How Executives Can Avoid Analytics Mistakes

- Data Scientists tend to be well-versed in math, probabilities, and statistics.
- Even so, humans in general are susceptible to cognitive biases, especially when interpreting data
- In this session we show you how to avoid some of these failings

“We have a problem with customer loyalty”

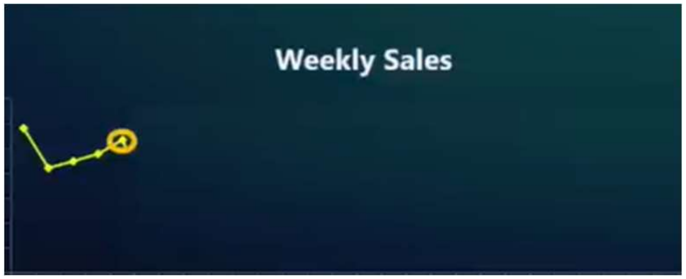
- “Our best customers (top 10%) in 2016 bought 30% less in 2017”

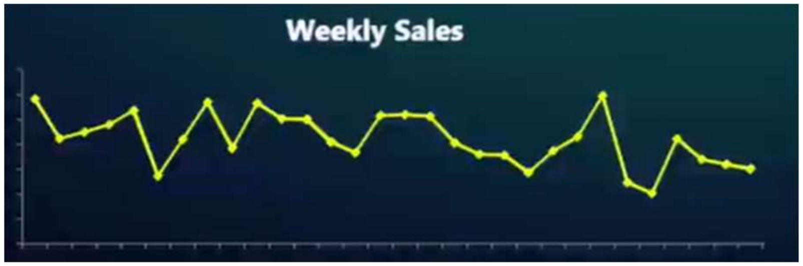
Issue 1: Regression to the Mean

- First witnessed by Sir Francis Galton in 1886

“Sales have increased for the past 4 weeks. We’re on an upswing!!”

- Is this a valid conclusion?
- this is an issue of “spotting trends too early”
- Issue 2: Invalid trends





Issue 3: Learning Something That Isn't True

- "Superstitious learning occurs when the connection between the cause of an action and the outcomes experienced aren't clear or misattributed."
- This can be due to:
  - regression to the mean issues (Issue 1)
  - trends that are really random (Issue 2)
  - faulty case studies ("one-off occurrences")
  - causation inferred from correlation

A statistically significant correlation between two variables may be due to:

- chance
  - the usual statistical significance burden of proof is 5%. If there is no relationship between 2 variables then we would be concluding there IS a statistical significance 1 in 20 times.
  - If you look at relationships among 15 variables (by looking at pairs), 5 correlations will be statistically significant simply by chance.
- underlying (hidden) factor
- a true cause-effect relationship (but which causes which)

Issue 4: Most observational studies tend to be wrong

- chance
  - the usual statistical significance burden of proof is 5%. If there is no relationship between 2 variables then we would be concluding there IS a statistical significance 1 in 20 times.
  - If you look at relationships among 15 variables (by looking at pairs), 5 correlations will be statistically significant simply by chance.
- underlying (hidden) factor
- a true cause-effect relationship (but which causes which)

Wrong results reported from an observational study could be due to:

- innocence
  - someone found a nugget of (fool's) gold
- not so innocent
  - we continue to torture the data until it confesses, all the while ignoring all other signals or common sense
  - don't be pressured by management or customers to do this

Solution:

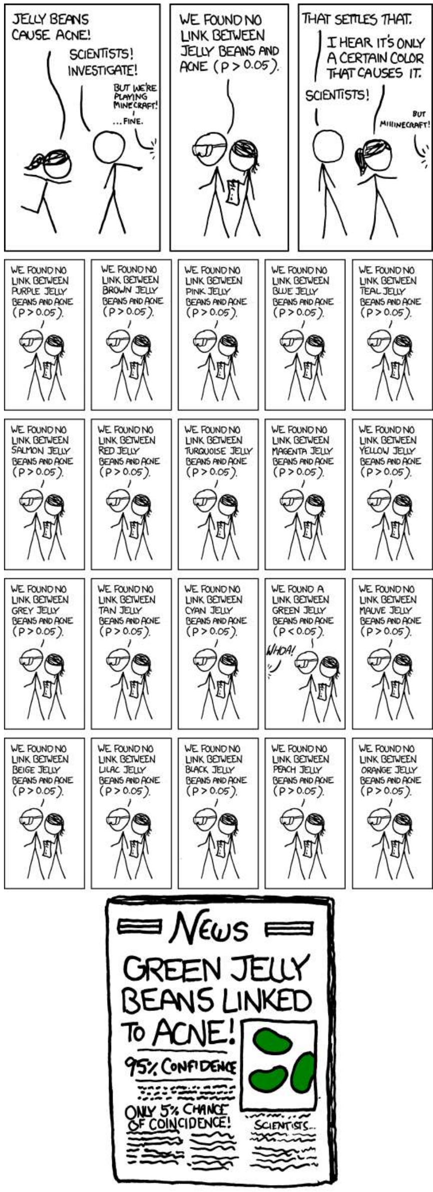
Test it. (clinical trial, Design of Experiments, A/B Test, Champion/Challenger)

Look at independent data sets

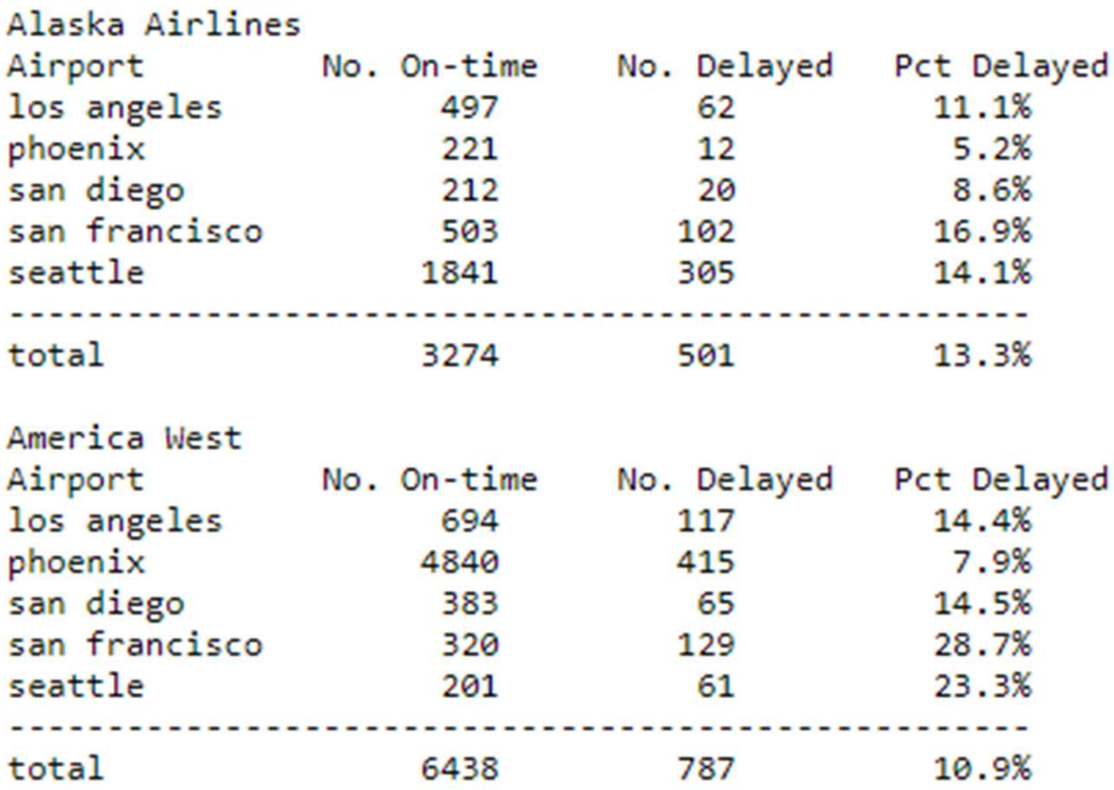
Never allow a dataset to suggest a relationship AND validate it

