

Should Security Researchers Experiment More and Draw More Inferences?*

* With thanks to Walter Tichy's "Should Computer Scientists Experiment More?" (1998)

Kevin Killourhy
with Roy Maxion

Carnegie Mellon University
CSET 2011 (August 8)

Should Security Researchers
Experiment More and Draw More
Inferences?

YES!

Security researchers rarely conduct experiments and draw inferences

- 101 keystroke dynamics papers surveyed
- 80 papers evaluated a classifier

Comparative experiments:	43 / 80	(53.75%)
Inferential statistics:	6 / 80	(7.5%)

<http://www.cs.cmu.edu/~keystroke/cset-2011>

- Similar experience in IDS and Insider-Threat research

One-off evaluations confound detector and data

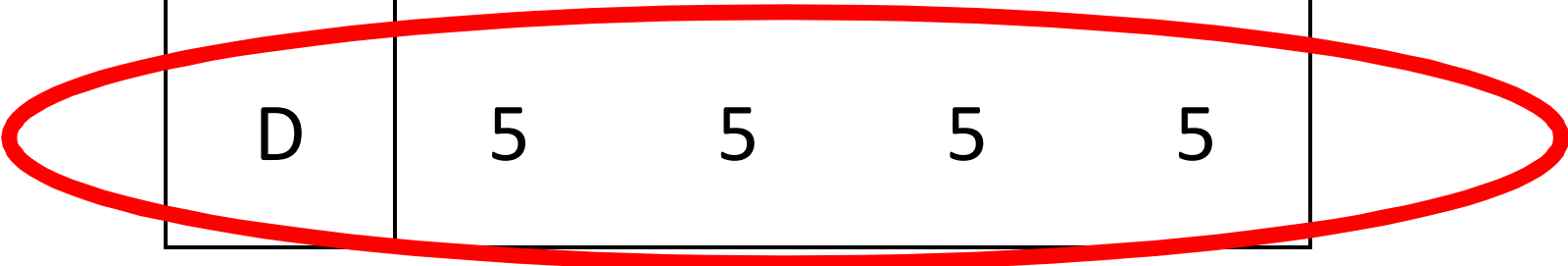
Researcher	Detector	Data Set	Error Rate (percentage)
Alice	A	1	20
Bob	B	2	15
Carol	C	3	10
Dave	D	4	5

One-off evaluations reveal diagonals of a matrix

		Data Set			
		1	2	3	4
Detector	A	20			
	B		15		
	C			10	
	D				5

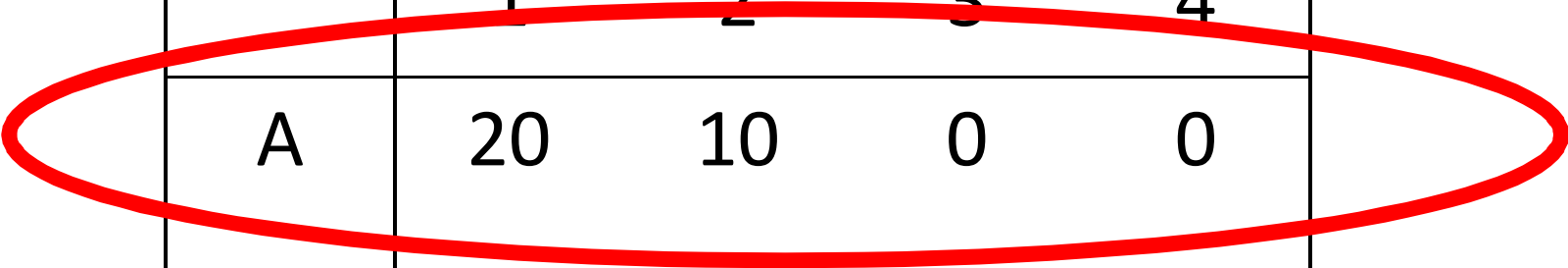
Case 1: No Data Effect

		Data Set			
		1	2	3	4
Detector	A	20	20	20	20
	B	15	15	15	15
	C	10	10	10	10
	D	5	5	5	5



Case 2: Data Effect

		Data Set			
		1	2	3	4
Detector	A	20	10	0	0
	B	25	15	5	0
	C	30	20	10	0
	D	35	25	15	5



Case 3: Data/Detector Interaction

		Data Set			
		1	2	3	4
Detector	A	20	10	5	15
	B	5	15	20	10
	C	10	5	10	20
	D	15	20	15	5

Which case holds for security research?

