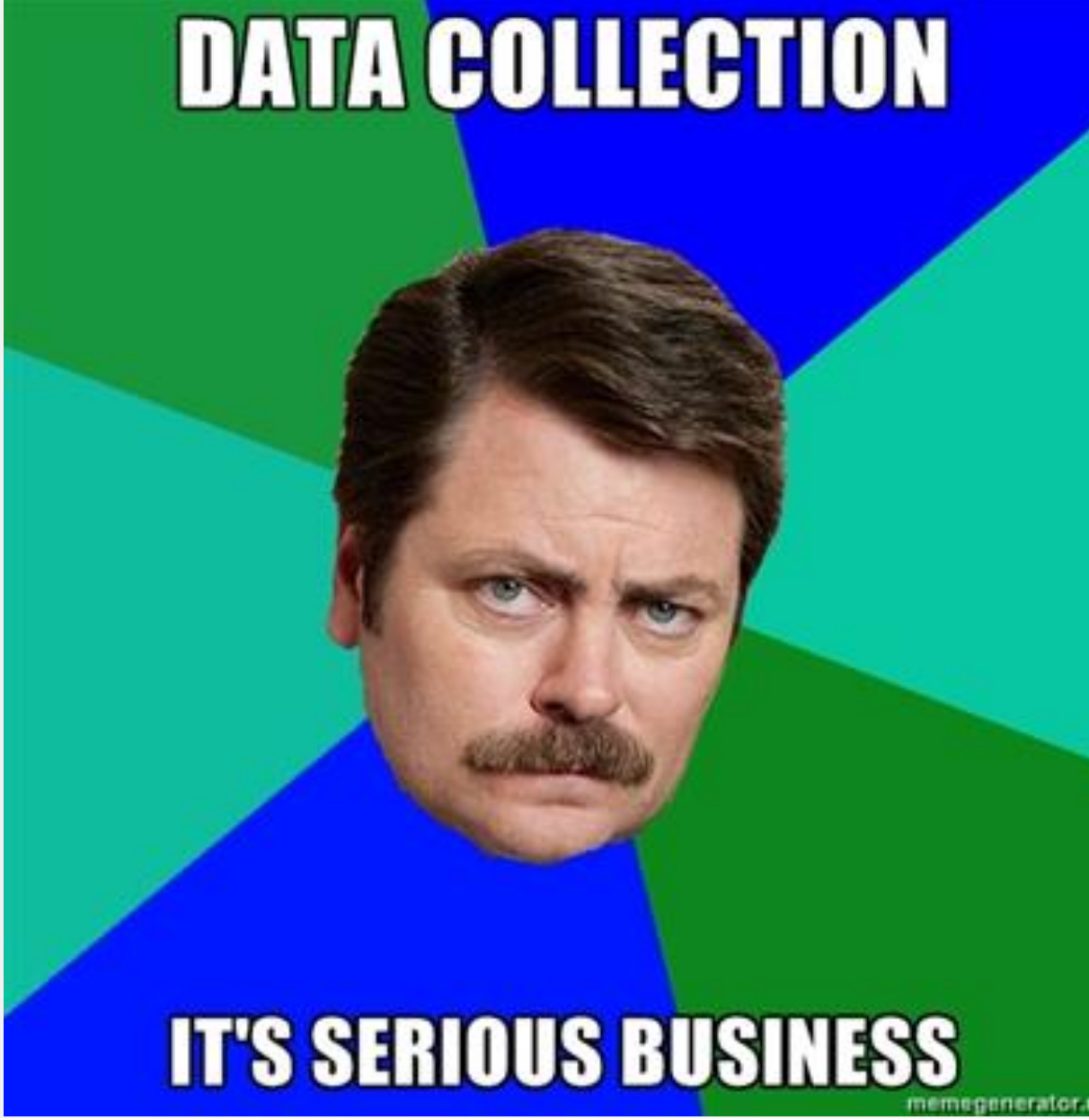
