Data Science 1

STAT/CS 287

Jim Bagrow, UVM Dept of Math and Statistics

LECTURE 08

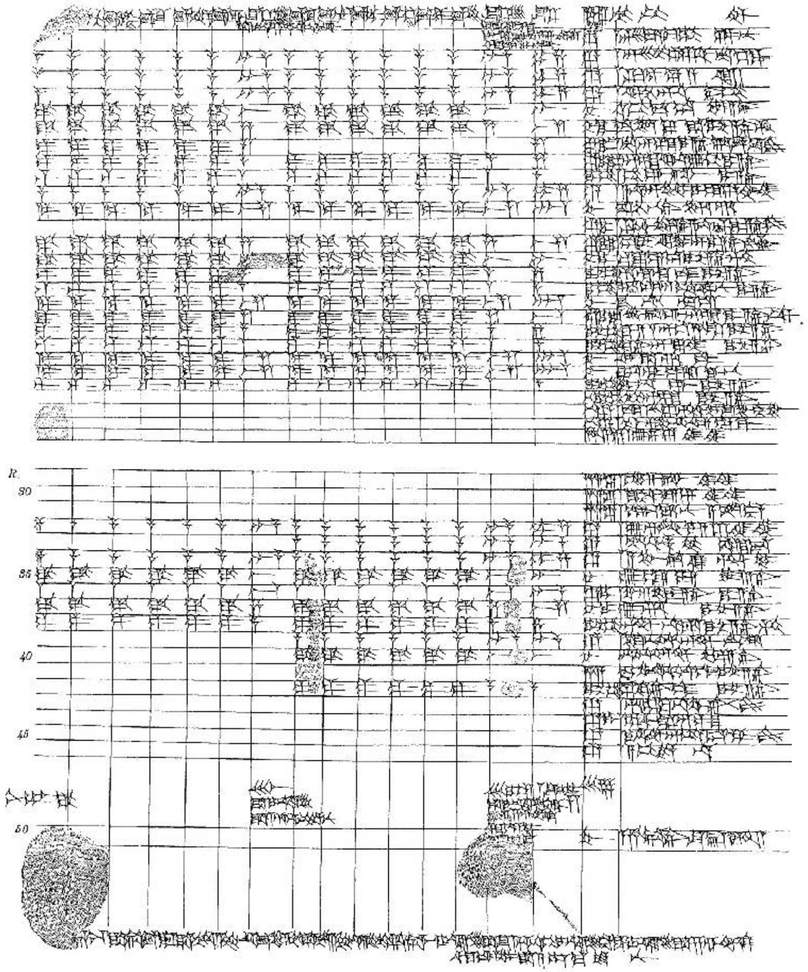
Spreadsheets considered harmful

Long history of data in tables

Tables and tabular formatting in Sumer, Babylonia, and Assyria, 2500 BCE–50 CE

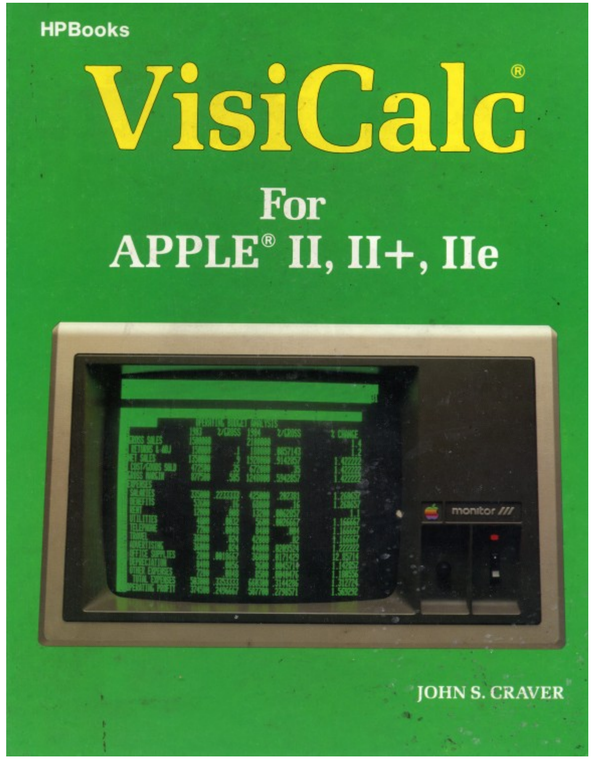
ELEANOR ROBSON

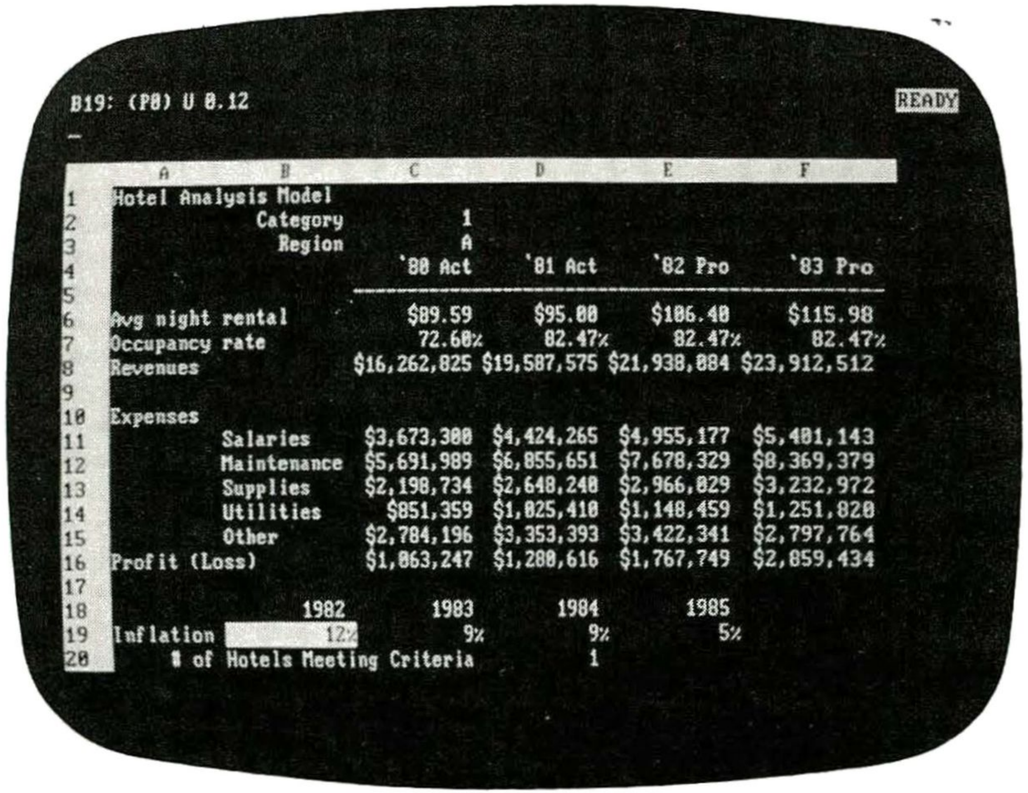
Fig. 1.1 Month-by-month wage account for the temple of Enlil at Nippur, for the year 1295 BCE. This tablet, recording the monthly salaries of forty-five temple personnel, exhibits most of the



