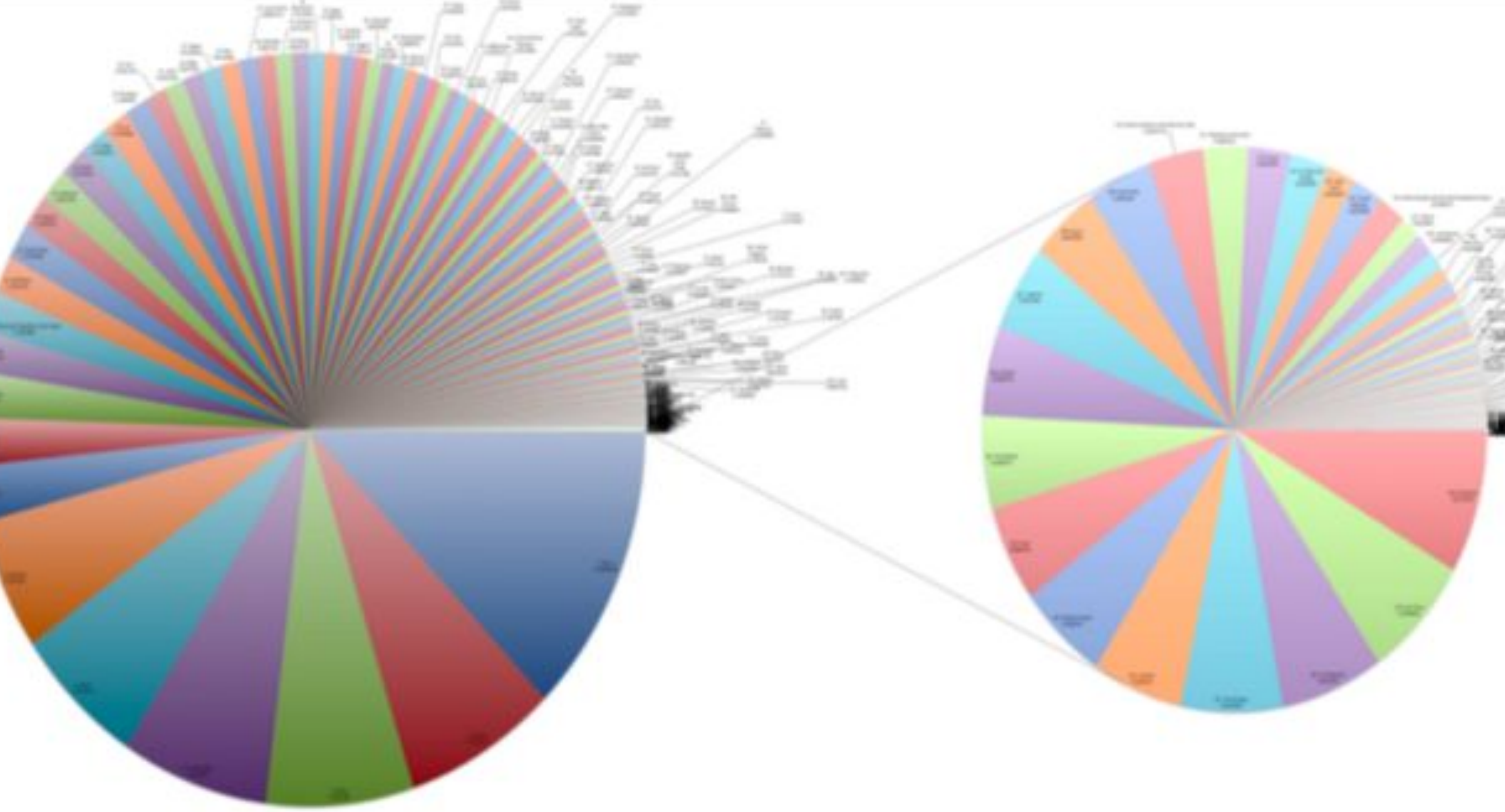
Data visualization

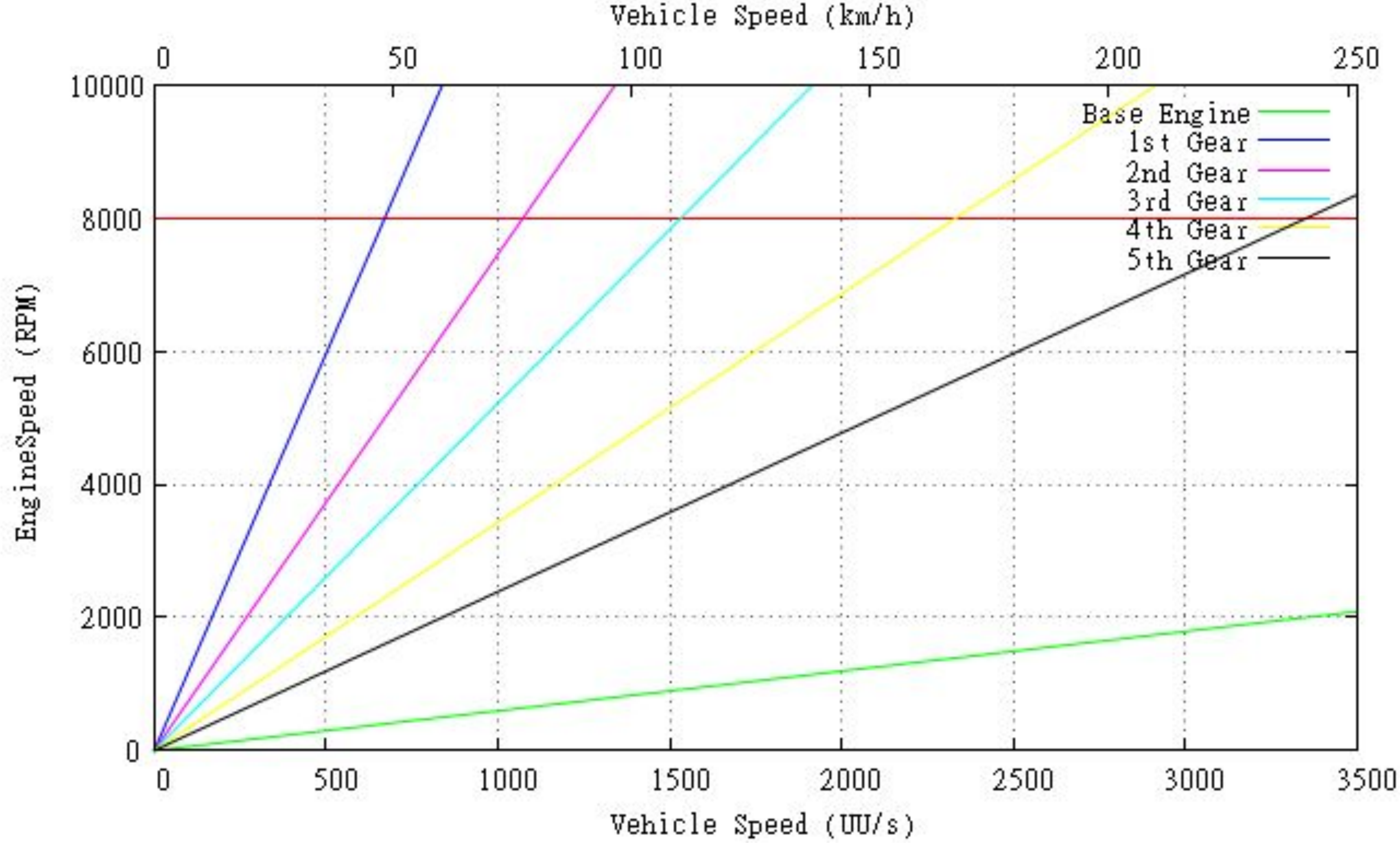
Quiz

- What do you find interesting in today's VotW?
- Explain "pre-attentive processing". Can you imagine and explain an example of incorporating this principle into your visualizations?
- What are the pros and cons of the minimalism design principle (e.g., Tufte's data-ink maximization)?

Visualization maxims

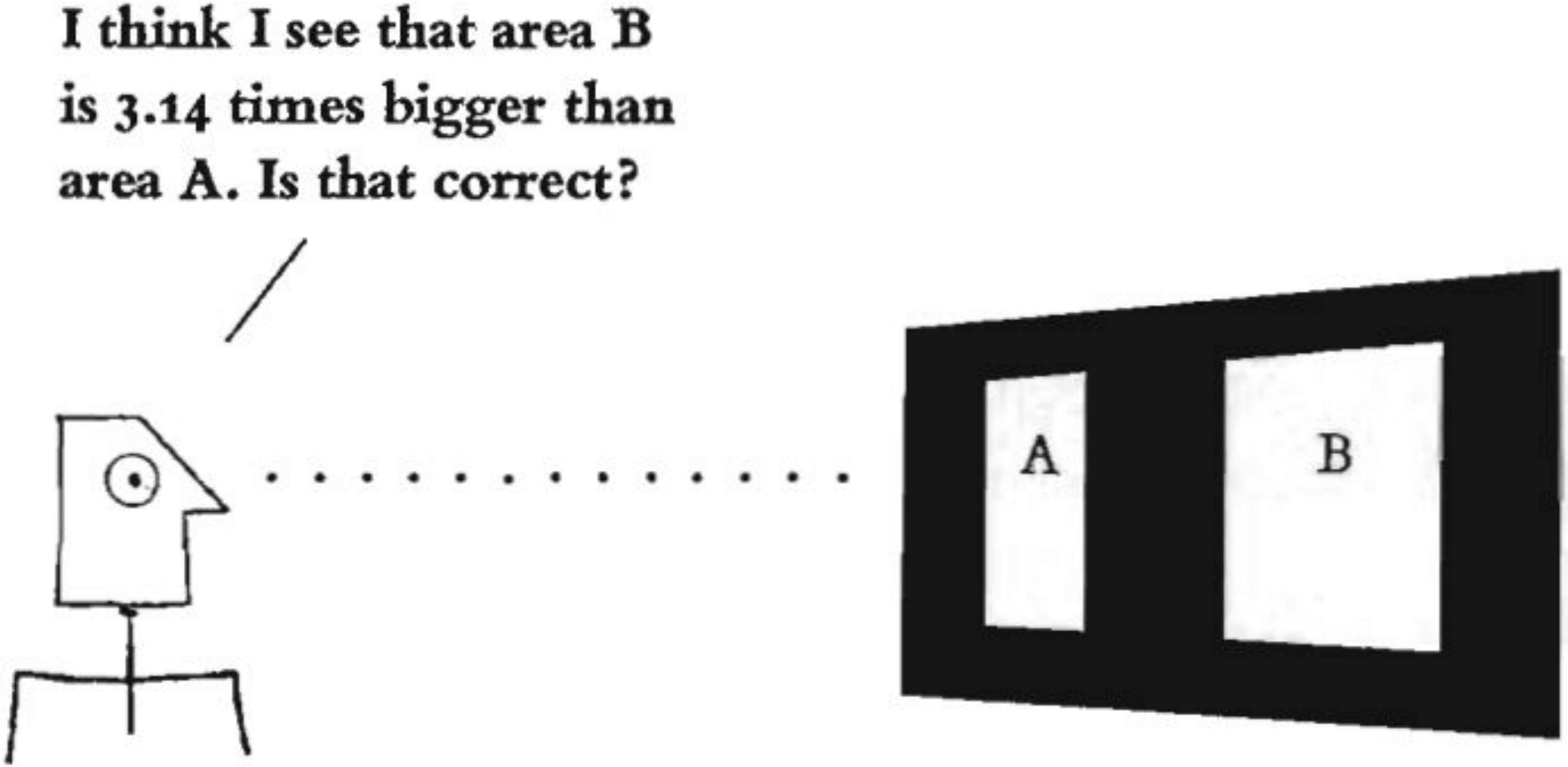
What is visualized should be visible.




