Pitfalls in Data Analytics: Biases and “Weird Stuff”

Will Doyle

Survivorship Bias

Concentrating on successful cases while overlooking those that didn't make it.

Example: Mutual funds that survive and advertise market-beating performance after eliminating underperforming funds.

“Success Studies”

Regression to the Mean

Extreme values of random events are often followed by less extreme values.

Example: An NFL player having an outstanding season followed by a less stellar performance the next year

Israeli Fighter Pilots

Simpson's Paradox

A trend or association between variables is reversed after the addition of a third variable.

Example: Success rates for kidney stone removal procedures appearing to favor the new procedure overall, but when broken down by stone size, the traditional procedure performs better.

Do small class sizes affect student learning?

Confirmation Bias

Interpreting data to confirm existing beliefs while dismissing conflicting evidence.

Example: Data analysis conducted to validate decisions already made.

Forgetting Baseline Rates

The baseline rate for a disease is 1%

A new vaccine offers “95% Efficacy”

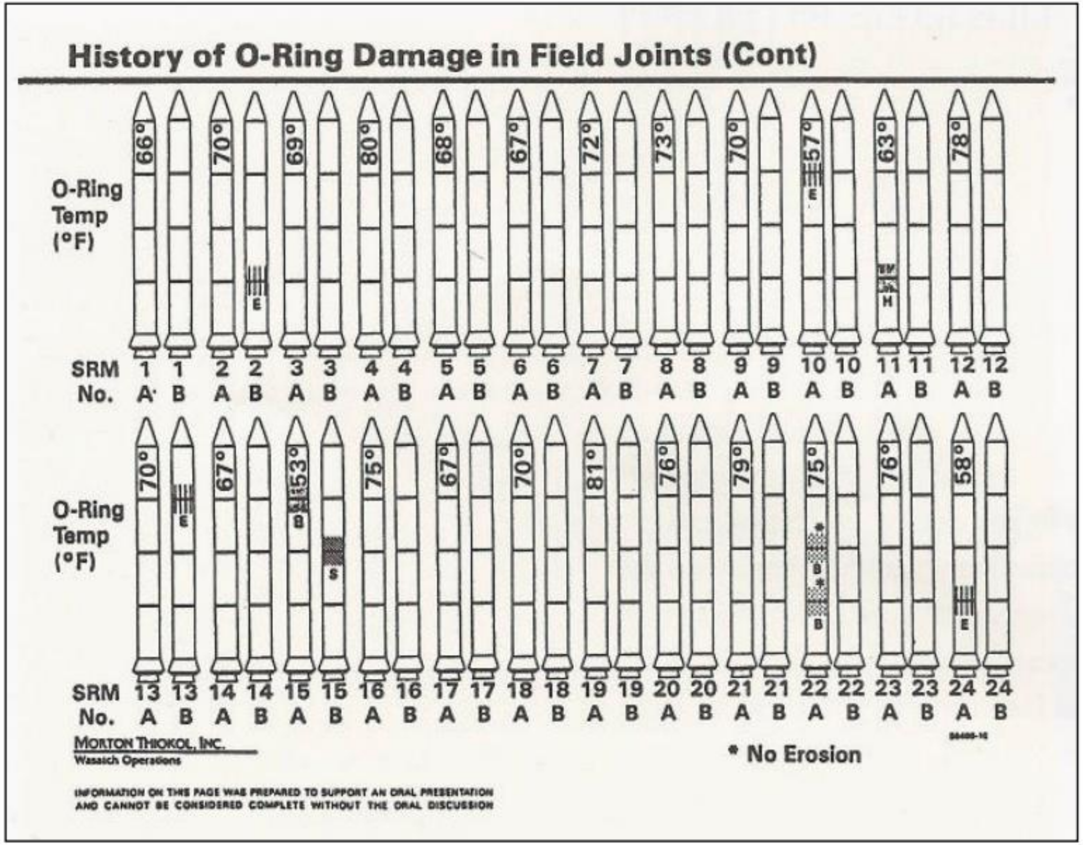
What does this mean?