	1	2	3	4
A	Case 1: No Data Effect			
B				
C				
D				

	1	2	3	4
A	Case 2: Data Effect			
B				
C				
D				

	1	2	3	4
A	Case 3: Data/Detector Interaction			
B				
C				
D				

Keystroke dynamics:

	1	2
A	19.5	46.8
B	1.0	85.9

(Cho et al., 2000)
(Killourhy & Maxion, 2009)

Worm detection:

	1	2	3
A	0	1	1
B	3	0	2
C	5	5	1

(Stafford & Li, 2010)

Inferential statistics focus our efforts

Security technologies do not have *an* error rate; they have *many* error rates, depending on factors in the operating environment.

Keystroke Dynamics:

Worm Detection:

Malware Scanning:

- Operating system
- File format
- Packer
- Environment (home/office)
- Web browser
- User habits

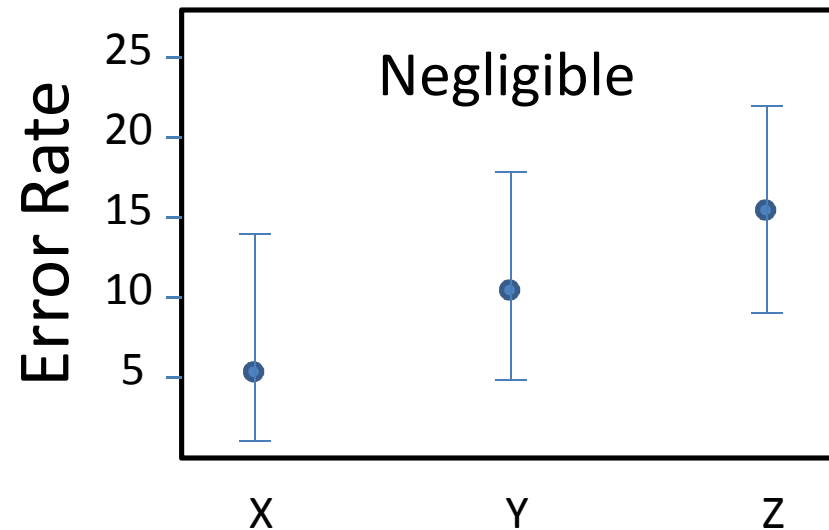
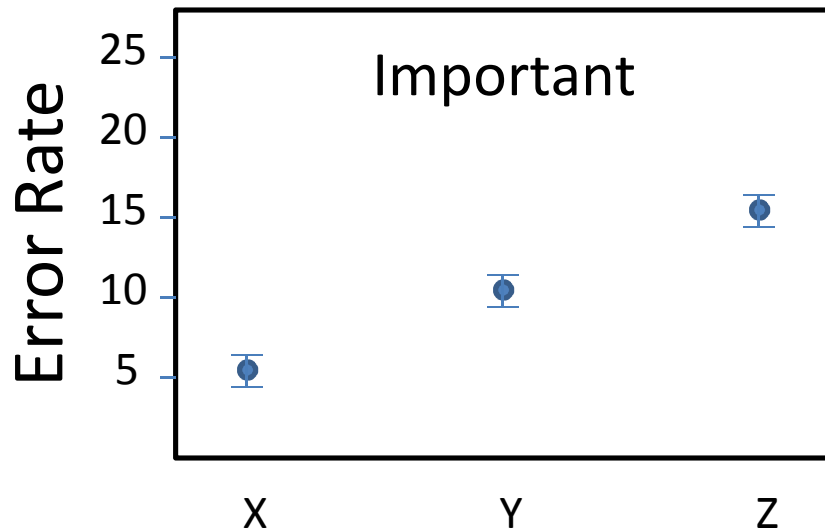
...

The number of potentially important factors can be overwhelming

Empirical averages only tell part of the story

Factor (value)	Error Rate (percentage)
X	5
Y	10
Z	15

Is the factor important or not?



Outline

What?

- Security researchers rarely conduct experiments and draw inferences.

So What?

- Current results are not very meaningful.
- They cannot answer important research questions.
- There is no direction for future work.
- A lot of research effort is wasted.

Now What? (Issues)

- [Gathering and sharing good data](#)
- [Establishing a standard methodology](#)
- [Security-specific challenges](#)
- [Changing the culture](#)
- [Beyond experiments and inferences](#)

