Data management

Owen Jones
jones@biology.sdu.dk

Overview

- Why you should care.
- Collecting and organising.
- Storing.
- Using.

Why you should care

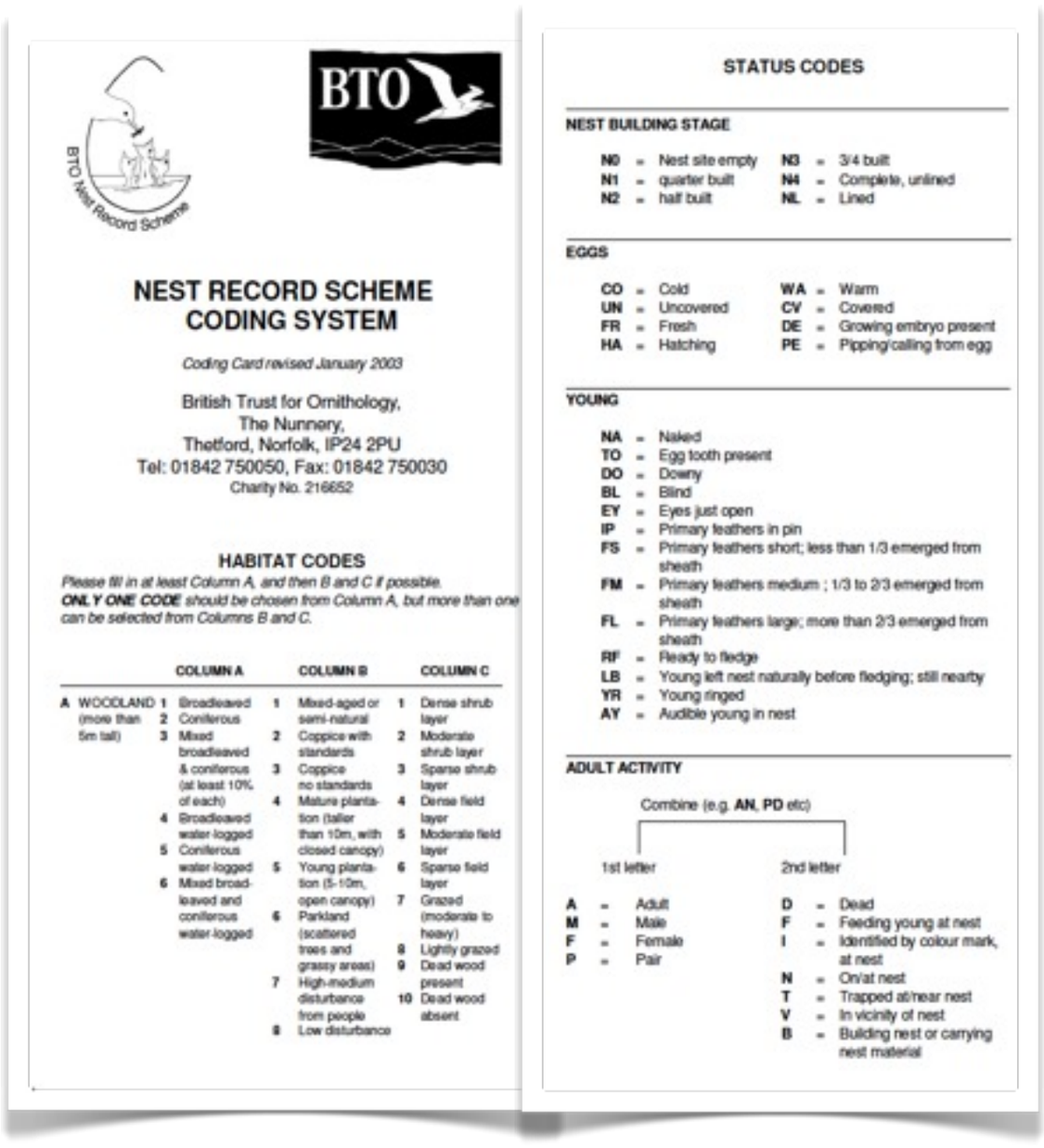
- Data are the raw material of scientific studies.
- Typically 80% of effort on analysis is spent cleaning data (getting it ready to analyse).
- We should minimise this work!
- Data are valuable and should be preserved.

Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices

Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices



BTO Nest Record Scheme

BTO

NEST RECORD SCHEME CODING SYSTEM

Coding Card revised January 2003

British Trust for Ornithology,
The Nunnery,
Thetford, Norfolk, IP24 2PU
Tel: 01842 750050, Fax: 01842 750030
Charity No. 216652

HABITAT CODES

Please fill in at least Column A, and then B and C if possible.
ONLY ONE CODE *should be chosen from Column A, but more than one can be selected from Columns B and C.*

	COLUMN A		COLUMN B		COLUMN C
A	WOODLAND (more than 5m tall)	1 Broadleaved	1 Mixed-aged or semi-natural	1	Dense shrub layer
		2 Coniferous	2 Coppice with standards	2	Moderate shrub layer
		3 Mixed broadleaved & coniferous (at least 10% of each)	3 Coppice no standards	3	Sparse shrub layer
		4 Broadleaved water-logged	4 Mature plantation (taller than 10m, with closed canopy)	4	Dense field layer
		5 Coniferous water-logged	5 Young plantation (5-10m, open canopy)	5	Moderate field layer
		6 Mixed broad-leaved and coniferous water-logged	6 Parkland (scattered trees and grassy areas)	6	Sparse field layer
			7 High-medium disturbance from people	7	Grazed (moderate to heavy)
			8 Low disturbance	8	Lightly grazed
				9	Dead wood present
				10	Dead wood absent

STATUS CODES

NEST BUILDING STAGE

N0 = Nest site empty		**N3** = 3/4 built	
N1 = quarter built		**N4** = Complete, unlined	
N2 = half built		**NL** = Lined	

EGGS

CO = Cold		**WA** = Warm	
UN = Uncovered		**CV** = Covered	
FR = Fresh		**DE** = Growing embryo present	
HA = Hatching		**PE** = Pipping/calling from egg	

YOUNG

- **NA** = Naked
- **TO** = Egg tooth present
- **DO** = Downy
- **BL** = Blind
- **EY** = Eyes just open
- **IP** = Primary feathers in pin
- **FS** = Primary feathers short; less than 1/3 emerged from sheath
- **FM** = Primary feathers medium ; 1/3 to 2/3 emerged from sheath
- **FL** = Primary feathers large; more than 2/3 emerged from sheath
- **RF** = Ready to fledge
- **LB** = Young left nest naturally before fledging; still nearby
- **YR** = Young ringed
- **AY** = Audible young in nest

