REDJACK

What does it mean to do science? Are we doing it?

John McHugh

University of North Carolina

RedJack, LLC

Why should we care about good science?

- Aside from being the proper thing to do,
  - Bad science impedes progress. It's not just a Computer Security or Computer Science problem
    - Lincoln 98-99 data is still damaging IDS work
    - Publication of results not achieved in SIFT has had a long term effect on the formal methods community
    - "Cold Fusion" diverted effort in physics for years
  - Bad/badly-reported science is not a suitable basis for building blocks.
    - Not clear how possibly-useful results were obtained
    - Refinements/improvements on existing work difficult
- Nonetheless, we tolerate bad work/reporting

What is Science, anyhow?

- Dictionary definition (dictionary.app, OS X)

  science |ˈsīəns| noun

  The intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment

  – Usually the study is undertaken to answer some question about the objects of the study.

- The operational aspects here are:
  – systematic study
  – observation and experiment

Systematic study

- *Systematic* implies an orderly and well-defined process.
  - for a given study, it should be possible …
    - to describe the study process clearly, so it can be understood and followed by others.
    - repeat the study process, obtaining similar results with similar objects of study
    - organize the findings, so as to develop broader insights into the objects or areas of study

Objects of study

- The dictionary definition is couched in terms of structure and behavior of the *physical* and *natural* world.
  - We need to expand this to include artifacts, such as computer systems and networks, as legitimate objects of study.
  - This is necessary because we have created systems that are so complex that their behaviors are not immediately obvious from examination of their designs or implementations.
  - From a practical standpoint, the theoretical analysis of such systems to determine properties of interest, e.g., those related to security, is not feasible.

Observation

- Observational studies help to characterize or explain properties or behaviors of the studied objects.
  - objects of study are viewed passively without intervention
  - lead to theories or generalizations
  - explain differences among individuals or groups studied.
- Good observational studies will record all aspects of the study objects and their environment that might affect the answers to the question that drives the study.
  - In the absence of a driving question, *everything* may be relevant.
  - Conversely, a narrowly-focused question may limit the detail to be observed, but fail to inform related questions.

Kinds of observational studies

- Case studies / Snapshots
  - systematic description and cataloging of what was observed at a place/time, e.g., Darwin's Beagle voyage.
- Cross sectional / Population
  - record distributions of observed characteristics among individuals in a population – e.g., defects in ICs from a fab
- Longitudinal
  - Differences within a population over time, e.g., Framingham study of risk factors for heart disease. (6+ decades)
- Comparative
  - Some groups or individuals are better equipped to deal with a given situation than others. Why?

Experimentation

- Experiments alter the study environment in measurable ways; the effects of the alterations (interventions) on the objects of study are observed.
  - Usually driven by a hypothesis that predicts a specific change, or by a theory that conditions observed behaviors on the observations.
  - Objective is to explore the effect of changing conditions on the observed behaviors, to provide evidence that supports or challenges the behaviors predicted by the theory.
  - Experiments are often classified by the degree of control that can be maintained in isolating the effects of the intervention(s).

Experimental Flavors

- Case studies
  - A context is established, and effects observed. Fairly weak as a basis of comparison; may be the best we can do.
- Controlled studies
  - Intervention (experimental) / non-intervention (control) results compared.
  - Random assignment
    - Subjects randomly assigned to exptl/ctrl groups. All factors except intervention identical between groups
  - Balanced assignment
    - Subjects assigned to groups so as to balance factors that might affect outcome among the groups.

Validity, Reproducibility, …

- It is also important that experiments be valid.
  - Absence of confounding factors that give rise to misleading results.
  - Applicability outside the experimental setting.
    - Overly simplified experimental settings often produce results that are not applicable in the wider world
- Repeatability and reproducibility are also hallmarks of good experimentation
  - A deterministic system should be trivially reproducible
    - Given inputs always give the same result – uninteresting
  - Similar inputs should give similar results – within population variation and measurement errors, even for non-deterministic systems.

Etc., etc.

- More can be said about types of observations and experiments, but courses and texts abound. The most important thing is to realize that convincing results require:
  - An approach that is appropriate to the subjects and objectives of the study, and …
  - The strengths of the conclusions are functions of:
    - the approach / design,
    - the degree to which factors that can affect the results are controlled or accounted for,
    - the number of observations / subjects, and
    - whether the physical world is consistent with the effects measured; e.g., light travels 1ft in about 1ns.

So, how do we do good science?

