Lecture #22

How to be wrong

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

Slides in this presentation are from material kindly provided by
Shannon Ellis and Brad Voytek

Errors of measurement

Errors of analysis

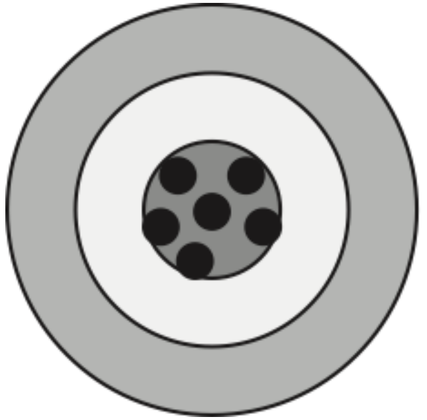
Errors of borked tools

Errors of human cognition

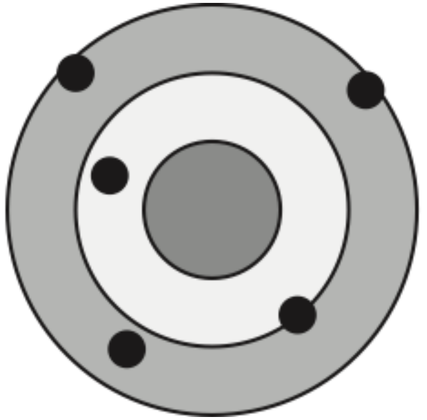
Errors of communication

Errors of measurement

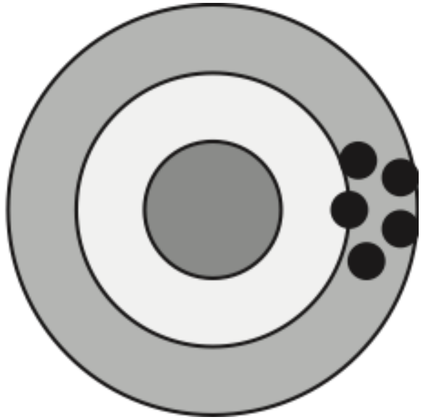
Accurate and precise



Accurate but not precise

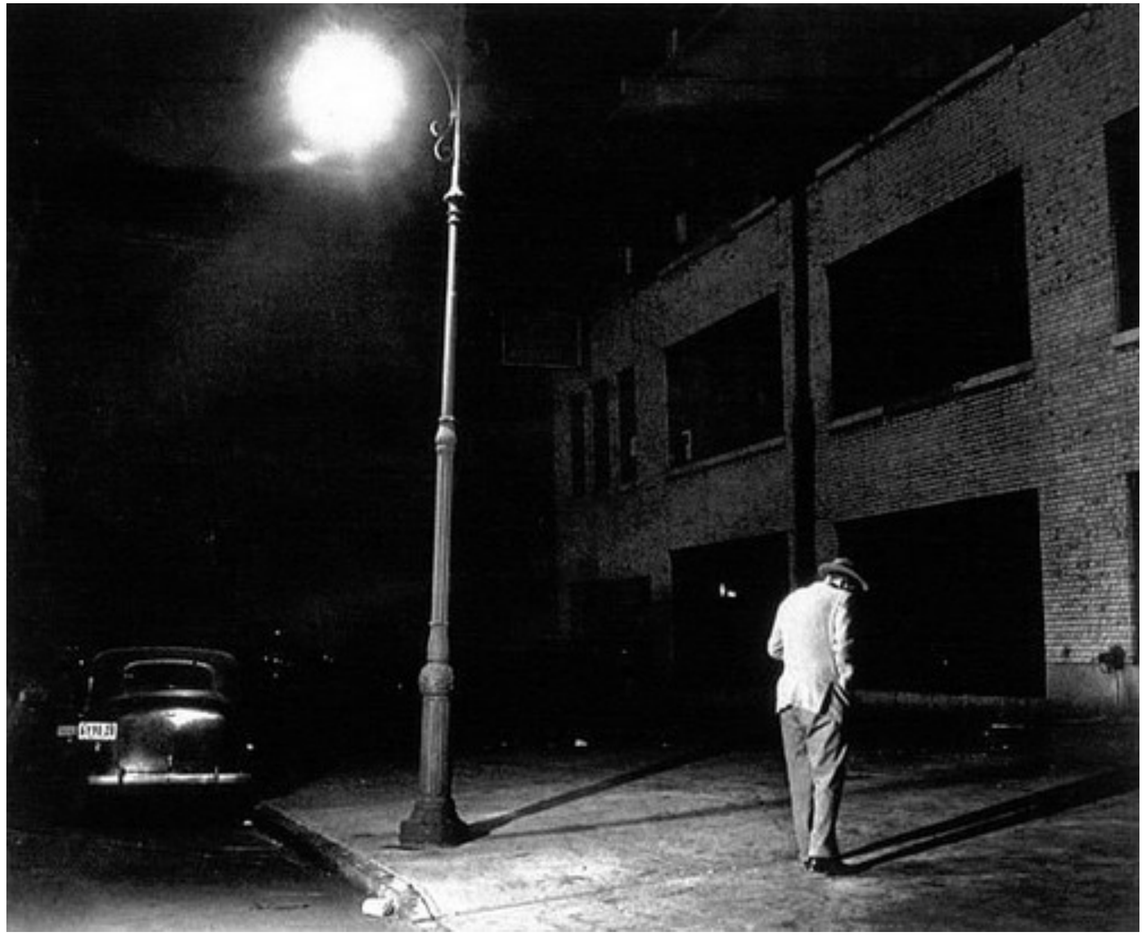


Precise but not accurate



The Lamppost Problem

aka. Streetlight Effect
aka. The Drunkard's Search



Do we understand what we are measuring?

Are we measuring what's relevant?

Academic Life at University of California--San Diego

The student-faculty ratio at University of California--San Diego is 19:1, and the school has 47.4% of its classes with fewer than 20 students. The most popular majors at University of California--San Diego include: Biology, General; Mathematics; Economics; International/Global Studies; and Computer Science. The average freshman retention rate, an indicator of student satisfaction, is 94%.

Class Sizes

Classes with fewer than 20 students **47.4%** 20-49 **25.8%** 50 or more **26.8%**