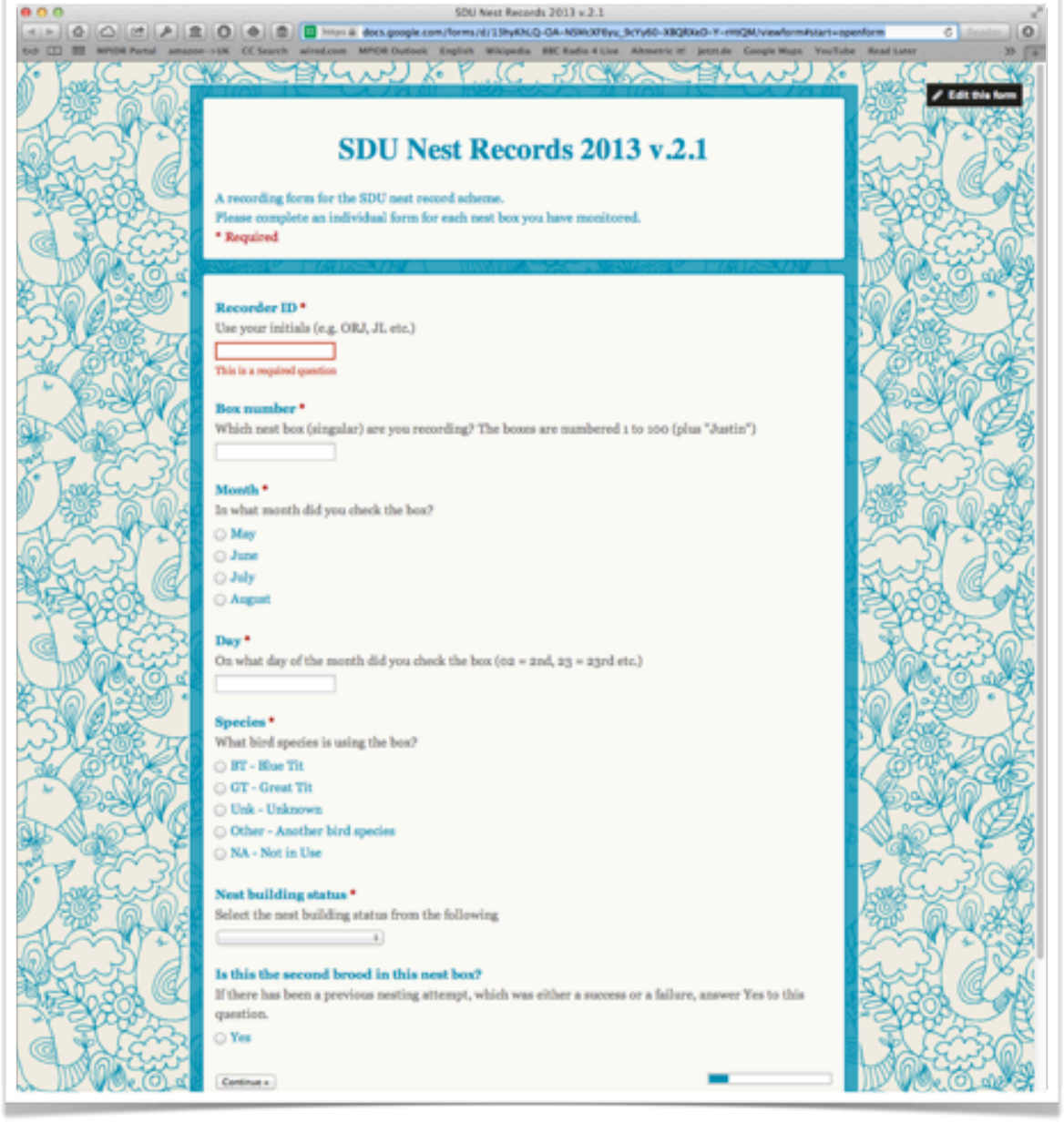
ADULT ACTIVITY

Combine (e.g. **AN**, **PD** etc)

1st letter			2nd letter	
A	=	Adult	**D**	= Dead
M	=	Male	**F**	= Feeding young at nest
F	=	Female	**I**	= Identified by colour mark, at nest
P	=	Pair	**N**	= On/at nest
			T	= Trapped at/near nest
			V	= In vicinity of nest
			B	= Building nest or carrying nest material

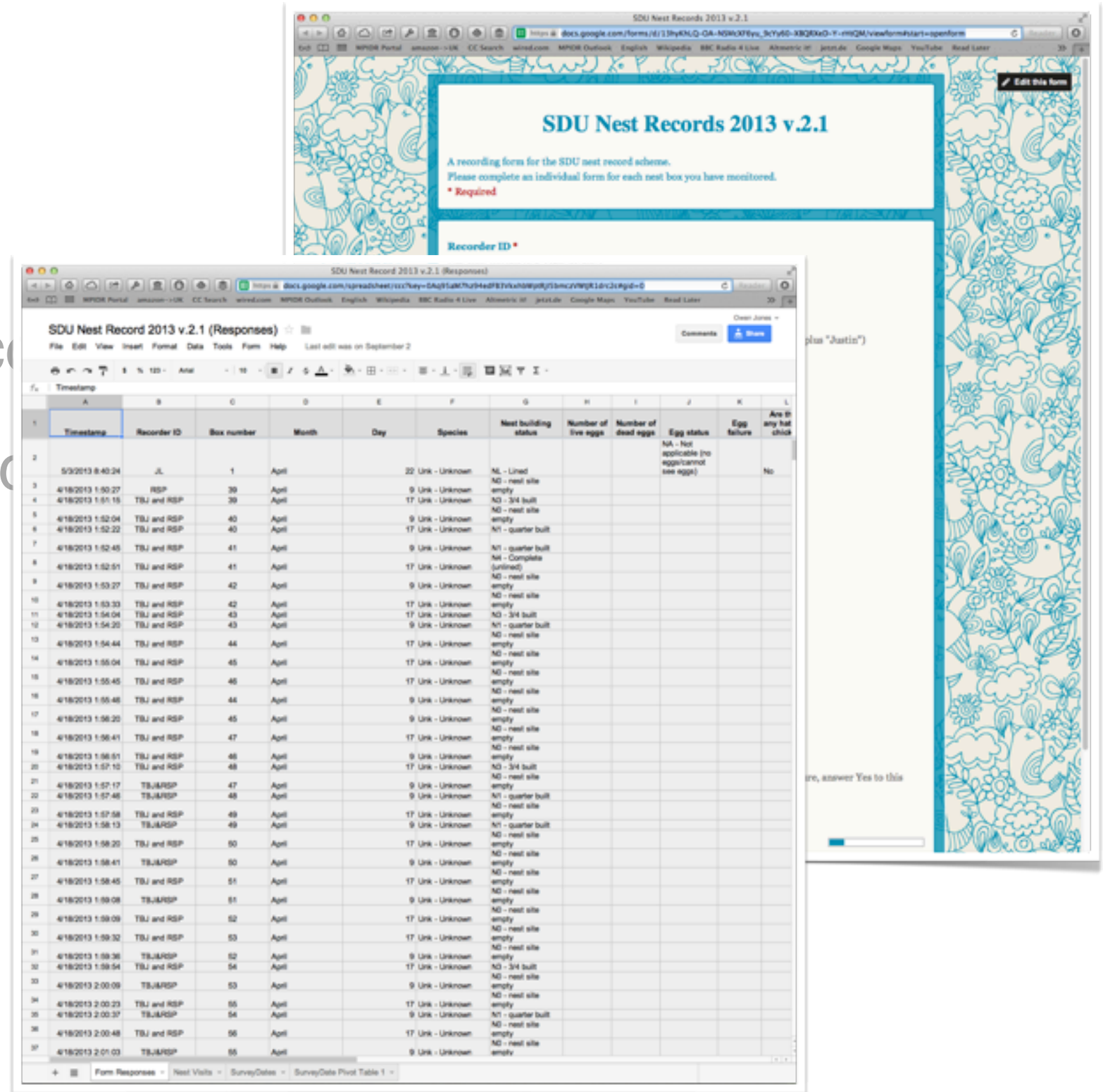
Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices



