EXPO 2023 DOHA QATAR إكسبو 2023 الدوحة قطر

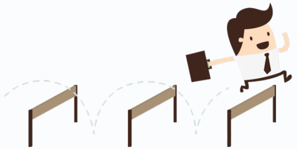
جهاز التخطيط والإحصاء
Planning and Statistics Authority
State of Qatar • دولة قطر

Common Mistakes in **Data science** and **AI**

Dr. Khaled Alqahtani

INTRODUCTION

- The path to becoming a data scientist is one riddled with **traps** and **challenges**.



- If you’re **not careful**, you might just fall into one of them!

- There are so many **resources** out there which aim to help aspiring data scientists become experts, but most of them are half-baked efforts that leave a gaping hole in your data scientist journey.

- Many a times, **mistakes** have led to some fantastic discoveries – **penicillin**, an antibiotic that saves millions of lives, is one such example. This holds true in case of **data science** as well.

PENICILLIN
PENICILLI

- But **not** all mistakes lead to new discoveries; in fact, most of them lead to a dead end – which translates into an insight-less or misleading study for a data scientist.

- IBM estimates that bad data costs the U.S. economy around **$3.1 trillion dollars** each year.

IBM

$3.1 trillion

- Additional research from Experian Data found that bad data has a direct impact on the bottom line of **88%** of American companies, with the average company losing around **12%** of its total revenue

Experian

Data Quality

88%

12%

1985
new coke

The infamous flop of a reformulation of Coca-Cola, unofficially referred to as “New Coke.” was meant to compete with Pepsi’s sweeter taste and ever-growing market share during the Instead the product ended up becoming a textbook example of market research data gone awry. and resulted in tens of millions of dollars in losses for Coca-Cola.

Coca-Cola

What were the **BAD DATA?**

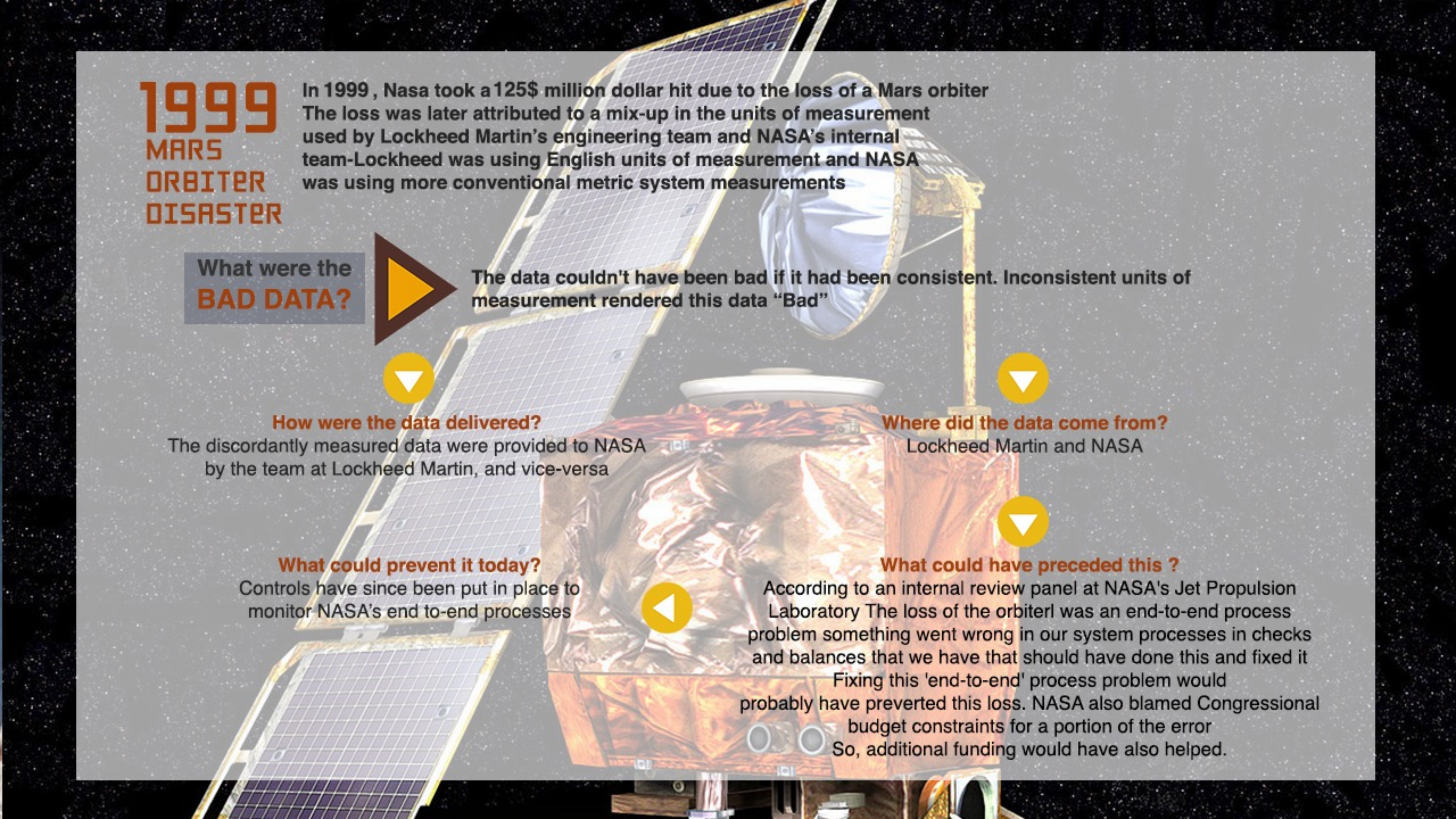
Coca-cola tested the New Coke formula on 200.000 subjects. It beat Pepsi and “Old Coce” Time after time in a series of taste tests. Turns out, there were several factors the market research didn’t account for, and instead relied on the single data point of taste to introduce New Coke to the market. However, customers are motivated by more than just taste. Marketers didn’t consider the classic formula’s relation to the larger brand. and the blind taste test didn’t adress the fact that launching New Coke would result in Coca-Cola pulling the original formula from the shelves.