The History of Mathematical Tables: From Sumer to Spreadsheets, 2003

Spreadsheets are awesome

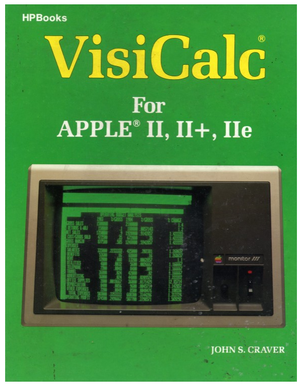


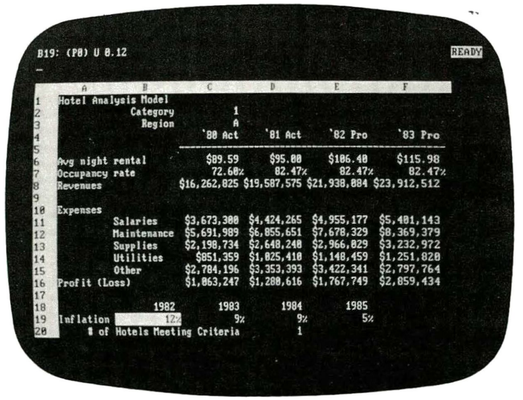


Spreadsheets are awesome

The *original* killer app

VisiCalc launched the
personal computer revolution





Graphical interface is intuitive, interactive

Fast learning curve

Great for data entry

I use spreadsheets *all the time*:

- Task lists
- Scheduling and planning
- Budgets
- Grading

Spreadsheets are bad for (rigorous) data analysis

Anecdotes

Business

Excel snafu costs firm $24m

Some cleric, some error

By Drew Cullen 19 Jun 2003 at 09:27 SHARE ▼

A simple spreadsheet error cost a firm a whopping US$24m.

The mistake led to TransAlta, a big Canadian power generator, buying more US power transmission hedging contracts in May at higher prices than it should have.

In a conference call, chief executive Steve Snyder said the snafu was "literally a cut-and-paste error in an Excel spreadsheet that we did not detect when we did our final sorting and ranking bids prior to submission," Reuters reports.

theregister.co.uk

How The London Whale Debacle Is Partly The Result Of An Error Using Excel

Linette Lopez Feb. 12, 2013, 2:04 PM

It's all in JP Morgan's 129 page report on the $6 billion trading loss. In an appendix on page 127, the report talks about how one London-based quant was working on a new VaR (Value at Risk) model for the Chief Investment Office.

During the review process, additional operational issues became apparent. For example, **the model operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another...** in a January 23, 2012 e-mail to the modeler, the trader to whom the modeler reported wrote that he should "keep the pressure on our friends in Model Validation and [Quantitative Research]." **There is some evidence the Model Review Group accelerated its review as a result of this pressure, and in so doing it may have been more willing to overlook the operational flaws apparent during the approval process.**

businessinsider.com

Anecdotes

Business

Excel snafu costs firm $24m

Some cleric, some error

By Drew Cullen 19 Jun 2003 at 09:27 SHARE ▼

A simple spreadsheet error cost a firm a whopping US$24m.

The mistake led to TransAlta, a big Canadian power generator, buying more US power transmission hedging contracts in May at higher prices than it should have.

In a conference call, chief executive Steve Snyder said the snafu was "literally a cut-and-paste error in an Excel spreadsheet that we did not detect when we did our final sorting and ranking bids prior to submission," Reuters reports.

theregister.co.uk

How The London Whale Debacle Is Partly The Result Of An Error Using Excel

Linette Lopez Feb. 12, 2013, 2:04 PM

It's all in JP Morgan's 129 page report on the $6 billion trading loss. In an appendix on page 127, the report talks about how one London-based quant was working on a new VaR (Value at Risk) model for the Chief Investment Office.

During the review process, additional operational issues became apparent. For example, **the model operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another...** in a January 23, 2012 e-mail to the modeler, the trader to whom the modeler reported wrote that he should "keep the pressure on our friends in Model Validation and [Quantitative Research]." **There is some evidence the Model Review Group accelerated its review as a result of this pressure, and in so doing it may have been more willing to overlook the operational flaws apparent during the approval process.**

businessinsider.com

Anecdotes

Fixing this Excel error transforms high-debt countries from recession to growth

By Tim Fernholz • April 16, 2013

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

"It looks like the most frequently cited justification for fiscal austerity in the aftermath of the global financial crisis is based on a boneheaded Microsoft Excel error."

qz.com

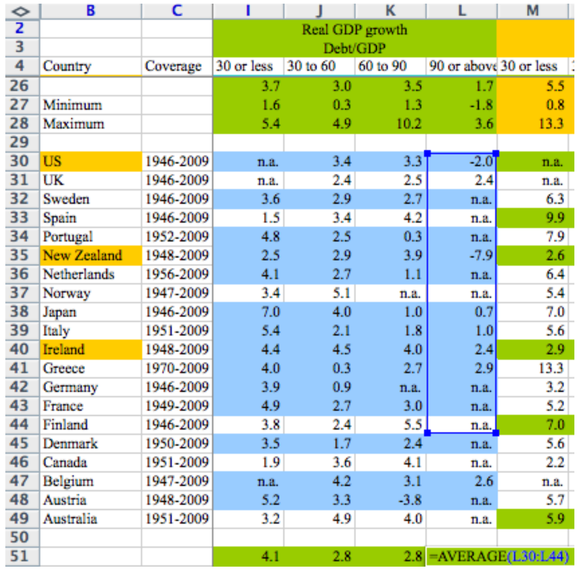
"Advocates of [reducing government debt] often cited an exhaustive survey by economists Kenneth Rogoff and Carmen Reinhart (pdf), who found that economic growth slows after a country's debt grows to 90% of its GDP"

(emphasis added)

Anecdotes

Fixing this Excel error transforms high-debt countries from recession to growth

By Tim Fernholz • April 16, 2013



	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

"But new research suggests that the magic 90% number isn't so magic after all […] Three economists at the University of Massachusetts, Amherst, set out to replicate the Rogoff-Reinhart findings and instead *found the spreadsheet equivalent of a typo*, which you can see in the image above. That **blue grid in column L should go down five more cells**, the new research claims."

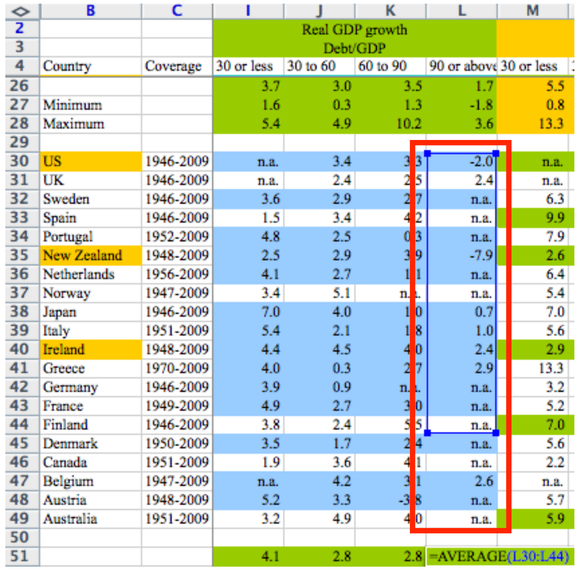
qz.com

(emphasis added)

Anecdotes

Fixing this Excel error transforms high-debt countries from recession to growth

By Tim Fernholz • April 16, 2013



	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.1	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	3.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

"But new research suggests that the magic 90% number isn't so magic after all […] Three economists at the University of Massachusetts, Amherst, set out to replicate the Rogoff-Reinhart findings and instead *found the spreadsheet equivalent of a typo*, which you can see in the image above. That **blue grid in column L should go down five more cells**, the new research claims."

qz.com

(emphasis added)

Anecdotes

Excel errors are life-and-death:

An Excel spreadsheet error that wiped nearly 16,000 English Covid cases from national statistics may have led to more than 1,500 preventable deaths, according to a paper from Warwick University.

