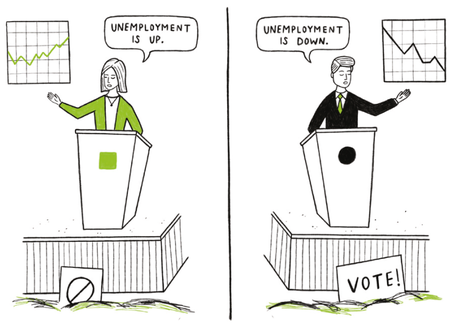
DATA FALLACIES TO AVOID

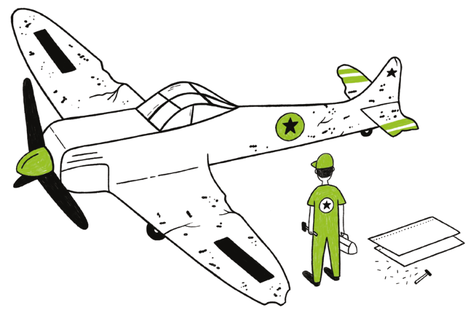


CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.

DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

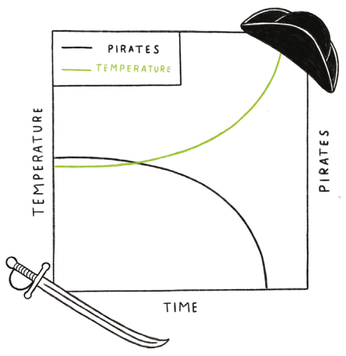


SURVIVORSHIP BIAS

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

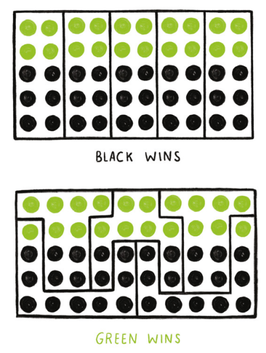
COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



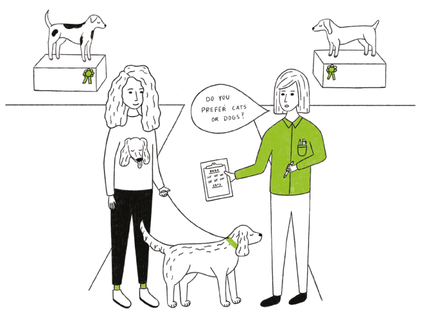
FALSE CAUSALITY

Falsely assuming when two events appear related that one must have caused the other.



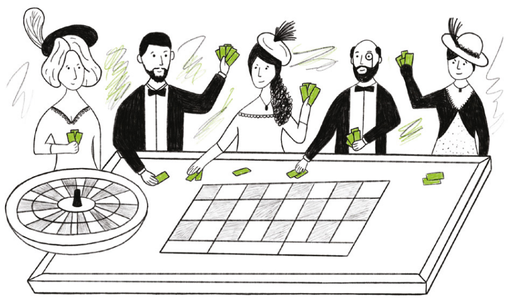
GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.



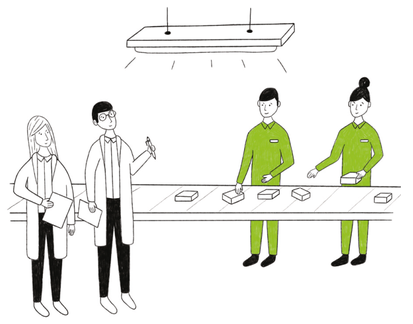
SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.



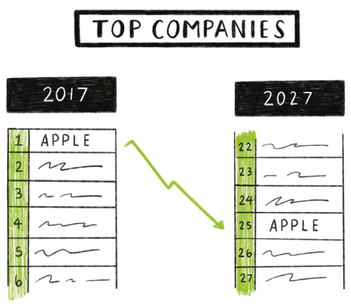
GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).



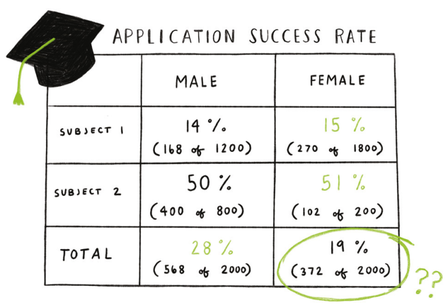
HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



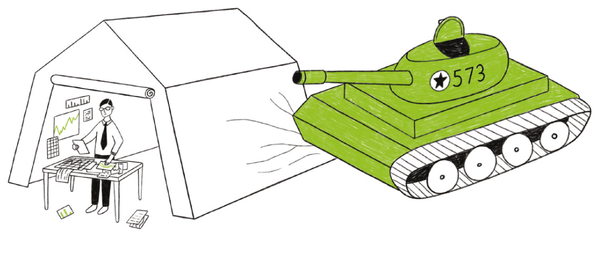
REGRESSION TOWARDS THE MEAN

When something happens that's unusually good or bad, it will revert back towards the average over time.



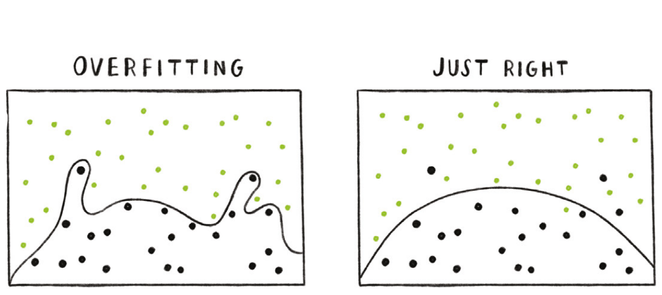
SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



MCNAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.

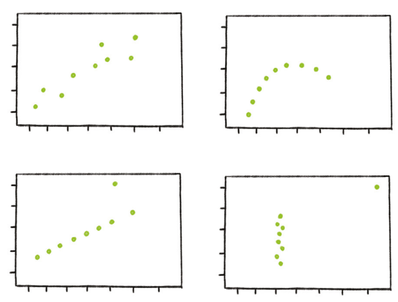


OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.

PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.