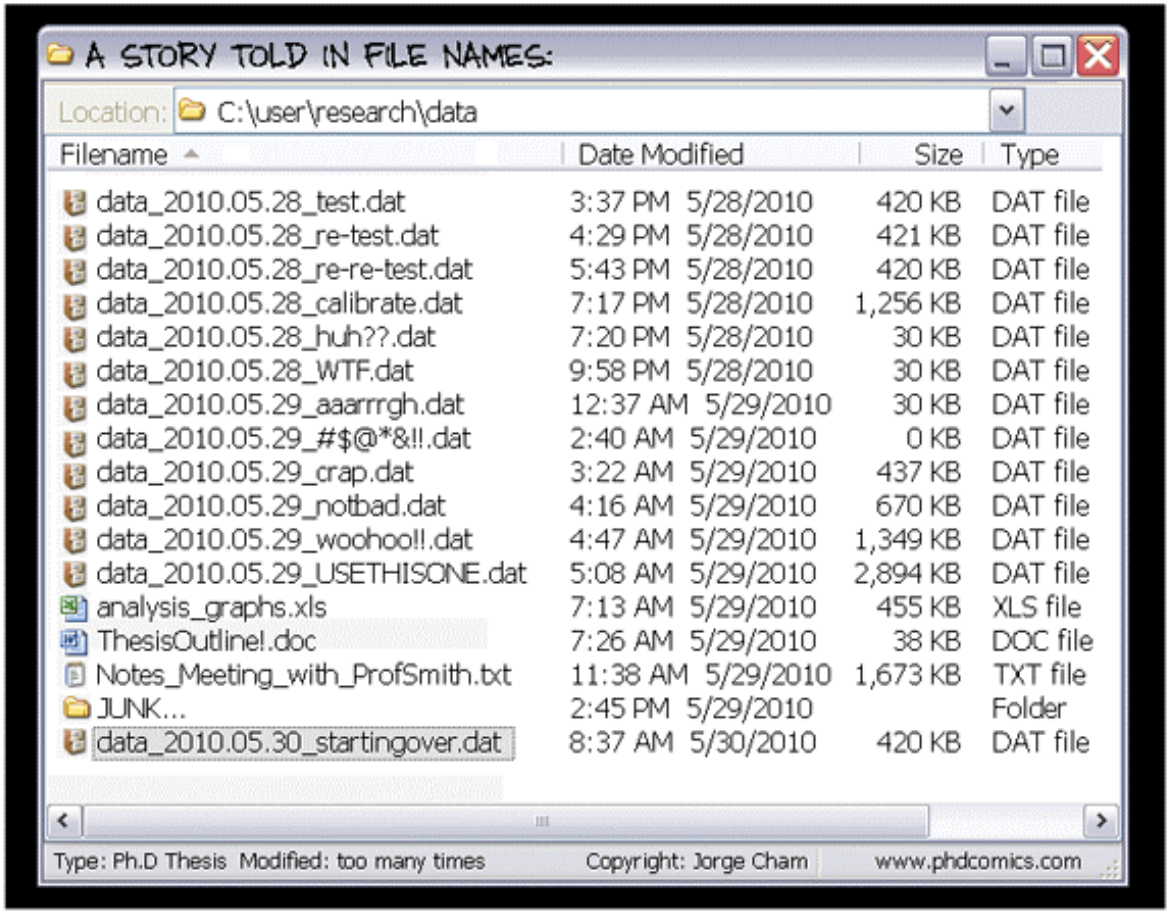
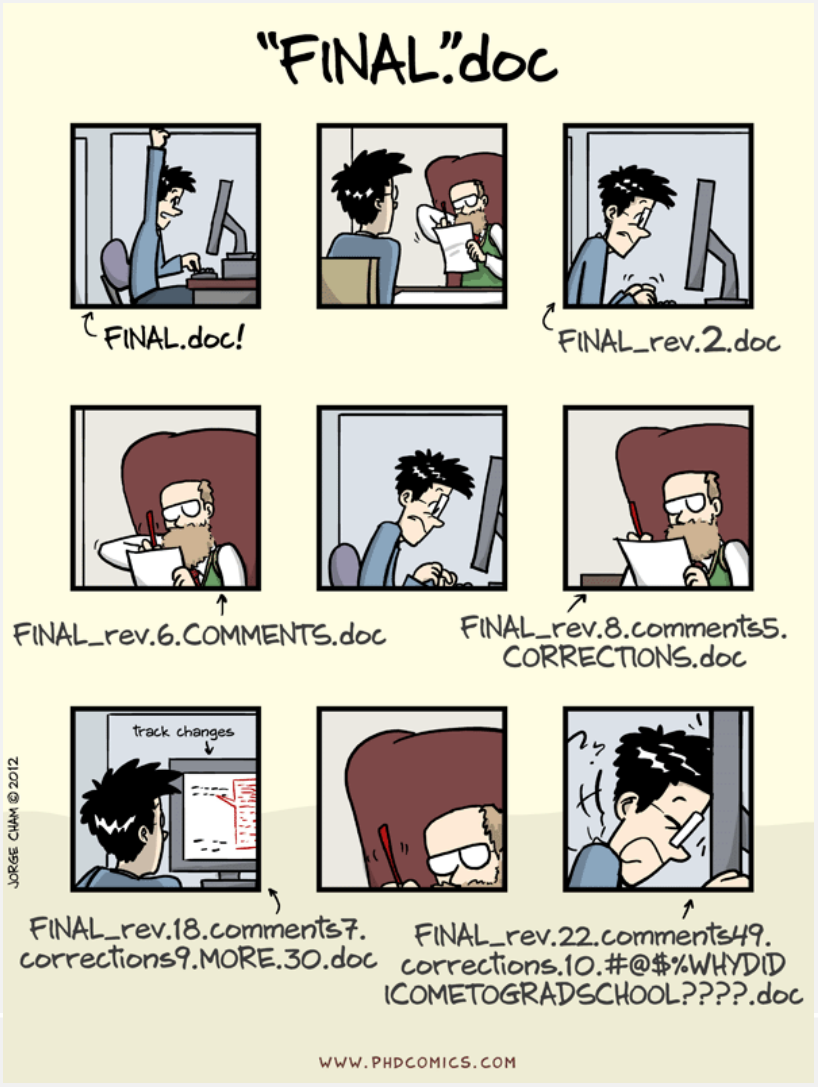
Data packaging:
A practical approach


