GG606

Data transformations and wrangling

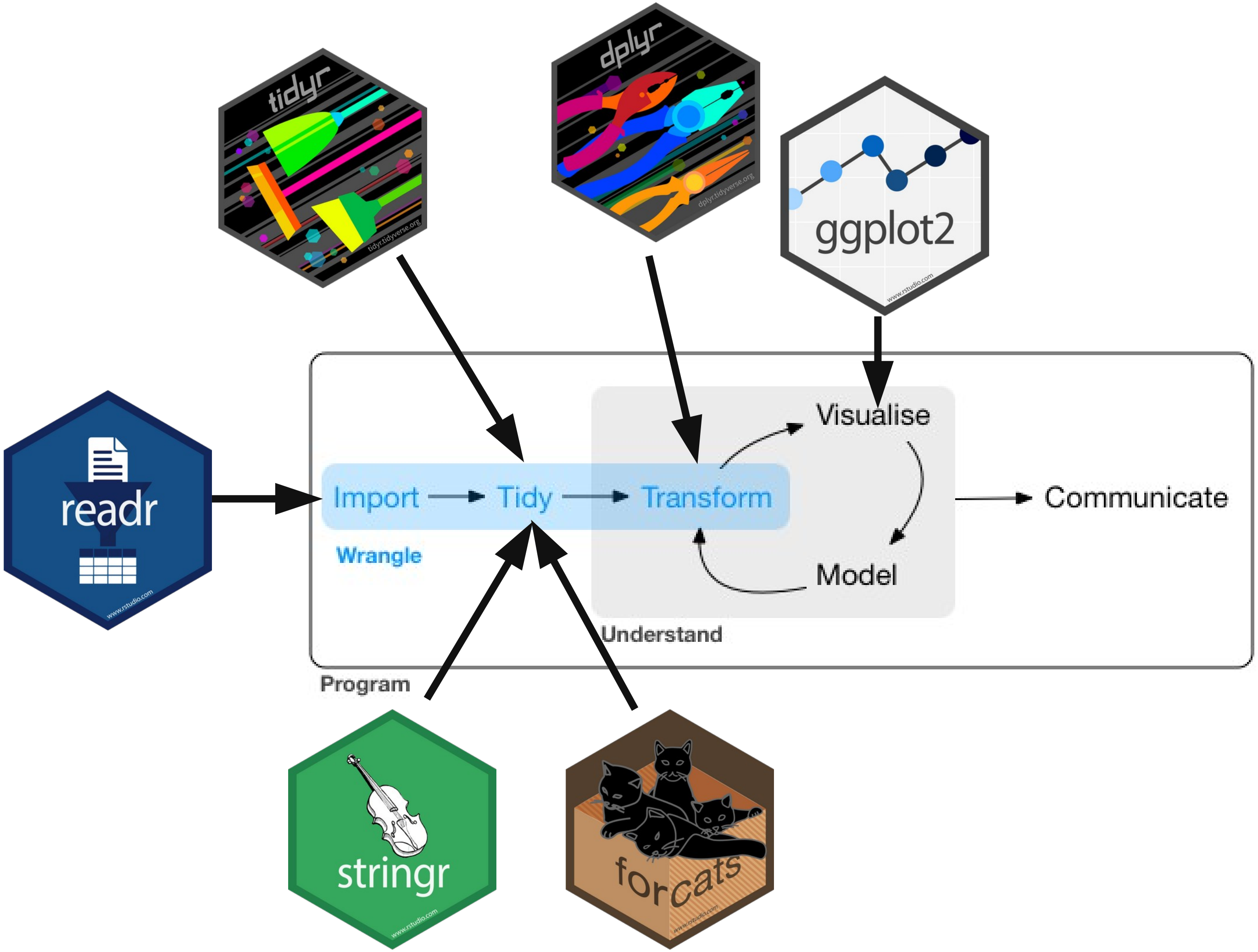
Homework

- folder structure for the workflow
  - we spoke about keeping raw data separate from processed data and keeping figures and/or tables together)
- use the `here` package and function
- R script to load data from pangaea.de or www.frdr-dfdr.ca
  - example, `read_csv(here("folder", "file"))`
- create and save a figure to an appropriate folder
  - hint, use the `ggsave` and `here` functions
- look at the metadata fields & make list of metadata required to deposit data

Homework

Transformations

- Organisation
- Reproducible
- Inputs & Outputs
- R-Script vs R-Markdown vs Function



Import → Tidy → Transform

Visualise

Model

→ Communicate

Wrangle

Understand

Program

Intro

- `vector`, `matrix`, `data.frame`, `tibble`
- import -> tidy -> save?
- So many data types

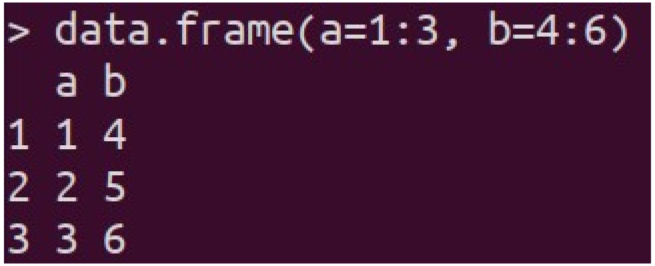
- MATLAB array: `[1 2 3 4]` (row vector)
- MATLAB vector: `[1 2 3; 4 5 6; 7 8 9]`
-
-
-
-

- MATLAB array: `[1 2 3 4]` (row vector)
- MATLAB vector: `[1 2 3; 4 5 6; 7 8 9]`
- R array: `array(1, 2, 3)` (⩾1 dimensions)
- R vector: `c(1, 2, 3)` (fixed size, same type)
-
-

- MATLAB array: `[1 2 3 4]` (row vector)
- MATLAB vector: `[1 2 3; 4 5 6; 7 8 9]`
- R array: `array(1, 2, 3)` (⩾1 dimensions)
- R vector: c(1, 2, 3) (fixed size, same type)
- R matrix: 2-dimensional vector
- R data.frame (table): (1 type per column, header names)
  `data.frame(a=1:3, b=4:6)`

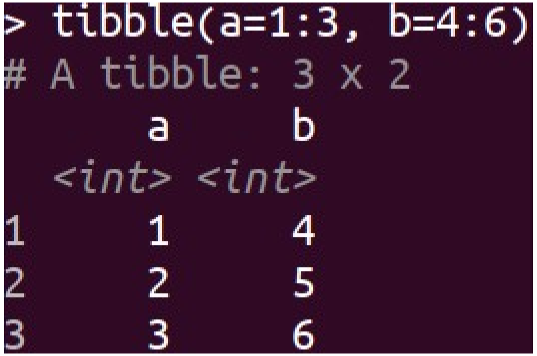
tibbles

- `data.frame(a=1:3, b=4:6)`



```
> data.frame(a=1:3, b=4:6)
  a b
1 1 4
2 2 5
3 3 6
```

- `tibble(a=1:3, b=4:6)`



```
> tibble(a=1:3, b=4:6)
# A tibble: 3 x 2
      a     b
  <int> <int>
1     1     4
2     2     5
3     3     6
```

tibbles

- Prints 10 rows
- Column type
- (default options can be changed)
- Strict(er) behaviour can be useful
- Tools for identifying data types
- Can easily convert:

tibbles

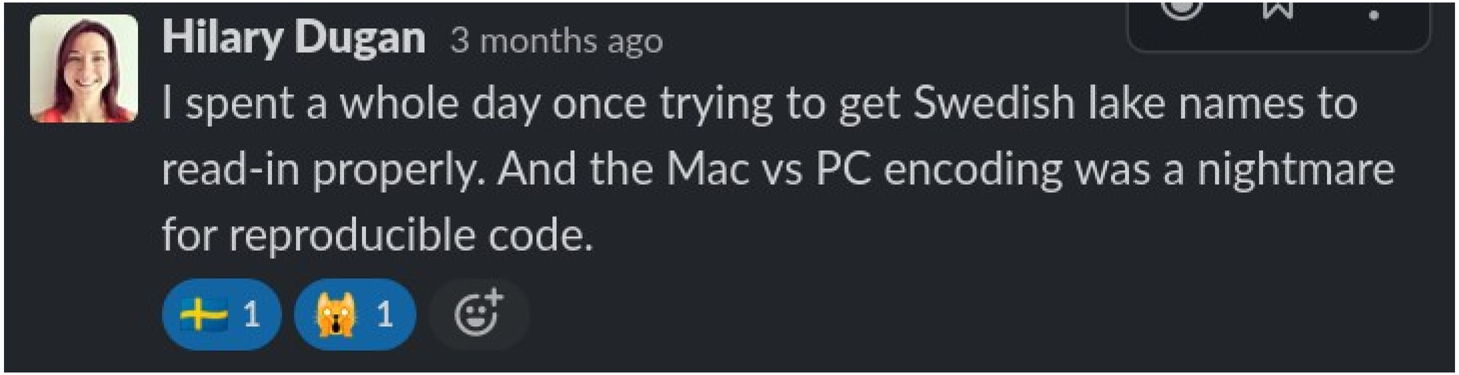
- Can easily convert:

```
x ← tibble(a=1:3, b=4:6)
y ← as.data.frame(x)
y
```

