

File and data management



Sergio Martínez Cuesta



Aims

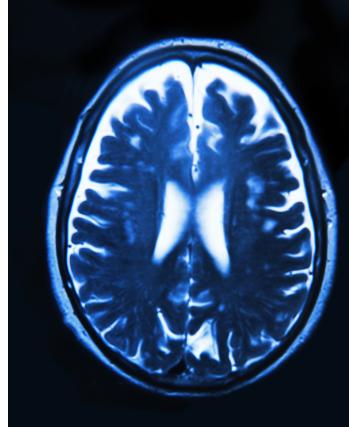
- What are research files and data?
- Why managing your data can be useful for you and others?
- Challenges
- Gaining confidence to organise your data well
- Existing resources and tools

What is research data?

Pieces of information that are descriptive of the research object ... or are the object itself

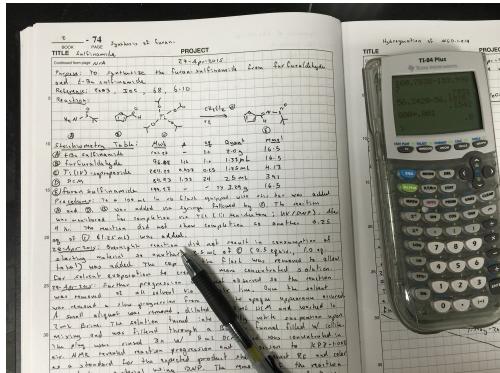
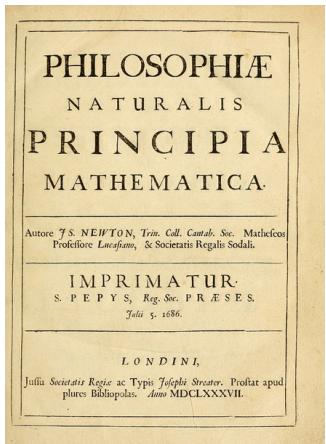
- ✓ Raw/processed data produced at a research facility
- ✓ Published dataset

Necessary to validate research findings



Experimental data

Images



Primary sources

Notebooks and diaries



Questionaries and surveys



Databases



Email



Audio

... and many others ...

Data formats



... and more ...

How to manage your data?

The everlasting external disks



But are they really permanent? What if ...?

How do you manage your research data?

Talk to the person next to you for 5 min and exchange information:

- What kind of research do you do?
- Do you do any data **backup**?
- How **often**?
- How do you **share** files/data with collaborators?



Why the need to manage it ?

It can be a **painful** process ... but definitely **worth** long-term

*“My field is very competitive and I can’t just risk **wasting time** with all of this, so I’d rather do **real** research than tidy up my data”*

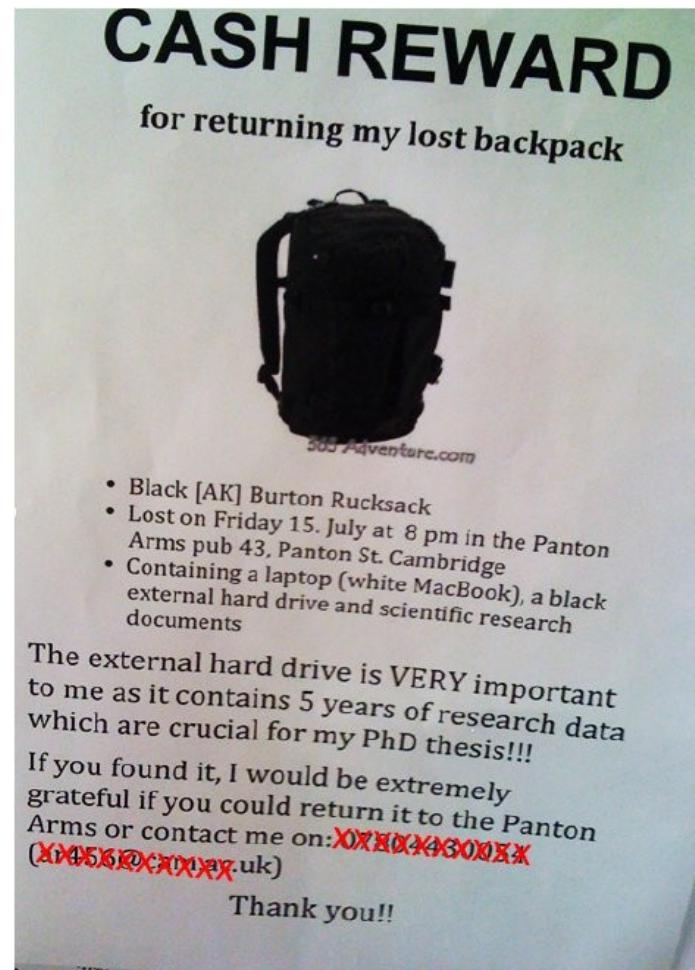
*“My data **are spread** over so many hard drives and directories that it would just be too much work to collect them all in one place”*

