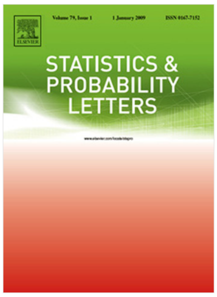

STATISTICS & PROBABILITY LETTERS

Statistics for big data: A perspective

Peter Bühlmann, Sara van de Geer

PII: S0167-7152(18)30061-0
DOI: https://doi.org/10.1016/j.spl.2018.02.016
Reference: STAPRO 8138

To appear in: *Statistics and Probability Letters*

Please cite this article as: Bühlmann P., van de Geer S., Statistics for big data: A perspective. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.016

Statistics for Big Data: A Perspective

Peter Bühlmann and Sara van de Geer

Seminar for Statistics, ETH Zürich

Abstract

We look at the role of statistics in data science. Two statisticians, two views. Besides the need of developing appropriate concepts, methodology and algorithms, the first one makes in Section 3 a case for validation and carefully designed simulation studies, while the second one writes in Section 4 that a mathematical underpinning of methods is fundamental. Both views converge to the same point: there should be more room for publishing negative findings.

Keywords: Heterogeneity, Large-scale data, Lasso, Learning theory, Mathematical theory, Negative results, Replicability, Reproducibility, Validation

2010 MSC: 62-01 (primary), 68-01 (secondary)

1. A short introduction

"Big Data" is perhaps not a well-defined terminology. Wikipedia (`https://en.wikipedia.org/wiki/Big_da` states the following: "Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time."

Computation, open software, open data and reproducibility. The computational issue mentioned above is certainly a relevant one. We publicize open source soft-

Email address: `buhlmann,geer @stat.math.ethz.ch` (Peter Bühlmann and Sara van de Geer)