The 10 Commandments of Internet Measurement

Prof. Matthew Roughan
matthew.roughan@adelaide.edu.au
http://www.maths.adelaide.edu.au/matthew.roughan/

UoA

March 30th, 2017

THE UNIVERSITY of ADELAIDE

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Acks

Much of this has arisen out of conversations with, and the work of many others. A very incomplete list is: Mark Allman, Paul Barford, Randy Bush, Mark Crovella, Christophe Diot, Anja Feldmann, Albert Greenberg, Tim Griffin, Balachander Krishnamurthy, Olaf Maennel, Jennifer Rexford, Jono Tuke, Darryl Veitch, Walter Willinger, Jennifer Yates and Yin Zhang (sorry if I left someone out).

But more than that, these grew because I have been burned! Maybe I can save you some scars.

I write to find out what I think about something.

Neil Gaiman, The View From the Cheap Seats

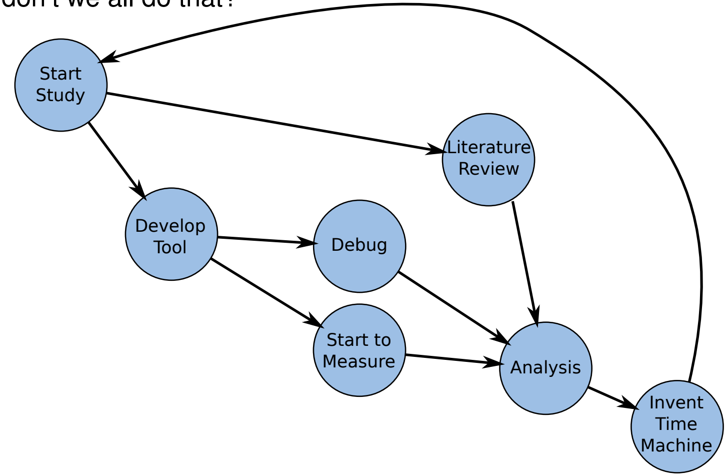
1 – Thou Shalt Plan Your Experiment

1 – Thou Shalt Plan Your Experiment

- Seems obvious?
    - don’t we all do that?

1 – Thou Shalt Plan Your Experiment

- Seems obvious?
  - don't we all do that?



- Actually, we usually concentrate on the technical difficulties, and getting the data, and ignore important pieces

1 – Thou Shalt Plan Your Experiment

The sort of thing to think about at the start

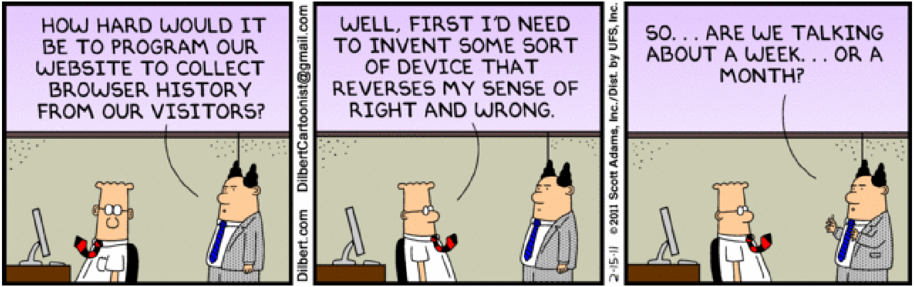
- What about statistical sample size calculations?
- What about a "control" experiment?
- What technique will we use to analyse the data?

Deliberate focus here on statistics

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.
>
> *Ronald Fisher*

2 – Thou Shalt Do No Harm

2 – Thou Shalt Do No Harm


HOW HARD WOULD IT BE TO PROGRAM OUR WEBSITE TO COLLECT BROWSER HISTORY FROM OUR VISITORS?
WELL, FIRST I'D NEED TO INVENT SOME SORT OF DEVICE THAT REVERSES MY SENSE OF RIGHT AND WRONG.
SO. . . ARE WE TALKING ABOUT A WEEK. . . OR A MONTH?
Dilbert.com DilbertCartoonist@gmail.com
2-15-11 ©2011 Scott Adams, Inc./Dist. by UFS, Inc.

