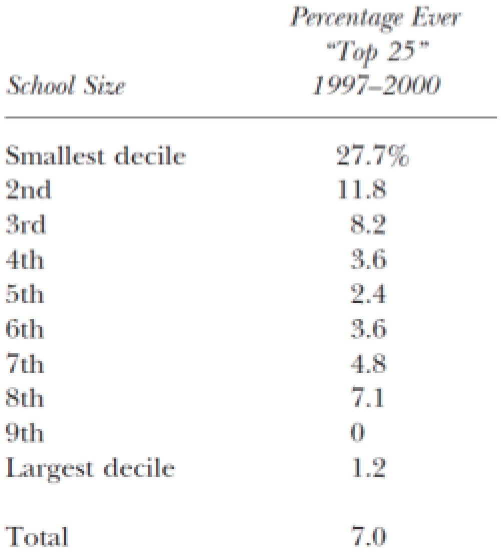
Pitfalls in Predictive Analytics

Thinking Correlation is Causation

Gates Foundation and Small Schools



School Size	*Percentage Ever "Top 25" 1997–2000*
Smallest decile	27.7%
2nd	11.8
3rd	8.2
4th	3.6
5th	2.4
6th	3.6
7th	4.8
8th	7.1
9th	0
Largest decile	1.2
Total	7.0

"The lead author concluded, 'I'm afraid we have done a terrible disservice to kids.' "

Source: http://assets.press.princeton.edu/chapters/s8863.pdf

p-hacking

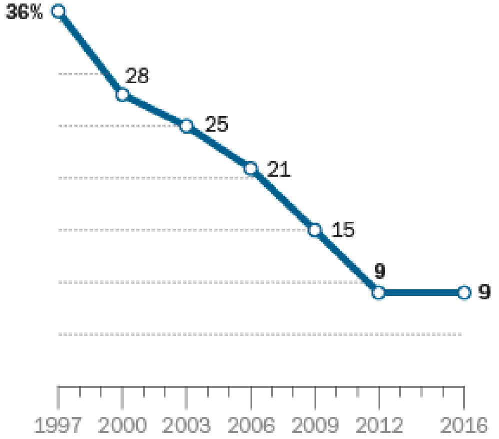
Canonical Source: https://xkcd.com/882/

Non-representative Sample

Here’s the rub: no sample is representative

Despite overall decrease, response rates have stabilized over past four years

Response rate by year (%)



Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates typical for surveys conducted in each year.
Source: Pew Research Center surveys conducted 1997-2016.
“What Low Response Rates Mean for Telephone Surveys”

PEW RESEARCH CENTER

Is this really god news?

Data Leakage

Sounds unpleasant!

“It is difficult to make predictions, especially about the future.” - A Danish Parliamentarian, Evidently (https://quoteinvestigator.com/2013/10/20/no-predict/)

But, it’s easy to predict the past! Make sure that analyses don’t use data from the future to predict the past (happens more often than you’d think)

Overfitting

When a model performs well on training data but poorly on new observations.

Non-representative Training Data

Using data for training that doesn't reflect the real-world circumstances of model application.