*“I can always sort out all my data **after submission** anyway”*

Markowitz F., 2015

To avoid data *disasters* ...

What would you do if you'd lose your data tomorrow?



What would you do if you'd lose your data tomorrow?

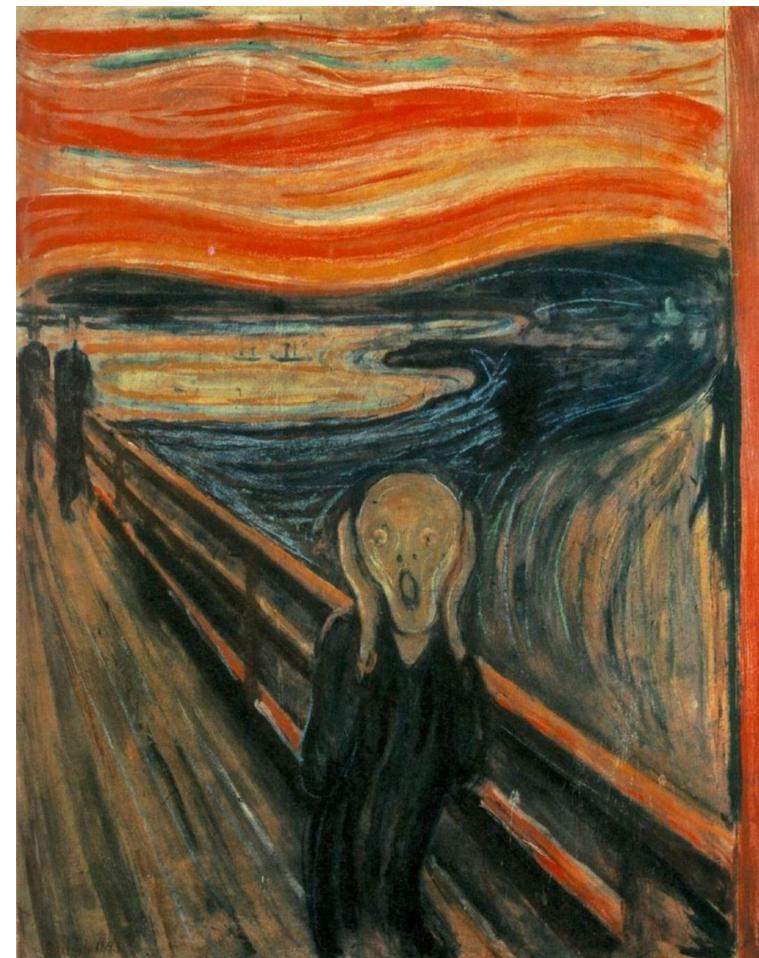


Cancer Research UK – University of Manchester – 27 April 2017

<http://www.itv.com/news/granada/2017-04-26/fire-breaks-out-at-manchester-christie-cancer-research-building/>

What would you do if you'd lose your data tomorrow?