- Much of the *how* follows from *careful* consideration of the definitions.
  - Document every step carefully – there are reasons for the lab-notebook practice in the physical sciences.
  - Think carefully about the questions you are asking. – You want to make sure they are precise, possible to prove (or falsify) and should be answerable using available resources.
    - This leads us into both physical limitations and experimental power analyses, given expected effects.
  - Pick an observational or experimental approach that is as strong as possible, given the topic.
  - Narrow the question if necessary
  - Pay attention to details and don't make things up

Convincing others we did good science

- In a bit, we will discuss some papers that ought to represent good science, but fail to convince.
  - Making convincing arguments is a skill that can be taught, but is often not.
  - Prescribed structures are liberating, but often not used or understood.
    - In "The Mythical Man Month," Fred Brooks notes the liberating effect of structure, as it allows effort to be spent worrying about content, not form.
    - By imposing a structure on the reporting of results, these also impose a structure on the scientific process
    - Roy will address issues of structure in a later talk.

Doing and convincing become confounded

- Even if we do good science, we have to do it in a way that will convince others.
  - Exploratory work is often haphazard, especially when the unexpected occurs.
    - There is nothing wrong with this, but it may be necessary to repeat the work in a systematic fashion
      - to confirm the observation and
      - to ensure that we understand possibly confounding factors and
      - to assemble the convincing narrative.
  - Parnas and Clements note that presenting results through the lens of a well-structured process is useful.
    - Even if the results were not obtained that way
    - This is not good science, but may guide the repetition.

Transparency – a historical note

- Charles Darwin's name is most closely associated with the theory of evolution
  - Darwin was not the first to set forth such a theory
  - His work was the first to gain acceptance because
    - His meticulous observations provided evidence to support the theory and
    - His writings clearly exposed the chain of reasoning that led to his conclusions.
    - The combination of clearly-recorded observations and clearly-articulated reasoning provided most readers with a convincing argument in favor of the theory

Is that all?

- No - the devil is in the details. Once you have the questions, you need to turn them into a plan, execute the plan, and report the results. But …

"Do you know how to do good science?"
"No, but if you'll hum a few bars, I can fake it."

A bit of structure

- The next few slides provide a general framework for the good reporting of good science.
  - In most cases, the reports, conference and journal papers, and the often-overlooked technical report are the only basis that we have for judging the quality of the science.
  - While it is possible to do science well, and report it poorly, the reader of a poor report cannot know whether or not the science was done well.
    - If the quality of the work is not easily seen from the report, the conservative assumption is to assume that the work was done badly.
      - Fortunately, bad reporting is easier to fix than bad science.
- THINK first, DO later.

Writing for the Reviewer

- In order present our results, we must publish.
- In order to publish (other than TRs?) we must pass a review process.
- The reviewer's task can be simplified by providing a coherent, well-structured narrative that reflects the scientific process.
- The narrative should convince the reviewer that the reported work meets our criteria for good science
  - Clear statement of the questions investigated
  - Complete description of methods, equipment, instruments, etc.
  - Transparent design / decision process
  - Complete and thorough analysis and results

Reading Critically to Access Science

- Just as the writer has an obligation to express the science clearly and coherently, the reader / reviewer has an obligation to read critically.
- What is the question being addressed?
  - Is it clearly stated? Is it something I can review?
    - What is its import? (Do I or the community care?)
- Are the methods, equipment, instruments, etc. adequate and appropriate?
  - Could I carry out the work using the paper's description?
- Is the description transparent?
  - Do I understand why things were done as reported?
- Is the analysis appropriate and convincing?

This is sort of a dance

- The researcher / writer engages the reviewer.
  - If done well, the review process helps to improve both the presentation of the current work and the execution of future work.
  - Conference reviews often short circuit the process
- Reviewing is both hard and time consuming.
  - A lot of inadequately described or simply bad work gets through.
  - Criticism is not accepted as a legitimate topic for publication
    - Contrast with other fields
- We present critical reviews of some recent papers

A case study

- We evaluated papers from a recent cyber security conference. Candidates were:
  - Oakland(Security and Privacy)
  - ACSAC
  - USENIX Security Symposium
  - ACM Computer & Communications Security (CCS)
- None of these had an explicit directive in the call requiring papers to be scientific.
  - Hot SoS (Hot Science of Security) was chosen, because its call for papers had explicit requirements for scientific reporting.


WTF
Hot
SoS

Hot SoS call for papers - 1

The following are excerpted verbatim from the Hot SoS call for papers

Scope.

The Science of Security considers not just computational artifacts, but incorporates the human, social, and organizational aspects of computing within its purview.

Approach.

The Science of Security takes a decidedly scientific approach, based on the understanding of empirical evaluation and theoretical foundations as developed in the natural and social sciences, but adapted as appropriate for the artificial science (in Herb Simon's term) that is computing.

Hot SoS call for papers - 2

Emphasis on Science: Required section in each submission.

