Miao Cai[[1]](#footnote-20)

3/15/2019

Apart from the three-V standards (volume, velocity, and variety), a more down-to-earth definition of big data is that the analytical tools or environment have to be changed or improved only because the size of the data is beyond the current the analytical tools or environment.

Inferential modeling focuses on the inference and estimation about the parameters in the population. In comparison, predictive modeling aims at predicting the future and focuses on the accuracy of prediction.

Two clarifying examples on inferential and predictive modeling:

* Inferential: What are the significant factors associated with today’s weather?
* Predictive: what is the weather like tomorrow based on the weather of the previous week?

In my opinion, machine learning and traditional statistical inference are two streams of data science, and there are no superior in the current era.

The choice of machine learning or traditional statistics highly depends on the field and the types of question. In the field of public health, traditional ‘data models’ are viewed as the de facto way of statistics since researchers care more about odds ratio, intuition, and interpretability. By contrast, modern machine learning including consulting and recommending systems focus on the accuracy of prediction, therefore the ‘algorithm modeling’ is the de facto way of doing statistics.

Data leakage is adding accidentally adding variables that are strongly associated with the outcome variable while these variables are impossible to obtain in real-life application (Wiens and Shenoy [2017](#ref-wiens2017machine)).

Below is some symptoms of data leakage:

* The predictive accuracy seems to be too high but it becomes very low when used in real-world situation,
* A strong association between a few of the predictors and the outcome variable
* Large sample size: real world evidence (RWE) are often generated by administrative database
* More external validity:
* Relatively low costs:
* Randomisation to eliminate selection bias
* One treatment is directly compared to another to establish superiority.

Wiens, Jenna, and Erica S Shenoy. 2017. “Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology.” *Clinical Infectious Diseases* 66 (1): 149–53.

1. Department of Epidemiology and Biostatistics, College for Public Health and Social Justice, Saint Louis University [↑](#footnote-ref-20)