Design of Experiments

Justin Zobel

University of Melbourne, Australia

Semester 2, 2017

Planning research

Write down a specific question that you are investigating.

Define what is to be evaluated or measured – what system, what properties of the system.

Sketch a plausible experiment or trial:

- To get output to be used as evidence in a paper.
- To gather information for your own learning.

Brainstorm different ways of investigating the questions.

Consider simple preliminary investigations, or case studies, before embarking on a large-scale implementation.

Search for definitive experiments – small differences are not convincing.

Distinguish exploratory work from substantial, confirming experiments.

Planning research . . .

Identify baselines. Exactly what can existing approaches do, at their best?

Consider how the results of the experiments are likely to relate to the hypothesis:

- Does success mean that the hypothesis holds?
- Does failure mean that the hypothesis is wrong?

A great many published experiments have unrealistic results.

- Yes, they're published.
- But they're probably not cited.

Planning research . . .

Keep a notebook:

- Software used and how to use it.
- Locations of data and logfiles.
- Intentions.
- Expected outcomes.

Hand-written, portable notebooks meet ethical guidelines and are an effective way to work.

Planning research . . .

Construct the expected form of the results – a table of figures, a graph showing a trend.

- What are the axes (or column&row headings)?
- What are the parameters? – For example, hash table size is a tunable parameter that controls space and time.
- Which parameters are dependent on others? Which are truly fundamental?
- What lines will the graph need?
- What is the likely scale on each axis?
- Is the scaling likely to be geometric (logarithmic axes)? Linear?

Planning research . . .

Again: consider how to measure your system or process.

- Is a human needed to gather the results?
- Are observers or experimental subjects needed to evaluate the results?
- Will the experiments require a particular kind of data?
- What code needs to be written?
- What platform is required?

Again: only undertake work if you are confident the outcomes can be measured.

Are users needed?

How carefully will the experiment have to be planned? Can it be run again if the results are unusable?

How to make the experiment blind, or double-blind?

Will the users know whether there is a human or an algorithm behind the curtain?

Will they know what the study is intended to find out?

Users allow many forms of measurement:

- Qualitative assessment of a system – ease of use, satisfaction, preference, etc.
- Subjective (self-reporting by subjects) vs objective (based on data that is gathered silently behind the scenes).

Are users needed . . .

Does it need to be a large, controlled experiment, or are case studies sufficient?

How many users will be needed, and who will they be?

- Will payment be required?
- Example – are undergraduates typical of naïve search engine users?
- What steps need to be taken to avoid communicating the desired outcomes to the users?
- What instructions will they be given?
- Are you aware of the relevant ethical guidelines?
  (Test: why can't you use the students in your undergraduate tutorial group as experimental subjects?)

Obtaining data?

How much data is required for realism?

Does real data need to be obtained, or is simulated or artificial data sufficient?

- What are the compromises inherent in the artificial data? Example – is a sample of words in text a good substitute for a search engine query log? Are randomly generated strings a good substitute for real web pages?
- Any scheme for generating artificial data relies on a *model*; what is the model representative of?
- Can the artificial data be validated against a sample of real data?

Obtaining data . . .

Have you considered the relevant statistics?

- Is a single data set sufficient?
- Are subsamples of a data set independent of each other?
- Is the 'test' data set sufficiently independent of the 'observation' data set?
- Do you have sufficient data, or data sets, to reliably observe the effect you are hoping for?
  That is, does the test have sufficient experimental power?

You need to know you can get the data *before* wasting effort on coding; if possible, keep a static snapshot.

Equipment, code, tools?

You may need to trade off ease of implementation against realism of the result. Computational examples:

- If you plan to measure algorithmic efficiency, implement in a suitable language.
- Coding for a day followed by execution for a month is a lot less efficient than coding for five days followed by execution for 20 minutes . . . especially if you have to run again.

Don't overimplement or invest in unnecessary effort.

- Download codebanks, use libraries.
- Leave off unnecessary features.
- Decompose into simple, separate pieces of code.

Equipment, code, tools? . . .

Make your assumptions explicit. Computational examples again:

- Can all data be held in memory, or is it necessary, for realism, to manage data on disk?
- Hard-coding of data structures, input formats, etc., may allow for rapid implementation; does it lead to unrealistic behaviour or simplifications?

Construct baselines to the highest possible standard.

Test remorselessly.

Example: if you think you've successfully compressed some data, prove it: write a working decompressor.

For long-running processes, consider incremental dumps so that weeks or more of work isn't lost.

Running a computational experiment

Always run code from a script, even if it is only a couple of lines long. This is reproducible, and means the process will still work later on.

Your script should:

- Recompile code if necessary.
- Write all output to the logfile.
- Set parameters, and make sure they are recorded with the results.
- Record the data file name.
- Include all steps necessary for producing the output; this may include creating a graph (or a table in LaTeX or Excel).
- Undertake averaging and other statistical steps.
- Include loops to ensure that timings are meaningful (times of less than a minute are unreliable).

Running the experiment . . .

Keep logfiles that contain the output of experiments.

- Include dates, parameter settings.
- Digest results into a human-readable form; keep raw output separately, if necessary.
- Choose formats that allow simple processing, for example to allow a script to build a graph from the results.
- Never hand-edit a log file.
- Tip: date formats such as YYYYMMDD allow for easy browsing.

Archive all text, logs, data, scripts, and source code in a durable, unlosable way.

Running the experiment . . .

Vary one variable; fix the others.

Identify *factorial designs* that separate the contributions of each parameter.

If there are users involved, randomly allocate them to pools that correspond to different environments.

Use pilot studies to identify the parameters that appear to be of the greatest interest – but beware of confirmation fallacies, such as expecting a strong effect in the first study to be strong in the second.

Running the experiment . . .

Does your method decompose into a series of stages? Are there choices at each stage? Does your innovation consist of a chain of independent decisions? If so,

- Separately test the impact of each of the decisions.
- Often all of the effect is due to a single decision.
- Consider independent tests of each stage – impact on the
- For example, measure stemming of text by its effect on the vocabulary and its ability to correctly bring together words, not just its effect on a search engine.

Running the experiment . . .

Explore each variable in detail.

- If increasing a value has an effect, what happens with a further increase?
- A decrease?
- Where do trends asymptote to?

Take advantage of every opportunity to visualise results.

Learn the simplest scripting language that will do the job.

Use the right tool, not the most convenient tool.

Don't stop after the first successful result; it may be a fluke.

What to report

Readers need to know or see:

- How the data was gathered.
- How the data might be obtained, or recreated.
- What the results look like.
- Behaviour as different parameters are varied.
- A multi-perspective analysis of the data using appropriate methods or tools.
- What the shortcomings of the data are: uncertainty, incompleteness, unreliability.
- What aspects of the research question are not tested by the data.
- What the results mean or imply – what new knowledge can be inferred from the results.

Observe that raw data and massive listings of intermediate outcomes are not in this list!

What to report . . .

Make a reasonable attempt to allow reproduction.

- That is, give the readers enough information to run an experiment that confirms your results.

Most experiments are not reproducible . . .

- The barrier to entry is too high (e.g., the software platform is unavailable, comparable data cannot be collected).
- The authors don't provide enough information. Or the information they do provide is wrong.
- Arbitrary choices are unstated.

. . . but truly unreproducible work is worthless.