- Just a joke, but how many Internet Measurement experiments have done something similar without thinking?

P.S. I am aware of the contradiction in stealing a Dilbert for this part of the talk.

Measurement Ethics

Internet measurements are ultimately measurements of humans (in most cases), and so ethics **MUST** be considered **BEFORE** starting an experiment, but ethics should be part of all activities regardless.

- Most often, this won't be a big deal.
- Starting point: **do no harm**
- But there are more extensive discussions, *e.g.,* [1], or see the ACM code of Ethics `https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct`.
- Ethics is an evolving conversation – participate.

3 – Thou Shalt Do Reproducible Research

3 – Thou Shalt Do Reproducible Research

Science **requires** reproducibility. We don't believe X because Matt R said so. We believe it because A,B,C, ..., and Z confirmed it.

Why?

- Authors and review processes are flawed
- Single data sets aren't always representative
- Things change over time (quickly in the Internet)
- Some (rare) authors are dishonest

The idea isn't new – it goes back to Aristotle – and is **core** to the philosophy of science.

Reprise of [2]

Quoting Mark Allman: “Among my least favorite review comments ...

- data comes only from one university ...
- ...
- the user population is small ...

... so, the study is not representative.
These review comments are simultaneously ...

- correct
- vacuous”

The point is that no single study will ever be completely, convincingly representative.

- In important areas, they repeat studies.
- In important areas, they do “meta-analysis” of groups of studies.

3 – Reproducible Research TO-DO list

- Distinguish raw data from cleaned/analysed data, and link them with a repeatable/reproducible pipeline
  - don't hand edit data
  - record **exactly what you do**
- Reuse and share data [3]
  - Public data is best, but if you can't consider sharing under NDA
  - Too many studies do "yet another data set" without using existing data as a baseline – **reuse data**
- Publish code, even if you can't publish data

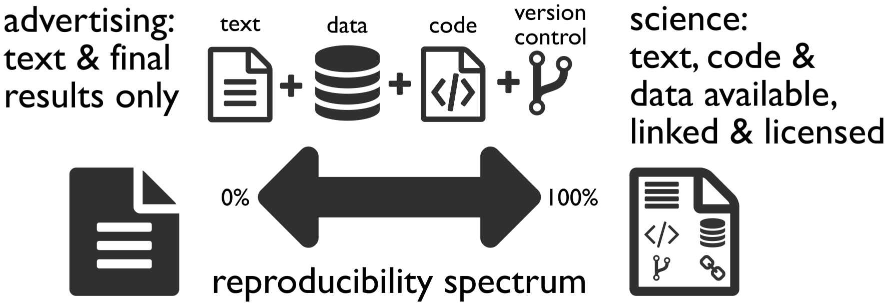
> An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the Figures.
>
> *Buckheit and Donoho [4]*

- Don't automatically reject replication studies

3 – Reproducibility spectrum

This is all HARD and SCARY, but we need to get better.

Push yourself towards the right of the spectrum



Adapted with permission from Rodríguez-Sánchez F, Pérez-Luque AJ, Bartomeus I, Varela S (2016) Ciencia reproducible: qué, por qué, cómo. Ecosistemas, 25(2): 83-92. http://doi.org/10.7818/ECOS.2016.25-2.11. See also Marwick, B. (2016). Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. Journal of Archaeological Method and Theory 23(2): 1–27. http://doi.org/10.1007/s10816-015-9272-9. This figure is CC-BY

In keeping with the Biblical Theme: Go forth and reproduce!