Where did the data come from?
Coke’s market research team

How were the data delivered?
The data were delivered from Coke’s market rsearch firm to the product and branding teams.

What could have prevented this?
The marketing experts at Coke should have known that purchasing decisions for soft drinks are based on more than just taste. The team could have released New Coke as a product option instead of releasing it as a product replacement to the original formula.

What could prevent it today?
Marketing reseatch firms now use both quantitave and qualitative data when launching new products to guarantee they’re not the new “New Coke”



1999 MARS ORBITER DISASTER

In 1999 , Nasa took a 125$ million dollar hit due to the loss of a Mars orbiter
The loss was later attributed to a mix-up in the units of measurement
used by Lockheed Martin's engineering team and NASA's internal
team-Lockheed was using English units of measurement and NASA
was using more conventional metric system measurements

What were the BAD DATA?

The data couldn't have been bad if it had been consistent. Inconsistent units of measurement rendered this data "Bad"

How were the data delivered?

The discordantly measured data were provided to NASA by the team at Lockheed Martin, and vice-versa

Where did the data come from?

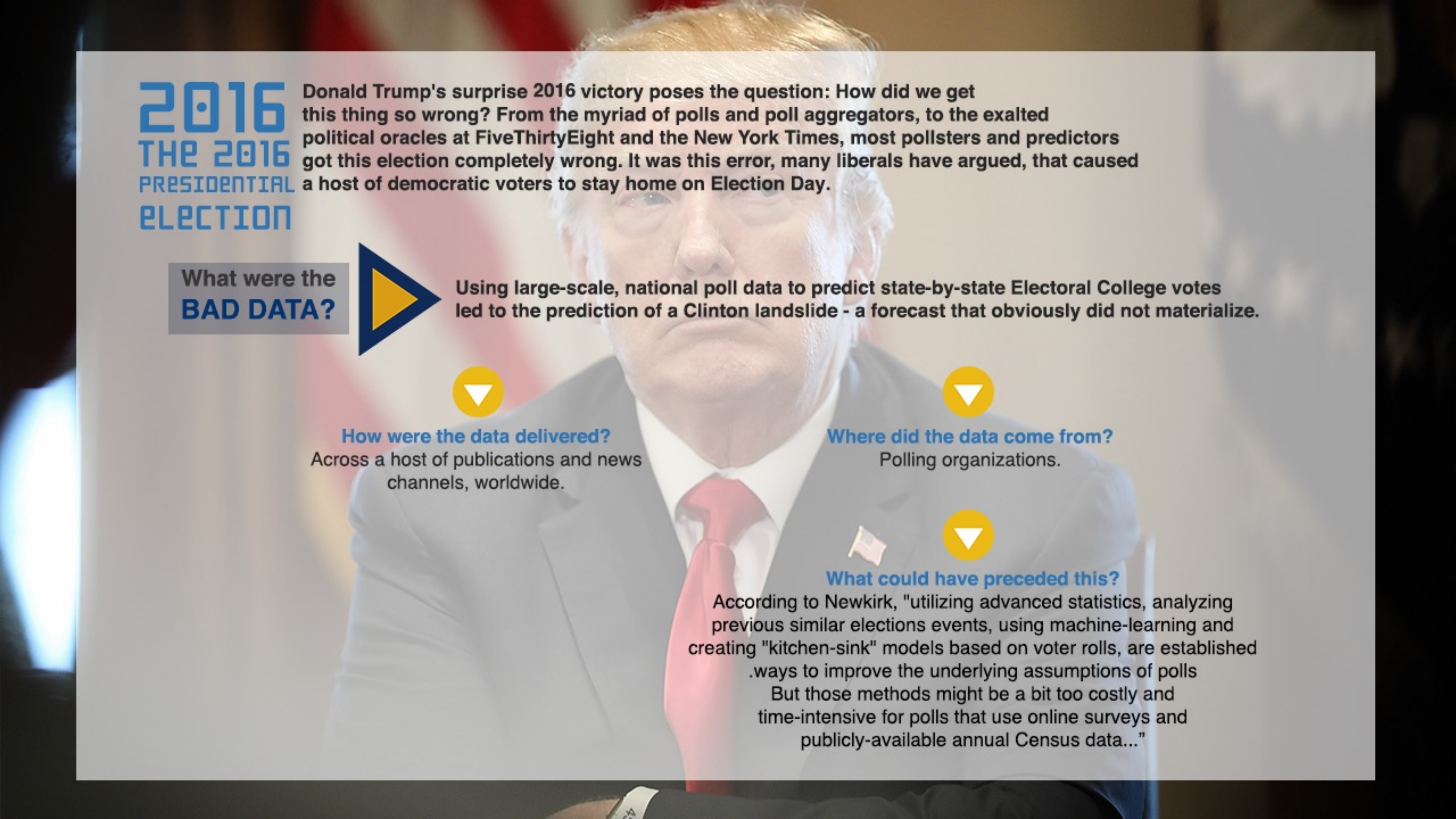
Lockheed Martin and NASA

What could have preceded this ?

According to an internal review panel at NASA's Jet Propulsion Laboratory The loss of the orbiterl was an end-to-end process problem something went wrong in our system processes in checks and balances that we have that should have done this and fixed it Fixing this 'end-to-end' process problem would probably have preverted this loss. NASA also blamed Congressional budget constraints for a portion of the error
So, additional funding would have also helped.

What could prevent it today?

Controls have since been put in place to monitor NASA's end to-end processes



2016 THE 2016 PRESIDENTIAL ELECTION

Donald Trump's surprise 2016 victory poses the question: How did we get this thing so wrong? From the myriad of polls and poll aggregators, to the exalted political oracles at FiveThirtyEight and the New York Times, most pollsters and predictors got this election completely wrong. It was this error, many liberals have argued, that caused a host of democratic voters to stay home on Election Day.

