Data Management

Karl Broman

Biostatistics & Medical Informatics
Univ. Wisconsin–Madison

kbroman.org

Slides: bit.ly/datamgmt2018

Outline

- Organizing data within spreadsheets
- File organization and naming
- Data cleaning

Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

The results in Table 1 don’t seem to correspond to those in Figure 2.

Where did we get this data file?

Why did I omit those samples?

How did I make that figure?

1. Organizing data in spreadsheets

Organize data for computers

Improve this arrangement?

	A	B	C	D	E	F	G
1							
2	1min						
3			Normal			Mutant	
4		10−05−16	10−12−16	10−19−16	10−05−16	10−12−16	10−19−16
5	B6	146.6	138.6	155.6	166	179.3	186.9
6	BTBR	245.7	240	243.1	177.8	171.6	188.1
7							
8	5min						
9			Normal			Mutant	
10		10−05−16	10−12−16	10−19−16	10−05−16	10−12−16	10−19−16
11	B6	333.6	353.6	408.8	450.6	474.4	423.8
12	BTBR	514.4	610.6	597.9	412.1	447.4	446.5

Improved arrangement

	A	B	C	D	E
1	strain	genotype	treatment_time	date	response
2	B6	Normal	1min	2016−10−05	146.6
3	B6	Normal	1min	2016−10−12	138.6
4	B6	Normal	1min	2016−10−19	155.6
5	B6	Mutant	1min	2016−10−05	166
6	B6	Mutant	1min	2016−10−12	179.3
7	B6	Mutant	1min	2016−10−19	186.9
8	BTBR	Normal	1min	2016−10−05	245.7
9	BTBR	Normal	1min	2016−10−12	240
10	BTBR	Normal	1min	2016−10−19	243.1

Organizing data in spreadsheets

- Make it a rectangle (rows = observations, cols=variables)
- Use a single header row; avoid spaces.
- Be consistent.
- Use care about dates.
- Put just one thing in a cell.
- Fill in all cells.
- Explicit code for missing values (e.g. `-` or `N/A`)
- No calculations/graphs in the raw data files.
- Don't use font color or highlighting as data.
- Make backups.
- Use data validation to avoid data entry mistakes.
- Save the data in plain text files.

Broman & Woo, Am Stat 78:2–10, 2018

2. Organizing projects

File organization and naming
are powerful weapons against chaos.

– Jenny Bryan

2. Organizing projects

Your closest collaborator is you six months ago,
but you don’t reply to emails.

(paraphrasing Mark Holder)

2. Organizing projects

Have sympathy for your future self.

How to organize these files?

```
Raw phenotype data

    CPL_Rosetta_Lipids_FINAL.xlsx
    Complete F2 Liver TG Set.xlsx
    D2O_Summary_of_All_F2_Samples_MF_30July2009.xlsx
    FINAL_RBM_DATA_102989_26Sep2007.xlsx
    Mapped_Urine_Plasma_Data_to_Statgen.xlsx
    Necropsy_Tracking_Report_rk61412.xlsx
    Necropsy_Tracking_Report_rk_052912_atb.xlsx
    Necropsy_Tracking_Report_rk_2011-04-26.xlsx
    Original_Necropsy_Tracking_Report_rk.xlsx
    RBM_Tube_Number_Key.xlsx

Raw genotype data

    Final Fit1_Filtered_Assay_Allele_Signals_and_Genotypes_18Sep.txt

Converted data

    clinpheno.csv
    detailed_genotypes.csv
    genotypes4rqtl.csv
    genotypes_karl.csv
```

```
R scripts to organize data

    check_necropsy_files.R
    check_necropsy_files_2012-06-02.R
    combine_pheno.R
    combine_pheno2.R
    combine_pheno3.R
    compareData.R
    func.R
    prepData.R

Analysis

    fig1.png
    fig2.png
    fig3.png
    fig4.png
    fig5.png
    fig6.png
    fig7.png
    fig8.png
    scanone_clinphe.Rmd
    scanone_clinphe.html
```