Vehicle Speed (km/h)
0
50
100
150
200
250
10000
8000
6000
4000
2000
0
EngineSpeed (RPM)
Base Engine
1st Gear
2nd Gear
3rd Gear
4th Gear
5th Gear
0
500
1000
1500
2000
2500
3000
3500
Vehicle Speed (UU/s)

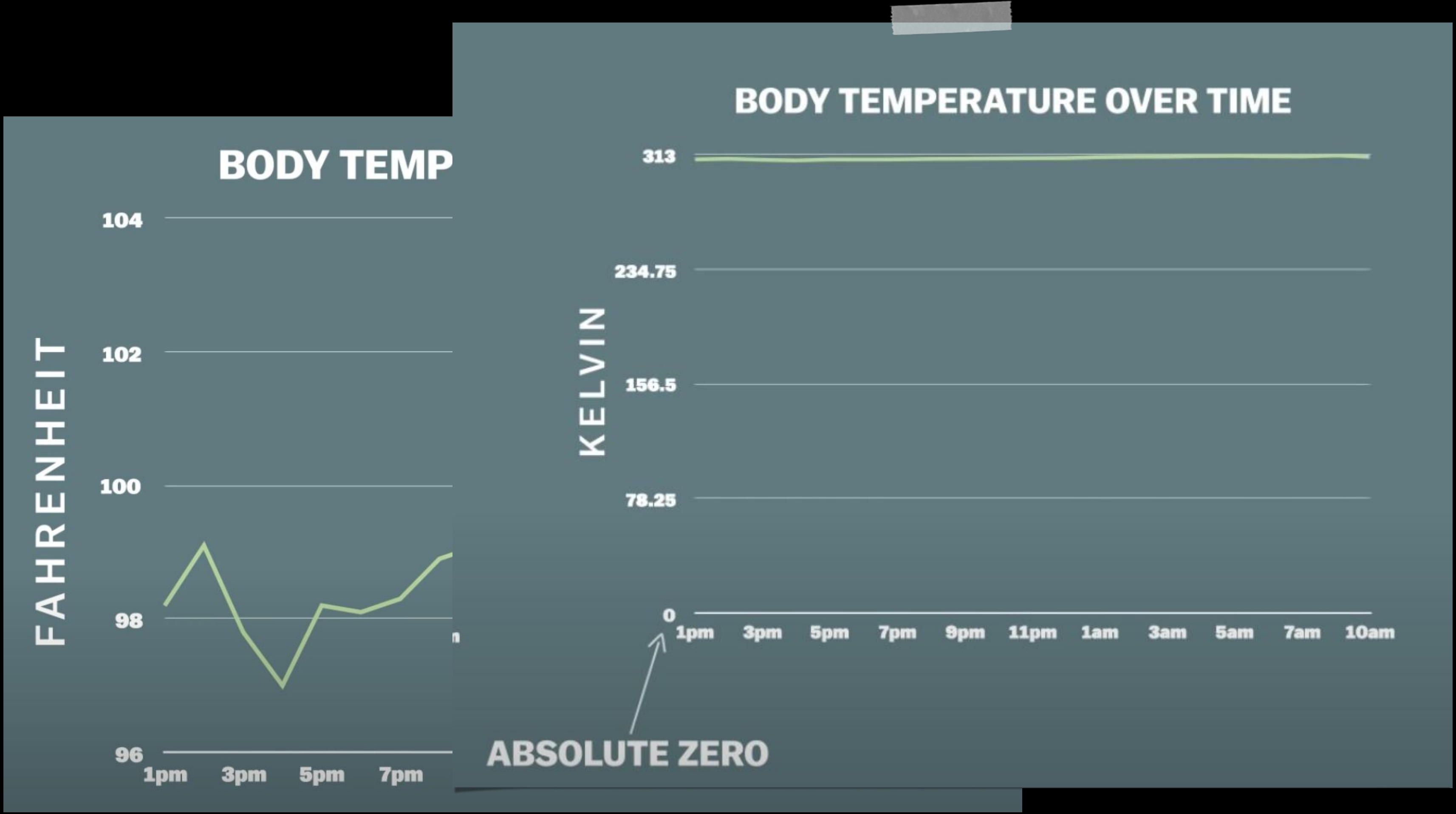
Create yours

Keep asking yourself, “does my visualization correctly represent the data?”



I think I see that area B is 3.14 times bigger than area A. Is that correct?
A
B

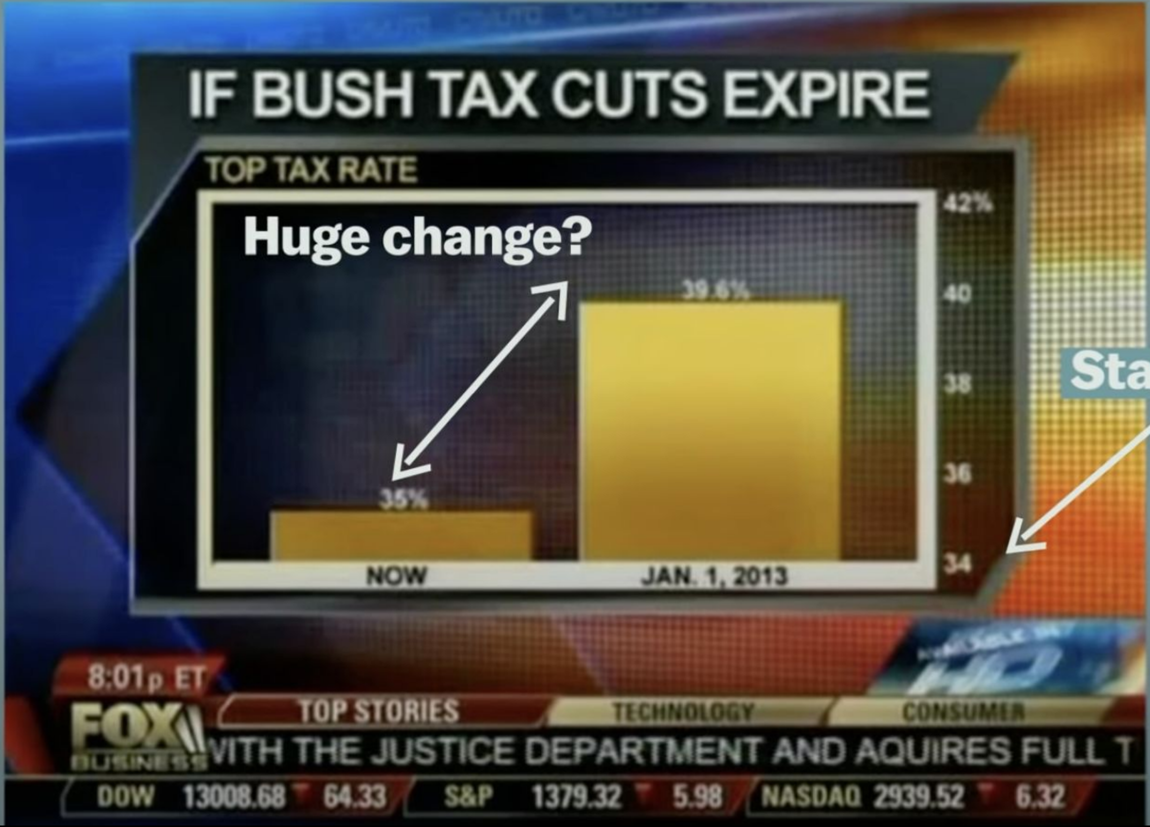
Show proper, enough contexts.