Student-faculty ratio	**19:1**
4-year graduation rate	**65%**

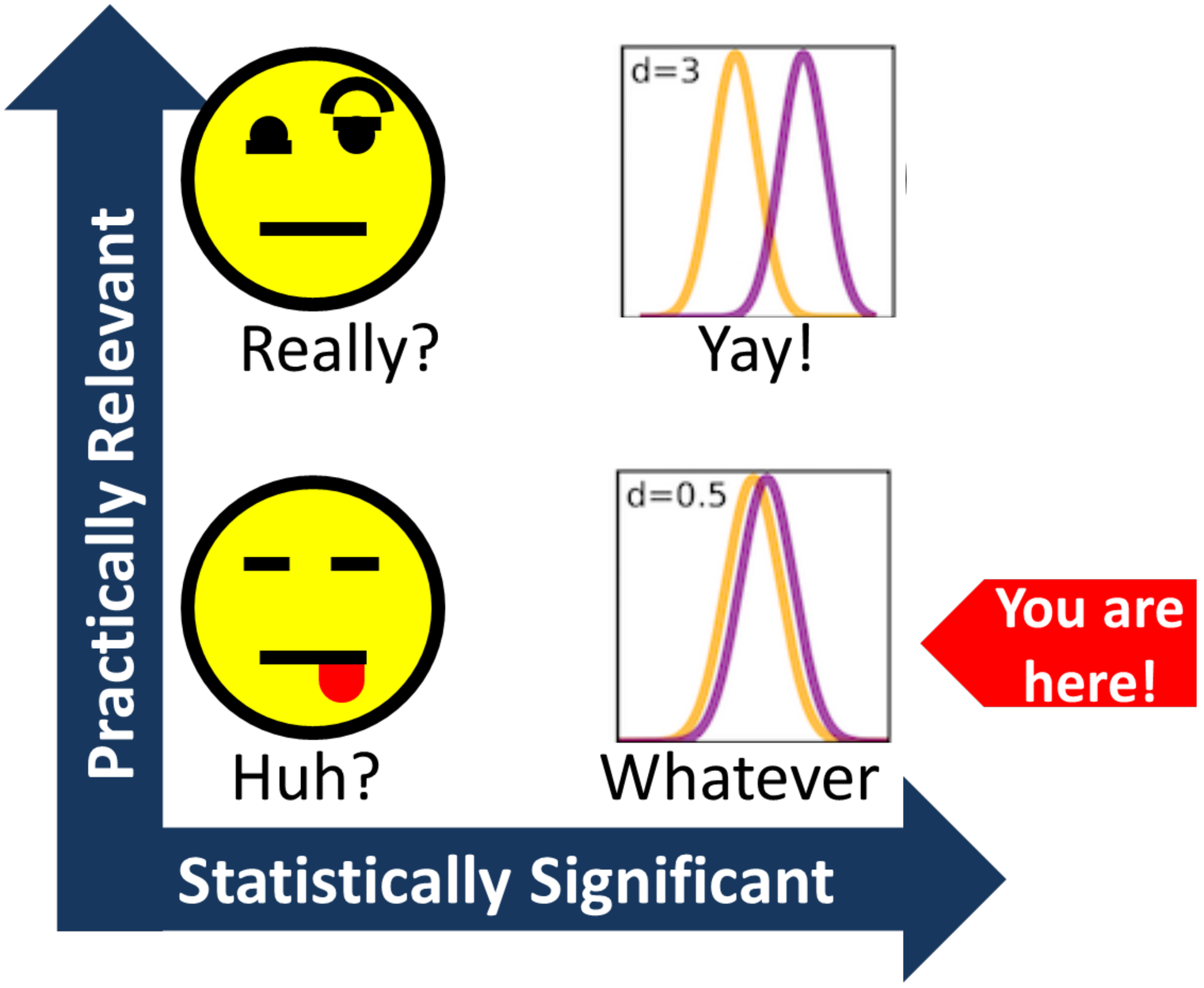
US News & World Report rankings

UCSD median class size vs median student experience

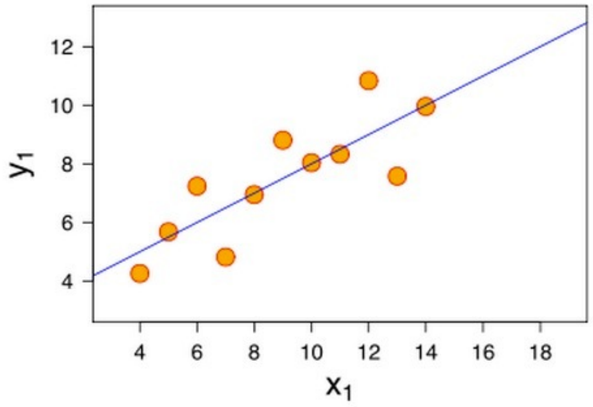
		% of classes with this many students	Cumulative %	Fraction of classes with this many students * min number of students in that class type	% of students in these classes (normalized version of column to the left)	
	2-9 students:	12%	12%	0.24	0.67%	
	10-19 students	32%	44%	3.2	8.95%	
Median class size as experienced by faculty	20-29 students:	14%	58%	2.8	7.83%	
	30-39 students:	8%	66%	2.4	6.72%	
	40-49 students:	4%	70%	1.6	4.48%	
	50-99 students:	11%	81%	5.5	15.39%	
Median class size as experienced by students	Over 100 students:	20%	101%	20	55.96%	
			Sum:	35.74		
	Data from https://www.collegedata.com/college/University-of-California-San-Diego/					

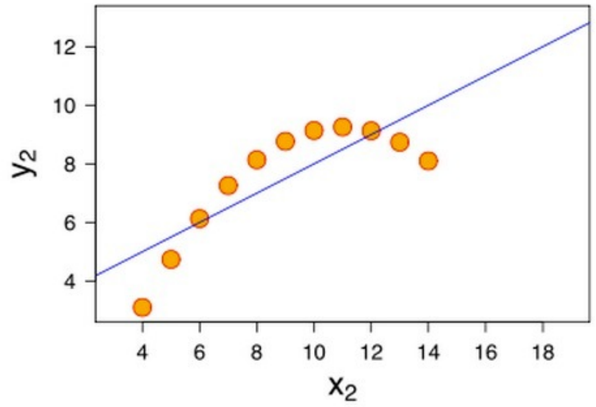

1
Follow
justsaysinmice
@justsaysinmice
The world's hardest working little scientist. Run by @jamesheathers

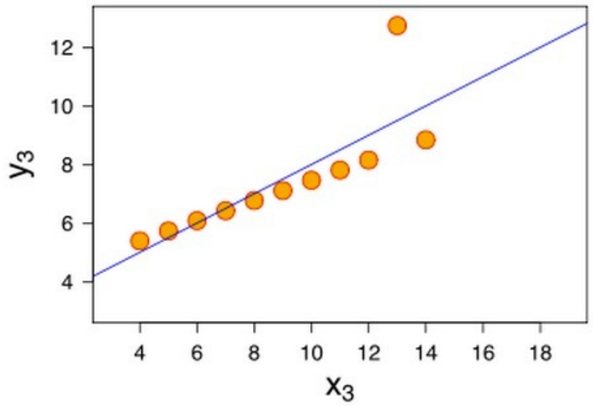
Errors of analysis

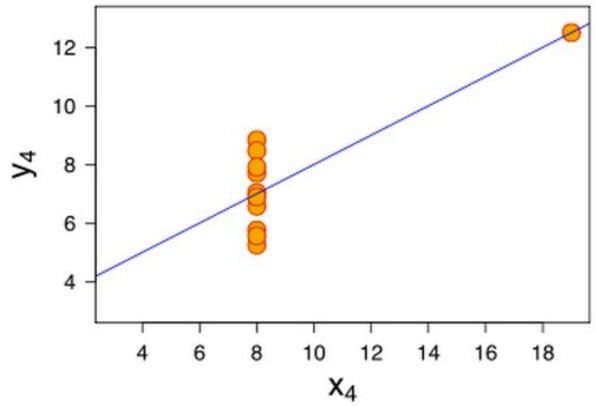

Practically Relevant
Really?
d=3
Yay!
Huh?
d=0.5
Whatever
You are here!
Statistically Significant

Anscombe's quartet









Property	Value
Mean of *x* in each case	9 (exact)
Variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