Cases that were removed from the record due to the spreadsheet error were also not referred to the NHS test-and-trace operation, meaning people who had been exposed to a Covid sufferer were not told to self-isolate.

— The Guardian, 25 Nov 2020

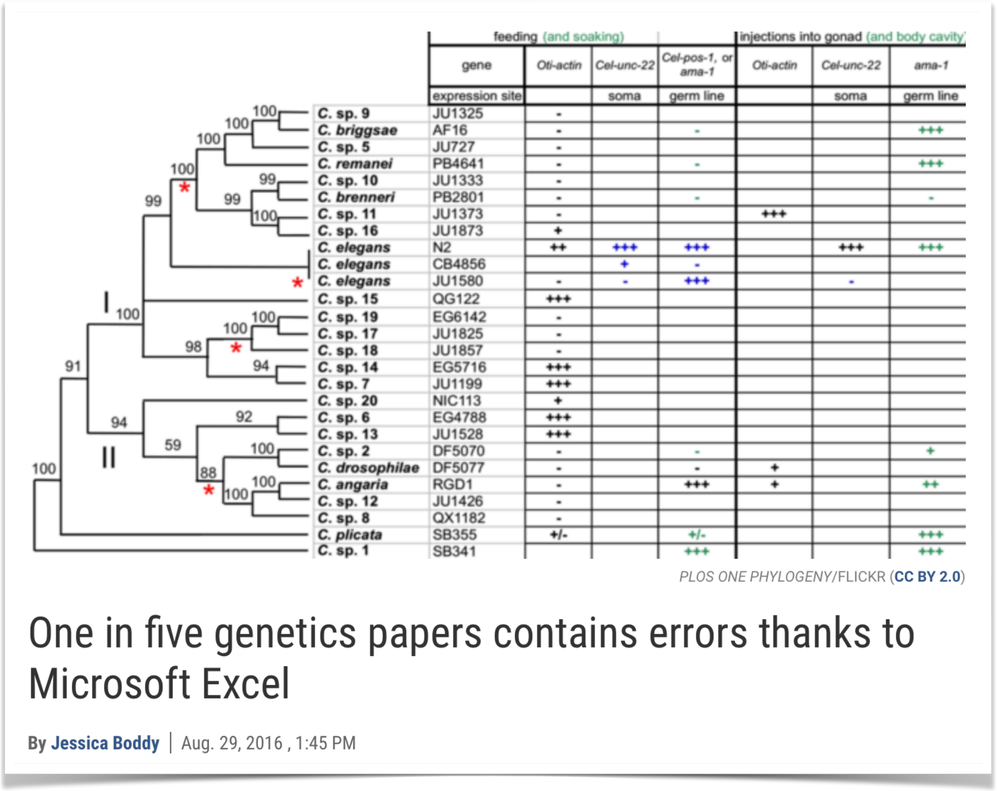
Data glitch 'may have led to more than 1,500 Covid deaths in England'

Public Health England disputes Warwick University economists' findings as 'misleading'

- **Coronavirus - latest updates**
- **See all our coronavirus coverage**

Anecdotes

Does it affect science? Absolutely!



PLOS ONE PHYLOGENY/FLICKR (CC BY 2.0)

One in five genetics papers contains errors thanks to Microsoft Excel

By Jessica Boddy | Aug. 29, 2016 , 1:45 PM

sciencemag.org

Anecdotes

Does it affect science? Absolutely!

Genome Biology

Home About Articles Submission Guidelines

Comment | Open Access | Published: 23 August 2016

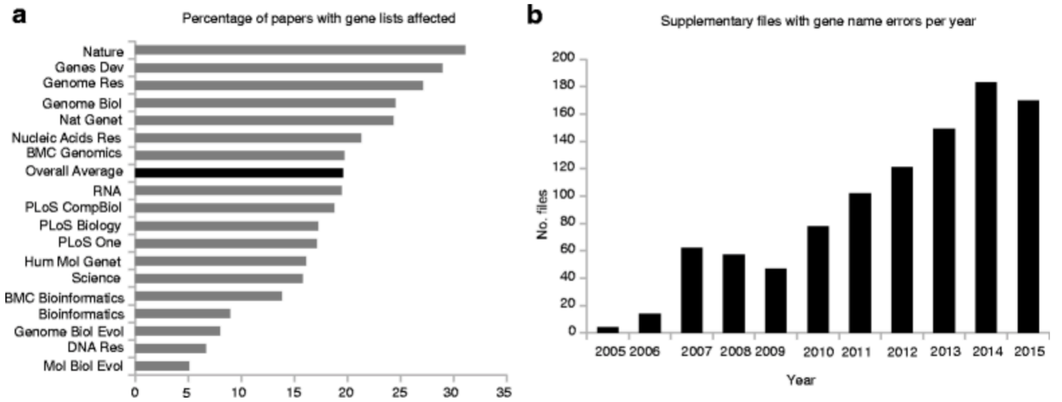
Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Download Citation

Fig. 1

From: Gene name errors are widespread in the scientific literature



Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

Anecdotes

Does it affect science? Absolutely!

Genome Biology

Home About Articles Submission Guidelines

Comment | Open Access | Published: 23 August 2016

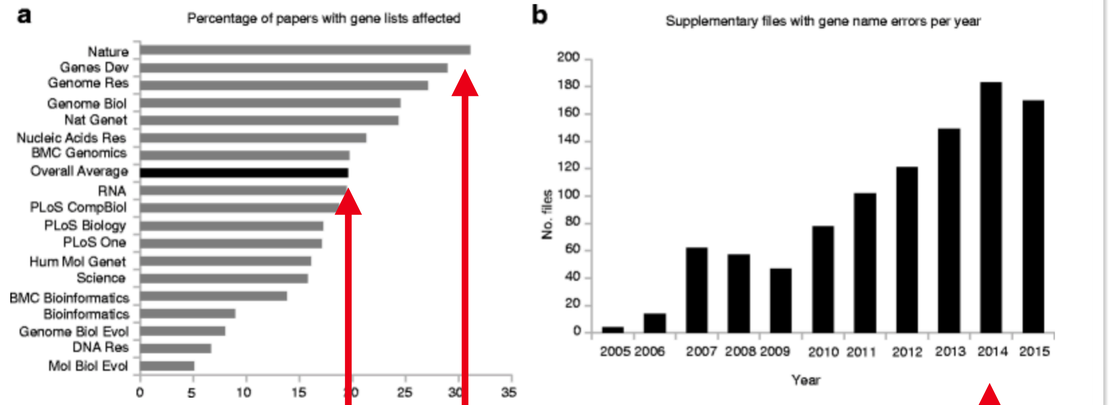
Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Download Citation

