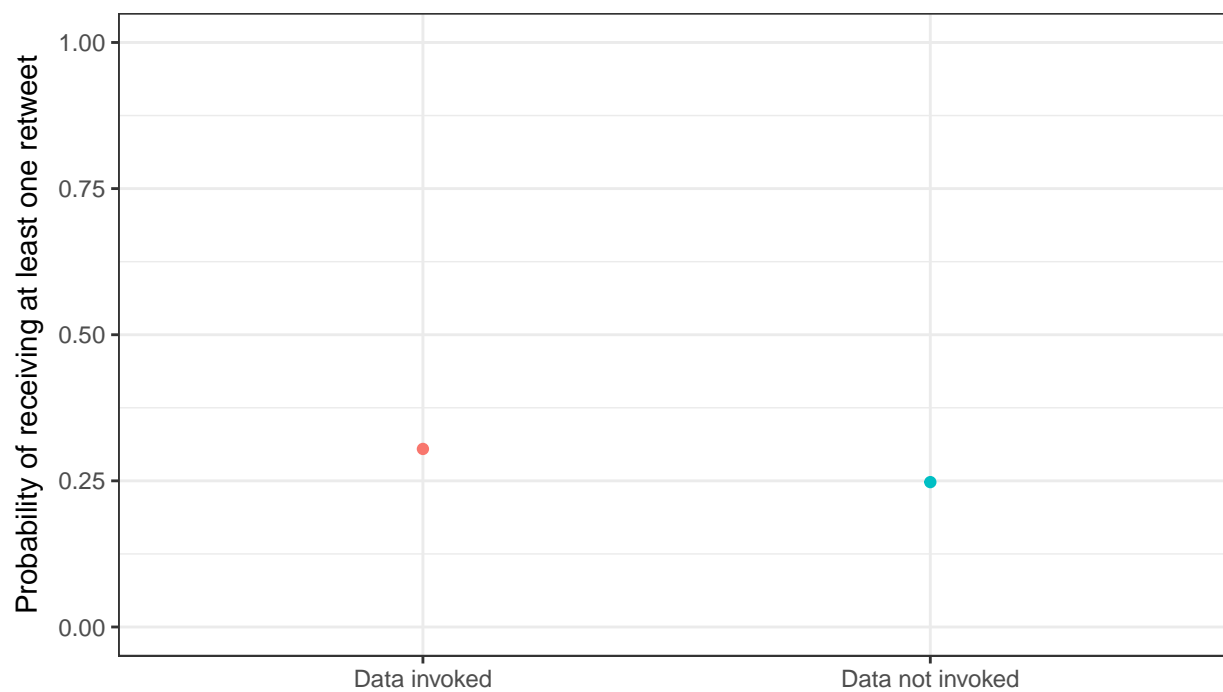
Datastories.co.uk, @datastoriesuk

Figure 2: Figure 2

Figure 3: Figure 3

Figure 4: Figure 4