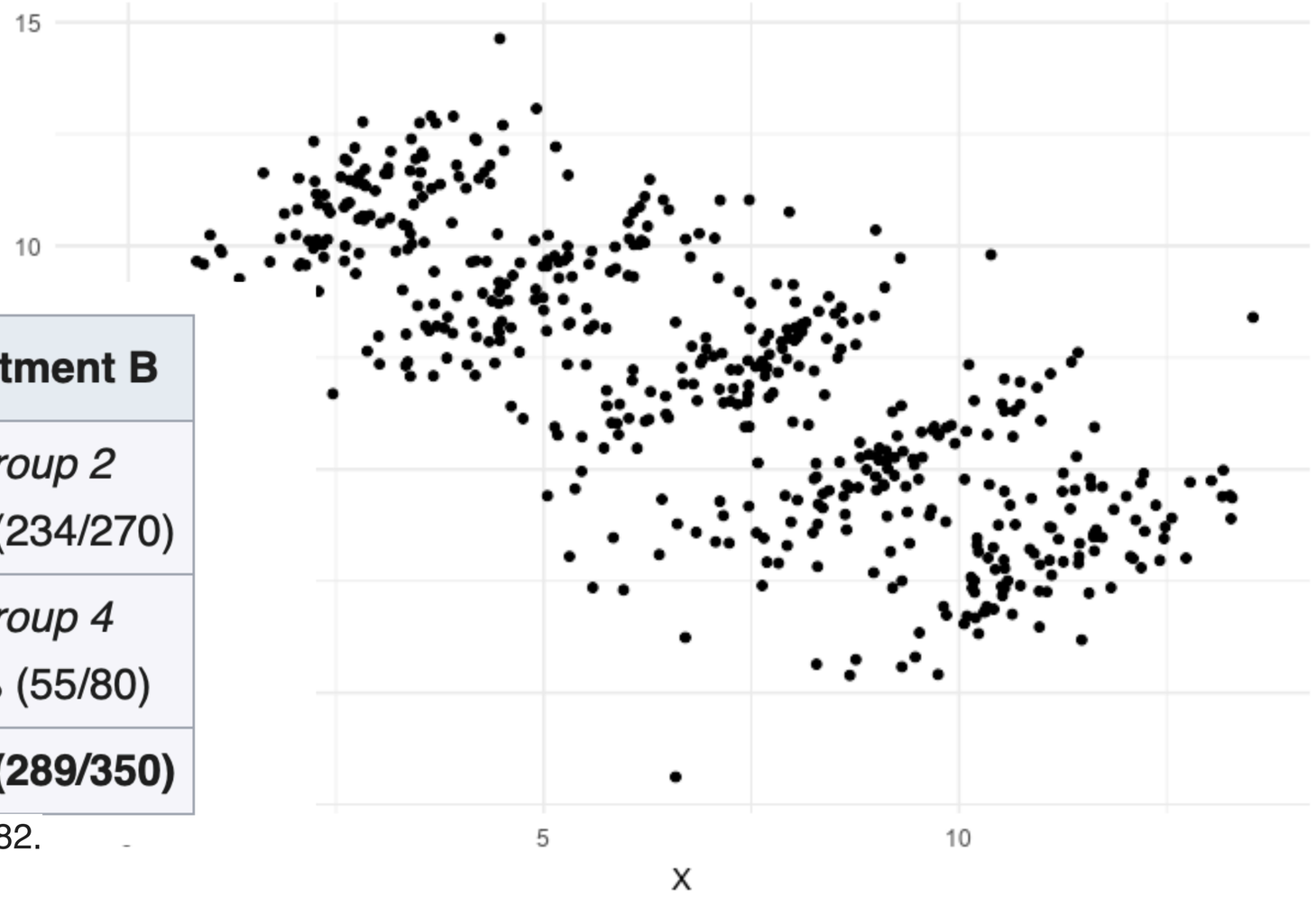
Simpson’s paradox

Answers change when you include subgroups in your analysis

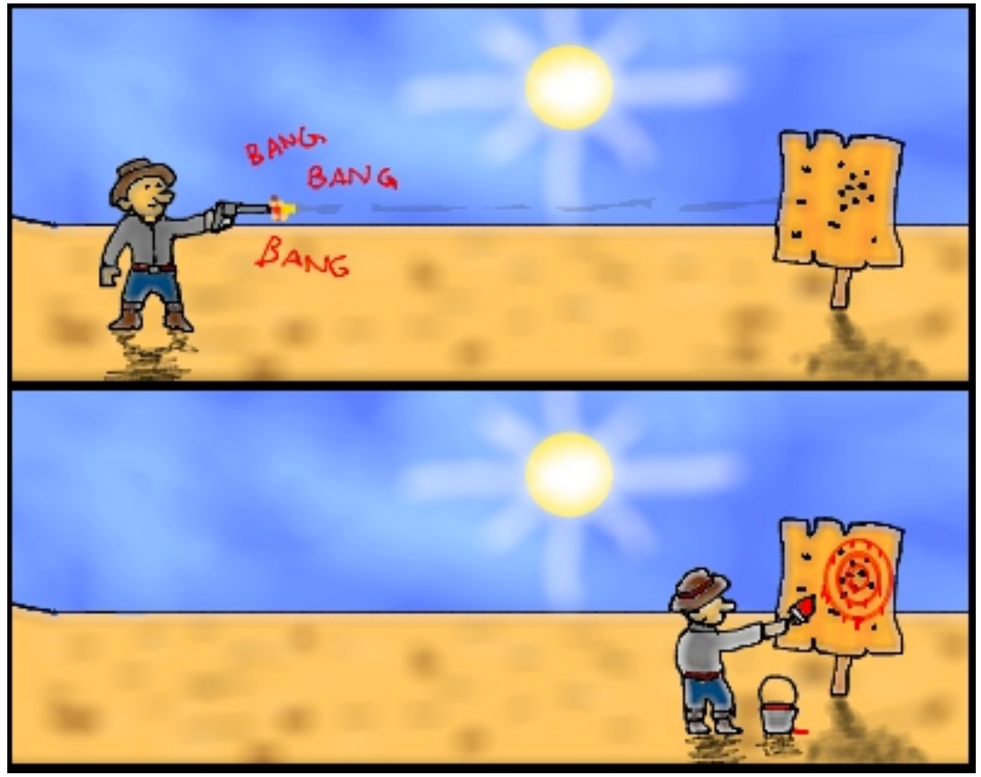
Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Br Med J (Clin Res Ed). **292** (6524): 879–882.

Correlation:



Bad methods



The Texas Sharpshooter fallacy is characterized by a lack of a specific hypothesis prior to the gathering of data, or the formulation of a hypothesis only after data have already been gathered and examined.

Preregistration and publishing negative results

“Data available upon reasonable request”

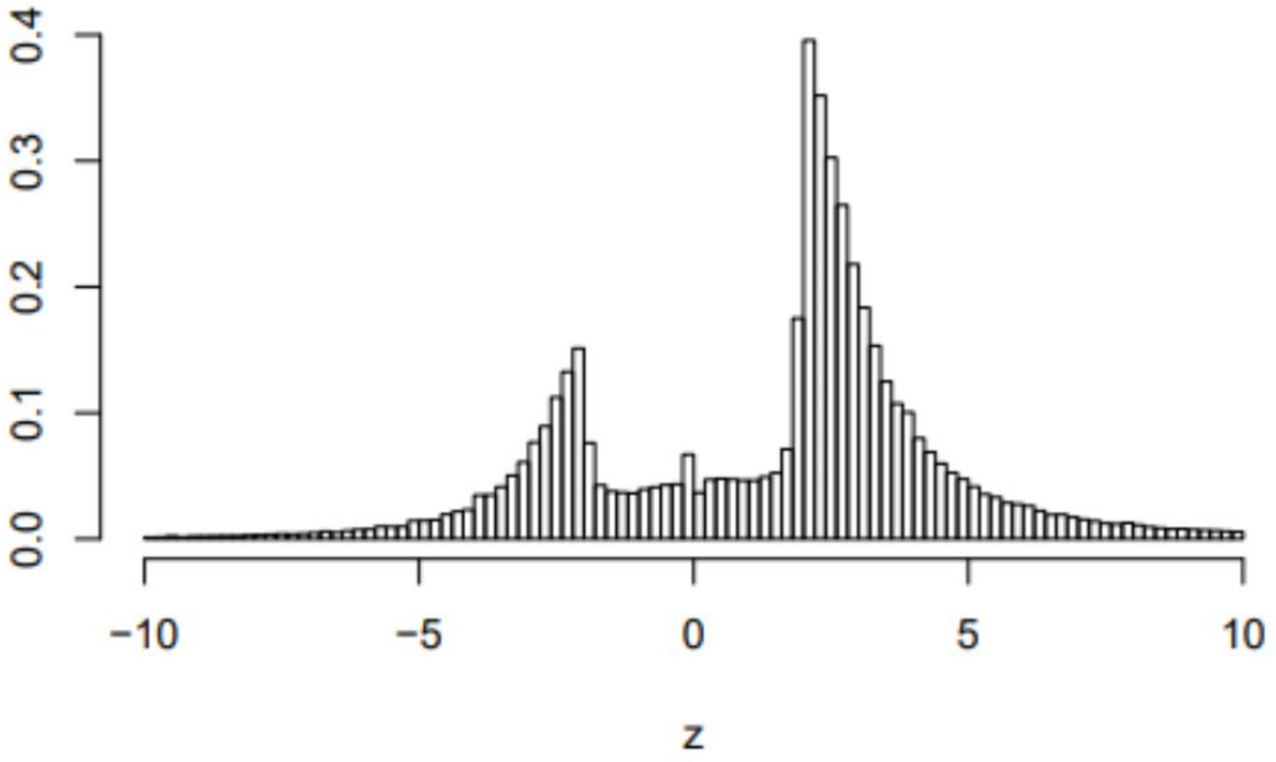


Figure 1: The distribution of more than one million z-values from Medline (1976–2019).

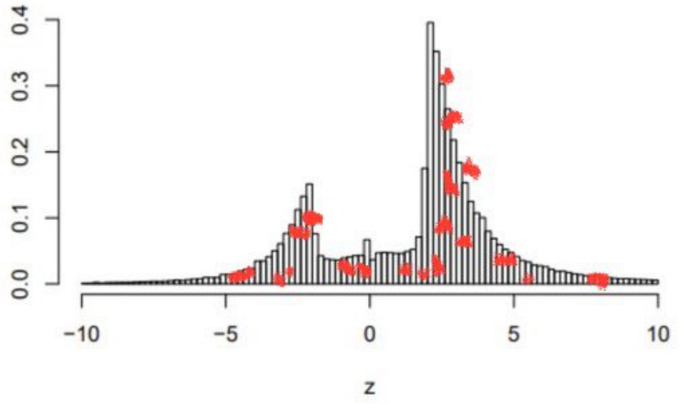
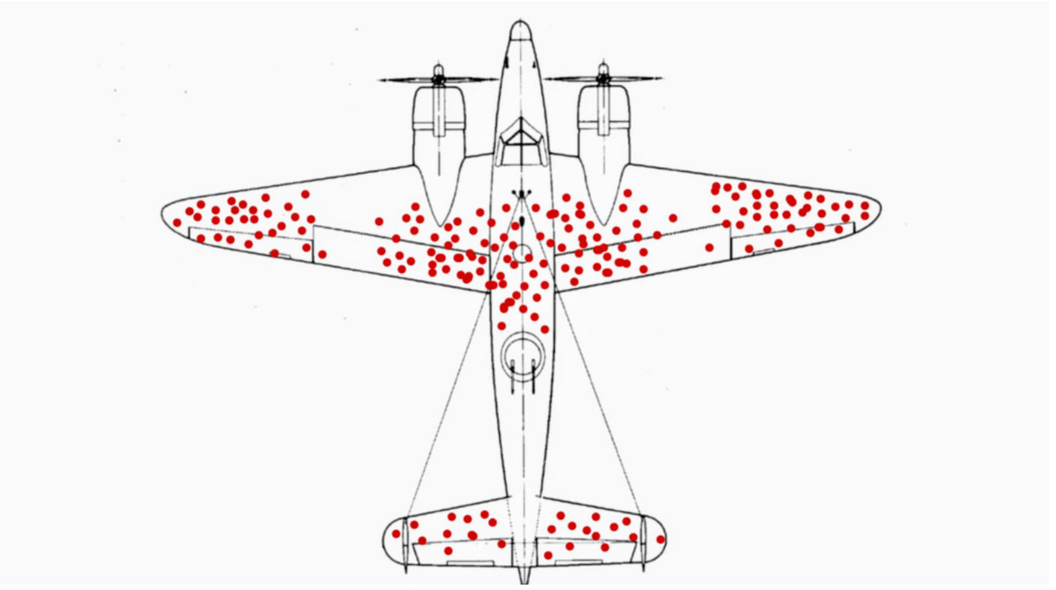
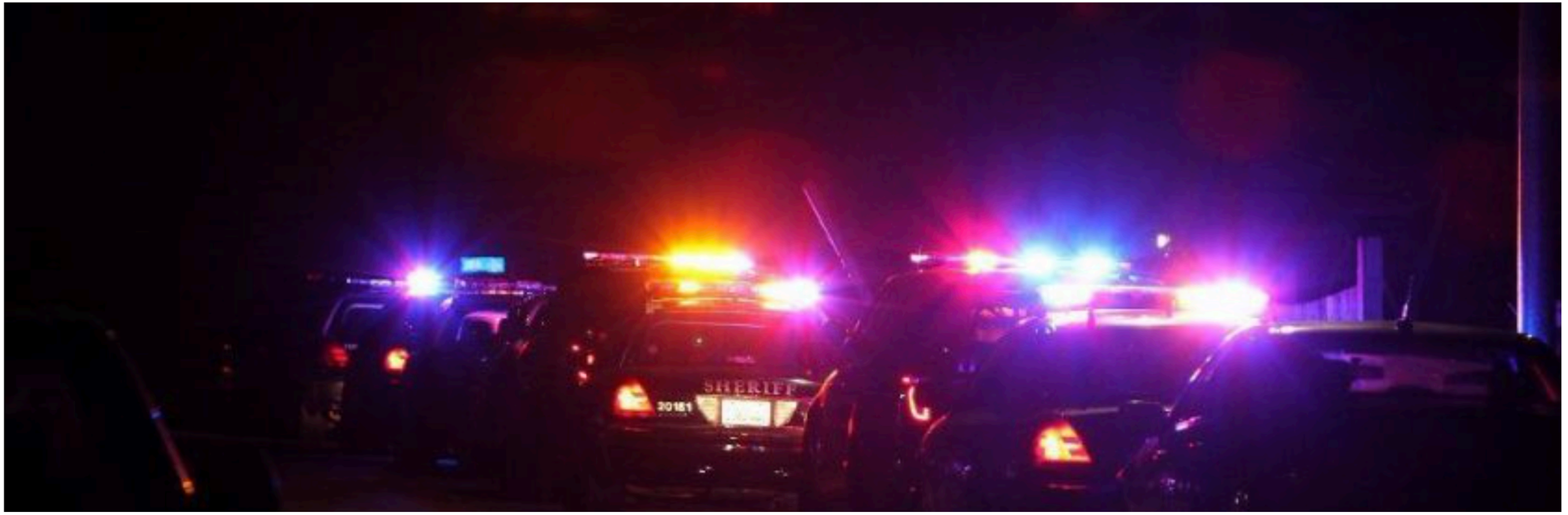