BODY TEMP

BODY TEMPERATURE OVER TIME

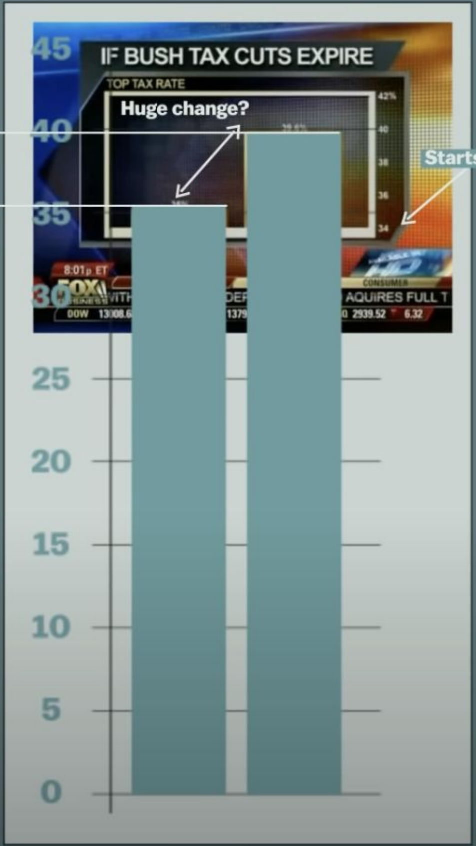
ABSOLUTE ZERO



Huge change?

Starts at 34%

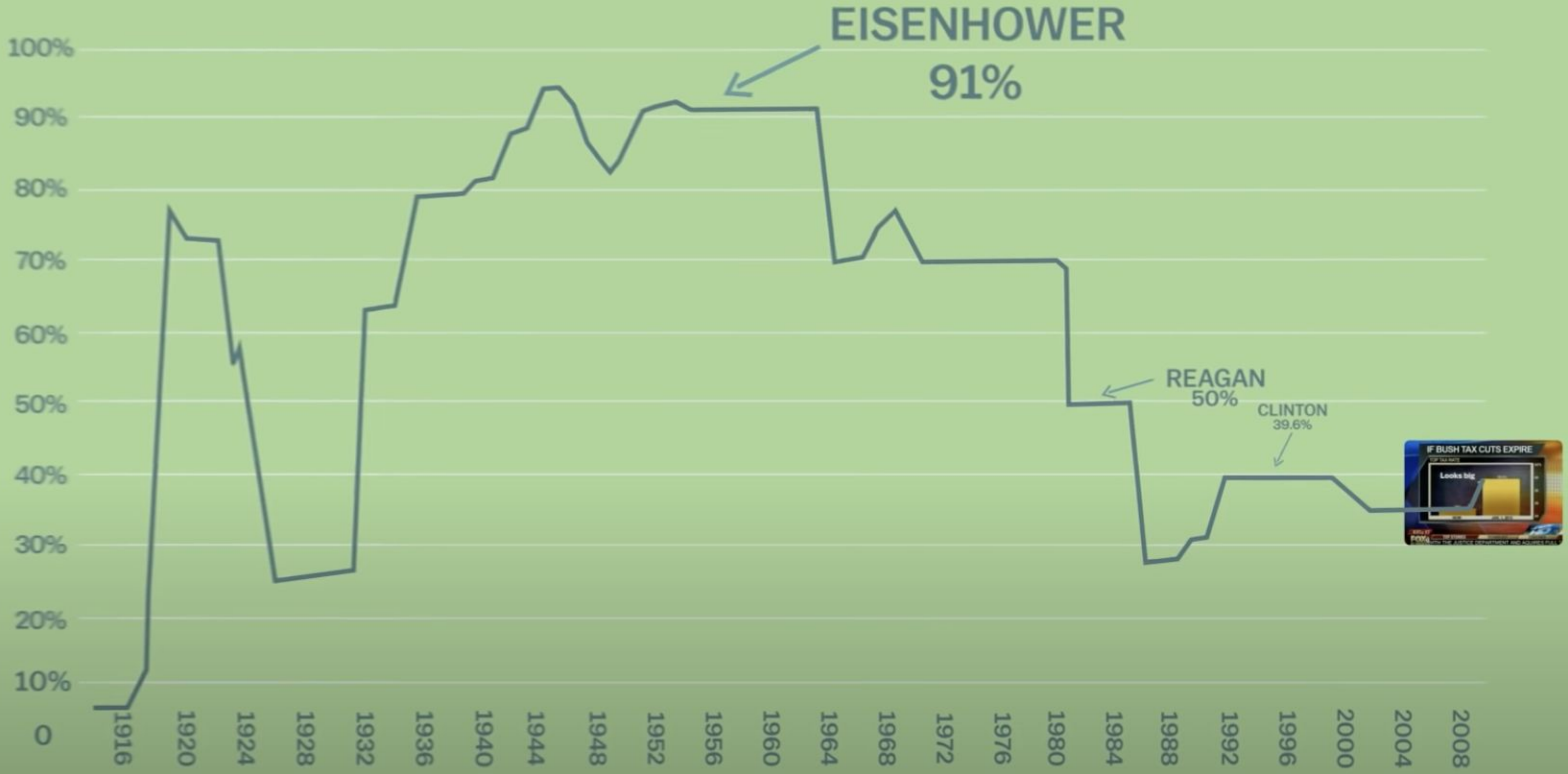
Nope!