Fig. 1

From: Gene name errors are widespread in the scientific literature



Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

20% overall average

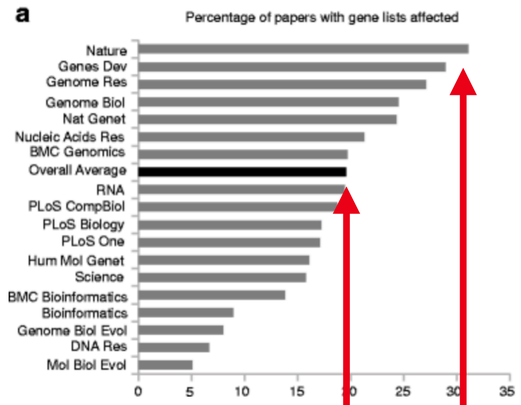
30% in Nature!

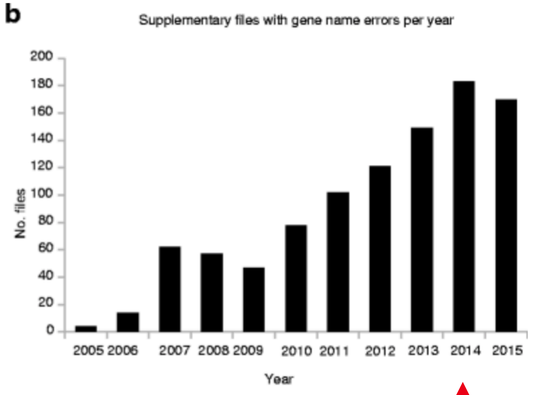
More errors over time

Anecdotes

Fig. 1

From: Gene name errors are widespread in the scientific literature





Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

20% overall average

30% in Nature!

More errors over time

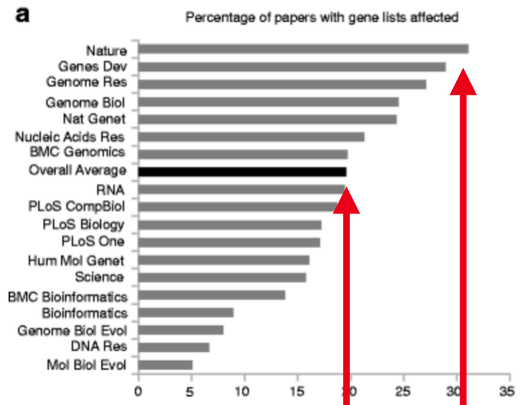
Scientific aside

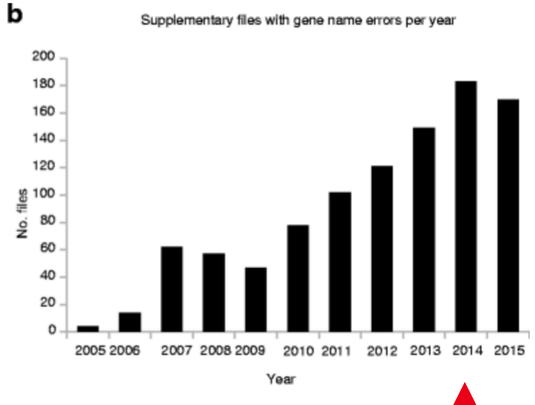
Let's unpack this last statement

Anecdotes

Fig. 1

From: Gene name errors are widespread in the scientific literature





Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

20% overall average

30% in Nature!

More errors over time

Scientific aside

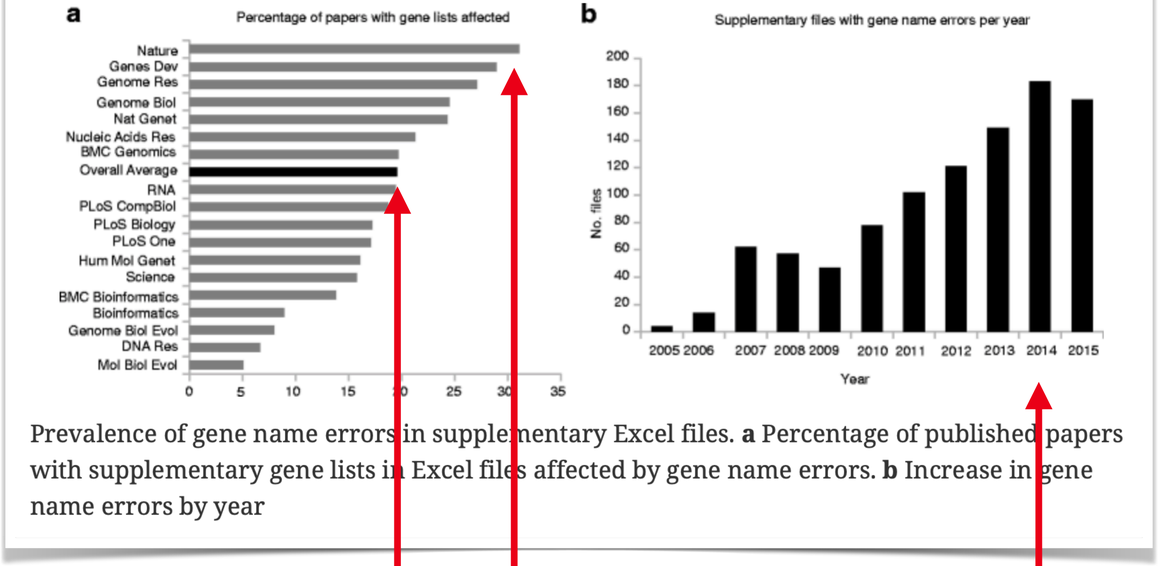
Let's unpack this last statement

More errors over time? But what is the error *rate*? Aren't there just more files over time?

Anecdotes

Fig. 1

From: Gene name errors are widespread in the scientific literature



Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

20% overall average

30% in Nature!

More errors over time

Scientific aside

Let's unpack this last statement

More errors over time? But what is the error *rate*? Aren't there just more files over time?