- Chapter 10 for more details & exercises: https://r4ds.had.co.nz/tibbles.html

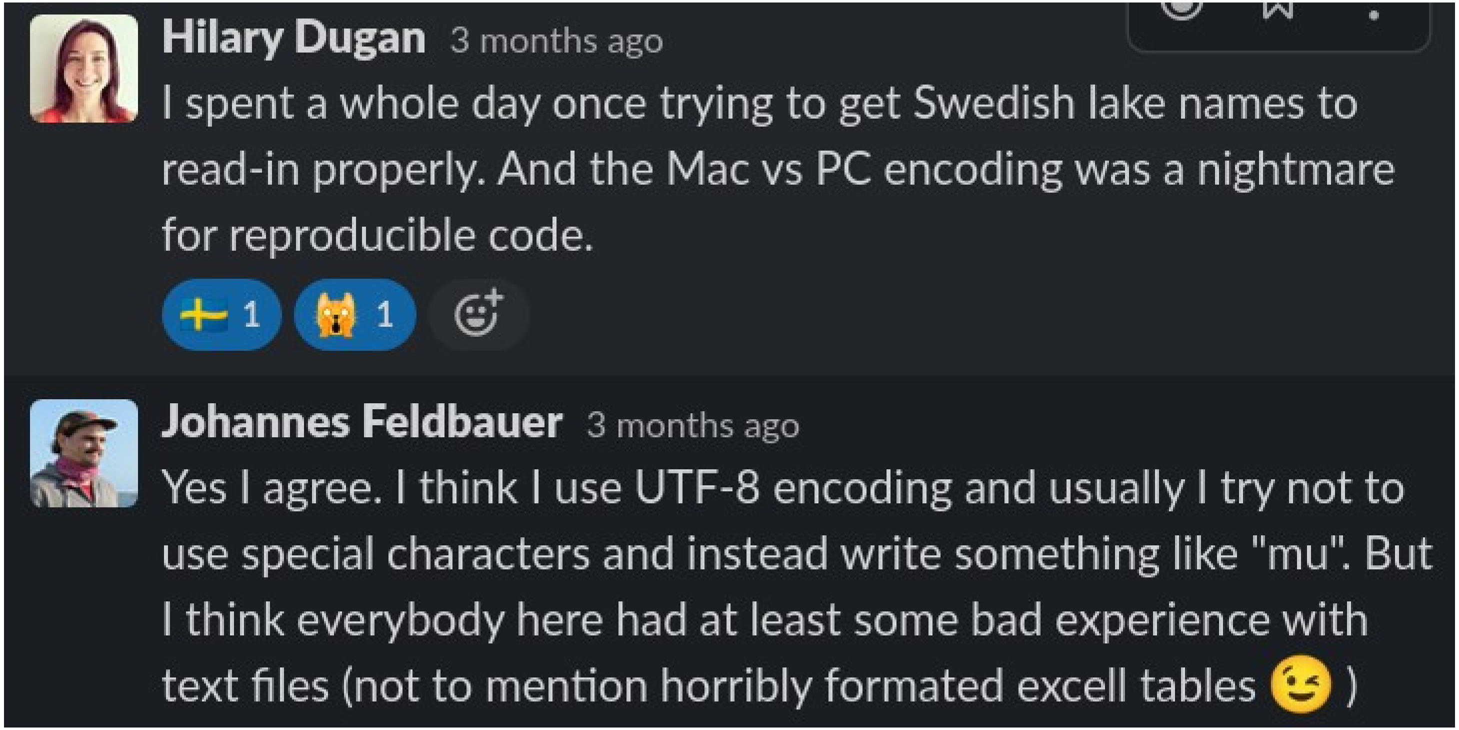
Data import

- `read_csv` vs `read.csv`
- Your examples
- A few examples



Hilary Dugan 3 months ago

I spent a whole day once trying to get Swedish lake names to read-in properly. And the Mac vs PC encoding was a nightmare for reproducible code.



Hilary Dugan 3 months ago

I spent a whole day once trying to get Swedish lake names to read-in properly. And the Mac vs PC encoding was a nightmare for reproducible code.

🇸🇪 1 🙀 1

Johannes Feldbauer 3 months ago

Yes I agree. I think I use UTF-8 encoding and usually I try not to use special characters and instead write something like "mu". But I think everybody here had at least some bad experience with text files (not to mention horribly formated excell tables 😉)

Parsing

- `parse_` functions
  - `parse_logical()` `parse_integer()`
  - `parse_double()` `parse_number()`
  - `parse_character()`
  - `parse_factor()`
  - `parse_datetime()` `parse_date()` `parse_time()`
  - `guess_parser()`

Parsing Numbers

- `parse_double("1.23")`
- `parse_double("1,23", locale = locale(decimal_mark = ","))`
-
-

Parsing Numbers

- `parse_double("1.23")`
- `parse_double("1,23", locale = locale(decimal_mark = ","))`
- `parse_number("$100")`
- `parse_number("123.456.789", locale = locale(grouping_mark = "."))`

Parsing Dates

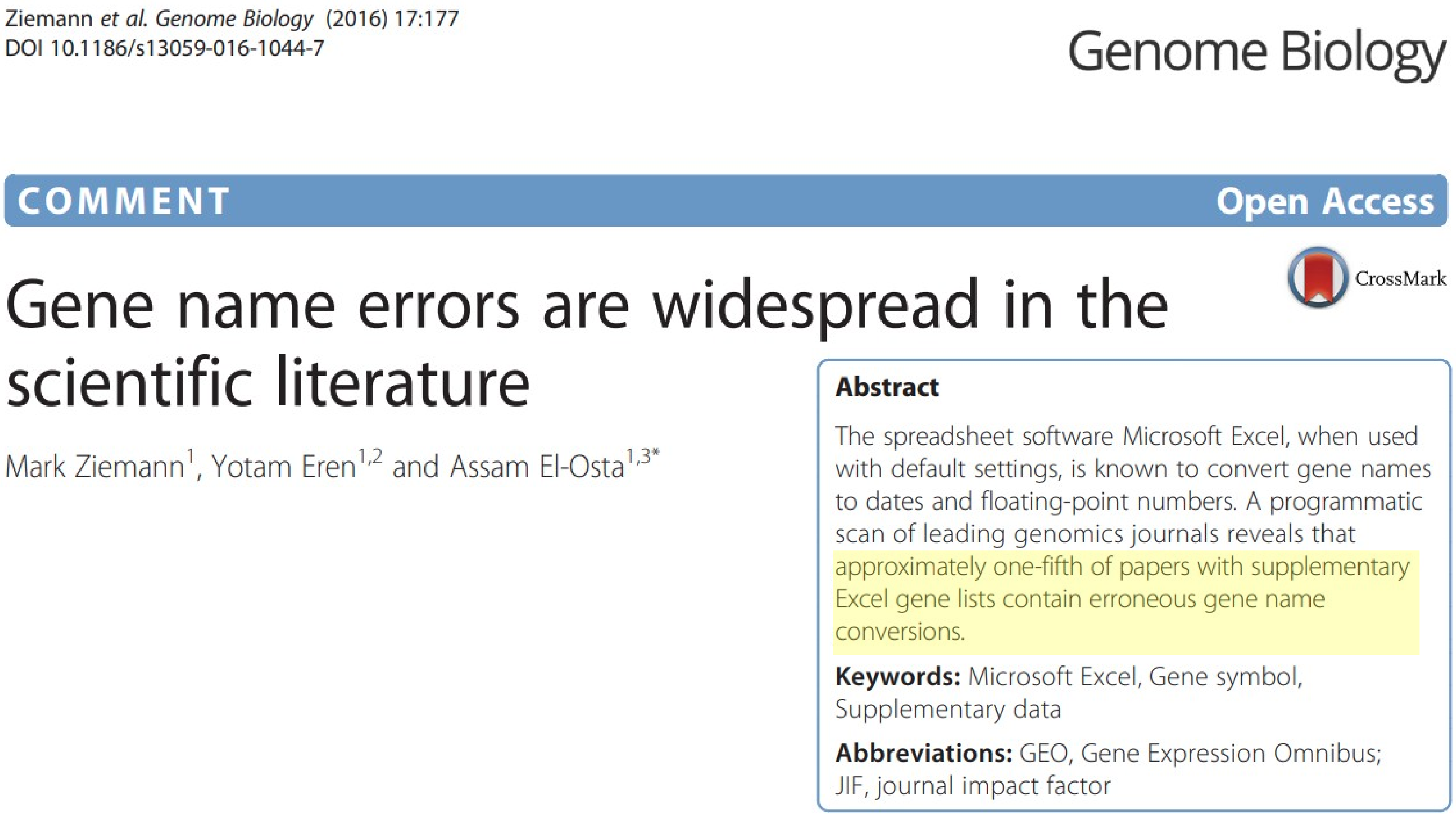
- `parse_datetime("2021-01-01T0001")`
- ISO8601 by default
- What if no time?
- `hms` & `lubridate` packages