Figure 1: The distribution of more than one million z-values from Medline (1976–2019).



WATCHDOG

Faulty test leads Sheriff's Department to mistakenly claim no evidence of bias in policing



https://twitter.com/jasongfleischer/status/1282730972178350086?s=20

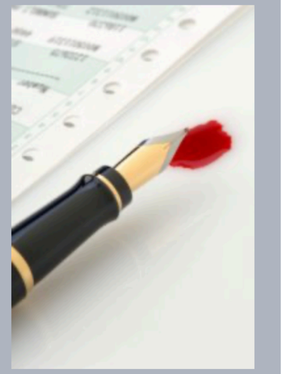
Errors of borked tools


X
Excel

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
Release Date:	08 October 2020
Risk:	Lives put at risk because the contact-tracing process had been delayed
Discrepancy:	16,000 test cases in a week

Excel: Why using Microsoft's tool caused Covid-19 results to be lost
"The badly thought-out use of Microsoft's Excel software was the reason nearly 16,000 coronavirus cases went unreported in England. [The labs] filed their [result logs] results in the form of text-based lists - known as CSV files - without issue. PHE had set up an automatic process to pull this data together into Excel templates so that it could then be uploaded to a central system. The problem is that [Public Health England] PHE's own developers picked an old file format to do this - known as XLS. As a consequence, each template could handle only about 65,000 rows of data rather than the one million-plus rows that Excel is actually capable of. And since each test result created several rows of data, in practice it meant that each template was limited to about 1,400 cases. When that total was reached, further cases were simply left off. To handle the problem, PHE is now

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

MICROSOFT REPORT SCIENCE

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

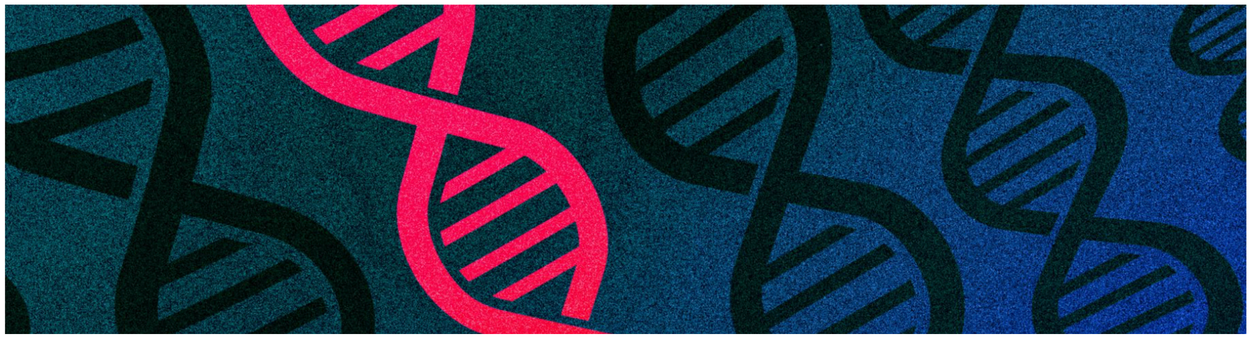
99

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

Listen to this article

SHARE



https://twitter.com/jasongfleischer/status/1291898958747299840?s=20

What if Excel is used as intended?

15k spreadsheets

97M cells

20M formulas

Enron's Spreadsheets and Related Emails: A Dataset and Analysis

Felienne Hermans
Delft University of Technology
Mekelweg 4
2628 CD Delft, the Netherlands
f.f.j.hermans@tudelft.nl

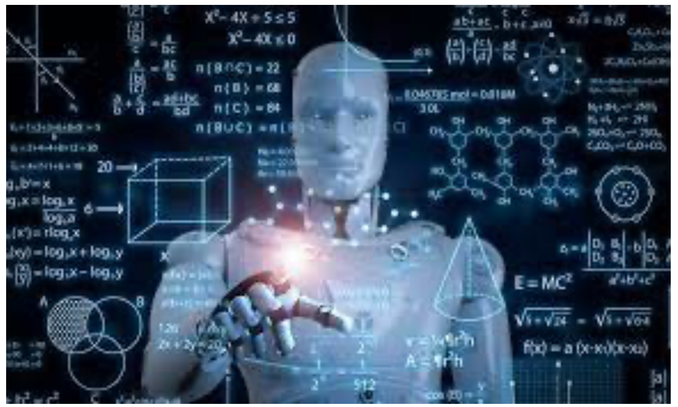
Emerson Murphy-Hill
North Carolina State University
890 Oval Drive
Raleigh, North Carolina, USA
emerson@csc.ncsu.edu

TABLE III
SPREADSHEETS CONTAINING EXCEL ERRORS IN THE ENRON SET

Error type	Spreadsheets	Formulas	Unique Ones
#DIV/0!	580	76,656	4,779
#N/A	635	948,194	6,842
#NAME?	297	33,9365	29,422
#NUM!	52	4,087	178
#REF!	931	18,3014	6824
#VALUE!	423	11,1024	1751
Total	2,205	1,662,340	49,796

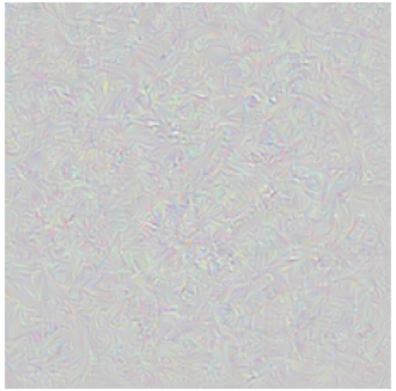
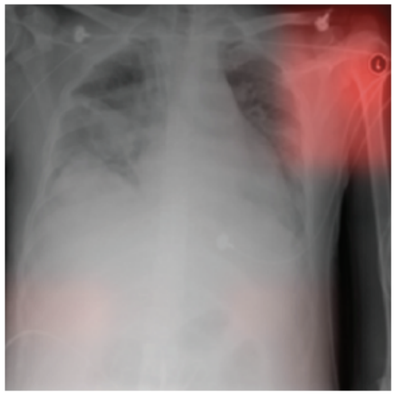
24% of spreadsheets with formulas had errors!