Be consistent

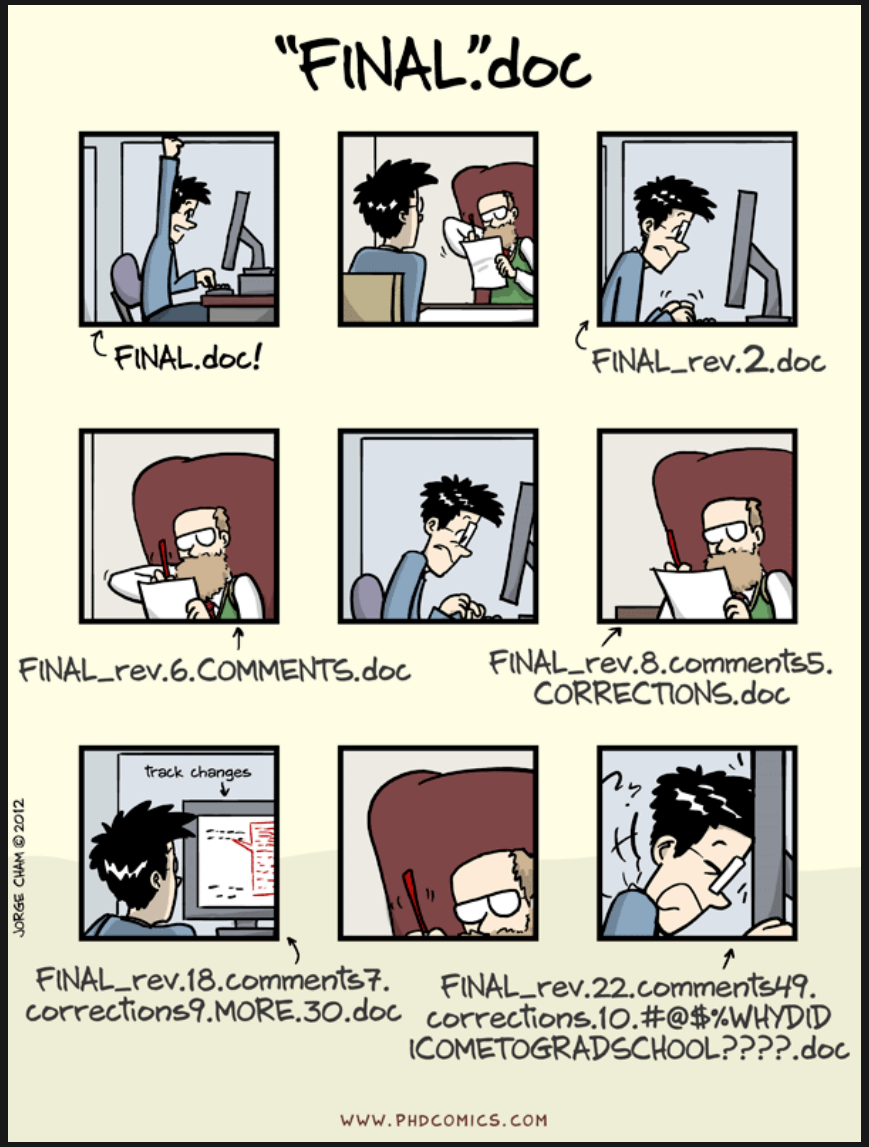
```
RawData/                    Notes/
CleanData/                  Refs/

Python/                     ReadMe.txt
R/                          ToDo.txt
Ruby/
```

Chaos

```
AimeeNullSims/     Deuterium/              Ping/
AimeeResults/      ExtractData4Gary/       Ping2/
AnnotationFiles/   FromAimee/              Ping3/
Brian/             GoldStandard/           Ping4/
Chr6_extrageno/    HumanGWAS/              Play/
Chr6_segdis/       Insulin/                Prdm9/
ChrisPlaisier/     Int2_for_Mark/          RBM_PlasmaUrine_2012-03-08/
Code4Aimee/        Islet_2011-05/          Slco1a6/
CompAnnot/         MappingProbes/          StudyLineupMethods/
CondScans/         MultiProbes/            kidney_chr6.R
D2O_2012-02-14/    NewMap/                 pck2_sucla2.R
D2O_cellcycle/     Notes/                  penalties.txt
D2Ocorr/           NullSims/               transeQTL4Lude/
Data4Aimee/        NullSims_2009-09-10/
Data4Tram/         PepIns_2012-02-09/
```