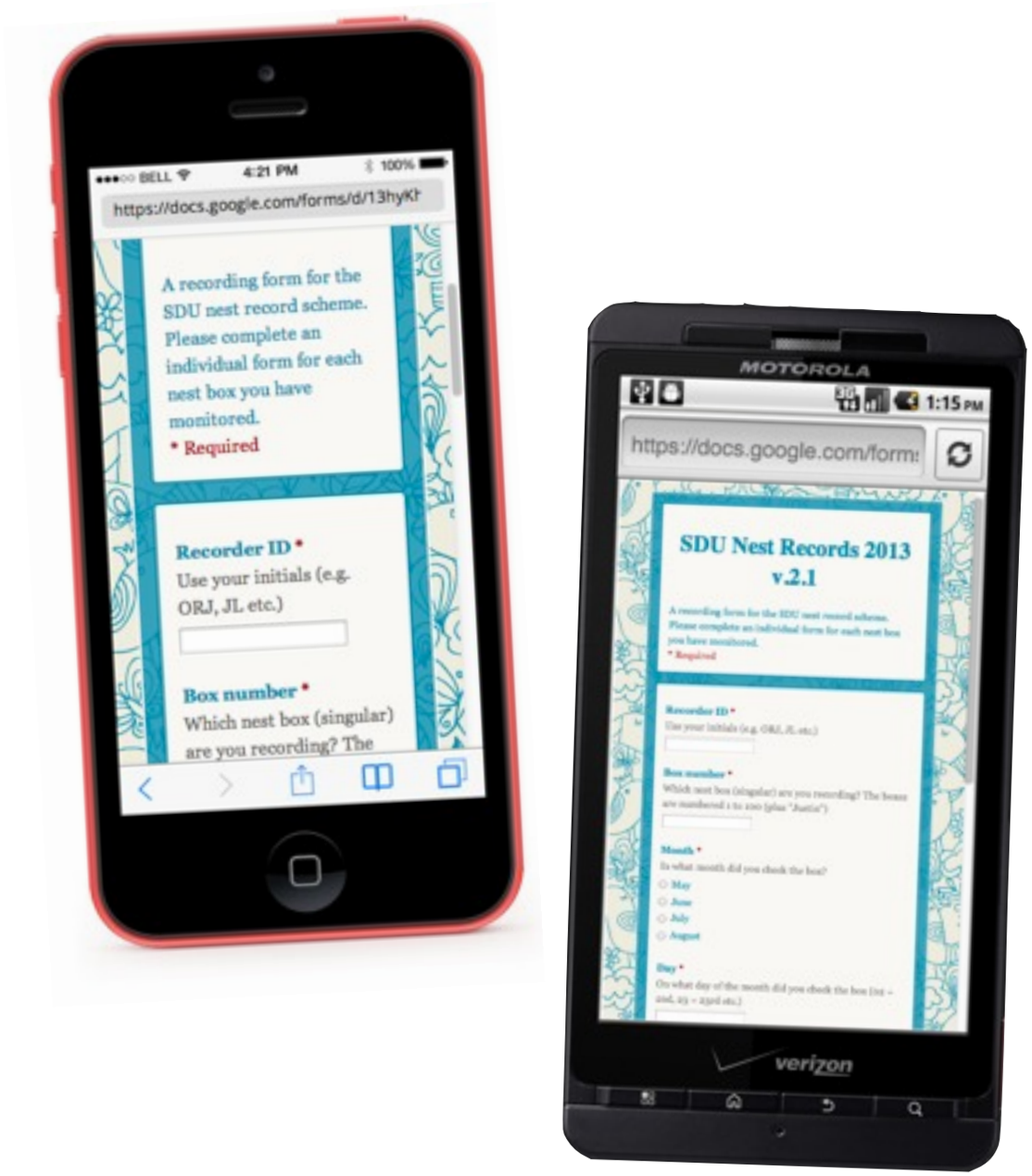
Collecting data

- Use printed forms/rec
- Standardised inputs (c
- Google forms
- Mobile devices



Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices



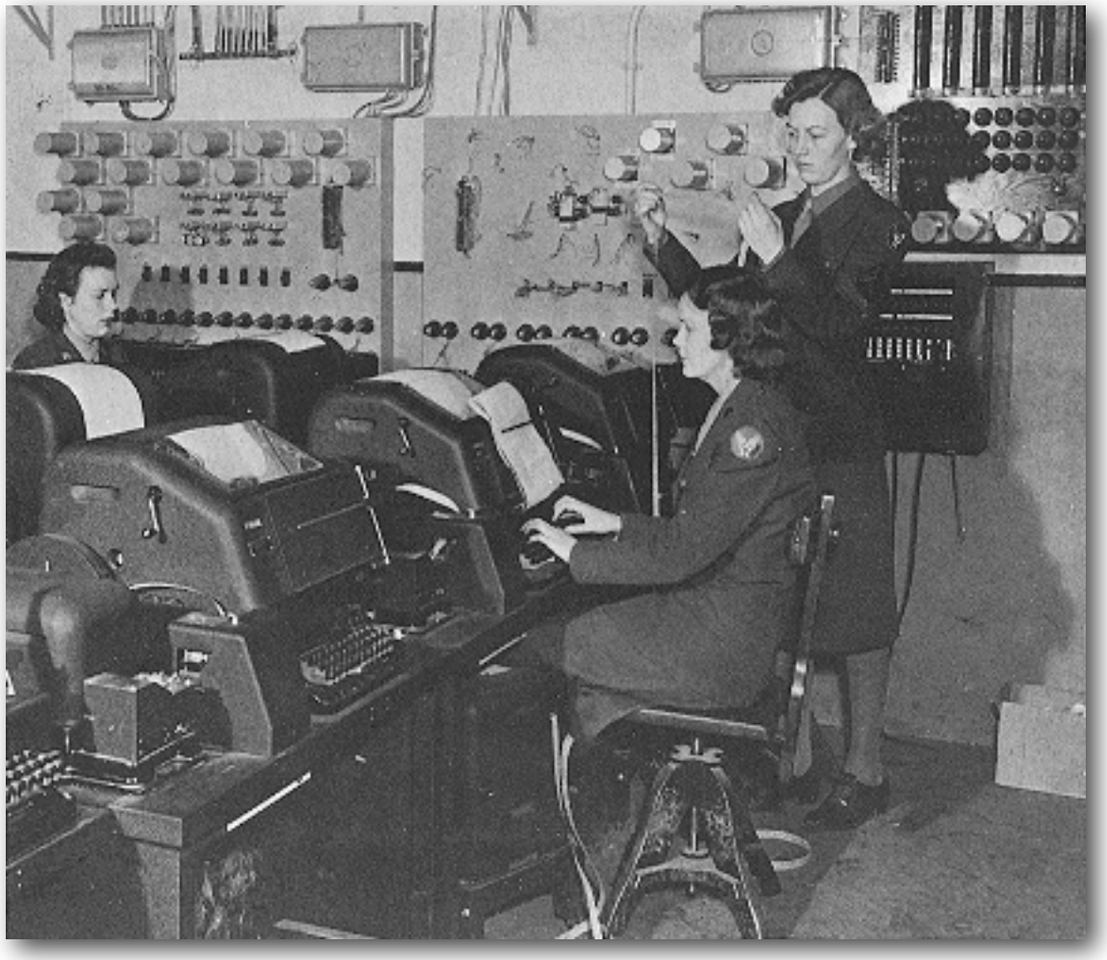
Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices

Privacy concerns
No emergency hard copy
Battery life in field

Organising data

- 98%* of data entry done in Excel
- Excel is useful but can lead to bad habits



*I made this number up

Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

	A
1	Length
2	15
3	62
4	45
5	
6	54
7	76
8	86

Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

	A
1	Date
2	31-Jan-76
3	27/01/82
4	27/1/82
5	14.1.70
6	05/08/1990
7	09 April 1979
8	03/04/2001
9	

Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

	A	B	C
1	ID	Length	Weight
2	1	15	3.00
3	2	62	7.00
4	3	45	4.00
5	4	did not meaure	
6	5	54	9.00
7	6	76	5.00
8	7	86	7.80