DATA COLLECTION
IT'S SERIOUS BUSINESS
memegenerator


A STORY TOLD IN FILE NAMES:
Location: C:\user\research\data
Filename
Date Modified
Size
Type
data_2010.05.28_test.dat 3:37 PM 5/28/2010 420 KB DAT file
data_2010.05.28_re-test.dat 4:29 PM 5/28/2010 421 KB DAT file
data_2010.05.28_re-re-test.dat 5:43 PM 5/28/2010 420 KB DAT file
data_2010.05.28_calibrate.dat 7:17 PM 5/28/2010 1,256 KB DAT file
data_2010.05.28_huh??.dat 7:20 PM 5/28/2010 30 KB DAT file
data_2010.05.28_WTF.dat 9:58 PM 5/28/2010 30 KB DAT file
data_2010.05.29_aaarrrgh.dat 12:37 AM 5/29/2010 30 KB DAT file
data_2010.05.29_#$@*&!!.dat 2:40 AM 5/29/2010 0 KB DAT file
data_2010.05.29_crap.dat 3:22 AM 5/29/2010 437 KB DAT file
data_2010.05.29_notbad.dat 4:16 AM 5/29/2010 670 KB DAT file
data_2010.05.29_woohoo!!.dat 4:47 AM 5/29/2010 1,349 KB DAT file
data_2010.05.29_USETHISONE.dat 5:08 AM 5/29/2010 2,894 KB DAT file
analysis_graphs.xls 7:13 AM 5/29/2010 455 KB XLS file
ThesisOutline!.doc 7:26 AM 5/29/2010 38 KB DOC file
Notes_Meeting_with_ProfSmith.txt 11:38 AM 5/29/2010 1,673 KB TXT file
JUNK... 2:45 PM 5/29/2010 Folder
data_2010.05.30_startingover.dat 8:37 AM 5/30/2010 420 KB DAT file
Type: Ph.D Thesis Modified: too many times
Copyright: Jorge Cham
www.phdcomics.com


