Collaboration in the 21st century: tips and tools for dealing with data

(or, how I learned to stop worrying and love version control)

Nick Waters

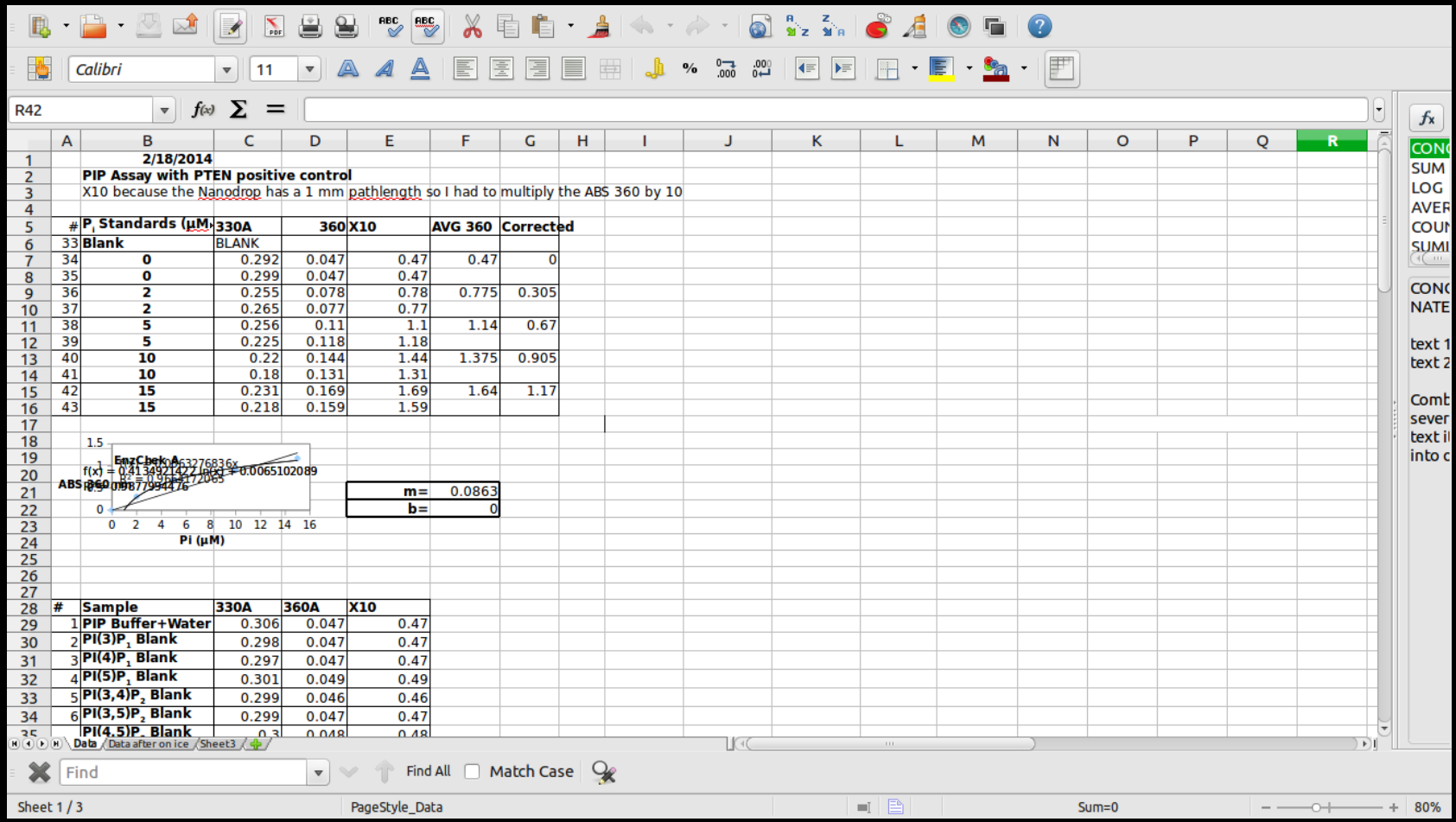
Hot Topic: March 7, 2017

National University of Ireland, Galway
James Hutton Institute, Dundee

The Problem: other people’s data

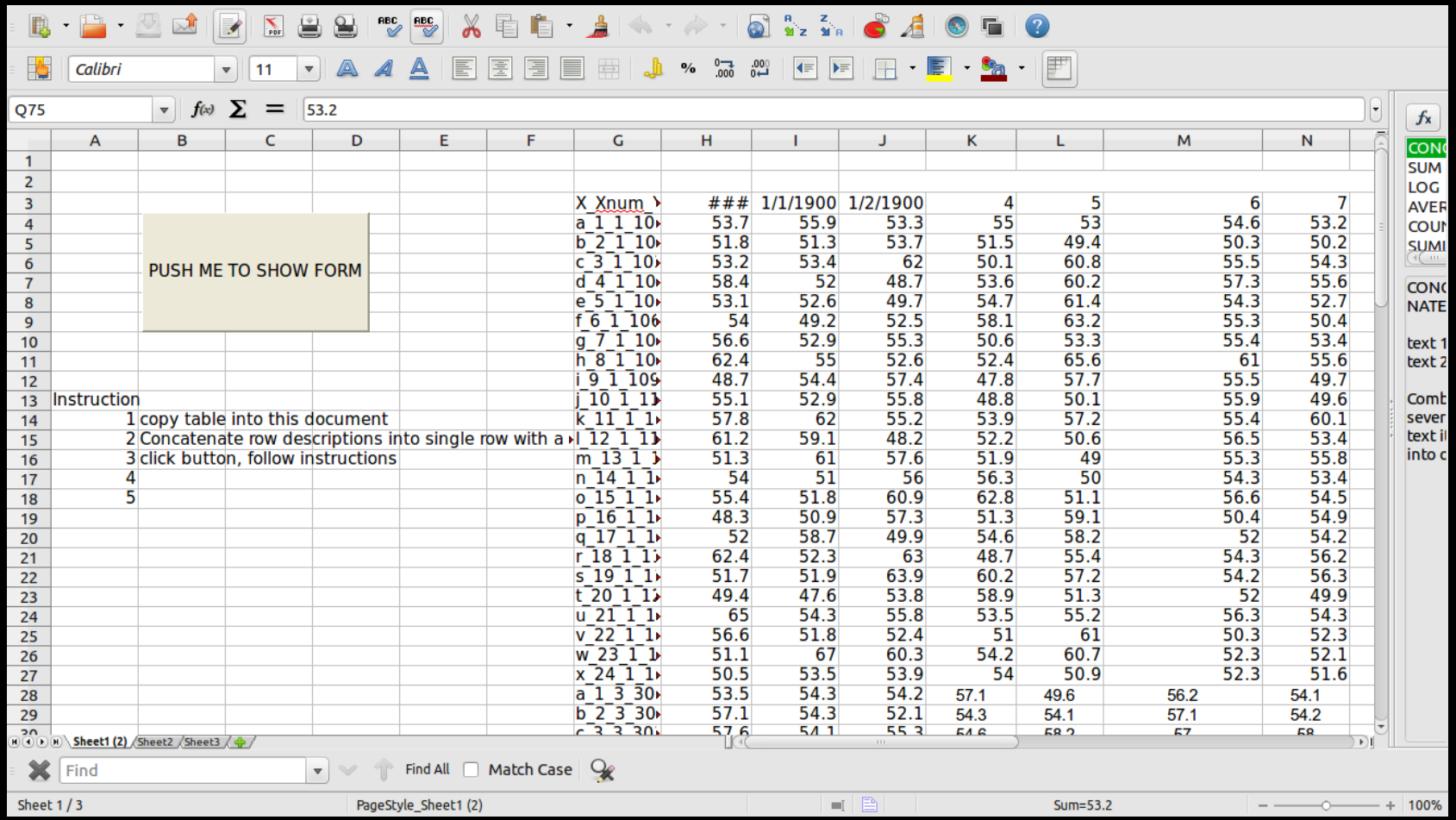
The Problem: my data

Bad Data



	A	B	C	D	E	F	G
1		**2/18/2014**					
2		**PIP Assay with PTEN positive control**					
3		X10 because the Nanodrop has a 1 mm pathlength so I had to multiply the ABS 360 by 10					
4							
5	**#**	**P_i Standards (µM)**	**330A**	**360**	**X10**	**AVG 360**	**Corrected**
6	33	**Blank**	BLANK				
7	34	**0**	0.292	0.047	0.47	0.47	0
8	35	**0**	0.299	0.047	0.47		
9	36	**2**	0.255	0.078	0.78	0.775	0.305
10	37	**2**	0.265	0.077	0.77		
11	38	**5**	0.256	0.11	1.1	1.14	0.67
12	39	**5**	0.225	0.118	1.18		
13	40	**10**	0.22	0.144	1.44	1.375	0.905
14	41	**10**	0.18	0.131	1.31		
15	42	**15**	0.231	0.169	1.69	1.64	1.17
16	43	**15**	0.218	0.159	1.59		
17							
18							
19							
20							
21						**m=**	0.0863
22						**b=**	0
23							
24							
25							
26							
27							
28	**#**	**Sample**	**330A**	**360A**	**X10**		
29	1	**PIP Buffer+Water**	0.306	0.047	0.47		
30	2	**PI(3)P_1 Blank**	0.298	0.047	0.47		
31	3	**PI(4)P_1 Blank**	0.297	0.047	0.47		
32	4	**PI(5)P_1 Blank**	0.301	0.049	0.49		
33	5	**PI(3,4)P_2 Blank**	0.299	0.046	0.46		
34	6	**PI(3,5)P_2 Blank**	0.299	0.047	0.47		
35		**PI(4,5)P_2 Blank**	0.3	0.048	0.48		

