Veracity of Big Data

Machine Learning and Other Approaches to Verifying Truthfulness

Vishnu Pendyala

Apress®

Veracity of Big Data

Vishnu Pendyala
San Jose, California, USA

ISBN-13 (pbk): 978-1-4842-3632-1 ISBN-13 (electronic): 978-1-4842-3633-8
https://doi.org/10.1007/978-1-4842-3633-8

Library of Congress Control Number: 2018945464

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Celestin Suresh John
Development Editor: Laura Berendson
Coordinating Editor: Divya Modi

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit http://www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/978-1-4842-3632-1. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

I dedicate this book to the loving memory of my father,
Pendyala Srinivasa Rao.

Table of Contents

About the Author

Vishnu Pendyala is a Senior Member of IEEE and of the Computer Society of India (CSI), with over two decades of software experience with industry leaders such as Cisco, Synopsys, Informix (now IBM), and Electronics Corporation of India Limited. He is on the executive council of CSI, Special Interest Group on Big Data Analytics, and is the founding editor of its flagship publication, *Visleshana*. He recently taught a short-term course on "Big Data Analytics for Humanitarian Causes," which was sponsored by the Ministry of Human Resources, Government of India under the GIAN scheme; and delivered multiple keynotes in IEEE-sponsored international conferences. Vishnu has been living and working in the Silicon Valley for over two decades. More about him at: https://www.linkedin.com/in/pendyala.

Acknowledgments

The story of this book starts with Celestin from Apress contacting me to write a book on a trending topic. My first thanks therefore go to Celestin. Thanks to the entire editorial team for pulling it all together with me – it turned out to be a much more extensive exercise than I expected, and the role you played greatly helped in the process.

Special thanks to the technical reviewer, Oystein, who provided excellent feedback and encouragement. In one of the emails, he wrote, "Just for your information, I learnt about CUSUM from your 4th chapter and tested it out for the audio-based motion detector that is running on our …. From metrics I could see that it worked really well, significantly better than the exponentially weighted moving average (EWMA) method, and it is now the default change detection algorithm for the motion detector in all our products!"

Introduction

Topics on Big Data are growing rapidly. From the first 3 V's that originally characterized Big Data, the industry now has identified 42 V's associated with Big Data. The list of how we characterize Big Data and what we can do with it will only grow with time. Veracity is often referred to as the 4th V of Big Data. The fact that it is the first V after the notion of Big Data emerged indicates how significant the topic of Veracity is to the evolution of Big Data. Indeed, the quality of data is fundamental to its use. We may build many advanced tools to harness Big Data, but if the quality of the data is not to the mark, the applications will not be of much use. Veracity is a foundation block of data and, in fact, the human civilization.

In spite of its significance striking at the roots of Big Data, the topic of its veracity has not been initiated sufficiently. A topic really starts its evolution when there is a printed book on it. Research papers and articles, the rigor in their process notwithstanding, can only help bring attention to a topic. But the study of a topic at an industrial scale starts when there is a book on it. It is sincerely hoped that this book initiates such a study on the topic of Veracity of Big Data.

The chapters cover topics that are important not only to the veracity of Big Data but to many other areas. The topics are introduced in such a way that anyone with interest in math and technology can understand, without needing the extensive background that some other books on the same topics often require. The matter for this book evolved from the various lectures, keynotes, and other invited talks that the author delivered over the last few years, so they are proven to be interesting and insightful to a live audience.

The book is particularly useful to managers and practitioners in the industry who want to get quick insights into the latest technology. The book has made its impact in the industry even before it was released, as the technical reviewer acknowledged in one of his emails that one of the techniques explained in this book that he reviewed, turned out to be better than the one they were using, so much so that the new technique from the book became the default for all the products in the product line.

The book can be used to introduce not only Veracity of Big Data, but also topics in Machine Learning; Formal Methods; Statistics; and the revolutionary technology, Blockchain, all in one book. It can serve as a suggested reading for graduate and undergraduate courses. The exercises at the end of each chapter are hoped to provoke critical thinking and stoke curiosity. The book can also be used by researchers, who can evaluate the applicability of the novel techniques presented in the book, for their own research. The use of some of the techniques described in the book for the problem of veracity are based on the author's own research.

The chapters can be read in any order, although there are some backward and forward references. Chapter 3, on the approaches to the Veracity of Big Data, gives a good overview of the remaining textbook and is a must for someone short on time. Blockchain is being touted as the ultimate truth machine that can solve a number of trust and veracity problems in the real world. The last chapter briefly introduces the topic of Blockchain as a trend that deserves to be keenly watched. It is sincerely hoped that the book will initiate a thorough study of the topic of veracity in the long run. Happy reading!