# Data analysis tools for statistical non-experts

Eunice Jun

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2023

*Reading Committee:*
Jeffrey Heer, Co-Chair
René Just, Co-Chair
Tyler H. McCormick

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

University of Washington

## Abstract

Data analysis tools for statistical non-experts

Eunice Jun

Co-chairs of the Supervisory Committee:

Jerre D. Noe Endowed Professor Jeffrey Heer
Paul G. Allen School of Computer Science & Engineering

Associate Professor René Just
Paul G. Allen School of Computer Science & Engineering

Data analysis is critical to science, public policy, and business. Despite their importance, statistical analyses are difficult to author, especially for researchers with expertise outside of statistics. Existing statistical tools, prioritizing mathematical expressivity and computational control, are low-level while researchers' motivating questions and hypotheses are high-level. Researchers need to translate their questions and hypotheses into low-level statistical code in an error-prone process that involves grappling with their domain knowledge, statistics, and programming.

In this talk, I will introduce two tools that embody a new way of authoring analyses: Tea and Tisane. Researchers directly express their domain knowledge through higher level abstractions, and the tools will validate the data, select a statistical analysis, and implement it, all while educating analysts about why a statistical approach is valid. Tea helps analysts author statistical tests. Tea's key insight is that statistical test selection can be cast as a constraint satisfaction problem. Tisane enables analysts to author generalized linear models with or without mixed effects, which are difficult for even statistical experts to author. Using Tisane, analysts can express their conceptual models using a high-level domain specific language. Tisane translates these conceptual models into causal DAGs and engages analysts in a disambiguation process to arrive at an output statistical model. Real-world researchers have already used these tools to conduct analyses in

published research that push their own disciplines forward. I will also introduce "hypothesis formalization," a series of cognitive and operational steps analysts take to translate their research questions into statistical implementations. Hypothesis formalization retrospectively explains why Tea improves statistical testing and directly inspired the design of Tisane.

Tea and Tisane serve as platforms for further research into computational support for statistical analysis. This talk also exemplifies how combining human-computer interaction with other areas in and outside of computer science leads to software tools that impact real-world users.

# Acknowledgements

Insert acknowledgments here.

# DEDICATION

To ????

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Insert introduction here.

This chapter typically includes:

- a brief overview
- a background section
- a challenges section
- a section about your approach
- a section (or subsection in the approach section) giving a dissertation outline (a roadmap of the rest of the thesis)

Alternatively: you could choose to remove the background section and make a separate background chapter directly following this introduction chapter.

# Chapter 2

# Example Project

This chapter covers material from a single project. There are typically 3-4 of these chapters total (but any number is fine as long as your advisor is on board).

## 2.1 Updating things if you've copied and pasted

If you are copying and pasting material from one of your papers, then remember to:

- Remove the abstract and instead add a little overview of the chapter and how it ties in to the rest of the thesis. You should also mention the original paper's source like: "This chapter includes materials originally published in \citet{myownppr}"
- Make sure the formatting still works – this is single column now!
- Consider rephrasing conference-paper-style language:
  - Find every place you mention some variation of "in this paper" and say "in this chapter" instead.
  - Remove or rephrase the parts where you talk about "our main contributions".
  - Rephrase the language describing code and data releases.
- Replace the conclusion section with a summary section. Again, you should tie this chapter back to the main themes of the thesis.

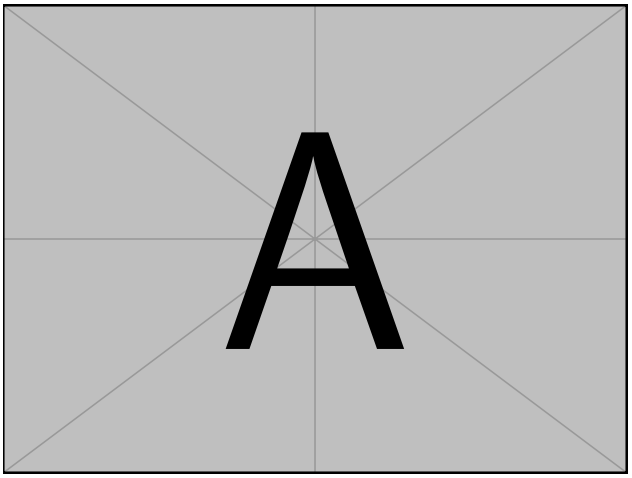


**Figure 2.2:** Full caption, make it as long as you want.

| Latex command | Citation |
|---|---|
| \citep example | [Grice, 1975] |
| \citet example | Grice [1975] |

**Table 2.1:** The style files that come included in this latex template use \citep and \citet.

## 2.2 Formatting tips

Here are a few notes about the layout and usage of this latex template:

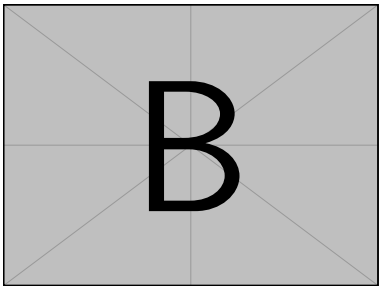


**Figure 2.1:** You can also make inline figures.

**Captions** In order for the captions from figures and tables to show up cleanly in the lists of figures and tables, you should use the caption command with the bracket argument to create a short title for the list of figures/tables, like:

```
\caption [shortened title]{full caption}
```

as in Figure 2.2 and Table 2.1.

**Citations** This latex was given to me with an old NAACL style file that uses the standard `\citep` and `\citet` commands as in Table 2.1. I don't think it supports `\cite` or `\newcite`. Feel free to add commands as needed for you.

# Chapter 3

# Add Project 2

## Chapter 4

# Add Project 3 ... or as many as needed

# Chapter 5

# Conclusion

Insert conclusion here. It seems common to include a future work section to the end of this chapter.

# Bibliography

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58.

# Chapter A

# Appendix One

## A.1 Appendix section 1

**Table A.1:** Table in the Appendix