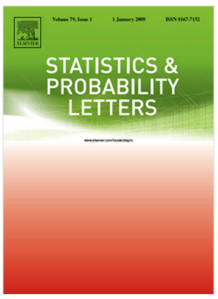
Accepted Manuscript



A practical guide to big data

Ekaterina Smirnova, Andrada Ivanescu, Jiawei Bai, Ciprian M. Crainiceanu

PII: S0167-7152(18)30059-2
DOI: https://doi.org/10.1016/j.spl.2018.02.014
Reference: STAPRO 8136

To appear in: *Statistics and Probability Letters*

Please cite this article as: Smirnova E., Ivanescu A., Bai J., Crainiceanu C.M., A practical guide to big data. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.014

A practical guide to big data

Ekaterina Smirnova* Andrada Ivanescu† Jiawei Bai‡ Ciprian M. Crainiceanu§

November 9, 2017

Abstract

Big Data is increasingly prevalent in science and data analysis. We provide a short tutorial for adapting to these changes and making the necessary adjustments to the academic culture to keep Biostatistics truly impactful in scientific research.

1 Introduction

Big Data has been analyzed for a long time. Indeed, in a 1938 landmark paper, Raymond Pearl [9] used data on 6,813 men (2,094 non-smokers, 2,814 moderate smokers, and 1,905 heavy smokers) to show that tobacco smoking was "statistically associated with the impairment of life duration, and the amount of this impairment increased as the habitual amount of smoking increased". It took until January 11, 1964, for Luther L. Terry, M.D., Surgeon General of the U.S. Public to officially acknowledge that "cigarette smoking was cause of lung cancer and laryngeal cancer in men, a probable cause of lung cancer in women, and the most important cause of chronic bronchitis". In 1982 Allan Gittelsohn [5] published results on the distribution of underlying causes of death in the US using 21 million death records from 1968 to 1978. Currently, Biostatisticians routinely work with hundreds of Terabytes of data from genomics, brain imaging, and wearable sensors. Thus, one could think that the "Big Data" phenomenon is not new and is just a clever rebranding of the analysis of ever larger datasets generated by increasingly sophisticated new technologies. However, this would not explain the explosion in popularity of Big Data. What could explain it is the large amount of money it can generate when analyzing who will click a "like" button, what advertising to provide to a net surfer, or what smart phone to recommend to an online shopper. The sheer sexiness of money makes Big Data cool. We believe that this excitement should be captured, embraced, and directed to solving important societal problems. In this short paper we try to provide a practical guide to doing that, mention a few tautologies, and identify a few arbitrage opportunities. The recipe is simple, though the implementation is difficult because it requires actual work.

*Assistant Professor, Department of Mathematical Sciences, University of Montana, 32 Campus Dr, Missoula, MT 59812. E-mail: ekaterina.smirnova@mso.umt.edu

†Assistant Professor, Department of Mathematical Sciences, Montclair University, 1 Normal Avenue Montclair, NJ 07043; E-mail: ivanescua@montclair.edu

‡Assistant Scientist, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. Baltimore, MD 21205 USA. E-mail: jiawei.bai@jhu.edu

§Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. Baltimore, MD 21205 USA. E-mail: ccraini1@jhu.edu