Data organisation practices

Introduction to Data Management Practices course

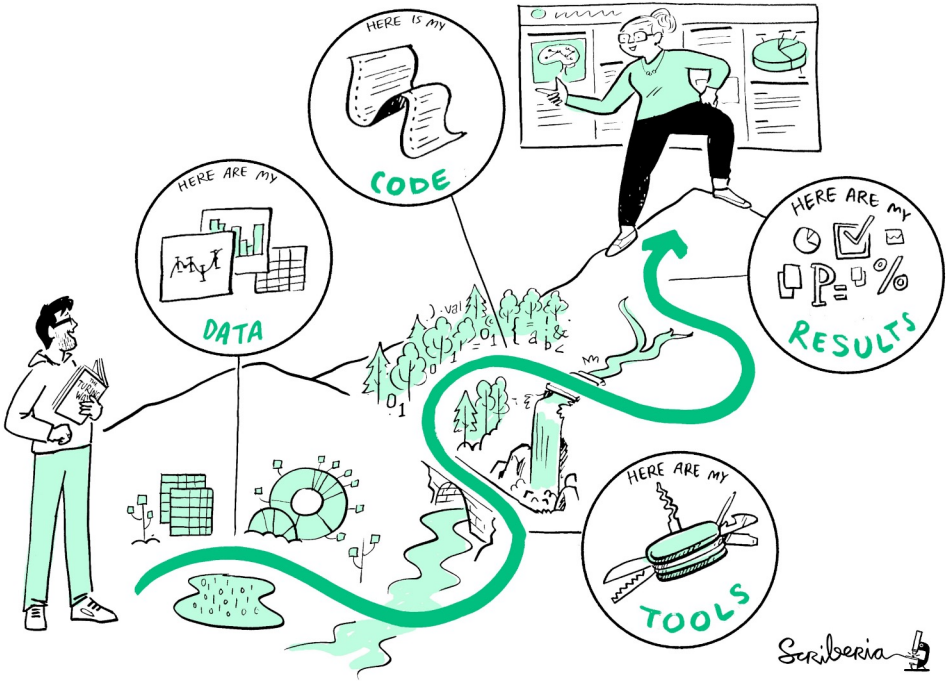
NBIS DM Team

data@nbis.se

https://nbisweden.github.io/module-organising-data-dm-practices/

Objectives

- What to consider for maintaining data organization strategies in a project
- What to consider when settling for a file structure
- Understanding good practices for data storage, processing and documentation (**FAIR-ification**)



Credit: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

Welcome to Science!

You have been recruited to the Famous lab!

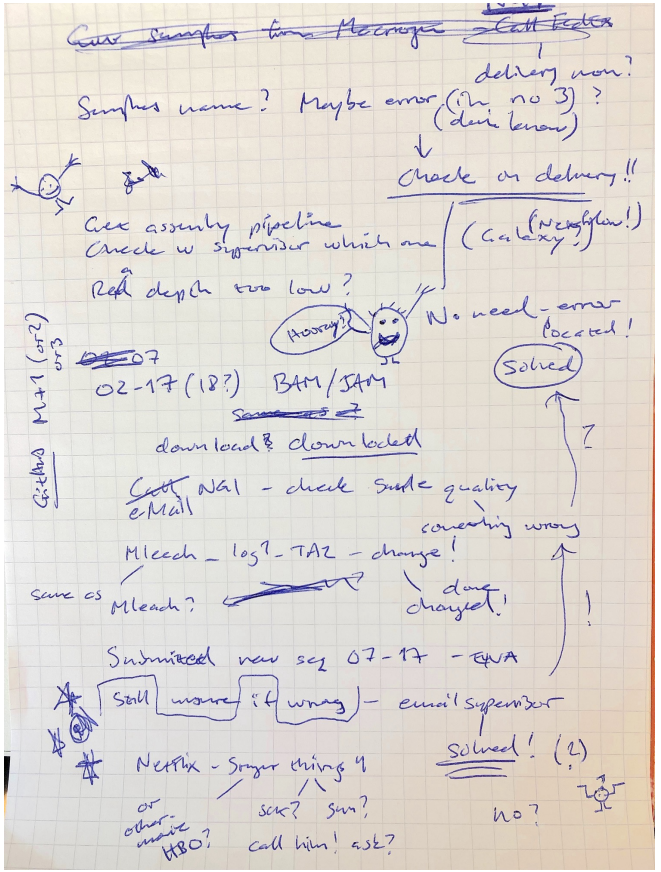
Your research project is a continuation of previous work by PhD, Wang Fang (王芳).

You inherit a zipped folder, and a digital copy of the lab notes.

The road to success is open!

Welcome to Science!

… And this is what you get…

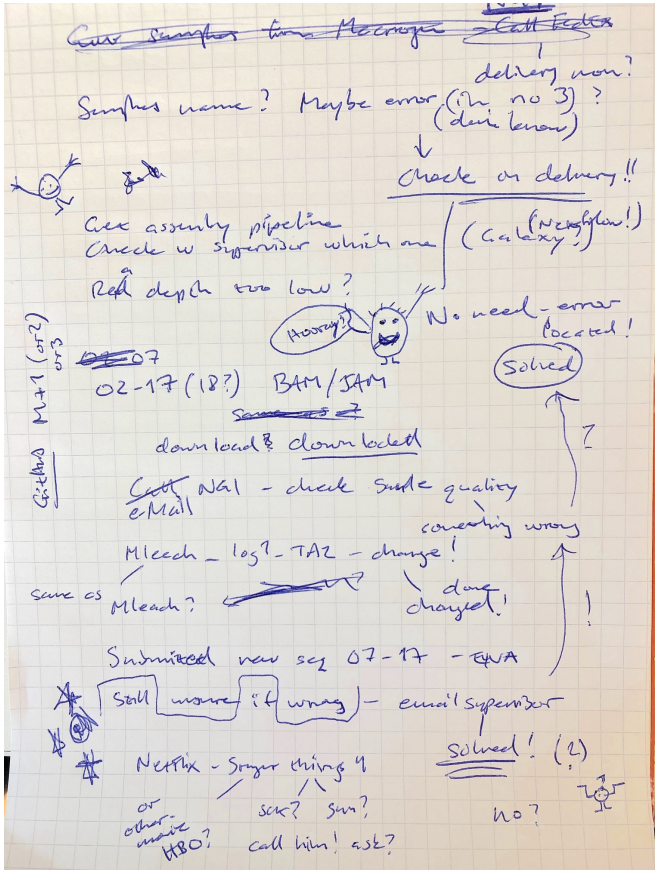


Exercise 1

Can you list at least five major issues with the lab documentation in the image?