Starts at 34%

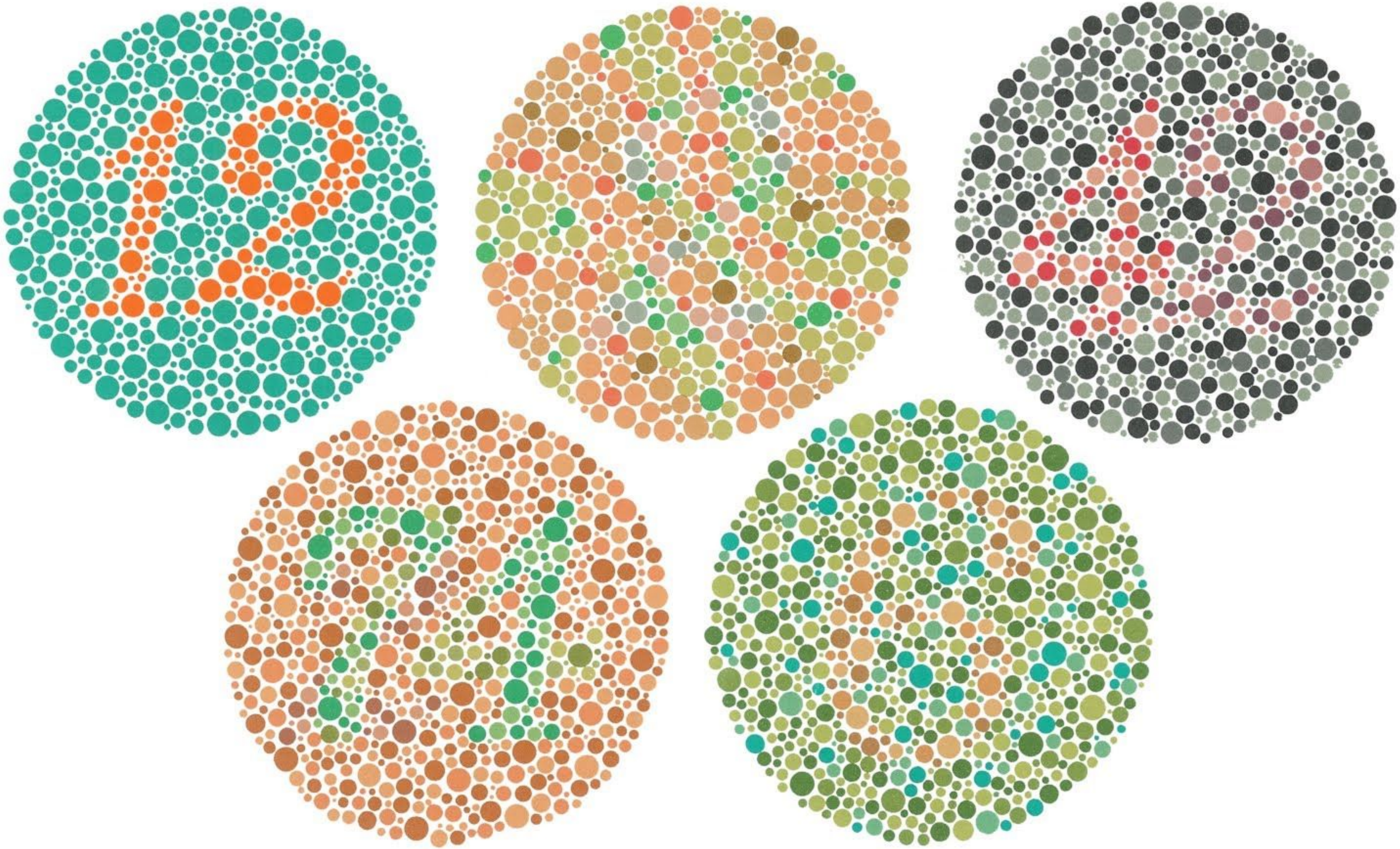
Starts at 0%

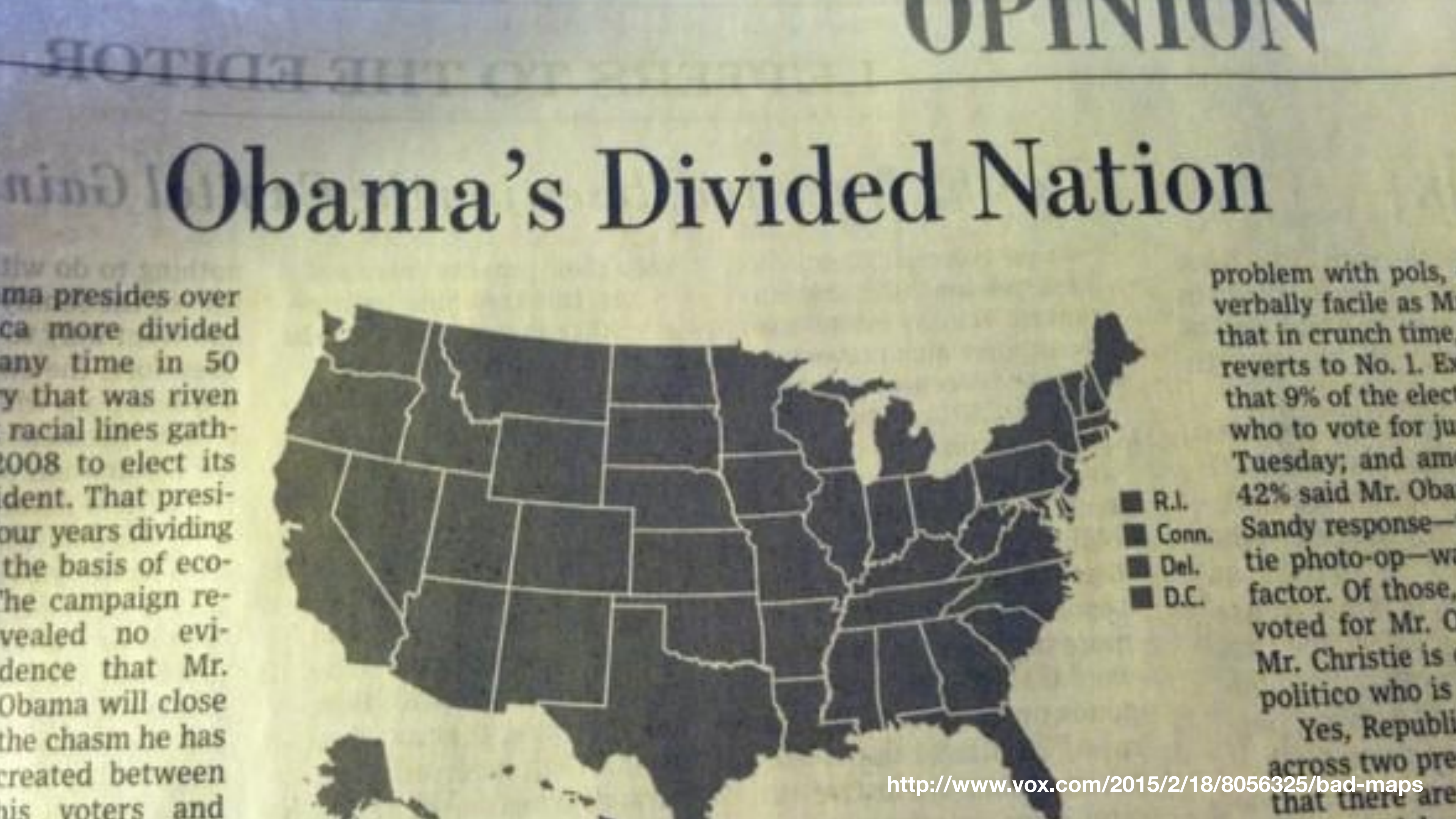
(Marginal) TAX RATE FOR TOP EARNERS



Know your audience

Make accessible and robust visualizations





OPINION

Obama's Divided Nation

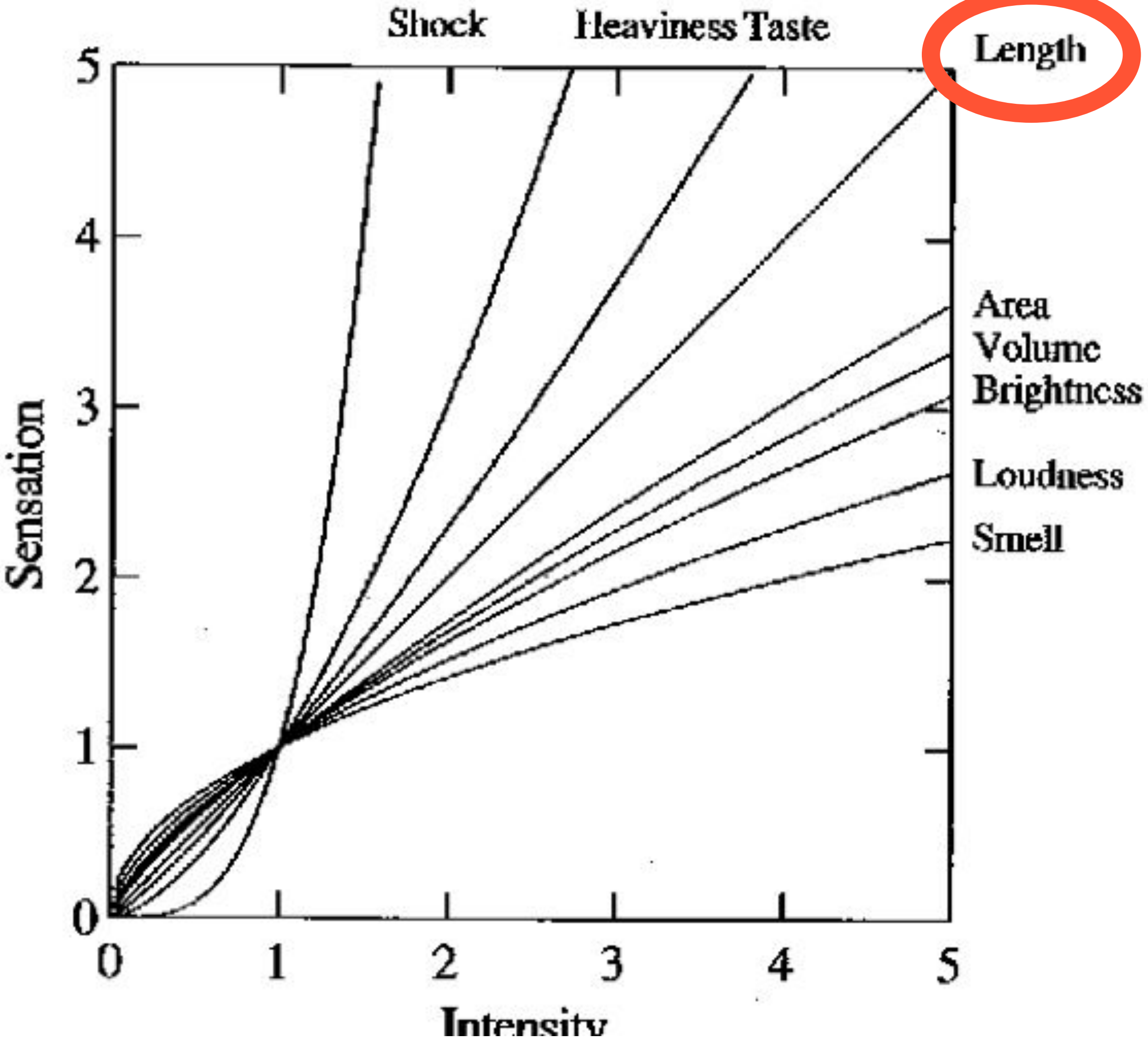
ma presides over ca more divided any time in 50 y that was riven racial lines gath- 2008 to elect its ident. That presi- our years dividing the basis of eco- The campaign revealed no evidence that Mr. Obama will close the chasm he has created between his voters and

problem with pols, verbally facile as M that in crunch time reverts to No. 1. Ex that 9% of the elect who to vote for ju Tuesday; and am 42% said Mr. Oba Sandy response— tie photo-op—wa factor. Of those, voted for Mr. O Mr. Christie is politico who is

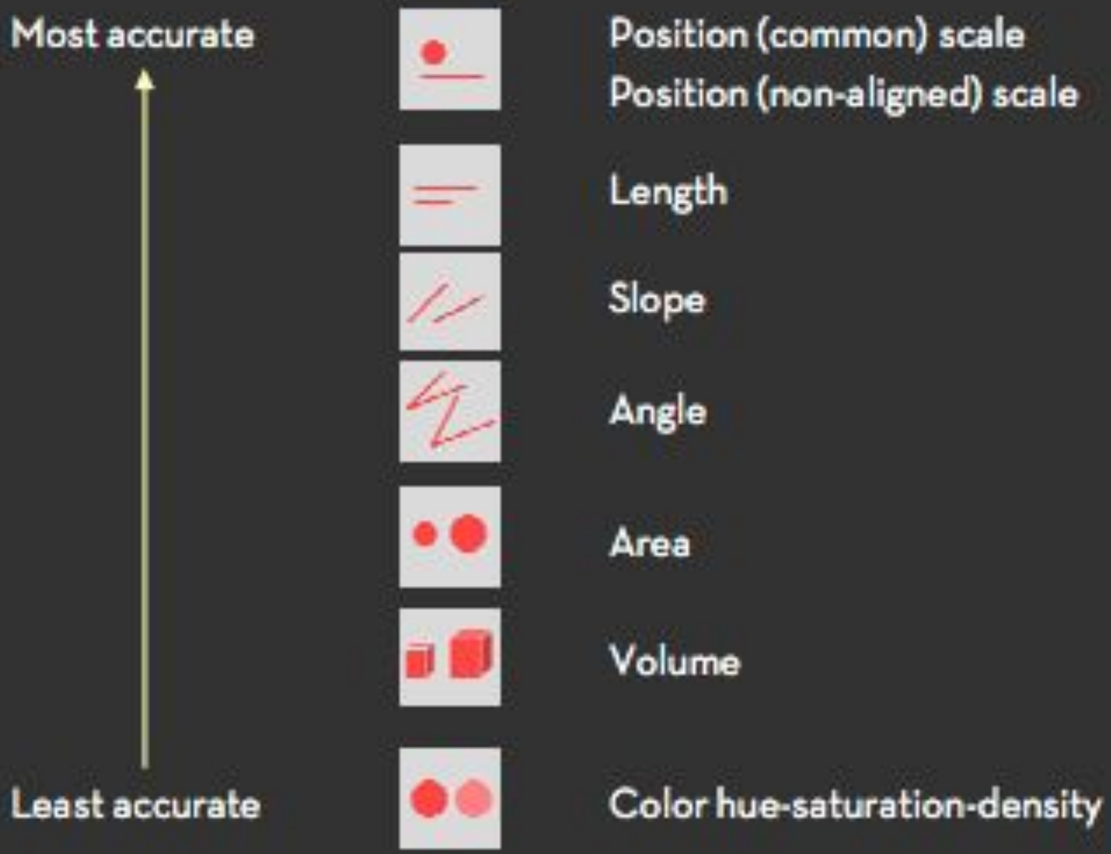
Yes, Republi across two pre that there are

http://www.vox.com/2015/2/18/8056325/bad-maps

Use accurate visual encodings when possible.