What were the **BAD DATA?**

Using large-scale, national poll data to predict state-by-state Electoral College votes led to the prediction of a Clinton landslide - a forecast that obviously did not materialize.

How were the data delivered?

Across a host of publications and news channels, worldwide.

Where did the data come from?

Polling organizations.

What could have preceded this?

According to Newkirk, "utilizing advanced statistics, analyzing previous similar elections events, using machine-learning and creating "kitchen-sink" models based on voter rolls, are established .ways to improve the underlying assumptions of polls But those methods might be a bit too costly and time-intensive for polls that use online surveys and publicly-available annual Census data..."

⭐ Mistakes in Research

- Kim et al.(2011) reviewed 418 papers published between 1995 and 2009 in ten well-established dental journals. The reported proportion of erroneous articles has been about 30 – 50%.

30 - 50%

- 95% of the approximately 5000 submitted manuscripts every year to the New England Journal of Medicine are rejected before they reach the final review by statistical experts who then in turn reject on average 20% of the manuscripts.

20%

🌟 Daily data science mistakes

Be Specific

„1”

Be specific

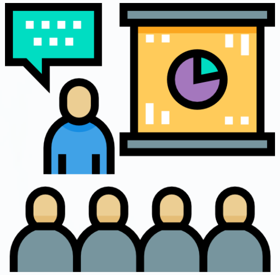
- Your professor hands back the midterm. The grades are distributed as follows:

Grade	100	98	95	63	58
Received	4	5	2	4	6

What Do you **Think** about this test ?

⭐ Be specific

The professor felt that the test must have been too easy, because the average grade was a 95.



When a colleague asked the professor about how the midterm grades came out, he answered, knowing that his classes were gaining a reputation for being too easy, that the average grade was an 80.

When your parents ask you how you can justify doing so poorly on the midterm, you answer, Don’t Worry about my 63. It is not as bad as it sounds. The average grade was a 58.

The lesson here?

Specify which of the three forms of average you are using.

Be
Clear.“^{2}”

Be clear

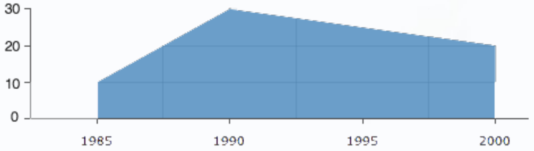


CHART A

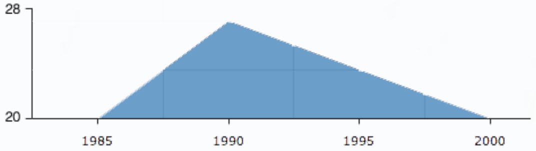


CHART B

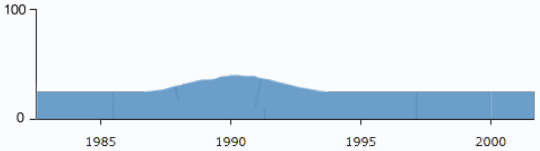


CHART C

A. Chart ? reveals a steep jump, with a sharp drop-off immediately following.

B. Chart ? seems to demonstrate that there was virtually no change over time.

C. Chart ? shows a mild increase, followed by a slow decline.

The lesson here?

Show the entire picture

Be Fair „3”

Be Fair

- Suppose we have two cities, **Leeds** and **Coventry**. In Coventry, the murder rate has gone up **75%**, while in Leeds, the rate has only increased by **10%**. Which city is having a bigger murder problem?

75%

Coventry

10%

Leeds

⭐ Be Fair

- This is actually much less clear than it looks. In order to really know which city has a worse problem, we have to look at the actual numbers.

- if I told you that **Coventry** had **4** murders last year and **7** this year, and **Leeds** had **30** murders last year and **33** murders this year, would you change your answer?

⭐ Be Fair

Maybe, since 33 murders are significantly more than 7. One would certainly feel safer in Coventry, right?

Not so fast, because we still do not have all the facts!!

We have to make the comparison between the two based on equivalent standards.

Be Fair

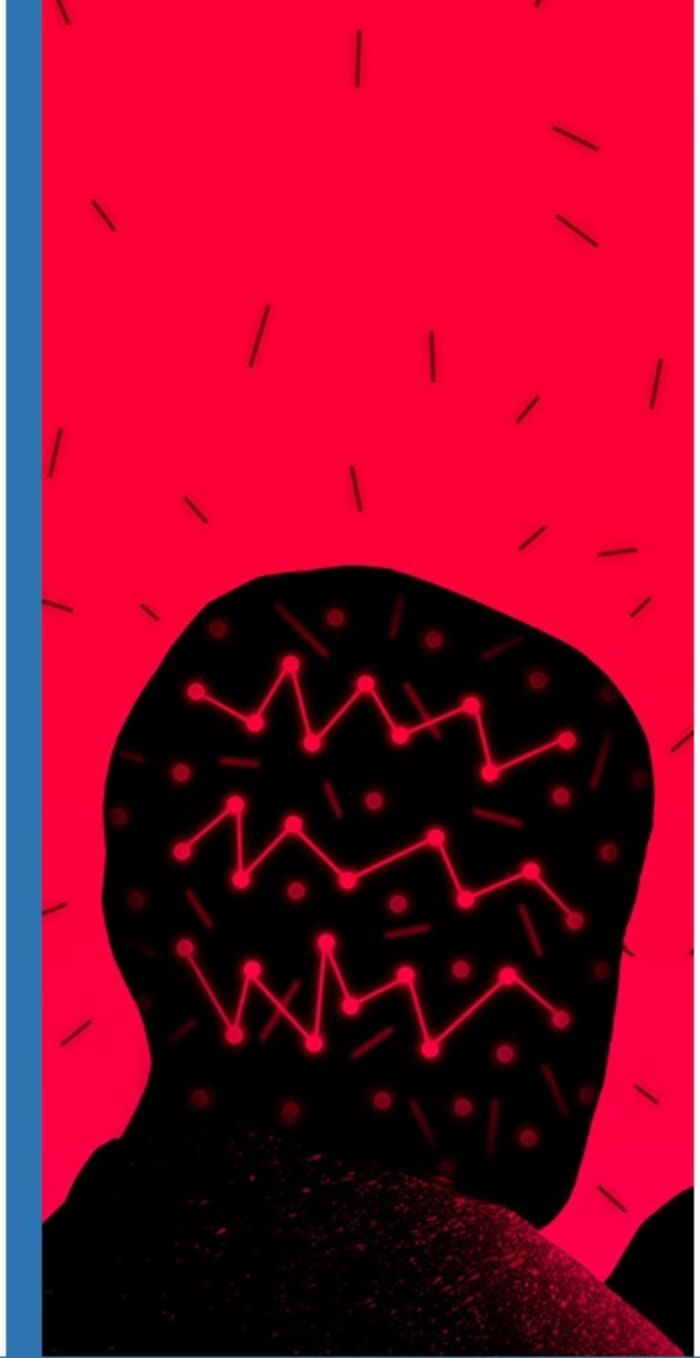
- To do that, we have to look at the per capita rate (often given in rates per **100,000** people per year)