Parsing Dates

- `parse_date("01/02/15", "%m/%d/%y")`
- `parse_date("01/02/15", "%d/%m/%y")`
- `parse_date("01/02/15", "%y/%m/%d")`
- Can be infuriating

Parsing Dates

- `parse_date("1 janvier 2015", "%d %B %Y", locale = locale("fr"))`
-
- Can be infuriating
- Also time zones



Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

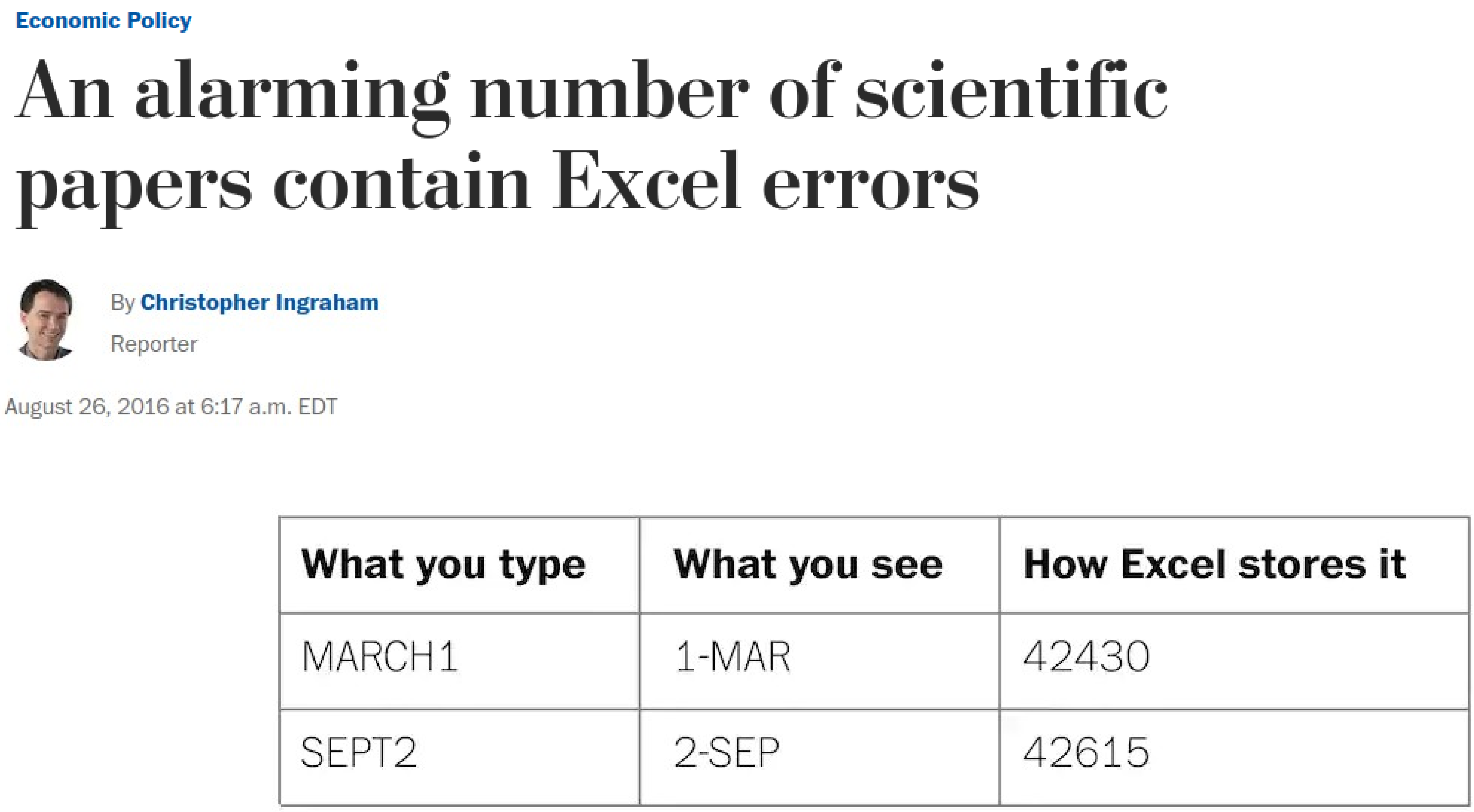
Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor



Economic Policy

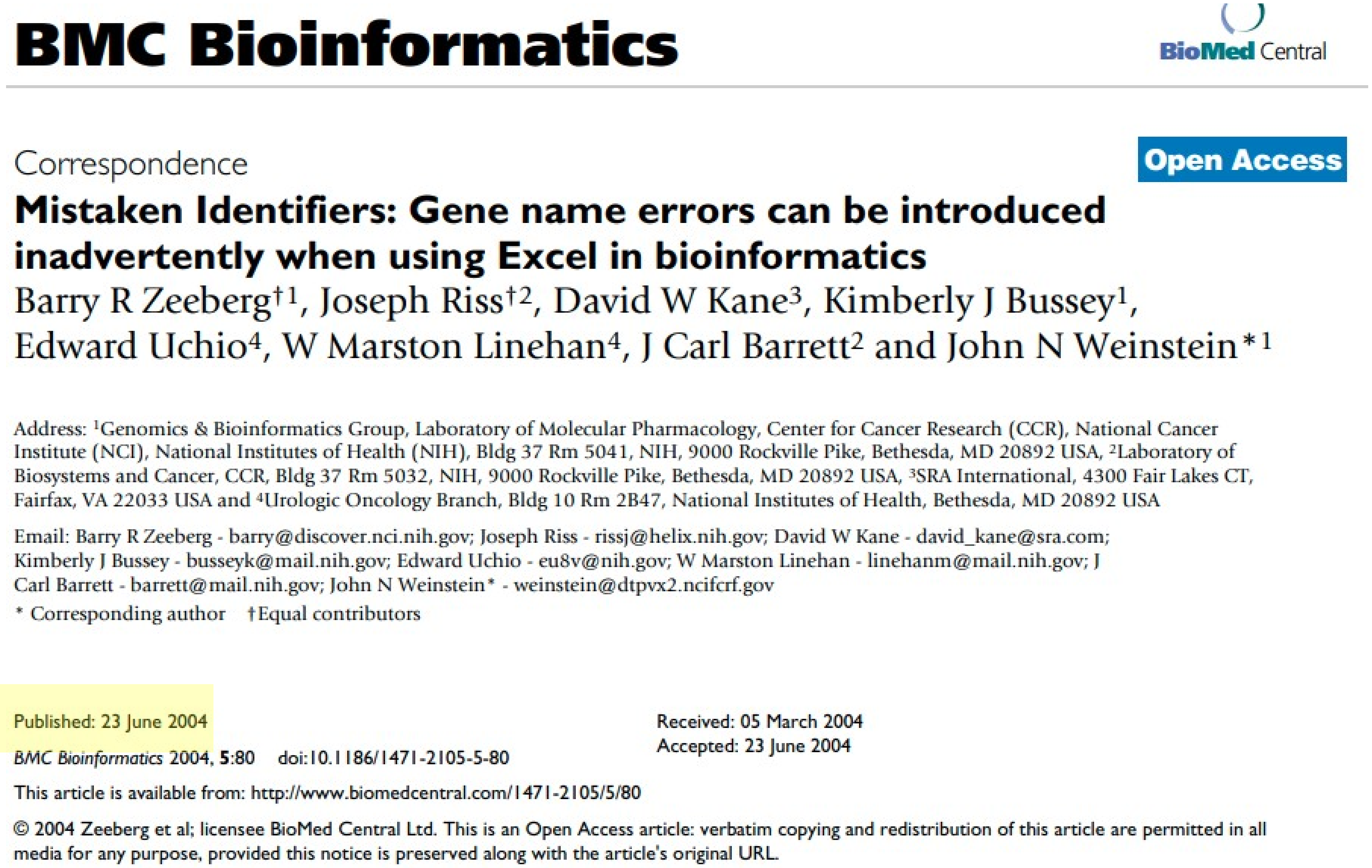
An alarming number of scientific papers contain Excel errors

By **Christopher Ingraham**

Reporter

August 26, 2016 at 6:17 a.m. EDT

What you type	What you see	How Excel stores it
MARCH1	1-MAR	42430
SEPT2	2-SEP	42615