Gathering and sharing good data

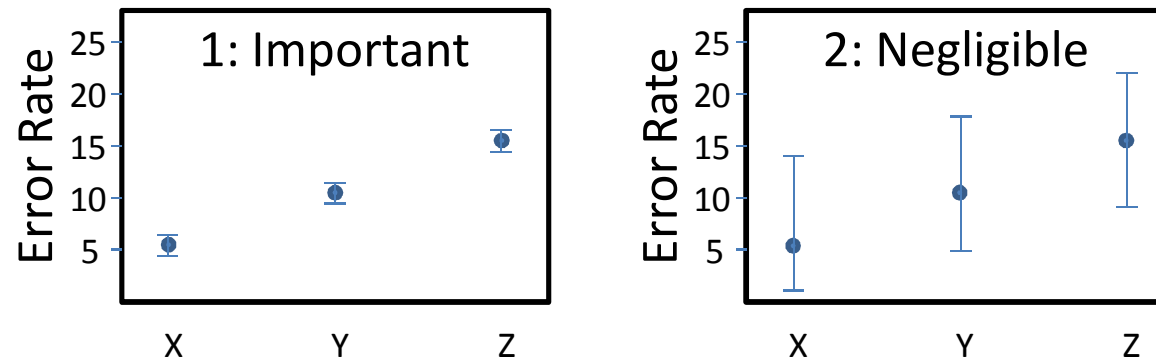
- Gathering and sharing good data is hard!
 - Ground truth, artifacts, and realism are recurring problems
 - Confidential or sensitive information limit willingness to share

	1	2	3	4
A	Case 3: Data/Detector Interaction ☹️☹️☹️			
B				
C				
D				

- Good science without comparative experiments is also hard
 - The problem does not go away because the solution is inconvenient.
- Possible solutions:
 - Repositories like PREDICT can protect shared data
 - Testbeds like DETER can generate non-sensitive data
 - One shared data set, even if perfect, would not be enough
 - Detectors could be shared instead of data

Establishing a standard methodology

- Choosing the right inferential technique can be hard!
 - Statistical hypothesis tests vs. confidence intervals
 - Threshold significance levels vs. p-values
 - Classical, non-parametric, or Bayesian methods



- They may disagree on the details, but all statisticians make inferences
- Additional thoughts:
 - Practically, different techniques lead to similar conclusions
 - Consult with statisticians and discuss the right techniques for our data or domain
 - My suggestion is to start with classical methods and confidence intervals

Security-specific challenges

- Dealing with a malicious and intelligent adversary is hard!
 - A lot of other sciences deal with averages; we deal with worst cases

For certain areas of computer security, experiments seem useful, and the community will benefit from better experimental infrastructure, datasets, and methods. For other areas, it seems difficult to do meaningful experiments without developing a way to model a sophisticated, creative adversary.

(Stolfo, Bellovin, & Evans, 2011)

- Possible solutions:
 - Identify where experiments and inferences would be useful; start doing them
 - Establish the ratio of useful to difficult (e.g., 80:20, 50:50, 20:80)
 - Study adversaries and build a model (possibly using experiments and inferences)

Changing the culture

- Fine! We could and should do experiments and inferences. How?
 - Despite the magnitude of the problem, inertia is strong
 - Comparative experiments are sometimes done, inferences never
- Change starts at home
 - Where “home” is our own research and peer reviews
- Additional thoughts:
 - Conferences can and do offer a “carrot” for shared data
 - Perhaps a “stick” is sometimes necessary (e.g., archival journals)
 - Reviewer guidelines for what constitute acceptable methods
 - Decide when promising exploratory work is acceptable

Beyond experiments and inferences

- The limits of comparative experiments and inferences
 - Is it enough to do comparative experiments and inferential statistics?
- Experiments and inferences are necessary, not sufficient:
 - Invalid experiments that test the wrong things
 - Unrealistic evaluation data
 - Research that cannot be reproduced
 - Inferential techniques that are inappropriate for the data
- Bad science can be done with experiments and inferences. Can good science be done without them?

Thank you!

- NSF, CyLab, ARO, CERT, and USENIX
- David Banks, Shing-hon Lao, Soojung Ha, Chao Shen, and Pat Loring
- CSET organizers, reviewers, and participants