- If **Coventry** has **700** residents while Leeds has **3.3** million, then **Coventry** has a murder rate of **1,000 per 100,000** people, and **Leeds**’ rate is merely **1 per 100,000**

- I think I will stick with nice, safe **Leeds**.

The lesson here?

Give bases of all percentages



Don’t ***Overstate*** the **results** of an observational study.

“4”

Don’t *Overstate* the results of an observational study

- There is a study found that the more firefighters sent to put out a fire, the more damage the fire did. What is your explanation?

Don’t *Overstate* the results of an observational study

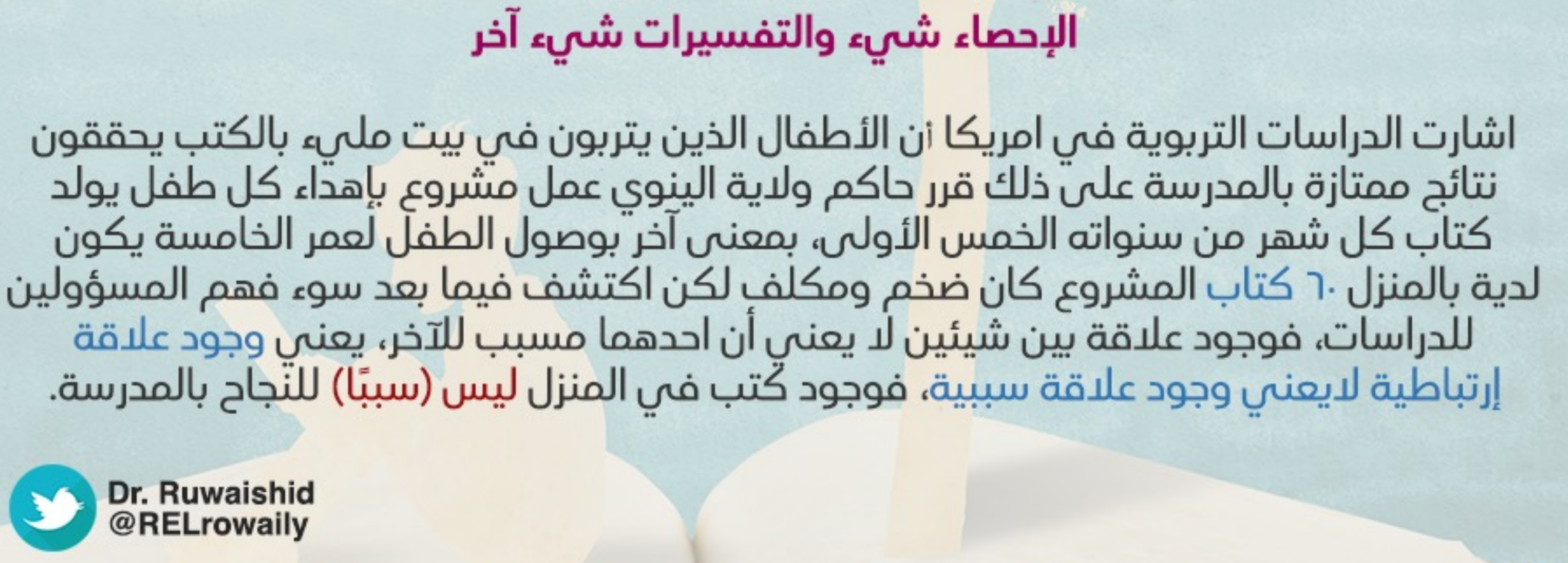
- This seemingly contradictory finding can be easily explained by pointing to a third factor that causes both: the size of the fire.

The lesson here?

Correlation does not equal causation.

Confusing correlation with causation. Just because two things vary together does not mean that one of them is causing the other.

Correlation does not mean causation



الإحصاء شيء والتفسيرات شيء آخر

اشارت الدراسات التربوية في امريكا أن الأطفال الذين يتربون في بيت مليء بالكتب يحققون
نتائج ممتازة بالمدرسة على ذلك قرر حاكم ولاية الينوي عمل مشروع بإهداء كل طفل يولد
كتاب كل شهر من سنواته الخمس الأولى، بمعنى آخر بوصول الطفل لعمر الخامسة يكون
لدية بالمنزل ٦٠ كتاب المشروع كان ضخم ومكلف لكن اكتشف فيما بعد سوء فهم المسؤولين
للدراسات، فوجود علاقة بين شيئين لا يعني أن احدهما مسبب للآخر، يعني وجود علاقة
إرتباطية لايعني وجود علاقة سببية، فوجود كتب في المنزل ليس (سببًا) للنجاح بالمدرسة.

