Hui Lin[1] *and Ming Li*

Introduction to Data Science

[1] http://scientistcafe.com

Contents

List of Tables

List of Figures

```r
# Sys.setenv(TZ="UTC")
options(formatR.indent = 2, width = 55)
#bookdown::render_book("index.Rmd", "bookdown::gitbook")
#bookdown::render_book("index.Rmd", "bookdown::pdf_book")
```

Copyright Statement

About the Authors

Hui Lin is currently Data Scientist at DuPont Pioneer. She is a leader within DuPont at applying advanced data science to enhance Marketing

[1] https://creativecommons.org/licenses/by-nc-sa/3.0/us/

[2] https://github.com/happyrabbit/IntroDataScience/blob/master/CONDUCT.md

[3] https://creativecommons.org/licenses/by-nc-sa/3.0/us/

[4] https://github.com/happyrabbit/IntroDataScience/issues

and Sales effectiveness. She has been providing statistical leadership for a broad range of predictive analytics and market research analysis since 2013. She is the co-founder of Central Iowa R User Group, blogger of scientistcafe.com and program Chair of Statistics in Marketing Section of ASA for 2018. She enjoys making analytics accessible to a broad audience and teaches tutorials and workshops for practitioners on data science.

She holds MS and Ph.D. in statistics from Iowa State University, BS in mathematical statistics from Beijing Normal University.

Ming Li is currently a Sr. Data Scientist at Amazon. He was Data Scientist at Wal-Mart and an Adjunct Faculty of Department and Marketing and Business Analytics in TAMU – Commerce. He is also the Chair of Quality & Productivity Section of ASA for 2016. He was a Statistical Leader at General Electric Global Research Center and Research Statistician at SAS Institute. He obtained his Ph.D. in Statistics from Iowa State University at 2010. With deep statistics background and a few years' experience in data science, he has trained and mentored numerous junior data scientist with different backgrounds such as statistics, computer science, and business analytics.

Acknowledgements

We want to give special thanks to Alex Shum and David Body for their editing and comments on the sections of this book.

0.1 The art of data science

Data science and data scientist have become buzz words. Allow me to reiterate what you may have already heard a million times in the media: **data scientists are in demand and demand continues to grow**. A study by the McKinsey Global Institute concludes,

> “a shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.).”

You may expect that statisticians and graduate students from traditional statistics departments are great data scientist candidates. But the situation is that the majority of current data scientists do not have a statistical background. As David Donoho pointed out:

> “statistics is being marginalized here; the implicit message is that statistics is a part of what goes on in data science but not a very big part.” (from “50 years of Data Science[5]”).

What is wrong? The activities that preoccupied statistics over centuries are now in the limelight, but those activities are claimed to belong to a new discipline and are practiced by professionals from various backgrounds. Various professional statistics organizations are reacting to this confusing situation. (Page 5-7, “50 Years of Data Science”) From those discussions, Donoho summarizes the main recurring “Memes” about data sciences:

1. The ‘Big Data’ Meme
2. The ‘Skills’ Meme
3. The ‘Jobs’ Meme

The first two are linked together which leads to statisticians’ current position on data science. We assume everyone has heard the 3V (volume, variety and velocity) definition of big data. The media hasn’t taken a minute break from touting “big” data. Data science trainees now need the skills to cope with such big data sets. What are those skills? You may hear about:

[5] http://pages.cs.wisc.edu/~anhai/courses/784-fall15/50YearsDataScience.pdf

Hadoop, system using Map/Reduce to process large data sets distributed across a cluster of computers. The new skills are for dealing with organizational artifacts of large-scale cluster computing but not for better solving the real problem. A lot of data on its own is worthless. It isn't the size of the data that's important. It's what you do with it. The big data skills that so many are touting today are not skills for better solving the real problem of inference from data.

Some media think they sense the trends in hiring and government funding. We are transiting to universal connectivity with a deluge of data filling telecom servers. But these facts don't immediately create a science. The statisticians have been laying the groundwork of data science for at least 50 years. Today's data science is an enlargement of traditional academic statistics rather than a brand new discipline.

0.1.1 What is data science?

This question is not new. When you tell people "I am a data scientist". "Ah, data scientist!" Yes, who doesn't know that data scientist is the sexist job in 21th century? If they ask further what is data science and what exactly do data scientists do, it may effectively kill the conversation.

Data Science doesn't come out of the blue. Its predecessor is data analysis. Back in 1962, John Tukey wrote in "The Future of Data Analysis":

> For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. … All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

It deeply shocked his academic readers. Aren't you supposed to present something mathematically precise, such as definitions, theorems and proofs? If we use one sentence to summarize what John said, it is:

> data analysis is more than mathematics.

0.2 References

Bibliography