Welcome to Science!

… And this is what you get…



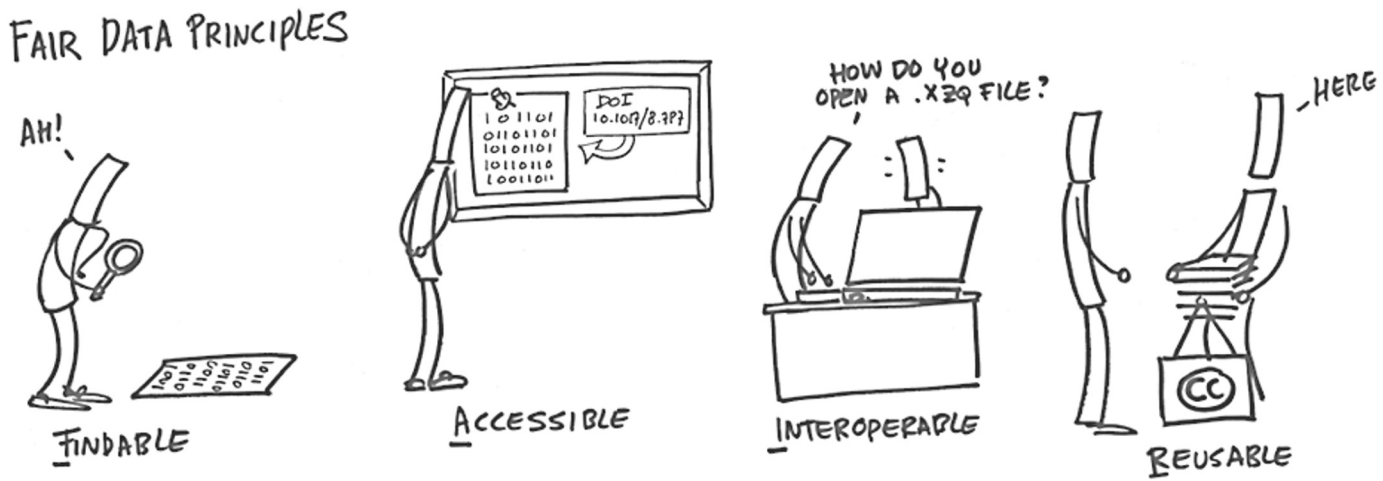
Exercise 2

What kind of general questions does the information raise about the work done in the lab?

Importance of good records

- Why do we need to keep good quality records?
- Ensures data, analysis and results to be transparent, reproducible and traceable – Accountability!
- Keeping good records prevents issues, misunderstandings. Quality of subsequent research
  In cumulative science mistakes can result in cascade effects
- Reduces the risk of data mistakes, data manipulation and research fraud
- Promotes open science and safeguards integrity of science itself
- Good records promote data and documentation being …

FAIR!



Adopting good practices for data organization, makes research data more **FAIR**

Image: https://book.fosteropenscience.eu/

Data Management Recipients

- Colleagues
- Scientific community
- Society
- Yourself

People I collaborate with must understand what I do with the data

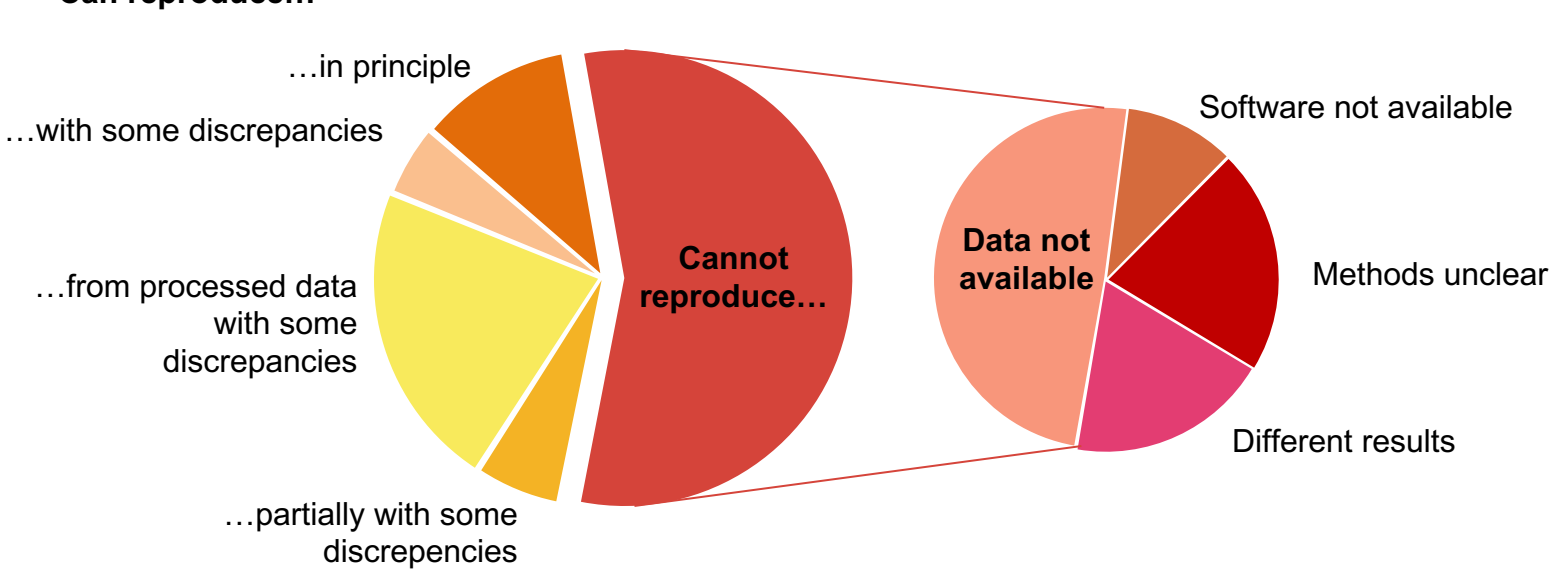
Scientists wanting to reuse or review my data can find and understand the data

