Data Management

Karl Broman

Biostatistics & Medical Informatics
Univ. Wisconsin–Madison

`kbroman.org`

Slides: `bit.ly/datamgmt2019`

Outline

- Organizing and naming files
- Organizing data within spreadsheets
- Cleaning data

> 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?

Which image goes with which experiment?

How did I make that figure?

In what order do I run these scripts?

“Your script is now giving an error.”

“The attached is similar to the code we used.”

1. Organizing projects

> File organization and naming
> are powerful weapons against chaos.
>
> – Jenny Bryan

1. Organizing projects

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

(paraphrasing Mark Holder)

1. Organizing projects

Have sympathy for your future self.

How to organize these files?

```
Raw phenotype data                                                      R scripts to organize data

    CPL_Rosetta_Lipids_FINAL.xlsx                                           check_necropsy_files.R
    Complete F2 Liver TG Set.xlsx                                           check_necropsy_files_2012-06-02.R
    D2O_Summary_of_All_F2_Samples_MF_30July2009.xlsx                        combine_pheno.R
    FINAL_RBM_DATA_102989_26Sep2007.xlsx                                    combine_pheno2.R
    Mapped_Urine_Plasma_Data_to_Statgen.xlsx                                combine_pheno3.R
    Necropsy_Tracking_Report_rk61412.xlsx                                   compareData.R
    Necropsy_Tracking_Report_rk_052912_atb.xlsx                             func.R
    Necropsy_Tracking_Report_rk_2011-04-26.xlsx                             prepData.R
    Original_Necropsy_Tracking_Report_rk.xlsx
    RBM_Tube_Number_Key.xlsx                                            Analysis

Raw genotype data                                                           fig1.png
                                                                            fig2.png
    Final Fit1_Filtered_Assay_Allele_Signals_and_Genotypes_18Sep.txt        fig3.png
                                                                            fig4.png
Converted data                                                              fig5.png
                                                                            fig6.png
    clinpheno.csv                                                           fig7.png
    detailed_genotypes.csv                                                  fig8.png
    genotypes4rqtl.csv                                                      scanone_clinphe.Rmd
    genotypes_karl.csv                                                      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/
```

Choose good names for things

```
betw_tissue_corr.R     expr_scatterplot_allprobes.R  gve_similarity_alltissues.R
coatcolor_lod.R        expr_scatterplots_dup.R       gve_similarity.R
colors.R               expr_scatterplots_mix.R       gve_supp.R
cover_fig.R            expr_scatterplots_swap.R      insulin_lod.R
eqtl_counts_10.R       expr_swaps.R                  local_eqtl_locations.R
eqtl_counts.R          func.R                        my_plot_map.R
eve_hist.R             genotype_plates.R             my_plot_scanone.R
eve_scheme.R           gve_hist.R                    sex_vs_X.R
eve_similarity.R       gve_new.R                     xchr_fig.R
eve_similarity_supp.R  gve.R                         xist_and_y.R
expr_corr_dup.R        gve_scheme.R
expr_corr_mix.R        gve_similarity_2ndbest.R
```

Choose good names for things

```
fig1.png        fig5.png
fig10.png       fig6.png
fig2.png        fig7.png
fig3.png        fig8.png
fig4.png        fig9.png
```

Choose good names for things

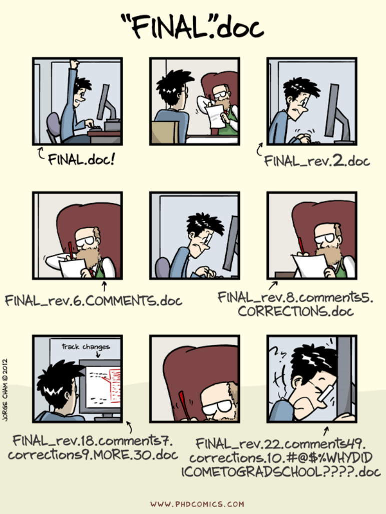
- Machine readable
  - No spaces
  - No special characters except _ and -
- Human readable
  - Explain the contents
- Consistent
  - Name similar files in a similar way
- Make use of computer's sorting
  - pad numbers with 0's (e.g., `01`, `02`, ...)
  - start with general grouping, then more specific
  - dates like `2019-05-14`

xkcd.com/1179

Choose good names for things

```
0_vcf2db.R
1_prep_geno.R
2_prep_pheno_clin.R
2_prep_pheno_otu.R
3_prep_covar.R
4_prep_analysis_pheno_clin.R
4_prep_analysis_pheno_otu.R
5_scans.R
6_grab_peaks.R
7_find_nearby_peaks.R
```

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
```

Choose good names for things