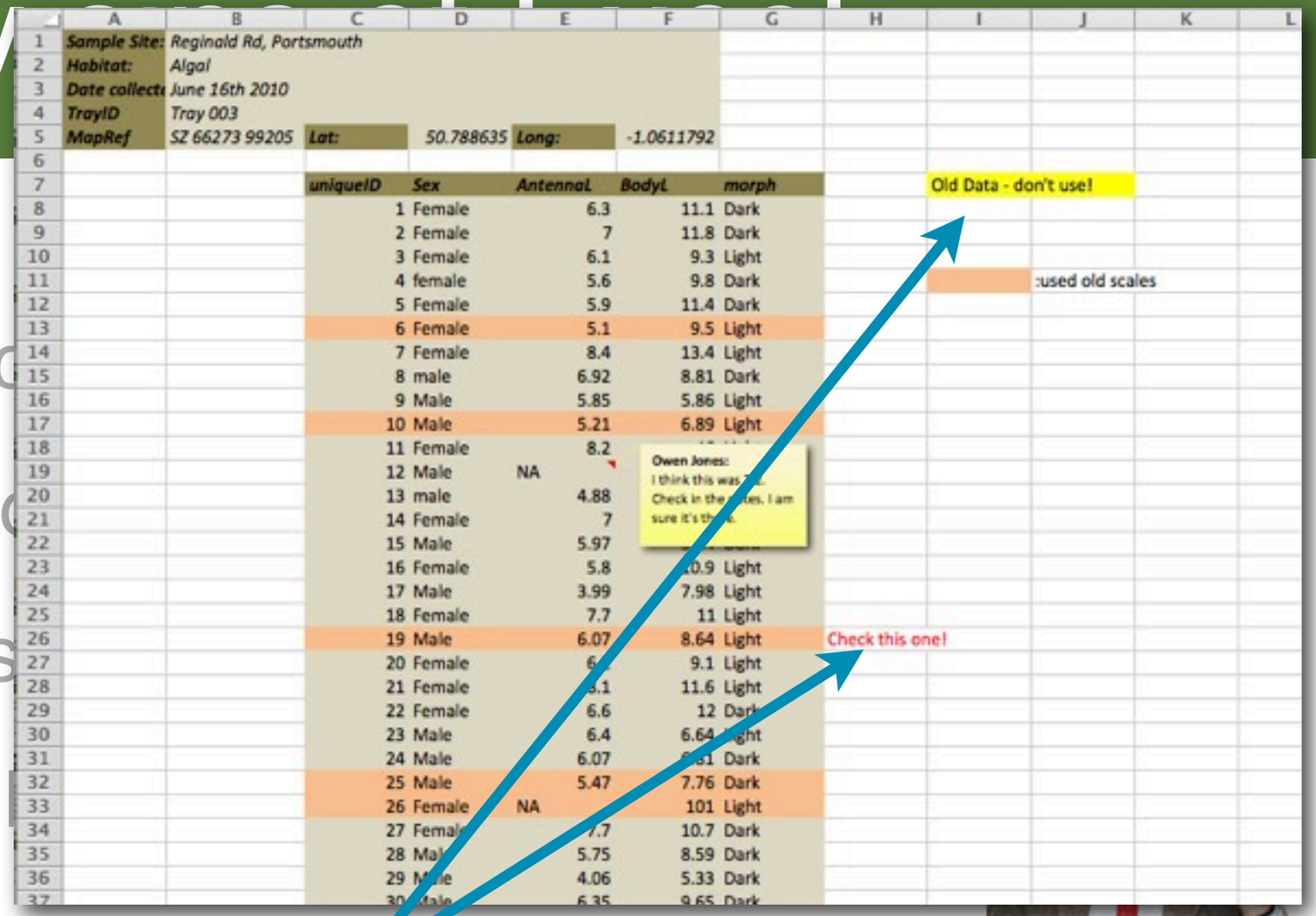
Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Ins
- Ins
- Us
- Us

	A	B	C	D	E	F	G	H
1	*Sample Site:*	*Reginald Rd, Portsmouth*						
2	*Habitat:*	*Algal*						
3	*Date collecte*	*June 16th 2010*						
4	*TrayID*	*Tray 003*						
5	*MapRef*	*SZ 66273 99205*	*Lat:*	*50.788635*	*Long:*	*-1.0611792*		
6								
7		*uniqueID*	*Sex*	*AntennaL*	*BodyL*	*morph*		
8		1	Female	6.3	11.1	Dark		
9		2	Female	7	11.8	Dark		
10		3	Female	6.1	9.3	Light		
11		4	female	5.6	9.8	Dark		
12		5	Female	5.9	11.4	Dark		
13		6	Female	5.1	9.5	Light		
14		7	Female	8.4	13.4	Light		
15		8	male	6.92	8.81	Dark		
16		9	Male	5.85	5.86	Light		
17		10	Male	5.21	6.89	Light		
18		11	Female	8.2	12	Light		
19		12	Male	NA	9.29	Dark		
20		13	male	4.88	7.38	Dark		
21		14	Female	7	9.6	Light		

Deadly sins of Excel

- Using blank c
- Inconsistent c
- Merging cells
- Including mul
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns



	A	B	C	D	E	F	G	H	I	J	K	L
1	Sample Site:	Reginald Rd, Portsmouth										
2	Habitat:	Algal										
3	Date collecte	June 16th 2010										
4	TrayID	Tray 003										
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792						
6												
7			uniqueID	Sex	AntennaL	BodyL	morph		Old Data - don't use!			
8			1	Female	6.3	11.1	Dark					
9			2	Female	7	11.8	Dark					
10			3	Female	6.1	9.3	Light					
11			4	female	5.6	9.8	Dark			used old scales		
12			5	Female	5.9	11.4	Dark					
13			6	Female	5.1	9.5	Light					
14			7	Female	8.4	13.4	Light					
15			8	male	6.92	8.81	Dark					
16			9	Male	5.85	5.86	Light					
17			10	Male	5.21	6.89	Light					
18			11	Female	8.2							
19			12	Male	NA							
20			13	male	4.88							
21			14	Female	7							
22			15	Male	5.97							
23			16	Female	5.8		Light					
24			17	Male	3.99	7.98	Light					
25			18	Female	7.7	11	Light					
26			19	Male	6.07	8.64	Light	Check this one!				
27			20	Female	6	9.1	Light					
28			21	Female	6.1	11.6	Light					
29			22	Female	6.6	12	Dark					
30			23	Male	6.4	6.64	ght					
31			24	Male	6.07	8.81	Dark					
32			25	Male	5.47	7.76	Dark					
33			26	Female	NA	101	Light					
34			27	Female	7.7	10.7	Dark					
35			28	Male	5.75	8.59	Dark					
36			29	Male	4.06	5.33	Dark					
37			30	Male	6.35	9.65	Dark					

Owen Jones:
I think this was [illegible]
Check in the notes. I am sure it's the [illegible].