4 – Thou Shalt Have a Data Management Plan

4 – Thou Shalt Have a Data Management Plan

Data Management Issues

- **release:** the data could be confidential, or covered by an NDA or acceptable use conditions:
  - How do you ensure that your data isn't inappropriately released?
- **retention:** this is a key to reproducibility, and a measure against academic fraud

  The ARC and NHMRC's "Australian Code for the Responsible Conduct of Research", requires, amongst many others issues:

  > In general, the minimum recommended period for retention of research data is 5 years from the date of publication.
  >
  > *Section 2.1.1*

  This is just the Australian example – there are similar in other countries.
- ...

Data Management: Retention

> Any code of your own that you haven't looked at for six or more months might as well have been written by someone else.
>
> *Eagleson's law*

> Always code as if the guy who ends up maintaining your code will be a violent psychopath who knows where you live.
>
> *Rick Osborne*

> Data is code.
>
> *MattR*

Data retention is not just about storing the data, its about being able to access it:

- need software (or preferably a portable format)
- need documentation so you understand it

4 – Data Management Plan

6Ws of Data Management

- What data is involved?
- How will it be formatted, stored and analysed?
- Who provides/collects it, and who has access?
- When will you get it?
  To when must it be retained?
- Why are you collecting it, and are you allowed to use it for other purposes?
- Where will it be stored and worked on?
  Where will it be backed up?

5 – Thou Shalt Not Accept Data on Face Value

5 – Thou Shalt Not Accept Data on Face Value

All (non-trivial) data/measurements contain errors

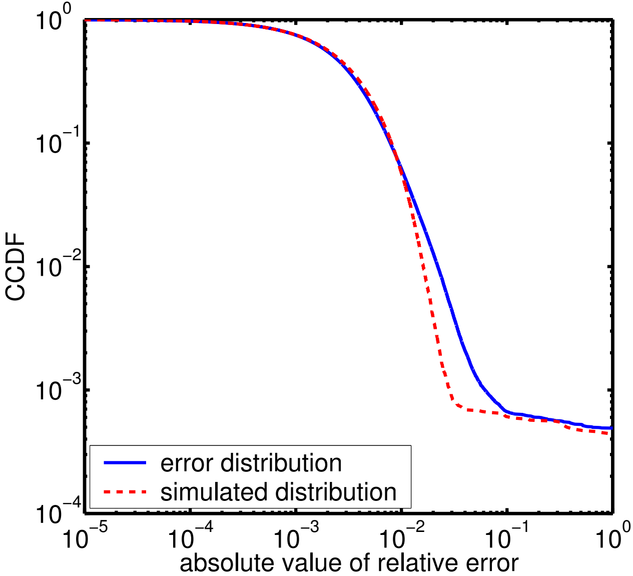
- formatting errors
- missing data
- ambiguous data
- “noise”
- artefacts

You need to deal with these problems

- Calibration
- Cleaning

Even if you got the data from X and they said it is fine!

SNMP – Noise vs Artefacts [5]



Calibration [6]

Calibration: is a key process in a good study. Its part of understanding what is good and bad about your data.

- Examine outliers – often they signify a problem.
- Self-consistency checks
  - simple: non-negativity
  - better: compare different views of same data
- Compare to ground truth(iness)
- Use synthetic test data: emulation, simulation, ...

Data Cleaning

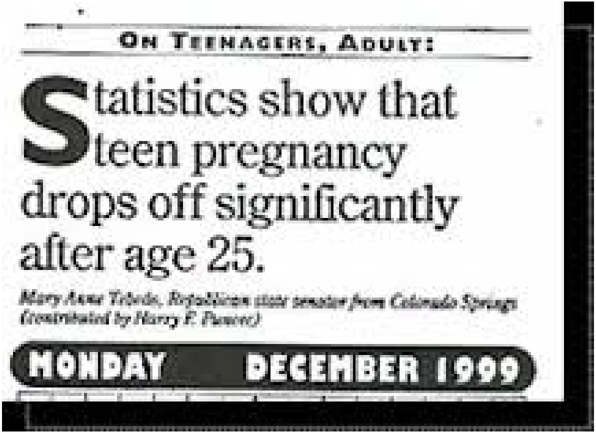
> Its often said that 80% of the effort in a data analysis is spent on data cleaning, the process of getting the data ready to analyse.
>
> *Hadley Wickham [7]*

> Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's Law.
>
> *Douglas Hofstadter, Gödel, Escher, Bach: An Eternal Golden Braid*

- Apply Hofstadter's Law twice to data cleaning.
- Remember about reproducibility!