Issue 4: Avoid Other data dredging techniques

- Throw out the "outliers" until we get the desired result
  - we can throw out outliers sometimes if they truly are not representative, but don't do it because the data doesn't fit our narrative
- Slice and dice the data until you find a subset that gives you the desired result
- conduct many hypothesis tests (checking for many correlations). →
  - significance by chance 5% of the time
- Ignore negative results
- Question negative results and don't question positive results


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

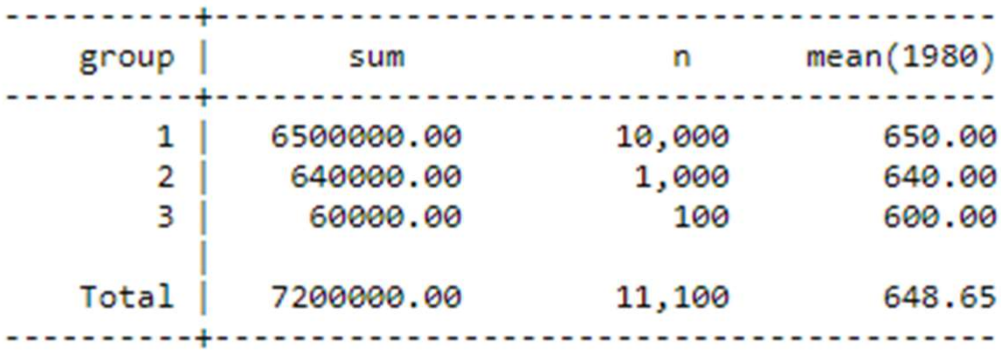
Issue 5: Simpson’s Paradox



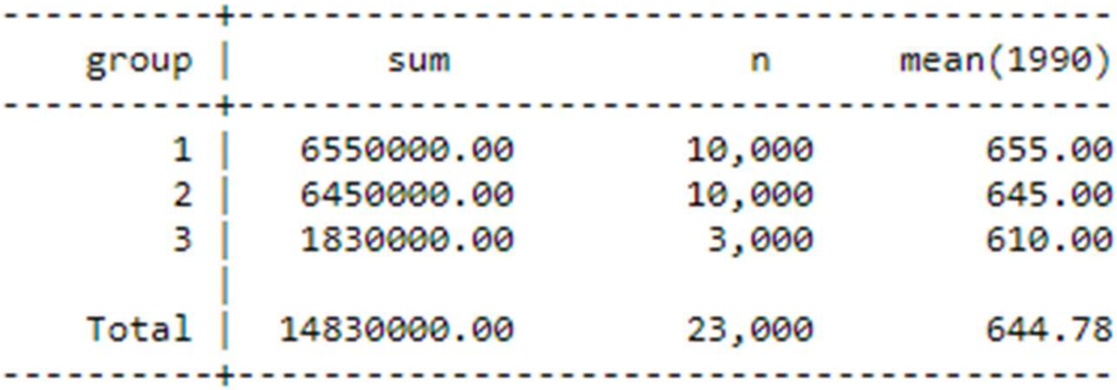
```
Alaska Airlines
Airport          No. On-time    No. Delayed   Pct Delayed
los angeles          497            62           11.1%
phoenix              221            12            5.2%
san diego            212            20            8.6%
san francisco        503           102           16.9%
seattle             1841           305           14.1%
-----------------------------------------------------
total               3274           501           13.3%

America West
Airport          No. On-time    No. Delayed   Pct Delayed
los angeles          694           117           14.4%
phoenix             4840           415            7.9%
san diego            383            65           14.5%
san francisco        320           129           28.7%
seattle              201            61           23.3%
-----------------------------------------------------
total               6438           787           10.9%
```