Excel



Learning the irrelevant

ML loves the shortcut

				**Article:** Super Bowl 50 **Paragraph:** "*Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.*" **Question:** "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*" **Original Prediction:** John Elway **Prediction under adversary:** Jeff Dean
Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

https://thegradient.pub/shortcuts-neural-networks-love-to-cheat/

Errors of human cognition

COGNITIVE BIAS CODEX

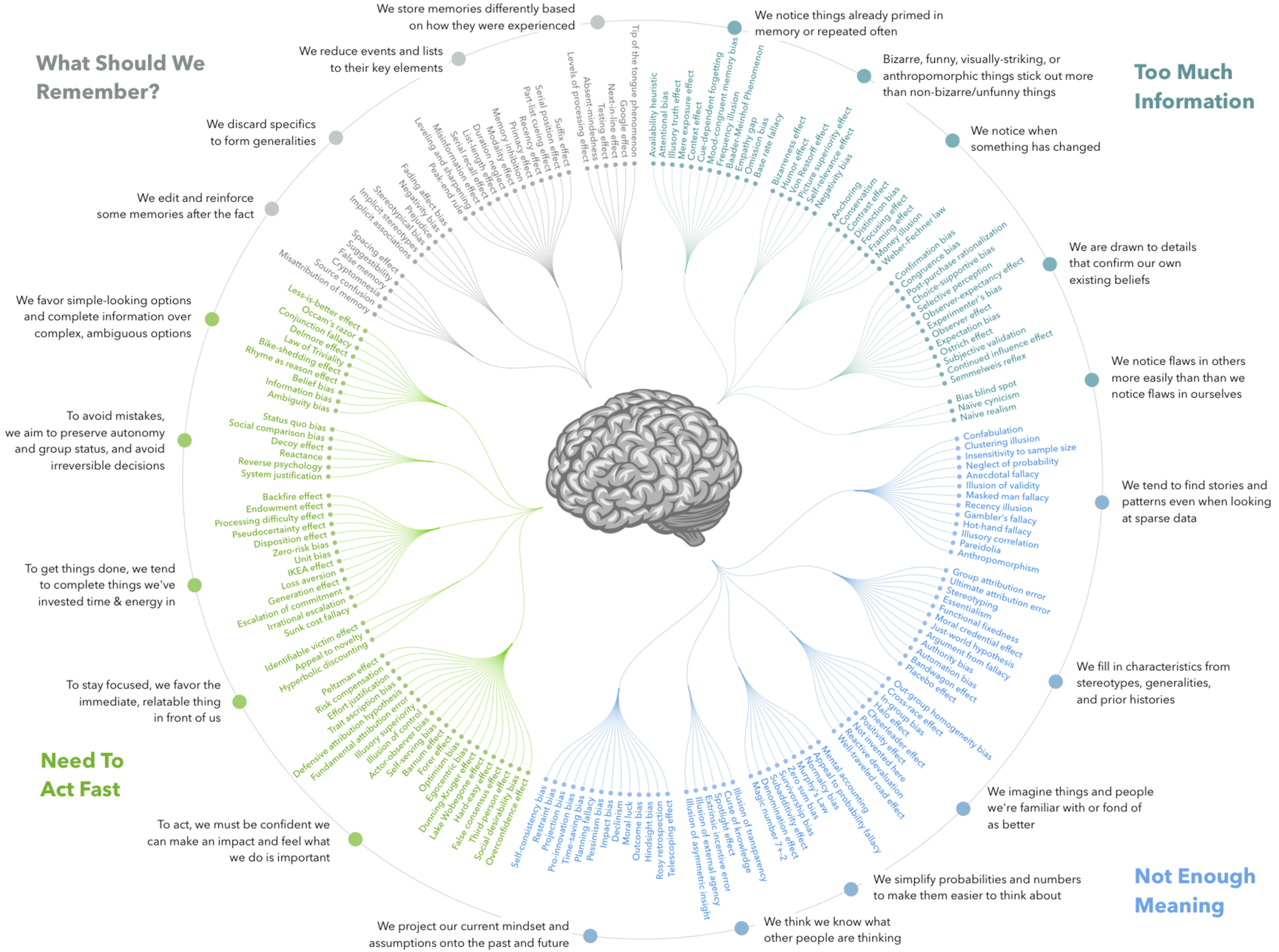
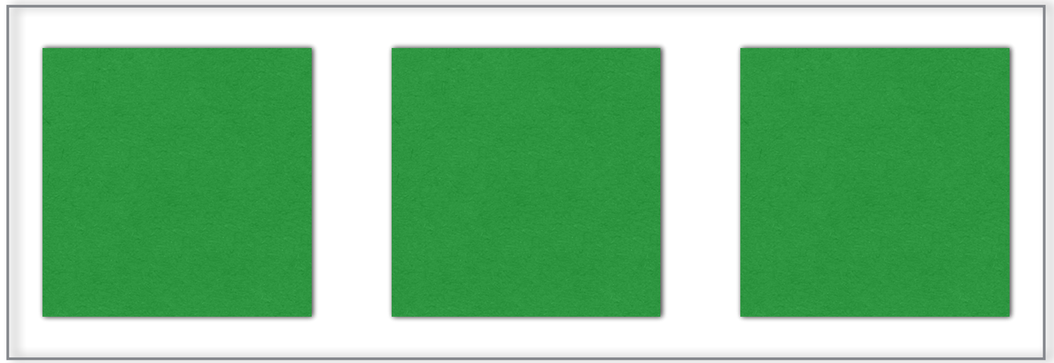
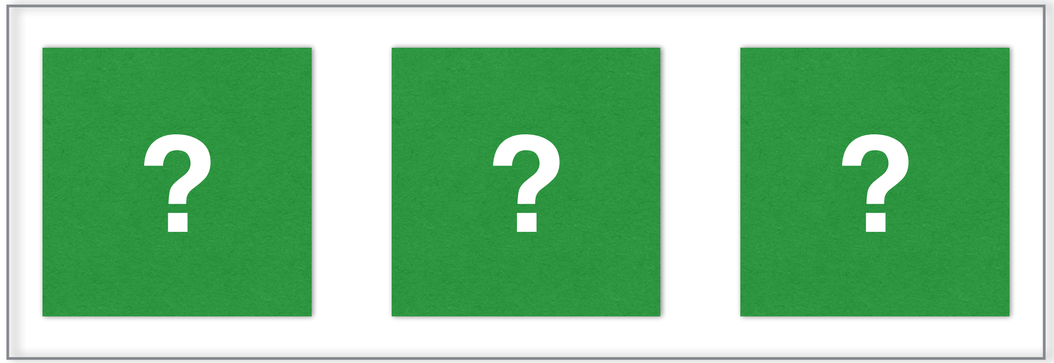


Figure out the rule





I T I V E B I A S C O D E X

ently based
xperienced

Tip of the tongue phenomenon
Google effect
Next-in-line effect
Testing effect
Absent-mindedness
Levels of processing effect
Suffix effect
Serial position effect
Part-list cueing effect
Recency effect
Primacy effect

We notice things already primed in memory or repeated often

Availability heuristic
Attentional bias
Illusory truth effect
Mere exposure effect
Context effect
Cue-dependent forgetting
Mood-congruent memory bias
Frequency illusion
Baader-Meinhof Phenomenon
Empathy gap
Omission bias
Base rate fallacy

Bizarre, funny, visually-striking, or anthropomorphic things stick out more than non-bizarre/unfunny things

Bizarreness effect
Humor effect
Von Restorff effect
Picture superiority effect
Self-relevance effect
Negativity bias

We notice when something has changed

Anchoring
Conservatism
Contrast effect
Distinction bias
Focusing effect
Framing effect
Money illusion
Weber-Fechner law

Too Much Information

We are drawn to details that confirm our own existing beliefs

Confirmation bias
Congruence bias
Post-purchase rationalization
Choice-supportive bias
Selective perception
Observer-expectancy effect
Experimenter's bias
Observer effect
Expectation bias
Ostrich effect
Subjective validation
Continued influence effect
Semmelweis reflex