BMC Bioinformatics

BioMed Central

Correspondence

Open Access

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg†[1], Joseph Riss†[2], David W Kane[3], Kimberly J Bussey[1], Edward Uchio[4], W Marston Linehan[4], J Carl Barrett[2] and John N Weinstein*[1]

Address: [1]Genomics & Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research (CCR), National Cancer Institute (NCI), National Institutes of Health (NIH), Bldg 37 Rm 5041, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA, [2]Laboratory of Biosystems and Cancer, CCR, Bldg 37 Rm 5032, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA, [3]SRA International, 4300 Fair Lakes CT, Fairfax, VA 22033 USA and [4]Urologic Oncology Branch, Bldg 10 Rm 2B47, National Institutes of Health, Bethesda, MD 20892 USA

Email: Barry R Zeeberg - barry@discover.nci.nih.gov; Joseph Riss - rissj@helix.nih.gov; David W Kane - david_kane@sra.com; Kimberly J Bussey - busseyk@mail.nih.gov; Edward Uchio - eu8v@nih.gov; W Marston Linehan - linehanm@mail.nih.gov; J Carl Barrett - barrett@mail.nih.gov; John N Weinstein* - weinstein@dtpvx2.ncifcrf.gov

* Corresponding author †Equal contributors

Published: 23 June 2004

BMC Bioinformatics 2004, **5**:80 doi:10.1186/1471-2105-5-80

Received: 05 March 2004
Accepted: 23 June 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/80

BMC Bioinf

BioMed Central

Open Access

Correspondence

Mistaken Identifiers … duced
inadvertently when

Barry R Zeeberg†[1], Jose … Bussey[1],
Edward Uchio[4], W Ma … hn N Weinstein*[1]

Address: [1]Genomics & Bioinformatics Gro … h (CCR), National Cancer
Institute (NCI), National Institutes of Hea … 20892 USA, [2]Laboratory of
Biosystems and Cancer, CCR, Bldg 37 Rm … rnational, 4300 Fair Lakes CT,
Fairfax, VA 22033 USA and [4]Urologic Onc … da, MD 20892 USA

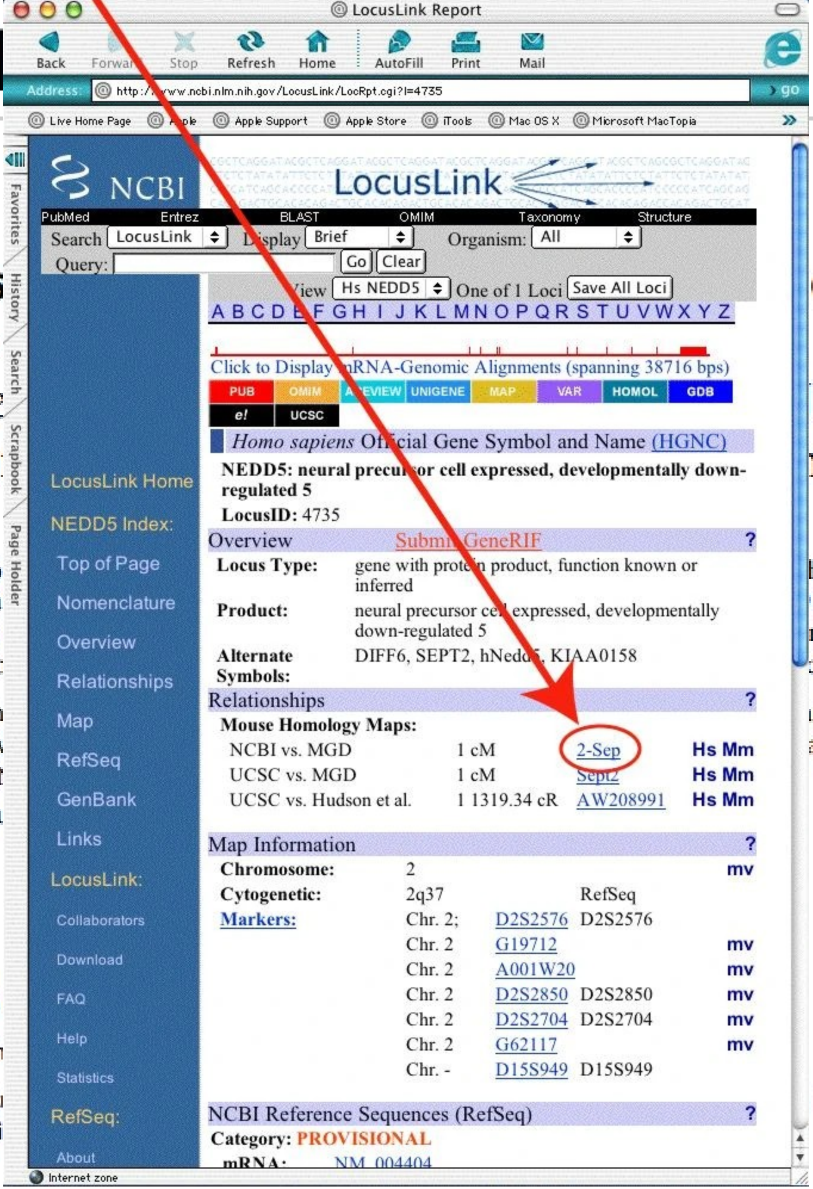
Email: Barry R Zeeberg - barry@discover.n … ne@sra.com;
Kimberly J Bussey - busseyk@mail.nih.gov … ail.nih.gov; J
Carl Barrett - barrett@mail.nih.gov; John I

* Corresponding author †Equal contribu

Published: 23 June 2004

BMC Bioinformatics 2004, **5**:80 doi:10.1186/

This article is available from: http://www.bion



http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=4735

Search: LocusLink Display: Brief Organism: All

Query:

View: Hs NEDD5 One of 1 Loci

Click to Display mRNA-Genomic Alignments (spanning 38716 bps)

Homo sapiens Official Gene Symbol and Name (HGNC)

NEDD5: neural precursor cell expressed, developmentally down-regulated 5

LocusID: 4735

Overview Submit GeneRIF ?

Locus Type:	gene with protein product, function known or inferred
Product:	neural precursor cell expressed, developmentally down-regulated 5
Alternate Symbols:	DIFF6, SEPT2, hNedd5, KIAA0158

Relationships ?

Mouse Homology Maps:

NCBI vs. MGD	1 cM	2-Sep	Hs Mm
UCSC vs. MGD	1 cM	Sept2	Hs Mm
UCSC vs. Hudson et al.	1 1319.34 cR	AW208991	Hs Mm

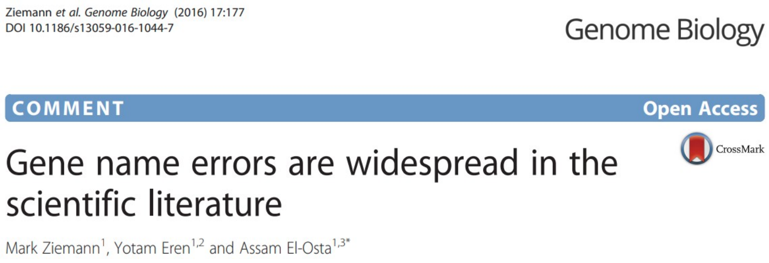
Map Information ?

Chromosome:	2			mv
Cytogenetic:	2q37		RefSeq	
Markers:	Chr. 2;	D2S2576	D2S2576	
	Chr. 2	G19712		mv
	Chr. 2	A001W20		mv
	Chr. 2	D2S2850	D2S2850	mv
	Chr. 2	D2S2704	D2S2704	mv
	Chr. 2	G62117		mv
	Chr. -	D15S949	D15S949	

NCBI Reference Sequences (RefSeq) ?