The society funding my research have a right to know what happens to the data

Your future You will not always remember what Your present You decided today

Reprocucibility

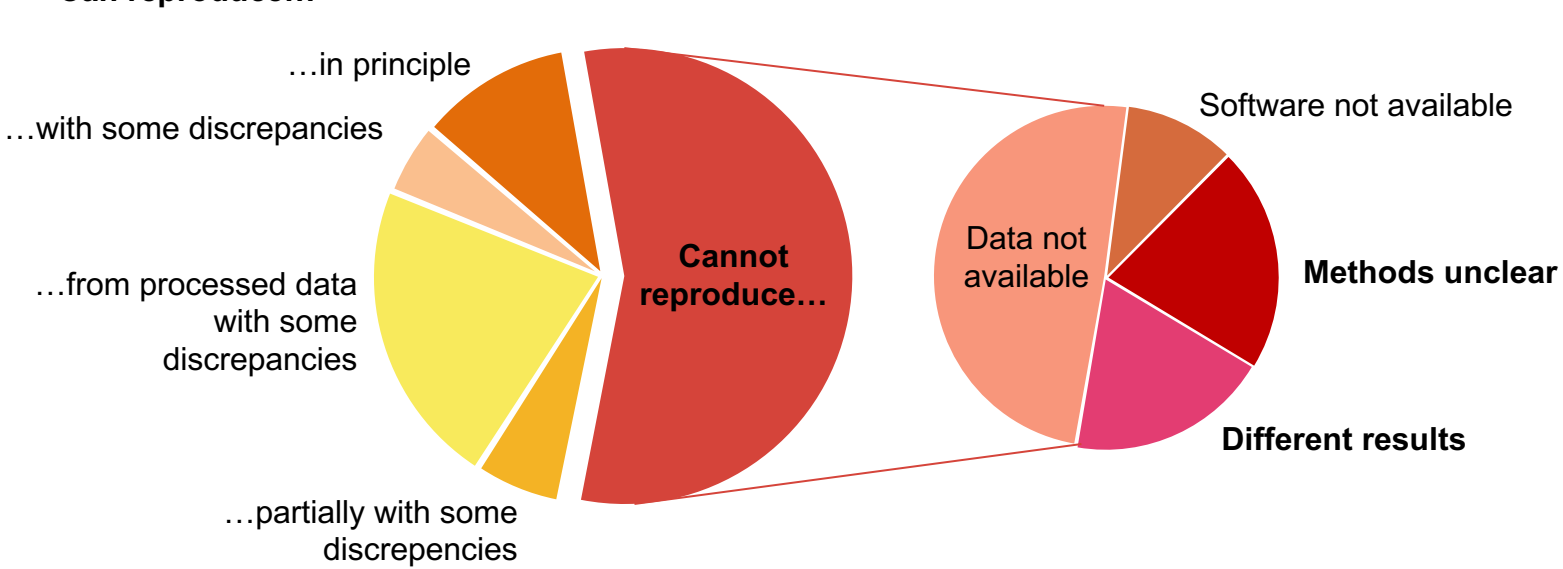
Reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:



Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* **41** (2009) doi:10.1038/ng.295

Reprocucibility

Reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:



Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* **41** (2009) doi:10.1038/ng.295

Principles for good records

Contents in **protocols** can include

- Name, affiliation and contact information
- Originator of protocol (if not you)
- Information on why and how experiment was done
- Health and safety advice (and technical advice)
- Required software, materials and instruments
- Being self-explanatory
- Describe mistakes (for others to avoid repeatning)
- Reference ethical application (if applicable)

Protocols and lab notes should both be…

- Detailed
- Up to date
- Accurate
- Easy to understand

Contents in **lab notes** can include

- Your name and affiliation
- Details on what, when and how
- What project the experiment is part of
- Lot and batch numbers for consumables
- Information on metadata collected
- Post-outcome treatment of data
- Interpretation of outcome and outlook/plans

Exercise 3

Test yourself on record keeping statements

1. Analogue and digital records makes information equally findable.
2. New information in digital records can be easily shared with other users.
3. Analogue records can be kept safe from any physical accidents.
4. All researchers in a shared lab should have access to the same platform for keeping records and taking notes.
5. Digital records should follow the same backup strategy as the data they describe.

Exercise 3

Test yourself on record keeping statements

1. Analogue and digital records makes information equally findable. (F)
2. New information in digital records can be easily shared with other users. (T)
3. Analogue records can be kept safe from any physical accidents. (F)
4. All researchers in a shared lab should have access to the same platform for keeping records and taking notes. (T)
5. Digital records should follow the same backup strategy as the data they describe. (T)

FAIR by README

A file usually defined as the starting point of information about something (attracts attention!)

FAIRify your research by using them as documentation files for:

Folder level – Explaining folder contents, naming, file history, organisation/structure etc

Data – Explaining file names and contents

README in Markdown (.md)

- Allows text and content formatting without interference
- Highly compatible with e.g. GitHub
- Allows inclusion of comments without having to visualize them
- Easily editable and versatile
- Does not require particular skills

Exercise 4

Discussion

Think of an example where you would have benefited from having access to a README-file when working with data.

Describe to your neighbor what you would have wanted such a file to contain.

Backup

Data and hardware failure is always a threat. Plan early for potential failure!

Good to know for backup planning purposes:

- Data sensitivity
- Ease of access
- File sizes
- Overall data volumes
- Data life cycle in project

- Nearly all data, metadata and project information necessary to understand your analysis and results require some sort of backup strategy.
- Try to keep backup in three separate locations, on at least two different kind of media (server, portable hard drive, cloud). Consider off-site backups.
- Never backup your data on portable drives only (SSD or ATA), and particularly not on USB sticks!
- Robust backups need to be automated.

Exercise 5

Discussion

Discuss in pairs the validity of the following statements on data backup:

1. I have my most important data backed up on my laptop. I have never experienced a hard drive failure, and my current laptop has a new state-of-the-art hard drive. Therefore, I don't need external backups.
2. All my data is stored in a cloud service.
3. My data is on a portable hard drive. There is a backup of the most important files on a shared USB stick in my research group.
4. My data is on a departmental backup administered by my University. Additionally we have a server for all the data stored in our project.
5. We have no shared backup at all. All members in our research group are responsible for their own data.

Exercise 5

Discussion

Discuss in pairs the validity of the following statements on data backup:

1. Unsafe and not recommended. All hard drives can be subject to failure. In case of failure, all data will be lost.
2. Cloud services can be sufficient as backup, but are not fail safe. It can be sufficient in combinaiton with a secondary backup on e.g. a shared server. For certain types of data (e.g. sensitive information), a cloud service may be outright inappropriate.
3. Not a good solution. Both portable hard drives as well as USB sticks are prone to failure.
4. A good solution in general. Data is stored independently in two separate systems. Centrally administerd services are usually organised in such a way that partial failures does not affect the users.
5. Worst possible alternative. A disaster waiting to happen.