Dr. Ruwaishid
@RELrowaily

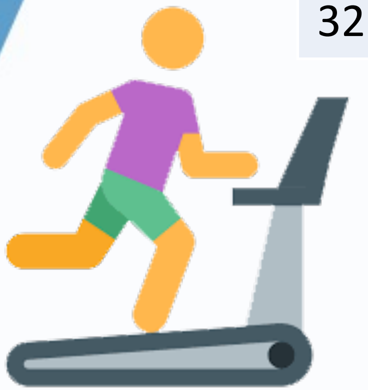
WRONG DECISION
Due to
WRONG ANALYSIS

„5”

Wrong Decision Due to Wrong Analysis

- Pulse Rates **Before** and **After Marching**

Participants	Before	After	Difference
1	60	78	18
2	56	66	10
3	90	96	6
.	.	.	.
.	.	.	.
32	78	88	10



Wrong Decision Due to Wrong Analysis

- Two sample T test for **AFTER** vs **BEFORE**

	N	Mean	SD
After	32	82	13
Before	32	72	15.9

Hypothesis: H0 : μ_After = μ_Before
P – Value = 0.33
Conclude **no difference in** mean pulse rates before and after marching.

Wrong Decision Due to Wrong Analysis

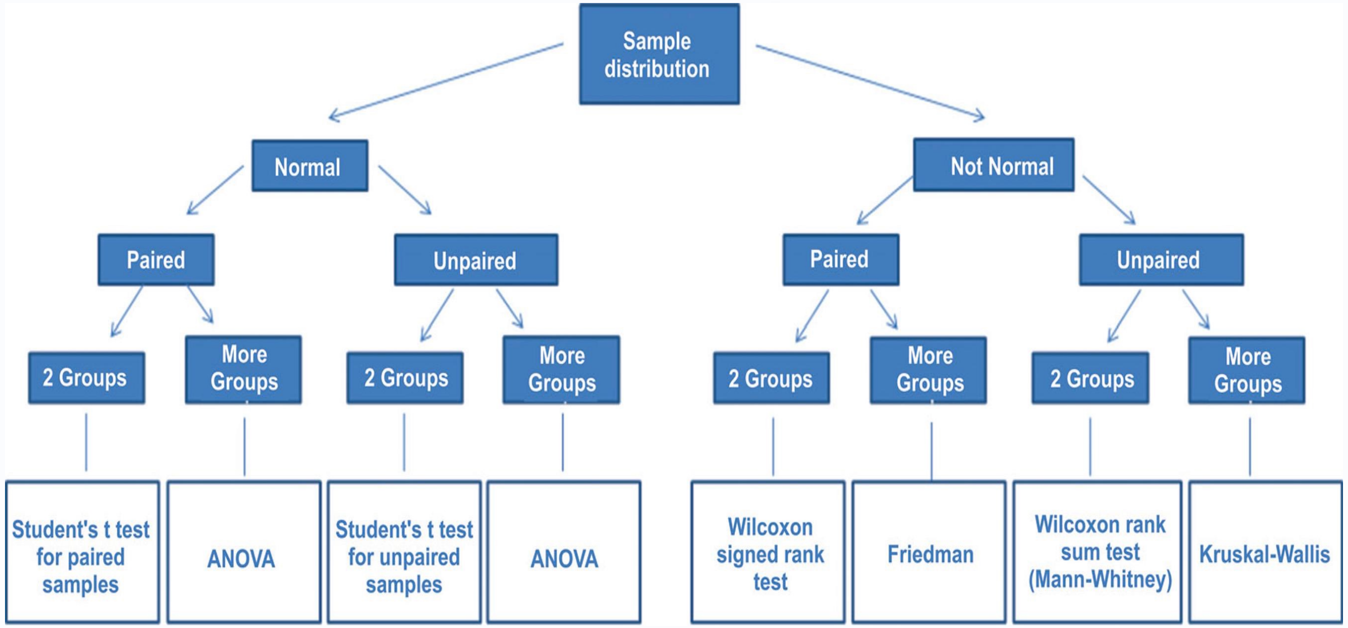
- Paired T for **AFTER – BEFORE**

	N	Mean	SD
After	32	82	13
Before	32	72	15.9
Difference	32	11	5.03

P – Value = 0.02

Conclude mean pulse rate after **is greater than** mean pulse rate before.