Category: PROVISIONAL

mRNA: NM_004404



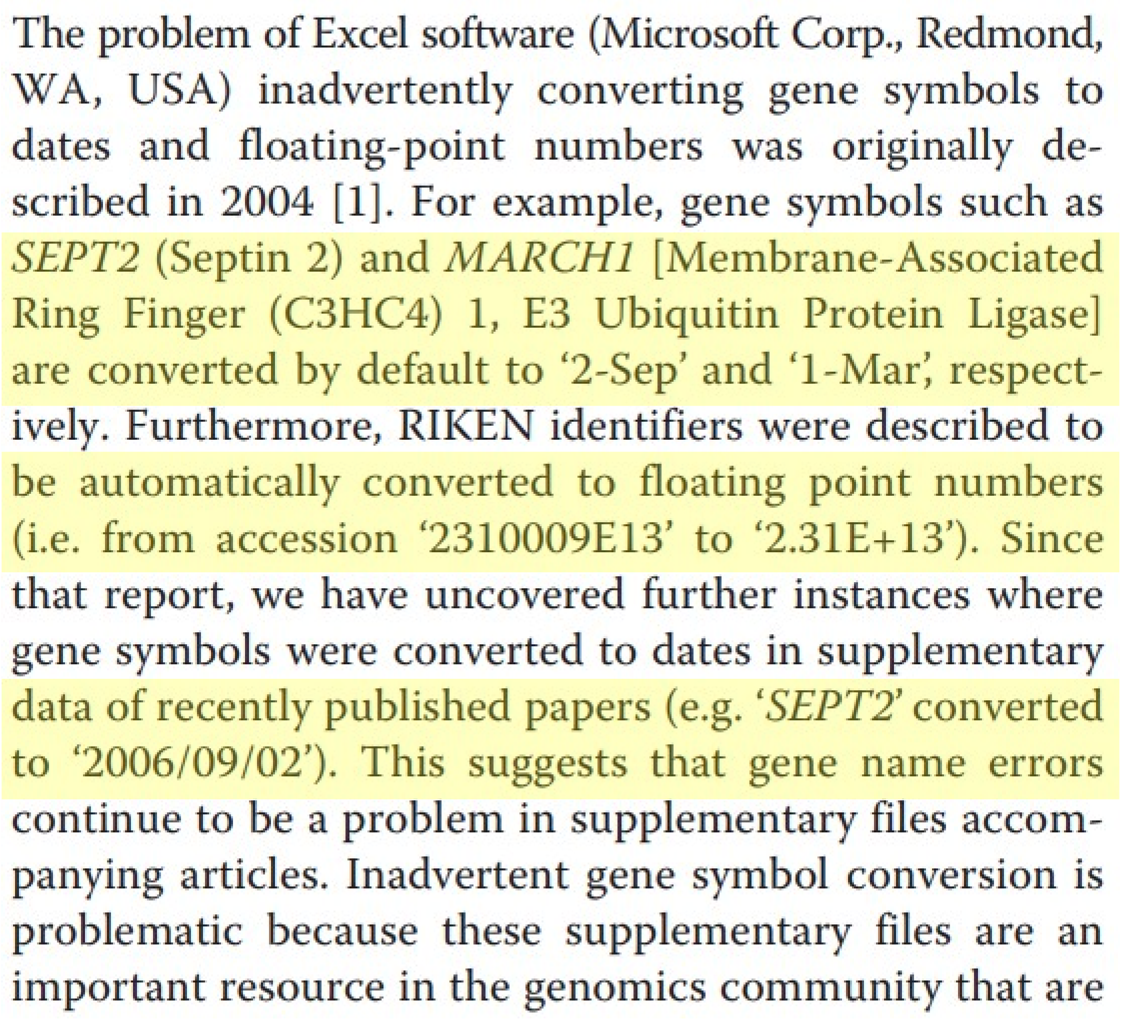
Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

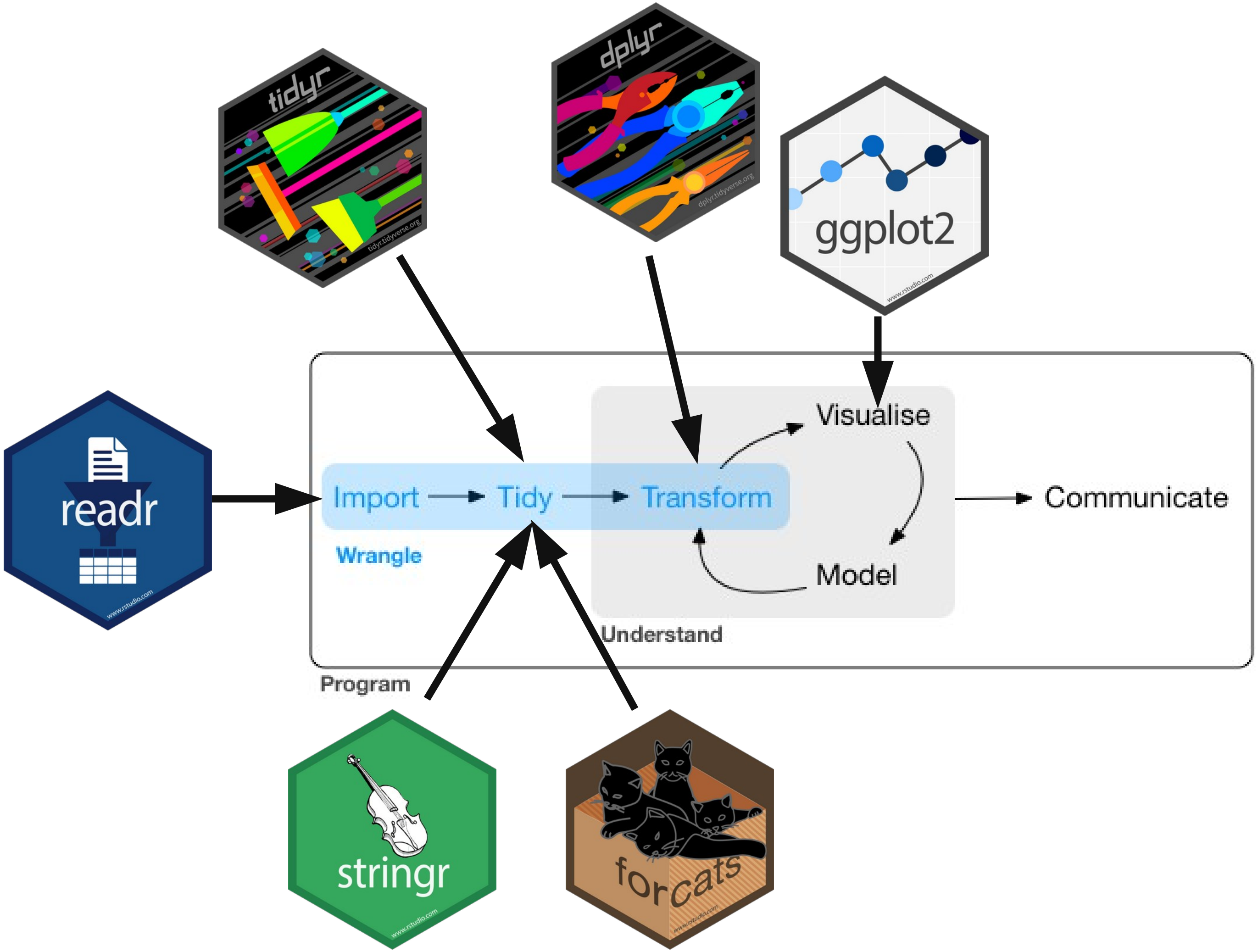
Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]



The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since that report, we have uncovered further instances where gene symbols were converted to dates in supplementary data of recently published papers (e.g. '*SEPT2*' converted to '2006/09/02'). This suggests that gene name errors continue to be a problem in supplementary files accompanying articles. Inadvertent gene symbol conversion is problematic because these supplementary files are an important resource in the genomics community that are

Strategies

- readr uses heuristic over first 1000 rows
- guess_parser()
- homework


tidyr
dplyr
ggplot2
readr
Import
Tidy
Transform
Visualise
Model
Communicate
Wrangle
Understand
Program
stringr
forcats

Save

- Import
- Tidy
- Save `write_csv()` or `write_rds()`
- How would you organise multiple R scripts to:
  import, tidy, save &
  load, continue

12 Points

- Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, https://doi.org/10.1080/00031305.2017.1375989