“A 95% vaccine efficacy means that instead of 1000 cases in a population of 100,000 without vaccine we would expect 50 cases (99.95% of the population is disease-free, at least for 3 months).”

95% efficacy in this case means a drop from 1000 cases to 50 cases.

Source: https://doi.org/10.1016/S1473-3099(21)00075-X

What’s the Pattern in the Data?



History of O-Ring Damage in Field Joints (Cont)

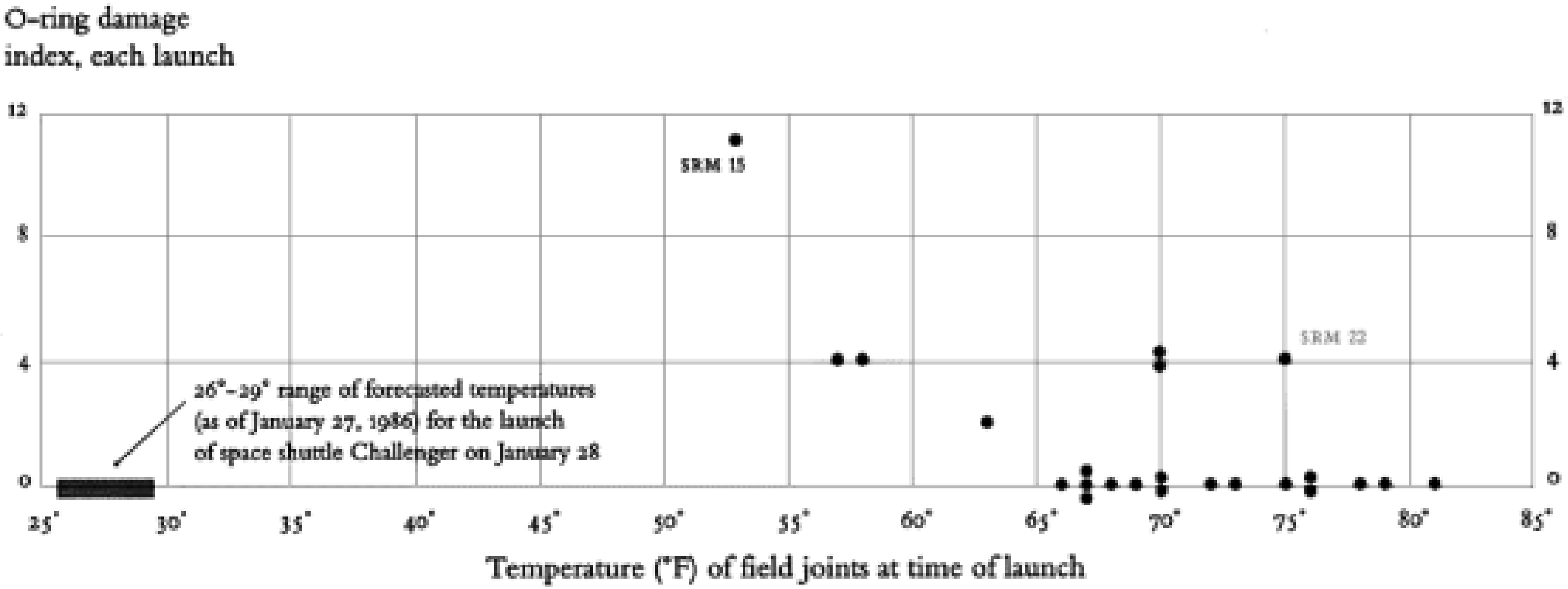
MORTON THIOKOL, INC.
Wasatch Operations

* No Erosion

INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

Source: Tufte

What’s the Pattern in the Data?



The Garden of Forking Paths

“A multiple comparisons problem does not have to come from ‘fishing’ but can arise more generally from reasonable processing and analysis decisions that are contingent on data.”

http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Noise, Noise Everywhere

“No matter how sophisticated our choices, how good we are at dominating the odds, randomness will have the last word.”

> Nicholas Nassim Taleb *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*