Deadly sins of Excel

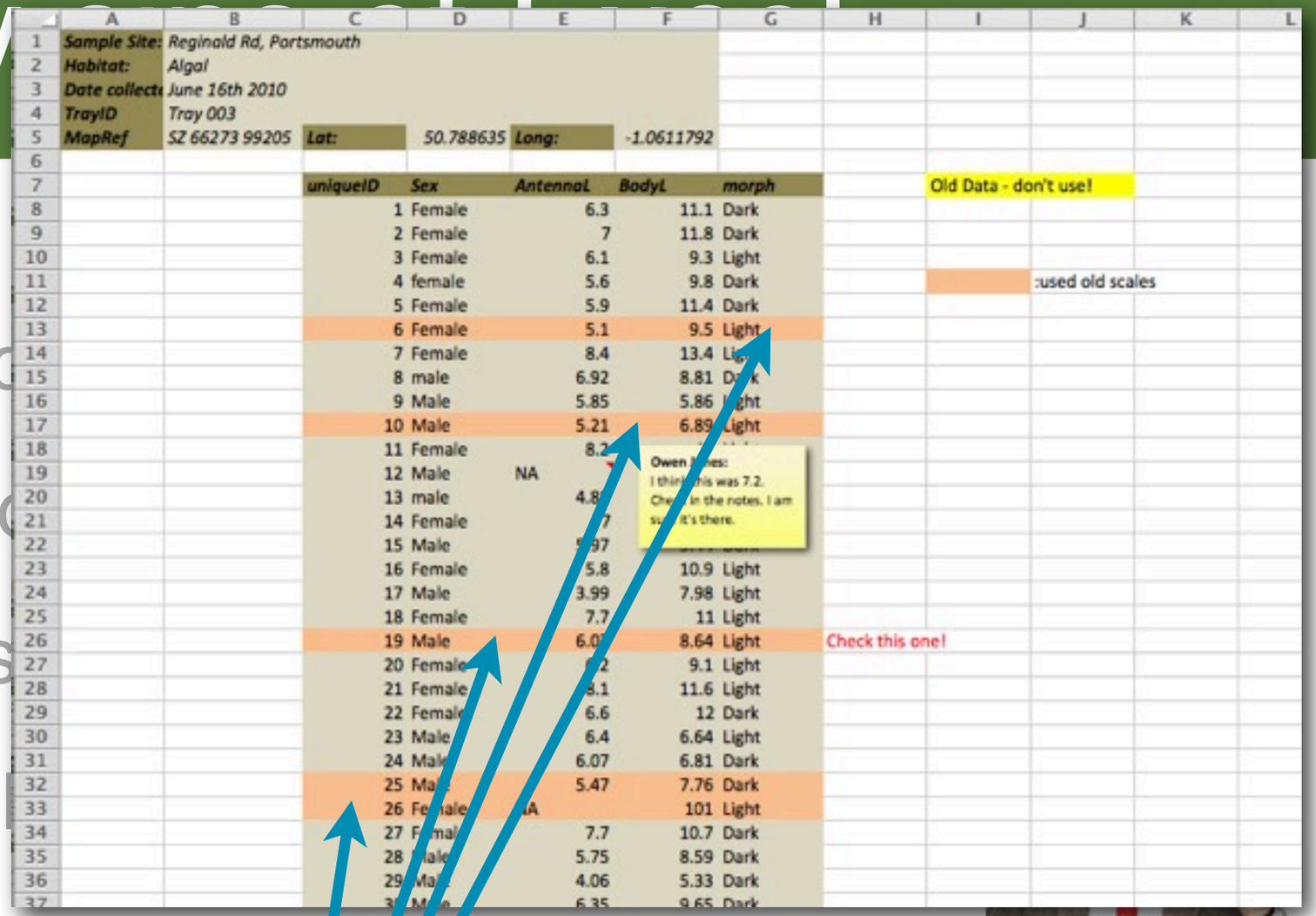
- Using blank c
- Inconsistent c
- Merging cells
- Including mul
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sample Site:	Reginald Rd, Portsmouth										
2	Habitat:	Algal										
3	Date collecte	June 16th 2010										
4	TrayID	Tray 003										
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792						
6												
7			uniqueID	Sex	AntennaL	BodyL	morph		Old Data - don't use!			
8			1	Female	6.3	11.1	Dark					
9			2	Female	7	11.8	Dark					
10			3	Female	6.1	9.3	Light					
11			4	female	5.6	9.8	Dark			used old scales		
12			5	Female	5.9	11.4	Dark					
13			6	Female	5.1	9.5	Light					
14			7	Female	8.4	13.4	Light					
15			8	male	6.92	8.81	Dark					
16			9	Male	5.85	5.86	Light					
17			10	Male	5.21	6.89	Light					
18			11	Female	8.2							
19			12	Male	NA							
20			13	male	4.88							
21			14	Female	7							
22			15	Male	5.97							
23			16	Female	5.8	10.9	Light					
24			17	Male	3.9	7.98	Light					
25			18	Female	7	11	Light					
26			19	Male	6.07	8.64	Light	Check this one!				
27			20	Female	6.2	9.1	Light					
28			21	Female	8.1	11.6	Light					
29			22	Female	6.6	12	Dark					
30			23	Male	6.4	6.64	Light					
31			24	Male	6.07	6.81	Dark					
32			25	Male	5.47	7.76	Dark					
33			26	Female	N	101	Light					
34			27	Female	7.7	10.7	Dark					
35			28	Male	5.75	8.59	Dark					
36			29	Male	4.06	5.33	Dark					
37			30	Male	6.35	9.65	Dark					

Owen Jones:
I think this was 7.2. Check in the notes. I am sure it's there.

Deadly sins of Excel

- Using blank c
- Inconsistent c
- Merging cells
- Including mul
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns



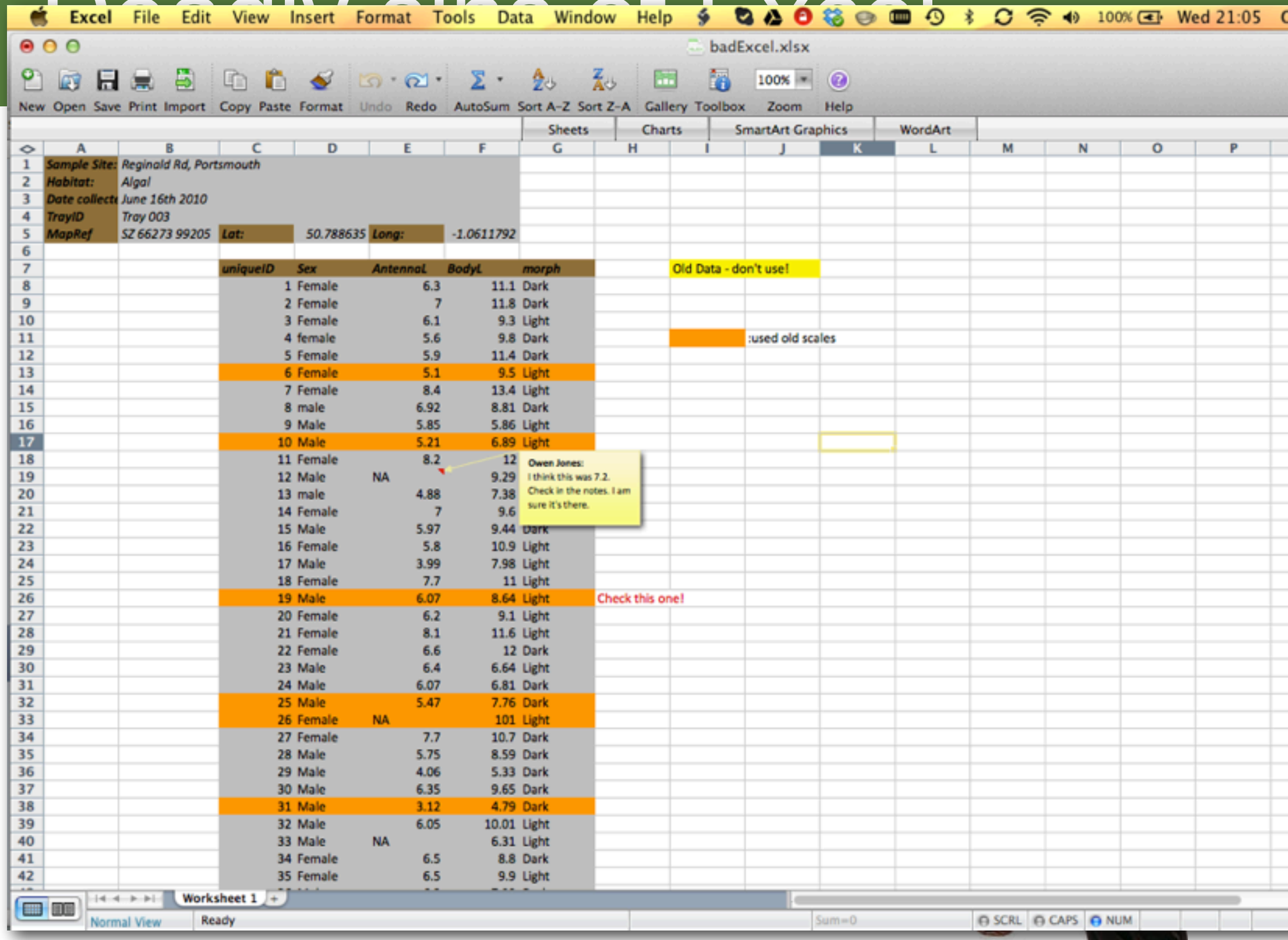
	A	B	C	D	E	F	G	H	I	J	K	L
1	Sample Site:	Reginald Rd, Portsmouth										
2	Habitat:	Algal										
3	Date collecte	June 16th 2010										
4	TrayID	Tray 003										
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792						
6												
7			uniqueID	Sex	AntennaL	BodyL	morph		Old Data - don't use!			
8			1	Female	6.3	11.1	Dark					
9			2	Female	7	11.8	Dark					
10			3	Female	6.1	9.3	Light					
11			4	female	5.6	9.8	Dark			used old scales		
12			5	Female	5.9	11.4	Dark					
13			6	Female	5.1	9.5	Light					
14			7	Female	8.4	13.4	Li					
15			8	male	6.92	8.81	D					
16			9	Male	5.85	5.86	ght					
17			10	Male	5.21	6.89	Light					
18			11	Female	8.2							
19			12	Male	NA							
20			13	male	4.8							
21			14	Female		7						
22			15	Male		97						
23			16	Female	5.8	10.9	Light					
24			17	Male	3.99	7.98	Light					
25			18	Female	7.7	11	Light					
26			19	Male	6.0	8.64	Light	Check this one!				
27			20	Female		2	9.1	Light				
28			21	Female		8.1	11.6	Light				
29			22	Female		6.6	12	Dark				
30			23	Male		6.4	6.64	Light				
31			24	Male		6.07	6.81	Dark				
32			25	Ma		5.47	7.76	Dark				
33			26	Female	NA		101	Light				
34			27	Female		7.7	10.7	Dark				
35			28	Male		5.75	8.59	Dark				
36			29	Ma		4.06	5.33	Dark				
37			30	Male		6.35	9.65	Dark				