We notice flaws in others more easily than than we notice flaws in ourselves

Bias blind spot
Naïve cynicism
Naïve realism

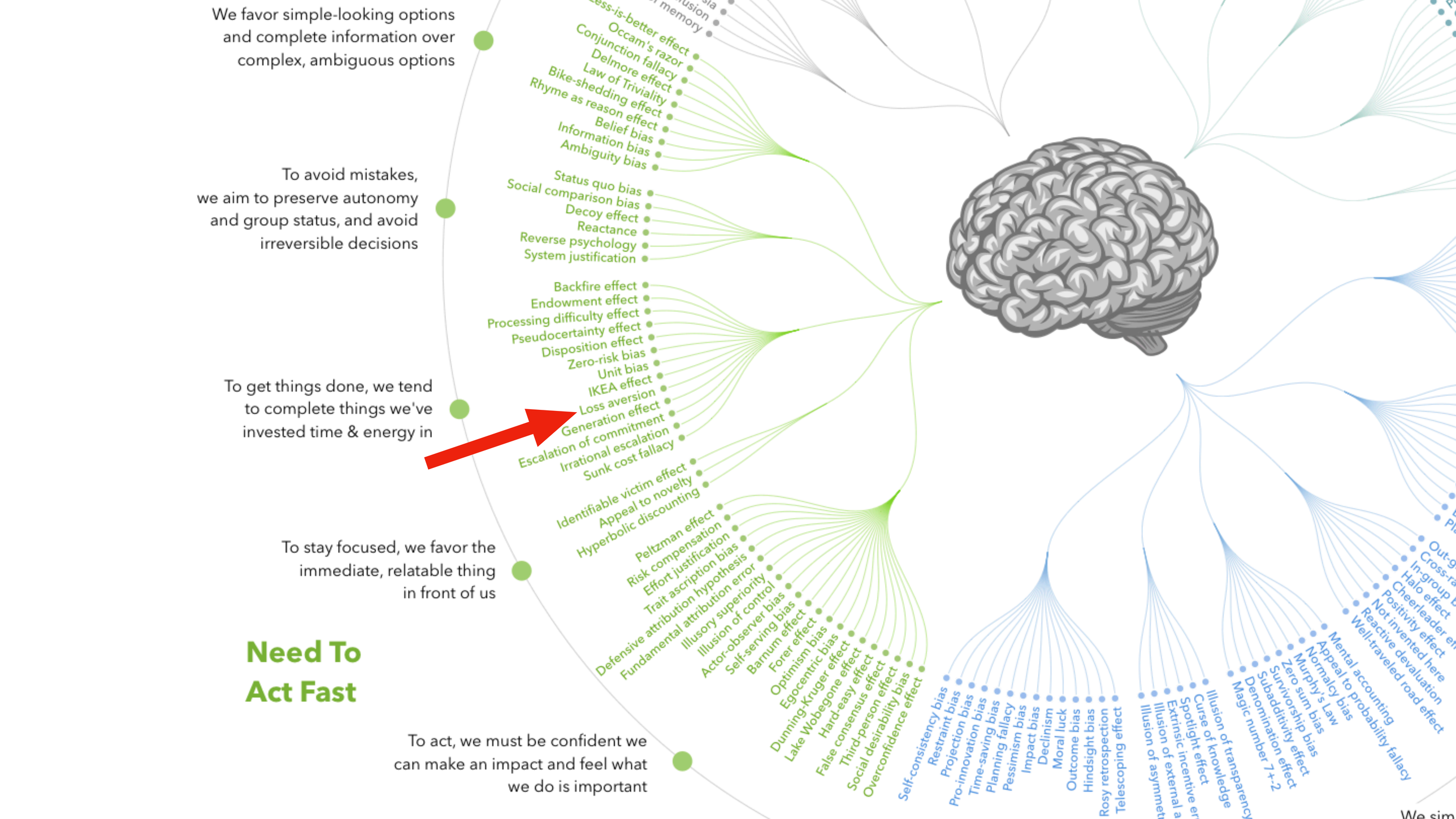
We tend to find stories and patterns even when looking

Confabulation
Clustering illusion
Insensitivity to sample size
Neglect of probability
Anecdotal fallacy
Illusion of validity
Masked man fallacy
Recency illusion

Confirmation bias

The only way to show you're right is to try to show you're wrong!

Confirmation bias…
and programming?


We favor simple-looking options and complete information over complex, ambiguous options
To avoid mistakes, we aim to preserve autonomy and group status, and avoid irreversible decisions
To get things done, we tend to complete things we've invested time & energy in
To stay focused, we favor the immediate, relatable thing in front of us
Need To Act Fast
To act, we must be confident we can make an impact and feel what we do is important
Less-is-better effect
Occam's razor
Conjunction fallacy
Delmore effect
Law of Triviality
Bike-shedding effect
Rhyme as reason effect
Belief bias
Information bias
Ambiguity bias
Status quo bias
Social comparison bias
Decoy effect
Reactance
Reverse psychology
System justification
Backfire effect
Endowment effect
Processing difficulty effect
Pseudocertainty effect
Disposition effect
Zero-risk bias
Unit bias
IKEA effect
Loss aversion
Generation effect
Escalation of commitment
Irrational escalation
Sunk cost fallacy
Identifiable victim effect
Appeal to novelty
Hyperbolic discounting
Peltzman effect
Risk compensation
Effort justification
Trait ascription bias
Defensive attribution hypothesis
Fundamental attribution error
Illusory superiority
Illusion of control
Actor-observer bias
Self-serving bias
Barnum effect
Forer effect
Optimism bias
Egocentric bias
Dunning-Kruger effect
Lake Wobegone effect
Hard-easy effect
False consensus effect
Third-person effect
Social desirability bias
Overconfidence effect
Self-consistency bias
Restraint bias
Projection bias
Pro-innovation bias
Time-saving bias
Planning fallacy
Pessimism bias
Impact bias
Declinism
Moral luck
Outcome bias
Hindsight bias
Rosy retrospection
Telescoping effect
Illusion of asymmetric
Illusion of external a
Extrinsic incentive er
Spotlight effect
Curse of knowledge
Illusion of transparency
Magic number 7+-2
Denomination effect
Subadditivity effect
Survivorship bias
Zero sum bias
Murphy's Law
Normalcy bias
Appeal to probability fallacy
Mental accounting
Well-traveled road effect
Reactive devaluation
Not invented here
Positivity effect
Halo effect
In-group b
Cross-ra
Out-g

Should I try base jumping? It could be fun!

Loss aversion

If you tell someone they have a 99.5% of surviving they think it's fine

…but tell them 1 in 200 people die and they think it’s too scary

- Bias blind spot
- Naïve cynicism
- Naïve realism

We notice flaws in others more easily than than we notice flaws in ourselves

- Confabulation
- Clustering illusion
- Insensitivity to sample size
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity
- Masked man fallacy
- Recency illusion
- Gambler's fallacy
- Hot-hand fallacy
- Illusory correlation
- Pareidolia
- Anthropomorphism

We tend to find stories and patterns even when looking at sparse data

- Group attribution error
- Ultimate attribution error
- Stereotyping
- Essentialism
- Functional fixedness
- Moral credential effect
- Just-world hypothesis
- Argument from fallacy
- Authority bias
- Automation bias
- Bandwagon effect
- Placebo effect

We fill in characteristics from stereotypes, generalities, and prior histories

- Out-group homogeneity bias
- Cross-race effect
- In-group bias
- Halo effect
- Cheerleader effect
- Positivity effect
- Not invented here
- Reactive devaluation
- Well-traveled road effect

We imagine things and people we're familiar with or fond of as better

- Mental accounting
- Appeal to probability fallacy
- Normalcy bias
- Murphy's Law
- Zero sum bias
- Survivorship bias
- Subadditivity effect
- Denomination effect
- Magic number 7+-2

We simplify probabilities and numbers to make them easier to think about

- Illusion of transparency
- Curse of knowledge
- Spotlight effect
- Extrinsic incentive error
- Illusion of external agency
- Illusion of asymmetric insight

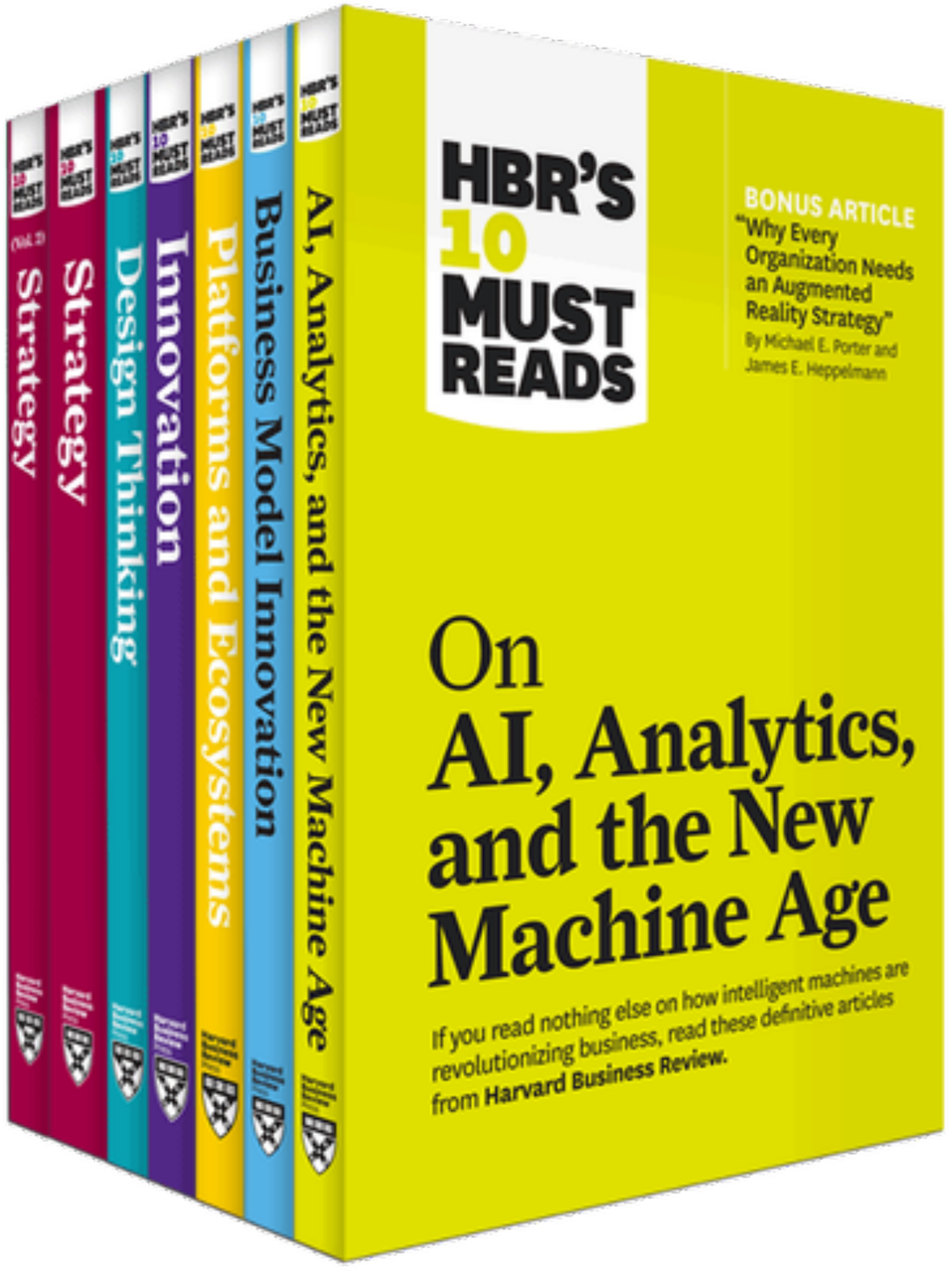
We think we know what other people are thinking

