Set01 - Data Management

STAT 401 (Engineering) - Iowa State University

January 11, 2017

Duke Breast Cancer Clinical Trial Fraud

`http://cancerletter.com/articles/20150522_1/`:

> *...fraudulent data...irregularities in handling of the data...problems with the data*

`http://www.nature.com/nm/journal/v13/n11/full/nm1107-1276b.html`

> *We report here our inability to reproduce their findings.*
>
> 1. *We cannot reproduce their selection of cell lines.*
> 2. *lists of genes ... are wrong because of an 'off-by-one' indexing error*
> 3. *Using their software and lists of cell lines, we [could not reproduce their findings] ...*
> 4. *For docetaxel, their software yields only 31 of their 50 reported genes... We do not know how these 19 can be obtained from the training data, and we suspect that they were included by mistake.*
> 5. *Their software does not maintain the independence of training and test sets ...*
> 6. *suggesting that most labels are reversed. If the labels are reversed, the model suggests administering the drug only to the patients it would not benefit.*
> 7. *When we apply the same methods but maintain the separation of training and test sets, predictions are poor*
>
> *We believe that this situation may be improved by an approach that allows a complete, auditable trail of data handling and statistical analysis.*

KISS Data Management

Create a process and stick to it!

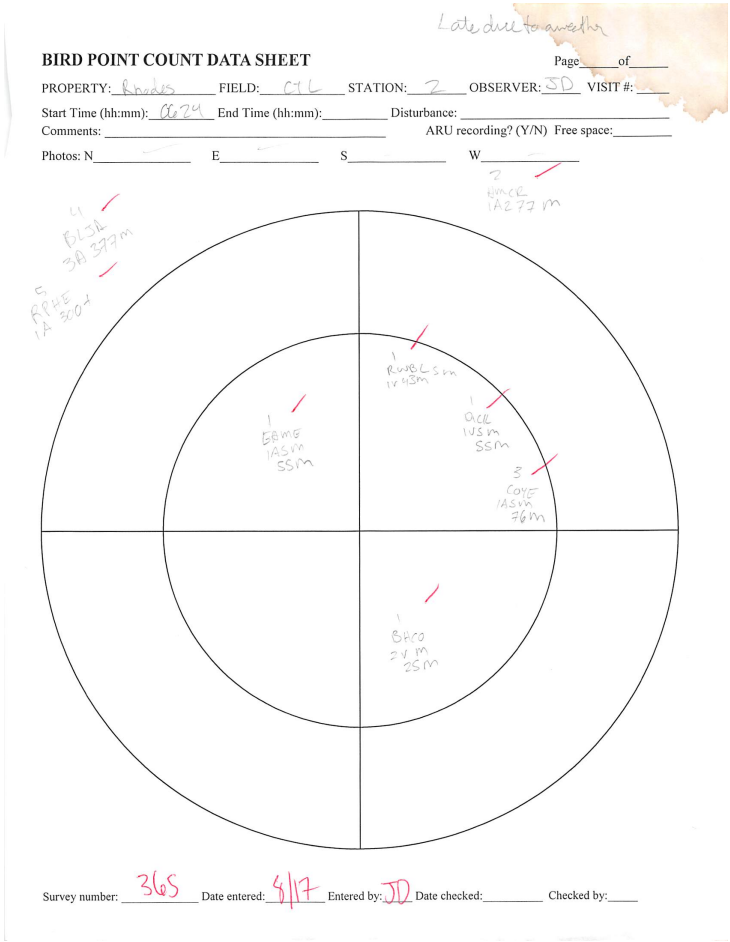
Suggested process:

1. Take a picture/scan/etc of raw non-digital data
2. Digitize raw non-digital data
3. Back up digital and non-digital raw data
4. Use scripts to create tidy data
5. Use scripts to perform statistical analyses

Do steps 1-3 routinely, e.g. every day.

Take a picture/scan/etc

To make sure you always have access to the raw non-digitized data, take a picture/scan/etc and save it wherever you will be saving the digitized version. For example,



Late due to weather

BIRD POINT COUNT DATA SHEET

Page_____of_____

PROPERTY: Brooks FIELD: CTL STATION: 2 OBSERVER: JD VISIT #: ____

Start Time (hh:mm): 0624 End Time (hh:mm): _____ Disturbance: _____

Comments: _____ ARU recording? (Y/N) Free space: _____

Photos: N_____ E_____ S_____ W_____

Survey number: 365 Date entered: 4/17 Entered by: JD Date checked: _____ Checked by: ____

KISS Digitize data

Either

- your data is already digital or
- you need to make it digital.