- Your laptop got stolen
- Your office/house burnt
- Your USB stick is lost
- Your portable hard disk is damaged
- Data copied to Dropbox disappeared



https://en.wikipedia.org/wiki/The_Scream

Data backup and file sharing

Never work directly on the raw data

Leave it intact

Make a copy, and work on the copy

Data backup

At least 2 backups at 2 different locations

External disks



Online backup



Servers

Department
College
IT



Cheap
£10-15 / TB (1024GB)



Failure rate
1.5%/year

Accessibility
Free (limit)

Personal data
Hacking

Managed by
experts

Moving between
institutions

Data backup

Manual



Copying files to relevant folders



Automated

Install software
e.g. Time machine
(Mac users)



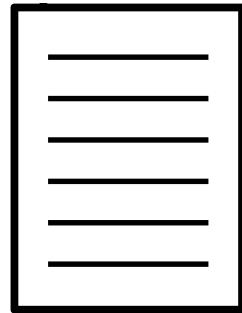
Automatically upload files to the cloud when any changes are saved

If manual ... how often?



How much would you be willing to lose?

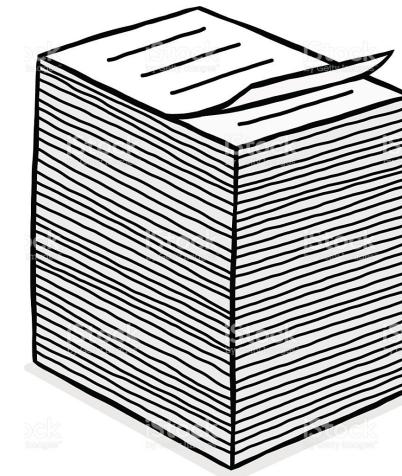
1 day



1 week



1 month-year



*Software allows you to set up **backup time automatically***

Data backup and file sharing



Space/price	2 GB (free) Unlimited (£55/year)	15 GB (free) 1 TB (~£80/year)	1 TB (free)
File history and recovery	Yes, unlimited	Yes	Last 90 days
File size limit	None	5 GB	15 GB
Support	UIS	Unsupported	UIS
OS	Windows, Mac, Linux, Android, iOS	Windows, Mac, Android, iOS	Windows, Mac, Android, iOS
Accessibility	Sync anywhere on any devices	Live editing	Integration with Microsoft Office

More ... file sharing



Email



Website



FTP

Data organisation and file naming

How do you **organise** your research data?

Talk to the person next to you for 5 min and exchange information:

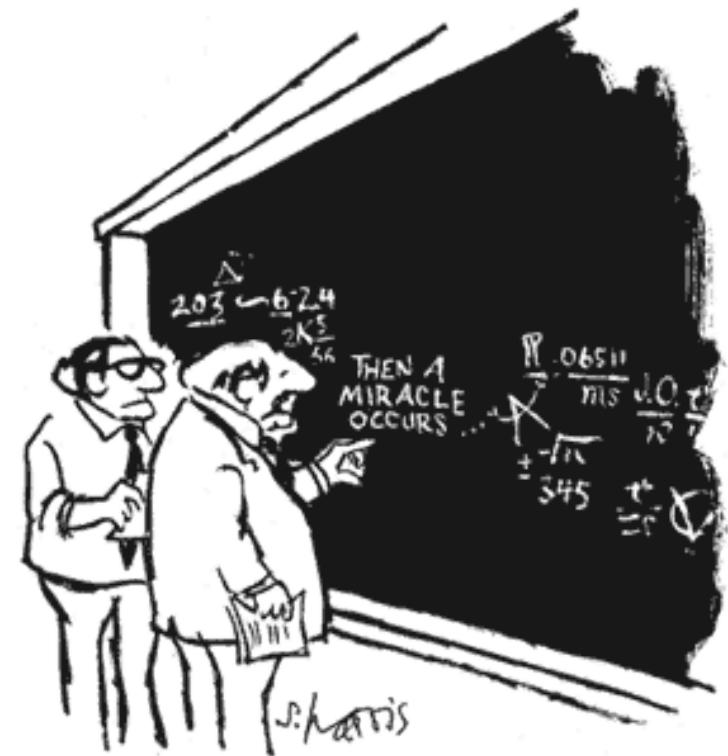
- How do you **organise** your files?
- How do you **name** your files?
- How do you **share** your files?