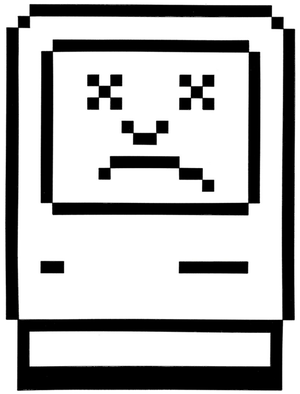
From the paper:

"Linear-regression estimates show gene name errors in supplementary files have increased at an **annual rate of 15%** over the past five years, outpacing the increase in published papers (**3.8%** per year)"

(emphasis added)

Problems with spreadsheets*

- Interactivity is error-prone
  - Interactivity leads to a poor paper trail!
- Automatic routines used wrongly by spreadsheet applications
- Difficult to track changes across different versions of a dataset
- Unable to separate code and data
- Formatting stops being metadata and becomes data
- Poor scalability



*and spreadsheet ***applications***

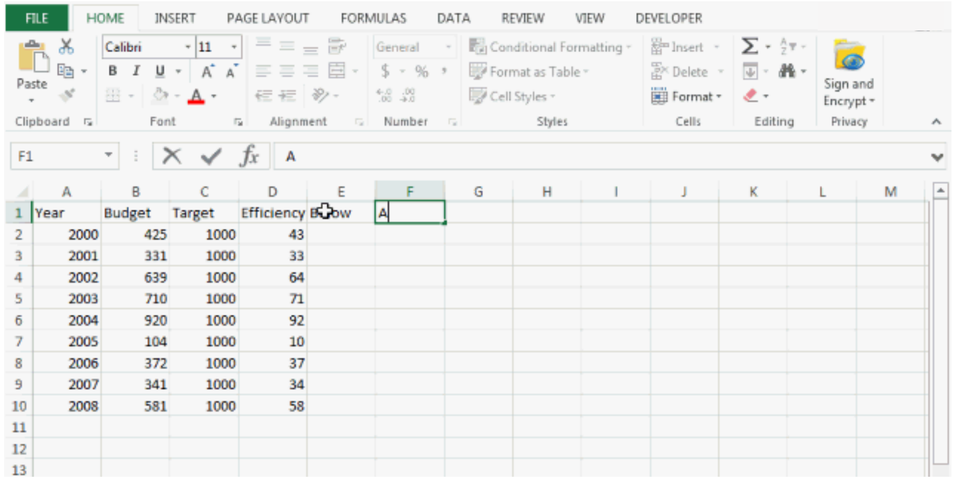
Interactivity is error-prone

We are all human

It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected



VERY HARD to keep a record of all the modifications occurring to the data when you are working on it

Harms reproducibility in the long run

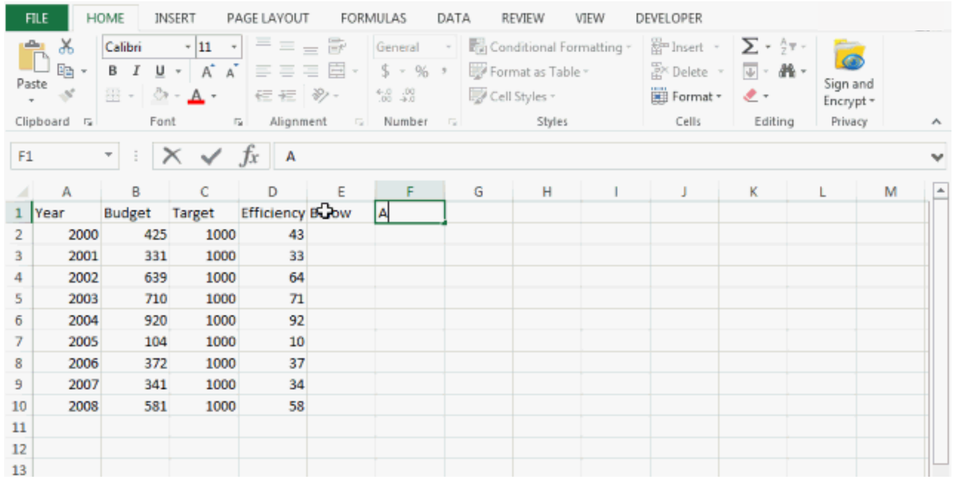
Interactivity is error-prone

We are all human

It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected



VERY HARD to keep a record of all the modifications occurring to the data when you are working on it

Harms reproducibility in the long run

Interactivity is error-prone

We are all human

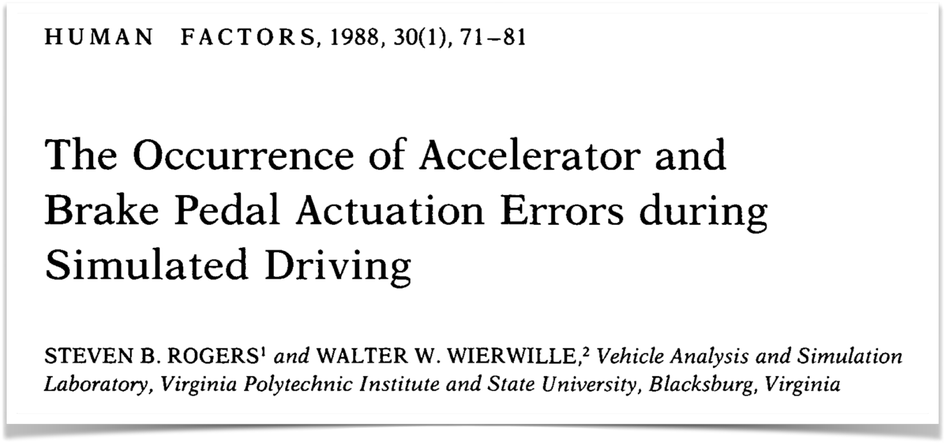
It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected

We do not like being confronted with our true error rates

How often does a driver press the wrong pedal when driving? Approximately 1-4 times per hour of driving



HUMAN FACTORS, 1988, 30(1), 71–81

The Occurrence of Accelerator and Brake Pedal Actuation Errors during Simulated Driving

STEVEN B. ROGERS[1] *and* WALTER W. WIERWILLE,[2] *Vehicle Analysis and Simulation Laboratory, Virginia Polytechnic Institute and State University, Blacksburg, Virginia*

Interactivity is error-prone

We are all human

It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected

What about spreadsheets?

Spreadsheet Errors:
What We Know.
What We Think We Can Do.

Dr. Raymond R. Panko
University of Hawaii
panko@hawaii.edu, http://panko.cba.hawaii.edu

Interactivity is error-prone

We are all human

It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected

Panko, 1998, 2005

Table 1: Studies of Spreadsheet Errors **(audits)**

Authors	Year	Number of SSs Audited	Average Size (Cells)	Percent of SSs with Errors	Cell Error Rate	Comment
Hicks	1995	1	3,856	100%	1.2%	One omission error would have caused an error of more than a billion dollars.
Coopers & Lybrand	1997	23	More than 150 rows	91%		Off by at least 5%
KPMG	1998	22		91%		Only significant errors
Lukasic	1998	2	2,270 & 7,027	100%	2.2%, 2.5%	In Model 2, the investment's value was overstated by 16%. Quite serious.
Butler	2000	7		86%	0.4%**	Only errors large enough to require additional tax payments**
Clermont, Hanin, & Mittermeier	2002	3		100%	1.3%, 6.7%, 0.1%	Computed on the basis of non-empty cells
Interview I*	2003	~36 / yr		100%		Approximately 5% had *extremely* serious errors
Interview II*	2003	~36 / yr		100%		Approximately 5% had *extremely* serious errors
Lawrence and Lee	2004	30	2,182 unique formulas	100%	6.9%	30 most financially significant SSs audited by Mercer Finance & Risk Consulting in previous year.
Total since 1995		**88**		**94%**	**5.2%**	