- Self-consistency bias
- Restraint bias
- Projection bias
- Pro-innovation bias
- Time-saving bias
- Planning fallacy
- Pessimism bias
- Impact bias
- Declinism
- Moral luck
- Outcome bias
- Hindsight bias
- Rosy retrospection
- Telescoping effect

We project our current mindset and assumptions onto the past and future

- Peltzman effect
- Risk compensation
- Effort justification
- Trait ascription bias
- Fundamental attribution error
- Illusory superiority
- Illusion of control
- Actor-observer bias
- Self-serving bias
- Barnum effect
- Forer effect
- Optimism bias
- Egocentric bias
- Dunning-Kruger effect
- Lake Wobegone effect
- Hard-easy effect
- False consensus effect
- Third-person effect
- Social desirability bias
- Overconfidence effect

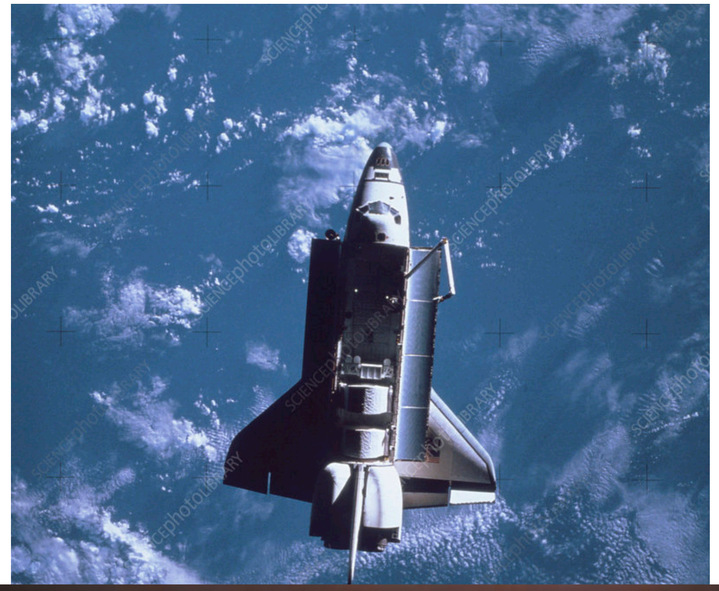
Not Enough Meaning



Anecdata

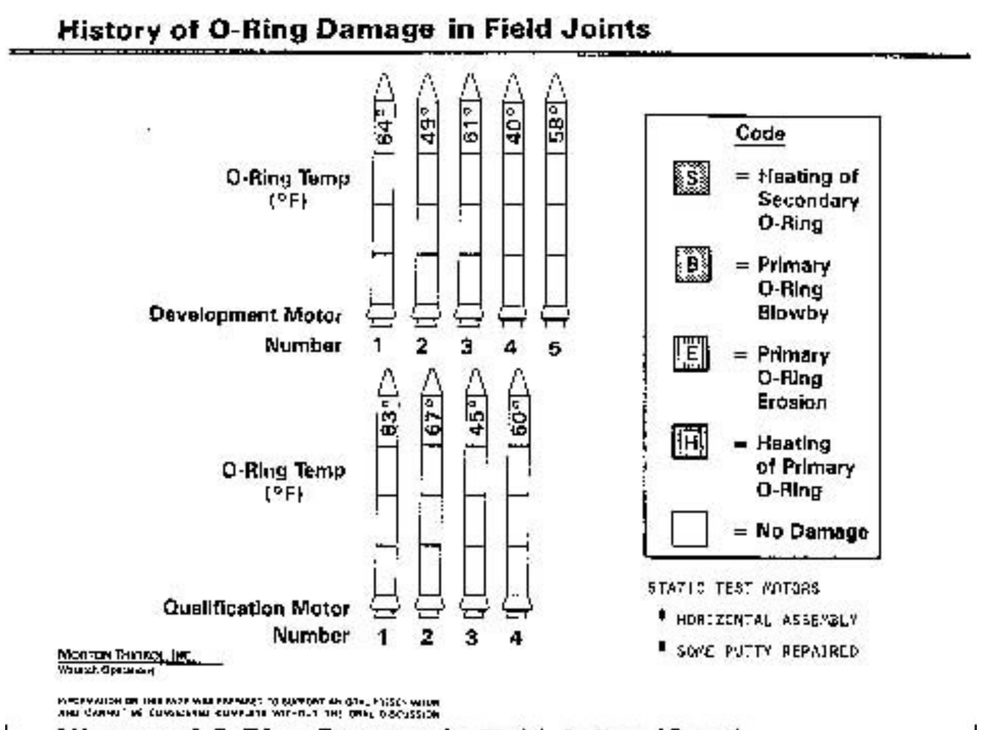
Rigorous analysis supported by anecdote

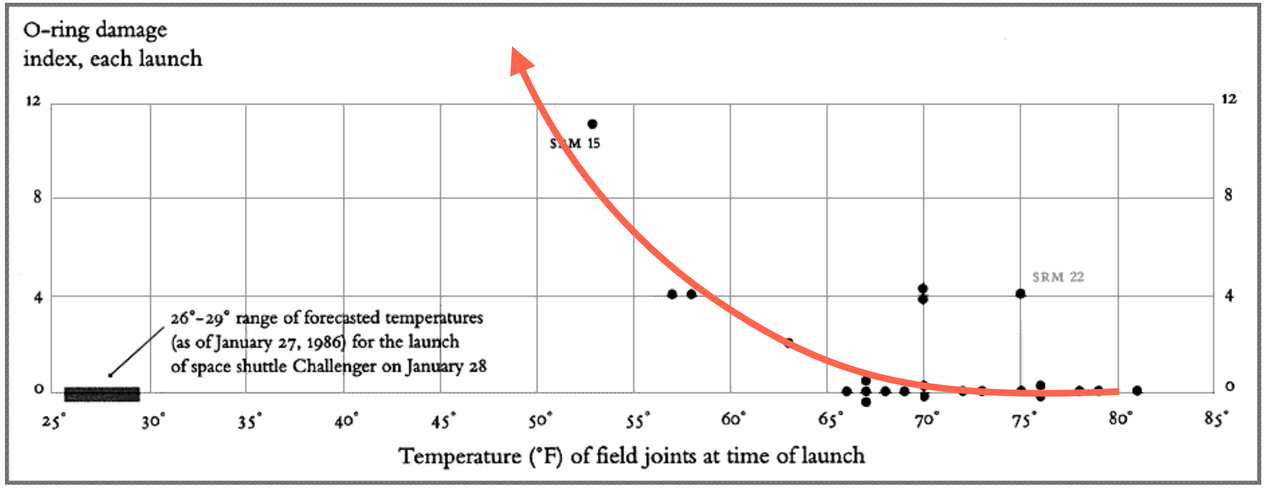
Errors of communication



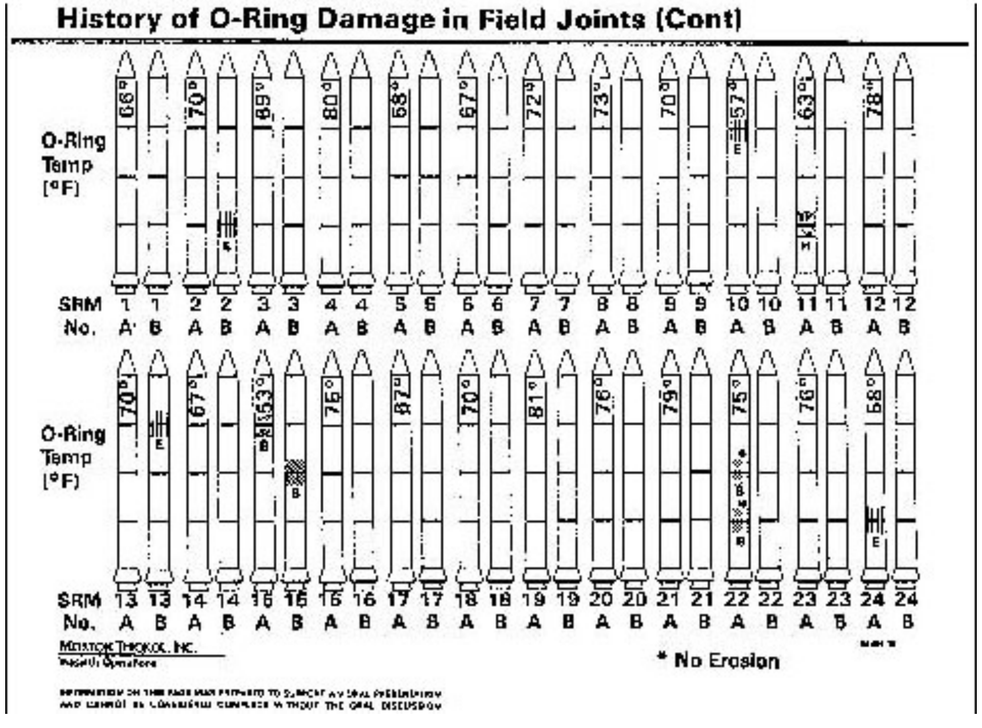
Jan 28, 1986

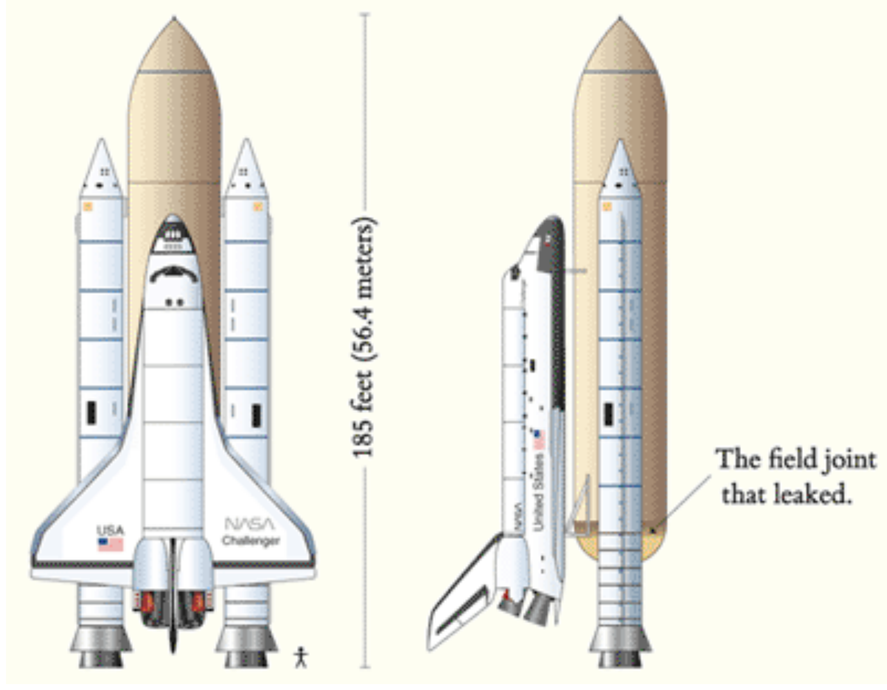
Feb 1, 2003





Images and graphs from Edward Tufte





On this one Columbia slide, a PowerPoint festival of bureaucratic hyper-rationalism, 6 different levels of hierarchy are used to display, classify, and arrange 11 phrases:

Level 1	Title of Slide
Level 2	● Very Big Bullet
Level 3	– big dash
Level 4	◆ medium-small diamond
Level 5	• tiny square bullet
Level 6	( ) parentheses ending level 5

The analysis begins with the dreaded Executive Summary, with a conclusion presented as a headline: "Test Data Indicates Conservatism for Tile Penetration." This turns out to be unmerited reassurance. Executives, at least those who don't want to get fooled, had better read far beyond the title.

The "conservatism" concerns the *choice of models* used to predict damage. But why, after 112 flights, are foam-debris models being calibrated during a crisis? How can "conservatism" be inferred from a loose comparison of a spreadsheet model and some thin data? Divergent evidence means divergent evidence, not inferential security. Claims of analytic "conservatism" should be viewed with skepticism by presentation consumers. Such claims are often a rhetorical tactic that substitutes verbal fudge factors for quantitative assessments.

