Datasets

Marcio Augusto Diniz

Cedars Sinai Medical Center

23 August, 2022

Why is important to handle data correctly?

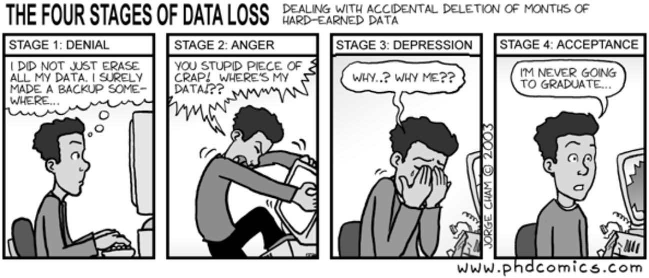

THE FOUR STAGES OF DATA LOSS
DEALING WITH ACCIDENTAL DELETION OF MONTHS OF HARD-EARNED DATA
STAGE 1: DENIAL
I DID NOT JUST ERASE ALL MY DATA. I SURELY MADE A BACKUP SOMEWHERE...
STAGE 2: ANGER
YOU STUPID PIECE OF CRAP! WHERE'S MY DATA!??
STAGE 3: DEPRESSION
WHY..? WHY ME??
STAGE 4: ACCEPTANCE
I'M NEVER GOING TO GRADUATE...
JORGE CHAM © 2003
www.phdcomics.com

Figure 1: Do different backups every day you work on your dataset.

Why is important to handle data correctly?

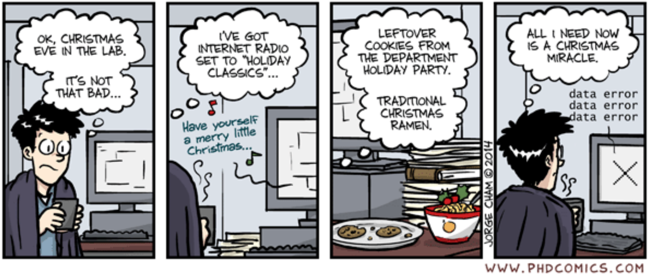

OK, CHRISTMAS EVE IN THE LAB.
IT'S NOT THAT BAD...
I'VE GOT INTERNET RADIO SET TO "HOLIDAY CLASSICS"...
Have yourself a merry little Christmas...
LEFTOVER COOKIES FROM THE DEPARTMENT HOLIDAY PARTY.
TRADITIONAL CHRISTMAS RAMEN.
JORGE CHAM © 2014
ALL I NEED NOW IS A CHRISTMAS MIRACLE.
data error
data error
data error
WWW.PHDCOMICS.COM

Figure 2: Do different backups every day you work on your dataset.

UppSala Scandal

- Lönnstedt, O.M. and Eklöv, P., 2016. Environmentally relevant concentrations of microplastic particles influence larval fish ecology. Science, 352(6290), pp.1213-1216.
- Authors reported experiments showing that fish that ate tiny 'microplastics' grew more slowly and were more likely to be eaten by predators.
- A group of researchers raised a complaint about the study that not all the data underlying the results were available.
- The only computer containing the study's raw data was allegedly stolen and no backups existed on another machine or an online repository.
- Uppsala University investigated those allegations last year and found no evidence of misconduct.
- The paper was retracted by the authors. http://science.sciencemag.org/content/356/6340/812.1

Duke Scandal

- Smoking and carcinoma of the lung. British medical journal. 1950 Sep 30;2(4682):739.
- Mortality in relation to smoking: ten years' observations of British doctors. British medical journal. 1964 May 30;1(5395):1399.

Duke Scandal

Personalized Medicine

- Aim: Establish whether a patient's genetic make-up can be used to identify therapeutic regimes that would provide better responses.
- Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D. Genomic signatures to guide the use of chemotherapeutics. Nature medicine. 2006 Nov 1;12(11):1294-300.
- Major breakthrough: oncologists could choose the most adequate chemotherapeutic regime based on the cancer patient's insensitivity.

Duke Scandal

Keith Baggerly and Kevin Coombes

- Biostatisticians at MD Anderson Cancer Center;
- Data that they could not understand;
- Mislabelled data;
- Non-reproducible steps of the analysis;
- Results seems the opposite;
- Duke researchers did not listen to them and stopped replying them;
- Medical journals published some communications, but they did not give much attention.

Duke Scandal

Clinical Trial

- Started in 2007 based on Potti's research;
- Accrued 109 patients;
- Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. The Annals of Applied Statistics. 2009 Dec 1:1309-34.
- The trial was suspended by Duke;
- The final Duke's report was that the data had been subject to modification carried out in a non-random way.