Another Example: Test Scores

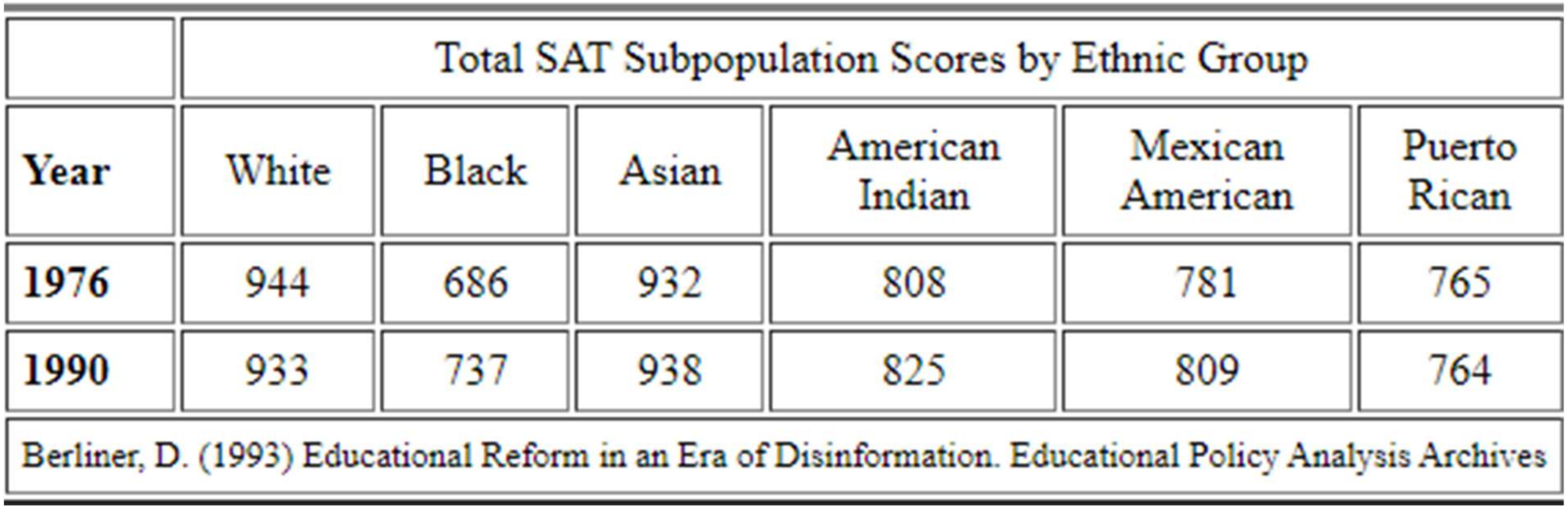


```
----------+-------------------------------------
    group |        sum            n    mean(1980)
----------+-------------------------------------
        1 |  6500000.00       10,000       650.00
        2 |   640000.00        1,000       640.00
        3 |    60000.00          100       600.00
          |
    Total |  7200000.00       11,100       648.65
----------+-------------------------------------
```



```
---------+----------------------------------------
   group |        sum               n   mean(1990)
---------+----------------------------------------
       1 |  6550000.00       10,000       655.00
       2 |  6450000.00       10,000       645.00
       3 |  1830000.00        3,000       610.00
         |
   Total | 14830000.00       23,000       644.78
---------+----------------------------------------
```

