Data Dredging [1]

Siddharth Kalla [2]

Data dredging, also called as data snooping or data fishing, refers to the practice of misusing data mining techniques to show misleading scientific 'research'.

Data dredging is usually followed by the researcher who wants to try and 'prove' a point of view that might not hold or might not be shown in by the actual data. There are a number of reasons for data snooping and it is a matter of grave concern as it uses statistical principles for the purpose of drawing misleading and false conclusions.

The Way Data Snooping Works is as Follows:

Suppose there is a given data set and there are a huge number of hypotheses [3] about this data set [4]. If the data is totally random, then say all the hypotheses are actually false. However, owing to the sheer number of hypotheses on a limited data set, it is possible to see some very highly correlated [5] data that are statistically significant [6]. In such cases, data dredging is said to have taken place. Industries that are heavy on data mining are many times involved in data dredging [7]

For example, a drug company might spend millions of dollars on a drug but it may not show the kind of results that were initially expected. However, it needs to market the drug in order to make profits from it. Therefore the company uses data snooping to project claims that are not actually true, even though the data confirms the claim. This is done by taking a representative sample [8] and collecting huge number of parameters related to the test subjects, so that the drug can be claimed and correlated [5] to the problem in some form or the other.

Data fishing can also be done by narrowing down the sample size [9] to include those results that bear out the intended hypothesis. Thus the drug might be tested on 1000 patients and the results might not show a statistically significant positive result for a given problem. However, by narrowing down the study to 500 people and using a selection bias [10] towards those who showed favorable results by using the drug, the company can claim something that is not actually true.

If there is no effect between variables and your confidence level [11] is at .05 (5%), 1 of 20 tests will show that there is an effect even though this is not true, due to random error [12].

However, most data dredging is intentional. Many times, researchers are simply misled by the apparent correlations that they see. This happens most frequently when the researchers themselves are not sure what exactly they are looking for. Therefore it is important to form a hypothesis before starting and conducting the experiment [13] in order to prevent any

accidental cases of data dredging.

If not, the researchers might stumble upon some correlation that doesn't actually exist but shows strongly in their data. Thus researchers working in data mining need to be aware of this as it can be a serious mislead and divert valuable resources to some claims that are not really true.

Links
[1] https://explorable.com/data-dredging
[2] https://explorable.com/users/siddharth
[3] https://explorable.com/research-hypothesis
[4] https://explorable.com/statistical-data-sets
[5] https://explorable.com/statistical-correlation
[6] https://explorable.com/statistically-significant-results
[7] http://en.wikipedia.org/wiki/Data-snooping_bias
[8] https://explorable.com/sample-group
[9] https://explorable.com/sample-size
[10] https://explorable.com/sampling-error
[11] https://explorable.com/statistics-confidence-interval
[12] https://explorable.com/random-error
[13] https://explorable.com/conducting-an-experiment