Reproducible vs Replicable

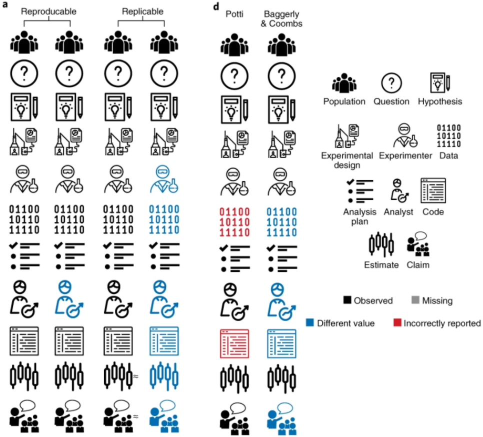


Figure 3: Patil P, Peng RD, Leek JT. A visual tool for defining reproducibility and replicability. Nature human behaviour. 2019 Jul;3(7):650-2.

Data Policies

Nature

- Authors must deposit their data in an approved data repository as part of the manuscript submission process; manuscripts will not otherwise be sent to review.
- During the peer-review process, Editors, Editorial Board Members and referees are asked to evaluate whether the data repository(s) selected by the authors is appropriate, and may deem it necessary for authors to archive their data in additional repositories prior to publication.
- More details: https://www.nature.com/sdata/policies/data-policies
- Data repositories: https://www.nature.com/sdata/policies/repositories

Data Policies

Science

- Before publication, large data sets (including microarray data, protein or DNA sequences, atomic coordinates or electron microscopy maps for molecular and macromolecular structures, and climate data) must be deposited in an approved database and an accession number or a specific access address must be included in the published paper.
- After publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of a Science Journal.
- More details: https://www.sciencemag.org/authors/science-journals-editorial-policies

Why is important to handle data correctly?

- To avoid to redo work;
- To reduce the probability to make mistakes;
- To make the interaction with other researchers more efficient;
- To make your statistical analysis to be reproducible.

What is wrong with this dataset?

Basic rules to organize data sets

- The preferred format are either Excel or .csv;
- Do not include Protected health information (any information in the medical record or designated record set that can be used to identify an individual and that was created, used, or disclosed in the course of providing a health care service such as diagnosis or treatment);
- Provide a dictionary of variables that describes each variable in more detail and indicates any coding scheme used for categorical variables (e.g., 0 = 'Female'; 1 = 'Male')
- Each row should contain all the information for one sample unit (patient, mouse, . . .). Avoid having multiple rows for the same patient unless the data is collected repeatedly over time;

What is wrong with this dataset?

Basic rules to organize data sets

- Each column (variable) must contain only one piece of information describing the sample unit. If necessary, create new variables;
- Use a specific code to indicate missing data (for example, NA);
- The first character of a variable name must be an alphabetic character. Subsequent characters can be alphabetic characters, numeric digits, or underscores. Special characters, except for the underscore or dot, are not allowed;
- Do not use colors or highlighting to distinguishing patient characteristics. Instead create a column (variable) to indicate the characteristic;

What is wrong with this dataset?

Basic rules to organize data sets

- All notes should be made in a separate column;
- Outcomes should all be converted to one specific unit and unit measure listed in the variable dictionary rather than in the data;
- Do not include blank/hidden rows or columns;
- Do not include calculations and graphics;
- Variables which are mutually exclusive listed as separate variables should be concatenate into one variable (column);
- If you have already sent the data to the statistician, but you need to add new information, do not change your previous format. If you have to add new variables, add the news variables in columns after the existing ones.

What does R expect whem importing a dataset?

- R expects a rectangle dataset with n rows and p columns.

Dataset formats

Wide

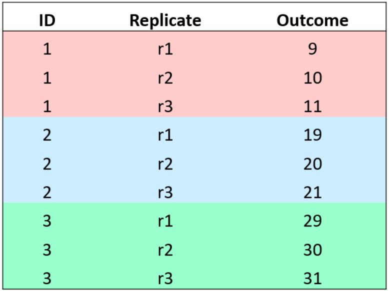
- Each row is a subject;
- Each column is a different variable with its first row labeled with the variable name.

Long

- A subject can have more than one row;
- Each column still is a variable, but there is one variable indicating the repeated measures.

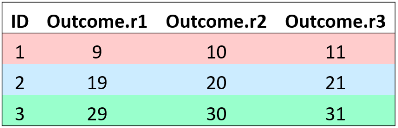
Why does the data format matter?

- R sometimes require data in different formats to plot data and perform tests of hypotheses.



ID	Replicate	Outcome
1	r1	9
1	r2	10
1	r3	11
2	r1	19
2	r2	20
2	r3	21
3	r1	29
3	r2	30
3	r3	31

Figure 4: Long format



ID	Outcome.r1	Outcome.r2	Outcome.r3
1	9	10	11
2	19	20	21
3	29	30	31

Figure 5: Wide format