```
batches_raw_v1.rds               geneexpr_mlratio_gastroc_v2.rds
batches_v1.rds                   geneexpr_mlratio_hypo_v1.rds
clinical_cpl_v2.rds              geneexpr_mlratio_hypo_v2.rds
clinical_d2o_v2.rds              geneexpr_mlratio_islet_v2.rds
clinical_lipomics_v4.rds         geneexpr_mlratio_kidney_v2.rds
clinical_liverTG_v2.rds          geneexpr_mlratio_liver_v2.rds
clinical_mirna_v2.rds            geneexpr_mlratio_nqrank_adipose_v2.rds
clinical_necropsy_v4.rds         geneexpr_mlratio_nqrank_gastroc_v2.rds
clinical_plasmaurine_v3.rds      geneexpr_mlratio_nqrank_hypo_v1.rds
clinical_rbm_v2.rds              geneexpr_mlratio_nqrank_hypo_v2.rds
Deprecated/                      geneexpr_mlratio_nqrank_islet_v2.rds
geneexpr_int1_adipose_v2.rds     geneexpr_mlratio_nqrank_kidney_v2.rds
geneexpr_int1_gastroc_v2.rds     geneexpr_mlratio_nqrank_liver_v2.rds
geneexpr_int1_hypo_v2.rds        geneexpr_omit_hypo.rds