Owen [illegible]:
I think this was 7.2. Check in the notes. I am sure it's there.

Deadly sins of Excel

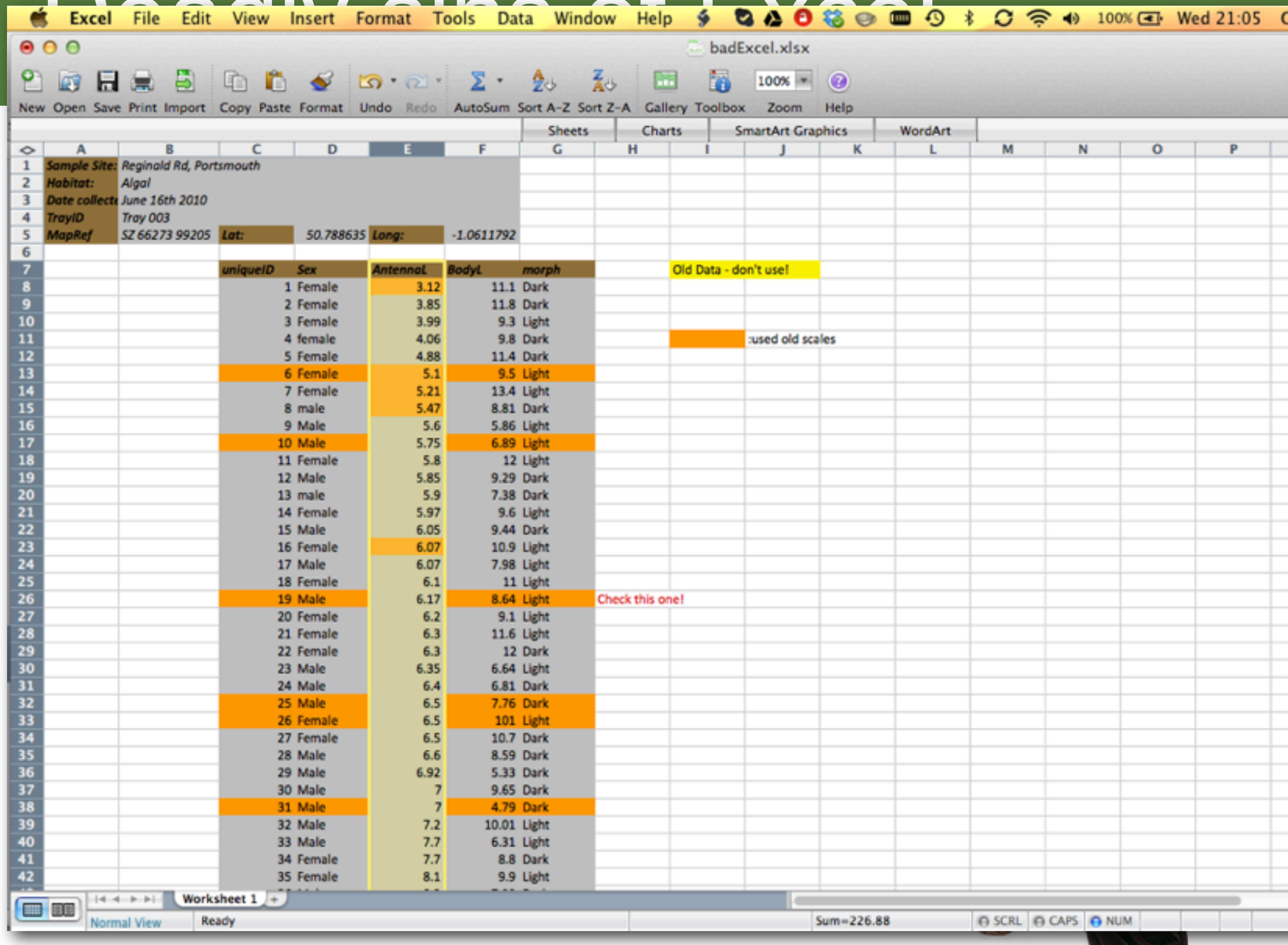
- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

Deadly sins of Excel



- Using “sort” on only some columns

Deadly sins of Excel



- Using “sort” on only some columns

Good practice in Excel

- Use NA for missing data
- Dates: use columns for day, month, year
- Include only one table per worksheet
- Have a “comments” column if necessary
- Use a data column instead of colour coding
- Be very careful with “sort”

Data format

- Each variable should form a column.
- Each observation should form a row.

Bad

	treatmentA	treatmentB
John Smith	—	5
Jane Doe	1	4
Mary Johnson	2	3

Good

name	treatment	result
Jane Doe	a	1
Jane Doe	b	4
John Smith	a	—
John Smith	b	5
Mary Johnson	a	2
Mary Johnson	b	3

Data format

Problem: Column headers are values not variable names

Bad

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k
Agnostic	27	34	60	81	76	137	122
Atheist	12	27	37	52	35	70	73
Buddhist	27	21	30	34	33	58	62
Catholic	418	617	732	670	638	1116	949
DonâĂŹt know/refused	15	14	15	11	10	35	21
Evangelical Prot	575	869	1064	982	881	1486	949
Hindu	1	9	7	9	11	34	47
Historically Black Prot	228	244	236	238	197	223	131
Jehovah's Witness	20	27	24	24	21	30	15
Jewish	19	19	25	25	30	95	69

Good

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Data format

Problem: Variables stored in both rows and columns

Bad

shown in Table 8. This dataset comes from the World Health Organisation, and records the counts of confirmed tuberculosis cases by country, year, and demographic group. The demographic groups are broken down by sex (m, f) and age (0–14, 15–25, 25–34, 35–44, 45–54, 55–64, unknown).

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Table 8: Original TB dataset. Corresponding to each 'm' column for males, there is also an 'f' column for females: `f1524`, `f2534` and so on. These are not shown to conserve space. Note the mixture of 0s and missing values (—): this is due to the data collection process and the distinction is important for this dataset.

Data format

Problem: Variables stored in both rows and columns

Good!

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Data format

Problem: multiple data types in each column

Bad

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—	—	34.5
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—	—	16.8
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—	—	—

Table 10: Original weather dataset. There is a column for each possible day in the month. Columns `d11` to `d31` have been omitted to conserve space.