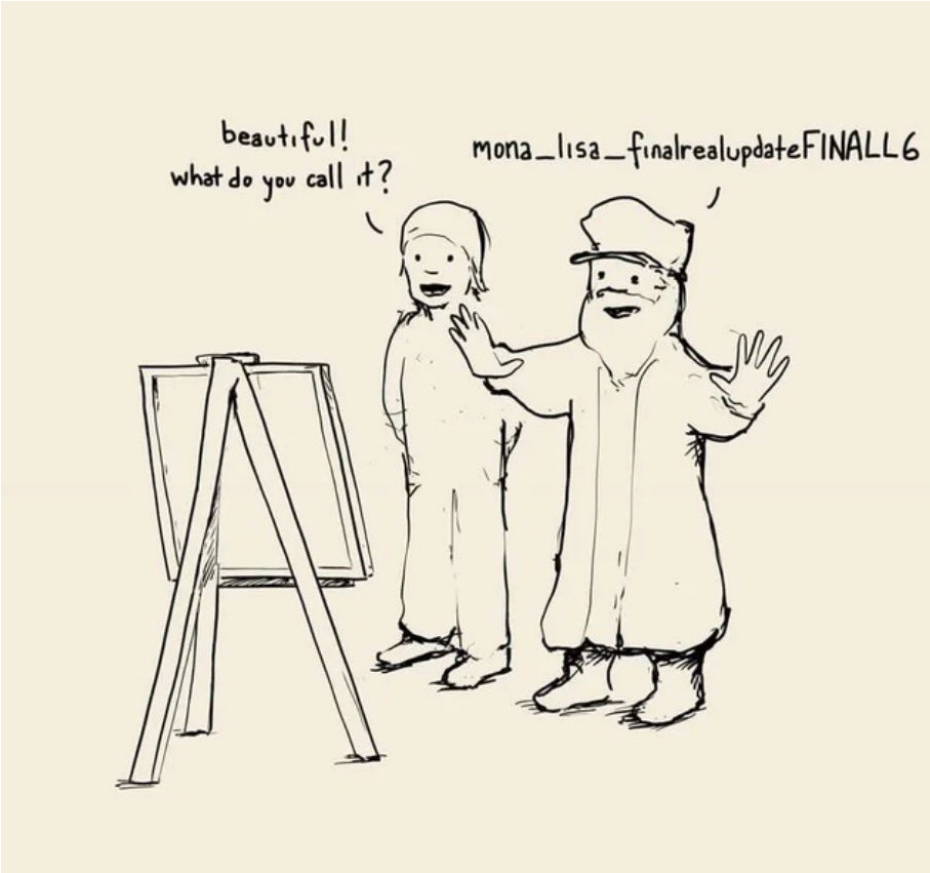
Backup

Creating a backup strategy in 10 steps

1. Find out whether your institution has a backup strategy
2. Determine what you want to back up
3. Decide how many backups you will need and how frequently to back up
4. Decide where backups will be stored
5. Determine how much storage capacity will be needed
6. Determine if there are tools you could use to automate backup
7. Determine how long backups will be kept and how they will be destroyed
8. Determine how personal data will be protected
9. Devise a disaster recovery plan
10. Assign responsibilities

Files and Folders

Why is file organisation important for data management?



https://twitter.com/nathanwpyle/status/1108902487203958784

What level of data organisation will work for me and my project/ team?

File organisation

Benefits of systematically organising research and data files:

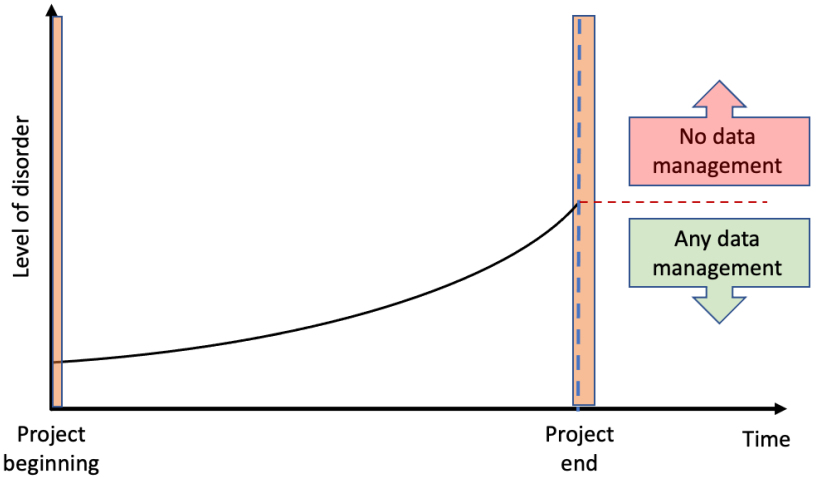
- Easier to locate a file
- Find similar files together
- Moving files becomes much easier
- Easy to identify which files you want to back up
- Keep organised in the long-run
- Increases productivity
- Helps you to keep and maintain a record of the project
- Projects can easily be understood by others (including your future self)

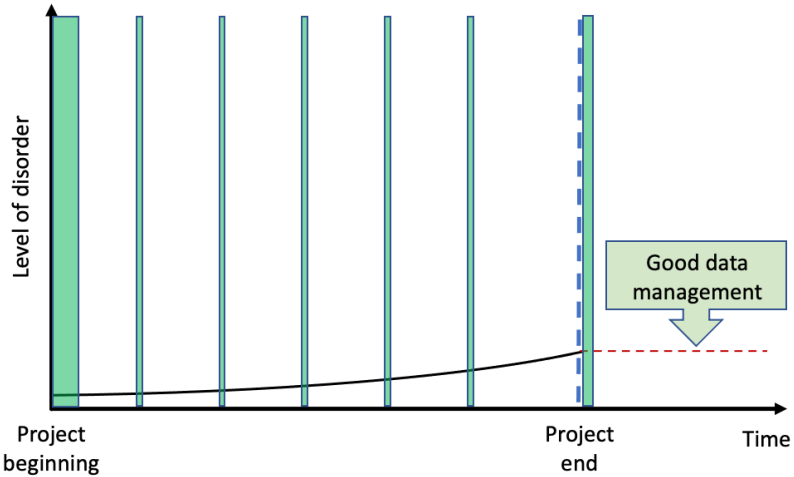
File organisation

Files will become unorganised over time (particularly downloads and/or desktop folders)

Files can multiply across folders and versions, decreasing findability

Organising will reduce clutter and maintanence requirements over time





File and Folder naming

Names for files and folders should be *consistent* and *meaningful to yourself and collaborators*, allow for *easy tracking/searching*, and be *somewhat descriptive of content*.

