Interfacing R with web technologies for interactive statistical graphics and computing with data

by

Carson Sievert

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Heike Hofmann, Major Professor

Dianne Cook

Jarad Niemi

Ulrike Genschel

Grayson Calhoun

Iowa State University

Ames, Iowa

2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This thesis would not be possible without many people. First and foremost, special thanks to my major professor Heike Hofmann. I would not made it to this point without such a warm and friendly mentor who was often more confident in my abilities than I was of my own. Thank you for always supporting me no matter how many things I had going on to distract me from research. I aspire to inherit the same empathy and support that you show to your students on a daily basis.

Another special thanks goes to Di Cook. In the fourth year of my PhD, I was experiencing burnout (and growing tired of Ames), when Di took me to lunch, told me she was transferring to Monash University in Australia, and invited me to join her. Of course, I said yes, and when I arrived, I immediately felt welcomed and a part of the group – all thanks to Di. She strategically assigned me to assist her students with their thesis projects, run R workshops for the university, and most importantly, work on my tennis game. Through this "work" I met so many amazing people and had many memorable experiences. Those 6 months gave me a new perspective on life in general and I am forever grateful for being blessed enough to take the opportunity.

Thank you to many of Heike and Di's former students who came before me (just to name a few: Hadley Wickham, Michael Lawrence, Yihui Xie, Xiaoyue Cheng, Barrett Schloerke, Susan VanderPlas). Your work has not only inspired and enabled my work, but it has also enabled an entire community of people working with data to do amazing things. Without this strong history and community at Iowa State, I would not have had the courage or the vision to follow such a "non-traditional" research path. I hope the University continues to value this type of work as it teaches students skills that are in high demand and generally improves the way data-driven research is performed.

Thank you to all my collaborators, especially Toby Dylan Hocking. Toby and Susan Van-

derPlas laid the initial framework for **animint** – which I first worked on as a Google Summer of Code student under Toby's guidance. Toby later went on write the initial version of the `ggplotly()` function in **plotly**, borrowing a lot of ideas from **animint**. As Toby became busy with other things, he introduced me to the plotly team, and eventually handed over the reigns on the project, which has helped to financially support the last year or so of grad school.

Thank you also to the plotly team, and in particular, the software engineers who work on the open source project plotly.js. My work has benefited greatly from your responsiveness to my questions, feature requests, and bug reports. I have a great amount of respect for the work that you do, and I hope this project keeps improving at its current break neck pace.

Finally, thank you to my family for their encouragement and keeping me grounded throughout this experience. Thank you to my father for the initial encouragement to pursue a PhD, conversations surrounding work-life balance, and also pushing me to "graduate before I'm 40". Thank you to my mother for her unconditional love, endless care, and predenting to understand what I do for a living. Thank you to my brothers for providing me with shelter, beer, and Twins tickets. Thank you all for your willingness to drop everything to help me at any given moment. I can't say I've done the same, as I have been selfish with my time in pursuing a PhD, but I hope to change that after graduation.

ABSTRACT

The following describes a collection of software interfaces for data acquisiton and visualization. All of these interfaces are freely available as extension packages to the R language and leverage web technologies to achieve accessible, portable, and reproducible workflows. The majority of this work (LDAvis, animint, and plotly) focuses on interactive visualization. These interfaces fall roughly into two categories: (1) domain-specific (LDAvis) and (2) general purpose tools for interactive data visualization (animint and plotly). More specifially, the LDAvis package produces an interactive visualization to aid interpretation of Latent Dirichlet Allocation (LDA) model output. The animint and plotly packages are more general, and build upon principles from the grammar of graphics, but extend those principles in slightly different ways to enable interactivity, such as animation and brushing a scatterplot matrix.

CHAPTER 1. OVERVIEW

This is the opening paragraph to my thesis which explains in general terms the concepts and hypothesis which will be used in my thesis.

With more general information given here than really necessary.

1.1 Introduction

Here initial concepts and conditions are explained and several hypothesis are mentioned in brief.

1.1.1 Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

1.1.1.1 Parts of the hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

1.1.2 Second Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

1.1.2.1 Parts of the second hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

1.2 Criteria Review

Here certain criteria are explained thus eventually leading to a foregone conclusion.

CHAPTER 2. REVIEW OF LITERATURE

This is the opening paragraph to my thesis which explains in general terms the concepts and hypothesis which will be used in my thesis.

