File 20100210.0755: Notes from meeting with Dr Martin this morning:

I did the exercise recommended by [1] and it improved my thesis problem. Here is the current question:

> Is there a difference in the post-CT&E software defect rate (measured in terms of the number of findings of various severity) between different versions of the same system in subsequent rounds of CT&E by different DAAs?

I think this thesis question passes the tests in [1] for a good thesis statement: there are three clearly defined possible outcomes, each one tells us something important if it turns out to be true, I can think of persuasive arguments why each possible outcome is likely, but I don't know which way to bet. If I did, then the problem would be trivial.

I am concentrating this week on the teaching/lecturing aspect of the talk I have to give next week: how to explain things clearly, communicate to the audience what an interesting problem it is, and how to show at every step that my methodology is both necessary and sufficient to answer the question.

I will send Dr Martin a picture of my slides by Friday.

Three possible outcomes from the problem (to come out of the data collection and analysis box):

1. More findings: this could occur because of:

   - better tools
   - more eyes looking at the product
   - better criteria/standards; eg NIST SP 800-53 replacing DCID 6/3
   - new attacks being developed all the time

   I have an example of this occurring in my first case study: a system that had been evaluated many times by NSA (and they found things, too), but the first time CESG looked at it, they found a new vulnerabilty that had been overlooked for years, because they used different tools or looked at it in a different way.

2. Fewer findings: this could occur because of:

   - The software gets better with more testing.

   This is perhaps the way the smart money would bet; what you'd expect if the system is working the way it is supposed to be working.

3. No change in the number of findings (NOTE: what about the derivative? Are we looking at number, rate, or trend?) This could occur because of:

   - ?

   I need to think about this case and whether it is a subset of the 'fewer findings' case. I think it might be degenerate.

Cross domain systems are a good example of systems that get repeatedly CT&E'd because they go into a new situation with new data owners nearly every time. What about safety-critical systems, however? Do they get repeatedly tested under new criteria because they are being installed in a new aircraft? Do different countries have different DO 178-B/C criteria? What about the difference between DO 178-B and -C, does it trigger a new round of testing for already deployed systems? Probably not, but those existing systems get upgrades and new features, and those new features get tested under the current standards—or do they?—each year when they come out.

Notes: findings is a multidemensional metric: Category I, II, II, IV where Cat I findings are show-stoppers and Cat IVs are typically documented and never fixed.

Cost, measured in £is another metric, but one that is going to be hard to gather data for. I have not seen it clearly broken out in budgets and project managers are loathe to disclose.

Dr Martin, playing Devil's Advocate, asked: how do you know you're not just measuring noise? Is it feasible to conduct enough experiments? The counter-argument is that in this field, people don't get to conduct enough experiments.

Statistical significance:

I can calculate the number of experiments necessary to achieve a 95 percent confidence interval, in the spirit of thoroughness and demonstrating to the external examiner that I recognise the need for and know how to do a confidence interval analysis, and have gone to the trouble in my dissertation. Put it in an appendix.

I can also calculate the expected confidence interval given the experiments and cases that I have. Dr Martin pointed out that in this field, people don't get to do enough experiments to really establish a traditional 95 per cent confidence interval. But I can do the statistical maths to show it anyway, even if it turns out to be 20 percent, I can show that I recognise the need to do it, know how to do it, and bothered to do it.

Dr Martin says I am making progress, but I need to watch out for too much introspection. Showing thoroughness and looking at the problem from all angles and understanding it is all well and good, but there needs to be results as well.

Next meeting: Wednesday after Reading Group. After that I get on a plane and fly out.

Problem: I fly out a day earlier than I thought. Have to reschedule the meeting!

Call ended 0708.

References

[1] Gordon Rugg and Marian Petre. *The Unwritten Rules of PhD Research*. Open University Press, 2004.