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
JORGE CHAM © 2012
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
WWW.PHDCOMICS.COM

Responsible research conduct: data management/packaging

‘Inadequate record keeping or data management related to research projects’

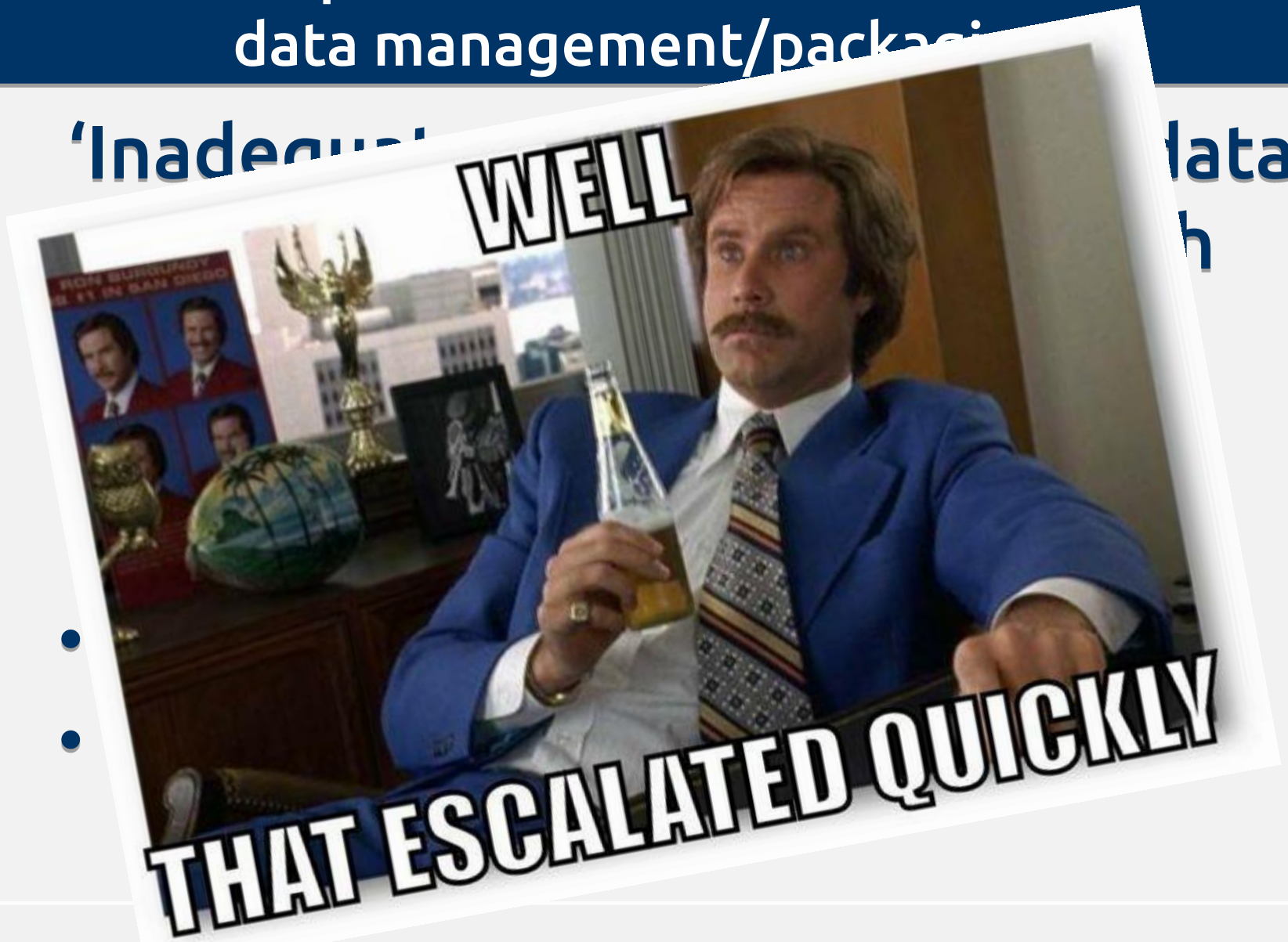
Responsible research conduct: data management/packaging