Shock
Heaviness
Taste
Length
Area
Volume
Brightncss
Loudness
Smell
Sensation
0
1
2
3
4
5
0
1
2
3
4
5
Intensity

Relative magnitude estimation



Most accurate

Position (common) scale

Position (non-aligned) scale

Length

Slope

Angle

Area

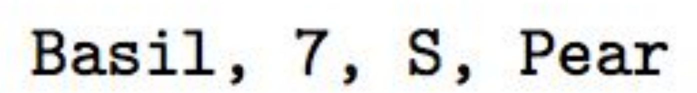
Volume

Color hue-saturation-density

Least accurate

Break rules if necessary, but be fully aware of their pitfalls.

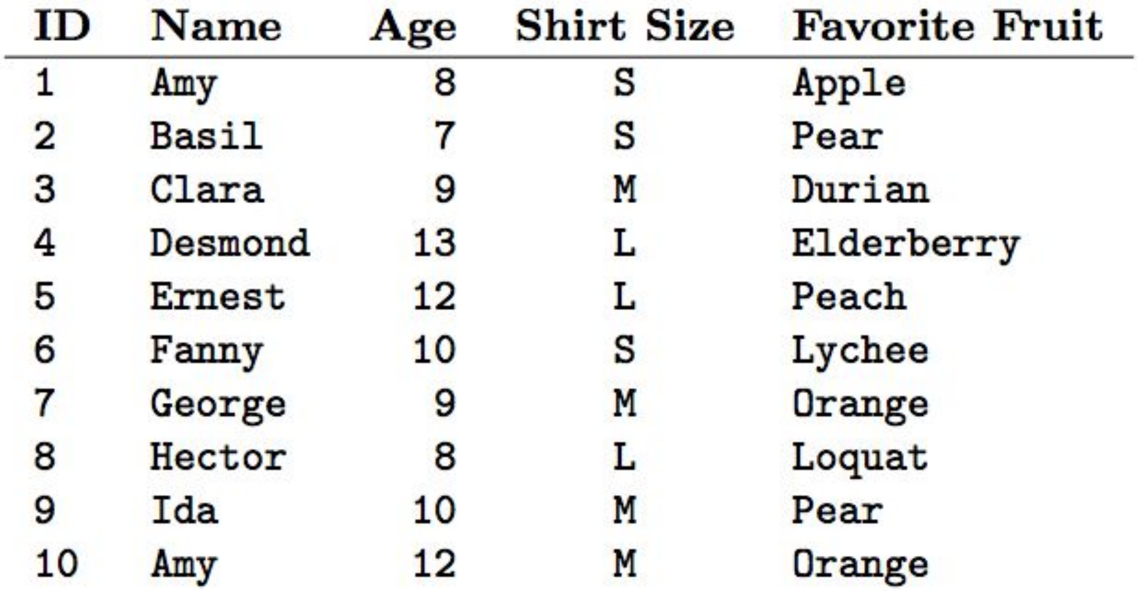
Data types



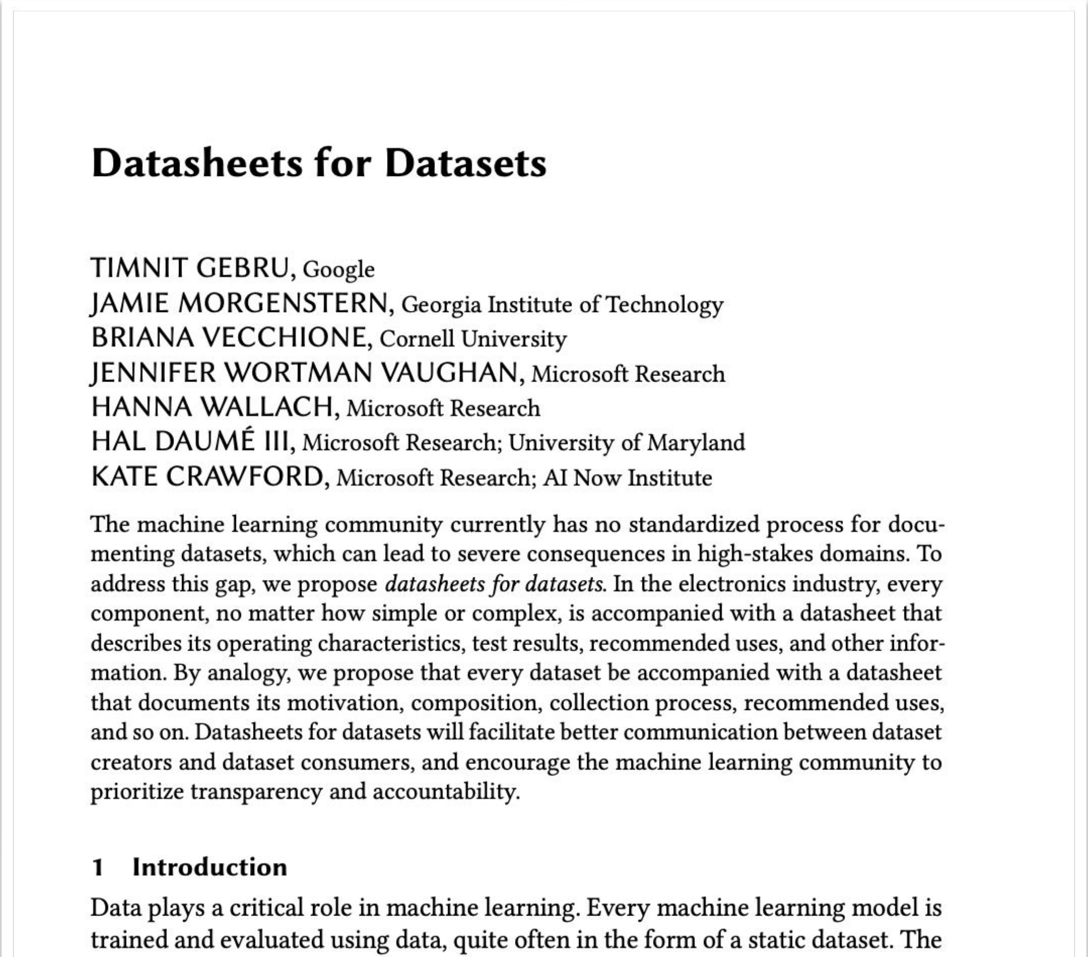
```
Basil, 7, S, Pear
```

?

Data should always be accompanied by “**data dictionary**” that contains details about the data



ID	Name	Age	Shirt Size	Favorite Fruit
1	Amy	8	S	Apple
2	Basil	7	S	Pear
3	Clara	9	M	Durian
4	Desmond	13	L	Elderberry
5	Ernest	12	L	Peach
6	Fanny	10	S	Lychee
7	George	9	M	Orange
8	Hector	8	L	Loquat
9	Ida	10	M	Pear
10	Amy	12	M	Orange



Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of a static dataset. The

Without a proper documentation, a dataset is **incomplete**!

Worse, it can lead to **disasters!**

https://arxiv.org/abs/1803.09010

What are the data types out there?

Nominal vs. Ordinal

Are all nominal variables

categorical?

Names, tweets, … are nominal, yet not categorical.

Are all ordinal variables

quantitative?

Small, medium, large,

…

Monday, Tuesday, …

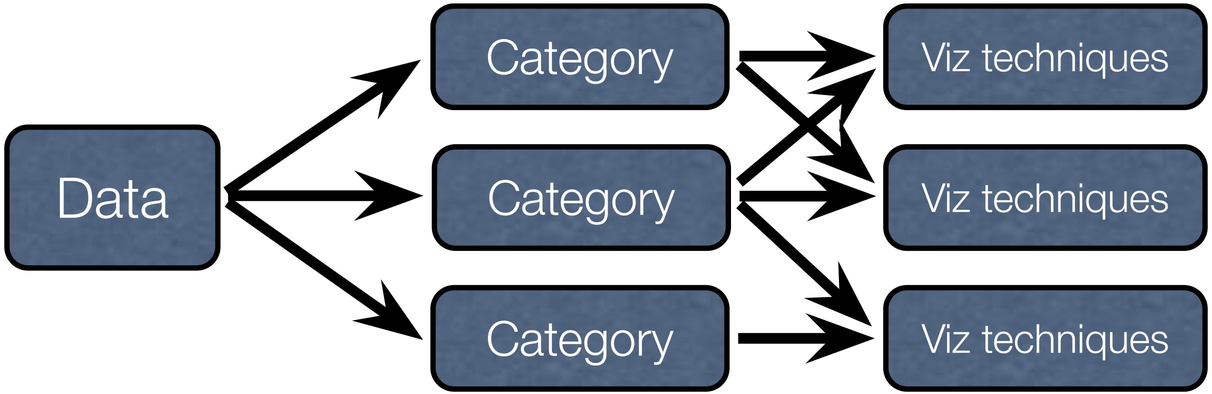
…

What are the other data types?