The key motivation behind Hot SoS is to bring out and promote the science underlying security. Therefore, we require that each submission include a section called "The Science" in which the authors should **describe in what ways their contribution constitutes science**. We interpret science in the broadest sense of the systematization of knowledge to uncover foundational principles through theory-driven inquiry. Thus methods reminiscent of the natural, life, social, or behavioral sciences are all acceptable. Our motivation behind asking for this section is not to impose any ideology on researchers, but to give a spotlight to, and thus promote, the science of security.

Dividing the Hot SoS papers

- Long *vs.* short paper
- Science statement
  - Is the science statement any good?
- Experimental *vs.* theoretical
  - Fully evaluated only experimental papers
  - Is the experimental work any good?
- 47 papers
  - 12 full papers
  - 35 short papers (2 pages) (Only 2 with a science section -- #18 and #30)
- Evaluated only full papers

Hot SoS paper evaluation criteria

- Long papers only (there are 12 of them)
- Classified 6 experimental, 4 pure theory, 2 hybrid (limited experiment / simulation)
  - Usually obvious from the paper that the objective was experimental or theoretical, but …
  - Few papers had a clear problem statement or hypothesis, which would have made the task much easier.
- The 2 hybrid papers are problematic
  1. new implementation of an analysis system with results of a few examples as a sanity check
  2. model with validation by simulation
     - Could it be implemented on the real network?

Restricted Study to Experimental Papers

- Fully evaluated experimental papers only
  - Did science statement meet CFP criteria?
  - Did research description pass muster?
    - Complete, transparent?
- What about the theory papers?
    - Theory papers likely to state assumptions.
      - Some models based on over-simplified assumptions.
      - Examined for link to empiricism or the real world
    - Theory should lead to experimental work
      - Almost completely lacking, either for theory validation or for investigating consequences.

Science section criteria

- Is there a numbered section that addresses the required “emphasis on science?” If not, is there a section or paragraph anywhere that addresses the required “emphasis on science?”
  - We consider “the authors should describe in what ways their contribution constitutes science” to be the operational imperative. This is qualified by CFP language.
    - “We interpret science in the broadest sense of the systematization of knowledge to uncover foundational principles through theory-driven inquiry.”
    - "Thus methods reminiscent of the natural, life, social, or behavioral sciences are all acceptable."
  - Only one paper met these criteria, and its science was bad.

Criteria for good (experimental) science

- We used the criteria developed earlier in the presentation to evaluate the science as presented in the experimental papers.
- The problems are all over the map; only one of the experimental papers presents the process and results in a way that can withstand scrutiny.
- The following table summarizes the analysis of the full papers.
  - The papers were examined by both Roy Maxion and myself. We reached consensus.
  - We summarize the kinds of problems we found on subsequent slides.

Hot SoS paper evaluation: summary

#	Theory	Exptl	Sci Stmt Present	Sci Stmt Good	Good/bad Science
1		Yes	No	N/A	Bad
2		Yes	Yes	0	Bad
3	Mostly	Limited	Yes	0	Decent
4	Yes		Yes[1]	0	N/A
5		Yes	Yes	0	Decent
6	Yes		Yes	0	N/A
7		Yes	No	N/A	Bad
8		Yes	Yes	0	Bad
9		Yes	Yes	Yes[2]	Bad
10	Yes		Yes	0	N/A
11	Yes	Simulation	Yes	0	Decent / N/A
12	Yes		Yes	0	N/A

Example problems with the science

- Confounds – results that could have been due to differences in training, selection, etc.
- Failure to show external validity – subjects selected via Mechanical Turk not representative of target population
- Failure to perform power analysis - only one paper has such, albeit post hoc; 500 subjects needed, 80 used
- Statistics fishing (do many tests; 1 might be significant)
- Ill-formed research statements
- Unexplained participant dropouts
- Bad/misnamed designs, e.g., “Controlled” w/o ctrl grp
- Poor and/or confusing presentation
- Decisions / actions not justified

Summary

- What does it mean to do science?
  - doing observation or experimentation …
    - in a complete and transparent fashion
    - driven by a clearly-stated objective or hypothesis
    - with appropriate methods, design, and analysis
    - to provide valid and replicable results
- Are we doing it?
  - The evidence from the published literature in our field is a resounding NO for the most part.
    - The Hot SoS papers should be exemplars of good science, but are not.
    - Other conference venues, like Oakland, ACSAC, and Usenix, seem to be as bad or worse.

I think there is a problem

- Even when good science is explicitly called for, the community does not do well.
  - Why?
  - What can we do?
  - What, if anything *should* we do?
- The remainder of the workshop will be devoted to answering these questions, and to developing ways to move forward.
- If you want to discuss any of these issues further, talk to me during the workshop, or contact me as

mchugh at cs dot unc dot edu