As the bullet points march on, the seemingly reassuring headline fades away. Lower-level bullets at the end of the slide undermine the executive summary. This third-level point notes that "Flight condition [that is, the debris hit on the Columbia] is significantly outside of test database." How far outside? The final bullet will tell us.

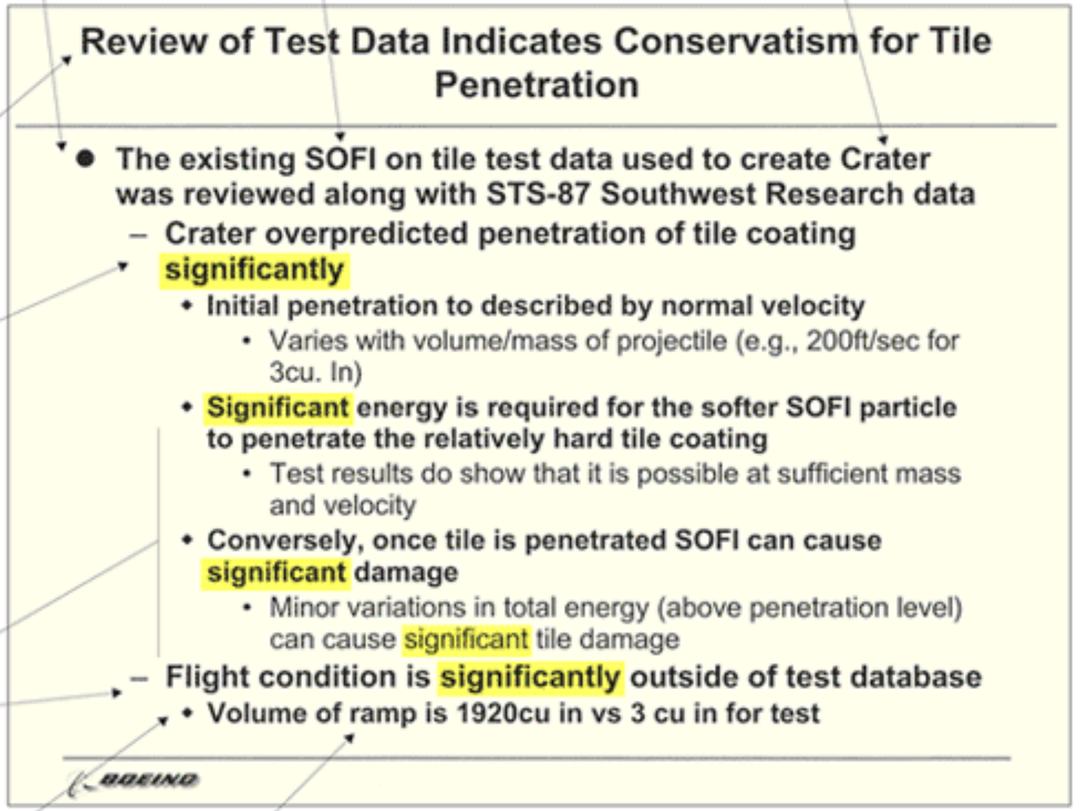
This fourth-level bullet concluding the slide reports that the debris hitting the Columbia is estimated to be 1920/3 = 640 times larger than data used in the tests of the model! The correct headline should be "Review of Test Data Indicates Irrelevance of Two Models." This is a powerful conclusion, indicating that pre-launch safety standards no longer hold. The original optimistic headline has been eviscerated by the lower-level bullets.

Note how close readings can help consumers of presentations evaluate the presenter's reasoning and credibility.

The Very-Big-Bullet phrase fragment does not seem to make sense. No other VBB's appear in the rest of the slide, so this VBB is not necessary.

Spray On Foam Insulation, a fragment of which caused the hole in the wing

A model to estimate damage to the tiles protecting flat surfaces of the wing



Here "ramp" refers to foam debris (from the bipod ramp) that hit Columbia. Instead of the cryptic "Volume of ramp," say "estimated volume of foam debris that hit the wing." Such clarifying phrases, which may help upper level executives understand what is going on, are too long to fit on low-resolution bullet outline formats. PP demands the shorthand of acronyms, phrase fragments, and clipped jargon in order to get at least some information into the tight format.

Edward Tufte

Our models are irrelevant

Debris hitting the wing was 640x larger than the experimental data used to build these models

We have no clue what will happen on re-entry

Communication is key

Identify audience & setting

Identify key insight, main points of evidence, and assumptions

Organize into a story focussed on ⬆️

Create supporting visualizations

Revise to be as precise and concise as possible

Errors of measurement - are we measuring what we think we are?

Errors of analysis - did we use the right methods to address the question?

Errors of borked tools - choosing the wrong tools or using them poorly leads to bad results

Errors of human cognition - data science is a human endeavor with all the usual frailties and foibles

Errors of communication - sometimes you get everything right, but the group and the decision makers never understand properly

https://forms.gle/VscVbHePDBNmJKXv7