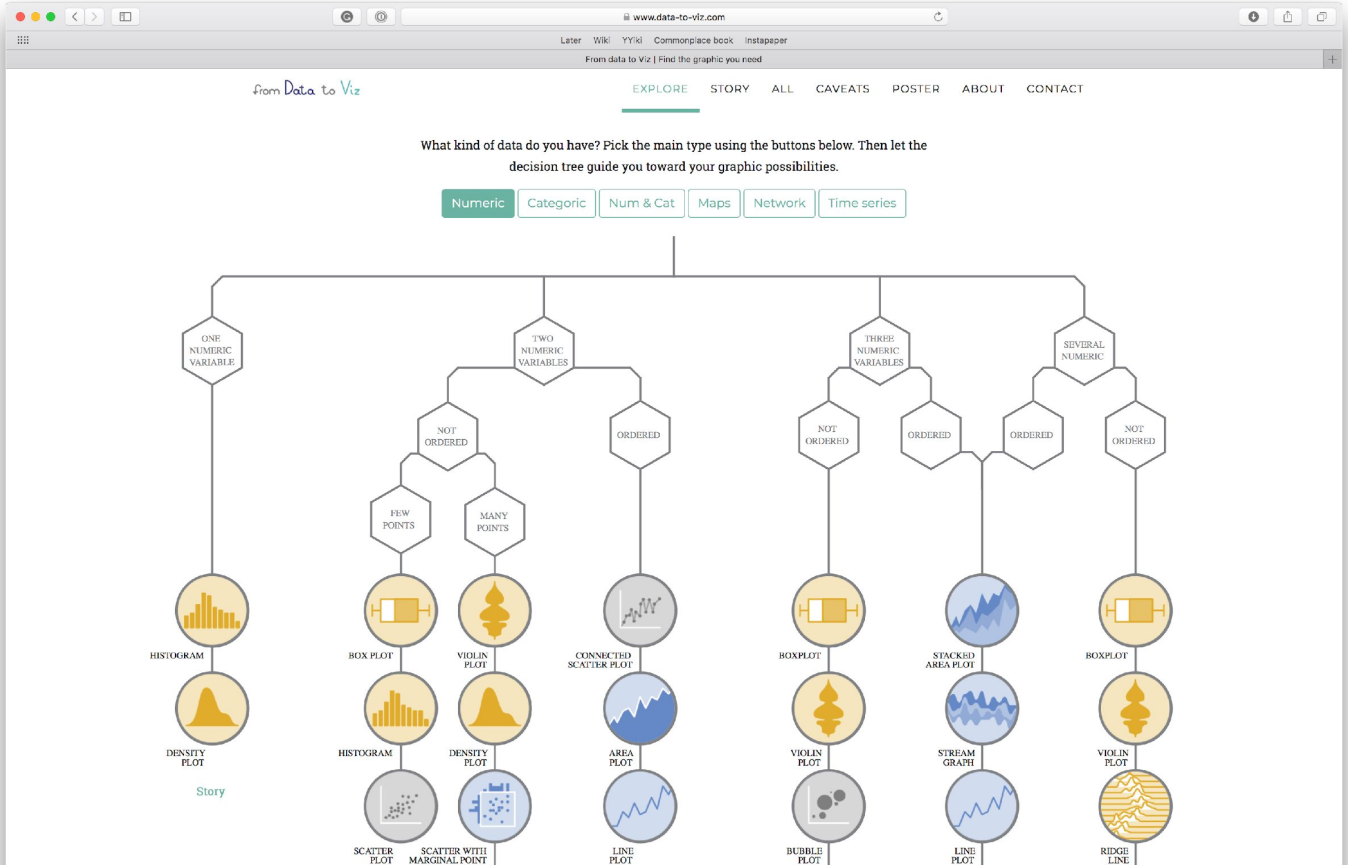
Why should we care about data types?

Examples: zipcode (some starting from zero), long patient IDs that should be read as a string not a number, …

Data types are closely linked to visualization/analysis techniques that you can apply.



https://www.data-to-viz.com

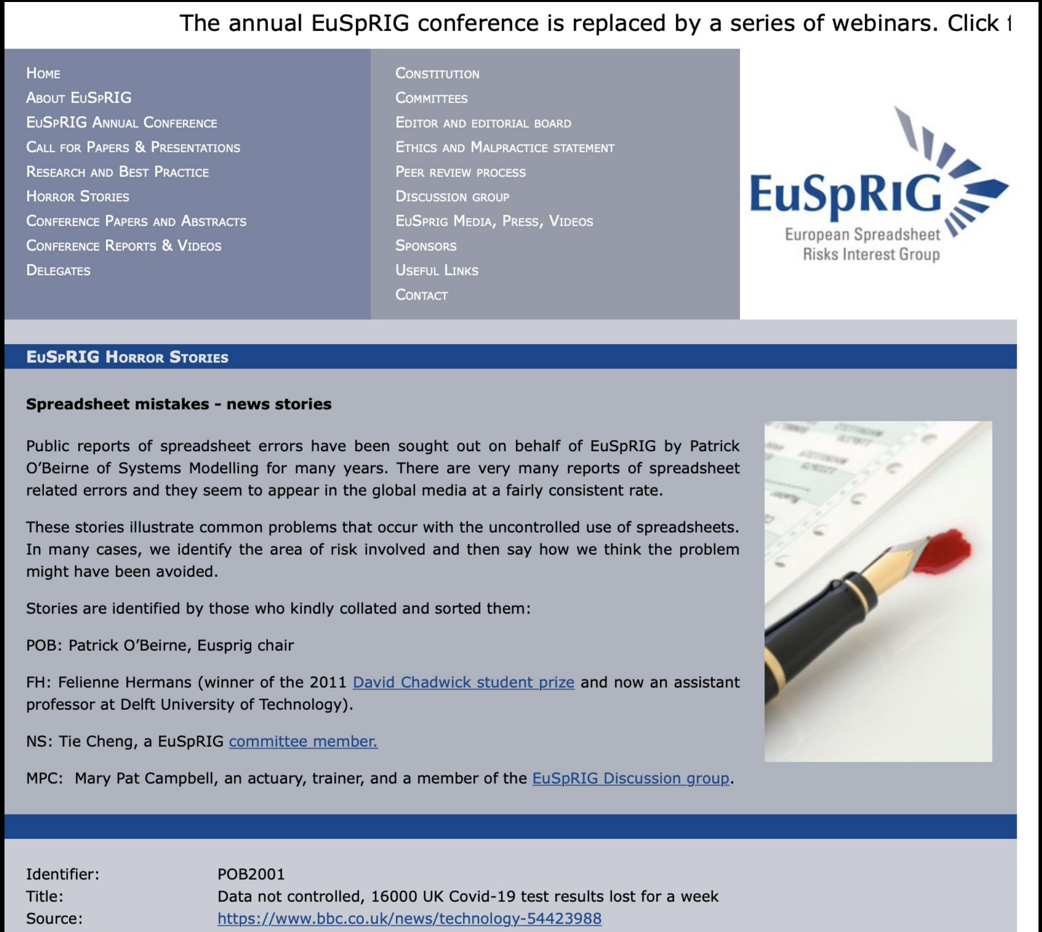


What kind of data do you have? Pick the main type using the buttons below. Then let the decision tree guide you toward your graphic possibilities.

Spreadsheets (Excel) considered harmful

EuSpRIG
European Spreadsheet
Risks Interest Group

http://www.eusprig.org/horror-stories.htm



The annual EuSpRIG conference is replaced by a series of webinars. Click t
Home
About EuSpRIG
EuSpRIG Annual Conference
Call for Papers & Presentations
Research and Best Practice
Horror Stories
Conference Papers and Abstracts
Conference Reports & Videos
Delegates
Constitution
Committees
Editor and editorial board
Ethics and Malpractice statement
Peer review process
Discussion group
EuSprig Media, Press, Videos
Sponsors
Useful Links
Contact
EuSpRIG
European Spreadsheet
Risks Interest Group

EuSpRIG Horror Stories

Spreadsheet mistakes - news stories

Public reports of spreadsheet errors have been sought out on behalf of EuSpRIG by Patrick O'Beirne of Systems Modelling for many years. There are very many reports of spreadsheet related errors and they seem to appear in the global media at a fairly consistent rate.

These stories illustrate common problems that occur with the uncontrolled use of spreadsheets. In many cases, we identify the area of risk involved and then say how we think the problem might have been avoided.

Stories are identified by those who kindly collated and sorted them:

POB: Patrick O'Beirne, Eusprig chair

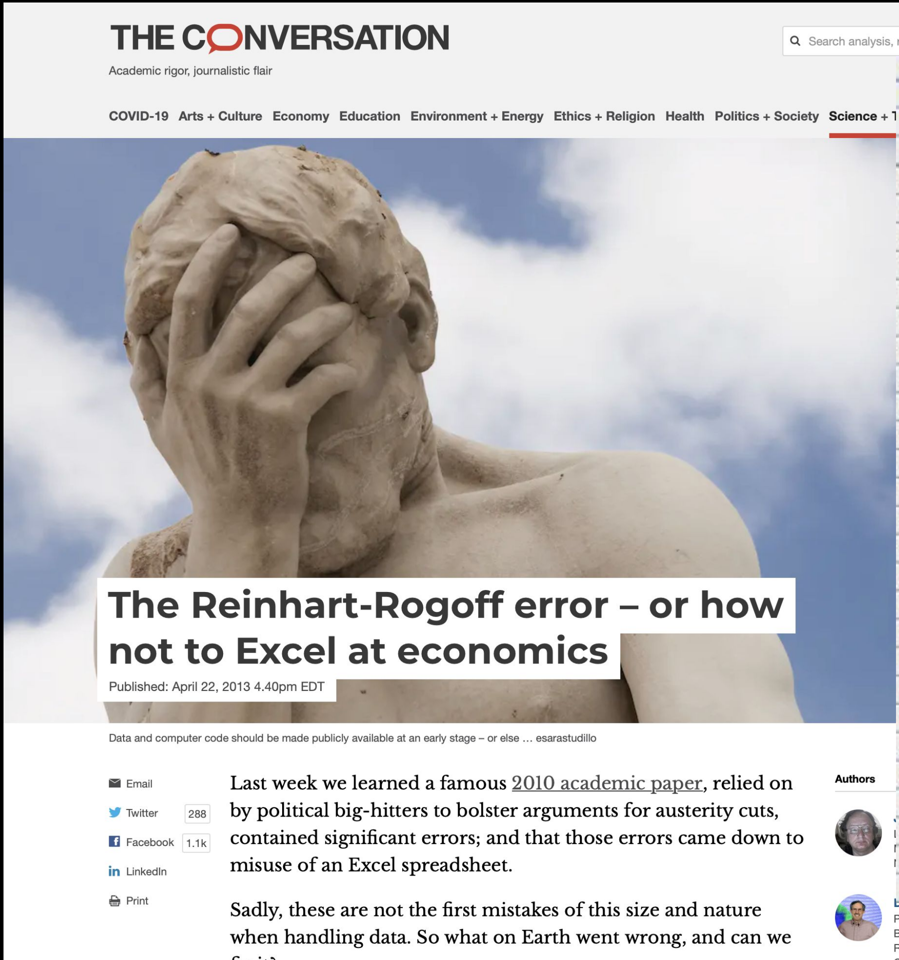
FH: Felienne Hermans (winner of the 2011 David Chadwick student prize and now an assistant professor at Delft University of Technology).