geneexpr_int1_islet_v2.rds       geneexpr_prcomp_adipose_v2.rds
geneexpr_int1_kidney_v2.rds      geneexpr_prcomp_gastroc_v2.rds
geneexpr_int1_liver_v2.rds       geneexpr_prcomp_hypo_v2.rds
geneexpr_int2_adipose_v2.rds     geneexpr_prcomp_islet_v2.rds
geneexpr_int2_gastroc_v2.rds     geneexpr_prcomp_kidney_v2.rds
geneexpr_int2_hypo_v2.rds        geneexpr_prcomp_liver_v2.rds
geneexpr_int2_islet_v2.rds       geno_aligned_w_pmap.rds
geneexpr_int2_kidney_v2.rds      geno_pmark.rds
geneexpr_int2_liver_v2.rds       ReadMe.txt
geneexpr_mlratio_adipose_v2.rds
```

Document your work

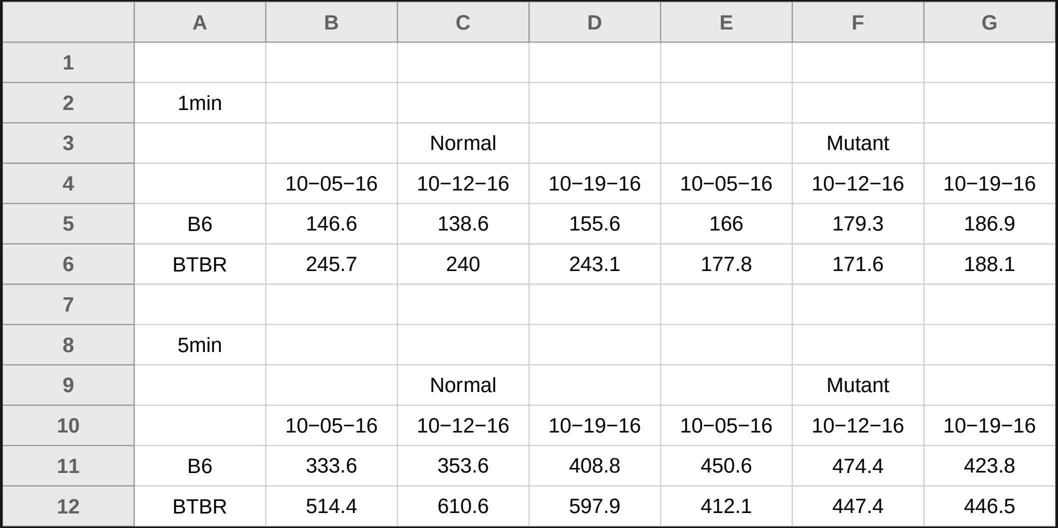
- What is all of this stuff?
- What was your analysis process?

→ `ReadMe` files

2. Organizing data in spreadsheets

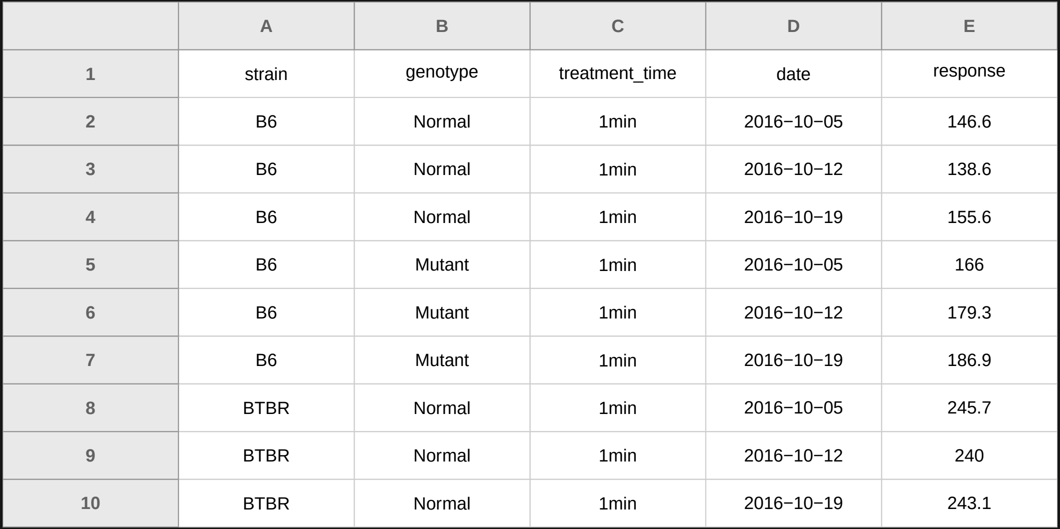
Organize data for computers

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.

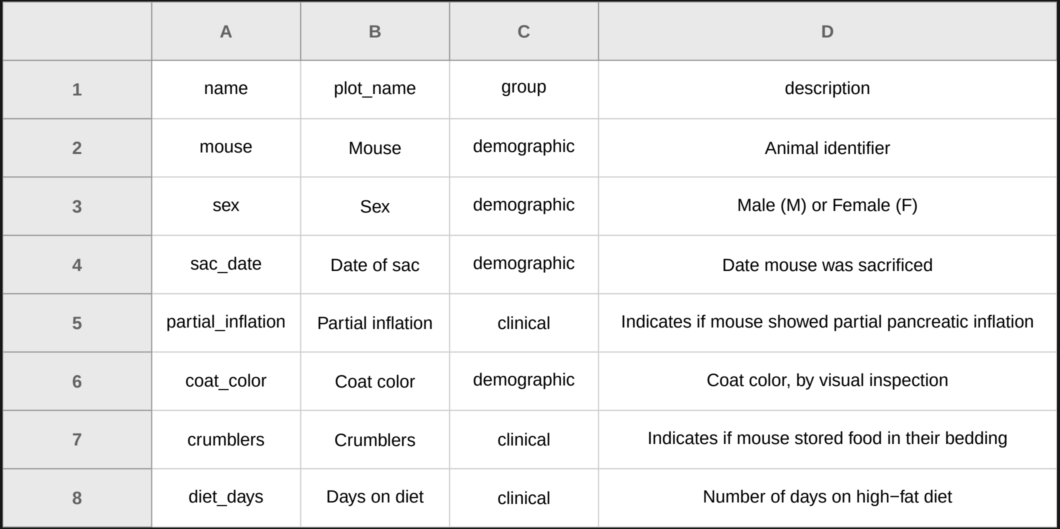
Broman & Woo, Am Stat 78:2–10, 2018