Interactivity is error-prone

We are all human

It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected

Table 2: Studies of Spreadsheet Errors **(experiments)**

Study	SSs	% with Errors	Cell Error Rate (CER)	Remarks
				spreadsheets Differences were nonsignificant
No training in design	30		16.8%	
Training in design	58		8.4%	
Kreie, et al. (post test) (2000)	73	42%	2.5%	
Teo & Tan (1997)	168	42%	2.1%	Wall task.
Panko & Halverson (1997)				Galumpke task. Undergraduates Based on all text and number cells
Working alone	42	79%	5.6%	
Dyads (Groups of 2)	46	78%	3.8%	
Tetrads (Groups of 4)	44	64%	1.9%	
Panko & Halverson (2001)	35	86%	4.6%	Undergraduates
Panko & Halverson (2001)				
Panko & Sprague (1998)	26	35%	2.1%	Wall Task. MBA students with little or no SS development experience
Panko & Sprague (1998)	17	24%	1.1%	Wall Task. MBA students with 250 hours or more SS development experience

Panko, 1998, 2005

Interactivity is error-prone

We are all human

It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected

Panko, 1998, 2005

Table 3: Code Inspection Experiments

Study	Subjects	Sample Size	% of Errors Detected	Remarks
Galletta et al. (1993)	MBA students & CPAs Taking a Continuing Education Course			Budgeting task containing seeded errors
Total sample		60	56%	
CPA novices	<100 hours of work experience with SSs	15	57%	
CPA experts	<100 hours of work experience with SSs	15	66%	
MBA students, novices	>250 hours of work experience with SSs	15	52%	
MBA students, experienced	>250 hours of work experience with SSs	15	48%	
Domain (logic) errors corrected			46%	
Device (mechanical) errors corrected			65%	
Galletta et al. (1997)	MBA students		51%	Same task used 1993 study
Overall		113	51%	
On-Screen		45	45%	
On Paper		68	55%	
Panko (1999)				Modified version of Galletta wall task.
Undergrads working alone		60	63%	
Undergrads working in groups of three		60	83%	
Panko & Sprague (1998)	Undergrads	23	16%	Students who made errors in the Wall task who then went on to inspect their own spreadsheets.

≈46% (on average) of ***seeded errors*** went ***undetected***

Interactivity is error-prone

We are all human

It is very easy to make a mistake when working interactively

Typos and stray key presses, copy-and-paste mistakes, Missed mouse actions, …

Not all mistakes will be noticed and corrected

Panko, 1998, 2005

Table 3: Code Inspection Experiments

Study	Subjects	Sample Size	% of Errors Detected	Remarks
Galletta et al. (1993)	MBA students & CPAs Taking a Continuing Education Course			Budgeting task containing seeded errors
Total sample		60	56%	
CPA novices	<100 hours of work experience with SSs	15	57%	
CPA experts	<100 hours of work experience with SSs	15	66%	
MBA students, novices	>250 hours of work experience with SSs	15	52%	
MBA students, experienced	>250 hours of work experience with SSs	15	48%	
Domain (logic) errors corrected			46%	
Device (mechanical) errors corrected			65%	
Galletta et al. (1997)	MBA students		51%	Same task used 1993 study
Overall		113	51%	
On-Screen		45	45%	
On Paper		68	55%	
Panko (1999)				Modified version of Galletta wall task.
Undergrads working alone		60	63%	
Undergrads working in groups of three		60	83%	
Panko & Sprague (1998)	Undergrads	23	16%	Students who made errors in the Wall task who then went on to inspect their own spreadsheets.

≈46% (on average) of ***seeded errors*** went ***undetected***

Interactivity is error-prone

"Hang on!" you may be thinking

Typos and stray key presses?
Copy-and-paste mistakes?
Missed mouse actions?

All these happen when coding
too!

Interactivity is error-prone

"Hang on!" you may be thinking

Typos and stray key presses?
Copy-and-paste mistakes?
Missed mouse actions?

All these happen when coding too!

Indeed:

[Spreadsheet error rates are]
Consistent with Other Human Error Data

When most people look at Tables 1 2, and 3, their first reaction is that such high error rates are impossible. In fact, they are not only possible. They are entirely consistent with data on human error rates from other work domains. The Human Error Website (Panko, 2005a) presents data from a number of empirical studies. Broadly speaking, when humans do simple mechanical tasks, such as typing, they make undetected errors in about 0.5% of all actions. When they do more complex logical activities, such as writing programs, the error rate rises to about 5%. These are not hard and fast numbers,

Panko, 1998, 2005

Interactivity is error-prone

"Hang on!" you may be thinking

Typos and stray key presses?
Copy-and-paste mistakes?
Missed mouse actions?

All these happen when coding too!

Indeed:

[Spreadsheet error rates are]
Consistent with Other Human Error Data

When most people look at Tables 1 2, and 3, their first reaction is that such high error rates are impossible. In fact, they are not only possible. They are entirely consistent with data on human error rates from other work domains. The Human Error Website (Panko, 2005a) presents data from a number of empirical studies. Broadly speaking, when humans do simple mechanical tasks, such as typing, they make undetected errors in about 0.5% of all actions. When they do more complex logical activities, such as writing programs, the error rate rises to about 5%. These are not hard and fast numbers,

Panko, 1998, 2005

The difference: I argue that it is *less likely* (not impossible, but less likely) for errors to go unnoticed in computer code than in spreadsheets. The code still has to run.

Automatic routines used wrongly by spreadsheet applications

Excel in particular has lots of handy automatic features

For example, auto formatting of dates

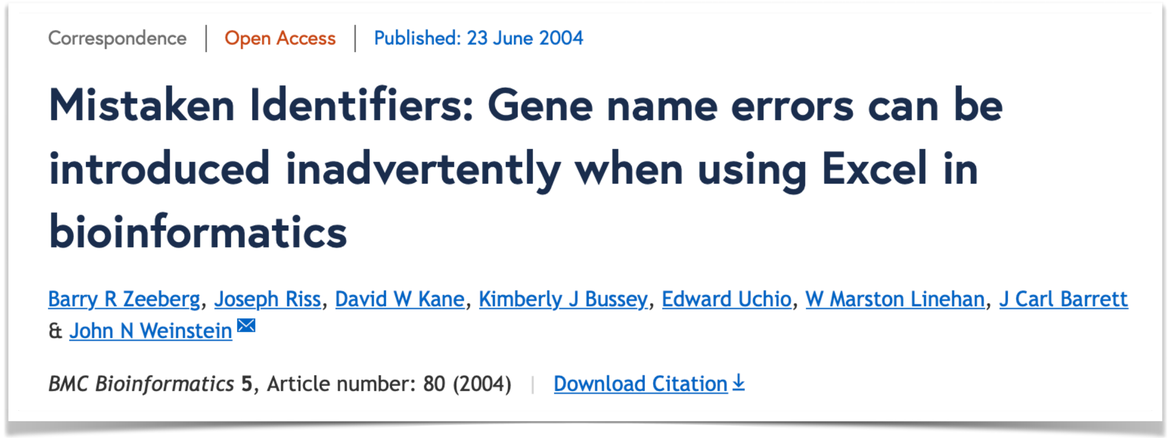
Trouble begins when the **data are changed without our knowledge**, especially when changes occur inconsistently

Automatic routines used wrongly by spreadsheet applications

Excel in particular has lots of handy automatic features

For example, auto formatting of dates

Trouble begins when the **data are changed without our knowledge**, especially when changes occur inconsistently