Data format

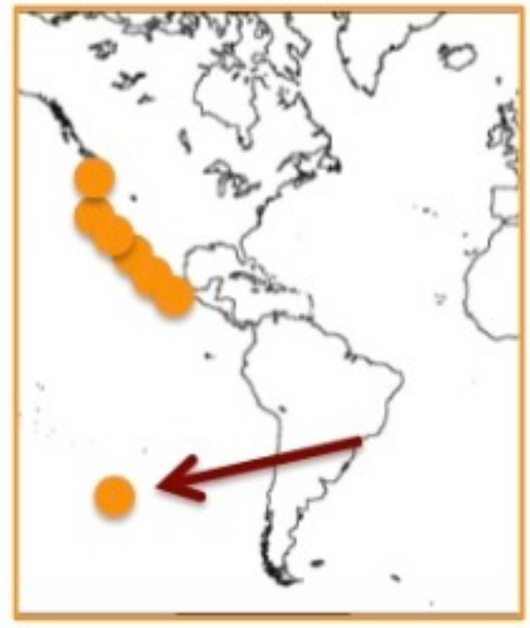
Problem: multiple data types in each column

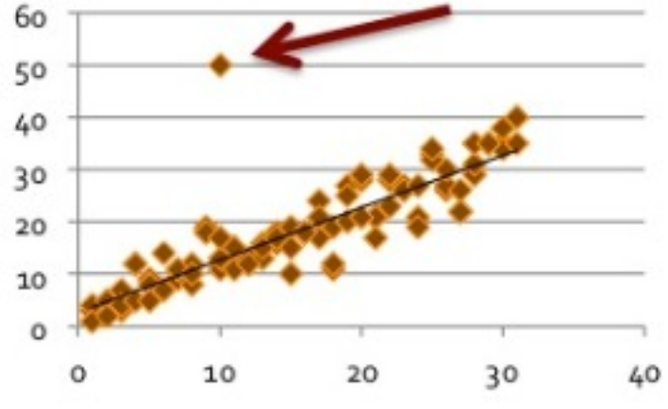
Good!

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

Quality control

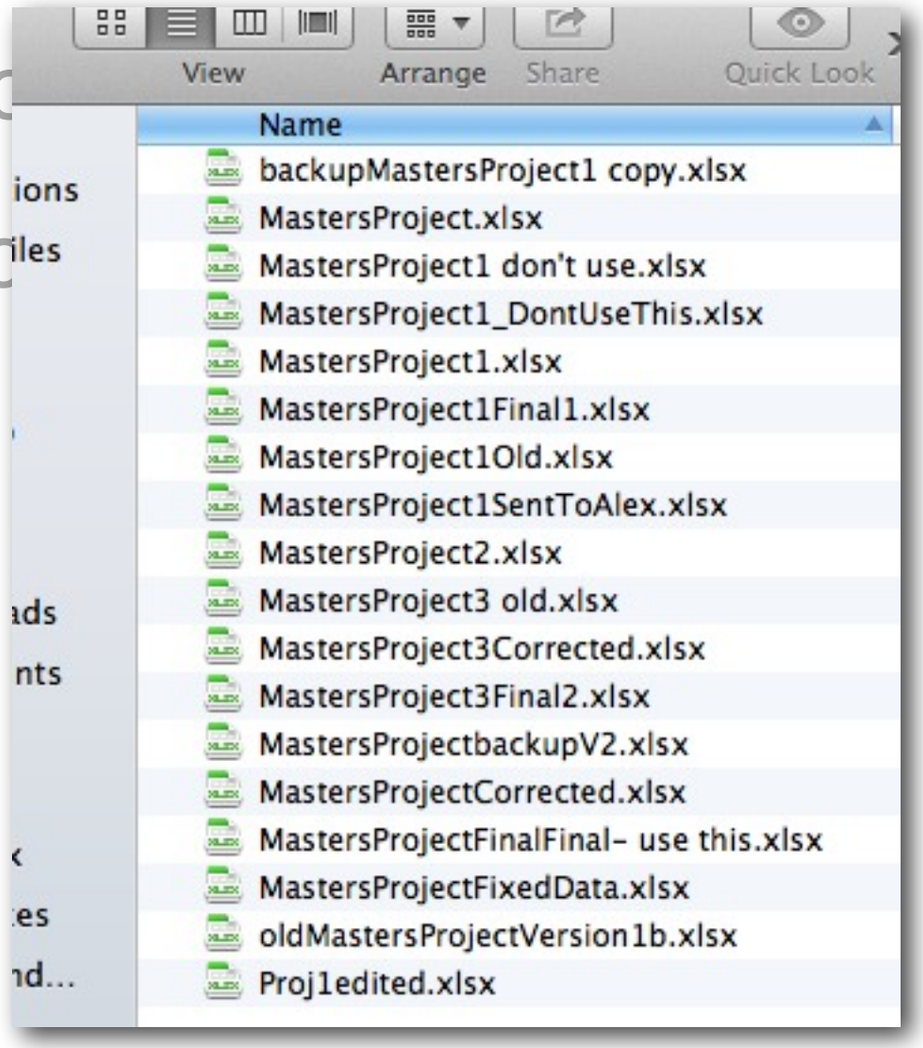
- Check data for outliers, weird data as soon as possible.
- Aim to make work flows repeatable.
- Keep raw data raw - use scripts to process.
  - Gold standard: from raw data to analysis and plots with script(s).
  - Work is repeatable at click of a button.
  - Easy to correct.





Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable fo
- Backups (inc



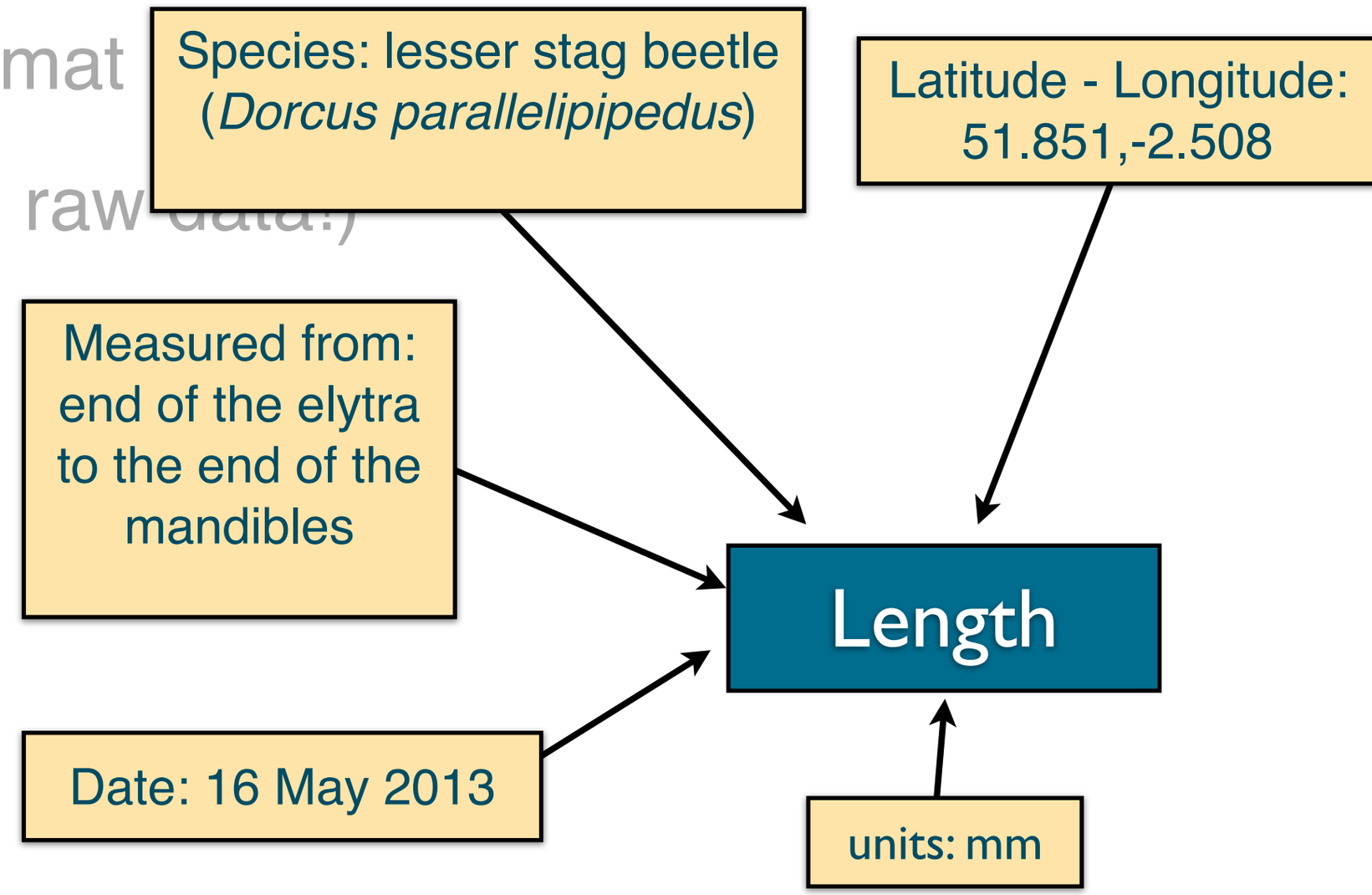
Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format (csv, txt)
- Backups (incl. raw data!)