To allow **continuity** of your work ...

*"I obtained the data 6 months ago. I am too busy. Of course I **can't remember** all the details of all my projects after such a long time"*

*"My supervisor said I should continue the project of a previous student, but that student is long gone and **hasn't saved any data or documentation**"*



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Copyright Sidney Harris

Markowitz F., 2015

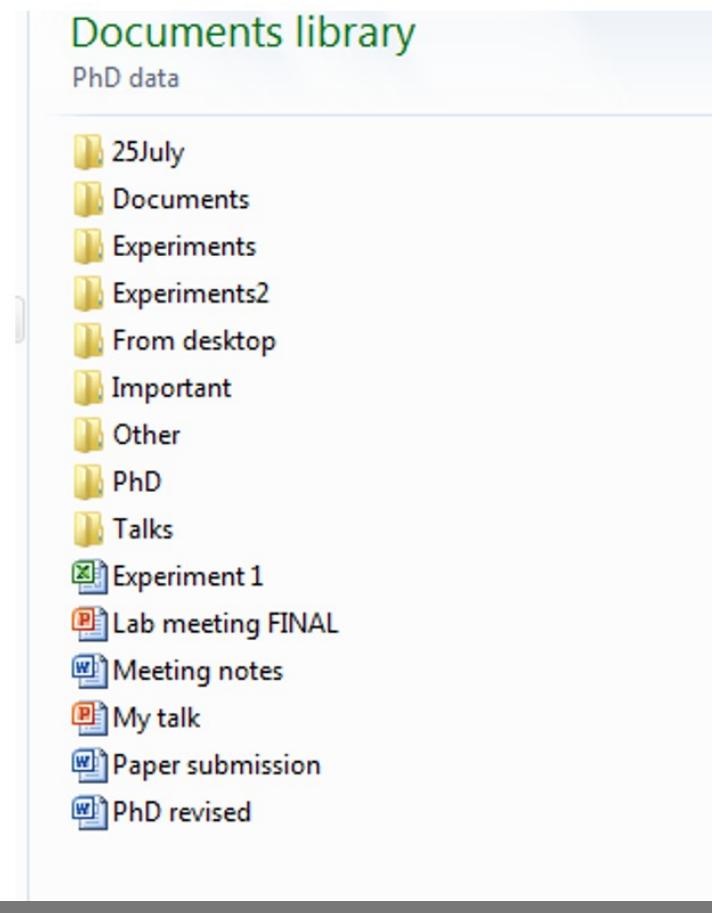
Data organisation



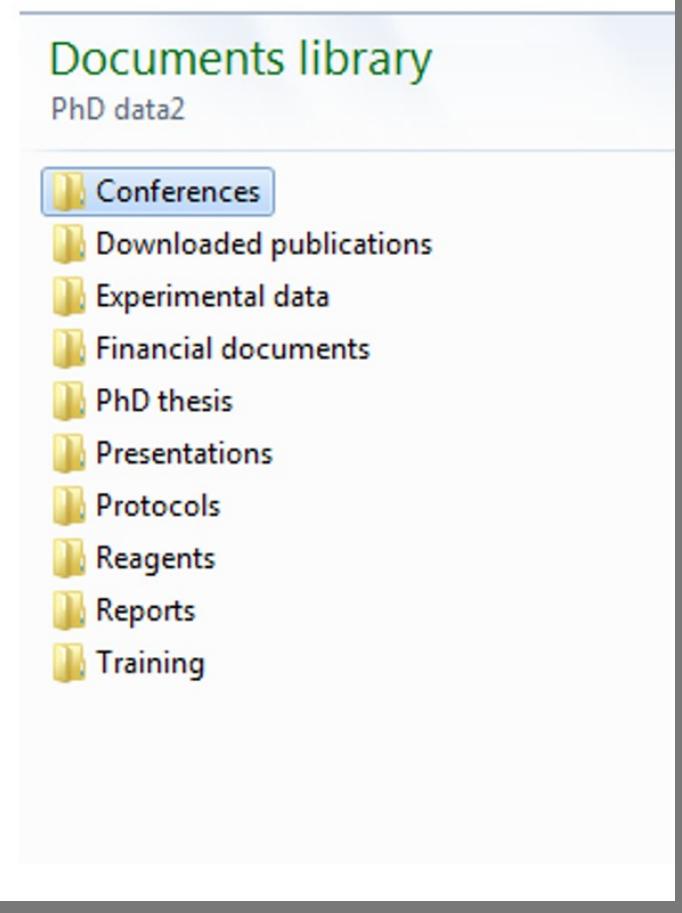
<https://www.wired.com/2013/04/desktop-cluttered-help/>

Data organisation

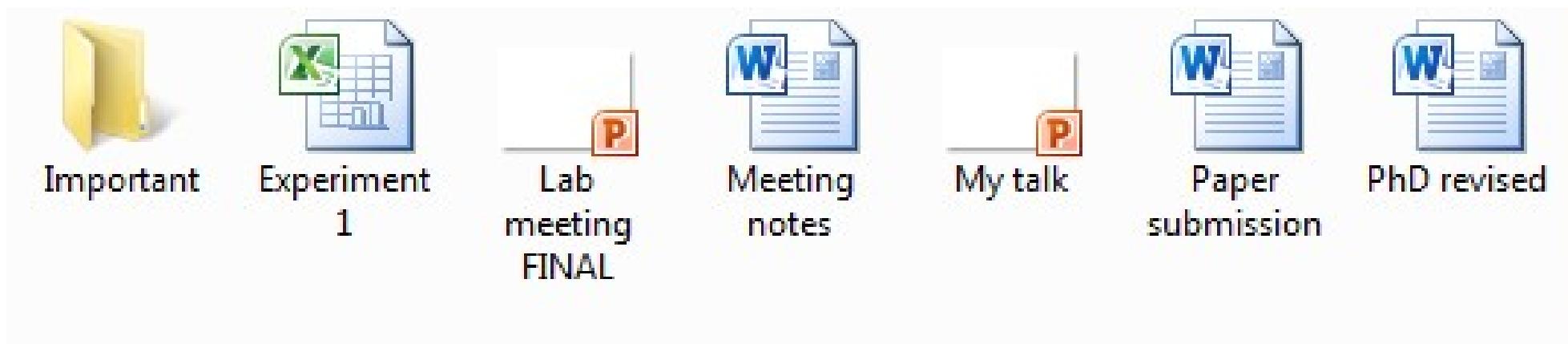
Example A¶



Example B¶



Data organisation



- Consistent and descriptive
- Meaningful to you and your colleagues
- Allow you to find files easily

File naming

Create a naming **convention** that works for you and your collaborators that allows to **distinguish** files from one another

fk468_PCF-b2-oxhyd-1.fastq.gz

.../20170601_RDM_Wolfson/...

Collaborator: fk (Fumiko Kawasaki)

Date: 20170601

Experiment number: 468

Topic: RDM
(Research Data Management)

Life cycle stage: PCF

Location: Wolfson college

Batch: b2

Treatment: oxhyd

Replicate: 1

Tips

Avoid **special characters** ~ ! @ # \$ % ^ & * () ` ; < > ? , [] { } ' " |

A good format for **dates** is **YYYY-MM-DD** or **YYMMDD**

All of your files will always stay in chronological order

Use **leading zeros** for clarity and to make sure files sort in sequential order

E.g. "001, 002 ... 010, 011 ..." instead of "1, 2, ...10, 11 ..."