Correspondence | Open Access | Published: 23 June 2004

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

Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett & John N Weinstein

BMC Bioinformatics **5**, Article number: 80 (2004) | Download Citation

"A default date conversion feature in Excel (Microsoft Corp., Redmond, WA) was altering gene names that it considered to look like dates. For example, the **tumor suppressor DEC1 [Deleted in Esophageal Cancer 1]** [3] was being converted to '1-DEC.' [a date]

(emphasis added)

Automatic routines used wrongly by spreadsheet applications

Correspondence | Open Access | Published: 23 June 2004

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

Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett & John N Weinstein

BMC Bioinformatics **5**, Article number: 80 (2004) | Download Citation

"A default date conversion feature in Excel (Microsoft Corp., Redmond, WA) was altering gene names that it considered to look like dates. For example, the **tumor suppressor DEC1 [Deleted in Esophageal Cancer 1]** [3] was being converted to '1-DEC.' [a date]

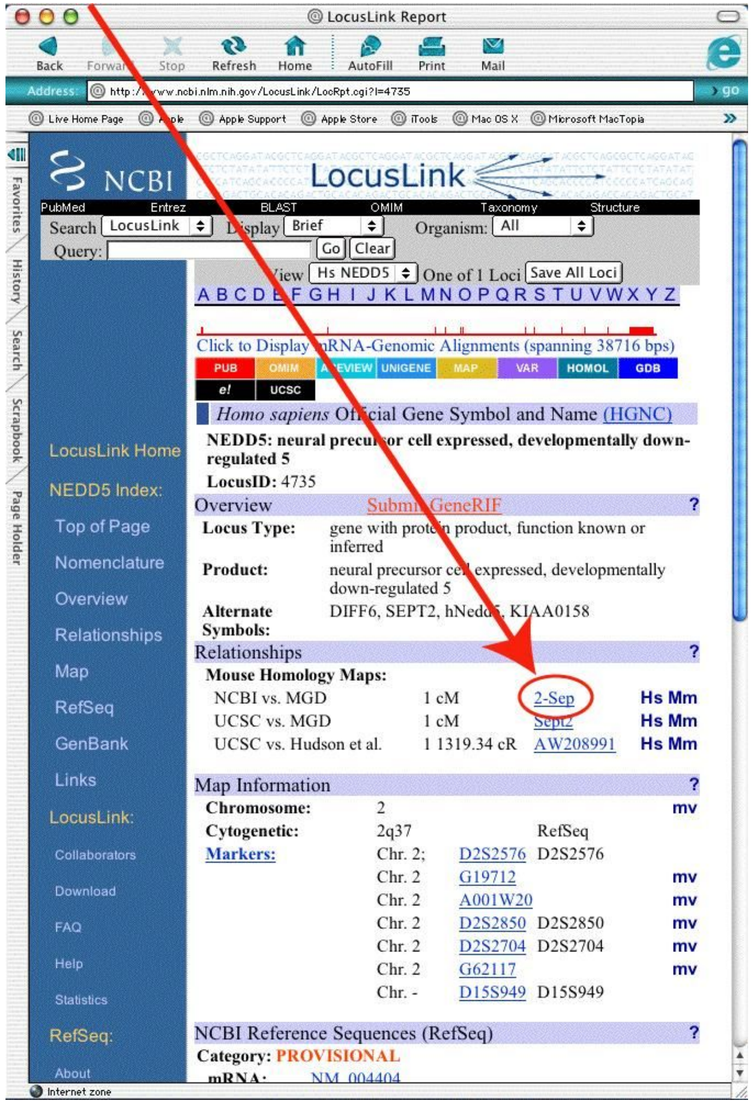


Figure 2 Screen shot of LocusLink from November 12, 2002 illustrating an error caused by default conversion of a gene name to date that had propagated from the human-mouse homology map data

Inability to separate code and data

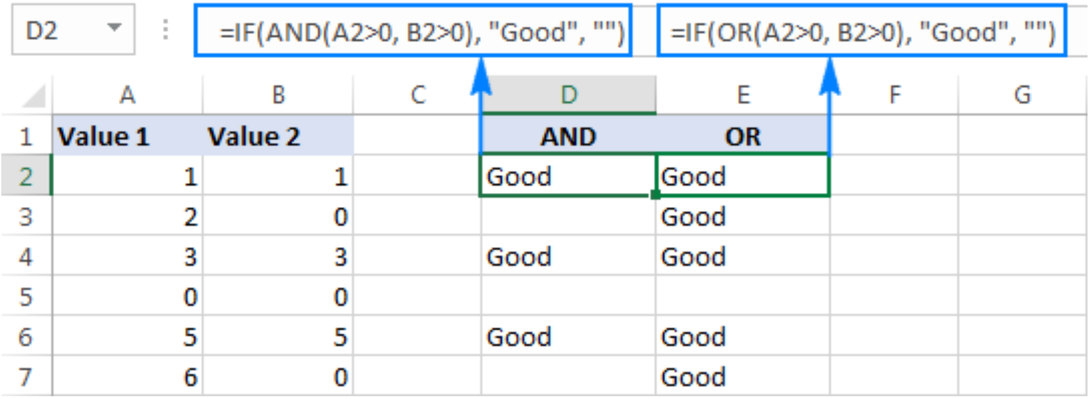
Good practice of scientific computing is to **isolate** code and data from one another as much as possible. Make code portable across datasets

Avoid having parts of the data (ex: *variable name strings*) scattered about the code

Instead: contain the details specific to the current data within variables that are then used in the code

Minimizes the need to change the code when the data change

Spreadsheets—by definition—cannot do this



	A	B	C	D	E	F	G
1	Value 1	Value 2		AND	OR		
2	1	1		Good	Good		
3	2	0			Good		
4	3	3		Good	Good		
5	0	0					
6	5	5		Good	Good		
7	6	0			Good		

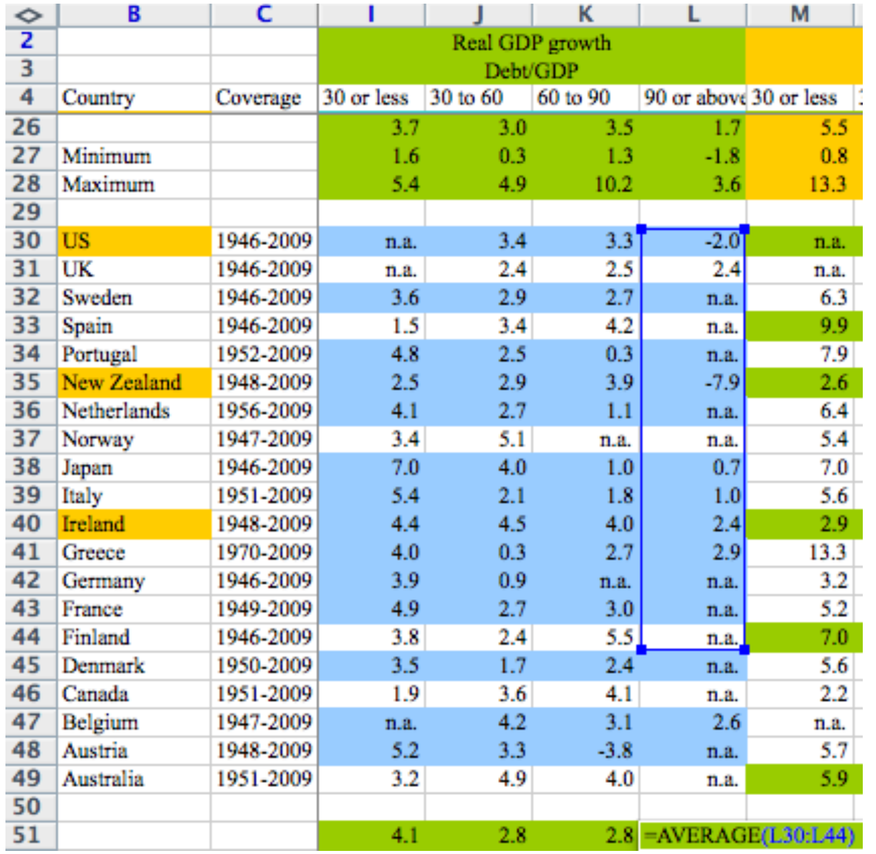
- Spreadsheet code (formula cells) intermingled with data
- Hard to isolate, see all the code
- Hard to debug