Wrong Decision Due to Wrong Analysis

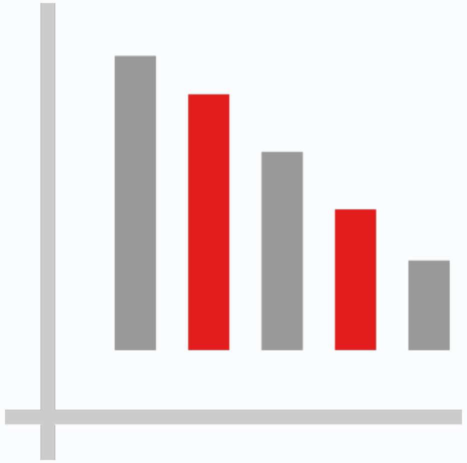


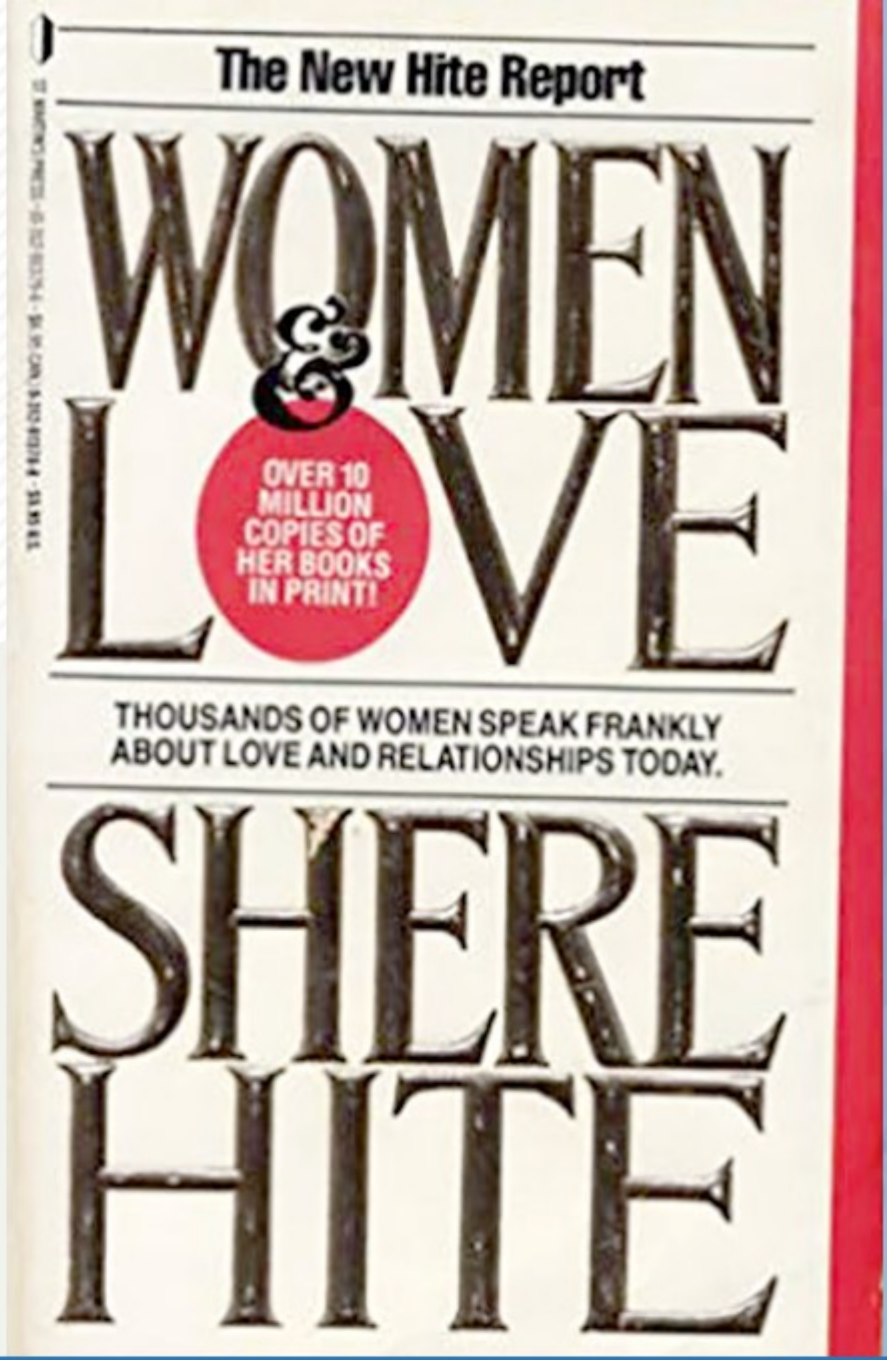
DO YOU PREFER CATS OR DOGS?
DOGS
CATS

Sampling Bias “6”

Sampling Bias

- Shere Hite wrote a book in **1987** called "Women in Love"
- **100,000** questionnaires about love, and relationships sent to women's group.
- Entire book devoted to results of survey.





The New Hite Report

WOMEN & LOVE

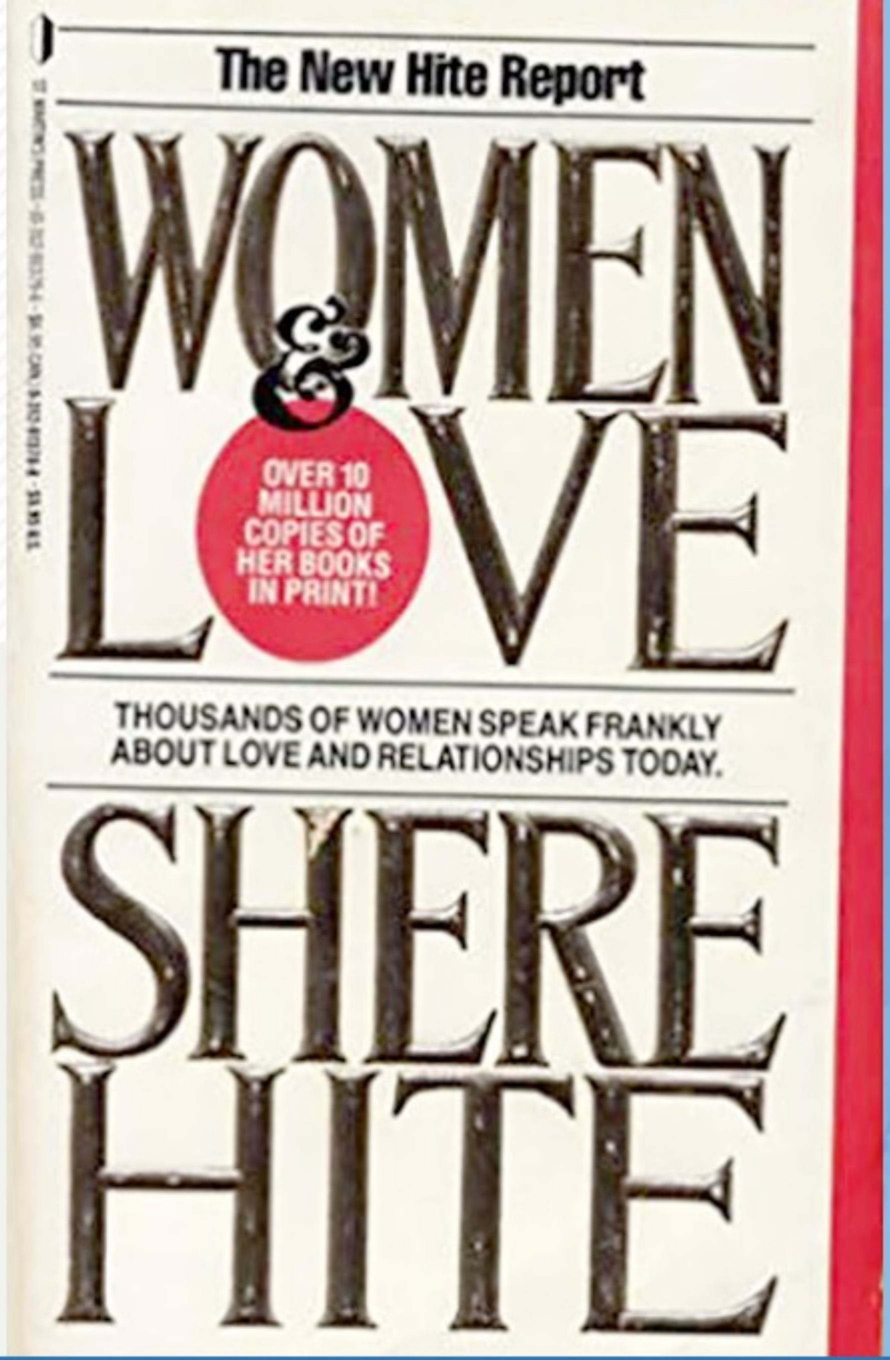
OVER 10 MILLION COPIES OF HER BOOKS IN PRINT!

