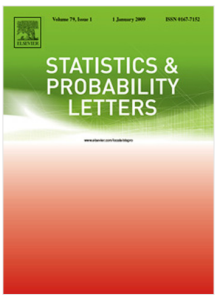
Accepted Manuscript



Statistics in the big data era: Failures of the machine

David B. Dunson

PII: S0167-7152(18)30073-7
DOI: https://doi.org/10.1016/j.spl.2018.02.028
Reference: STAPRO 8150

To appear in: *Statistics and Probability Letters*

Please cite this article as: Dunson D.B., Statistics in the big data era: Failures of the machine. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.028

Statistics in the big data era: Failures of the machine

David B. Dunson

Department of Statistical Science, Duke University, dunson@duke.edu

There is vast interest in automated methods for complex data analysis. However, there is a lack of consideration of (1) interpretability, (2) uncertainty quantification, (3) applications with limited training data, and (4) selection bias. Statistical methods can achieve (1)-(4) with a change in focus.

Key Words: Deep learning; High-dimensional data; Large p, small n; Machine learning; Scientific inference; Selection bias; Uncertainty quantification

1. Introduction

1.1 Different cultures

The culture and ways in which the statistical community thinks of analyzing and interpreting data have been rapidly evolving in recent years, with the machine learning and signal processing communities having a fundamental impact on the rate and direction of this evolution. To set the stage for this discussion article, it is helpful to first comment on the culture and background of the machine learning and statistical communities. These comments are meant to give a "cartoon" of a complex reality, with this cartoon helpful as a starting point for discussion.

Machine learning (ML) community: tends to have its roots in engineering, computer science, and to a certain extent neuroscience – growing out of artificial intelligence (AI). The main publication outlets tend to be peer-reviewed conference proceedings, such as *Neural Information Processing Systems (NIPS)*, and the style of research is very fast paced, trendy, and driven by performance metrics in prediction and related tasks. One measure of "trendiness" is the fact that there is a strong auto-correlation in the main focus areas that are represented in the papers accepted to NIPS and other top conferences. For example, in the past several years much of the focus has been on deep neural network methods. The ML community also has a tendency towards marketing and salesmanship, posting talks and papers on social media and attempting to sell their ideas to the broader public. This feature of the research seems to reflect a desire or tendency to want to monetize the algorithms in the near term, perhaps leading to a focus on industry problems over scientific problems, where the road to monetization is often much longer and less assured. ML marketing has been quite successful in recent years, and there is abundant interest and discussion in the general public about ML/AI, along with increasing success in start-ups and industrial sector high paying jobs partly fueled by the hype.

Statistical (Stats) community: made up predominantly of researchers who received their initial degree(s) in mathematics followed by graduate training in statistics. The main publication outlets are peer-reviewed journals, most of which have a long drawn out review process, and the style of research tends to be careful, slower paced, intellectual as