Bad Data

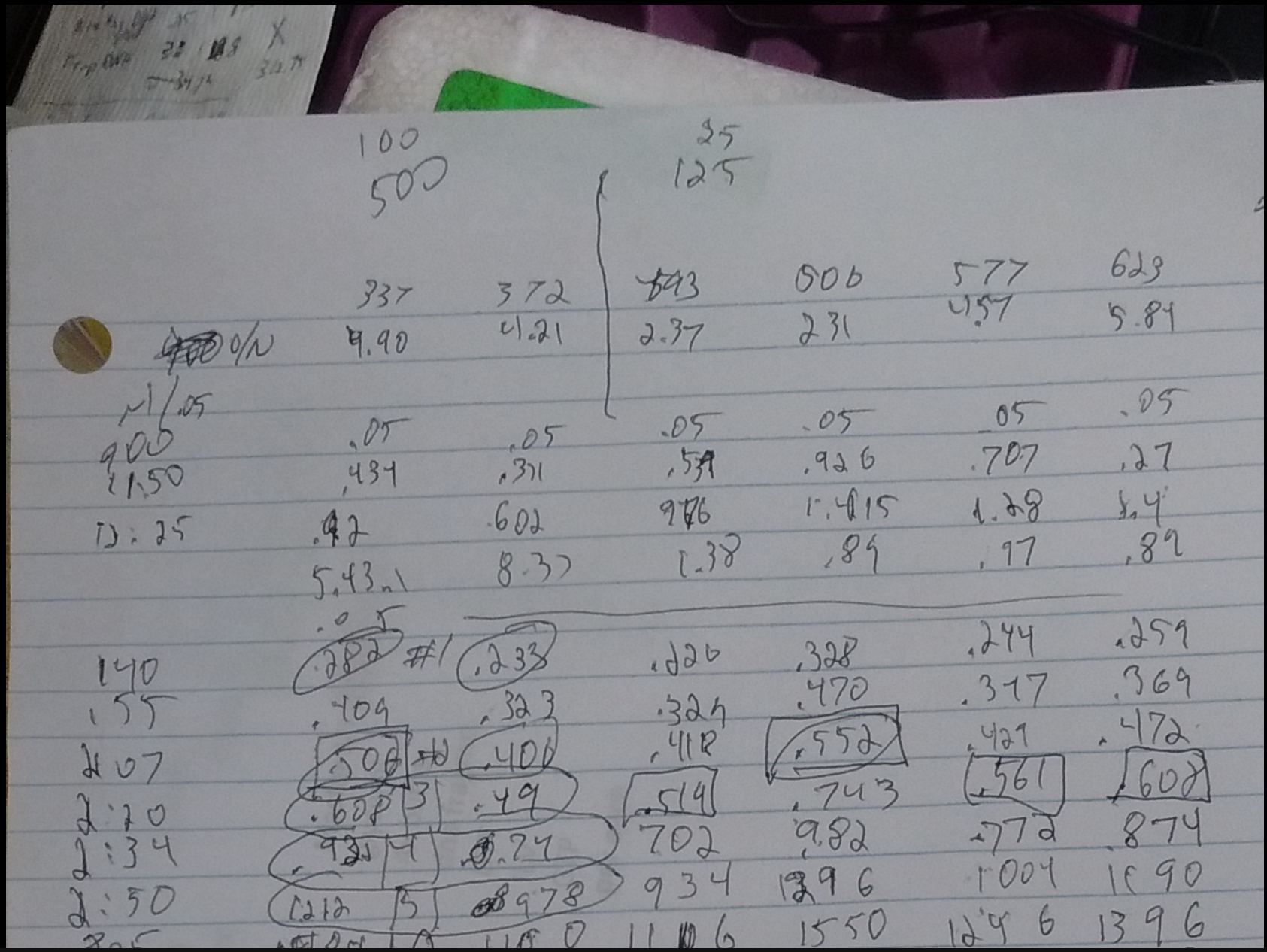


PUSH ME TO SHOW FORM

Instruction		
1	copy table into this document	
2	Concatenate row descriptions into single row with a	
3	click button, follow instructions	
4		
5		