I suggest a 1 for 1 principle: make the digital version an exact (as best you can) copy of the non-digital version.

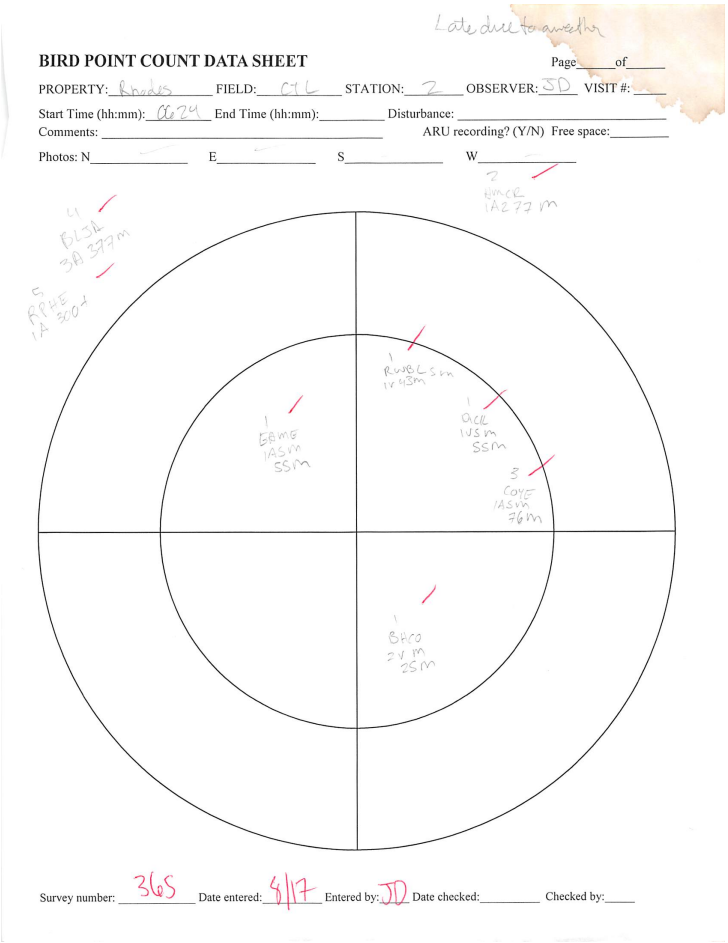
When making it digital, BE CONSISTENT!

- Directory structure
- File names
- Data file structure
- Column names in data file

It is okay if it isn't perfect (as long is it is consistent). Once it is digital, you can change it later. As long as you were consistent.

KISS Digitize data - example

bpc/2015/06/25/JD/0624.pdf:


BIRD POINT COUNT DATA SHEET

Page____of____

PROPERTY: Rhodes FIELD: CTL STATION: 2 OBSERVER: JD VISIT #: ____

Start Time (hh:mm): 0624 End Time (hh:mm):_____ Disturbance: _____

Comments: _____ ARU recording? (Y/N) Free space:_____

Photos: N_____ E_____ S_____ W_____

Late due to weather

Survey number: 365 Date entered: 4/17 Entered by: JD Date checked:_____ Checked by:____

bpc/2015/06/25/JD/0624.csv:

```
read.csv("0624.csv")
```

	minute	species	code	distance	angle
1	1	RWBL	1VSM	43	15
2	2	HMCR	1A	277	35
3	1	DICK	1VSM	55	45
4	3	COYE	1ASM	76	75
5	1	BHCO	2VM	25	170
6	5	RPHE	1A	300	315
7	1	EAME	1ASM	55	320
8	4	BLJA	3A	377	325

Backup raw data

Definition

The photo/scan/etc and the digital version are your raw data.

Your raw data should be

- in 2 physically different locations and, separately,
- routinely given to your PI.

