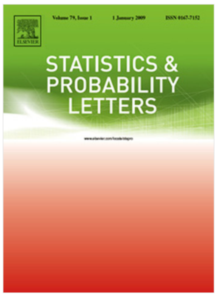

Big data and biostatistics: The death of the asymptotic Valhalla

Ernst C. Wit

PII: S0167-7152(18)30084-1
DOI: https://doi.org/10.1016/j.spl.2018.02.039
Reference: STAPRO 8161

To appear in: *Statistics and Probability Letters*

Please cite this article as: Wit E.C., Big data and biostatistics: The death of the asymptotic Valhalla. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.039

Big data and Biostatistics: the death of the asymptotic Valhalla

Ernst C. Wit

Johann Bernoulli Institute, Nijenborgh 9, 9747 AG Groningen, NL

Abstract

Despite the ubiquity of Big Data in the modern scientific discourse, most references describe storage and query considerations and rarely full-flexed analyses. In this article, we propose another definition with particular relevance to biometrics. We argue that the complexity of the generating measure of biological process means that the model complexity of any statistical model will have to be smaller. Only, when the model is used for prediction can we have any hope that the number of available features reasonably outnumbers the desired complexity of the model.

Keywords:
biostatistics, big data, high-dimensional inference, model complexity

1. Introduction

Biostatistics stood at the craddle of the modern era of statistics about a century ago. It is often said that the work or Pearson and Fisher revolutionized statistical thinking and put it on a firm mathematical foundation, however it is also true that their work had a general biometric flavour. A century later, we ask ourselves how biostatistics is shaping up in response to the next challenge: Big Data.

Although from recent discussions it is clear that there is no universally accepted definition of Big Data, we will still attempt to define what we will understand by Big Data in a biometric setting, especially because this is different from more computer science definitions. In particular, we focus on two separate interpretations. The first harks back to the high-dimensional interpretation of big data with the aim of doing prediction. Twenty years ago microarrays revolutionized genomic screening. Large numbers of gene