THOUSANDS OF WOMEN SPEAK FRANKLY ABOUT LOVE AND RELATIONSHIPS TODAY.

SHERE HITE

Sampling Bias

- 84% of women **not satisfied** with their relationship.
- 70% of women married 5 years committed adultery.
- 95% of women report psychological and physical harassment from their partners.



The New Hite Report

WOMEN & LOVE

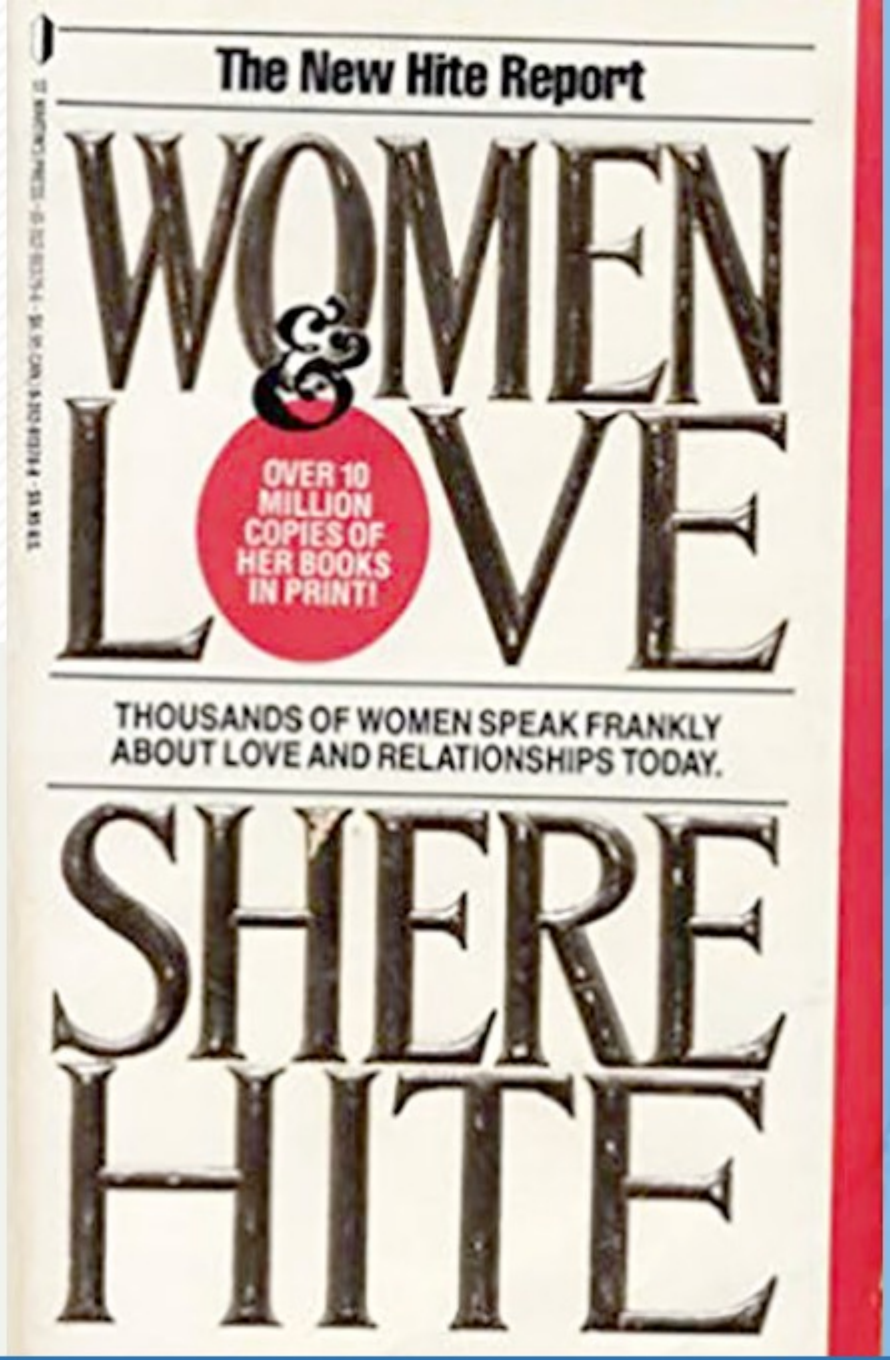
OVER 10 MILLION COPIES OF HER BOOKS IN PRINT!

THOUSANDS OF WOMEN SPEAK FRANKLY ABOUT LOVE AND RELATIONSHIPS TODAY.