With more general information given here than really necessary.

2.1 Introduction

Here initial concepts and conditions are explained and several hypothesis are mentioned in brief.

Allen (1984), Bruner (1960) and Cox (1974) did the initial work in this area. But in Struss' work [Struss (1996)] the definitive model is seen.

2.1.1 Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

2.1.1.1 Parts of the hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

2.1.2 Second Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

2.1.2.1 Parts of the second hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

2.2 Criteria Review

Here certain criteria are explained thus eventually leading to a foregone conclusion.

CHAPTER 3. METHODS AND PROCEDURES

This is the opening paragraph to my thesis which explains in general terms the concepts and hypothesis which will be used in my thesis.

With more general information given here than really necessary.

3.1 Introduction

Here initial concepts and conditions are explained and several hypothesis are mentioned in brief.

As can be seen in Table 3.1 it is truly obvious what I am saying is true.

Table 3.1 This table shows a standard empty table

3.1.1 Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

This can also be seen in Figure 3.1 that the rest is obvious.

Figure 3.1 This table shows a standard empty figure

3.1.1.1 Parts of the hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

3.1.2 Second Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

3.1.2.1 Parts of the second hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

3.2 Criteria Review

Here certain criteria are explained thus eventually leading to a foregone conclusion as can be seen in Table 3.2.

Table 3.2 This table shows a standard empty table with a limited captionwidth

CHAPTER 4. RESULTS

This is the opening paragraph to my thesis which explains in general terms the concepts and hypothesis which will be used in my thesis.

With more general information given here than really necessary.

4.1 Introduction

Here initial concepts and conditions are explained and several hypothesis are mentioned in brief.

Of course, data on this as seen in Table 4.1 is few and far between.

Table 4.1 Moon Data

Element	Control	Experimental
Moon Rings	1.23	3.38
Moon Tides	2.26	3.12
Moon Walk	3.33	9.29

4.1.1 Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

Or graphically as seen in Figure 4.1 it is certain that my hypothesis is true.

4.1.1.1 Parts of the hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

Figure 4.1 Durham Centre

4.1.2 Second Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

4.1.2.1 Parts of the second hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

4.2 Criteria Review

Here certain criteria are explained thus eventually leading to a foregone conclusion.

CHAPTER 5. SUMMARY AND DISCUSSION

This is the opening paragraph to my thesis which explains in general terms the concepts and hypothesis which will be used in my thesis.

With more general information given here than really necessary.

5.1 Introduction

Here initial concepts and conditions are explained and several hypothesis are mentioned in brief.

Or graphically as seen in Figure 5.1 it is certain that my hypothesis is true.

5.1.1 Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

As can be seen in Table 5.1 it is truly obvious what I am saying is true.

5.1.1.1 Parts of the hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

5.1.2 Second Hypothesis

Here one particular hypothesis is explained in depth and is examined in the light of current literature.

Table 5.1 This table shows almost nothing but is a sideways table and takes up a whole page by itself

Element	**Control**	**Experimental**
Moon Rings	1.23	3.38
Moon Tides	2.26	3.12
Moon Walk	3.33	9.29

5.1.2.1 Parts of the second hypothesis

Here one particular part of the hypothesis that is currently being explained is examined and particular elements of that part are given careful scrutiny.

5.2 Criteria Review

Here certain criteria are explained thus eventually leading to a foregone conclusion.

Figure 5.1 Durham Centre— Another View

APPENDIX A. ADDITIONAL MATERIAL

This is now the same as any other chapter except that all sectioning levels below the chapter level must begin with the *-form of a sectioning command.

More stuff

Supplemental material.

APPENDIX B. STATISTICAL RESULTS

This is now the same as any other chapter except that all sectioning levels below the chapter level must begin with the *-form of a sectioning command.

Supplemental Statistics

More stuff.

BIBLIOGRAPHY

Allen, B. S. (1984). System-assigned learning strategies and cbi. *Journal of Instructional Computing Research*, 1(1):3–18.

Bruner, J. (1960). *The process of education*. Random House, New York.

Cox, S. R. (1974). Computer-assisted instruction and student performance in macroeconomic principles. *The Journal of Economic Education*, 6(1):29–37.

Struss, J. A. (1996). An investigation of the sequence of utilizing a simulation in an introductory programming course. Master's thesis, Iowa State University.