Tips

Do not use spaces. Some softwares do not recognize file names with spaces.

e.g. data table.xls

Other options include:

Underscores, e.g. data_table.xls

Dashes, e.g. data-table.xls

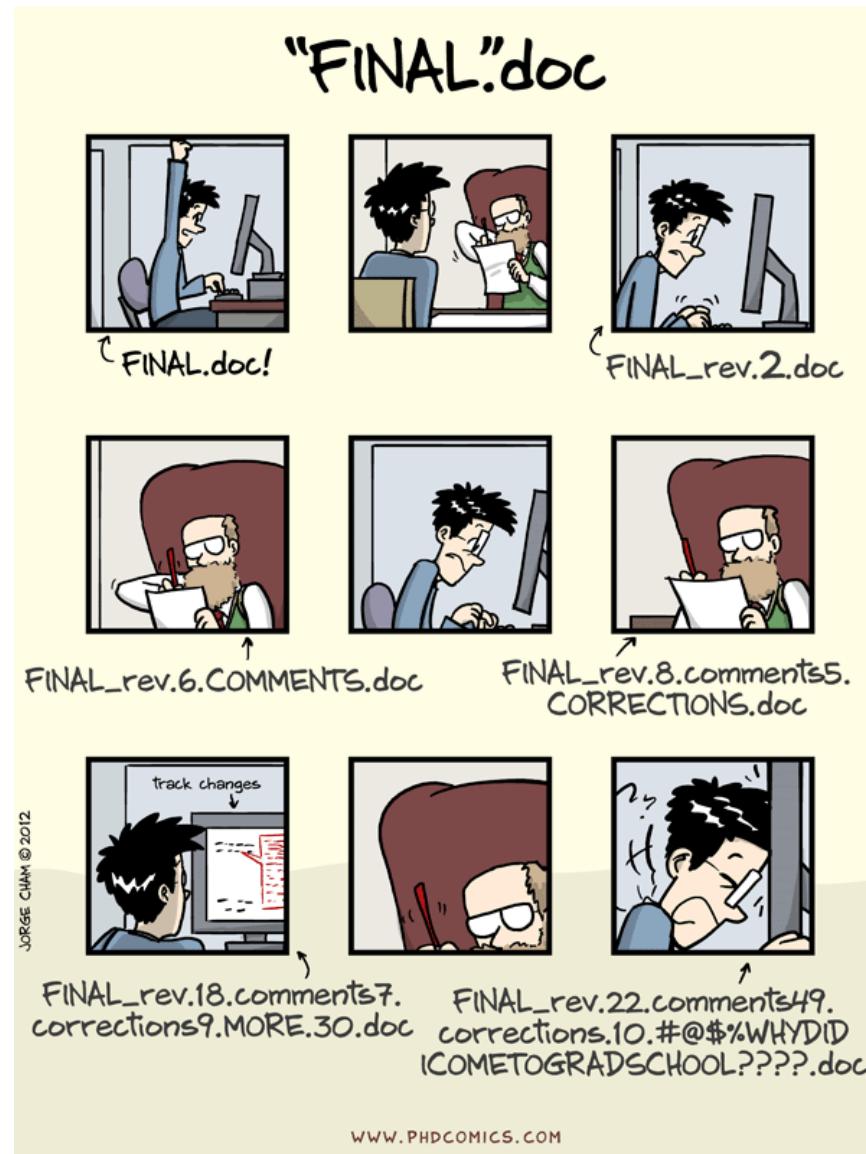
No separation, e.g. datatable.xls

Camel case, e.g. DataTable.xls

File naming

- ▶  20160706_Presentations
- ▶  20160726_RDMworkshopForLibrarians_MillLane
- ▶  20160802_Altmetrics
- ▶  20160804_Bibliometrics
- ▶  20160810_RDMworkshopForLibrarians_MilsteinRoom
 -  GDL_DMP_V8_2016511.docx
 -  GDL_DMPForLibrarians_V9_20160726docx.docx
 -  GDL_ExampleDMP_V9_20160511.docx
 -  LST_RDMforLibsAttendees_V1_20160808.pdf
 -  MEM_RDMforLibs2Feedback_V1_20160811.docx
 -  PRE_DataLossScenarios_V5_20160726.docx
 -  PRE_RDMforLibsSigns_V1_20160808.pptx
 -  PRE_RDMWorkshopForLibrarians_V3_20160726.pptx
 -  PRE_RDMWorkshopForLibrarians_V4_20160810.pdf
 -  PRE_RDMWorkshopForLibrarians_V4_20160810.pptx
- ▶  20160914_RDMWorkshopforGLS
- ▶  20161004_PhDTraining

File version control

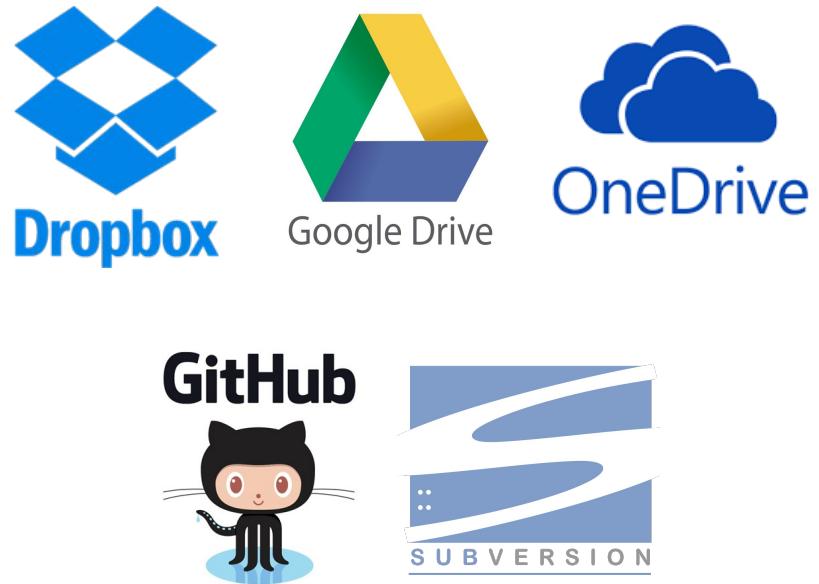


File version control

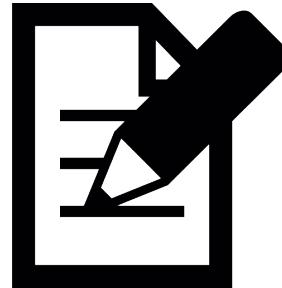
Manual

- Save new versions when you make significant changes
- Include a version number in the file name,
e.g. v01, v02 or v02.1
- Challenging if keeping lots of versions or working with many collaborators

Software options



Easier to *write papers* ...



Build your *reputation* ...

The FAIR principles:



Findability
Accesibility
Interoperability
Reusability

When do I need to worry about managing my data?

 ***Always!*** 

Before you start the project

Before collecting the data

While you do the analysis

When writing/co-authoring your paper/thesis

When reviewing the work of others

Overview

Try to keep your projects ***organised***

Name files and directories ***consistently*** using
some ***informative*** way

Store your data at a ***single backed-up***
location

Questions?

sm848@cam.ac.uk



Materials

Research Data Team
<http://www.data.cam.ac.uk/>

A friendly introduction to GitHub
<https://kirstiejane.github.io/friendly-github-intro/>

Five selfish reasons to work reproducibly
<https://dx.doi.org/10.1186/s13059-015-0850-7>