‘Inadequate record keeping or data management related to research projects’

- 27,5% (Martinson et al., 2005)
- 48%! (Godecharle et al., 2017)

Responsible research conduct: data management/packa

‘Inadequ … data … h



-
-

Say that you wake up one morning with full amnesia related to your current project, but you are still a very capable scientist. Are you able to understand what you did the day before?

Data package: Metadata and materials

- 1. Metadata and data collection
  - Author roles
  - Who collected data/where/when/how
  - ERB protocol number

- 2. Material
  - All digital material that can be used to replicate project.
    - Surveys/Stimuli/interview protocols/experiment leader protocol/computer scripts/etc. etc.

Data package: Raw data and analyses

- 3. Raw database
  - Raw, time-marked datafile
  - “Any information that could lead to the identification of participants (direct and indirect identifiers) must be removed from the data files.”
  - In case you are not allowed to store these data in a repository: clearly explain and document the reasons for absence
- 4. Data processing and analyses
  - Contain all information that allows others to replicate your study
    - Coding schemes, syntaxes, computer scripts, statistical logbooks of raw data processing
    - A sufficiently documented processed database

How does one achieve all these easily?

How does one achieve all these easily?

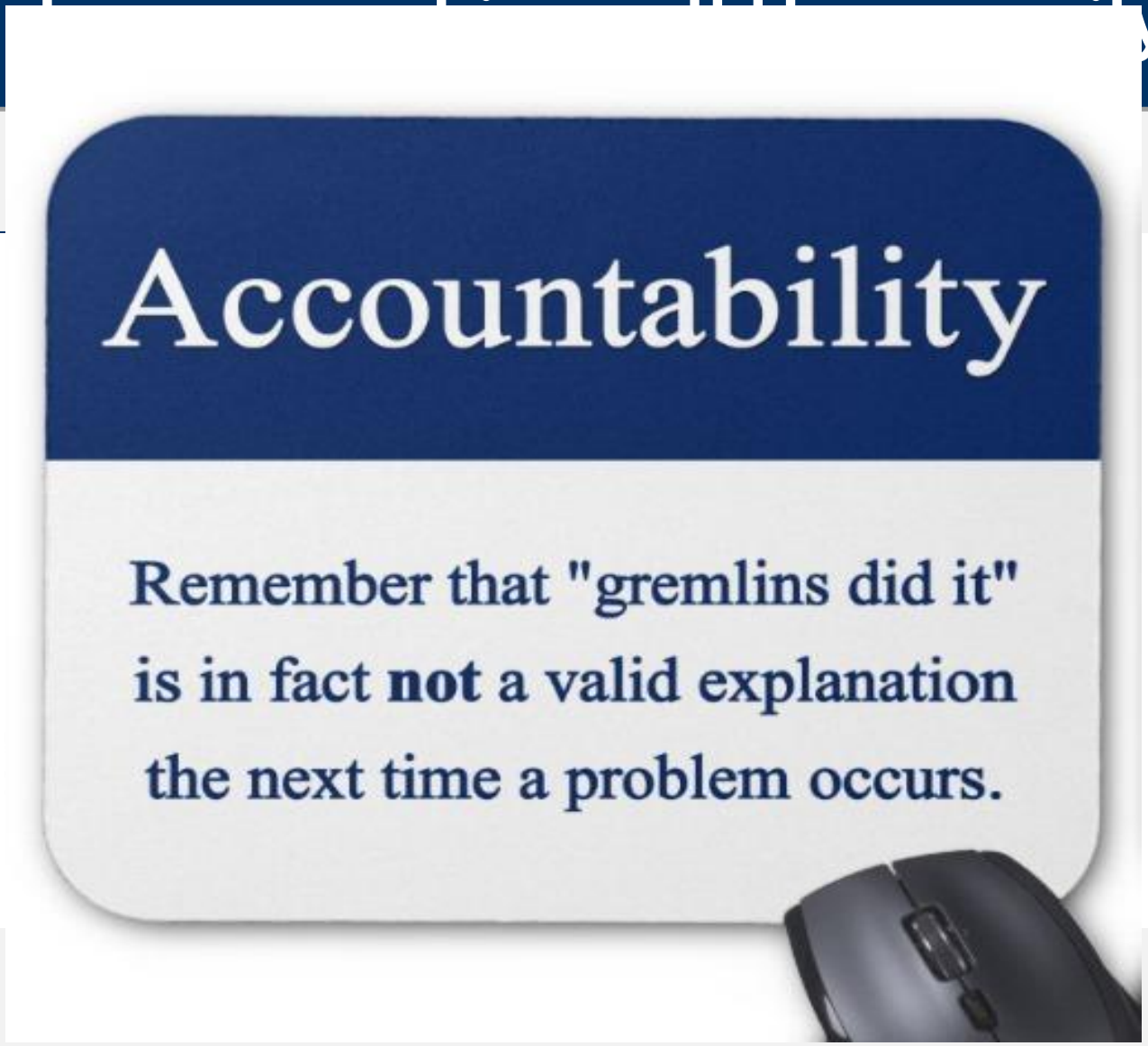
Transparency & Accountability

How does one achieve all these easily?

High transparency

No transparency

How


Accountability
Remember that "gremlins did it"
is in fact **not** a valid explanation
the next time a problem occurs.

How does one achieve T&A?

Today, we'll focus on two things:

- **Log changes**
- **Reproducibility**

How does one achieve T&A?

Version control

=

Track changes for files

Recent Activity

Marino van Zelst added file S5. Meta-regression operationalization.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S1. Papers by journal.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S3. Results of HOMA.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst removed file S1. Appendix Outlier Analyses.docx from OSF Storage in supplements
2016-05-10 06:02 PM

Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-10 06:02 PM

Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-09 08:26 PM

< 1 ... 4 5 6 ... 75 >

Recent Activity

Marino van Zelst added file S5. Meta-regression operationalization.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S1. Papers by journal.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst added file S3. Results of HOMA.pdf to OSF Storage in supplements
2016-05-10 06:07 PM

Marino van Zelst removed file S1. Appendix Outlier Analyses.docx from OSF Storage in supplements
2016-05-10 06:02 PM