NS: Tie Cheng, a EuSpRIG committee member.

MPC: Mary Pat Campbell, an actuary, trainer, and a member of the EuSpRIG Discussion group.

Identifier:	POB2001
Title:	Data not controlled, 16000 UK Covid-19 test results lost for a week
Source:	https://www.bbc.co.uk/news/technology-54423988

https://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646



THE CONVERSATION

Academic rigor, journalistic flair

Search analysis, research, academics…

COVID-19 Arts + Culture Economy Education Environment + Energy Ethics + Religion Health Politics + Society Science + T

The Reinhart-Rogoff error – or how not to Excel at economics

Published: April 22, 2013 4.40pm EDT

Data and computer code should be made publicly available at an early stage – or else … esarastudillo

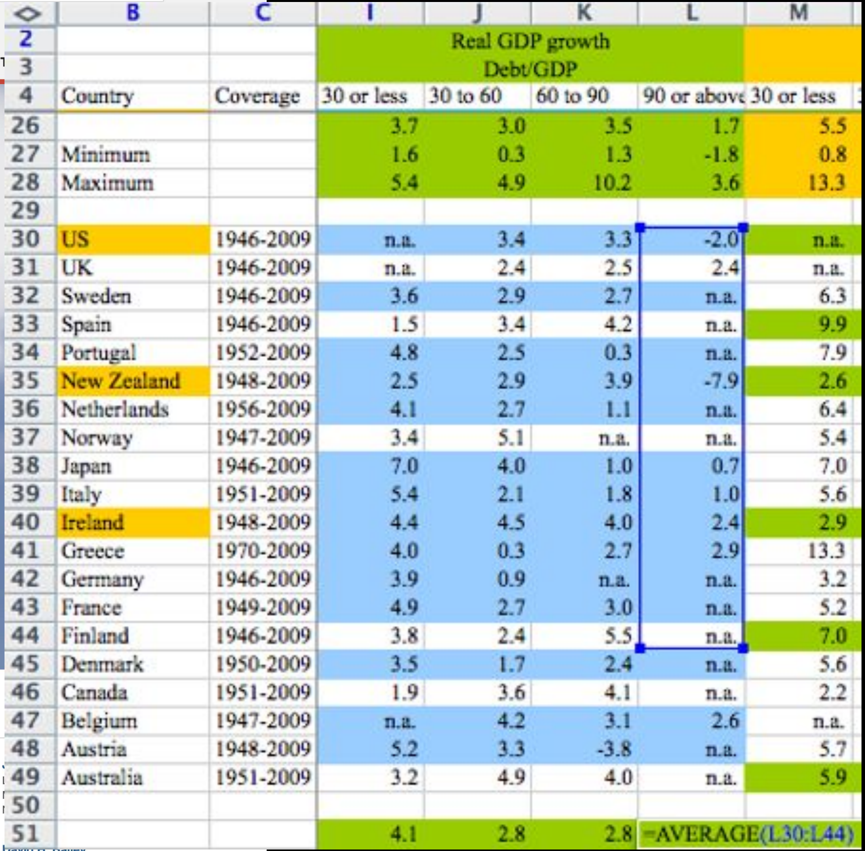
Email
Twitter 288
Facebook 1.1k
LinkedIn
Print

Last week we learned a famous 2010 academic paper, relied on by political big-hitters to bolster arguments for austerity cuts, contained significant errors; and that those errors came down to misuse of an Excel spreadsheet.

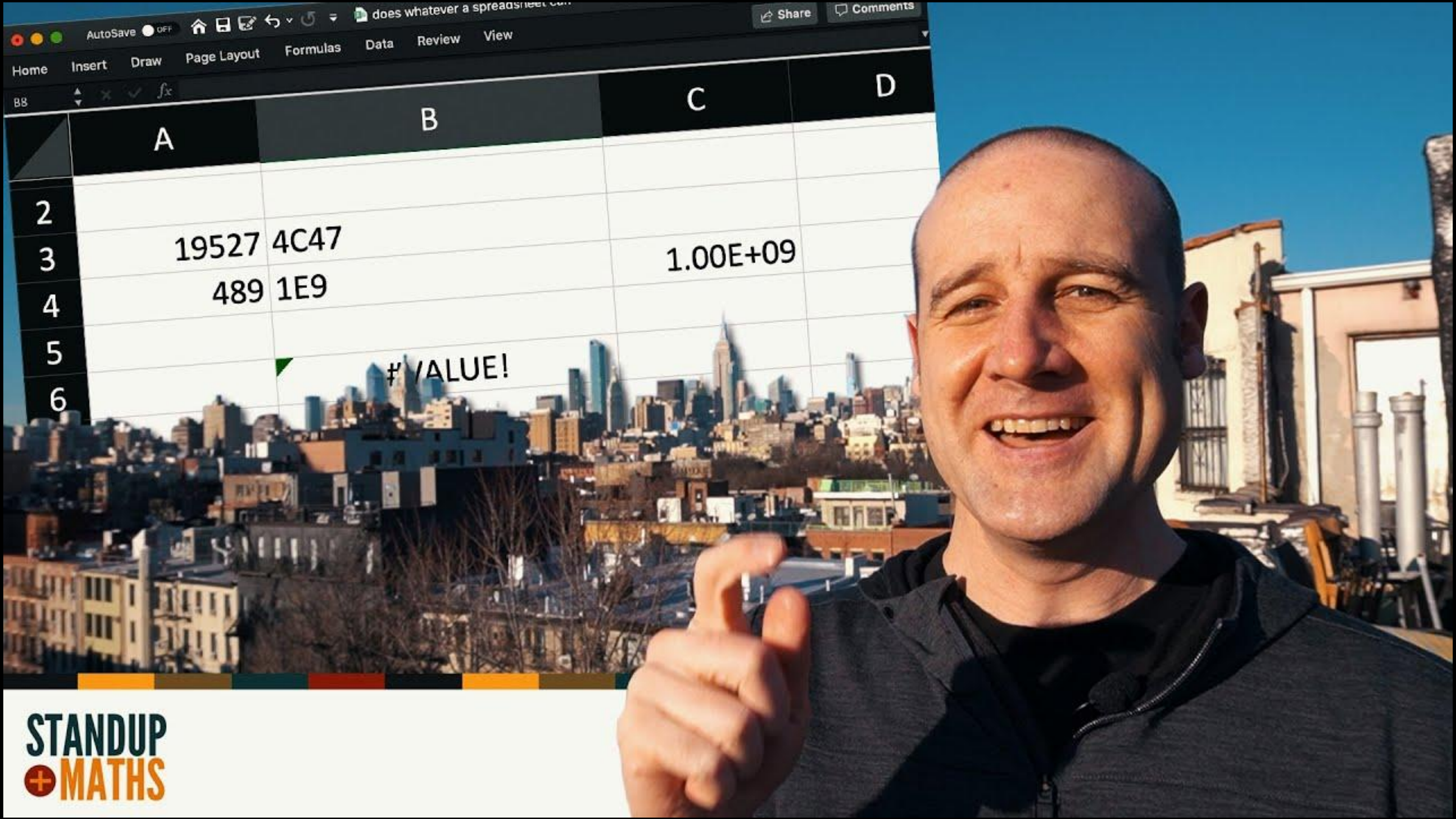
Sadly, these are not the first mistakes of this size and nature when handling data. So what on Earth went wrong, and can we fix it?

Authors

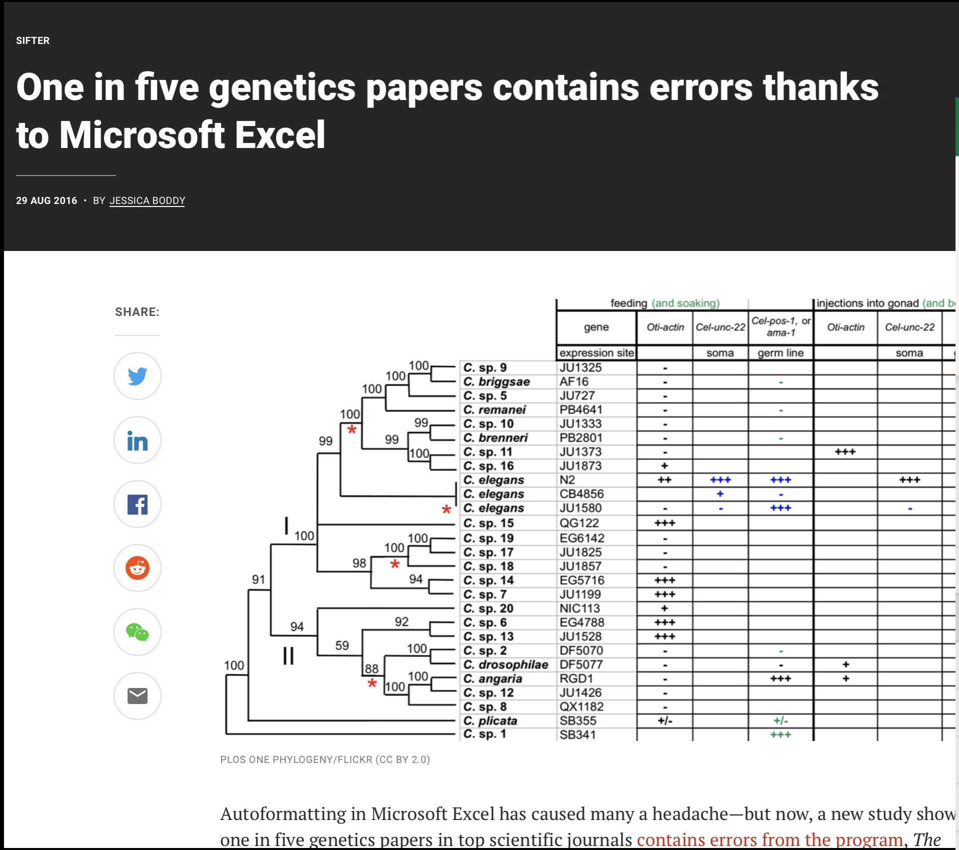
David H. Bailey
PhD; Senior Scientist, Lawrence Berkeley Laboratory (retired) and Research Fellow, University of