No “final” in file names



No “final” in file names

```
Deprecated/                              hypo_prcomp.RData
ReadMe.txt                               islet_int1_final.RData
adipose_int1_final.RData                 islet_int2_final.RData
adipose_int2_final.RData                 islet_mlratio_final.RData
adipose_mlratio_final.RData              islet_mlratio_nqrank_final.RData
adipose_mlratio_nqrank_final.RData       islet_prcomp.RData
adipose_prcomp.RData                     kidney_int1_final.RData
aligned_geno_with_pmap.RData             kidney_int2_final.RData
batches_final.RData                      kidney_mlratio_final.RData
batches_raw_final.RData                  kidney_mlratio_nqrank_final.RData
cpl_final.RData                          kidney_prcomp.RData
d2o_final.RData                          lipomics_final_rev2.RData
gastroc_int1_final.RData                 liverTG_final.RData
gastroc_int2_final.RData                 liver_int1_final.RData
gastroc_mlratio_final.RData              liver_int2_final.RData
gastroc_mlratio_nqrank_final.RData       liver_mlratio_final.RData
gastroc_prcomp.RData                     liver_mlratio_nqrank_final.RData
hypo_int1_final.RData                    liver_prcomp.RData
hypo_int2_final.RData                    mirna_final.RData
hypo_mlratio_final.RData                 necropsy_final_rev2.RData
hypo_mlratio_final_old.RData             plasmaurine_final_rev.RData
hypo_mlratio_nqrank_final.RData          pmark.RData
hypo_mlratio_nqrank_final_old.RData      rbm_final.RData
hypo_omit.RData
```

No “final” in file names

```
Deprecated/                                hypo_prcomp.RData
ReadMe.txt                                 islet_int1_final.RData
adipose_int1_final.RData                   islet_int2_final.RData
adipose_int2_final.RData                   islet_mlratio_final.RData
adipose_mlratio_final.RData                islet_mlratio_nqrank_final.RData
adipose_mlratio_nqrank_final.RData         islet_prcomp.RData
adipose_prcomp.RData                       kidney_int1_final.RData
aligned_geno_with_pmap.RData               kidney_int2_final.RData
batches_final.RData                        kidney_mlratio_final.RData
batches_raw_final.RData                    kidney_mlratio_nqrank_final.RData
cpl_final.RData                            kidney_prcomp.RData
d2o_final.RData                            lipomics_final_rev2.RData
gastroc_int1_final.RData                   liverTG_final.RData
gastroc_int2_final.RData                   liver_int1_final.RData
gastroc_mlratio_final.RData                liver_int2_final.RData
gastroc_mlratio_nqrank_final.RData         liver_mlratio_final.RData
gastroc_prcomp.RData                       liver_mlratio_nqrank_final.RData
hypo_int1_final.RData                      liver_prcomp.RData
hypo_int2_final.RData                      mirna_final.RData
hypo_mlratio_final.RData                   necropsy_final_rev2.RData
hypo_mlratio_final_old.RData               plasmaurine_final_rev.RData
hypo_mlratio_nqrank_final.RData            pmark.RData
hypo_mlratio_nqrank_final_old.RData        rbm_final.RData
hypo_omit.RData
```

Naming files

- Concise
- Meaningful
- No spaces or special characters except underscores and hyphens
- Dates like 2018-09-05 (or 20180905)
- Sortable
- Be consistent