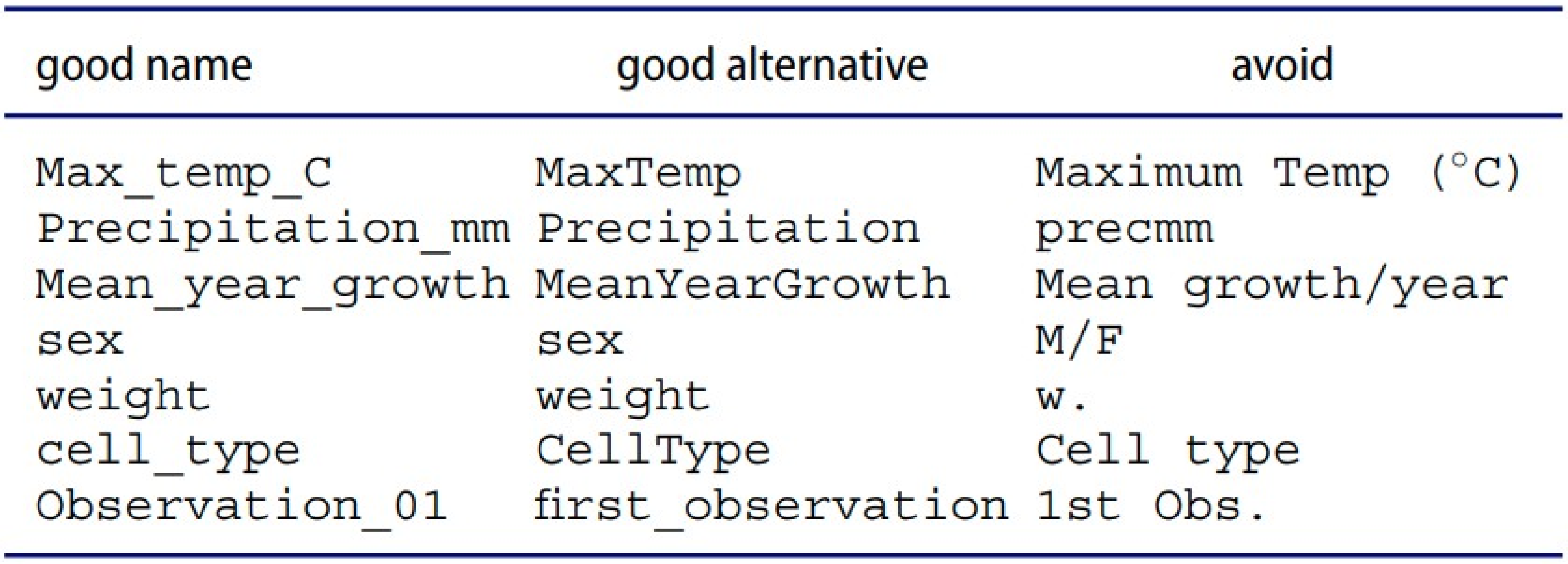
Be Consistent

- Names
- Codes
- Identifiers
- Extra spaces

Good Names

- Avoid spaces
- No special characters or symbols

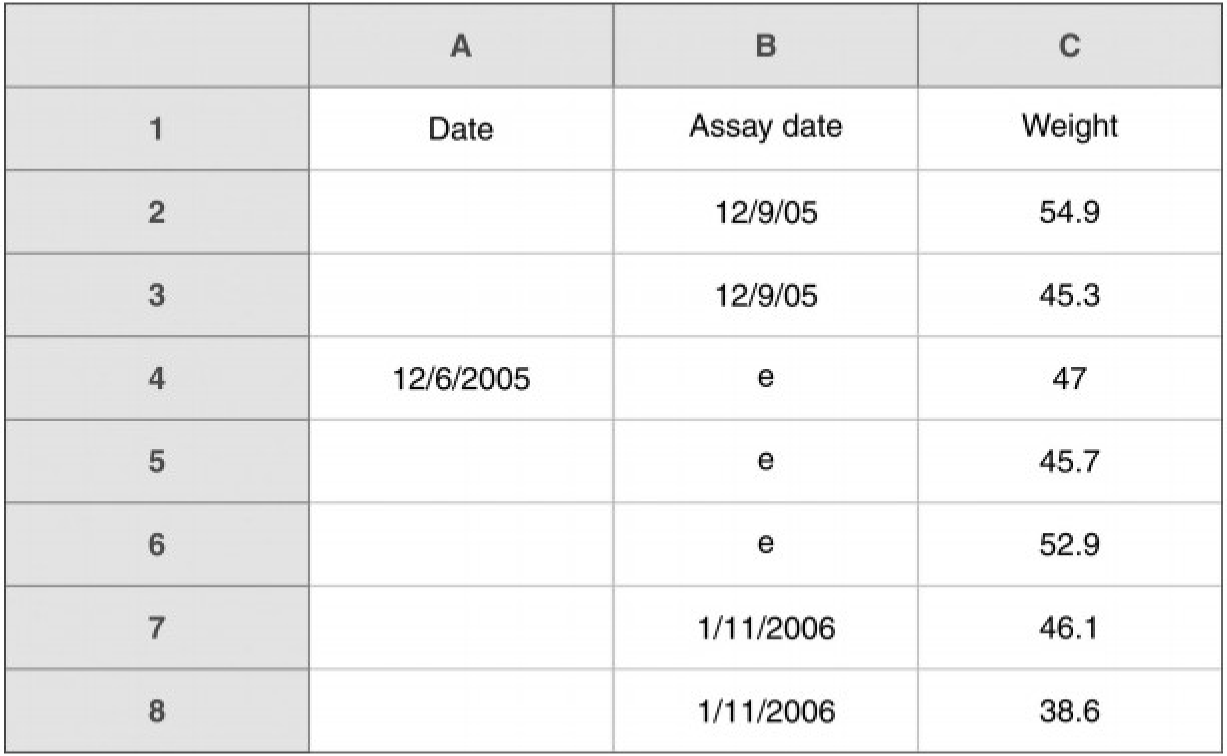
Table 1. Examples of good and bad variable names.



good name	good alternative	avoid
`Max_temp_C`	`MaxTemp`	`Maximum Temp (°C)`
`Precipitation_mm`	`Precipitation`	`precmm`
`Mean_year_growth`	`MeanYearGrowth`	`Mean growth/year`
`sex`	`sex`	`M/F`
`weight`	`weight`	`w.`
`cell_type`	`CellType`	`Cell type`
`Observation_01`	`first_observation`	`1st Obs.`

Dates

- YYYY-MM-DD ISO8601



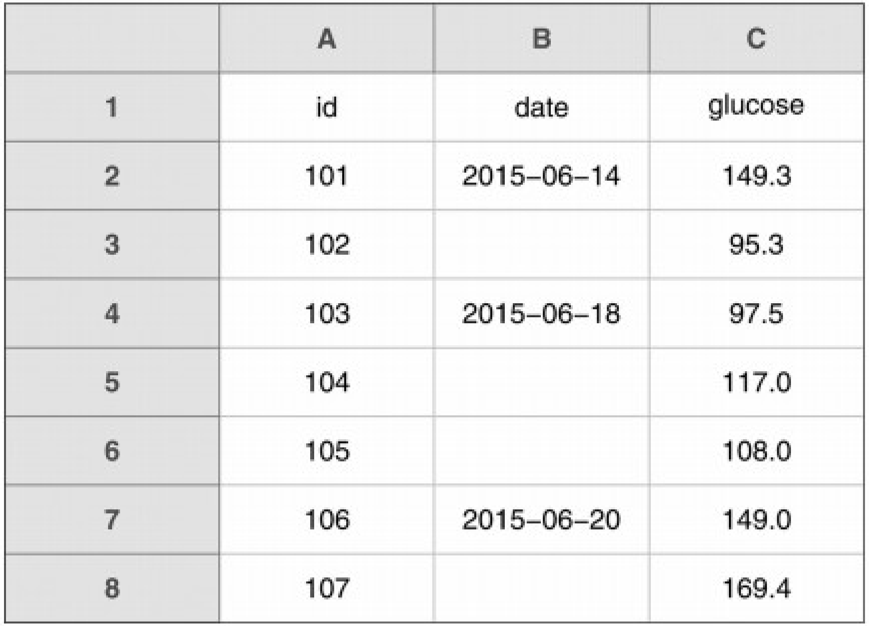
	A	B	C
1	Date	Assay date	Weight
2		12/9/05	54.9
3		12/9/05	45.3
4	12/6/2005	e	47
5		e	45.7
6		e	52.9
7		1/11/2006	46.1
8		1/11/2006	38.6

Figure 1. A spreadsheet with inconsistent date formats. This spreadsheet does not adhere to our recommendations for consistency of date format.