Formatting stops being metadata and becomes data

Being so visual, spreadsheet formatting, such as cell colors, text colors, and font changes, all help improve the readability and organization of the data

Some spreadsheet apps even support conditional formatting, using formulas to alter formatting programmatically

But, formatting has a tendency to become intrinsic to the data itself, yet it's often not portable across file formats. Won't be captured in a .csv file, for example



	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

Poor scalability

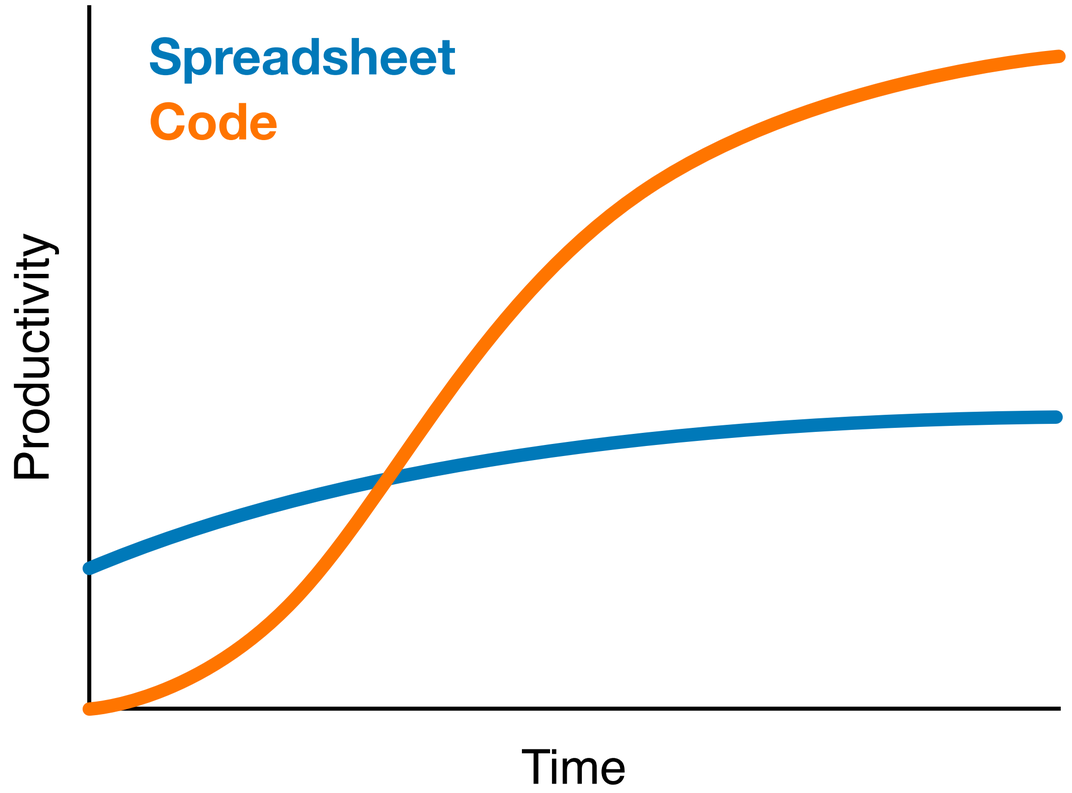
Typical apps like Excel can't handle 1M+ rows

Spreadsheets have a fast learning curve, quick to get started

Sufficiently complex tasks can become too time-consuming, much harder to automate

- Imagine having to change a single formula in 1400 .xlsx files?

Writing code takes longer to learn and get started, but scales better to large numbers of files and rows and columns



If spreadsheets are so bad, what about code?

Spreadsheets

Code

Code for data analysis

Code

Code acts as a safety barrier between you and the data
Error logging
Time stamping
Record keeping
Explicit testing
Automated pipelines

Even consoles like IPython can enable some of the great interactivity of a graphical interface

You do not get these things for free, however
Careful data handling requires well designed, meticulous code
Simple, reliable code, no need to be clever
A defensive posture ("*anything that can go wrong, will*")

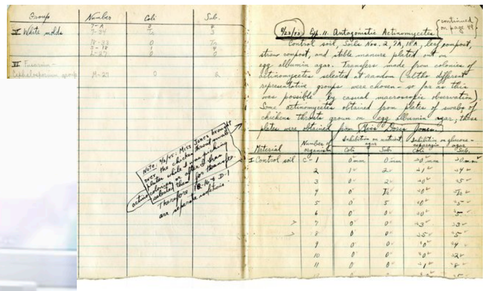
The computer is your lab

Code acts as a safety barrier between you and the data
Error logging
Time stamping
Recording keeping
Explicit testing
Automated pipelines

Even consoles like IPython can enable some of the great interactivity of a graphical interface

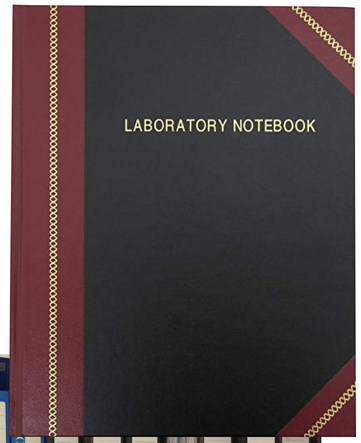
You do not get these things for free, however
Careful data handling requires well designed, meticulous code
Simple, reliable code, no need to be clever
A defensive posture ("*anything that can go wrong will*")

CULTURA CREATIVE (RF) / ALAMY STOCK PHOT



How to keep a lab notebook

By Elisabeth Pain | Sep. 3, 2019 , 1:45 PM



Summary

Problems with spreadsheets

- Interactivity is error-prone
  - Interactivity leads to a poor paper trail!
- Automatic routines used wrongly by spreadsheet applications
- Difficult to track changes across different versions of a dataset
- Inability to separate code and data
- Formatting stops being metadata and becomes data
- Poor scalability

Recall that I said:

I use spreadsheets all the time:

- Task lists
- Scheduling and planning
- Budgets
- Grading

When do I use spreadsheets? When I have one, maybe two *screens* of data

Never for research!