“What the heck is ‘`FAD_NAD SI 8.3_3.3G`’?”

Metadata

- Create a data dictionary
  - Explain each column
  - Include different versions of the variable names (compact vs descriptive)
  - Units
  - Allowable values
- The metadata are data
  - Make it a rectangle

Data dictionary

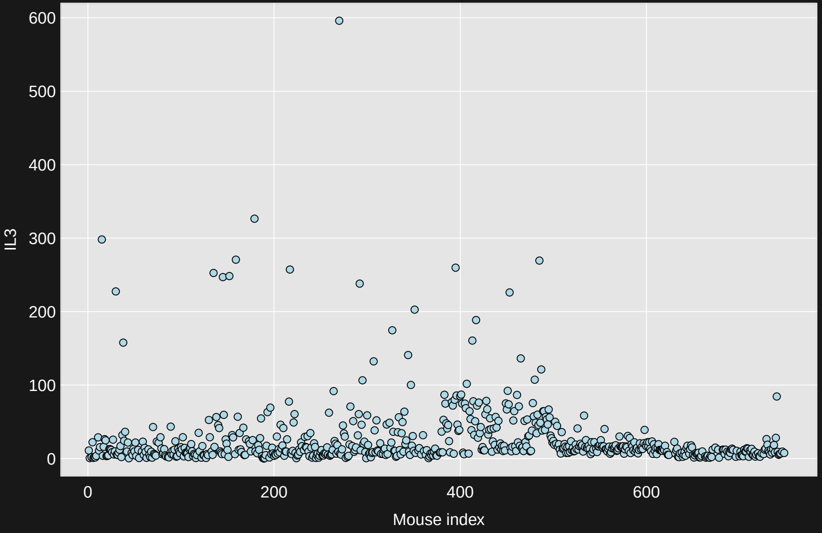


	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high−fat diet

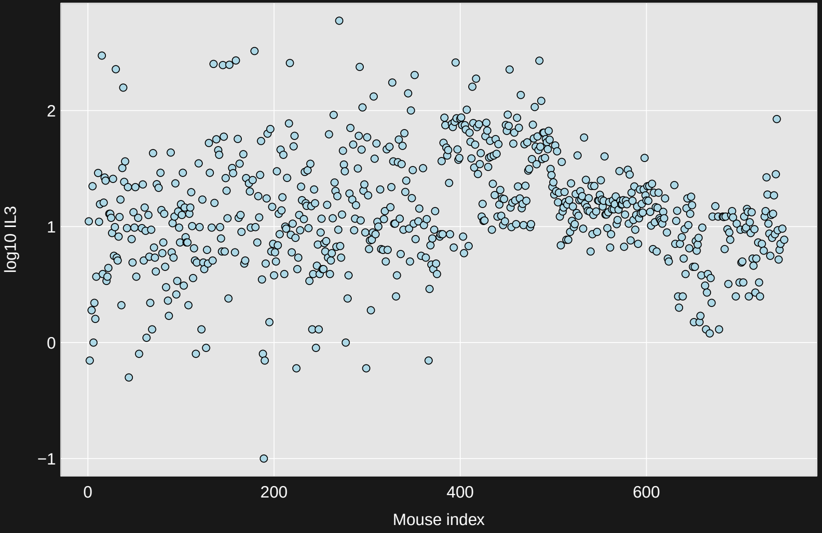
3. Data cleaning

- 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
- Outliers
  - Real or error?
  - Are the results affected?

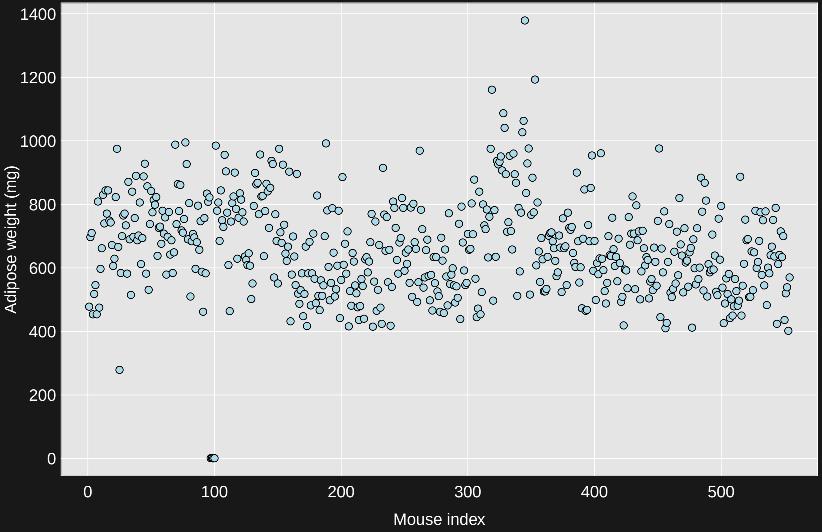
Batch effect



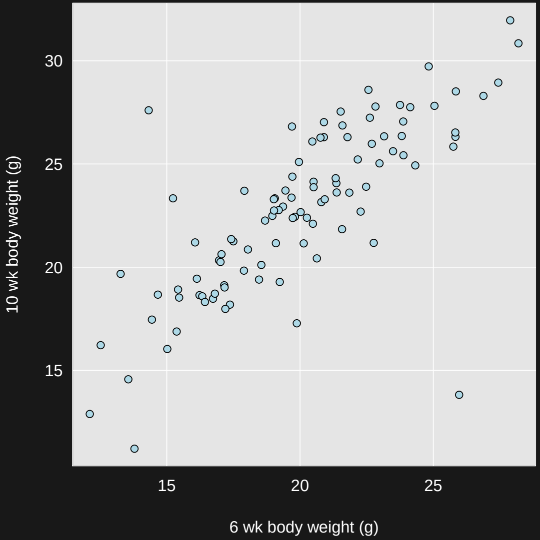
Batch effect



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/datamgmt2019`
- Briney (2015) Data management for researchers
- Research Data Services, `researchdata.wisc.edu`
- Data Science Hub, `datascience.wisc.edu`
- Data Carpentry workshops, `datacarpentry.org`