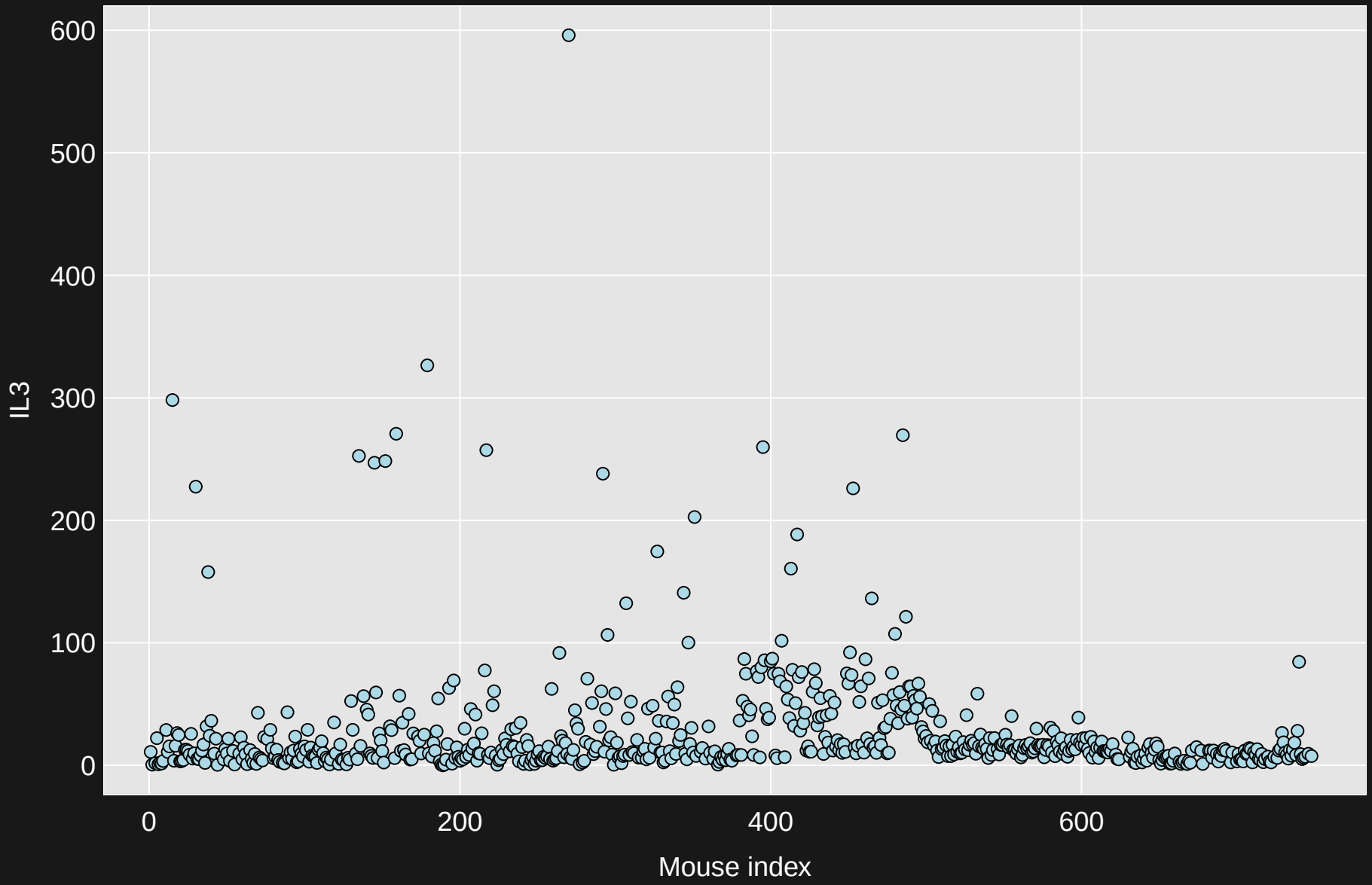
A few more things

- Documentation
  - Detailed electronic notes
  - Have sympathy for your future self
  - Maintained to match the material
- Meta-data
  - Data about the data
  - Yet another spreadsheet
  - Data dictionary: explain variables, abbreviations, units, codes
- Versions
  - Save everything
  - Number files like `_v1`, `_v2`, ...
- Backups
  - Multiple places (external drives, cloud, server)
  - Needs to be automatic