Example: LD_phyA_off_t04_2020-08-12_norm.xlsx

Based on the name, the file could contain information about:

LD	- Long day sampling, of the
phyA	- Phytochrome A genotype, in a
off	- Medium without sucrose, at
t04	- Time point 4,
2020-08-12	- Sampled on Aug 12th, 2020, with
norm	- Normalised data

But! Not obvious from the letters and words alone. Explanation is required - README.md

Exercise 6

Group discussion

The following example contain files from an imaginary project

- *phyA/phyB* - genotypes
- *sXX* - sample number
- *LD/SD* - light conditions (Long Day, Short Day)
- *on/off* - different growth media (on sucrose, off sucrose)
- *date format* - sample date
- *tXX* - sample timepoint
- *raw*, *norm* - raw or normalised data

2020-07-14_s12_phyB_on_SD_t04.raw.xlsx
2020-07-14_s1_phyA_on_LD_t05.raw.xlsx
2020-07-14_s2_phyB_on_SD_t11.raw.xlsx
2020-08-12_s03_phyA_on_LD_t03.raw.xlsx
2020-08-12_s12_phyB_on_LD_t01.raw.xlsx
2020-08-13_s01_phyB_on_SD_t02.raw.xlsx
2020-7-12_s2_phyB_on_SD_t01.raw.xlsx
AUG-13_phyB_on_LD_s1_t11.raw.xlsx
JUL-31_phyB_on_LD_s1_t03.raw.xlsx
LD_phyA_off_t04_2020-08-12.norm.xlsx
LD_phyA_on_t04_2020-07-14.norm.xlsx
LD_phyB_off_t04_2020-08-12.norm.xlsx
LD_phyB_on_t04_2020-07-14.norm.xlsx
SD_phyB_off_t04_2020-08-13.norm.xlsx
SD_phyB_on_t04_2020-07-12.norm.xlsx
SD_phya_off_t04_2020-08-13.norm.xlsx
SD_phya_ons_t04_2020-07-12.norm.xlsx
ld_phyA_ons_t04_2020-08-12.norm.xlsx

Exercise 6

1. Should dates be put first, and if not, why?
2. What is the difference between using leading 0 (zero) and not?
3. Is there a difference between using upper and lower case letters?
4. What is the difference between using two letters for *on* compared to three letters *ons*?
5. What are the effects if we, as in the above example, mix naming conventions?

- *phyA/phyB* - genotypes
- *sXX* - sample number
- *LD/SD* - light conditions (Long Day, Short Day)
- *on/off* - different growth media (on sucrose, off sucrose)
- *date format* - sample date
- *tXX* - sample timepoint
- *raw*, *norm* - raw or normalised data

2020-07-14_s12_phyB_on_SD_t04.raw.xlsx
2020-07-14_s1_phyA_on_LD_t05.raw.xlsx
2020-07-14_s2_phyB_on_SD_t11.raw.xlsx
2020-08-12_s03_phyA_on_LD_t03.raw.xlsx
2020-08-12_s12_phyB_on_LD_t01.raw.xlsx
2020-08-13_s01_phyB_on_SD_t02.raw.xlsx
2020-7-12_s2_phyB_on_SD_t01.raw.xlsx
AUG-13_phyB_on_LD_s1_t11.raw.xlsx
JUL-31_phyB_on_LD_s1_t03.raw.xlsx
LD_phyA_off_t04_2020-08-12.norm.xlsx
LD_phyA_on_t04_2020-07-14.norm.xlsx
LD_phyB_off_t04_2020-08-12.norm.xlsx
LD_phyB_on_t04_2020-07-14.norm.xlsx
SD_phyB_off_t04_2020-08-13.norm.xlsx
SD_phyB_on_t04_2020-07-12.norm.xlsx
SD_phya_off_t04_2020-08-13.norm.xlsx
SD_phya_ons_t04_2020-07-12.norm.xlsx
ld_phyA_ons_t04_2020-08-12.norm.xlsx

Exercise 6

1. Should dates be put first, and if not, why?
2. What is the difference between using leading 0 (zero) and not?
3. Is there a difference between using upper and lower case letters?
4. What is the difference between using two letters for *on* compared to three letters *ons*?
5. What are the effects if we, as in the above example, mix naming conventions?

2020-07-14_s12_phyB_on_SD_t04.raw.xlsx
2020-07-14_s1_phyA_on_LD_t05.raw.xlsx
2020-07-14_s2_phyB_on_SD_t11.raw.xlsx
2020-08-12_s03_phyA_on_LD_t03.raw.xlsx
2020-08-12_s12_phyB_on_LD_t01.raw.xlsx
2020-08-13_s01_phyB_on_SD_t02.raw.xlsx
2020-7-12_s2_phyB_on_SD_t01.raw.xlsx
AUG-13_phyB_on_LD_s1_t11.raw.xlsx
JUL-31_phyB_on_LD_s1_t03.raw.xlsx
LD_phyA_off_t04_2020-08-12.norm.xlsx
LD_phyA_on_t04_2020-07-14.norm.xlsx
LD_phyB_off_t04_2020-08-12.norm.xlsx
LD_phyB_on_t04_2020-07-14.norm.xlsx
SD_phyB_off_t04_2020-08-13.norm.xlsx
SD_phyB_on_t04_2020-07-12.norm.xlsx
SD_phya_off_t04_2020-08-13.norm.xlsx
SD_phya_ons_t04_2020-07-12.norm.xlsx
ld_phyA_ons_t04_2020-08-12.norm.xlsx

1. Using dates as leading information in file names makes finding data quickly harder as the more interesting information may be samples or timepoints (unless date is crucial to data).
2. Wihtout leading zeros, sorting will make 10 and 11 appear before 2.
3. Upper and lower cases may sort differently
4. Comparing files is easier if the file name lenghts are uniform.
5. Mixed naming conventions can make it difficult to locate particular files, and/or sort a large number of files.

File naming strategy

Two starting points for your file naming strategy are:

- A file name is a principal identifier of a file

Good file names contain useful clues to the content, status and version of a file, uniquely identify a file and help in classifying and sorting files. File names that reflect the file content also facilitate searching and discovering files. In collaborative research, it is essential to keep track of changes and edits to files via the file name.

- File naming strategy should be consistent in time and among different people

In both quantitative and qualitative research, file naming should be systematic and consistent across all files in the study. A group of cooperating researchers should follow the same file naming strategy and file names should be independent of the location of the file on a computer.