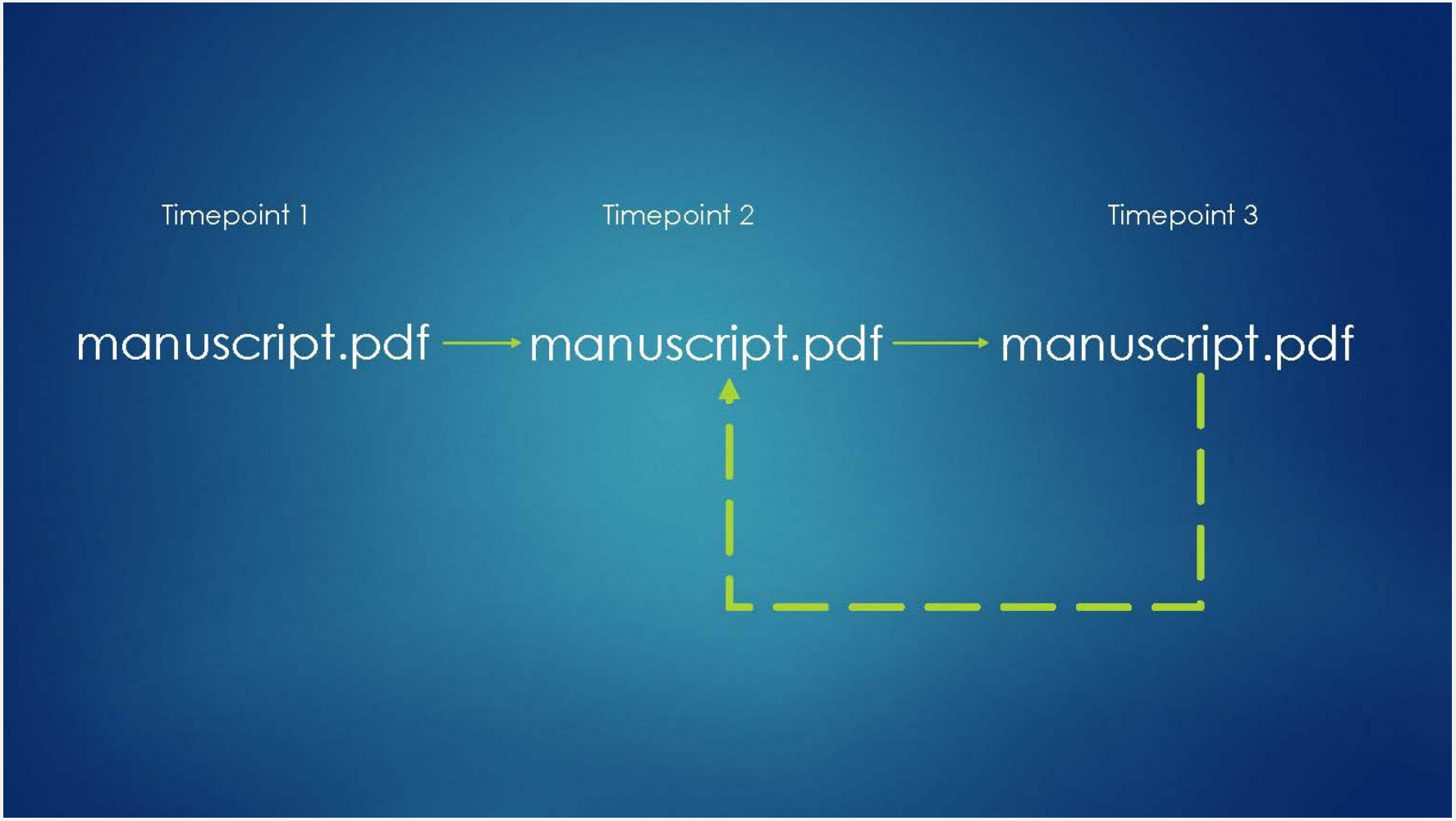
Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-10 06:02 PM

Marino van Zelst updated file 20150701thesis.docx in OSF Storage in Convenience-driven interorganizational tie formation: A meta-analysis.
2016-05-09 08:26 PM

< 1 ... 4 5 6 ... 75 >

Revisions

Version ID	Date	User	Download
13	2016-07-15 02:54 PM	Marino van Zelst	0
12	2016-05-27 09:13 PM	Marino van Zelst	2
11	2016-05-25 08:36 PM	Marino van Zelst	1
10	2016-05-10 06:09 PM	Marino van Zelst	3
9	2016-05-10 06:02 PM	Marino van Zelst	0
8	2016-05-09 08:26 PM	Marino van Zelst	0
7	2016-05-08 12:30 PM	Marino van Zelst	0
6	2016-05-04 09:18 PM	Marino van Zelst	0
5	2016-05-03 09:21 PM	Marino van Zelst	0
4	2016-05-03 11:29 AM	Marino van Zelst	0
3	2016-04-29 08:07 PM	Marino van Zelst	0
2	2016-04-25 07:38 PM	Marino van Zelst	0
1	2016-04-25 04:50 PM	Marino van Zelst	0



TILBURG UNIVERSITY

Adapted from Hartgerink (2016):
http://bit.ly/chrisPM101

How does one achieve transparency?