X_Xnum	###	1/1/1900	1/2/1900	4	5	6	7
a_1_10	53.7	55.9	53.3	55	53	54.6	53.2
b_2_10	51.8	51.3	53.7	51.5	49.4	50.3	50.2
c_3_10	53.2	53.4	62	50.1	60.8	55.5	54.3
d_4_10	58.4	52	48.7	53.6	60.2	57.3	55.6
e_5_10	53.1	52.6	49.7	54.7	61.4	54.3	52.7
f_6_106	54	49.2	52.5	58.1	63.2	55.3	50.4
g_7_10	56.6	52.9	55.3	50.6	53.3	55.4	53.4
h_8_10	62.4	55	52.6	52.4	65.6	61	55.6
i_9_109	48.7	54.4	57.4	47.8	57.7	55.5	49.7
j_10_11	55.1	52.9	55.8	48.8	50.1	55.9	49.6
k_11_1	57.8	62	55.2	53.9	57.2	55.4	60.1
l_12_1_11	61.2	59.1	48.2	52.2	50.6	56.5	53.4
m_13_1	51.3	61	57.6	51.9	49	55.3	55.8
n_14_1_1	54	51	56	56.3	50	54.3	53.4
o_15_1_1	55.4	51.8	60.9	62.8	51.1	56.6	54.5
p_16_1_1	48.3	50.9	57.3	51.3	59.1	50.4	54.9
q_17_1_1	52	58.7	49.9	54.6	58.2	52	54.2
r_18_1_1	62.4	52.3	63	48.7	55.4	54.3	56.2
s_19_1_1	51.7	51.9	63.9	60.2	57.2	54.2	56.3
t_20_1_1	49.4	47.6	53.8	58.9	51.3	52	49.9
u_21_1_1	65	54.3	55.8	53.5	55.2	56.3	54.3
v_22_1_1	56.6	51.8	52.4	51	61	50.3	52.3
w_23_1_1	51.1	67	60.3	54.2	60.7	52.3	52.1
x_24_1_1	50.5	53.5	53.9	54	50.9	52.3	51.6
a_1_3_30	53.5	54.3	54.2	57.1	49.6	56.2	54.1
b_2_3_30	57.1	54.3	52.1	54.3	54.1	57.1	54.2

Bad Data


100
500
25
125
337
372
993
006
577
623
4.57
4.90
4.21
2.37
231
5.84
.05
.05
.05
.05
.05
.05
.434
.311
.926
.707
.27
.602
.84
.97
.82
.282
#1
.233
.328
.244
.259
.404
.323
.324
.470
.317
.369
.506
.400
.552
.421
.472
.608
.49
.514
.743
.561
.608
.924
.74
.702
.982
.772
.874
.978
.934
1.396
1.004
1.390
1.212
1.106
1.550
1.246
1.396

The Real Problem: Separation of Concerns

Separation of Concerns

Separation of concerns: The practice of separating data from analysis.

Why do we need to separate data from analysis?

- Saves time for repeated analysis
- Reproducible results
- Prevents data loss or contamination
- Prevents user error
- Easier to track over time

Data

Raw data should never be changed.

- Store it in plain text format in utf-8
- Store it with metadata in the same directory also in plain text
- Make sure it is backed up

Analysis with Excel

Do's:

- Include a "metadata" sheet
- Make new columns/rows for each step in analysis
- Document each step
- Explicitly set data types

Don'ts:

- Link Excel workbooks or sheets
- Rely on color coding
- Use macros
- Copy and paste

Analysis with Literate Programming

Literate programming mixes text description with data analysis. Examples include Jupyter notebooks, knitr/sweave, emacs's org-mode.

Do's:

- Explicitly state where the data is coming from
- Keep the analysis scoped to a single folder
- Document each step
- Ensure compilation will fail with ill-formatted data
- When finished, zip the entire directory for stable storage.

Don'ts:

- Mix literate programming with hardcoding
- Leave out print statements
- Forget sanity checks

Analysis with Code

R, python, MATLAB, etc

Do's:

- Use version control
- Explicitly state where the data is coming from
- Write tests and comments for all the interesting steps
- Make the program write out a log
- Ensure compilation will fail with ill-formatted data
- Time-stamp results

Don'ts:

- Hard-code paths
- Write unhelpful comments
- Write similar code
- Forget sanity checks
- Rely on Stack Overflow

Version control

Overview

Version control keeps track of changes made to data and other files

- Dropbox/Google Drive
- Google docs
- Time Machine
- Lab wiki's
- Git, subversion, or murcurial

Things that should be version controlled

- Protocols
- Manuscripts
- Analysis pipeline

Things that shouldn’t be version controlled

- Data: backups are different from version control
- Temporary files
- Complex file formats, big files, etc

Takeaways

Moral of the Story

Write data for computers, write code for people

Homework

- Identify analyses that could be automated
- If you use excel, make a “metadata” sheet
- Never link Excel workbooks
- Restructure project folders