Length

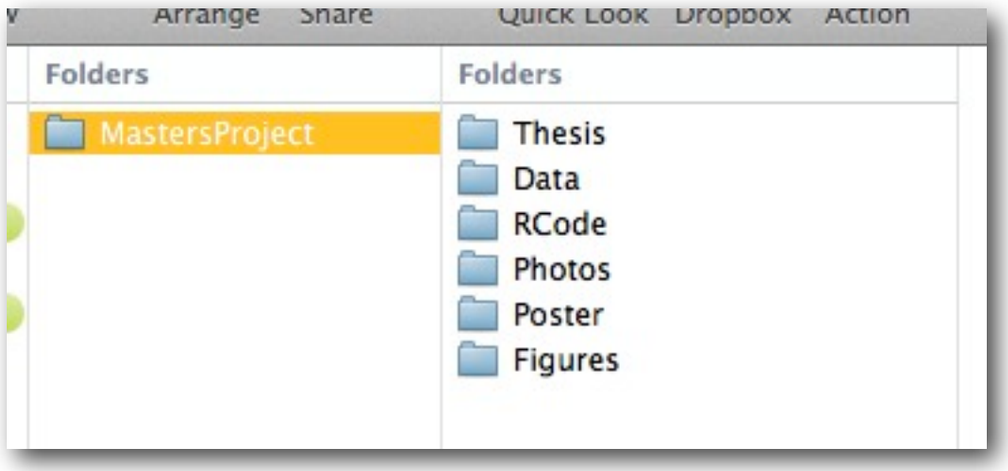
Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format
- Backups (incl. raw data!)



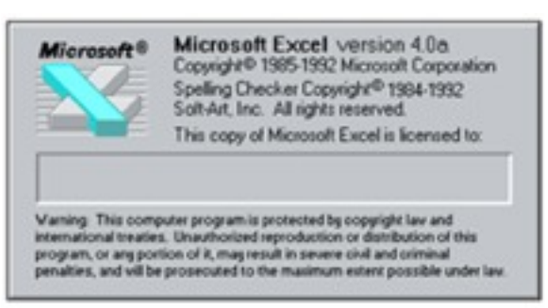
Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format (csv, txt)
- Backups (incl. raw data!)



Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format (csv, txt)
- Backups (incl. raw data!)



A fatal exception 0E has occurred at 0028:C00069F8 in VxD VMM(01) +
000059F8. The current application will be terminated.

* Press any key to terminate the current application.
* Press CTRL+ALT+DEL to restart your computer. You will
   lose any unsaved information in all applications.

Press any key to continue _

Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format (csv, txt)

- Backups (incl. raw data!)

SugarSync

Dropbox

Google Drive

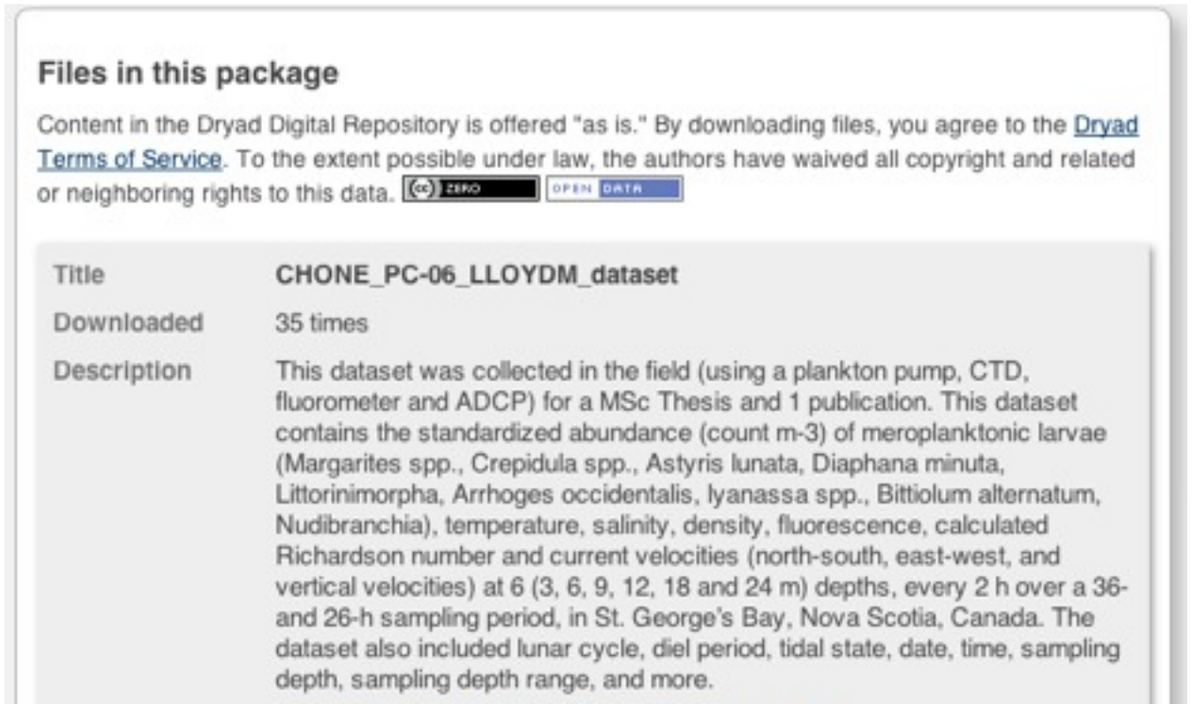
Publish your *data*

Data increasingly recognised as a first-class research product in its own right.

Can be published.

figshare
credit for **all** your research

DRYAD



Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the Dryad Terms of Service. To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.

Title	**CHONE_PC-06_LLOYDM_dataset**
Downloaded	35 times
Description	This dataset was collected in the field (using a plankton pump, CTD, fluorometer and ADCP) for a MSc Thesis and 1 publication. This dataset contains the standardized abundance (count m-3) of meroplanktonic larvae (Margarites spp., Crepidula spp., Astyris lunata, Diaphana minuta, Littorinimorpha, Arrhoges occidentalis, Iyanassa spp., Bittiolum alternatum, Nudibranchia), temperature, salinity, density, fluorescence, calculated Richardson number and current velocities (north-south, east-west, and vertical velocities) at 6 (3, 6, 9, 12, 18 and 24 m) depths, every 2 h over a 36- and 26-h sampling period, in St. George's Bay, Nova Scotia, Canada. The dataset also included lunar cycle, diel period, tidal state, date, time, sampling depth, sampling depth range, and more.

Data management plans

1. Project description
2. Method description
3. Metadata format and content
4. Access and use policies
5. Long-term storage
6. Budget

See example on Blackboard

Summary

- Why you should care.
- Collecting and organising.
- Quality control/assurance
- Storing.
- Using.

To do

- Work through *QualityControl.pdf*
- Read chapter in Gotelli and Ellison.
- Read about Data Management Plans (PDFs)
- Investigate Google Forms for collecting data (see video)
- Video about metadata

Links are on Blackboard