SHERE HITE

Sampling Bias

- Unrepresentative sample.
- Only **4,500** questionnaires returned.



The New Hite Report

WOMEN & LOVE

OVER 10 MILLION COPIES OF HER BOOKS IN PRINT!

THOUSANDS OF WOMEN SPEAK FRANKLY ABOUT LOVE AND RELATIONSHIPS TODAY.

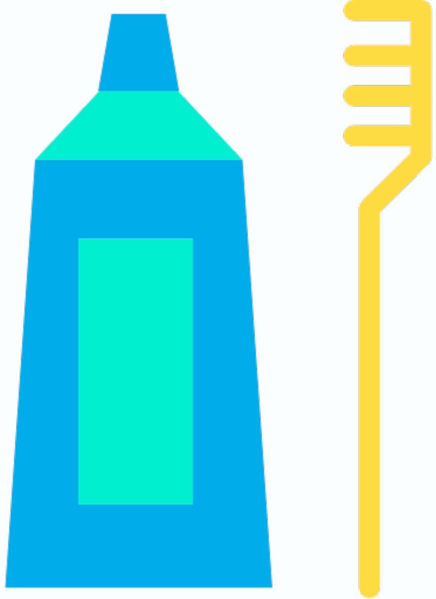
SHERE HITE

Sampling Bias

More than **80%** of Dentists recommend Colgate

Sampling Bias

Colgate was ordered by the Advertising Standards Authority (ASA) of the U.K. to abandon their claim. The slogan in question was positioned on an advertising billboard in the U.K., and was deemed to be in breach of U.K. advertising rules.



Sampling Bias

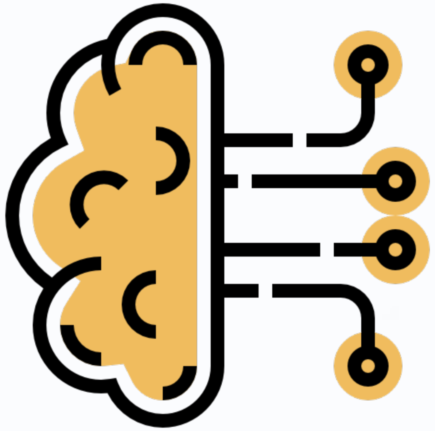
The claim, which was based on surveys of dentists and hygienists carried out by the manufacturer, was found to **be misrepresentative** as it allowed the participants to select one or more toothpaste brands.
The ASA stated that the claim would be understood by readers to mean that 80 percent of dentists recommend Colgate over and above other brands, and the remaining 20 percent would recommend different brands.
The ASA also claimed that the scripts used for the survey informed the participants that the research was being performed by an independent research company, which was inherently false.

GENERAL DATA SCIENCE MISTAKES

1 Learning Theoretical Concepts without Applying Them

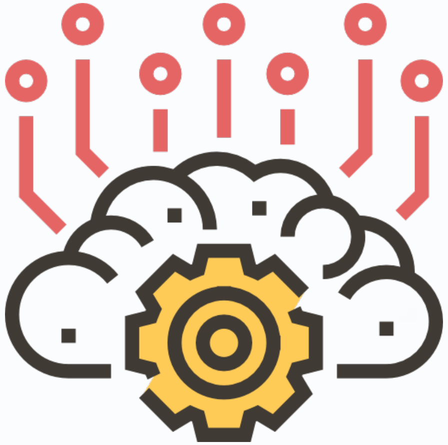
- **Learning process** should be a healthy balance between theoretical and practical.

- As soon as you learn a concept, **find** a dataset or problem where you can use it.



2 Using ML Techniques without Learning the Prerequisites

- ML components you should know about - Statistics, Probability, Linear Algebra, Calculus.
- Tons of resources available to learn/brush-up.



3 Relying Solely on Certifications and Degrees

- **Use** real-world datasets and ensure you write a blog about your work.
- **Ask** for feedback from the community.
- **Apply** for internships to understand how a data science project works.

4 Focusing on Model Accuracy over Interpretability

- Practice making simpler models and then explaining them to non-technical people.

- **Add** further complexity; this will teach you when to stop, and why simple models are always given preference.

Input → BLACK BOX → Output

5 Using Tools and Libraries without understanding the Problem

- **Read up** on how companies in your domain are using data science.

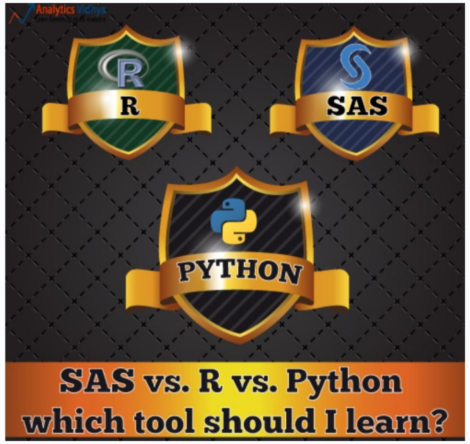
6 Not Spending Enough Time on Exploration and Visualization

- **Practice!** Next time you work on a dataset, spend more time on this step.

- **Ask questions!** Ask your manager, ask domain experts, search for solutions on the internet and if you don’t find any, ask on social media!

7 Trying to Learn Multiple Tools at the Same Time

- **Pick one tool** and stick to it until you have mastery over it.
- Each tool has a great user community which you can tap into whenever you get stuck.
- The aim is to learn data science through the tool, not the tool through data science.



8 Not Studying in a Consistent and Disciplined Manner

- **Set goals for yourself.** Map out a time table and stick it on your wall.
- **Plan how** and what you want to study and set deadlines for yourself (like learning a technique and applying it in a competition).
- **Be ready** to put in long hours.

9 Shying Away from Discussions and Competitions

- **Start participating** in discussions and competitions!
- If you learn a new technique out of the whole thing, you have won in your own right.

Do you have any questions?

Thank you.