Exercise 7

Group discussion

What are examples of potential benefits of agreeing on a File Naming Convention for a project?

- Easier to process - Team members will not have to over think the file naming process
- Easier to facilitate access, retrieval and storage of files
- Easier to browse through files, saving time and effort
- Harder to lose!
- Having logical and known naming conventions in place can also help you with version control.
- Check for obsolete or duplicate records

Principles for naming files

1. Consider file name lengths – beware of OS limitations and full path names!
2. Make names human readable – name describes content of file
3. Make names machine readable – Avoid spaces, punctations, accented characters etc.
4. Explain file naming in associated info files (README.md)

File naming

Examples of a **poor** file name**:**

”Honeybee project, experiment 2 done in Helsinki, data file created on the second of December 2020”

File name - Runnew_again_2NDTRY.xls

Explanation - N/A

File naming

Examples of a **good** file name**:**

”Honeybee project, experiment 2 done in Helsinki, data file created on the second of December 2020”

File name - 20201202_HB_EXP2_HEL_DATA_V03.xls

Explanation - Time_ProjectAbbreviation_ExperimentNumber_ Location_TypeOfData_VersionNumber

File naming Do's

- For dates use the YYYY-MM-DD standard and place at the end of the file UNLESS you need to organize your files chronologically
- Include version number (if applicable), use leading zeroes (i.e.: v005 instead of v5). make sure the end-letter file format extension is present at the end of the name (e.g. .doc, .xls, .mov, .tif)
- Add a README.md (or PROJECT_STRUCTURE.md) file in your top directory which details your naming convention, directory structure and abbreviations

Increasing file findability

Keyword tagging

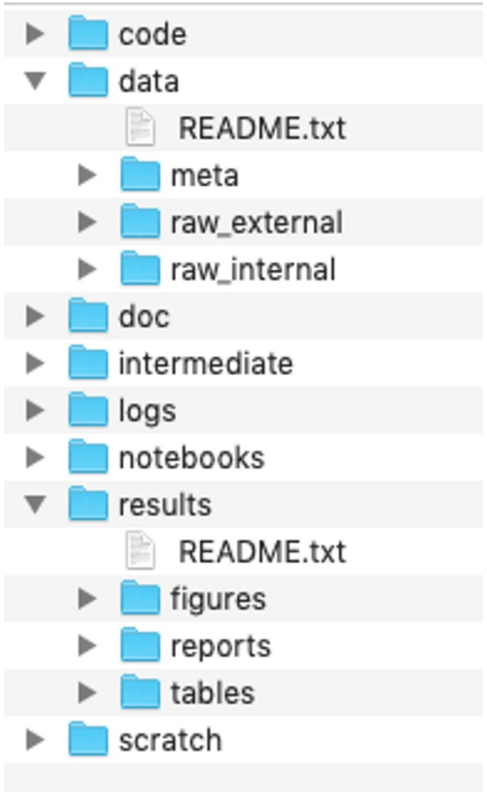
```
(Metadata.txt content)

20220115_MyFile_Project1_Location_Dataiteration1_V1.xml
First version of X data from Y, with additions of Z made by A and B on 20220110 includi
ng suggestions by C.
Keywords HumptyDumpty Genome_Assembly

20220115_MyFile_Project1_Location_Dataiteration2.xml
Contains X data from Y, with additions of Z made only by A on 20220111 not including su
ggestions by C.
Keywords Published
```

Associated metadata to increase findability of files over e.g. multiple projects

File naming Dont's

- Using spaces (use _ or - instead)
- Dots, commas and special characters (e.g. ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “)
- Using language specific characters (e.g óężé), unfortunately they still cause problems with most software or between operating systems (OS)
- Long names
- Repetition, e.g if directory name is Electron_Microscopy_Images, and file ELN_MI_IMG_20200101.img then ELN_MI_IMG is redundant
- Deep paths with long names (i.e. deeply nested folders with long names), as archiving or moving between OS may fail

Directory structure for a sample project



all code needed to go from input files to final results

raw and primary data, essentially all input files, **never** edit!

documentation for the study

output files from different analysis steps, *can be deleted*

logs from the different analysis steps

output from workflows and analyses

temporary files that can be safely *deleted or lost*

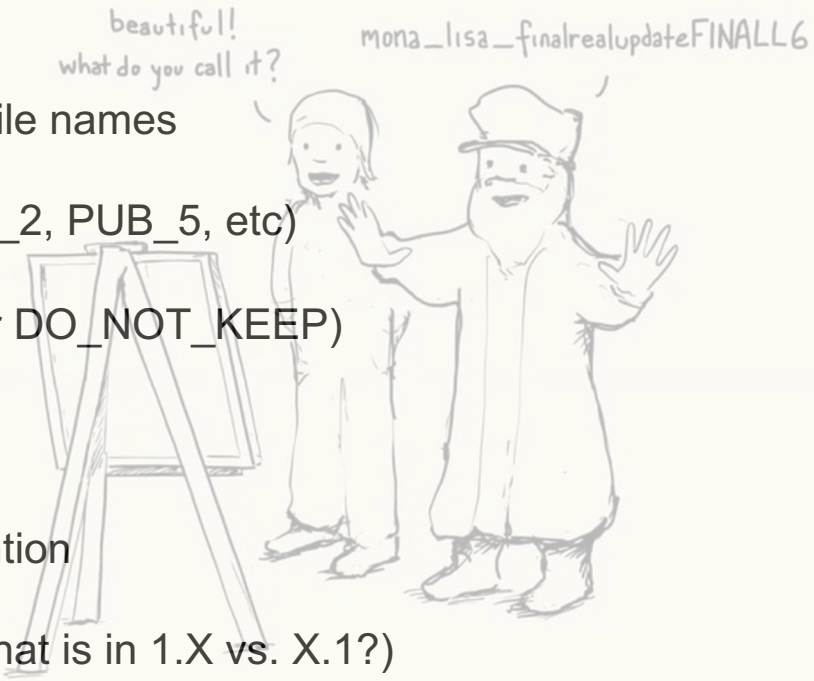
Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424.
http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000424

File versioning

The simple yet powerful Dont's and Do's of file versioning:

Dont's

- Add suffixes like FINAL, THIS_ONE, or PUB, to file names
- Add numbers to already bad suffixes (e.g. FINAL_2, PUB_5, etc)
- Add negative information (e.g. DELETE_THIS, or DO_NOT_KEEP)



Do's

- Explicitly include versioning in file naming convention
- Use version numbers, preferably consistently (What is in 1.X vs. X.1?)