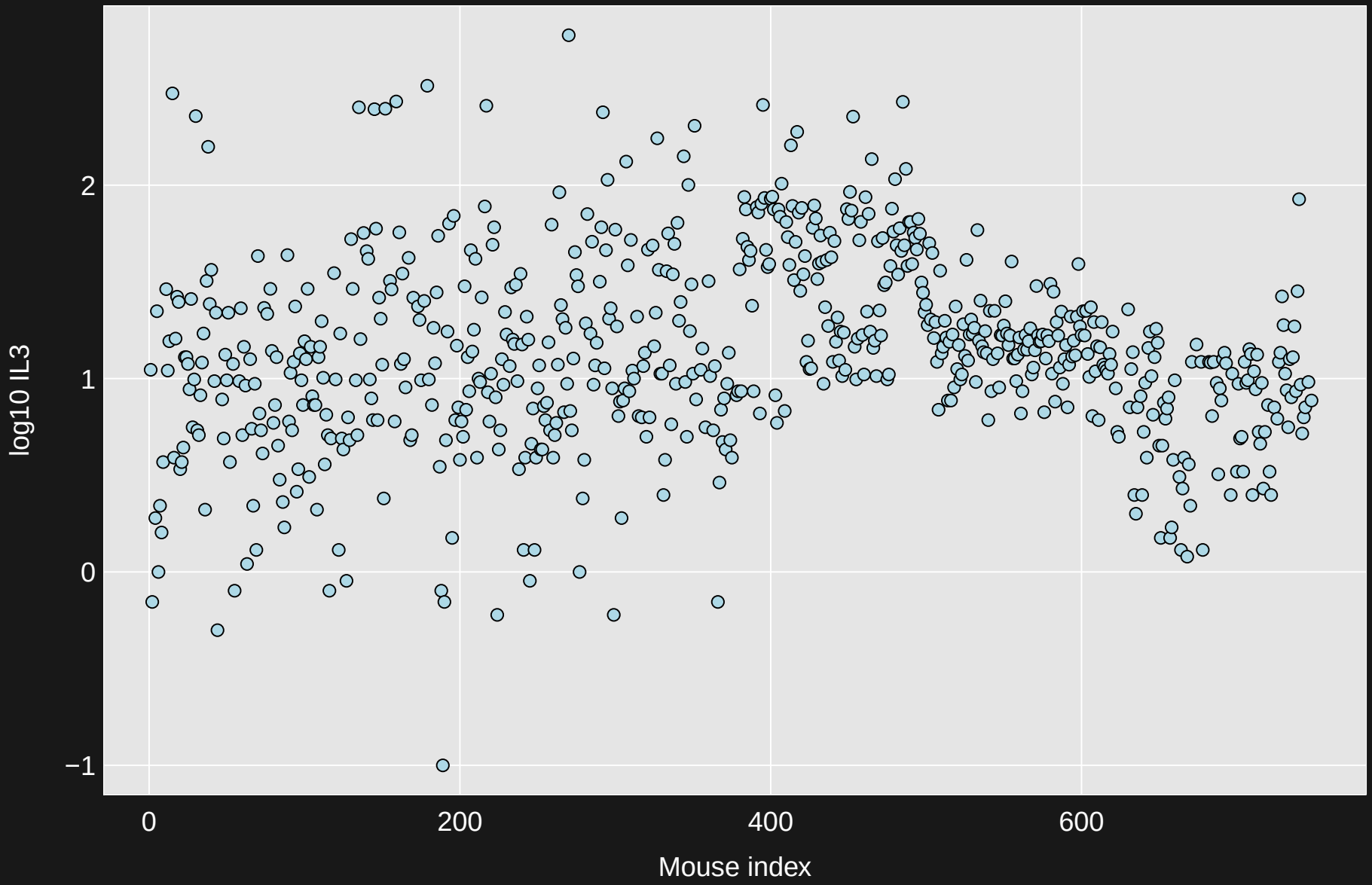
3. Data cleaning

- What might have gone wrong?
- How could it be revealed?
- Make lots of plots
  - scatterplots
  - plots against time
  - consider taking logs
- Check consistency between files

