Making lemonade from lemons (or spinning gold from straw): publishing papers from crappy data

Robert McCartney
University of Connecticut

February 7, 2005

Motivation

We often collect data with the best intentions: we are interested in how students learn, or which pedagogical techniques are most effective, or how to better retain students in computer science. Sadly, sometimes it doesn't work, and we find that we cannot do what we want with the data we collected, because the data suck[1]. The questions have obvious biases, or we didn't take a random sample, or we lost some of the files—there are lots of possible reasons why data may seem to have little value.

On the other hand, we *need* published results. Department chairs, deans, university administrators, people at agencies that fund your work: they all expect a long list of papers in respected journals and conferences. If we expect to go to conferences, meet people, and become important, we need to be noticed, which means presenting results[2].

Research questions:

1. How can I turn my crappy data into brilliant papers.
2. (less ambitious) How can I publish papers from my less-than-perfect data without being embarrassed.

Ground rules

The goal, then, is to turn bad data into good (enough) papers. This is to be done without obvious cheap tricks, such as choosing phantom co-authors

[1] Notice the correct usage of the plural
[2] Asking self-serving questions also will get you noticed, but not favorably

with reputations that overly impress reviewers (see, for example, [1, 2]). Making it more difficult, we are committed to accurately disclosing our data and how they were collected.

Approach

The first step in this study is to review the appropriate literature. We will examine papers to assess the quality of the supporting data and technique, looking specifically for those that "do the most with the least"—papers that are often referenced, but based on the flimsiest of data. We will then try and isolate the factors that contribute to the success of these and develop a general approach to effective publishing. To evaluate this approach, we will apply it to a number of less-than-perfect datasets, writing and submitting papers to a variety of venues.

I will need some collaborators with bad data, although I have some extras that I am willing to share.

What evidence would convince me?

If the papers *all* get published, I will be convinced that we have found effective ways to get papers in. If they are named "best papers", then I will be convinced that they are brilliant. If not all get in, but our accept rate is significantly higher than average, I will be convinced, but not as sure. If we use different methods for different papers, then we might determine which are "winners".

The other issue is whether I am embarrassed over getting in. For conference papers, if I am willing to go present the paper instead of co-author, I'll be convinced. If a prestige journal–hey, the reviewers must have liked it right?

Meta-question

Suppose we do this, but the data we get (papers submitted, papers accepted) are bad. Can our techniques turn *that* data into a publishable paper?

References

[1] God, Moses, and Robert McCartney. Ethical considerations in CS1: Ten rules. In *Proceedings of the Computer Hardware Extraordinary Ethics Society (CHEES 05)*, pages xii – xx, 2005.

[2] Robert McCartney, Bertrand Russell, and Ludwig Wittgenstein. Logical and philosophical bases of computer science education. *Info. Angelorum Hominibus*, 666(1):213–225, 2004.