Related efforts

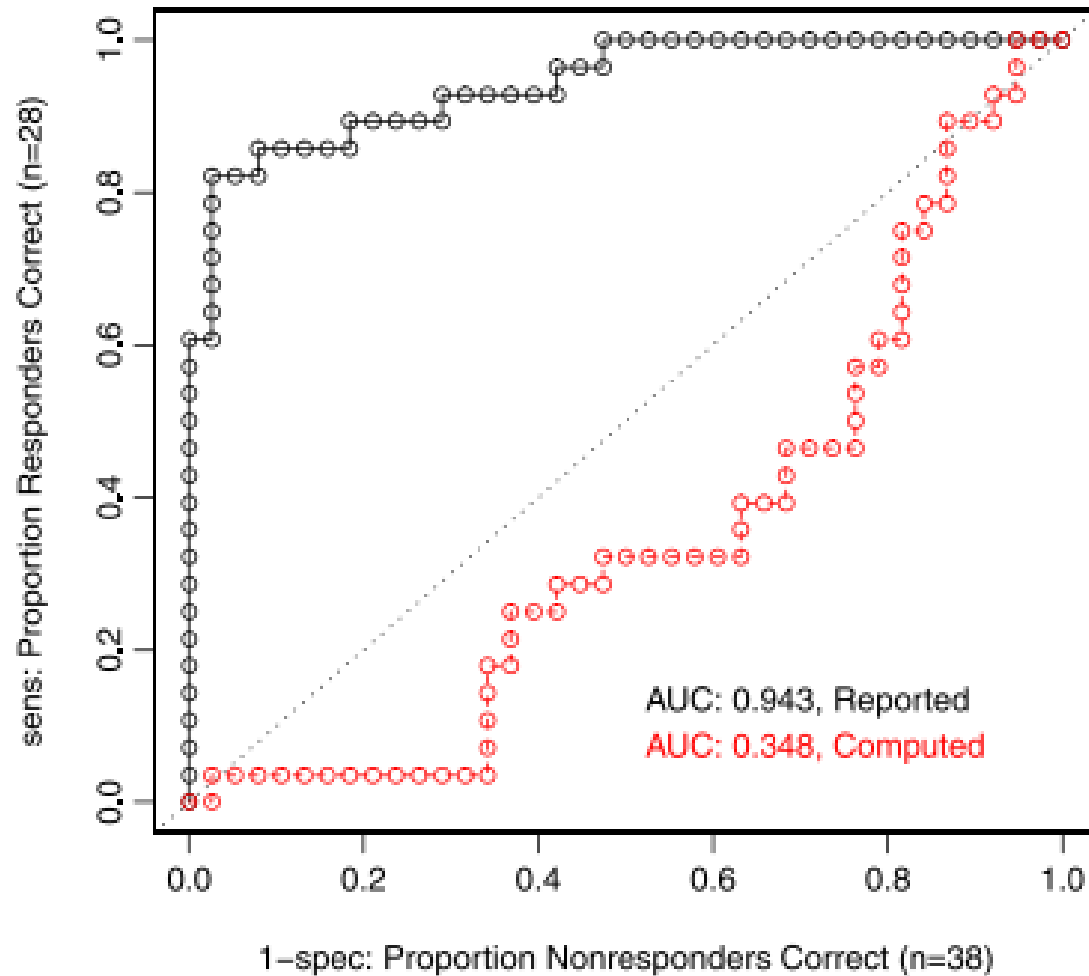
- Tichy (1998):
Computer science lags behind others in experimental methodology
- Kurkowski et al. (2005):
Similar problems exist in mobile network research
- Peisert and Bishop (2007):
Security experiments should be falsifiable, controlled, and reproducible
- Somayaji et al. (2009):
Adapted particular experimental and statistical methods (clinical trials) to security research
- Sommer and Paxson (2010):
More advice when using machine-learning in security domains

In closing ...

- In bioinformatics, researchers are trained to do comparative experiments and statistical inferences
- Government funding and journal publication require that the research data be shared and that statistical tests be significant
- The expectation is that someone can download researchers' data and scripts and "reproduce" all the tables and figures in their paper.
- For particularly promising results, forensic statisticians test this expectation.
- They often don't succeed:
 - Data sets contain duplicated and missing subjects
 - Class labels (e.g., diseased vs. healthy) have been reversed
 - Off-by-one errors identify the wrong factor as significant
 - Many times the failure cannot be adequately explained

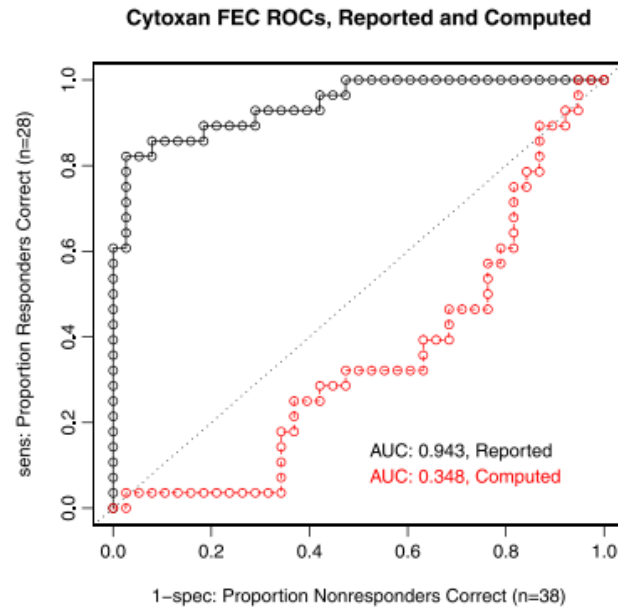
In closing ...

Cytoxan FEC ROCs, Reported and Computed



(Baggerly & Coombes, 2010)

In closing ...



- In a field where ...
 - comparative experiments are the status quo
 - inferential statistics are taught in research-methods courses
 - bad research is severely penalizedthey *still* discover problems.
- How concerned should we be about security research?