	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 abov	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)	


AutoSave OFF
does whatever a spreadsheet can
Share
Comments
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
B8
A
B
C
D
2
3
4
5
6
19527
4C47
1.00E+09
489
1E9
#VALUE!
STANDUP
MATHS

https://www.science.org/content/article/one-five-genetics-papers-contains-errors-thanks-microsoft-excel



SIFTER

One in five genetics papers contains errors thanks to Microsoft Excel

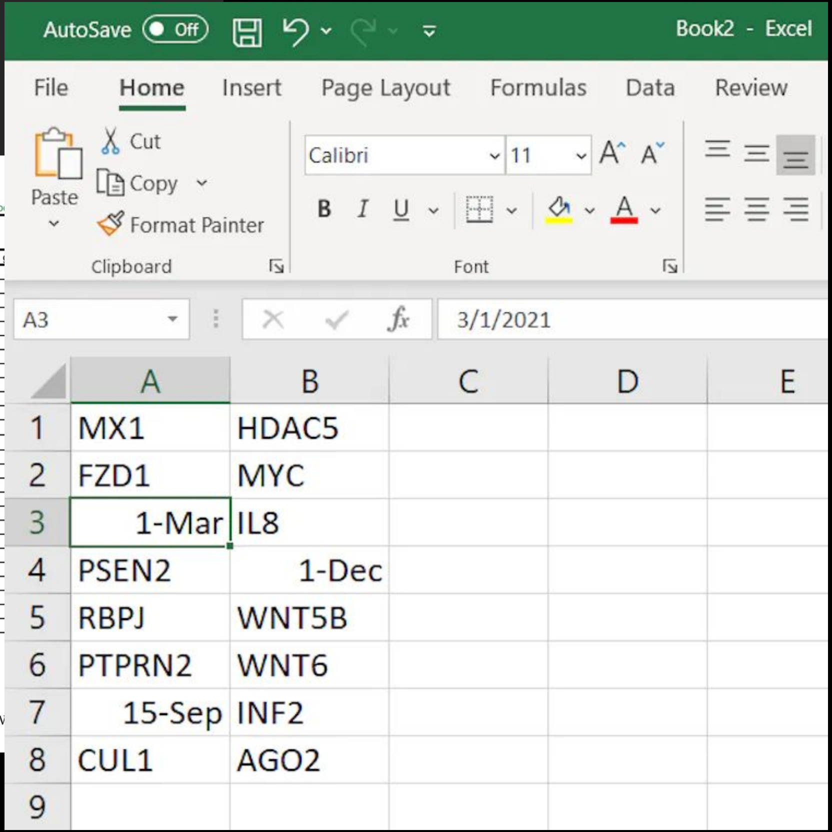
29 AUG 2016 · BY JESSICA BODDY

SHARE:

	feeding (and soaking)			injections into gonad (and b	
gene	Oti-actin	Cel-unc-22	Cel-pos-1, or ama-1	Oti-actin	Cel-unc-22
expression site		soma	germ line		soma
C. sp. 9 JU1325	-				
C. briggsae AF16	-		-		
C. sp. 5 JU727	-				
C. remanei PB4641	-		-		
C. sp. 10 JU1333	-				
C. brenneri PB2801	-		-		
C. sp. 11 JU1373	-			+++	
C. sp. 16 JU1873	+				
C. elegans N2	++	+++	+++		+++
C. elegans CB4856		+	-		
C. elegans JU1580	-	-	+++		-
C. sp. 15 QG122	+++				
C. sp. 19 EG6142	-				
C. sp. 17 JU1825	-				
C. sp. 18 JU1857	-				
C. sp. 14 EG5716	+++				
C. sp. 7 JU1199	+++				
C. sp. 20 NIC113	+				
C. sp. 6 EG4788	+++				
C. sp. 13 JU1528	+++				
C. sp. 2 DF5070	-		-		
C. drosophilae DF5077	-		-	+	
C. angaria RGD1	-		+++	+	
C. sp. 12 JU1426	-				
C. sp. 8 QX1182	-				
C. plicata SB355	+/-		+/-		
C. sp. 1 SB341			+++		

PLOS ONE PHYLOGENY/FLICKR (CC BY 2.0)

Autoformatting in Microsoft Excel has caused many a headache—but now, a new study show one in five genetics papers in top scientific journals contains errors from the program, *The*

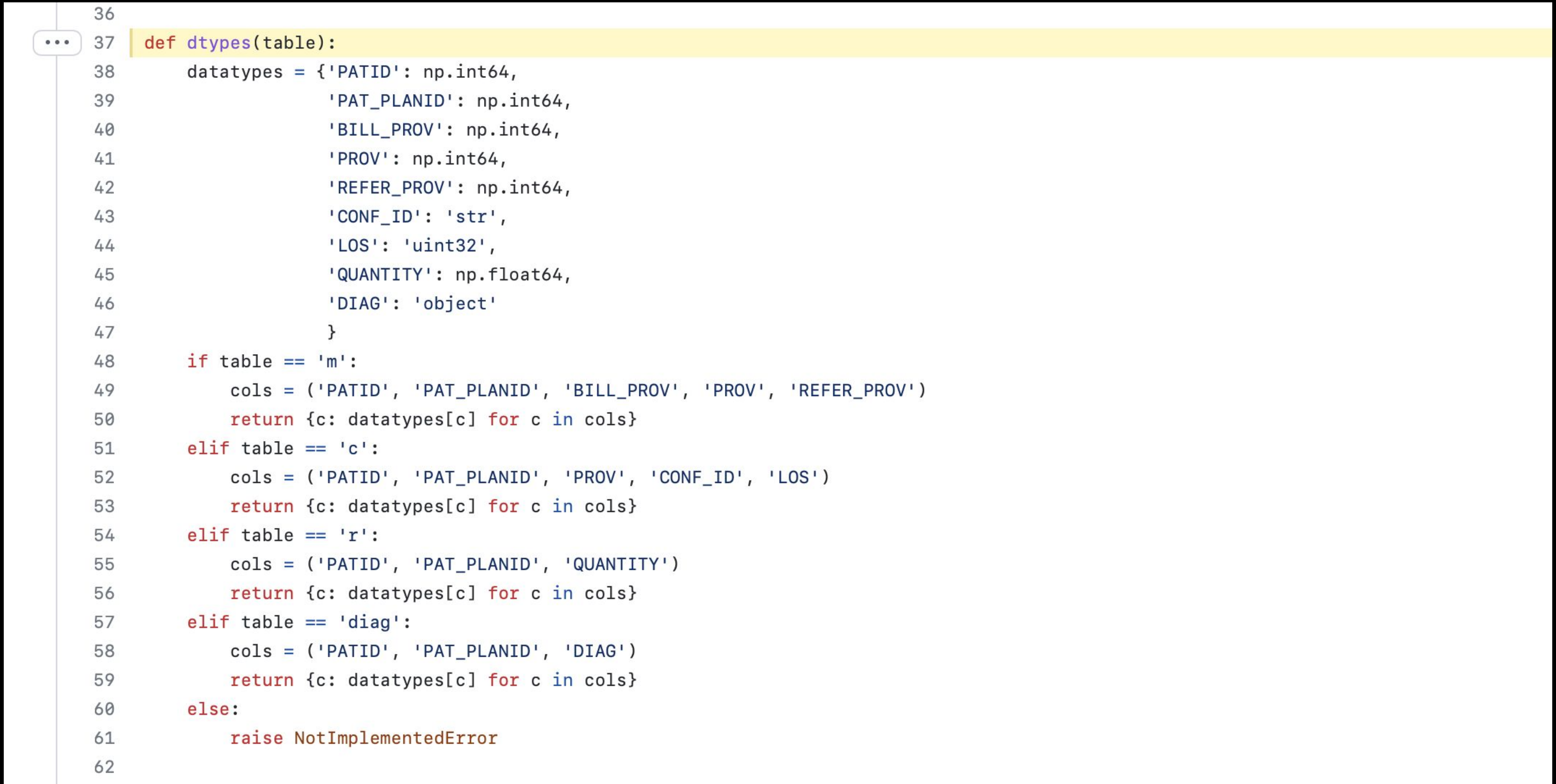


3/1/2021

	A	B	C	D	E
1	MX1	HDAC5			
2	FZD1	MYC			
3	1-Mar	IL8			
4	PSEN2	1-Dec			
5	RBPJ	WNT5B			
6	PTPRN2	WNT6			
7	15-Sep	INF2			
8	CUL1	AGO2			
9					

Data Science 💔 Excel

😭



```python

def dtypes(table):
    datatypes = {'PATID': np.int64,
                 'PAT_PLANID': np.int64,
                 'BILL_PROV': np.int64,
                 'PROV': np.int64,
                 'REFER_PROV': np.int64,
                 'CONF_ID': 'str',
                 'LOS': 'uint32',
                 'QUANTITY': np.float64,
                 'DIAG': 'object'
                 }
    if table == 'm':
        cols = ('PATID', 'PAT_PLANID', 'BILL_PROV', 'PROV', 'REFER_PROV')
        return {c: datatypes[c] for c in cols}
    elif table == 'c':
        cols = ('PATID', 'PAT_PLANID', 'PROV', 'CONF_ID', 'LOS')
        return {c: datatypes[c] for c in cols}
    elif table == 'r':
        cols = ('PATID', 'PAT_PLANID', 'QUANTITY')
        return {c: datatypes[c] for c in cols}
    elif table == 'diag':
        cols = ('PATID', 'PAT_PLANID', 'DIAG')
        return {c: datatypes[c] for c in cols}
    else:
        raise NotImplementedError

```

“Explicit is better than implicit.”