6 – Thou Shalt Learn Your Statistics Well

6 – Thou Shalt Learn Your Statistics Well



ON TEENAGERS, ADULT:

Statistics show that teen pregnancy drops off significantly after age 25.

Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Panetz)

MONDAY DECEMBER 1999

Lies, Damn Lies, and Internet Measurements [8]

"There are lies, damn lies, and statistics. We're looking for someone who can make all three of these work for us."

7 – Thou Shalt Collect and Keep Metadata

7 – Thou Shalt Collect and Keep Metadata [6]

Metadata: **Data about the data**

- Basics (descriptive metadata)
  - dates/times
  - where it was collected
  - parameter settings (*e.g.,* sampling rate)
  - resolution
  - units
- Equivalents of comments in code
  - file format information
  - non-standard details (e.g. I have put missing values as -1)
- Bread crumbs (tracking metadata)
  - what program (and version) created, processed or cleaned the data
  - date created (and altered)
  - parameter values used to generate data
  - links to other relevant data files (e.g. inputs)
  - the "author"
- Extras
  - known failures (*e.g.,* packet drops, session up/down times)
  - mappings (e.g., IP addresses to DNS names)

7 – Retain Your Metadata

Make sure it can't be lost!!! Perhaps include it in the data file itself.

8 – Thou Shalt Debug

8 – Thou Shalt Debug

Goes without saying you will debug your measurement code.

8 – Thou Shalt Debug

OK, maybe I am saying it.

8 – Thou Shalt Debug

Just do it. If you need motivation.

> It's not at all important to get it right the first time. It's vitally important to get it right the last time.
>
> *Andrew Hunt and David Thomas*

> Beware of bugs in the above code; I have only proved it correct, not tried it.
>
> *Donald Knuth*

> There are two ways to write error-free programs; only the third one works.
>
> *Epigrams in Programming, 40., Alan Perlis*

8 – Thou Shalt Debug

An Example:

- Code Red / Nimda were big worms back in Sept 2001 [9]
- They also had an affect on BGP (global routing) [10]
- But wait [6, 11], BGP multi-hop monitoring session (*e.g.,* Route Views at the time) are fragile. What they really saw was table resets, when the sessions were re-established.

Make sure you are measuring what you think you are measuring.

8 – Thou Shalt Debug

Debugging measurement code has a lot in common with calibration, but we should be thinking about calibration for **ANY** data set, not just when you are writing the code.

Also:

> Debugging is twice as hard as writing the code in the first place. Therefore, if you write the code as cleverly as possible, you are, by definition, not smart enough to debug it.
>
> *Brian Kernighan*

Doubly true for distributed systems, *e.g.,* Internet Measurements.

9 – Thou Shalt Make Mistakes

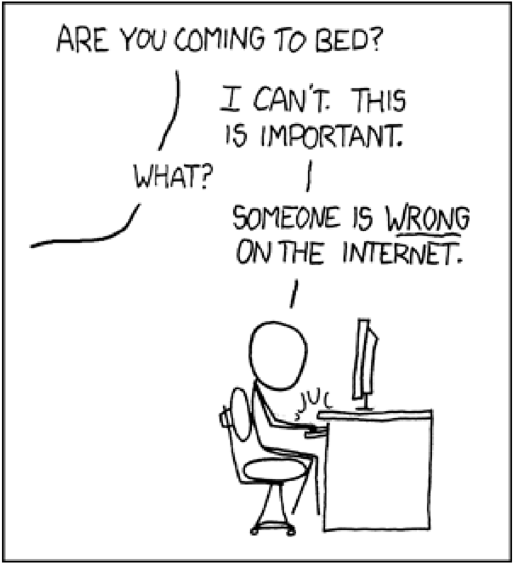
9 – Thou Shalt Make Mistakes

This not so much an instruction, as an observation, but maybe it should be an instruction as well.