`http://researchdata.wisc.edu/storing-data/top-5-data-management-tips-for-undergraduates/`:

> *This may be hard as a student with limited resources for storage. But if you can, try to practice 3-2-1. 3 copies of your data, in 2 different locations, on more than 1 type of storage hardware. This may seem excessive, but it can help protect you from the perfect storm of hardware malfunctions or physical accidents like flooding. UW offers Box and a number of other storage options depending on whether you are storing personal data or university data.*

`http://researchdata.wisc.edu/news/top-5-data-management-tips-for-graduate-students/`

> *Lets add on to that. 3-2-1-0. 0 USBs used as a form of storage hardware. A USB is easy to lose, misplace, and drop - it happens all the time. A USB is simply not a good form of backup.*

Backup raw data - options

IASTATE file storage: `https://www.it.iastate.edu/services/storage`

- CyBox `https://www.it.iastate.edu/services/storage/cybox`
- myfiles `https://www.it.iastate.edu/services/storage/myfiles`
- ResearchFiles `https://www.it.iastate.edu/services/storage/researchfiles`

Git/GitHub.com: Have the same repository (set of files) in multiple places.
Backup GitHub.com: `https://addyosmani.com/blog/backing-up-a-github-account/`

Use scripts to create tidy data

Definition

Tidy data are raw data that have been

- cleaned/munged/wrangled and
- collated/joined/processed

so that the data are ready for statistical analyses, e.g. making

- figures
- tables
- reports

Use scripts to create tidy data - example

Use this gist: `https://gist.github.com/jarad/8f3b79b33489828ab8244e82a4a0c5b3`:

Then for a particular set of files:

```
source("https://gist.githubusercontent.com/jarad/8f3b79b33489828ab8244e82a4a0c5b3/raw/494db9bffb10ed6d1928c1d13

bpc = read_dir(path = "../raw/bpc/2015",
               pattern = "*.csv",
               into = c(
                 "blank",
                 "raw",
                 "bpc",
                 "year",
                 "month",
                 "day",
                 "observer",
                 "property",
                 "field",
                 "station",
                 "start_time",
                 "extension")) %>%
  dplyr::select(-blank,-raw,-bpc,-extension)

readr::write_csv(bpc, path="bpc.csv")
```

Use scripts to perform analyses

Analysis scripts should use the tidy data to create

- figures,
- tables,
- reports, and/or
- manuscripts.

Use scripts to perform analyses - example

```
library(dplyr)

d <- read.csv("bpc.csv")

d %>%
  group_by(species) %>%
  summarize(count = n()) %>%
  arrange(-count)

# # A tibble: 21 x 2
#    species count
#     <fctr> <int>
# 1     DICK    11
# 2     RWBL     9
# 3     EAME     7
# 4     KILL     6
# 5     AMRO     4
# 6     COYE     4
# 7     RPHE     4
# 8     BHCO     2
# 9     INBU     2
# 10    NOCA     2
# # ... with 11 more rows
```

An iterative process

Although presented as a series of steps, data management is an iterative process. This usually only comes to light once you start doing (basic) statistical analyses. At that point you might need to

- fix errors in raw non-digital data (if you can)
- fix errors in raw digital data
- fix errors in tidying scripts
- fix errors analysis scripts
- update the raw non-digital format
- update the tidying scripts
- update the analysis scripts
- $\vdots$

You should also plan time to

- document your process
- review (annually) your process and make improvements.

My process and tools

As I'm not the one collecting the raw non-digital data (and typically not digitizing it), my job begins with the backup.

1. Use Git/GitHub for file storage and backup.
2. Create an R package (see `devtools`) for the data from each PI.
    - `data-raw/` contains the data from the PI and scripts to create tidy data
    - `data/` contains the tidy data in a binary R format (`.rda`)
    - `R/data.R` contains metadata for the data, e.g.
        - description including units
        - references
        - contact info
3. Use R to write all scripts.

Advantages:

- Using a version control system, e.g. Git, provides automatic documentation of changes and an ability to revert to a previous state at any time.
- Using an R package, allows R users to quickly access data, e.g.

```
devtools::install_github("ISU-STRIPS/STRIPS") # only need to do once
library(STRIPS)
```

Examples

STRIPS project:

- https://github.com/ISU-STRIPS/STRIPS
- https://github.com/ISU-STRIPS/STRIPSMeta
- https://github.com/ISU-STRIPS/STRIPSONeal
- https://github.com/ISU-STRIPS/STRIPSLiebman
- https://github.com/ISU-STRIPS/STRIPSSchulte/blob/master/tests/testthat/test-counts.R
- https://github.com/ISU-STRIPS/STRIPSSchulte/blob/master/R/data.R

Gas mileage:

- https://github.com/jarad/ToyotaSiennaGasMileage

Flash card data:

- https://github.com/jarad/flashcardData