

# Principles of data management and organization

Andy Lynch

CRUK CI

2016/02/02

# What are data disasters?

## (My) Definition

Anything that stops somebody analysing data in the way that they are supposed to be able so to do.

- Total destruction of data
- Inability to find data
- Corruption of data
- Forgetting what the data mean
- Being unable to reproduce your results
- Somebody else being unable to reproduce your results
- Unauthorized access-to/use-of data

# Some of these are Computing issues

## Principle 1

Ensure you have a secure back up of the raw data

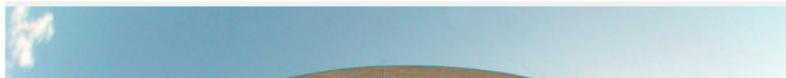
- Vulnerable until you have this - requires care
- Some degree of organization required
- Check that the backup is working
- Need to be clear what the raw data are
- There are costs associated with backing data up
- Check that the backup is future-proofed

# Loss of data

## Cambridge student 'devastated' as burglar steals the only copy of his PHD work

By **Cambridge News** | Posted: January 30, 2016

By Josh Thomas



“I had a whole year’s worth of PhD work stolen. There was work on the laptop which is not backed up”

[www.cambridge-news.co.uk](http://www.cambridge-news.co.uk)

# How are data corrupted?

- Data can be corrupted by hardware issues
- Data can maliciously be corrupted by a third party
- Data can deliberately (non-maliciously) be corrupted by the user
- Data can automatically be corrupted by 'helpful' software
- Data can accidentally be corrupted when using software

# Keith Baggerly's Duke example

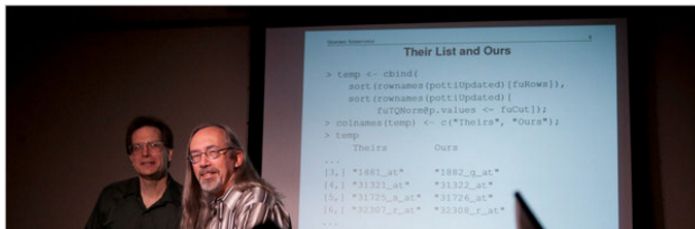
HOME SEARCH

The New York Times

## RESEARCH

# How Bright Promise in Cancer Testing Fell Apart

By GINA KOLATA JULY 7, 2011

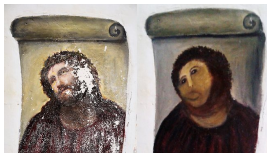


[Link to video](#)

# Don't touch-up a masterpiece

## Principle 2

Never work directly on the raw data



Ideally we adopt a practice of having a raw data file and recording the changes made to it

- Even better if the recording is 'automatic'
- This approach naturally makes research reproducible
- It can aid the understanding of the data
- It saves having to backup multiple large datasets

- Used to manipulate spreadsheet-like data in a recorded and reproducible manner
- Mark will demonstrate later if requested

[illegible]



# Use the right software

## Principle 3

Compose your 'raw' data with the right tools

If raw data are from genomics/proteomics et al. then the raw data are pre-defined. If assembling them yourself then care is required.

- For an **expert** user, Excel can be fine.
- Otherwise consider tools such as SPSS
- A simple database is not so great a cost, and can help with inconsistencies in data entry

# Use the right format

Format here doesn't mean .csv not .xls, nor fonts, colours etc.

## Principle 4

Set up your data with the right shape

You'll see a lot of this later, but:

- Aim for a rectangle
- No blank cells (but be careful how you represent missing data)
- Each row is a case, each column a variable (although not always)
- The key is deciding what constitutes a case
- If it isn't clear what a case is, you might be better off with two tables.

# Naming variables

## Principle 5

Give variables and cases sensible names

Case and variable names need to be

- unique
- lacking in exotic characters
- interpretable
- accurate

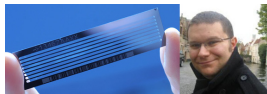


Figure 1: One of these is a person.

(flowcell image from [global.fncstatic.com](http://global.fncstatic.com))

# METABRIC example 1

In METABRIC we looked at breast cancer samples from  $\sim 2000$  patients

- Each patient had two breasts
- Each breast potentially had multiple tumours
- Each tumour potentially had multiple samples
- Each sample was run on multiple technologies
- Each technology could have been repeated

Two reasons not to label data from a microarray as if it was a person

- 1 if there are multiple arrays from the same person, then it leads to confusion

## METABRIC example 2

- 2 The only thing you know about the microarray data (i.e. the only metadata that are raw) is that they came from the microarray.

We had a problem with sample mixups within the project.

- Had we had 'raw' data where the cases were labelled by array name, we could simply have changed the file that mapped arrays to patients.
- It would have been easy to check that every analyst had the latest version
- By just changing the labels on the raw data, it became much harder to keep track of who was using what

# Meta data

## Principle 6

### Have thorough Meta Data

- Since you now have rectangular data with succinct variable names, you may have lost some detail of what those variables are.
- A document explaining the primary data is invaluable.
- It can be variously known as a Data Dictionary (KB), Glossary (GSS), or Variable View (SPSS),
- For variables that can only take fixed values it can define those levels.

# Versioning

## Principle 7

Be clear what data are being analysed

- The raw data shouldn't change (probably), but the working data could easily so do
- To reproduce results, it is important to be able to specify the version of the data that was used.
- Some form of versioning is therefore important.

# Summary

- 1 Ensure you have a secure back up of the raw data
- 2 Never work directly on the raw data
- 3 Compose your 'raw' data with the right tools
- 4 Set up your data with the right shape
- 5 Give variables and cases sensible names
- 6 Have thorough meta data
- 7 Be clear what data are being analysed

## Examples

Enjoy the rest of the course.