Open Science Framework

- Online environment
- Collaborative
- Preregistration
- Can make files public but not mandatory
- Watch out: US servers (privacy!)
- More advanced? Github/Bitbucket

How does one achieve transparency?

Ask ourselves continuously:

Can my future-self or anyone else reproduce this result within a reasonable amount of time?

Reproducibility

Data package requires you to:

- Be able to reproduce all your results in reasonable amounts of time

Reproducibility

Data package requires you to:

- **Be able to reproduce all your results in reasonable amounts of time**

- **Use syntax for everything!**
- **Log WHY you use specific syntax**

Reproducibility

```
1934 dat <- read.csv("masterdata.csv",header=TRUE)
1935 # Splined + separates #
1936 dat <- dat[ which(dat$perf_mix < 4 & dat$spline==1 & dat$spline_correct ==1),]
1937 dat$var <- 1/(dat$sample_size_firm-3)
1938
1939 # Preparing moderators to be included in models. Model-specific moderators are prepared within separate sections. #
1940 # Rescale median sample year by subtracting minimum year in sample for ease of interpretation #
1941 dat$medyear <- ((dat$sample_start+dat$sample_end)/2-(min(dat$sample_start,na.rm=TRUE)))
1942 dat$medyear <- dat$medyear - min(dat$medyear,na.rm=TRUE)
1943 dat$medyearsq <- dat$medyear^2
```

```
## Meta-regression for prior ties. Moderator analysis with tie purpose, median sample year, and full risk set. #
# Tests for difference in ES between R&D ties and investment ties (btt) #
dat$priorrtoz <- 0.5*log((1+dat$prior_form_r)/(1-dat$prior_form_r))
# Model 1. Includes omnibus-test (H0:B1=B2=B3=0) for R&D, manufacturing and investment ties #
prior_wls <- rma(priorrtoz,var,mods=~ resanddev + investment + manufacturing + medyear +fullrisk,method="DL",
  data=dat,btt=c(6,7))
summary(prior_wls,digits=3)

prior2_wls <- rma(priorrtoz,var,mods=~ priordicho + priorcount + published,method="DL",data=dat)
summary(prior2_wls,digits=3)
anova(prior2_wls,L=c(0,1,-1,0))

# Chisquare-test for difference between coefficients of manufacturing (2) and investment (3) ties #
anova(prior_wls,L=c(0,1,-1,0,0,0)) #Omnibus: R&D vs. Manufacturing
anova(prior_wls,L=c(0,0,1,-1,0,0)) #Omnibus: Manufacturing vs. Investment
anova(prior_wls,L=c(0,1,0,-1,0,0)) #Omnibus: Manufacturing vs. R&D
```

The future: dynamic documenting

- knitR, Rmarkdown

- **Code: Responses to historical performance feedback are heterogeneous, as the Q-statistic is** *`r round(all.hpfb_dv$QE,3)` (*p*-value = `r round(all.hpfb_dv$QEp,3)`)*
- **Text: Responses to historical performance feedback are heterogeneous, as the Q-statistic is 392.468 (p-value = 0.003)**

1) Will you be capable of reproducing your own results in half a year from now?

1) Will you be capable of reproducing your own results in half a year from now?

2) Will your co-authors be able to do this?

1) Will you be capable of reproducing your own results in half a year from now?

2) Will your co-authors be able to do this?

3) Will a colleague that is not a co-author be able to do this?

1) Will you be capable of reproducing your own results in half a year from now?

2) Will your co-authors be able to do this?

3) Will a colleague that is not a co-author be able to do this?

4) Will an independent researcher, who is in your area of expertise, be able to do this?

Assignment time

1) Will you be capable of reproducing your own results in half a year from now?

3) Will a colleague that is not a co-author be able to do this?

Data audit guidelines
https://www.tilburguniversity.edu/research/social-and-behavioral-sciences/download-guideline-datapackage-tsb/

Find me if you want help with making your work (more) reproducible!

- J.m.vanzelst@uvt.nl
- @mzelst
- S6.04