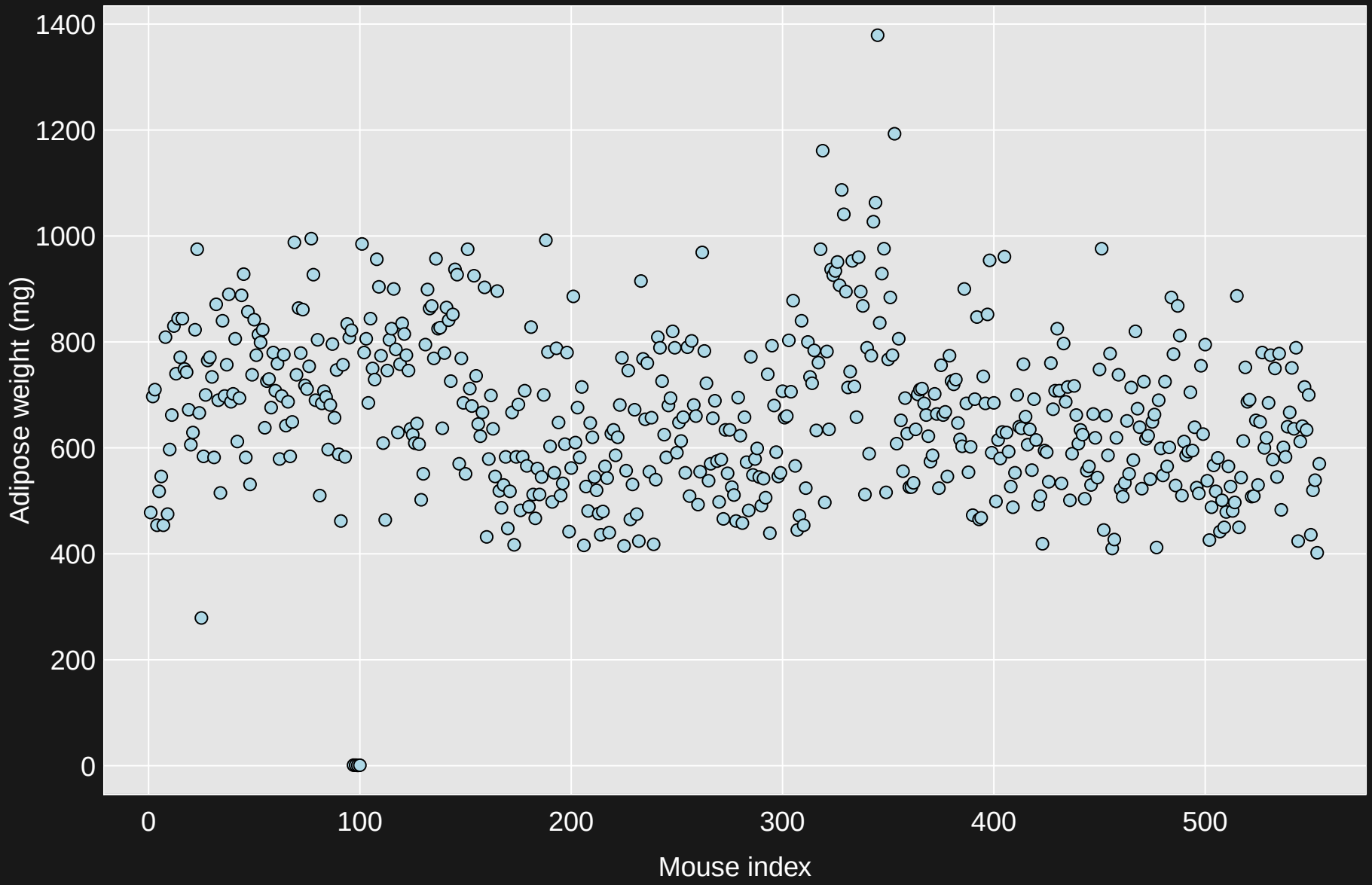
Batch effect



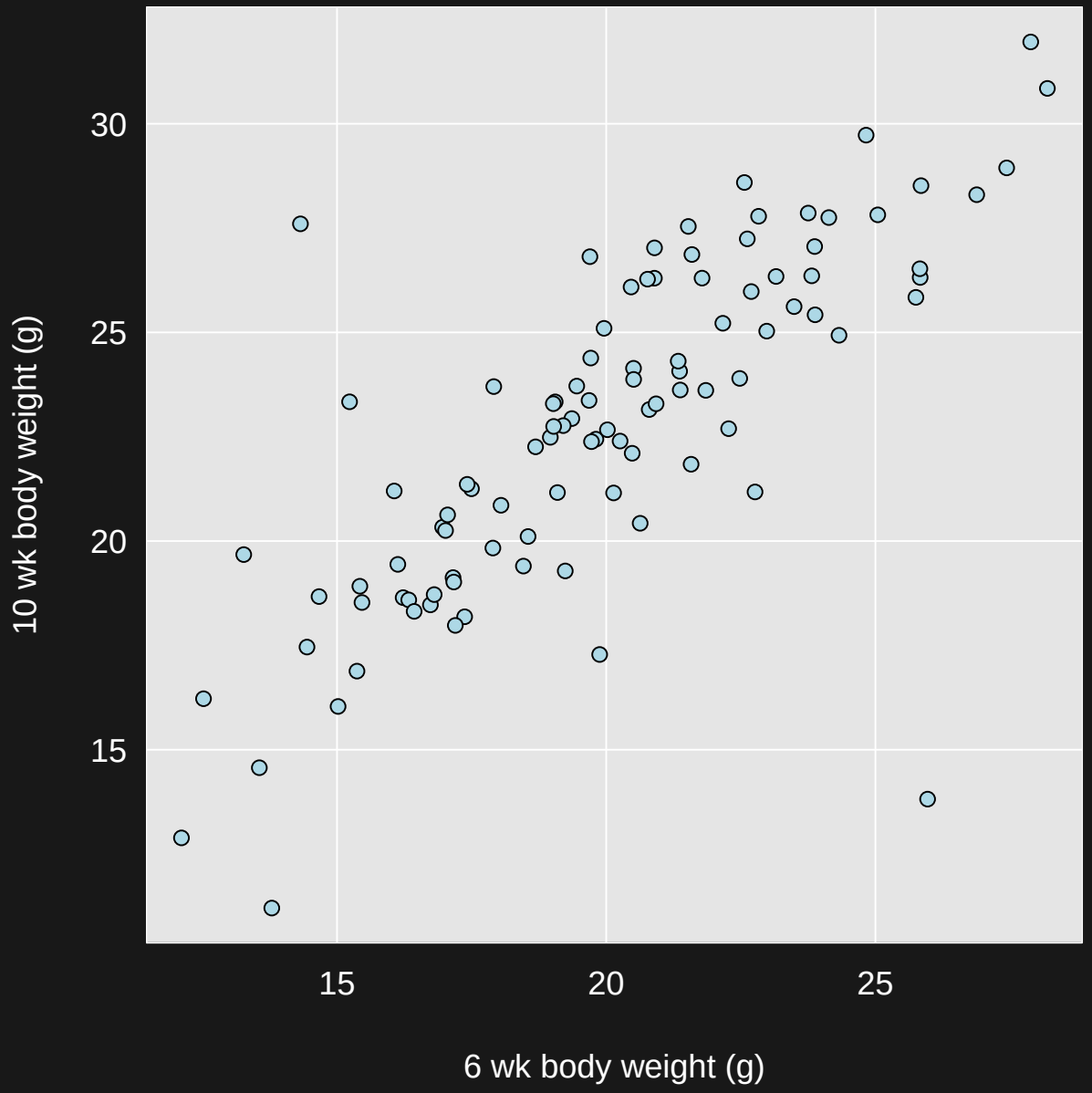
Batch effect



Messed up units



Outliers



The most important tool is the mindset,
when starting, that the end product
will be reproducible.

– Keith Baggerly

Resources

- These slides: `bit.ly/datamgmt2018`
- Briney (2015) Data management for researchers
- Research Data Services, `researchdata.wisc.edu`
- Data Science Hub, `datascience.wisc.edu`
- Data Carpentry workshops, `datacarpentry.org`