DIY File naming convention

Want to create your own File Naming Convention? Consider…

1. What group of files will this naming convention cover?
2. What information (metadata) is important about these files and makes each file distinct?
3. Do you need to abbreviate any of the metadata or encode it?
4. What is the order for the metadata in the file name?
5. What characters will you use to separate each piece of metadata in the file name?
6. Will you need to track different versions of each file?
7. Write down your naming convention pattern
8. Document this convention in a README.md (or save this worksheet) and keep it with your files

Tabular data

Spreadsheet data is very common, and equally misunderstood

Tabular data is not a data *type*, but …

- a way to *organize data*

- designed for machine readability

Long term storage, exporting, archiving and FAIRification by concerting to .CSV or .TSV

Tabular data

Good practice for structuring tabular data is to…

- Adopt good metadata and column header formats early in the data collection phase (Pre-adapting to publication of data)
- Think about how you want your data both from a data entry and data analysis point of view
- Consider how to document your work
- Separate raw data from the data used in analysis

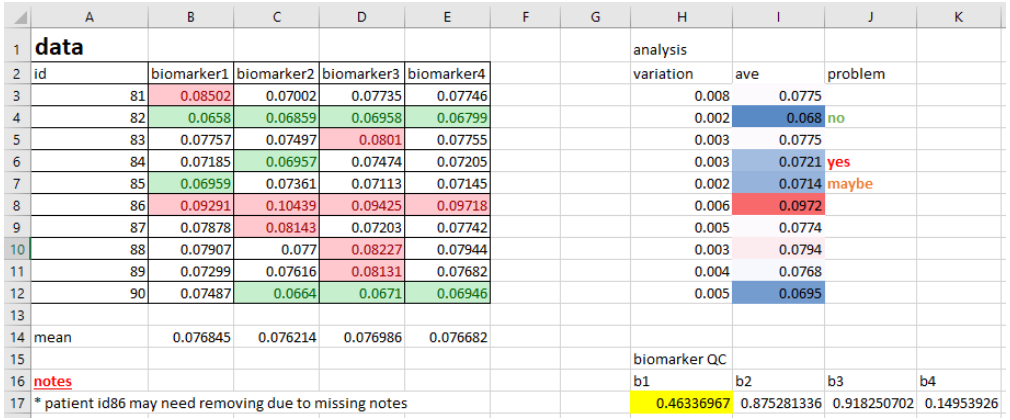
Tabular data

- Column = Variable
- Row = Observation
- Cell = Value

Open Access training					
Date	Length (hours)	Registered	Attended	Delivered by	Canceled
16/01/17	1	26	23	JM	N
05/02/17	1	38	26	JM	N
17/02/17	1	19	25	PG	N
07/03/17	1	27	17	JM	N
29/03/17	1	32	15	PG	N
02/04/17	1	41		PG	Y
24/04/17	2	44	44	JM	N
25/05/17	1	43	37	PG	N
16/06/17	1	15	15	JM	N

Tabular data

- ✓ Raw means raw!
- ✓ Tidy data tables
  - One cell–one value
  - One column–one variable
  - One row–one observation
- ✓ Beware of Excel “features”
  - Misguided ”auto-corrections” of dates, casing, numbers etc.
  - Misaligned formulas
  - Limited numerical precision
  - Limited number of rows/columns



	A	B	C	D	E	F	G	H	I	J	K
1	**data**							analysis			
2	id	biomarker1	biomarker2	biomarker3	biomarker4			variation	ave	problem	
3	81	0.08502	0.07002	0.07735	0.07746			0.008	0.0775		
4	82	0.0658	0.06859	0.06958	0.06799			0.002	0.068	no	
5	83	0.07757	0.07497	0.0801	0.07755			0.003	0.0775		
6	84	0.07185	0.06957	0.07474	0.07205			0.003	0.0721	yes	
7	85	0.06959	0.07361	0.07113	0.07145			0.002	0.0714	maybe	
8	86	0.09291	0.10439	0.09425	0.09718			0.006	0.0972		
9	87	0.07878	0.08143	0.07203	0.07742			0.005	0.0774		
10	88	0.07907	0.077	0.08227	0.07944			0.003	0.0794		
11	89	0.07299	0.07616	0.08131	0.07682			0.004	0.0768		
12	90	0.07487	0.0664	0.0671	0.06946			0.005	0.0695		
13											
14	mean	0.076845	0.076214	0.076986	0.076682						
15								biomarker QC			
16	notes							b1	b2	b3	b4
17	* patient id86 may need removing due to missing notes							0.46336967	0.875281336	0.918250702	0.14953926

Tabular data

Excel:	Is that a date?
Me:	57.39 is very much NOT a date
Excel:	Strong date vibes to me
Me:	H-how
Excel:	Fixed it
Me:	57/39/2020?
Excel:	You’re welcome
Me:	Please, please change it back to a number
Excel:	Ok! I think it was 57.389999999999999, right?

Tabular data

Zero vs. Missing data - How do you make explicit something that do not exist?

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

White et al, 2013, Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution

Tabular data

Do not:

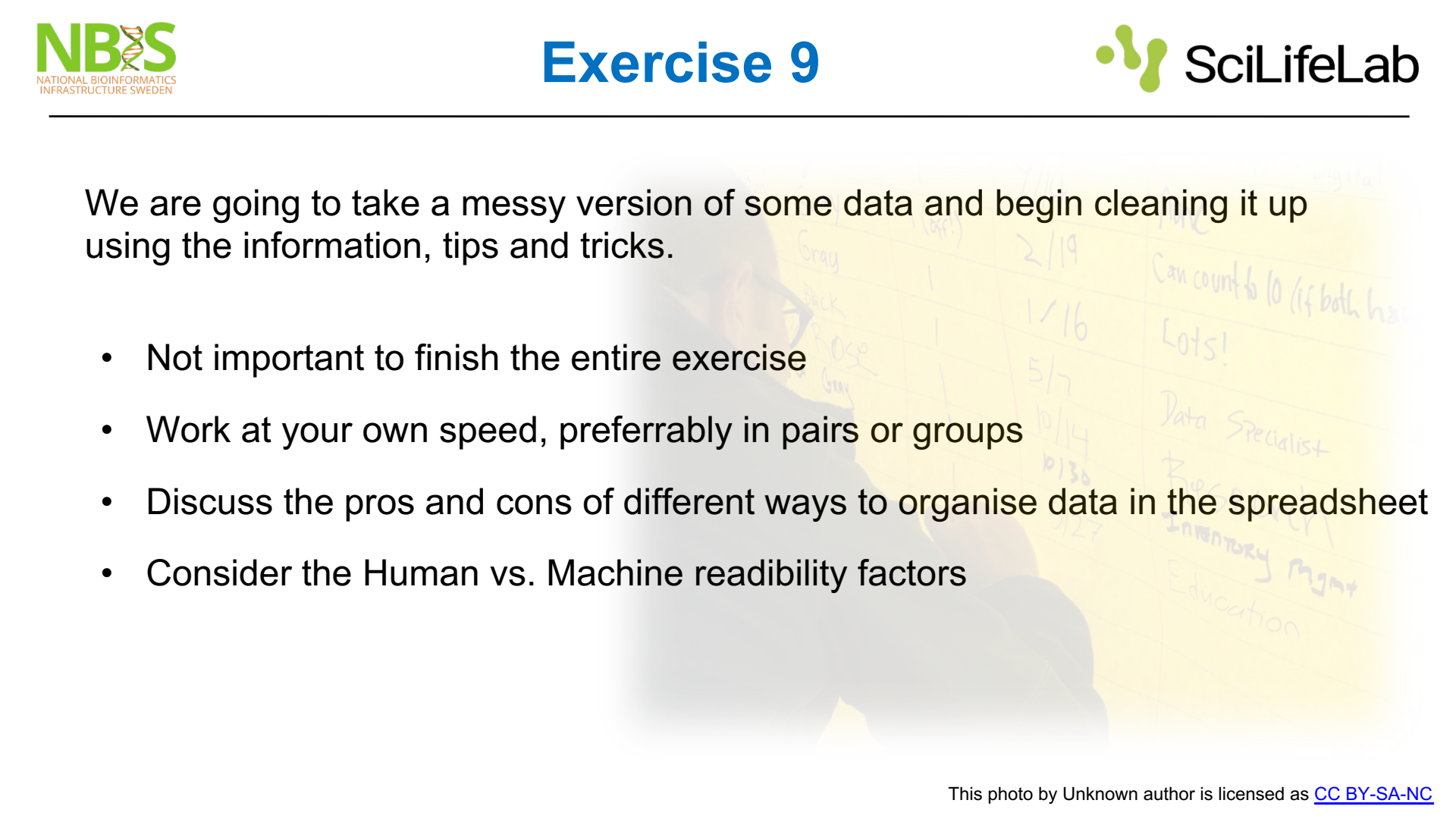
- Spatially distribute data
- Combine values in cells
- Split compatible data in tables
- Use colors
- Use cells for comments
- Mix metadata and data
- Use special characters

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2		RDM training					Open access						
3		Date	Length (hours)	PGR|PDRA|other	Delivered by		Date	Len	Attendees	Delivered by		
4		12 Jan	1.5	45|0|0	FG		8 Jan	1.5 hours	20	FG		
5		7 Feb	2	38|0|0	GH		13 Jan	1 hour	21	JM		
6		4 Mar	2	43|3|0	GH		22 Jan	1 hour	35	JM		
7		6 Mar	1	21|7|0	GH		2 Feb	1.5 hours	36	JM		cancelled
8		17 Mar	1.5	34|1|0	FG		3 Feb	1.5 hours	22	JM		
9		21 Mar	1	25|2|0	DQ		3 Feb	1 hours	30	JM		
10		23 Mar	2	32|10|0	FG		20 Feb	1.5 hours	36	FG		
11		19 Apr	1	34|0|0	GH		28 Feb	1.5 hours	28	JM		
12		30 Apr	1.5	37|0|0	FG		19 Mar	1.5 hours	33	FG		
13		4 Jun	1	45|0|0	GH		19 Mar	1 hour	39	JM		
14		12 Jun	2	36|0|0	DQ		4 Apr	1.5 hours	21	JM		
15		22 Jun	1.5	38|0|0	DQ		5 May	1.5 hours	25	JM		
16		25 Jun	1	35|4|0	GH		18 May	1 hour	22	JM		
17		30 Jun	1.5	44|3|0	FG		19 May	1.5 hours	20	FG		
18		1 Jul	1.5	40|0|4	FG		21 May	1.5 hours	21	JM		
19		6 Jul	1.5	21|0|0	GH		14 Jun	1.5 hours	37	JM		
20		7 Jul	1	37|4|1	DQ		18 Jun	1.5 hours	25	JM		
21		9 Jul	1	29|7|0	GH		4 Jul	1.5 hours	39	JM		
22		30 Jul	2	22|3|0	FG		6 Jul	1.5 hours	39	JM		
23		29 Aug	1.5	22|4|0	GH		10 Jul	1.5 hours	34	JM		
24		10 Sep	1	38|0|0	FG		13 Jul	1.5 hours	23	FG		
25		21 Sep	1	31|0|0	GH		17 Jul	1.5 hours	30	JM		
26		1 Oct	2	26|9|5	DQ		3 Aug	1.5 hours	28	JM		
27		25 Oct	1.5	20|4|0	DQ		20 Aug	1.5 hours	32	JM		
28		4 Nov	1.5	38|5|5	FG		26 Aug	1.5 hours	25	JM		
29		5 Nov	2	40|0|0	GH		28 Aug	1.5 hours	33	FG		
30		8 Nov	2	22|7|0	FG		1 Oct	1.5 hours	38	JM		
31		1 Dec	2	41|6|0	DQ		21 Oct	1.5 hours	34	JM		
32		19 Dec	2	39|9|1	GH		9 Nov	1.5 hours	32	JM		
33							15 Nov	1.5 hours	35	JM		
34							15 Nov	1.5 hours	27	JM		
35							2 Dec	1.5 hours	35	FG		
36							7 Dec	1.5 hours	23	JM		
37							11 Dec	1.5 hours	38	FG		
38							19 Dec	1.5 hours	20	FG		
39													

Exercise 9

We are going to take a messy version of some data and begin cleaning it up using the information, tips and tricks.

- Not important to finish the entire exercise
- Work at your own speed, preferrably in pairs or groups
- Discuss the pros and cons of different ways to organise data in the spreadsheet
- Consider the Human vs. Machine readibility factors



Example file structure

We are back in the Famous lab!

- Considering the very limited metadata we have access to, and the inherited files, what can we do in order to increase the level and quality of data organization?
    - Download the zip-file containing the inherited data structure
        - Consider the following:

            - File names
            - Folder structure
            - Documentation

            - Work in pairs or in smaller groups.
            - Focus on the discussion more than finishing the exercise.
            - Consider your own data and files from a third-person-view

Exercise 10