- We learn from mistakes
- Maybe, you will write your own version of these slides one day, based on your (corrected) mistakes.

> Sometimes the best way to learn something is by doing it wrong and looking at what you did.
>
> *Neil Gaiman*

- So take risks, do interesting things, and **learn from the things that go wrong.**



ARE YOU COMING TO BED?
I CAN'T. THIS IS IMPORTANT.
WHAT?
SOMEONE IS WRONG ON THE INTERNET.

https://xkcd.com/386/

10 – Thou Shalt ???

10 – Thou Shalt ???

Oops – I only came up with 9 commandments ...

10 – Thou Shalt ???

Oops – I only came up with 9 commandments ...

... Let’s go with the Donkey thing, that’s always a good one.

The 10

1. Thou Shalt Plan Your Experiment
2. Thou Shalt Do No Harm
3. Thou Shalt Do Reproducible Research
4. Thou Shalt Have a Data Management Plan
5. Thou Shalt Not Accept Data on Face Value
6. Thou Shalt Learn Your Statistics Well
7. Thou Shalt Collect and Keep Metadata
8. Thou Shalt Debug
9. Thou Shalt Make Mistakes
10. Thou Shalt Not Covet Your Neighbour's Donkey

I have been burned! Maybe I can save you some scars.

Conclusion

I don’t like endings, so here are some quotes to go on with.

> ”Measure what is measurable, and make measurable what is not so.”
>
> *Galilei, Galileo (1564 - 1642)*

> ”To measure is to know.”
>
> *Lord Kelvin (1824-1907)*

- C. Partridge and M. Allman, “Ethical considerations in network measurement papers,” in *Communications of the ACM*, vol. 59, pp. 58–64, 2016.

- M. Allman, “Towards better internet empiricalism,” in *Keynote talk at the Passive and Active Measurements Conference (PAM)*, (NY, NY, USA), 2015.

- J. Heidemann, “Sharing network data: Bright gray days ahead.” Keynote talk at the Passive and Active Measurements Conference (PAM), 2014.

- J. B. Buckheit and D. L. Donoho, *Wavelets and Statistics*, vol. 103, ch. WaveLab and Reproducible Research, pp. 55–81. Springer: Lecture Notes in Statistics, 1995.

- M. Roughan, “A case-study of the accuracy of SNMP measurements,” *Journal of Electrical and Computer Engineering*, vol. 2010, 2010. Article ID 812979, doi:10.1155/2010/812979. `http://www.hindawi.com/journals/jece/2010/812979.html`.

- V. Paxson, “Strategies for sound Internet measurement,” in *ACM Sigcomm Internet Measurement Conference (IMC)*, (Taormina, Sicily, Italy), October 2004.

- H. Wickham, “Tidy data,” *Journal of Statistical Software*, Submitted.

- M. Roughan, “Lies, damn lies, and Internet measurements,” in *Keynote talk at TMA*, 2016. `http://www.maths.adelaide.edu.au/matthew.roughan/talks.html`.

- D. Moore, C. Shannon, and J. Brown, “Code-Red: a case study on the spread and victims of an Internet worm,” in *ACM SIGCOMM Internet Measurement Workshop (IMW)*, 2002.

- L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, “Observation and analysis of BGP behavior under stress,” in *Proceedings of Internet Measurement Workshop*, 2002.

- B. Zhang, V. Kambhampati, M. Lad, D. Massey, and L. Zhang, “Identifying BGP routing table transfers,” in *ACM SIGCOMM Mining the Network Data (MineNet) Workshop*, 2005.

Bonus frames