No Empty Cells

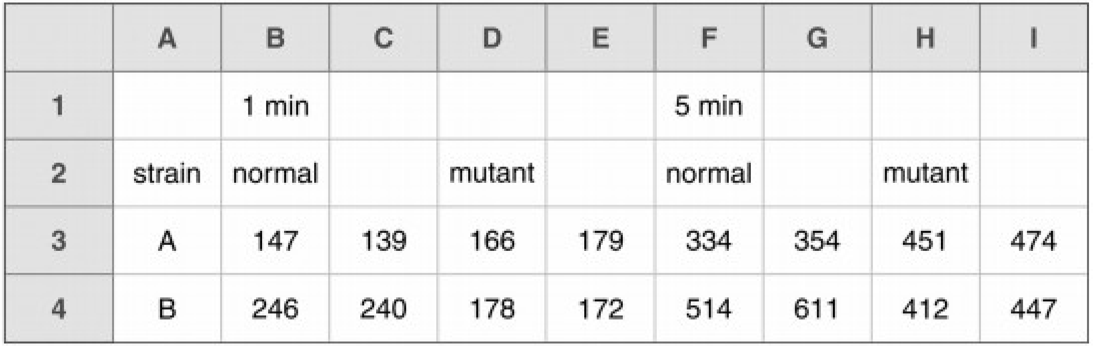
- Empty vs NA

A



	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4

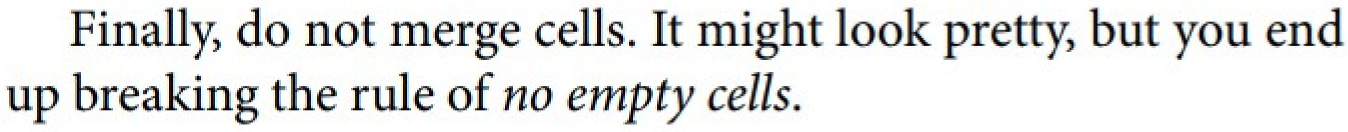
B



	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

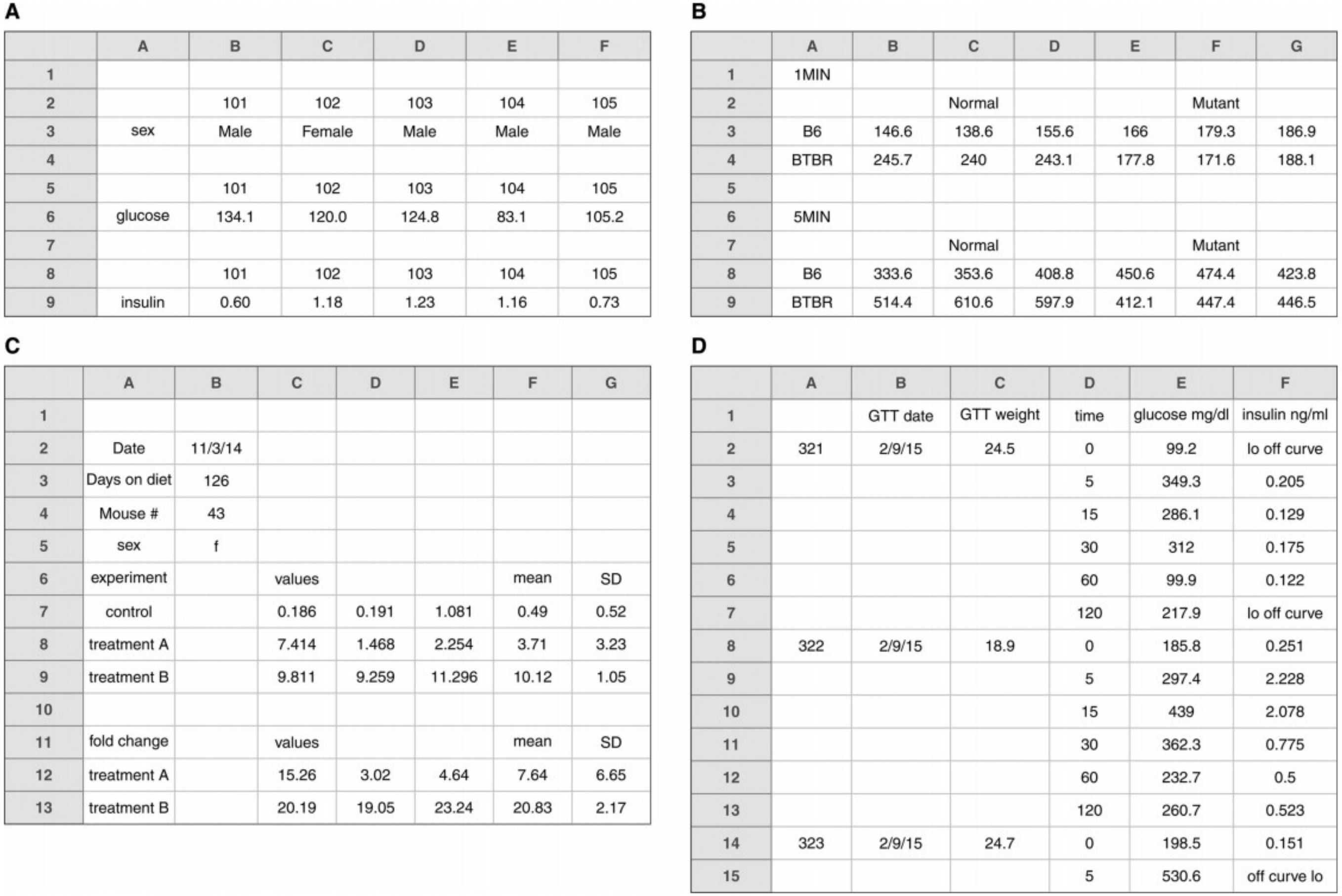
One Thing Per Cell

- a place for everything and everything in its place



Finally, do not merge cells. It might look pretty, but you end up breaking the rule of *no empty cells*.

Rectangle



A

	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Male
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73

B

	A	B	C	D	E	F	G
1	1MIN						
2			Normal			Mutant	
3	B6	146.6	138.6	155.6	166	179.3	186.9
4	BTBR	245.7	240	243.1	177.8	171.6	188.1
5							
6	5MIN						
7			Normal			Mutant	
8	B6	333.6	353.6	408.8	450.6	474.4	423.8
9	BTBR	514.4	610.6	597.9	412.1	447.4	446.5

C

	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

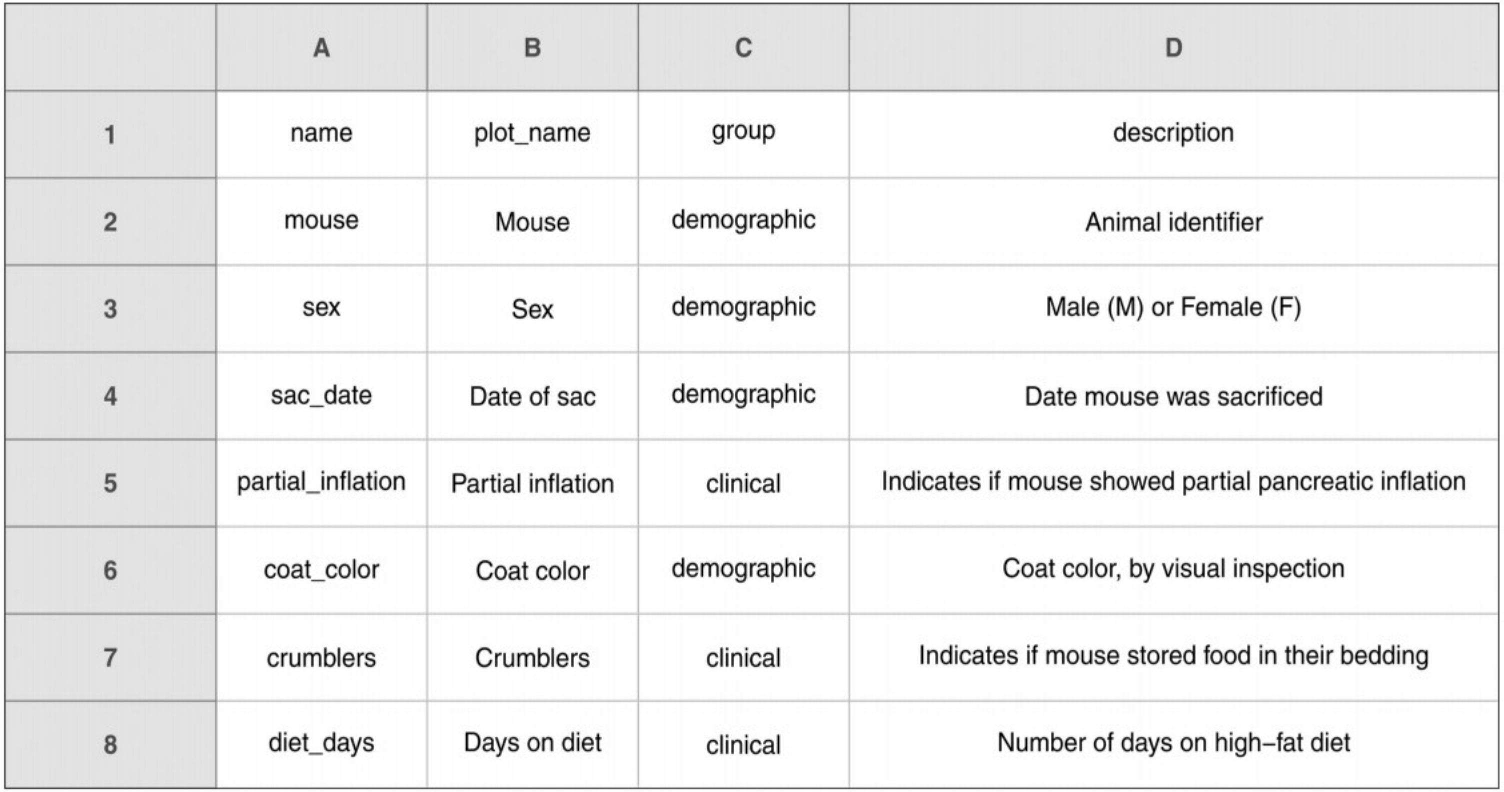
D

	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	286.1	0.129
5				30	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15				5	530.6	off curve lo

Figure 5. Examples of spreadsheets with nonrectangular layouts. These layouts are likely to cause problems in analysis.