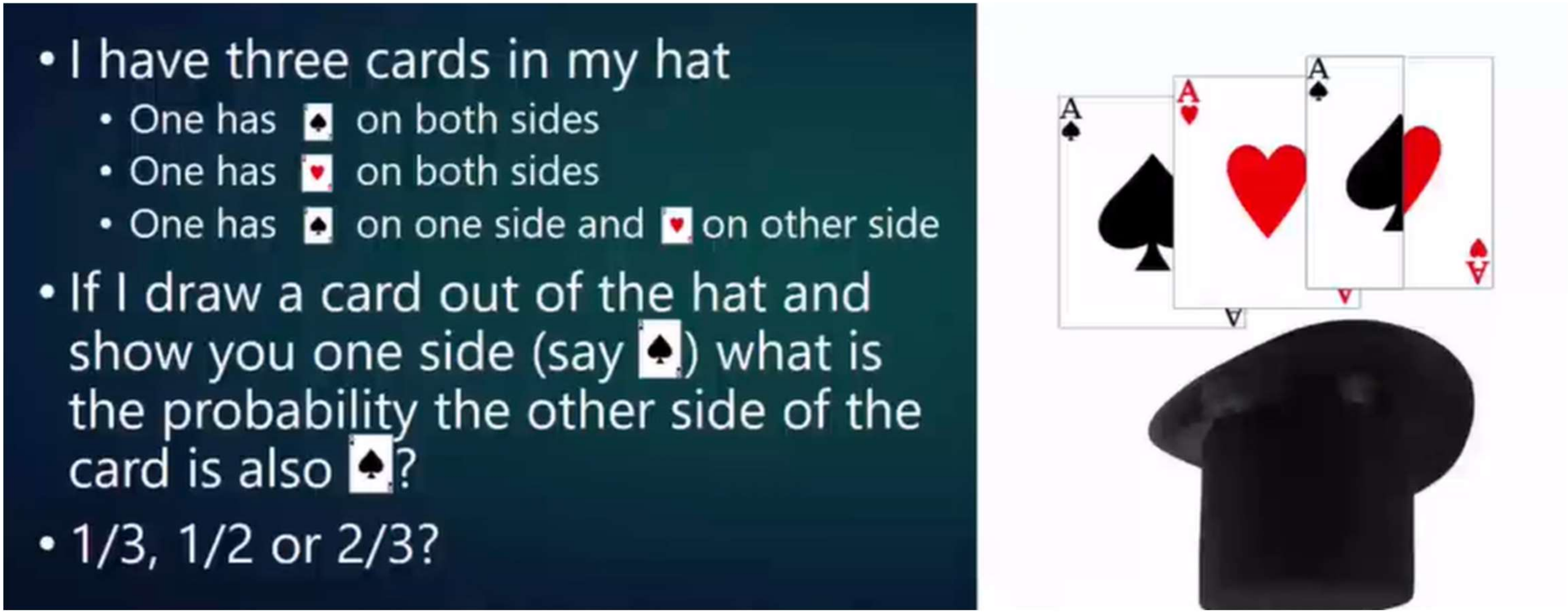
Another Example



	Total SAT Subpopulation Scores by Ethnic Group					
Year	White	Black	Asian	American Indian	Mexican American	Puerto Rican
1976	944	686	932	808	781	765
1990	933	737	938	825	809	764

Berliner, D. (1993) Educational Reform in an Era of Disinformation. Educational Policy Analysis Archives

Issue 6: Bertrand's Box Paradox/Monty Hall Paradox



- I have three cards in my hat
  - One has ♠ on both sides
  - One has ♥ on both sides
  - One has ♠ on one side and ♥ on other side
- If I draw a card out of the hat and show you one side (say ♠) what is the probability the other side of the card is also ♠?
- 1/3, 1/2 or 2/3?

Why is all of this hard for humans?

- we are poor at conditional probabilities
  - the likelihood of events
- we are fooled by randomness
  - misinterpret trends
  - make generalizations from small number of occurrences
- we are susceptible to fallacies

How can this be solved?

- Have staff that understand these issues
- Continue learning/get some training
- Recognize the situation
  - mistaking correlation/causation. Understand WHEN you need to prove causation and HOW to do it
  - regression to the mean
  - understand how to call a trend
- Use common sense and question all results