Data Dictionary

- Metadata

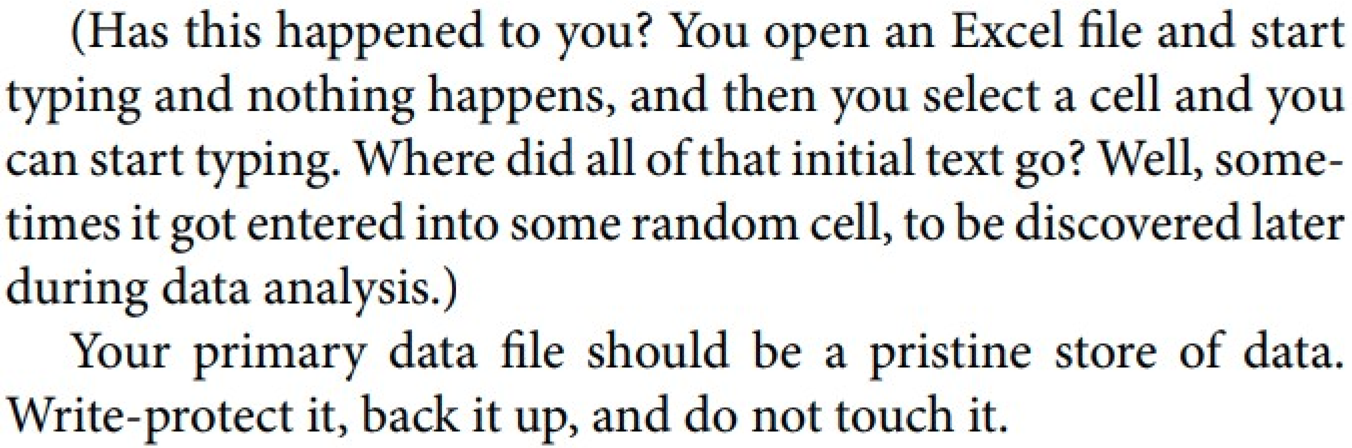


	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high–fat diet

Figure 9. An example data dictionary.

No Calcs in Data Files

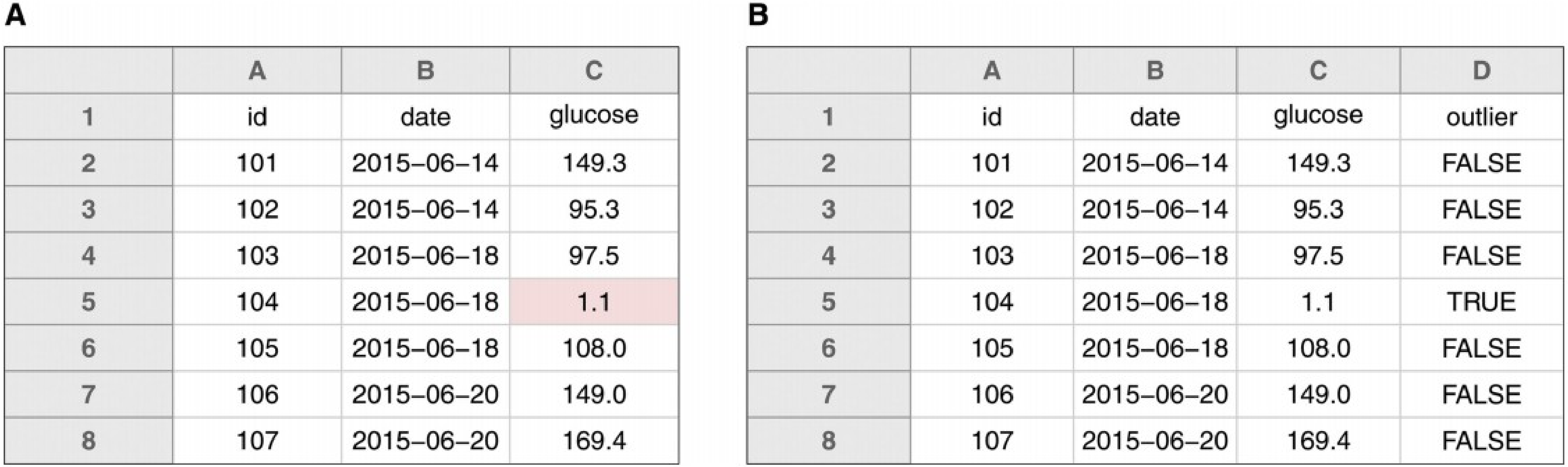
- Really
- This will be difficult for some people



(Has this happened to you? You open an Excel file and start typing and nothing happens, and then you select a cell and you can start typing. Where did all of that initial text go? Well, sometimes it got entered into some random cell, to be discovered later during data analysis.)

Your primary data file should be a pristine store of data. Write-protect it, back it up, and do not touch it.

Colour & Highlights Are Not Data



A

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102	2015–06–14	95.3
4	103	2015–06–18	97.5
5	104	2015–06–18	1.1
6	105	2015–06–18	108.0
7	106	2015–06–20	149.0
8	107	2015–06–20	169.4

B

	A	B	C	D
1	id	date	glucose	outlier
2	101	2015–06–14	149.3	FALSE
3	102	2015–06–14	95.3	FALSE
4	103	2015–06–18	97.5	FALSE
5	104	2015–06–18	1.1	TRUE
6	105	2015–06–18	108.0	FALSE
7	106	2015–06–20	149.0	FALSE
8	107	2015–06–20	169.4	FALSE

Figure 10. Highlighting in spreadsheets. (a) A potential outlier indicated by highlighting the cell. (b) The preferred method for indicating outliers, via an additional column.

Backups

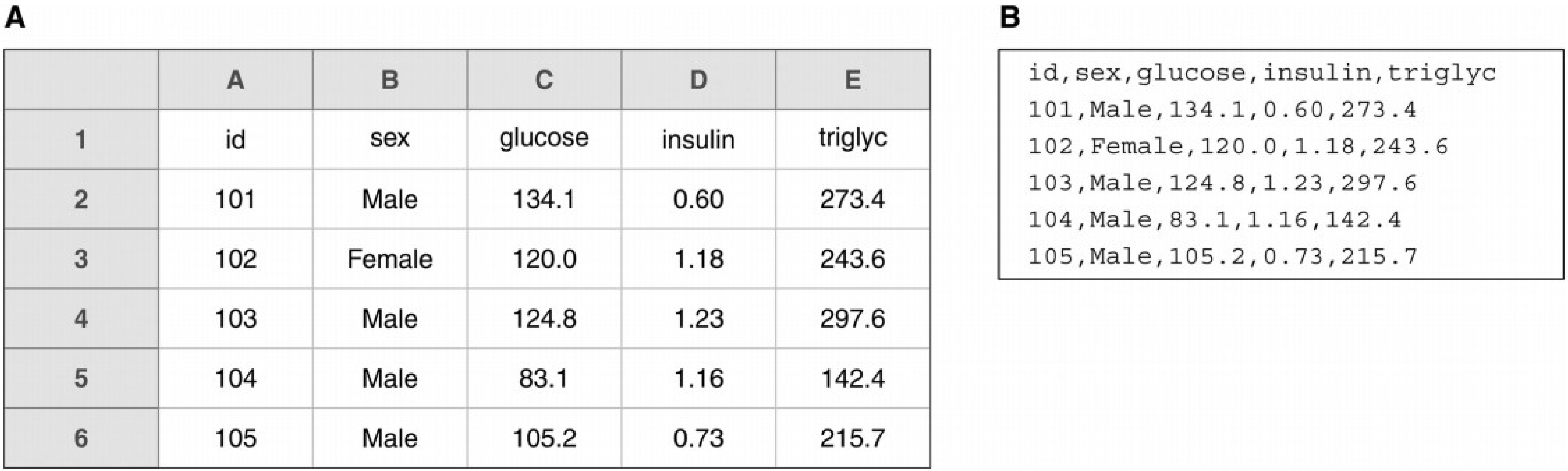
- March 31 is World Backup Day
- http://www.worldbackupday.com

Data Validation

- Feature in spreadsheets
  to help with data entry

Save Plain Text

- Good test



	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

```
id,sex,glucose,insulin,triglyc
101,Male,134.1,0.60,273.4
102,Female,120.0,1.18,243.6
103,Male,124.8,1.23,297.6
104,Male,83.1,1.16,142.4
105,Male,105.2,0.73,215.7
```

Figure 11. (a) An example spreadsheet. (b) The same data as a plain text file in CSV format.

Homework

- Pick a year: https://doi.org/10.5683/SP3/OUWVZ3 (physical, chemical, biological)
- Use this: https://doi.org/10.5683/SP2/TNYTQL “NW-20-C2-Chronology-Dspec50-2019-with-self-attenuation-SimpleView.xlsx” “NW-50-Chronology-Dspec649-2019-withdensity-SIMPLEVIEW with graphs v3.tab”